INTRODUCTION to ANALYSIS

John Hutchinson

email: John.Hutchinson@anu.edu.au

with earlier revisions by

Richard J. Loy

Pure mathematics have one peculiar advantage, that they occasion no disputes among wrangling disputants, as in other branches of knowledge; and the reason is, because the definitions of the terms are premised, and everybody that reads a proposition has the same idea of every part of it. Hence it is easy to put an end to all mathematical controversies by shewing, either that our adversary has not stuck to his definitions, or has not laid down true premises, or else that he has drawn false conclusions from true principles; and in case we are able to do neither of these, we must acknowledge the truth of what he has proved ...

The mathematics, he [Isaac Barrow] observes, effectually exercise, not vainly delude, nor vexatiously torment, studious minds with obscure subtlities; but plainly demonstrate everything within their reach, draw certain conclusions, instruct by profitable rules, and unfold pleasant questions. These disciplines likewise enure and corroborate the mind to a constant diligence in study; they wholly deliver us from credulous simplicity; most strongly fortify us against the vanity of scepticism, effectually refrain us from a rash presumption, most easily incline us to a due assent, perfectly subject us to the government of right reason. While the mind is abstracted and elevated from sensible matter, distinctly views pure forms, conceives the beauty of ideas and investigates the harmony of proportion; the manners themselves are sensibly corrected and improved, the affections composed and rectified, the fancy calmed and settled, and the understanding raised and excited to more divine contemplations.

Encyclopædia Britannica [1771]

Philosophy is written in this grand book—I mean the universe—which stands continually open to our gaze, but it cannot be understood unless one first learns to comprehend the language and interpret the characters in which it is written. It is written in the language of mathematics, and its characters are triangles, circles, and other mathematical figures, without which it is humanly impossible to understand a single word of it; without these one is wandering about in a dark labyrinth.

Galileo Galilei *Il Saggiatore* [1623]

Mathematics is the queen of the sciences.

Carl Friedrich Gauss [1856]

Thus mathematics may be defined as the subject in which we never know what we are talking about, nor whether what we are saying is true.

Bertrand Russell *Recent Work on the Principles of Mathematics,* International Monthly, vol. 4 [1901]

Mathematics takes us still further from what is human, into the region of absolute necessity, to which not only the actual world, but every possible world, must conform.

Bertrand Russell *The Study of Mathematics* [1902]

Mathematics, rightly viewed, possesses not only truth, but supreme beauty—a beauty cold and austere, like that of a sculpture, without appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of perfection such as only the greatest art can show.

Bertrand Russell *The Study of Mathematics* [1902]

The study of mathematics is apt to commence in disappointment.... We are told that by its aid the stars are weighed and the billions of molecules in a drop of water are counted. Yet, like the ghost of Hamlet's father, this great science eludes the

efforts of our mental weapons to grasp it.

Alfred North Whitehead *An Introduction to Mathematics* [1911]

The science of pure mathematics, in its modern developments, may claim to be the most original creation of the human spirit.

Alfred North Whitehead *Science and the Modern World* [1925]

All the pictures which science now draws of nature and which alone seem capable of according with observational facts are mathematical pictures .... From the intrinsic evidence of his creation, the Great Architect of the Universe now begins to appear as a pure mathematician.

Sir James Hopwood Jeans *The Mysterious Universe* [1930]

A mathematician, like a painter or a poet, is a maker of patterns. If his patterns are more permanent than theirs, it is because they are made of ideas.

G.H. Hardy *A Mathematician's Apology* [1940]

The language of mathematics reveals itself unreasonably effective in the natural sciences..., a wonderful gift which we neither understand nor deserve. We should be grateful for it and hope that it will remain valid in future research and that it will extend, for better or for worse, to our pleasure even though perhaps to our bafflement, to wide branches of learning.

Eugene Wigner [1960]

To instruct someone ... is not a matter of getting him (sic) to commit results to mind. Rather, it is to teach him to participate in the process that makes possible the establishment of knowledge. We teach a subject not to produce little living libraries on that subject, but rather to get a student to think mathematically for himself ... to take part in the knowledge getting. Knowing is a process, not a product.

J. Bruner *Towards a theory of instruction* [1966]

The same pathological structures that the mathematicians invented to break loose from 19-th naturalism turn out to be inherent in familiar objects all around us in nature.

Freeman Dyson *Characterising Irregularity,* Science 200 [1978]

Anyone who has been in the least interested in mathematics, or has even observed other people who are interested in it, is aware that mathematical work is work with ideas. Symbols are used as aids to thinking just as musical scores are used in aids to music. The music comes first, the score comes later. Moreover, the score can never be a full embodiment of the musical thoughts of the composer. Just so, we know that a set of axioms and definitions is an attempt to describe the main properties of a mathematical idea. But there may always remain as aspect of the idea which we use implicitly, which we have not formalized because we have not yet seen the counterexample that would make us aware of the possibility of doubting it ...

Mathematics deals with ideas. Not pencil marks or chalk marks, not physical triangles or physical sets, but ideas (which may be represented or suggested by physical objects). What are the main properties of mathematical activity or mathematical knowledge, as known to all of us from daily experience? (1) Mathematical objects are invented or created by humans. (2) They are created, not arbitrarily, but arise from activity with already existing mathematical objects, and from the needs of science and daily life. (3) Once created, mathematical objects have properties which are well-determined, which we may have great difficulty discovering,

but which are possessed independently of our knowledge of them.

Reuben Hersh Advances in Mathematics **31** [1979]

Don't just read it; fight it! Ask your own questions, look for your own examples, discover your own proofs. Is the hypothesis necessary? Is the converse true? What happens in the classical special case? What about the degenerate cases? Where does the proof use the hypothesis?

Paul Halmos *I Want to be a Mathematician* [1985]

Mathematics is like a flight of fancy, but one in which the fanciful turns out to be real and to have been present all along. Doing mathematics has the feel of fanciful invention, but it is really a process for sharpening our perception so that we discover patterns that are everywhere around.... To share in the delight and the intellectual experience of mathematics – to fly where before we walked – that is the goal of mathematical education.

One feature of mathematics which requires special care ... is its "height", that is, the extent to which concepts build on previous concepts. Reasoning in mathematics can be very clear and certain, and, once a principle is established, it can be relied upon. This means that it is possible to build conceptual structures at once very tall, very reliable, and extremely powerful. The structure is not like a tree, but more like a scaffolding, with many interconnecting supports. Once the scaffolding is solidly in place, it is not hard to build up higher, but it is impossible to build a layer before the previous layers are in place.

William Thurston, Notices Amer. Math. Soc. [1990]

Contents

CHAPTER 1

Introduction

1.1. Preliminary Remarks

These Notes provide an introduction to the methods of contemporary mathematics, and in particular to Mathematical Analysis, which roughly speaking is the "in depth" study of Calculus.

The notes arise from various versions of MATH2320 and previous related courses. They include most of the material from the current MATH2320, and some more. However, the treatment may not always be the same. The notes are not a polished text, and there are undoubtedly a few typos!

The mathematics here is basic to most of your subsequent mathematics courses (e.g. differential equations, differential geometry, measure theory, numerical analysis, to name a few), as well as to much of theoretical physics, engineering, probability theory and statistics. Various interesting applications are included; in particular to fractals and to differential and integral equations.

There are also a few remarks of a general nature concerning *logic* and the nature of *mathematical proof*, and some discussion of *set theory*.

There are a number of *Exercises* scattered throughout the text. The Exercises are usually simple results, and you should do them all as an aid to your understanding of the material.

Sections, Remarks, etc. marked with a * are "extension" material, but you should read them anyway. They often help to set the other material in context as well as indicating further interesting directions.

The dependencies of the various chapters are noted in Figure 1.

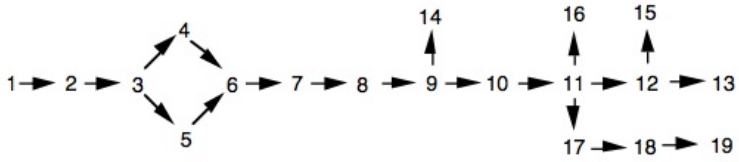

FIGURE 1. Chapter Dependencies.

There is a list of related books in the Bibliography.

The way to learn mathematics is by doing problems and by thinking *very* carefully about the material as you read it. Always ask yourself why the various assumptions in a theorem are made. It is almost always the case that if any particular assumption is dropped, then the conclusion of the theorem will no longer be true. Try to think of examples where the conclusion of the theorem is no longer valid if the various assumptions are changed. Try to see where each assumption is used in the proof of the theorem. Think of various interesting examples of the theorem.

1.2. History of Calculus

Calculus developed in the seventeenth and eighteenth centuries as a tool to describe various physical phenomena such as occur in astronomy, mechanics, and electrodynamics. But it was not until the nineteenth century that a proper understanding was obtained of the fundamental notions of *limit, continuity, derivative,* and *integral*. This understanding is important in both its own right and as a foundation for further deep applications to all of the topics mentioned in Section 1.1.

1.3. Why "Prove" Theorems?

A full understanding of a theorem, and in most cases the ability to apply it and to modify it in other directions as needed, comes only from knowing what really "makes it work", i.e. from an understanding of its proof.

1.4. "Summary and Problems" Book

There is an accompanying set of notes which contains a summary of all definitions, theorems, corollaries, etc. You should look through this at various stages to gain an overview of the material.

There is also a separate selection of problems and solutions available. The problems are at the level of the assignments which you will be required to do. They are not necessarily in order of difficulty. You should attempt, or at the very least think about, the problems before you look at the solutions. You will learn much more this way, and will in fact find the solutions easier to follow if you have already thought enough about the problems in order to realise where the main difficulties lie. You should also think of the solutions as examples of how to set out your own answers to other problems.

1.5. The approach to be used

Mathematics can be presented in a precise, logically ordered manner closely following a text. This may be an efficient way to cover the content, but bears little resemblance to how mathematics is actually done. In the words of Saunders Maclane (one of the developers of category theory) "intuition–trial–error–speculation–conjecture–proof is a sequence for understanding of mathematics." It is this approach which will be taken here, at least in part.

1.6. Acknowledgments

Thanks are due to many past students for suggestions and corrections, including Paulius Stepanas and Simon Stephenson, and to Maciej Kocan for supplying problems for some of the later chapters.

CHAPTER 2

Some Elementary Logic

In this Chapter we will discuss in an informal way some notions of logic and their importance in mathematical proofs. A very good reference is [Mo, Chapter I].

2.1. Mathematical Statements

In a mathematical proof or discussion one makes various assertions, often called *statements* or *sentences*.[1]
For example:

(1) $(x+y)^2 = x^2 + 2xy + y^2$.
(2) $3x^2 + 2x - 1 = 0$.
(3) if n (≥ 3) is an integer then $a^n + b^n = c^n$ has no positive integer solutions.
(4) the derivative of the function x^2 is $2x$.

Although a mathematical statement always has a very precise meaning, certain things are often assumed from the context in which the statement is made. For example, depending on the context in which statement (1) is made, it is probably an abbreviation for the statement

$$\text{for all real numbers } x \text{ and } y,\ (x+y)^2 = x^2 + 2xy + y^2.$$

However, it may also be an abbreviation for the statement

$$\text{for all complex numbers } x \text{ and } y,\ (x+y)^2 = x^2 + 2xy + y^2.$$

The precise meaning should always be clear from context; if it is not then more information should be provided.

Statement (2) probably refers to a particular real number x; although it is possibly an abbreviation for the (false) statement

$$\text{for all real numbers } x,\ 3x^2 + 2x - 1 = 0.$$

Again, the precise meaning should be clear from the context in which the statment occurs.

Statement (3) is known as Fermat's Last "Theorem".[2] An equivalent statement is

if n (≥ 3) is an integer and a, b, c are positive integers, then $a^n + b^n \neq c^n$.

Statement (4) is expressed informally. More precisely we interpret it as saying that the derivative of the function[3] $x \mapsto x^2$ is the function $x \mapsto 2x$.

Instead of the statement (1), let us again consider the more complete statement

[1] Sometimes one makes a distinction between sentences and statements (which are then certain types of sentences), but we do not do so.

[2] This was for a long time the best known open problem in mathematics; primarily because it is very simply stated and yet was incredibly difficult to solve. It was proved by Andrew Wiles in 1994, for which he received a knighthood and various other awards — but not a Fields Medal as he just exceeded the age requirement of ≤ 40. Wiles did his PhD under John Coates — Coates did his honours degree at ANU.

[3] By $x \mapsto x^2$ we mean the function f given by $f(x) = x^2$ for all real numbers x. We read "$x \mapsto x^2$" as "x maps to x^2".

$$\text{for all real numbers } x \text{ and } y,\ (x+y)^2 = x^2 + 2xy + y^2.$$

It is important to note that this has *exactly* the same meaning as

$$\text{for all real numbers } u \text{ and } v,\ (u+v)^2 = u^2 + 2uv + v^2,$$

or as

$$\text{for all real numbers } x \text{ and } v,\ (x+v)^2 = x^2 + 2xv + v^2.$$

In the previous line, the symbols u and v are sometimes called *dummy variables*. Note, however, that the statement

$$\text{for all real numbers } x\ ,\ (x+x)^2 = x^2 + 2xx + x^2$$

has a different meaning (while it is also true, it gives us "less" information).

In statements (3) and (4) the variables n, a, b, c, x are also dummy variables; changing them to other variables does not change the meaning of the statement. However, in statement (2) we are probably (depending on the context) referring to a *particular* number which we have denoted by x; and if we replace x by another variable which represents another number, then we do change the meaning of the statement.

2.2. Quantifiers

The expression *for all* (or *for every*, or *for each*, or (sometimes) *for any*), is called the *universal quantifier* and is often written $\forall$.

The following all have the same meaning (and are true)

(1) for all x and for all y, $(x+y)^2 = x^2 + 2xy + y^2$
(2) for any x and y, $(x+y)^2 = x^2 + 2xy + y^2$
(3) for each x and each y, $(x+y)^2 = x^2 + 2xy + y^2$
(4) $\forall x \forall y \Big((x+y)^2 = x^2 + 2xy + y^2 \Big)$

It is implicit in the above that when we say "for all x" or $\forall x$, we really mean for all *real numbers* x, etc. In other words, the quantifier $\forall$ "ranges over" the real numbers. More generally, we always quantify over some set of objects, and often make the *abuse of language* of suppressing this set when it is clear from context what is intended. If it is not clear from context, we can include the set over which the quantifier ranges. Thus we could write

$$\text{for all } x \in \mathbb{R} \text{ and for all } y \in \mathbb{R},\ (x+y)^2 = x^2 + 2xy + y^2,$$

which we abbreviate to

$$\forall x \in \mathbb{R}\, \forall y \in \mathbb{R} \Big((x+y)^2 = x^2 + 2xy + y^2 \Big).$$

Sometimes statement (1) is written as

$$(x+y)^2 = x^2 + 2xy + y^2 \text{ for all } x \text{ and } y.$$

Putting the quantifiers at the end of the statement can be very risky, however. This is particularly true when there are both existential and universal quantifiers involved. It is much safer to put quantifiers in front of the part of the statement to which they refer. See also the next section.

The expression *there exists* (or *there is*, or *there is at least one*, or *there are some*), is called the *existential quantifier* and is often written $\exists$.

The following statements all have the same meaning (and are true)

(1) there exists an irrational number
(2) there is at least one irrational number
(3) some real number is irrational
(4) irrational numbers exist
(5) $\exists x\, (x \text{ is irrational})$

The last statement is read as "there exists x such that x is irrational". It is implicit here that when we write $\exists x$, we mean that there exists a *real number* x. In other words, the quantifier $\exists$ "ranges over" the real numbers.

2.3. Order of Quantifiers

The order in which quantifiers occur is often critical. For example, consider the statements

$$\forall x \exists y (x < y) \tag{1}$$

and

$$\exists y \forall x (x < y). \tag{2}$$

We read these statements as

$$\text{for all } x \text{ there exists } y \text{ such that } x < y$$

and

$$\text{there exists } y \text{ such that for all } x,\ x < y,$$

respectively. Here (as usual for us) the quantifiers are intended to range over the real numbers. Note once again that the meaning of these statments is *unchanged* if we replace x and y by, say, u and v.[4]

Statement (1) is *true*. We can justify this as follows[5] (in somewhat more detail than usual!):
Let x be an arbitrary real number.
Then $x < x + 1$, and so $x < y$ is true if y equals (for example) $x + 1$.
Hence the statement $\exists y (x < y)$[6] is true.
But x was an *arbitrary* real number, and so the statement

$$\textit{for all } x \text{ there exists } y \text{ such that } x < y$$

is true. That is, (1) is *true*.

On the other hand, statement (2) is *false*.
It asserts that there exists some number y such that $\forall x (x < y)$.
But "$\forall x (x < y)$" means y is an upper bound for the set $\mathbb{R}$.
Thus (2) means "there exists y such that y is an upper bound for $\mathbb{R}$."
We know this last assertion is false.[7]

Alternatively, we could justify that (2) is false as follows:
Let y be an arbitrary real number.
Then $y + 1 < y$ is false.
Hence the statement $\forall x (x < y)$ is false.
Since y is an *arbitrary* real number, it follows that the statement

$$\textit{there exists } y \text{ such that for all } x,\ x < y,$$

is false.

There is much more discussion about various methods of proof in Section 2.6.3.

---

[4] In this case we could even be perverse and replace x by y and y by x respectively, without changing the meaning!

[5] For more discussion on this type of proof, see the discusion about the *arbitrary object method* in Subsection 2.6.3.

[6] Which, as usual, we read as "there exists y such that $x < y$.

[7] It is false because no matter which y we choose, the number $y + 1$ (for example) would be greater than y, contradicting the fact y is an upper bound for $\mathbb{R}$.

We have seen that reversing the order of consecutive existential and universal quantifiers can change the meaning of a statement. However, changing the order of consecutive existential quantifiers, or of consecutive universal quantifiers, does not change the meaning. In particular, if $P(x, y)$ is a statement whose meaning possibly depends on x and y, then

$$\forall x \forall y P(x, y) \quad \text{and} \quad \forall y \forall x P(x, y)$$

have the same meaning. For example,

$$\forall x \forall y \exists z (x^2 + y^3 = z),$$

and

$$\forall y \forall x \exists z (x^2 + y^3 = z),$$

both have the same meaning. Similarly,

$$\exists x \exists y P(x, y) \quad \text{and} \quad \exists y \exists x P(x, y)$$

have the same meaning.

2.4. Connectives

The *logical connectives* and the *logical quantifiers* (already discussed) are used to build new statements from old. The rigorous study of these concepts falls within the study of *Mathematical Logic* or the *Foundations of Mathematics*.

We now discuss the logical connectives.

2.4.1. Not. If p is a statement, then the *negation* of p is denoted by

$$\neg p \tag{3}$$

and is read as "not p".

If p is true then $\neg p$ is false, and if p is false then $\neg p$ is true.

The statement "not (not p)", i.e. $\neg\neg p$, means the same as "p".

Negation of Quantifiers

(1) The negation of $\forall x P(x)$, i.e. the statement $\neg\Big(\forall x P(x)\Big)$, is equivalent to $\exists x\Big(\neg P(x)\Big)$. Likewise, the negation of $\forall x \in \mathbb{R}\, P(x)$, i.e. the statement $\neg\Big(\forall x \in \mathbb{R}\, P(x)\Big)$, is equivalent to $\exists x \in \mathbb{R}\Big(\neg P(x)\Big)$; etc.

(2) The negation of $\exists x P(x)$, i.e. the statement $\neg\Big(\exists x P(x)\Big)$, is equivalent to $\forall x\Big(\neg P(x)\Big)$. Likewise, the negation of $\exists x \in \mathbb{R}\, P(x)$, i.e. the statement $\neg\Big(\exists x \in \mathbb{R}\, P(x)\Big)$, is equivalent to $\forall x \in \mathbb{R}\,\Big(\neg P(x)\Big)$.

(3) If we apply the above rules twice, we see that the negation of

$$\forall x \exists y P(x, y)$$

is equivalent to

$$\exists x \forall y \neg P(x, y).$$

Also, the negation of

$$\exists x \forall y P(x, y)$$

is equivalent to

$$\forall x \exists y \neg P(x, y).$$

Similar rules apply if the quantifiers range over specified sets; see the following Examples.

Examples

1 Suppose a is a fixed real number. The negation of

$$\exists x \in \mathbb{R}\,(x > a)$$

is equivalent to

$$\forall x \in \mathbb{R}\,\neg(x > a).$$

From the properties of inequalities, this is equivalent to

$$\forall x \in bR\,(x \leq a).$$

2 Theorem 3.2.10 says that

the set $\mathbb{N}$ of natural numbers is not bounded above.

The negation of this is the (false) statement

The set $\mathbb{N}$ of natural numbers is bounded above.

Putting this in the language of quantifiers, the Theorem says

$$\neg\Big(\exists y \forall x (x \leq y)\Big).$$

The negation is equivalent to

$$\exists y \forall x (x \leq y).$$

3 Corollary 3.2.11 says that

if $\epsilon > 0$ then there is a natural number n such that $0 < 1/n < \epsilon$.

In the language of quantifiers:

$$\forall \epsilon > 0\, \exists n \in \mathbb{N}\,(0 < 1/n < \epsilon).$$

The statement $0 < 1/n$ was only added for emphasis, and follows from the fact any natural number is positive and so its reciprocal is positive. Thus the Corollary is equivalent to

$$\forall \epsilon > 0\, \exists n \in \mathbb{N}\,(1/n < \epsilon). \tag{4}$$

The Corollary is proved by assuming it is false, i.e. by assuming the negation of (4), and obtaining a contradiction. Let us go through essentially the same argument again, but this time using quantifiers. This will take a little longer, but it enables us to see the logical structure of the proof more clearly.

Proof. The negation of (4) is equivalent to

$$\exists \epsilon > 0\, \forall n \in \mathbb{N}\, \neg(1/n < \epsilon). \tag{5}$$

From the properties of inequalities, and the fact ϵ and n range over certain sets of *positive* numbers, we have

$$\neg(1/n < \epsilon) \quad \text{iff} \quad 1/n \geq \epsilon \quad \text{iff} \quad n \leq 1/\epsilon.$$

Thus (5) is equivalent to

$$\exists \epsilon > 0\, \forall n \in \mathbb{N}\,(n \leq 1/\epsilon).$$

But this implies that the set of natural numbers is bounded above by $1/\epsilon$, and so is false by Theorem 3.2.10.

Thus we have obtained a contradiction from assuming the negation of (4), and hence (4) is true. □

4 The negation of

Every differentiable function is continuous (think of $\forall f \in D\, C(f)$)

is

Not (every differentiable function is continuous), i.e. $\neg\Big(\forall f \in D\, C(f)\Big)$,

and is equivalent to

> Some differentiable function is not continuous, i.e. $\exists f \in D\, \neg C(f)$.

or

> There exists a non-continuous differentiable function, which is also written $\exists f \in D\, \neg C(f)$.

5 The negation of "all elephants are pink", i.e. of $\forall x \in E\, P(x)$, is "not all elephants are pink", i.e. $\neg(\forall x \in E\, P(x))$, and an equivalent statement is "there exists an elephant which is not pink", i.e. $\exists x \in E\, \neg P(x)$.

The negation of "there exists a pink elephant", i.e. of $\exists x \in E\, P(x)$, is equivalent to "all elephants are not pink", i.e. $\forall x \in E\, \neg P(x)$.

This last statement is often confused in every-day discourse with the statement "not all elephants are pink", i.e. $\neg\Big(\forall x \in E\, P(x)\Big)$, although it has quite a different meaning, and is equivalent to "there is a non-pink elephant", i.e. $\exists x \in E\, \neg P(x)$. For example, if there were 50 pink elephants in the world and 50 white elephants, then the statement "all elephants are not pink" would be false, but the statement "not all elephants are pink" would be true.

2.4.2. And. If p and q are statements, then the *conjunction* of p and q is denoted by

$$p \wedge q \tag{6}$$

and is read as "p and q".

If both p and q are true then $p \wedge q$ is true, and otherwise it is false.

2.4.3. Or. If p and q are statements, then the *disjunction* of p and q is denoted by

$$p \vee q \tag{7}$$

and is read as "p or q".

If at least one of p and q is true, then $p \vee q$ is true. If both p and q are false then $p \vee q$ is false.

Thus the statement

$$1 = 1 \text{ or } 1 = 2$$

is true. This may seem different from common usage, but consider the following true statement

$$1 = 1 \text{ or I am a pink elephant.}$$

2.4.4. Implies. This often causes some confusion in mathematics. If p and q are statements, then the statement

$$p \Rightarrow q \tag{8}$$

is read as "p implies q" or "if p then q".

Alternatively, one sometimes says "q if p", "p only if q", "p" is a sufficient condition for "q", or "q" is a necessary condition for "p". But we will not usually use these wordings.

If p is true and q is false then $p \Rightarrow q$ is false, and in all other cases $p \Rightarrow q$ is true.

This may seem a bit strange at first, but it is essentially unavoidable. Consider for example the true statement

$$\forall x (x > 2 \Rightarrow x > 1).$$

Since in general we want to be able to say that a statement of the form $\forall x P(x)$ is true if and only if the statement $P(x)$ is true for every (real number) x, this leads us to saying that the statement

$$x > 2 \Rightarrow x > 1$$

is true, for *every* x. Thus we require that

$$\begin{aligned} 3 > 2 &\Rightarrow 3 > 1, \\ 1.5 > 2 &\Rightarrow 1.5 > 1, \\ .5 > 2 &\Rightarrow .5 > 1, \end{aligned}$$

all be true statements. Thus we have examples where p is true and q is true, where p is false and q is true, and where p is false and q is false; and in all three cases $p \Rightarrow q$ is true.

Next, consider the false statement

$$\forall x(x > 1 \Rightarrow x > 2).$$

Since in general we want to be able to say that a statement of the form $\forall x P(x)$ is false if and only if the statement $P(x)$ is false for some x, this leads us into requiring, for example, that

$$1.5 > 1 \Rightarrow 1.5 > 2$$

be false. This is an example where p is true and q is false, and $p \Rightarrow q$ is true.

In conclusion, if the truth or falsity of the statement $p \Rightarrow q$ is to depend only on the truth or falsity of p and q, then we cannot avoid the previous criterion in italics. See also the truth tables in Section 2.5.

Finally, in this respect, note that the statements

> If I am not a pink elephant then $1 = 1$
>
> If I am a pink elephant then $1 = 1$

and

> If pigs have wings then cabbages can be kings[8]

are true statements.

The statement $p \Rightarrow q$ is equivalent to $\neg(p \wedge \neg q)$, i.e. not(p and not q). This may seem confusing, and is perhaps best understood by considering the four different cases corresponding to the truth and/or falsity of p and q.

It follows that the negation of $\forall x\,(P(x) \Rightarrow Q(x))$ is equivalent to the statement $\exists x \neg\,(P(x) \Rightarrow Q(x))$ which in turn is equivalent to $\exists x\,(P(x) \wedge \neg Q(x))$.

As a final remark, note that the statement *all elephants are pink* can be written in the form $\forall x\,(E(x) \Rightarrow P(x))$, where $E(x)$ means x *is an elephant* and $P(x)$ means x *is pink*. Previously we wrote it in the form $\forall x \in E\, P(x)$, where here E is the *set* of pink elephants, rather than the property of being a pink elephant.

2.4.5. Iff. If p and q are statements, then the statement

$$p \Leftrightarrow q \tag{9}$$

is read as "p if and only if q", and is abbreviated to "p iff q", or "p is equivalent to q".

Alternatively, one can say "p is a necessary and sufficient condition for q".

If both p and q are true, or if both are false, then $p \Leftrightarrow q$ is true. It is false if (p is true and q is false), and it is also false if (p is false and q is true).

Remark In definitions it is conventional to use "if" where one should more strictly use "iff".

[8] With apologies to Lewis Carroll.

2.5. Truth Tables

In mathematics we require that the truth or falsity of $\neg p$, $p \wedge q$, $p \vee q$, $p \Rightarrow q$ and $p \Leftrightarrow q$ depend only on the truth or falsity of p and q.

The previous considerations then lead us to the following *truth tables*.

p	$\neg p$
T	F
F	T

p	q	$p \wedge q$	$p \vee q$	$p \Rightarrow q$	$p \Leftrightarrow q$
T	T	T	T	T	T
T	F	F	T	F	F
F	T	F	T	T	F
F	F	F	F	T	T

Remarks

(1) All the connectives can be defined in terms of $\neg$ and $\wedge$.
(2) The statement $\neg q \Rightarrow \neg p$ is called the *contrapositive* of $p \Rightarrow q$. It has the same meaning as $p \Rightarrow q$.
(3) The statement $q \Rightarrow p$ is the *converse* of $p \Rightarrow q$ and it does not have the same meaning as $p \Rightarrow q$.

2.6. Proofs

A mathematical proof of a theorem is a sequence of assertions (mathematical statements), of which the last assertion is the desired conclusion. Each assertion

(1) is an axiom or previously proved theorem, or
(2) is an assumption stated in the theorem, or
(3) follows from earlier assertions in the proof in an "obvious" way.

The word "obvious" is a problem. At first you should be very careful to write out proofs in full detail. Otherwise you will probably write out things which you think are obvious, but in fact are wrong. After practice, your proofs will become shorter.

A common mistake of beginning students is to write out the very easy points in much detail, but to quickly jump over the difficult points in the proof.

The problem of knowing "how much detail is required" is one which will become clearer with (much) practice.

In the next few subsections we will discuss various ways of proving mathematical statements.

Besides *Theorem*, we will also use the words *Proposition*, *Lemma* and *Corollary*. The distiction between these is not a precise one. Generally, "Theorems" are considered to be more significant or important than "Propositions". "Lemmas" are usually not considered to be important in their own right, but are intermediate results used to prove a later Theorem. "Corollaries" are fairly easy consequences of Theorems.

2.6.1. Proofs of Statements Involving Connectives. To prove a theorem whose conclusion is of the form "p and q" we have to show that both p is true and q is true.

To prove a theorem whose conclusion is of the form "p or q" we have to show that at least one of the statements p or q is true. Three different ways of doing this are:

- Assume p is false and use this to show q is true,
- Assume q is false and use this to show p is true,
- Assume p and q are both false and obtain a contradiction.

To prove a theorem of the type "p implies q" we may proceed in one of the following ways:

- Assume p is true and use this to show q is true,

- Assume q is false and use this to show p is false, i.e. prove the contrapositive of "p implies q",
- Assume p is true and q is false and use this to obtain a contradiction.

To prove a theorem of the type "p iff q" we usually

- Show p implies q and show q implies p.

2.6.2. Proofs of Statements Involving "There Exists". In order to prove a theorem whose conclusion is of the form "there exists x such that $P(x)$", we usually either

- show that for a certain explicit value of x, the statement $P(x)$ is true;

or more commonly

- use an *indirect argument* to show that some x with property $P(x)$ does exist.

  For example to prove

  $$\exists x \text{ such that } x^5 - 5x - 7 = 0$$

  we can argue as follows: Let the function f be defined by $f(x) = x^5 - 5x - 7$ (for all real x). Then $f(1) < 0$ and $f(2) > 0$; so $f(x) = 0$ for some x between 1 and 2 by the Intermediate Value Theorem[9] for continuous functions.

An alternative approach would be to

- assume $P(x)$ is false for all x and deduce a contradiction.

2.6.3. Proofs of Statements Involving "For Every". Consider the following trivial theorem:

THEOREM 2.6.1. *For every integer n there exists an integer m such that $m > n$.*

We cannot prove this theorem by individually examining each integer n. Instead we proceed as follows:

PROOF. Let n be any integer.

What this really means is—let n be a completely arbitrary integer, so that anything I prove about n applies equally well to any other integer. We continue the proof as follows:

Choose the integer

$$m = n + 1.$$

Then $m > n$.

Thus for every integer n there is a greater integer m. □

The above proof is an example of the *arbitrary object method*. We cannot examine every relevant object individually. So instead, we choose an arbitrary object x (integer, real number, etc.) and prove the result for this x. This is the same as proving the result for *every* x.

We often combine the *arbitrary object method* with *proof by contradiction*. That is, we often prove a theorem of the type "$\forall x P(x)$" as follows: Choose an *arbitrary* x and deduce a contradiction from "$\neg P(x)$". Hence $P(x)$ is true, and since x was arbitrary, it follows that "$\forall x P(x)$" is also true.

For example consider the theorem:

THEOREM 2.6.2. *$\sqrt{2}$ is irrational.*

From the definition of *irrational*, this theorem is interpreted as saying: "for all integers m and n, $m/n \neq \sqrt{2}$". We prove this equivalent formulation as follows:

[9] See later.

PROOF. Let m and n be arbitrary integers with $n \neq 0$ (as m/n is undefined if $n = 0$). Suppose that

$$m/n = \sqrt{2}.$$

By dividing through by any common factors greater than 1, we obtain

$$m^*/n^* = \sqrt{2}$$

where m^* and n^* have no common factors.

Then

$$(m^*)^2 = 2(n^*)^2.$$

Thus $(m^*)^2$ is even, and so m^* must also be even (the square of an odd integer is odd since $(2k+1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$).

Let $m^* = 2p$. Then

$$4p^2 = (m^*)^2 = 2(n^*)^2,$$

and so

$$2p^2 = (n^*)^2.$$

Hence $(n^*)^2$ is even, and so n^* is even.

Since both m^* and n^* are even, they must have the common factor 2, which is a contradiction. So $m/n \neq \sqrt{2}$. □

2.6.4. Proof by Cases. We often prove a theorem by considering various possibilities. For example, suppose we need to prove that a certain result is true for all pairs of integers m and n. It may be convenient to separately consider the cases $m = n$, $m < n$ and $m > n$.

CHAPTER 3

The Real Number System

3.1. Introduction

The *Real Number System* satisfies certain axioms, from which its other properties can be deduced. There are various slightly different, but equivalent, formulations.

DEFINITION 3.1.1. The *Real Number System* is a set[1] of objects called *Real Numbers* and denoted by $\mathbb{R}$ together with two binary operations[2] called *addition* and *multiplication* and denoted by $+$ and $\times$ respectively (we usually write xy for $x \times y$), a binary relation called *less than* and denoted by $<$, and two *distinct* elements called *zero* and *unity* and denoted by 0 and 1 respectively.

The axioms satisfied by these fall into three groups and are detailed in the following sections.

3.2. Algebraic Axioms

Algebraic properties are the properties of the four *operations*: *addition* $+$, *multiplication* $\times$, *subtraction* $-$, and *division* $\div$.

Properties of Addition If a, b and c are real numbers then:

A1: $a + b = b + a$
A2: $(a + b) + c = a + (b + c)$
A3: $a + 0 = 0 + a = a$
A4: there is exactly one real number, denoted by $-a$, such that $a + (-a) = (-a) + a = 0$

Property A1 is called the *commutative property of addition*; it says it does not matter how one commutes (interchanges) the order of addition.

Property A2 says that if we add a and b, and then add c to the result, we get the same as adding a to the result of adding b and c. It is called the *associative property of addition*; it does not matter how we associate (combine) the brackets. The analogous result is not true for subtraction or division.

Property A3 says there is a certain real number 0, called *zero* or the *additive identity*, which when added to any real number a, gives a.

Property A4 says that for any real number a there is a *unique* (i.e. exactly one) real number $-a$, called the *negative* or *additive inverse* of a, which when added to a gives 0.

Properties of Multiplication If a, b and c are real numbers then:

A5: $a \times b = b \times a$
A6: $(a \times b) \times c = a \times (b \times c)$
A7: $a \times 1 = 1 \times a = a$,and $1 \neq 0$.

[1] We discuss sets in the next Chapter.

[2] To say $+$ is a binary operation means that $+$ is a function such that $+ : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. We write $a + b$ instead of $+(a, b)$. Similar remarks apply to $\cdot$.

A8: if $a \neq 0$ there is exactly one real number, denoted by a^{-1}, such that $a \times a^{-1} = a^{-1} \times a = 1$

Properties A5 and A6 are called the *commutative* and *associative properties* for multiplication.

Property A7 says there is a real number $1 \neq 0$, called *one* or the *multiplicative identity*, which when multiplied by any real number a, gives a.

Property A8 says that for any *non-zero* real number a there is a *unique* real number a^{-1}, called the *multiplicative inverse* of a, which when multiplied by a gives 1.

Convention We will often write ab for $a \times b$.

The Distributive Property There is another property which involves both addition and multiplication:

A9: If a, b and c are real numbers then $a(b + c) = ab + ac$

The distributive property says that we can separately distribute multiplication over the two additive terms

Algebraic Axioms It turns out that one can prove all the *algebraic* properties of the real numbers from properties A1–A9 of addition and multiplication. We will do some of this in the next subsection.

We call A1–A9 the *Algebraic Axioms* for the real number system.

Equality One can write down various properties of equality. In particular, for all real numbers a, b and c:

(1) $a = a$
(2) $a = b \Rightarrow b = a$[3]
(3) $a = b$ and[4] $b = c \Rightarrow a = c$[5]

Also, if $a = b$, then $a + c = b + c$ and $ac = bc$. More generally, one can always replace a term in an expression by any other term to which it is equal.

It is *possible* to write down axioms for "=" and deduce the other properties of "=" from these axioms; but we do *not* do this. Instead, we take "=" to be a *logical notion* which means "is the same thing as"; the previous properties of "=" are then true from the *meaning* of "=".

When we write $a \neq b$ we will mean that a does *not* represent the same number as b; i.e. a represents a *different* number from b.

Other Logical and Set Theoretic Notions We do not attempt to axiomatise any of the *logical* notions involved in mathematics, nor do we attempt to axiomatise any of the properties of sets which we will use (see later). It *is* possible to do this; and this leads to some very deep and important results concerning the nature and foundations of mathematics. See later courses on the foundations mathematics (also some courses in the philosophy department).

3.2.1. Consequences of the Algebraic Axioms.

Subtraction and Division We first *define* subtraction in terms of addition and the additive inverse, by

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

---

[3] By $\Rightarrow$ we mean "implies". Let P and Q be two statements, then "$P \Rightarrow Q$" means "P implies Q"; or equivalently "if P then Q".

[4] We sometimes write "$\wedge$" for "and".

[5] Whenever we write "$P \wedge Q \Rightarrow R$", or "P and $Q \Rightarrow R$", the convention is that we mean "$(P \wedge Q) \Rightarrow R$", *not* "$P \wedge (Q \Rightarrow R)$".

Similarly, if $b \neq 0$ *define*

$$a \div b \left(= a/b = \frac{a}{b} \right) = ab^{-1}.$$

Some consequences of axioms A1–A9 are as follows. The proofs are given in the AH1 notes.

THEOREM 3.2.1 (Cancellation Law for Addition). *If a, b and c are real numbers and $a + c = b + c$, then $a = b$.*

THEOREM 3.2.2 (Cancellation Law for Multiplication). *If a, b and $c \neq 0$ are real numbers and $ac = bc$ then $a = b$.*

THEOREM 3.2.3. *If a, b, c, d are real numbers and $c \neq 0$, $d \neq 0$ then*

(1) $a0 = 0$
(2) $-(-a) = a$
(3) $(c^{-1})^{-1} = c$
(4) $(-1)a = -a$
(5) $a(-b) = -(ab) = (-a)b$
(6) $(-a) + (-b) = -(a + b)$
(7) $(-a)(-b) = ab$
(8) $(a/c)(b/d) = (ab)/(cd)$
(9) $(a/c) + (b/d) = (ad + bc)/cd$

Remark Henceforth (unless we say otherwise) we will assume all the usual properties of addition, multiplication, subtraction and division. In particular, we can solve simultaneous linear equations. We will also assume standard definitions including $x^2 = x \times x$, $x^3 = x \times x \times x$, $x^{-2} = \left(x^{-1}\right)^2$, etc.

3.2.2. Important Sets of Real Numbers. We define

$$2 = 1 + 1,\ 3 = 2 + 1\ ,\ \ldots\ ,\ 9 = 8 + 1\ ,$$
$$10 = 9 + 1\ ,\ldots,\ 19 = 18 + 1\ ,\ \ldots\ ,\ 100 = 99 + 1\ ,\ \ldots\ .$$

The set $\mathbb{N}$ of *natural numbers* is defined by

$$\mathbb{N} = \{1, 2, 3, \ldots\}.$$

The set $\mathbb{Z}$ of *integers* is defined by

$$\mathbb{Z} = \{m : -m \in \mathbb{N},\ \text{or}\ m = 0,\ \text{or}\ m \in \mathbb{N}\}.$$

The set $\mathbb{Q}$ of *rational numbers* is defined by

$$\mathbb{Q} = \{m/n : m \in \mathbb{Z}, n \in \mathbb{N}\}.$$

The set of all real numbers is denoted by $\mathbb{R}$.
A real number is *irrational* if it is not rational.

3.2.3. The Order Axioms. As remarked in Section 3.1, the real numbers have a natural ordering. Instead of writing down axioms directly for this ordering, it is more convenient to write out some axioms for the set P of *positive* real numbers. We then define $<$ in terms of P.

Order Axioms There is a subset[6] P of the set of real numbers, called the set of *positive numbers*, such that:

A10: For any real number a, exactly one of the following holds:

$$a = 0 \quad \text{or} \quad a \in P \quad \text{or} \quad -a \in P$$

A11: If $a \in P$ and $b \in P$ then $a + b \in P$ and $ab \in P$

[6]We will use some of the basic notation of set theory. Refer forward to Chapter 4 if necessary.

A number a is called *negative* when $-a$ is positive.

The "Less Than" Relation We now *define* $a < b$ to mean $b - a \in P$.
We also define:
$a \leq b$ to mean $b - a \in P$ or $a = b$;
$a > b$ to mean $a - b \in P$;
$a \geq b$ to mean $a - b \in P$ or $a = b$.

It follows that $a < b$ if and only if[7] $b > a$. Similarly, $a \leq b$ iff[8] $b \geq a$.

THEOREM 3.2.4. *If a, b and c are real numbers then*

(1) $a < b$ *and* $b < c$ *implies* $a < c$
(2) *exactly one of* $a < b$, $a = b$ *and* $a > b$ *is true*
(3) $a < b$ *implies* $a + c < b + c$
(4) $a < b$ *and* $c > 0$ *implies* $ac < bc$
(5) $a < b$ *and* $c < 0$ *implies* $ac > bc$
(6) $0 < 1$ *and* $-1 < 0$
(7) $a > 0$ *implies* $1/a > 0$
(8) $0 < a < b$ *implies* $0 < 1/b < 1/a$

Similar properties of $\leq$ can also be proved.

Remark Henceforth, in addition to assuming all the usual algebraic properties of the real number system, we will also assume all the standard results concerning inequalities.

Absolute Value The absolute value of a real number a is defined by

$$|a| = \begin{cases} a & \text{if } a \geq 0 \\ -a & \text{if } a < 0 \end{cases}$$

The following important properties can be deduced from the axioms; but we will not pause to do so.

THEOREM 3.2.5. *If a and b are real numbers then:*

(1) $|ab| = |a|\,|b|$
(2) $|a + b| \leq |a| + |b|$
(3) $\Big||a| - |b|\Big| \leq |a - b|$

We will use standard notation for *intervals*:

$$[a, b] = \{x : a \leq x \leq b\}, \quad (a, b) = \{x : a < x < b\},$$
$$(a, \infty) = \{x : x > a\}, \quad [a, \infty) = \{x : x \geq a\}$$

with similar definitions for $[a, b), (a, b], (-\infty, a], (-\infty, a)$. Note that ∞ is *not* a real number and there is no interval of the form $(a, \infty]$.

We only use the symbol ∞ as part of an expression which, when written out in full, does not refer to ∞.

3.2.4. Ordered Fields. Any set S, together with two operations $\oplus$ and $\otimes$ and two members $0_\oplus$ and $0_\otimes$ of S, and a subset $\mathcal{P}$ of S, which satisfies the corresponding versions of A1–A11, is called an *ordered field*.

Both $\mathbb{Q}$ and $\mathbb{R}$ are ordered fields, but finite fields are not.

Another example is the field of real algebraic numbers; where a real number is said to be *algebraic* if it is a solution of a polynomial equation of the form

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n = 0$$

[7] If A and B are statements, then "A if and only if B" means that A implies B *and* B implies A. Another way of expressing this is to say that A and B are either both true or both false.

[8] *Iff* is an abbreviation for *if and only if*.

for some integer $n > 0$ and integers $a_0, a_1, \ldots, a_n$. Note that any rational number $x = m/n$ is algebraic, since $m - nx = 0$, and that $\sqrt{2}$ is algebraic since it satisfies the equation $2 - x^2 = 0$. (As a nice exercise in algebra, show that the set of real algebraic numbers is indeed a field.)

3.2.5. Completeness Axiom. We now come to the property that singles out the real numbers from any other ordered field. There are a number of versions of this axiom. We take the following, which is perhaps a little more intuitive. We will later deduce the version in Adams.

Dedekind Completeness Axiom

A12 Suppose A and B are two (non-empty) sets of real numbers with the properties:

(1) if $a \in A$ and $b \in B$ then $a < b$
(2) every real number is in either A or B [9] (in symbols; $A \cup B = \mathbb{R}$).

Then there is a unique real number c such that:

: a. if $a < c$ then $a \in A$, and
: b. if $b > c$ then $b \in B$

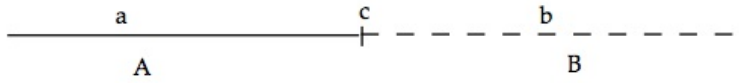

FIGURE 1. A Dedekind Cut $\{A, B\}$. Every $a < c$ belongs to A and every $b > c$ belongs to B.

Note that every number $< c$ belongs to A and every number $> c$ belongs to B. Moreover, either $c \in A$ or $c \in B$ by 2. Hence if $c \in A$ then $A = (-\infty, c]$ and $B = (c, \infty)$; while if $c \in B$ then $A = (-\infty, c)$ and $B = [c, \infty)$.

The pair of sets $\{A, B\}$ is called a *Dedekind Cut*.

The intuitive idea of A12 is that the Completeness Axiom says there are no "holes" in the real numbers.

Remark The analogous axiom is *not* true in the ordered field $\mathbb{Q}$. This is essentially because $\sqrt{2}$ is not rational, as we saw in Theorem 2.6.2.

More precisely, let

$$A = \{x \in \mathbb{Q} : x < \sqrt{2}\}, \; B = \{x \in \mathbb{Q} : x \geq \sqrt{2}\}.$$

(If you do not like to define A and B, which are sets of *rational* numbers, by using the *irrational* number $\sqrt{2}$, you could equivalently define

$$A = \{x \in \mathbb{Q} : x \leq 0 \text{ or } (x > 0 \text{ and } x^2 < 2)\}, \quad B = \{x \in \mathbb{Q} : x > 0 \text{ and } x^2 \geq 2\})$$

Suppose c satisfies a and b of A12. Then it follows from algebraic and order properties[10] that $c^2 \geq 2$ and $c^2 \leq 2$, hence $c^2 = 2$. But we saw in Theorem 2.6.2 that c cannot then be rational.

We next use Axiom A12 to prove the existence of $\sqrt{2}$, i.e. the existence of a number c such that $c^2 = 2$.

THEOREM 3.2.6. *There is a (positive)*[11] *real number c such that $c^2 = 2$.*

[9] It follows that if x is any real number, then x is in *exactly* one of A and B, since otherwise we would have $x < x$ from 1.

[10] Related arguments are given in a little more detail in the proof of the next Theorem.

[11] And hence a negative real number c such that $c^2 = 2$; just replace c by $-c$.

PROOF. Let

$$A = \{x \in \mathbb{R} : x \le 0 \text{ or } (x > 0 \text{ and } x^2 < 2)\},\ B = \{x \in \mathbb{R} : x > 0 \text{ and } x^2 \ge 2\}$$

It follows (from the algebraic and order properties of the real numbers; i.e. A1–A11) that every real number x is in exactly one of A or B, and hence that the two hypotheses of A12 are satisfied.
By A12 there is a unique real number c such that

(1) every number x less than c is either ≤ 0, or is > 0 and satisfies $x^2 < 2$
(2) every number x greater than c is > 0 and satisfies $x^2 \ge 2$.

See Figure 1.
From the Note following A12, either $c \in A$ or $c \in B$.
If $c \in A$ then $c < 0$ or ($c > 0$ and $c^2 < 2$). But then by taking $\epsilon > 0$ sufficiently small, we would also have $c + \epsilon \in A$ (from the definition of A), which contradicts conclusion b in A12.
Hence $c \in B$, i.e. $c > 0$ and $c^2 \ge 2$.
If $c^2 > 2$, then by choosing $\epsilon > 0$ sufficiently small we would also have $c - \epsilon \in B$ (from the definition of B), which contradicts a in A12.
Hence $c^2 = 2$. □

3.2.6. Upper and Lower Bounds.

DEFINITION 3.2.7. If S is a set of real numbers, then

(1) a is an *upper bound* for S if $x \le a$ for all $x \in S$;
(2) b is the *least upper bound* (or *l.u.b.* or *supremum* or sup) for S if b is an upper bound, and moreover $b \le a$ whenever a is any upper bound for S.

We write

$$b = \text{l.u.b. } S = \sup S$$

One similarly defines *lower bound* and *greatest lower bound* (or *g.l.b.* or *infimum* or inf) by replacing "$\le$" by "$\ge$".

A set S is *bounded above* if it has an upper bound[12] and is *bounded below* if it has a lower bound.

Note that if the l.u.b. or g.l.b. exists it is unique, since if b_1 and b_2 are both l.u.b.'s then $b_1 \le b_2$ and $b_2 \le b_1$, and so $b_1 = b_2$.

Examples

(1) If $S = [1, \infty)$ then any $a \le 1$ is a lower bound, and $1 = \text{g.l.b.}S$. There is no upper bound for S. The set S is bounded below but not above.
(2) If $S = [0, 1)$ then $0 = \text{g.l.b.}S \in S$ and $1 = \text{l.u.b.}S \notin S$. The set S is bounded above and below.
(3) If $S = \{1, 1/2, 1/3, \ldots, 1/n, \ldots\}$ then $0 = \text{g.l.b.}S \notin S$ and $1 = \text{l.u.b.}S \in S$. The set S is bounded above and below.

There is an equivalent form of the Completeness Axiom:

Least Upper Bound Completeness Axiom
A12′ Suppose S is a nonempty set of real numbers which is bounded above. Then S has a l.u.b. in $\mathbb{R}$.

A similar result follows for g.l.b.'s:

COROLLARY 3.2.8. *Suppose S is a nonempty set of real numbers which is bounded below. Then S has a g.l.b. in $\mathbb{R}$.*

[12] It follows that S has infinitely many upper bounds.

PROOF. Let

$$T = \{-x : x \in S\}.$$

Then it follows that a is a lower bound for S iff $-a$ is an upper bound for T; and b is a g.l.b. for S iff $-b$ is a l.u.b. for T.

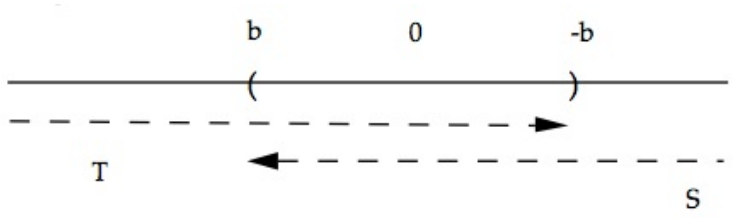

FIGURE 2. $T = \{-x : x \in S\}$ is obtained by reflecting S in the origin.

Since S is bounded below, it follows that T is bounded above. Moreover, T then has a l.u.b. c (say) by A12', and so $-c$ is a g.l.b. for S. □

Equivalence of A12 and A12′

1) Suppose A12 is true. We will deduce A12′.

For this, suppose that S is a nonempty set of real numbers which is bounded above.

Let

$$B = \{x : x \text{ is an upper bound for } S\}, \quad A = \mathbb{R} \setminus B.^{13}$$

Note that $B \neq \emptyset$; and if $x \in S$ then $x - 1$ is not an upper bound for S so $A \neq \emptyset$.
The first hypothesis in A12 is easy to check: suppose $a \in A$ and $b \in B$. If $a \geq b$ then a would also be an upper bound for S, which contradicts the definition of A, hence $a < b$.
The second hypothesis in A12 is immediate from the definition of A as consisting of every real number not in B.
Let c be the real number given by A12.

We claim that $c = l.u.b.\, S$.
If $c \in A$ then c is not an upper bound for S and so there exists $x \in S$ with $c < x$. But then $a = (c + x)/2$ is not an upper bound for S, i.e. $a \in A$, contradicting the fact from the conclusion of Axiom A12 that $a \leq c$ for all $a \in A$. Hence $c \in B$.
But if $c \in B$ then $c \leq b$ for all $b \in B$; i.e. c is $\leq$ any upper bound for S. This proves the claim; and hence proves A12′.

2) Suppose A12′ is true. We will deduce A12.

For this, suppose $\{A, B\}$ is a Dedekind cut.

Then A is bounded above (by any element of B). Let $c = l.u.b.\, A$, using A12'.
We claim that

$$a < c \Rightarrow a \in A, \quad b > c \Rightarrow b \in B.$$

Suppose $a < c$. Now every member of B is an upper bound for A, from the first property of a Dedekind cut; hence $a \notin B$, as otherwise a would be an upper bound for A which is *less* than the least upper bound c. Hence $a \in A$.

Next suppose $b > c$. Since c is an upper bound for A (in fact the *least* upper bound), it follows we cannot have $b \in A$, and thus $b \in B$.

This proves the claim, and hence A12 is true.

[13] $\mathbb{R} \setminus B$ is the set of real numbers x not in B

The following is a useful way to characterise the l.u.b. of a set. It says that $b = l.u.b.\, S$ iff b is an upper bound for S and there exist members of S arbitrarily close to b.

PROPOSITION 3.2.9. *Suppose S is a nonempty set of real numbers. Then $b = l.u.b.\, S$ iff*

(1) *$x \leq b$ for all $x \in S$, and*
(2) *for each $\epsilon > 0$ there exist $x \in S$ such that $x > b - \epsilon$.*

PROOF. Suppose S is a nonempty set of real numbers.

First assume $b = l.u.b.\, S$. Then 1 is certainly true.
Suppose 2 is *not* true. Then for some $\epsilon > 0$ it follows that $x \leq b - \epsilon$ for every $x \in S$, i.e. $b - \epsilon$ is an upper bound for S. This contradicts the fact $b = l.u.b.\, S$. Hence 2 *is* true.

Next assume that 1 and 2 are true. Then b is an upper bound for S from 1. Moreover, if $b' < b$ then from 2 it follows that b' is not an upper bound of S. Hence b' is the *least* upper bound of S. □

We will usually use the version Axiom A12′ rather than Axiom A12; and we will usually refer to either as the *Completeness Axiom*. Whenever we use the Completeness axiom in our future developments, we will explictly refer to it. The Completeness Axiom is essential in proving such results as the Intermediate Value Theorem[14].
Exercise: Give an example to show that the Intermediate Value Theorem does not hold in the "world of rational numbers".

3.2.7. *Existence and Uniqueness of the Real Number System. We began by assuming that $\mathbb{R}$, together with the operations $+$ and $\times$ and the set of positive numbers P, satisfies Axioms 1–12. But if we begin with the axioms for set theory, it is possible to *prove* the existence of a set of objects satisfying the Axioms 1–12.

This is done by first constructing the natural numbers, then the integers, then the rationals, and finally the reals. The natural numbers are constructed as certain types of sets, the negative integers are constructed from the natural numbers, the rationals are constructed as sets of ordered pairs as in Chapter II-2 of Birkhoff and MacLane. The reals are then constructed by the method of Dedekind Cuts as in Chapter IV-5 of Birkhoff and MacLane or Cauchy Sequences as in Chapter 28 of Spivak.

The structure consisting of the set $\mathbb{R}$, together with the operations $+$ and $\times$ and the set of positive numbers P, is uniquely characterised by Axioms 1–12, in the sense that any two structures satisfying the axioms are essentially the same. More precisely, the two systems are *isomorphic*, see Chapter IV-5 of Birkhoff and MacLane or Chapter 29 of Spivak.

3.2.8. The Archimedean Property. The fact that the set $\mathbb{N}$ of natural numbers is not bounded above, does not follow from Axioms 1–11. However, it does follow if we also use the Completeness Axiom.

THEOREM 3.2.10. *The set $\mathbb{N}$ of natural numbers is not bounded above.*

PROOF. Recall that $\mathbb{N}$ is defined to be the set

$$\mathbb{N} = \{1, 1+1, 1+1+1, \ldots\}.$$

[14]If a continuous real valued function $f : [a, b] \to \mathbb{R}$ satisfies $f(a) < 0 < f(b)$, then $f(c) = 0$ for some $c \in (a, b)$.

Assume that $\mathbb{N}$ is bounded above.[15] Then from the Completeness Axiom (version A12′), there is a *least* upper bound b for $\mathbb{N}$. That is,

$$n \in \mathbb{N} \text{ implies } n \leq b. \tag{10}$$

It follows that

$$m \in \mathbb{N} \text{ implies } m+1 \leq b, \tag{11}$$

since if $m \in \mathbb{N}$ then $m+1 \in \mathbb{N}$, and so we can now apply (10) with n there replaced by $m+1$.
But from (11) (and the properties of subtraction and of $<$) it follows that

$$m \in \mathbb{N} \text{ implies } m \leq b-1.$$

This is a contradiction, since b was taken to be the *least* upper bound of $\mathbb{N}$. Thus the assumption "$\mathbb{N}$ is bounded above" leads to a contradiction, and so it is false. Thus $\mathbb{N}$ is *not* bounded above. □

The following Corollary is often implicitly used.

COROLLARY 3.2.11. *If $\epsilon > 0$ then there is a natural number n such that $0 < 1/n < \epsilon$.*[16]

PROOF. *Assume* there is no natural number n such that $0 < 1/n < \epsilon$. Then for every $n \in \mathbb{N}$ it follows that $1/n \geq \epsilon$ and hence $n \leq 1/\epsilon$. Hence $1/\epsilon$ is an upper bound for $\mathbb{N}$, contradicting the previous Theorem.
Hence there *is* a natural number n such that $0 < 1/n < \epsilon$. □

We can now prove that between any two real numbers there is a rational number.

THEOREM 3.2.12. *For any two reals x and y, if $x < y$ then there exists a rational number r such that $x < r < y$.*

PROOF. (a) First suppose $y - x > 1$. Then there is an integer k such that $x < k < y$.
To see this, let l be the least upper bound of the set S of all integers j such that $j \leq x$. It follows that l itself is a member of S,and so in particular is an integer.[17])
Hence $l+1 > x$, since otherwise $l+1 \leq x$, i.e. $l+1 \in S$, contradicting the fact that $l = \text{lub}\, S$.
Moreover, since $l \leq x$ and $y - x > 1$, it follows from the properties of $<$ that $l+1 < y$. (Thus if $k = l+1$ then $x < k < y$. See Figure 3.
(b) Now just assume $x < y$.
By the previous Corollary choose a natural number n such that $1/n < y - x$.

---

[15] *Logical Point*: Our intention is to obtain a contradiction from this assumption, and hence to deduce that $\mathbb{N}$ is *not* bounded above.

[16] We usually use ϵ and δ to denote numbers that we think of as being small and positive. Note, however, that the result is true for *any* real number ϵ; but it is more "interesting" if ϵ is small.

[17] The least upper bound b of *any* set S of *integers* which is bounded above, must itself be a member of S. This is fairly clear, using the fact that members of S must be at least the fixed distance 1 apart.

More precisely, consider the interval $[b-1/2, b]$. Since the distance between any two integers is ≥ 1, there can be at most one member of S in this interval. If there is *no* member of S in $[b-1/2, b]$ then $b-1/2$ would also be an upper bound for S, contradicting the fact b is the least upper bound. Hence there is *exactly one* member s of S in $[b-1/2, b]$; it follows $s = b$ as otherwise s would be an upper bound for S which is $< b$; contradiction.

Note that this argument works for any set S whose members are all at least a fixed positive distance $d > 0$ apart. *Why?*

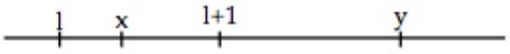

FIGURE 3. Here $y-x>1$ and l is the largest integer which is $\leq x$. It follows that $x<l+1<y$.

Hence $ny-nx>1$ and so by (a) there is an integer k such that $nx<k<ny$. Hence $x<k/n<y$, as required. □

A similar result holds for the irrationals.

THEOREM 3.2.13. *For any two reals x and y, if $x<y$ then there exists an irrational number r such that $x<r<y$.*

PROOF. First suppose a and b are rational and $a<b$.
Note that $\sqrt{2}/2$ is irrational (*why?*) and $\sqrt{2}/2<1$. Hence
$a<a+(b-a)\sqrt{2}/2<b$ and moreover $a+(b-a)\sqrt{2}/2$ is irrational[18].

To prove the result for general $x<y$, use the previous theorem twice to first choose a rational number a and then another rational number b, such that $x<a<b<y$.
By the first paragraph there is a rational number r such that $x<a<r<b<y$. □

COROLLARY 3.2.14. *For any real number x, and any positive number $\epsilon>$, there a rational (resp. irrational) number r (resp.s) such that $0<|r-x|<\epsilon$ (resp. $0<|s-x|<\epsilon$).*

[18]Let $r=a+(b-a)\sqrt{2}/2$. Hence $\sqrt{2}=2(r-a)/(b-a)$. So if r were rational then $\sqrt{2}$ would also be rational, which we know is not the case.

CHAPTER 4

Set Theory

4.1. Introduction

The notion of a *set* is fundamental to mathematics.

A *set* is, informally speaking, a *collection* of objects. We cannot use this as a definition however, as we then need to define what we mean by a *collection*.

The notion of a set is a basic or primitive one, as is membership $\in$, which are not usually defined in terms of other notions. Synonyms for *set* are *collection*, *class*[1] and *family*.

It is possible to write down axioms for the theory of sets. To do this properly, one also needs to formalise the logic involved. We will not follow such an axiomatic approach to set theory, but will instead proceed in a more informal manner.

Sets are important as it is possible to formulate all of mathematics in set theory. This is not done in practice, however, unless one is interested in the *Foundations of Mathematics*[2].

4.2. Russell's Paradox

It would seem reasonable to assume that for any "property" or "condition" P, there is a set S consisting of all objects with the given property.

More precisely, if $P(x)$ means that x has property P, then there should be a set S defined by

$$S = \{x : P(x)\}. \tag{12}$$

This is read as: "S is the set of all x such that $P(x)$ (is true)"[3].

For example, if $P(x)$ is an abbreviation for

$$x \text{ is an integer} > 5$$

or

$$x \text{ is a pink elephant,}$$

then there is a corresponding set (although in the second case it is the so-called *empty set*, which has no members) of objects x having property $P(x)$.

However, Bertrand Russell came up with the following property of x:

$$x \text{ is not a member of itself}^4,$$

or in symbols

$$x \notin x.$$

Suppose

$$S = \{x : x \notin x\}.$$

If there is indeed such a set S, then either $S \in S$ or $S \notin S$. But

[1]*Although we do not do so, in some studies of set theory, a distinction *is* made between *set* and *class*.

[2]There is a third/fourth year course Logic, Set Theory and the Foundations of Mathematics.

[3]Note that this is *exactly* the same as saying "S is the set of all z such that $P(z)$ (is true)".

[4]Any x we think of would normally have this property. Can you think of some x which is a member of itself? What about the set of weird ideas?

- if the first case is true, i.e. S is a member of S, then S *must* satisfy the defining property of the set S, and so $S \notin S$—contradiction;
- if the second case is true, i.e. S is not a member of S, then S *does not* satisfy the defining property of the set S, and so $S \in S$—contradiction.

Thus there is a contradiction in either case.

While this may seem an artificial example, there does arise the important problem of deciding which properties we should allow in order to describe sets. This problem is considered in the study of *axiomatic set theory*. We will not (hopefully!) be using properties that lead to such paradoxes, our construction in the above situation will rather be of the form "given a set A, consider the elements of A satisfying some defining property".

None-the-less, when the German philosopher and mathematician Gottlob Frege heard from Bertrand Russell (around the turn of the century) of the above property, just as the second edition of his two volume work *Grundgesetze der Arithmetik* (*The Fundamental Laws of Arithmetic*) was in press, he felt obliged to add the following acknowledgment:

> A scientist can hardly encounter anything more undesirable than to have the foundation collapse just as the work is finished. I was put in this position by a letter from Mr. Bertrand Russell when the work was almost through the press.

4.3. Union, Intersection and Difference of Sets

The members of a set are sometimes called *elements* of the set. If x is a member of the set S, we write

$$x \in S.$$

If x is not a member of S we write

$$x \notin S.$$

A set with a finite number of elements can often be described by explicitly giving its members. Thus

$$S = \left\{1, 3, \{1, 5\}\right\} \tag{13}$$

is the set with members 1,3 and $\{1, 5\}$. Note that 5 is *not* a member[5]. If we write the members of a set in a different order, we still have the same set.

If S is the set of all x such that $\ldots x \ldots$ is true, then we write

$$S = \{x : \ldots x \ldots\}, \tag{14}$$

and read this as "S is the set of x such that $\ldots x \ldots$". For example, if $S = \{x : 1 < x \leq 2\}$, then S is the interval of real numbers that we also denote by $(1, 2]$.

Members of a set may themselves be sets, as in (13).

If A and B are sets, their *union* $A \cup B$ is the set of all objects which belong to A or belong to B (remember that by the meaning of *or* this also includes those objects belonging to *both* A and B). Thus

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

The *intersection* $A \cap B$ of A and B is defined by

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

The *difference* $A \setminus B$ of A and B is defined by

$$A \setminus B = \{x : x \in A \text{ and } x \notin B\}.$$

[5]However, it is a member of a member; membership is generally not transitive.

It is sometimes convenient to represent this schematically by means of a *Venn Diagram.*

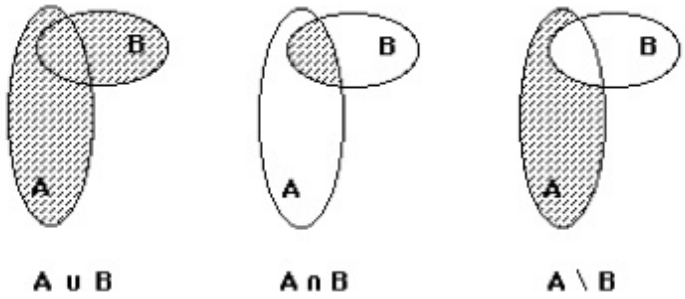

FIGURE 1. The union, intersection and difference of two sets.

We can take the union of more than two sets. If $\mathcal{F}$ is a family of sets, the *union* of all sets in $\mathcal{F}$ is defined by

$$\bigcup \mathcal{F} = \{x : x \in A \text{ for at least one } A \in \mathcal{F}\}. \tag{15}$$

The *intersection* of all sets in $\mathcal{F}$ is defined by

$$\bigcap \mathcal{F} = \{x : x \in A \text{ for every } A \in \mathcal{F}\}. \tag{16}$$

If $\mathcal{F}$ is finite, say $\mathcal{F} = \{A_1, \ldots, A_n\}$, then the union and intersection of members of $\mathcal{F}$ are written

$$\bigcup_{i=1}^{n} A_i \quad \text{or} \quad A_1 \cup \cdots \cup A_n \tag{17}$$

and

$$\bigcap_{i=1}^{n} A_i \quad \text{or} \quad A_1 \cap \cdots \cap A_n \tag{18}$$

respectively. If $\mathcal{F}$ is the family of sets $\{A_i : i = 1, 2, \ldots\}$, then we write

$$\bigcup_{i=1}^{\infty} A_i \quad \text{and} \quad \bigcap_{i=1}^{\infty} A_i \tag{19}$$

respectively. More generally, we may have a family of sets indexed by a set other than the integers— e.g. $\{A_\lambda : \lambda \in J\}$—in which case we write

$$\bigcup_{\lambda \in J} A_\lambda \quad \text{and} \quad \bigcap_{\lambda \in J} A_\lambda \tag{20}$$

for the union and intersection.

Examples

(1) $\bigcup_{n=1}^{\infty} [0, 1 - 1/n] = [0, 1)$
(2) $\bigcap_{n=1}^{\infty} [0, 1/n] = \{0\}$
(3) $\bigcap_{n=1}^{\infty} (0, 1/n) = \emptyset$

We say two sets A and B are *equal* iff they have the same members, and in this case we write

$$A = B. \tag{21}$$

It is convenient to have a set with no members; it is denoted by

$$\emptyset. \tag{22}$$

There is only one empty set, since any two empty sets have the same members, and so are equal!

If every member of the set A is also a member of B, we say A is a *subset* of B and write

$$A \subset B. \tag{23}$$

We include the possibility that $A = B$, and so in some texts this would be written as $A \subseteq B$. Notice the distinction between "$\in$" and "$\subset$". Thus in (13) we have $1 \in S$, $3 \in S$, $\{1,5\} \in S$ while $\{1,3\} \subset S$, $\{1\} \subset S$, $\{3\} \subset S$, $\{\{1,5\}\} \subset S$, $S \subset S$, $\emptyset \subset S$.

We usually prove $A = B$ by proving that $A \subset B$ and that $B \subset A$, c.f. the proof of (47) in Section 4.4.3.

If $A \subset B$ and $A \neq B$, we say A is a *proper subset* of B and write $A \subsetneq B$.

The sets A and B are *disjoint* if $A \cap B = \emptyset$. The sets belonging to a family of sets $\mathcal{F}$ are *pairwise disjoint* if any two distinctly indexed sets in $\mathcal{F}$ are disjoint.

The set of all subsets of the set A is called the *Power Set* of A and is denoted by

$$\mathcal{P}(A).$$

In particular, $\emptyset \in \mathcal{P}(A)$ and $A \in \mathcal{P}(A)$.

The following simple properties of sets are made plausible by considering a Venn diagram. We will prove some, and you should prove others as an exercise. Note that the proofs essentially just rely on the meaning of the logical words *and, or, implies* etc.

PROPOSITION 4.3.1. *Let A, B, C and B_λ (for $\lambda \in J$) be sets. Then*

$$\begin{aligned}
A \cup B &= B \cup A & A \cap B &= B \cap A \\
A \cup (B \cup C) &= (A \cup B) \cup C & A \cap (B \cap C) &= (A \cap B) \cap C \\
A &\subset A \cup B & A \cap B &\subset A \\
A \subset B &\text{ iff } A \cup B = B & A \subset B &\text{ iff } A \cap B = A \\
A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) & A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) \\
A \cap \bigcup_{\lambda \in J} B_\lambda &= \bigcup_{\lambda \in J} (A \cap B_\lambda) & A \cup \bigcap_{\lambda \in J} B_\lambda &= \bigcap_{\lambda \in J} (A \cup B_\lambda)
\end{aligned}$$

PROOF. We prove $A \subset B$ iff $A \cap B = A$ as an example.

First assume $A \subset B$. We want to prove $A \cap B = A$ (we will show $A \cap B \subset A$ and $A \subset A \cap B$). If $x \in A \cap B$ then certainly $x \in A$, and so $A \cap B \subset A$. If $x \in A$ then $x \in B$ by our assumption, and so $x \in A \cap B$, and hence $A \subset A \cap B$. Thus $A \cap B = A$.

Next assume $A \cap B = A$. We want to prove $A \subset B$. If $x \in A$, then $x \in A \cap B$ (as $A \cap B = A$) and so in particular $x \in B$. Hence $A \subset B$. □

If X is some set which contains all the objects being considered in a certain context, we sometimes call X a *universal* set. If $A \subset X$ then $X \setminus A$ is called the *complement* of A, and is denoted by

$$A^c. \tag{24}$$

Thus if X is the (set of) reals and A is the (set of) rationals, then the complement of A is the set of irrationals.

The complement of the union (intersection) of a family of sets is the intersection (union) of the complements; these facts are known as de Morgan's laws. More precisely,

PROPOSITION 4.3.2.

$$(A \cup B)^c = A^c \cap B^c \quad \textit{and} \quad (A \cap B)^c = A^c \cup B^c. \tag{25}$$

More generally,

$$\left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c \quad \textit{and} \quad \left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c, \tag{26}$$

and

$$\left(\bigcup_{\lambda\in J} A_\lambda\right)^c = \bigcap_{\lambda\in J} A_\lambda^c \quad \textit{and} \quad \left(\bigcap_{\lambda\in J} A_\lambda\right)^c = \bigcup_{\lambda\in J} A_\lambda^c. \tag{27}$$

4.4. Functions

We think of a function $f : A \to B$ as a way of assigning to each element $a \in A$ an element $f(a) \in B$. We will make this idea precise by defining functions as particular kinds of sets.

4.4.1. Functions as Sets. We first need the idea of an *ordered pair*. If x and y are two objects, the ordered pair whose *first* member is x and whose *second* member is y is denoted

$$(x, y). \tag{28}$$

The *basic property* of ordered pairs is that

$$(x, y) = (a, b) \quad \text{iff} \quad x = a \text{ and } y = b. \tag{29}$$

Thus $(x, y) = (y, x)$ iff $x = y$; whereas $\{x, y\} = \{y, x\}$ is always true. Any way of defining the notion of an ordered pair is satisfactory, provided it satisfies the *basic property*.

> One way to define the notion of an ordered pair in terms of sets is by setting
>
> $$(x, y) = \{\{x\}, \{x, y\}\}.$$
>
> This is natural: $\{x, y\}$ is the associated set of elements and $\{x\}$ is the set containing the first element of the ordered pair. As a non-trivial problem, you might like to try and prove the basic property of ordered pairs from this definition. HINT: consider separately the cases $x = y$ and $x \neq y$. The proof is in [**La**, pp. 42-43].

If A and B are sets, their *Cartesian product* is the set of all ordered pairs (x, y) with $x \in A$ and $y \in B$. Thus

$$A \times B = \{(x, y) : x \in A \text{ and } y \in B\}. \tag{30}$$

We can also define n-tuples $(a_1, a_2, \ldots, a_n)$ such that

$$(a_1, a_2, \ldots, a_n) = (b_1, b_2, \ldots, b_n) \quad \text{iff} \quad a_1 = b_1, a_2 = b_2, \ldots, a_n = b_n. \tag{31}$$

The *Cartesian Product* of n sets is defined by

$$A_1 \times A_2 \times \cdots \times A_n = \{(a_1, a_2, \ldots, a_n) : a_1 \in A_1, a_2 \in A_2, \ldots, a_n \in A_n\}. \tag{32}$$

In particular, we write

$$\mathbb{R}^n = \overbrace{\mathbb{R} \times \cdots \times \mathbb{R}}^{n}. \tag{33}$$

If f is a set of ordered pairs from $A \times B$ with the property that for every $x \in A$ there is exactly one $y \in B$ such that $(x, y) \in f$, then we say f is a *function* (or *map* or *transformation* or *operator*) from A to B. We write

$$f : A \to B, \tag{34}$$

which we read as: f sends (maps) A into B. If $(x, y) \in f$ then y is uniquely determined by x and for this particular x and y we write

$$y = f(x). \tag{35}$$

We say y is the *value* of f at x.

Thus if

$$f = \left\{(x, x^2) : x \in \mathbb{R}\right\} \tag{36}$$

then $f : \mathbb{R} \to \mathbb{R}$ and f is the function usually defined (somewhat loosely) by

$$f(x) = x^2, \tag{37}$$

where it is understood from context that x is a real number.

Note that it is exactly the same to define the function f by $f(x) = x^2$ for all $x \in \mathbb{R}$ as it is to define f by $f(y) = y^2$ for all $y \in \mathbb{R}$.

4.4.2. Notation Associated with Functions. Suppose $f : A \to B$. A is called the *domain* of f and B is called the *co-domain* of f.

The *range* of f is the set defined by

$$f[A] = \{y : y = f(x) \text{ for some } x \in A\} \tag{38}$$

$$= \{f(x) : x \in A\}. \tag{39}$$

Note that $f[A] \subset B$ but may not equal B. For example, in (37) the range of f is the set $[0, \infty) = \{x \in \mathbb{R} : 0 \leq x\}$.

We say f is *one-one* or *injective* or *univalent* if for every $y \in B$ there is *at most* one $x \in A$ such that $y = f(x)$. Thus the function $f_1 : \mathbb{R} \to \mathbb{R}$ given by $f_1(x) = x^2$ for all $x \in \mathbb{R}$ is not one-one, while the function $f_2 : \mathbb{R} \to \mathbb{R}$ given by $f_2(x) = e^x$ for all $x \in \mathbb{R}$ is one-one.

We say f is *onto* or *surjective* if every $y \in B$ is of the form $f(x)$ for some $x \in A$. Thus neither f_1 nor f_2 is onto. However, f_1 maps $\mathbb{R}$ onto $[0, \infty)$.

If f is both one-one and onto, then there is an *inverse function* $f^{-1} : B \to A$ defined by $f(y) = x$ iff $f(x) = y$. For example, if $f(x) = e^x$ for all $x \in \mathbb{R}$, then $f : \mathbb{R} \to [0, \infty)$ is one-one and onto, and so has an inverse which is usually denoted by ln. Note, incidentally, that $f : \mathbb{R} \to \mathbb{R}$ is not onto, and so strictly speaking does not have an inverse.

If $S \subset A$, then the *image* of S under f is defined by

$$f[S] = \{f(x) : x \in S\}. \tag{40}$$

Thus $f[S]$ is a subset of B, and in particular the image of A is the range of f.

If $S \subset A$, the *restriction* $f|_S$ of f to S is the function whose domain is S and which takes the same values on S as does f. Thus

$$f|_S = \{(x, f(x)) : x \in S\} \tag{41}$$

If $T \subset B$, then the *inverse image* of T under f is

$$f^{-1}[T] = \{x : f(x) \in T\}. \tag{42}$$

It is a subset of A. *Note* that $f^{-1}[T]$ is defined for *any* function $f : A \to B$. It is *not* necessary that f be one-one and onto, i.e. it is *not* necessary that the function f^{-1} exist.

If $f : A \to B$ and $g : B \to C$ then the *composition* function $g \circ f : A \to C$ is defined by

$$(g \circ f)(x) = g(f(x)) \ \forall x \in A. \tag{43}$$

For example, if $f(x) = x^2$ for all $x \in \mathbb{R}$ and $g(x) = \sin x$ for all $x \in \mathbb{R}$, then $(g \circ f)(x) = \sin(x^2)$ and $(f \circ g)(x) = (\sin x)^2$.

4.4.3. Elementary Properties of Functions. We have the following elementary properties:

PROPOSITION 4.4.1.

$$f[C \cup D] = f[C] \cup f[D] \qquad f\Big[\bigcup_{\lambda\in J} A_\lambda\Big] = \bigcup_{\lambda\in J} f[A_\lambda] \tag{44}$$

$$f[C \cap D] \subset f[C] \cap f[D] \qquad f\Big[\bigcap_{\lambda\in J} C_\lambda\Big] \subset \bigcap_{\lambda\in J} f[C_\lambda] \tag{45}$$

$$f^{-1}[U \cup V] = f^{-1}[U] \cup f^{-1}[V] \qquad f^{-1}\Big[\bigcup_{\lambda\in J} U_\lambda\Big] = \bigcup_{\lambda\in J} f^{-1}[U_\lambda] \tag{46}$$

$$f^{-1}[U \cap V] = f^{-1}[U] \cap f^{-1}[V] \qquad f^{-1}\Big[\bigcap_{\lambda\in J} U_\lambda\Big] = \bigcap_{\lambda\in J} f^{-1}[U_\lambda] \tag{47}$$

$$(f^{-1}[U])^c = f^{-1}[U^c] \tag{48}$$

$$A \subset f^{-1}\Big[f[A]\Big] \tag{49}$$

PROOF. The proofs of the above are straightforward. We prove (47) as an example of how to set out such proofs.

We need to show that $f^{-1}[U \cap V] \subset f^{-1}[U] \cap f^{-1}[V]$ and $f^{-1}[U] \cap f^{-1}[V] \subset f^{-1}[U \cap V]$.

For the first, suppose $x \in f^{-1}[U \cap V]$. Then $f(x) \in U \cap V$; hence $f(x) \in U$ and $f(x) \in V$. Thus $x \in f^{-1}[U]$ and $x \in f^{-1}[V]$, so $x \in f^{-1}[U] \cap f^{-1}[V]$. Thus $f^{-1}[U \cap V] \subset f^{-1}[U] \cap f^{-1}[V]$ (since x was an arbitrary member of $f^{-1}[U \cap V]$).

Next suppose $x \in f^{-1}[U] \cap f^{-1}[V]$. Then $x \in f^{-1}[U]$ and $x \in f^{-1}[V]$. Hence $f(x) \in U$ and $f(x) \in V$. This implies $f(x) \in U \cap V$ and so $x \in f^{-1}[U \cap V]$. Hence $f^{-1}[U] \cap f^{-1}[V] \subset f^{-1}[U \cap V]$. □

Exercise Give a simple example to show equality need not hold in (45).

4.5. Equivalence of Sets

DEFINITION 4.5.1. Two sets A and B are *equivalent* or *equinumerous* if there exists a function $f : A \to B$ which is one-one and onto. We write $A \sim B$.

The idea is that the two sets A and B *have the same number of elements*. Thus the sets $\{a, b, c\}$, $\{x, y, z\}$ and that in (13) are equivalent.

Some immediate consequences are:

PROPOSITION 4.5.2.

(1) $A \sim A$ *(i.e. $\sim$ is reflexive).*
(2) *If* $A \sim B$ *then* $B \sim A$ *(i.e. $\sim$ is symmetric).*
(3) *If* $A \sim B$ *and* $B \sim C$ *then* $A \sim C$ *(i.e. $\sim$ is transitive).*

PROOF. The first claim is clear.

For the second, let $f : A \to B$ be one-one and onto. Then the inverse function $f^{-1} : B \to A$, is also one-one and onto, as one can check (*exercise*).

For the third, let $f : A \to B$ be one-one and onto, and let $g : B \to C$ be one-one and onto. Then the composition $g \circ f : A \to B$ is also one-one and onto, as can be checked (*exercise*). □

DEFINITION 4.5.3. A set is *finite* if it is empty or is equivalent to the set $\{1, 2, \ldots, n\}$ for some natural number n. Otherwise it is *infinite*.

When we consider infinite sets there are some results which may seem surprising at first:

- The set E of even natural numbers is equivalent to the set $\mathbb{N}$ of natural numbers.
  To see this, let $f : E \to \mathbb{N}$ be given by $f(n) = n/2$. Then f is one-one and onto.
- The open interval (a, b) is equivalent to $\mathbb{R}$ (if $a < b$).
  To see this let $f_1(x) = (x - a)/(b - a)$; then $f_1 : (a, b) \to (0, 1)$ is one-one and onto, and so $(a, b) \sim (0, 1)$. Next let $f_2(x) = x/(1 - x)$; then $f_2 : (0, 1) \to (0, \infty)$ is one-one and onto[6] and so $(0, 1) \sim (0, \infty)$. Finally, if $f_3(x) = (1/x) - x$ then $f_3 : (0, \infty) \to \mathbb{R}$ is one-one and onto[7] and so $(0, \infty) \sim \mathbb{R}$. Putting all this together and using the transitivity of set equivalence, we obtain the result.

Thus we have examples where an apparently smaller subset of $\mathbb{N}$ (respectively $\mathbb{R}$) is in fact equivalent to $\mathbb{N}$ (respectively $\mathbb{R}$).

4.6. Denumerable Sets

DEFINITION 4.6.1. A set is *denumerable* if it is equivalent to $\mathbb{N}$. A set is *countable* if it is finite or denumerable. If a set is denumerable, we say it has *cardinality* d or *cardinal number* d [8].

Thus a set is denumerable iff it its members can be enumerated in a (non-terminating) sequence $(a_1, a_2, \ldots, a_n, \ldots)$. We show below that this fails to hold for infinite sets in general.

The following may not seem surprising but it still needs to be proved.

THEOREM 4.6.2. *Any denumerable set is infinite (i.e. is not finite).*

PROOF. It is sufficient to show that $\mathbb{N}$ is not finite (why?). But in fact any finite subset of $\mathbb{N}$ is bounded, whereas we know that $\mathbb{N}$ is not (Chapter 3). □

We have seen that the set of even integers is denumerable (and similarly for the set of odd integers). More generally, the following result is straightforward (the only problem is setting out the proof in a reasonable way):

THEOREM 4.6.3. *Any subset of a countable set is countable.*

PROOF. Let A be countable and let $(a_1, a_2, \ldots, a_n)$ or $(a_1, a_2, \ldots, a_n, \ldots)$ be an enumeration of A (depending on whether A is finite or denumerable). If $B \subset A$ then we construct a subsequence $(a_{i_1}, a_{i_2}, \ldots, a_{i_n}, \ldots)$ enumerating B by taking a_{i_j} to be the j'th member of the original sequence which is in B. Either this process never ends, in which case B is denumerable, or it does end in a finite number of steps, in which case B is finite. □

Remark This proof is rather more subtle than may appear. Why is the resulting function from $\mathbb{N} \to B$ onto? We should really prove that every non-empty set of natural numbers has a least member, but for this we need to be a little more precise in our definition of $\mathbb{N}$. See [**St**, pp 13–15] for details.

---

[6] This is clear from the graph of f_2. More precisely:
(i) if $x \in (0, 1)$ then $x/(1 - x) \in (0, \infty)$ follows from elementary properties of inequalities,
(ii) for each $y \in (0, \infty)$ there is a unique $x \in (0, 1)$ such that $y = x/(1 - x)$, namely $x = y/(1 + y)$, as follows from elementary algebra and properties of inequalities.

[7] As is again clear from the graph of f_3, or by arguments similar to those used for for f_2.

[8] See Section 4.8 for a more general discussion of cardinal numbers.

More surprising, at least at first, is the following result:

THEOREM 4.6.4. *The set $\mathbb{Q}$ is denumerable.*

PROOF. We have to show that $\mathbb{N}$ is equivalent to $\mathbb{Q}$.

In order to simplify the notation just a little, we first prove that $\mathbb{N}$ is equivalent to the set $\mathbb{Q}^+$ of *positive* rationals. We do this by arranging the rationals in a sequence, with no repetitions.

Each rational in $\mathbb{Q}^+$ can be uniquely written in the reduced form m/n where m and n are positive integers with no common factor. We write down a "doubly-infinite" array as follows:

> In the first row are listed all positive rationals whose reduced form is $m/1$ for some m (this is just the set of natural numbers);
> In the second row are all positive rationals whose reduced form is $m/2$ for some m;
> In the third row are all positive rationals whose reduced form is $m/3$ for some m;
> ...

The enumeration we use for $\mathbb{Q}^+$ is shown in the following diagram:

$$
\begin{array}{ccccccccccc}
1/1 & & 2/1 & \to & 3/1 & & 4/1 & \to & 5/1 & \dots \\
\downarrow & & \uparrow & & \downarrow & & \uparrow & & \downarrow & \\
1/2 & \to & 3/2 & & 5/2 & & 7/2 & & 9/2 & \dots \\
 & & & & \downarrow & & \uparrow & & \downarrow & \\
1/3 & \leftarrow & 2/3 & \leftarrow & 4/3 & & 5/3 & & 7/3 & \dots \\
\downarrow & & & & & & \uparrow & & \downarrow & \\
1/4 & \to & 3/4 & \to & 5/4 & \to & 7/4 & & 9/4 & \dots \\
 & & & & & & & & \downarrow & \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & \ddots
\end{array}
\tag{50}
$$

Finally, if $a_1, a_2, \dots$ is the enumeration of $\mathbb{Q}^+$ then $0, a_1, -a_1, a_2, -a_2, \dots$ is an enumeration of $\mathbb{Q}$. □

We will see in the next section that not all infinite sets are denumerable. However denumerable sets are the smallest infinite sets in the following sense:

THEOREM 4.6.5. *If A is infinite then A contains a denumerable subset.*

PROOF. Since $A \neq \emptyset$ there exists at least one element in A; denote one such element by a_1. Since A is not finite, $A \neq \{a_1\}$, and so there exists a_2, say, where $a_2 \in A$, $a_2 \neq a_1$. Similarly there exists a_3, say, where $a_3 \in A$, $a_3 \neq a_2$, $a_3 \neq a_1$. This process will never terminate, as otherwise $A \sim \{a_1, a_2, \dots, a_n\}$ for some natural number n.

Thus we construct a denumerable set $B = \{a_1, a_2, \dots\}$[9] where $B \subset A$. □

4.7. Uncountable Sets

There now arises the question

Are all infinite sets denumerable?

It turns out that the answer is *No*, as we see from the next theorem. Two proofs will be given, both are due to Cantor (late nineteenth century), and the underlying idea is the same.

THEOREM 4.7.1. *The sets $\mathbb{N}$ and $(0, 1)$ are not equivalent.*

[9] To be precise, we need the so-called *Axiom of Choice* to justify the construction of B by means of an infinite number of such choices of the a_i. See 4.10.1 below.

The first proof is by an ingenious *diagonalisation argument*. There are a couple of footnotes which may help you understand the proof.

PROOF. [10] We show that for any $f:\mathbb{N}\to(0,1)$, the map f *cannot* be onto. It follows that there is *no* one-one and onto map from $\mathbb{N}$ to $(0,1)$.

To see this, let $y_n = f(n)$ for each n. If we write out the decimal expansion for each y_n, we obtain a sequence

$$\begin{aligned} y_1 &= .a_{11}a_{12}a_{13}\ldots a_{1i}\ldots \\ y_2 &= .a_{21}a_{22}a_{23}\ldots a_{2i}\ldots \\ y_3 &= .a_{31}a_{32}a_{33}\ldots a_{3i}\ldots \\ &\vdots \\ y_i &= .a_{i1}\,a_{i2}\,a_{i3}\,\ldots a_{ii}\ldots \\ &\vdots \end{aligned} \tag{51}$$

Some rational numbers have two decimal expansions, e.g. $.14000\ldots = .13999\ldots$ but otherwise the decimal expansion is unique. In order to have uniqueness, we only consider decimal expansions which do *not* end in an infinite sequence of 9's.

To show that f cannot be *onto* we construct a real number z not in the above sequence, i.e. a real number z not in the range of f. To do this define $z = .b_1b_2b_3\ldots b_i\ldots$ by "going down the diagonal" as follows:

> Select $b_1 \neq a_{11}$, $b_2 \neq a_{22}$, $b_3 \neq a_{33}$, $\ldots$,$b_i \neq a_{ii}$,$\ldots$. We make sure that the decimal expansion for z does not end in an infinite sequence of 9's by also restricting $b_i \neq 9$ for each i; one explicit construction would be to set $b_n = a_{nn} + 1 \bmod 9$.

It follows that z is not in the sequence (51)[11], since for each i it is clear that z differs from the i'th member of the sequence in the i'th place of z's decimal expansion. But this implies that f is *not* onto. □

Here is the second proof.

PROOF. Suppose that (a_n) is a sequence of real numbers, we show that there is a real number $r \in (0,1)$ such that $r \neq a_n$ for every n.

Let I_1 be a closed subinterval of $(0,1)$ with $a_1 \notin I_1$, I_2 a closed subinterval of I_1 such that $a_2 \notin I_2$. Inductively, we obtain a sequence (I_n) of intervals such that $I_{n+1} \subseteq I_n$ for all n. Writing $I_n = [\alpha_n, \beta_n]$, the nesting of the intervals shows that $\alpha_n \leq \alpha_{n+1} < \beta_{n+1} \leq \beta_n$. In particular, (α_n) is bounded above, (β_n) is bounded below, so that $\alpha = \sup_n \alpha_n, \beta = \inf_n \beta_n$ are defined. Further it is clear that $[\alpha,\beta] \subseteq I_n$ for all n, and hence excludes all the (a_n). Any $r \in [\alpha,\beta]$ suffices. □

COROLLARY 4.7.2. *$\mathbb{N}$ is not equivalent to $\mathbb{R}$.*

PROOF. If $\mathbb{N} \sim \mathbb{R}$, then since $\mathbb{R} \sim (0,1)$ (from Section 4.5), it follows that $\mathbb{N} \sim (0,1)$ from Proposition (4.5.2). This contradicts the preceding theorem. □

A Common Error Suppose that A is an infinite set. Then it is not always correct to say "let $A = \{a_1, a_2, \ldots\}$". The reason is of course that this implicitly assumes that A is countable.

---

[10] We will show that any sequence ("list") of real numbers from $(0,1)$ cannot include *all* numbers from $(0,1)$. In fact, there will be an uncountable (see Definition 4.7.3) set of real numbers not in the list — but for the proof we only need to find one such number.

[11] First convince yourself that we really have constructed a number z. *Then* convince yourself that z is not in the list, i.e. z is not of the form y_n for any n.

DEFINITION 4.7.3. A set is *uncountable* if it is not countable. If a set is equivalent to $\mathbb{R}$ we say it has *cardinality* c (or *cardinal number* c)[12].

Another surprising result (again due to Cantor) which we prove in the next section is that the cardinality of $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ is also c.

Remark We have seen that the set of rationals has cardinality d. It follows[13] that the set of irrationals has cardinality c. Thus there are "more" irrationals than rationals.

On the other hand, the rational numbers are *dense* in the reals, in the sense that between any two distinct real numbers there is a rational number[14]. (It is also true that between any two distinct real numbers there is an irrational number[15].)

4.8. Cardinal Numbers

The following definition extends the idea of the number of elements in a set from finite sets to infinite sets.

DEFINITION 4.8.1. With every set A we associate a symbol called the *cardinal number* of A and denoted by $\overline{\overline{A}}$. Two sets are assigned the same cardinal number iff they are equivalent[16]. Thus $\overline{\overline{A}} = \overline{\overline{B}}$ iff $A \sim B$.

If $A = \emptyset$ we write $\overline{\overline{A}} = 0$.
If $A = \{a_1, \ldots, a_n\}$ (where $a_1, \ldots, a_n$ are all distinct) we write $\overline{\overline{A}} = n$.
If $A \sim \mathbb{N}$ we write $\overline{\overline{A}} = d$ (or $\aleph_0$, called "aleph zero", where $\aleph$ is the first letter of the Hebrew alphabet).
If $A \sim \mathbb{R}$ we write $\overline{\overline{A}} = c$.

DEFINITION 4.8.2. Suppose A and B are two sets. We write $\overline{\overline{A}} \leq \overline{\overline{B}}$ (or $\overline{\overline{B}} \geq \overline{\overline{A}}$) if A is equivalent to some subset of B, i.e. if there is a one-one map from A into B[17].

If $\overline{\overline{A}} \leq \overline{\overline{B}}$ and $\overline{\overline{A}} \neq \overline{\overline{B}}$, then we write $\overline{\overline{A}} < \overline{\overline{B}}$ (or $\overline{\overline{B}} > \overline{\overline{A}}$)[18].

PROPOSITION 4.8.3.

$$0 < 1 < 2 < 3 < \ldots < d < c. \tag{52}$$

PROOF. Consider the sets

$$\{a_1\}, \{a_1, a_2\}, \{a_1, a_2, a_3\}, \ldots, \mathbb{N}, \mathbb{R},$$

---

[12] c comes from *continuum*, an old way of referring to the set $\mathbb{R}$.

[13] We show in one of the problems for this chapter that if A has cardinality c and $B \subset A$ has cardinality d, then $A \setminus B$ has cardinality c.

[14] Suppose $a < b$. Choose an integer n such that $1/n < b - a$. Then $a < m/n < b$ for some integer m.

[15] Using the notation of the previous footnote, take the irrational number $m/n + \sqrt{2}/N$ for some sufficiently large natural number N.

[16] We are able to do this precisely because the relation of equivalence is reflexive, symmetric and transitive. For example, suppose 10 people are sitting around a round table. Define a relation between people by $A \sim B$ iff A is sitting next to B, or A is the same as B. It is not possible to assign to each person at the table a colour in such a way that two people have the same colour if and only if they are sitting next to each other. The problem is that the relation we have defined is reflexive and symmetric, but not transitive.

[17] This does not depend on the choice of sets A and B. More precisely, suppose $A \sim A'$ and $B \sim B'$, so that $\overline{\overline{A}} = \overline{\overline{A'}}$ and $\overline{\overline{B}} = \overline{\overline{B'}}$. Then A is equivalent to some subset of B iff A' is equivalent to some subset of B' (*exercise*).

[18] This is also independent of the choice of sets A and B in the sense of the previous footnote. The argument is similar.

where $a_1, a_2, a_3, \ldots$ are distinct from one another. There is clearly a one-one map from any set in this "list" into any later set in the list (*why?*), and so

$$1 \leq 2 \leq 3 \leq \ldots \leq d \leq c. \tag{53}$$

For any integer n we have $n \neq d$ from Theorem 4.6.2, and so $n < d$ from (53). Since $d \neq c$ from Corollary 4.7.2, it also follows that $d < c$ from (53).

Finally, the fact that $1 \neq 2 \neq 3 \neq \ldots$ (and so $1 < 2 < 3 < \ldots$ from (53)) can be proved by induction. □

PROPOSITION 4.8.4. *Suppose A is non-empty. Then for any set B, there exists a surjective function $g : B \to A$ iff $\overline{\overline{A}} \leq \overline{\overline{B}}$.*

PROOF. If g is onto, we can choose for every $x \in A$ an element $y \in B$ such that $g(y) = x$. Denote this element by $f(x)$[19]. Thus $g(f(x)) = x$ for all $x \in A$.

Then $f : A \to B$ and f is clearly one-one (since if $f(x_1) = f(x_2)$ then $g(f(x_1)) = g(f(x_2))$; but $g(f(x_1)) = x_1$ and $g(f(x_2)) = x_2$, and hence $x_1 = x_2$). Hence $\overline{\overline{A}} \leq \overline{\overline{B}}$.

Conversely, if $\overline{\overline{A}} \leq \overline{\overline{B}}$ then there exists a function $f : A \to B$ which is one-one. Since A is non-empty there is an element in A and we denote one such member by a. Now define $g : B \to A$ by

$$g(y) = \begin{cases} x & \text{if } f(x) = y, \\ a & \text{if there is no such } x. \end{cases}$$

Then g is clearly onto, and so we are done. □

We have the following important properties of cardinal numbers, some of which are trivial, and some of which are surprisingly difficult. Statement 2 is known as the *Schröder-Bernstein Theorem*.

THEOREM 4.8.5. *Let $\overline{\overline{A}}$, $\overline{\overline{B}}$ and $\overline{\overline{C}}$ be cardinal numbers. Then*

(1) $\overline{\overline{A}} \leq \overline{\overline{A}}$;
(2) $\overline{\overline{A}} \leq \overline{\overline{B}}$ *and* $\overline{\overline{B}} \leq \overline{\overline{A}}$ *implies* $\overline{\overline{A}} = \overline{\overline{B}}$;
(3) $\overline{\overline{A}} \leq \overline{\overline{B}}$ *and* $\overline{\overline{B}} \leq \overline{\overline{C}}$ *implies* $\overline{\overline{A}} \leq \overline{\overline{C}}$;
(4) *either* $\overline{\overline{A}} \leq \overline{\overline{B}}$ *or* $\overline{\overline{B}} \leq \overline{\overline{A}}$.

PROOF. The first and the third results are simple. The first follows from Theorem 4.5.2(1) and the third from Theorem 4.5.2(3).

The other two result are *not* easy.

**Proof of (2):* Since $\overline{\overline{A}} \leq \overline{\overline{B}}$ there exists a function $f : A \to B$ which is one-one (but not necessarily onto). Similarly there exists a one-one function $g : B \to A$ since $\overline{\overline{B}} \leq \overline{\overline{A}}$.

If $f(x) = y$ or $g(u) = v$ we say x is a *parent* of y and u is a *parent* of v. Since f and g are one-one, each element has exactly one parent, if it has any.

If $y \in B$ and there is a *finite* sequence $x_1, y_1, x_2, y_2, \ldots, x_n, y$ or $y_0, x_1, y_1, x_2, y_2, \ldots, x_n, y$, for some n, such that each member of the sequence is the parent of the next member, and such that the first member has no parent, then we say y has an *original ancestor*, namely x_1 or y_0 respectively. Notice that every member in the sequence has the same original ancestor. If y has no parent, then y is its own original ancestor. Some elements may have no original ancestor.

Let $A = A_A \cup A_B \cup A_\infty$, where A_A is the set of elements in A with original ancestor in A, A_B is the set of elements in A with original ancestor in B, and A_∞ is the set of elements in A with no original ancestor. Similarly let $B = B_A \cup B_B \cup B_\infty$, where B_A is the set of elements in B with original ancestor in A, B_B is the set of

[19]This argument uses the Axiom of Choice, see Section 4.10.1 below.

elements in B with original ancestor in B, and B_∞ is the set of elements in B with no original ancestor.

Define $h : A \to B$ as follows:

if $x \in A_A$ then $h(x) = f(x)$,
if $x \in A_B$ then $h(x) =$ the parent of x,
if $x \in A_\infty$ then $h(x) = f(x)$.

Note that every element in A_B must have a parent (in B), since if it did not have a parent in B then the element would belong to A_A. It follows that the definition of h makes sense.

If $x \in A_A$, then $h(x) \in B_A$, since x and $h(x)$ must have the same original ancestor (which will be in A). Thus $h : A_A \to B_A$. Similarly $h : A_B \to B_B$ and $h : A_\infty \to B_\infty$.

Note that h is one-one, since f is one-one and since each $x \in A_B$ has exactly one parent.

Every element y in B_A has a parent in A (and hence in A_A). This parent is mapped to y by f and hence by h, and so $h : A_A \to B_A$ is onto. A similar argument shows that $h : A_\infty \to B_\infty$ is onto. Finally, $h : A_B \to B_B$ is onto as each element y in B_B is the image under h of $g(y)$. It follows that h is onto.

Thus h is one-one and onto, as required. *End of proof of (2).*

**Proof of (4):* We do not really have the tools to do this, see Section 4.10.1 below. One lets

$$\mathcal{F} = \{f \mid f : U \to V,\ U \subset A,\ V \subset B,\ f \text{ is one-one and onto}\}.$$

It follows from *Zorn's Lemma*, see 4.10.1 below, that $\mathcal{F}$ contains a maximal element. Either this maximal element is a one-one function from A into B, or its inverse is a one-one function from B into A. □

COROLLARY 4.8.6. *Exactly one of the following holds:*

$$\overline{\overline{A}} < \overline{\overline{B}} \text{ or } \overline{\overline{A}} = \overline{\overline{B}} \text{ or } \overline{\overline{B}} < \overline{\overline{A}}. \tag{54}$$

PROOF. Suppose $\overline{\overline{A}} = \overline{\overline{B}}$. Then the second alternative holds and the first and third do not.

Suppose $\overline{\overline{A}} \neq \overline{\overline{B}}$. Either $\overline{\overline{A}} \leq \overline{\overline{B}}$ or $\overline{\overline{B}} \leq \overline{\overline{A}}$ from the previous theorem. Again from the previous theorem exactly one of these possibilities can hold, as both together would imply $\overline{\overline{A}} = \overline{\overline{B}}$. If $\overline{\overline{A}} \leq \overline{\overline{B}}$ then in fact $\overline{\overline{A}} < \overline{\overline{B}}$ since $\overline{\overline{A}} \neq \overline{\overline{B}}$. Similarly, if $\overline{\overline{B}} \leq \overline{\overline{A}}$ then $\overline{\overline{B}} < \overline{\overline{A}}$. □

COROLLARY 4.8.7. *If $A \subset \mathbb{R}$ and A includes an interval of positive length, then A has cardinality c.*

PROOF. Suppose $I \subset A$ where I is an interval of positive length. Then $\overline{\overline{I}} \leq \overline{\overline{A}} \leq \overline{\overline{\mathbb{R}}}$. Thus $c \leq \overline{\overline{A}} \leq c$, using the result at the end of Section 4.5 on the cardinality of an interval.

Hence $\overline{\overline{A}} = c$ from the Schröder-Bernstein Theorem. □

NB The converse to Corollary 4.8.7 is *false*. As an example, consider the following set

$$S = \{\sum_{n=1}^{\infty} a_n 3^{-n} : a_n = 0 \text{ or } 2\} \tag{55}$$

The mapping $S \to [0,1] : \sum_{n=1}^{\infty} \frac{a_n}{3^n} \mapsto \sum_{n=1}^{\infty} a_n 2^{-n-1}$ takes S onto $[0,1]$, so that S must be uncountable. On the other hand, S contains no interval at all. To see this, it suffices to show that for any $x \in S$, and any $\epsilon > 0$ there are points in $[x, x+\epsilon]$

lying outside S. It is a calculation to verify that $x+a3^{-k}$ is such a point for suitably large k, and suitable choice of $a = 1$ or 2 (*exercise*).

The set S above is known as the Cantor ternary set. It has further important properties which you will come across in topology and measure theory, see also Section 14.1.2.

We now prove the result promised at the end of the previous Section.

THEOREM 4.8.8. *The cardinality of* $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ *is* c.

PROOF. Let $f : (0,1) \to \mathbb{R}$ be one-one and onto, see Section 4.5. The map $(x, y) \mapsto (f(x), f(y))$ is thus (*exercise*) a one-one map from $(0,1) \times (0,1)$ onto $\mathbb{R} \times \mathbb{R}$; thus $(0,1) \times (0,1) \sim \mathbb{R} \times \mathbb{R}$. Since also $(0,1) \sim \mathbb{R}$, it is sufficient to show that $(0,1) \sim (0,1) \times (0,1)$.

Consider the map $f : (0,1) \times (0,1) \to (0,1)$ given by

$$(x, y) = (.x_1x_2x_3\ldots, .y_1y_2y_3\ldots) \mapsto .x_1y_1x_2y_2x_3y_3\ldots \tag{56}$$

We take the unique decimal expansion for each of x and y given by requiring that it does not end in an infinite sequence of 9's. Then f is one-one but not onto (since the number .191919... for example is not in the range of f). Thus $\overline{\overline{(0,1) \times (0,1)}} \leq \overline{\overline{(0,1)}}$.

On the other hand, there is a one-one map $g : (0,1) \to (0,1) \times (0,1)$ given by $g(z) = (z, 1/2)$, for example. Thus $\overline{\overline{(0,1)}} \leq \overline{\overline{(0,1) \times (0,1)}}$.

Hence $\overline{\overline{(0,1)}} = \overline{\overline{(0,1) \times (0,1)}}$ from the Schröder-Bernstein Theorem, and the result follows as $\overline{\overline{(0,1)}} = c$. □

The same argument, or induction, shows that $\mathbb{R}^n$ has cardinality c for each $n \in \mathbb{N}$. But what about $\mathbb{R}^{\mathbb{N}} = \{F : \mathbb{N} \to \mathbb{R}\}$?

4.9. More Properties of Sets of Cardinality c and d

THEOREM 4.9.1.

(1) *The product of two countable sets is countable.*
(2) *The product of two sets of cardinality c has cardinality c.*
(3) *The union of a countable family of countable sets is countable.*
(4) *The union of a cardinality c family of sets each of cardinality c has cardinality c.*

PROOF. (1) Let $A = (a_1, a_2, \ldots)$ and $B = (b_1, b_2, \ldots)$ (assuming A and B are infinite; the proof is similar if either is finite). Then $A \times B$ can be enumerated as follows (in the same way that we showed the rationals are countable):

$$\begin{array}{ccccccccc}
(a_1,b_1) & & (a_1,b_2) & \to & (a_1,b_3) & & (a_1,b_4) & \to & (a_1,b_5) & \ldots \\
\downarrow & & \uparrow & & \downarrow & & \uparrow & & \downarrow & \\
(a_2,b_1) & \to & (a_2,b_2) & & (a_2,b_3) & & (a_2,b_4) & & (a_2,b_5) & \ldots \\
 & & & & \downarrow & & \uparrow & & \downarrow & \\
(a_3,b_1) & \leftarrow & (a_3,b_2) & \leftarrow & (a_3,b_3) & & (a_3,b_4) & & (a_3,b_5) & \ldots \\
\downarrow & & & & & & \uparrow & & \downarrow & \\
(a_4,b_1) & \to & (a_4,b_2) & \to & (a_4,b_3) & \to & (a_4,b_4) & & (a_4,b_5) & \ldots \\
 & & & & & & & & \downarrow & \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & \ddots
\end{array} \tag{57}$$

(2) If the sets A and B have cardinality c then they are in one-one correspondence[20] with $\mathbb{R}$. It follows that $A \times B$ is in one-one correspondence with $\mathbb{R} \times \mathbb{R}$, and so the result follows from Theorem 4.8.8.

(3) Let $\{A_i\}_{i=1}^{\infty}$ be a countable family of countable sets. Consider an array whose first column enumerates the members of A_1, whose second column enumerates the members of A_2, etc. Then an enumeration similar to that in (1), but suitably modified to take account of the facts that some columns may be finite, that the number of columns may be finite, and that some elements may appear in more than one column, gives the result.

(4) Let $\{A_\alpha\}_{\alpha \in S}$ be a family of sets each of cardinality c, where the index set S has cardinality c. Let $f_\alpha : A_\alpha \to \mathbb{R}$ be a one-one and onto function for each α.

Let $A = \bigcup_{\alpha \in S} A_\alpha$ and define $f : A \to \mathbb{R} \times \mathbb{R}$ by $f(x) = (\alpha, f_\alpha(x))$ if $x \in A_\alpha$ (if $x \in A_\alpha$ for more than one α, choose one such α [21]). It follows that $\overline{\overline{A}} \leq \overline{\overline{\mathbb{R} \times \mathbb{R}}}$, and so $\overline{\overline{A}} \leq c$ from Theorem 4.8.8.

On the other hand there is a one-one map g from $\mathbb{R}$ into A (take g equal to the inverse of f_α for some $\alpha \in S$) and so $c \leq \overline{\overline{A}}$.

The result now follows from the Schröder-Bernstein Theorem. □

Remark The phrase "Let $f_\alpha : A_\alpha \to \mathbb{R}$ be a one-one and onto function for each α" looks like another invocation of the axiom of choice, however one could interpret the hypothesis on $\{A_\alpha\}_{\alpha \in S}$ as providing the maps f_α. This has implicitly been done in (3) and (4).

Remark It is clear from the proof that in (4) it is sufficient to assume that each set is countable or of cardinality c, provided that at least one of the sets has cardinality c.

4.10. *Further Remarks

4.10.1. The Axiom of choice. *For any non-empty set X, there is a function $f : \mathcal{P}(X) \to X$ such that $f(A) \in A$ for $A \in \mathcal{P}(X) \backslash \{\emptyset\}$.*

This axiom has a somewhat controversial history – it has some innocuous equivalences (see below), but other rather startling consequences such as the Banach-Tarski paradox[22]. It is known to be independent of the other usual axioms of set theory (it cannot be proved or disproved from the other axioms) and relatively consistent (neither it, nor its negation, introduce any new inconsistencies into set theory). Nowadays it is almost universally accepted and used without any further ado. For example, it is needed to show that any vector space has a basis or that the infinite product of non-empty sets is itself non-empty.

THEOREM 4.10.1. *The following are equivalent to the axiom of choice:*

(1) *If h is a function with domain A, there is a function f with domain A such that if $x \in A$ and $h(x) \neq \emptyset$, then $f(x) \in h(x)$.*
(2) *If $\rho \subseteq A \times B$ is a relation with domain A, then there exists a function $f : A \to B$ with $f \subseteq \rho$.*
(3) *If $g : B \to A$ is onto, then there exists $f : A \to B$ such that $g \circ f =$ identity on A.*

PROOF. These are all straightforward; (3) was used in 4.8.4. □

[20] A and B are in *one-one correspondence* means that there is a one-one map from A *onto* B.

[21] We are using the *Axiom of Choice* in simultaneously making such a choice for each $x \in A$.

[22] This says that a ball in $\mathbb{R}^3$ can be divided into five pieces which can be rearranged by rigid body motions to give two disjoint balls of the same radius as before!

For some of the most commonly used equivalent forms we need some further concepts.

DEFINITION 4.10.2. A relation $\leq$ on a set X is a *partial order* on X if, for all $x, y, z \in X$,

(1) $(x \leq y) \wedge (y \leq x) \Rightarrow x = y$ (*antisymmetry*), and
(2) $(x \leq y) \wedge (y \leq z) \Rightarrow x \leq z$ (*transitivity*), and
(3) $x \leq x$ for all $x \in X$ (*reflexivity*).

An element $x \in X$ is *maximal* if $(y \in X) \wedge (x \leq y) \Rightarrow y = x$, x is *maximum* (= greatest) if $z \leq x$ for all $z \in X$. Similar for minimal and minimum (= least), and upper and lower bounds.

A subset Y of X such that for any $x, y \in Y$, either $x \leq y$ or $y \leq x$ is called a *chain*. If X itself is a chain, the partial order is a *linear* or *total* order. A linear order $\leq$ for which every non-empty subset of X has a least element is a *well order*.

Remark Note that if $\leq$ is a partial order, then $\geq$, defined by $x \geq y := y \leq x$, is also a partial order. However, if both $\leq$ and $\geq$ are well orders, then the set is finite. (*exercise*).

With this notation we have the following, the proof of which is *not* easy (though some one way implications are).

THEOREM 4.10.3. *The following are equivalent to the axiom of choice:*

(1) **Zorn's Lemma** *A partially ordered set in which any chain has an upper bound has a maximal element.*
(2) **Hausdorff maximal principle** *Any partially ordered set contains a maximal chain.*
(3) **Zermelo well ordering principle** *Any set admits a well order.*
(4) **Teichmuller/Tukey maximal principle** *For any property of finite character on the subsets of a set, there is a maximal subset with the property*[23].

4.10.2. Other Cardinal Numbers. We have examples of infinite sets of cardinality d (e.g. $\mathbb{N}$) and c (e.g. $\mathbb{R}$).

A natural question is:

Are there other cardinal numbers?

The following theorem implies that the answer is YES.

THEOREM 4.10.4. *If A is any set, then $\overline{\overline{A}} < \overline{\overline{\mathcal{P}(A)}}$.*

PROOF. The map $a \to \{a\}$ is a one-one map from A into $\mathcal{P}(A)$.
If $f : A \to \mathcal{P}(A)$, let

$$X = \{a \in A : a \notin f(a)\}. \tag{58}$$

Then $X \in \mathcal{P}(A)$; suppose $X = f(b)$ for some b in A. If $b \in X$ then $b \notin f(b)$ (from the defining property of X), contradiction. If $b \notin X$ then $b \in f(b)$ (again from the defining property of X), contradiction. Thus X is not in the range of f and so f cannot be onto. □

Remark Note that the argument is similar to that used to obtain Russell's Paradox.

Remark Applying the previous theorem successively to $A = \mathbb{R}, \mathcal{P}(\mathbb{R}), \mathcal{P}(\mathcal{P}(\mathbb{R})), \ldots$ we obtain an increasing sequence of cardinal numbers. We can take the union S of all sets thus constructed, and it's cardinality is larger still. Then we can repeat

[23] A property of subsets is of *finite character* if a subset has the property iff all of its finite (sub)subsets have the property.

the procedure with $\mathbb{R}$ replaced by S, etc., etc. And we have barely scratched the surface!

It is convenient to introduce the notation $A \dot\cup B$ to indicate the union of A and B, considering the two sets to be disjoint.

THEOREM 4.10.5. *The following are equivalent to the axiom of choice:*

(1) *If A and B are two sets then either $\overline{\overline{A}} \le \overline{\overline{B}}$ or $\overline{\overline{B}} \le \overline{\overline{A}}$.*
(2) *If A and B are two sets, then $(\overline{\overline{A \times A}} = \overline{\overline{B \times B}}) \Rightarrow \overline{\overline{A}} = \overline{\overline{B}}$.*
(3) *$\overline{\overline{A \times A}} = \overline{\overline{A}}$ for any infinite set A.* (cf 4.9.1)
(4) *$\overline{\overline{A \times B}} = \overline{\overline{A \dot\cup B}}$ for any two infinite sets A and B.*

However $\overline{\overline{A \dot\cup A}} = \overline{\overline{A}}$ for all infinite sets $A \not\Rightarrow AC$.

4.10.3. The Continuum Hypothesis. Another natural question is:

Is there a cardinal number between c and d?

More precisely: Is there an infinite set $A \subset \mathbb{R}$ with *no* one-one map from A onto $\mathbb{N}$ and *no* one-one map from A onto $\mathbb{R}$? All infinite subsets of $\mathbb{R}$ that arise "naturally" either have cardinality c or d. The assertion that *all* infinite subsets of $\mathbb{R}$ have this property is called the *Continuum Hypothesis* (CH). More generally the assertion that for every infinite set A there is no cardinal number between $\overline{\overline{A}}$ and $\overline{\overline{\mathcal{P}(A)}}$ is the *Generalized Continuum Hypothesis* (GCH). It has been proved that the CH is an independent axiom in set theory, in the sense that it can neither be proved nor disproved from the other axioms (including the axiom of choice)[24]. Most, but by no means all, mathematicians accept at least CH. The axiom of choice is a consequence of GCH.

4.10.4. Cardinal Arithmetic. If $\alpha = \overline{\overline{A}}$ and $\beta = \overline{\overline{B}}$ are infinite cardinal numbers, we define their sum and product by

$$\alpha + \beta = \overline{\overline{A \dot\cup B}} \tag{59}$$

$$\alpha \times \beta = \overline{\overline{A \times B}}, \tag{60}$$

From Theorem 4.10.5 it follows that $\alpha + \beta = \alpha \times \beta = \max\{\alpha, \beta\}$.

More interesting is *exponentiation*; we define

$$\alpha^\beta = \overline{\overline{\{f \mid f : B \to A\}}}. \tag{61}$$

Why is this consistent with the usual definition of m^n and $\mathbb{R}^n$ where m and n are natural numbers?

For more information, see [**BM**, Chapter XII].

4.10.5. Ordinal numbers. Well ordered sets were mentioned briefly in 4.10.1 above. They are precisely the sets on which one can do (transfinite) induction. Just as cardinal numbers were introduced to facilitate the "size" of sets, ordinal numbers may be introduced as the "order-types" of well-ordered sets. Alternatively they may be defined explicitly as sets W with the following three properties.

(1) every member of W is a subset of W
(2) W is well ordered by $\subset$
(3) no member of W is an member of itself

Then $\mathbb{N}$, and its elements are ordinals, as is $\mathbb{N} \cup \{\mathbb{N}\}$. Recall that for $n \in \mathbb{N}$, $n = \{m \in \mathbb{N} : m < n\}$. An ordinal number in fact is equal to the set of ordinal numbers less than itself.

[24] We will discuss the Zermelo-Fraenkel axioms for set theory in a later course.

CHAPTER 5

Vector Space Properties of $\mathbb{R}^n$

In this Chapter we briefly review the fact that $\mathbb{R}^n$, together with the usual definitions of addition and scalar multiplication, is a vector space. With the usual definition of Euclidean inner product, it becomes an inner product space.

5.1. Vector Spaces

DEFINITION 5.1.1. A *Vector Space* (over the reals[1]) is a set V (whose members are called *vectors*), together with two operations called *addition* and *scalar multiplication*, and a particular vector called the *zero vector* and denoted by $\mathbf{0}$. The sum (addition) of $\mathbf{u}, \mathbf{v} \in V$ is a vector[2] in V and is denoted $\mathbf{u}+\mathbf{v}$; the scalar multiple of the scalar (i.e. real number) $c \in \mathbb{R}$ and $\mathbf{u} \in V$ is a vector in V and is denoted $c\mathbf{u}$. The following axioms are satisfied:

(1) $\mathbf{u}+\mathbf{v} = \mathbf{v}+\mathbf{u}$ for all $\mathbf{u}, \mathbf{v} \in V$ (*commutative law*)
(2) $\mathbf{u}+(\mathbf{v}+\mathbf{w}) = (\mathbf{u}+\mathbf{v})+\mathbf{w}$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ (*associative law*)
(3) $\mathbf{u}+\mathbf{0} = \mathbf{u}$ for all $\mathbf{u} \in V$ (*existence of an additive identity*)
(4) $(c+d)\mathbf{u} = c\mathbf{u}+d\mathbf{u}$, $\quad c(\mathbf{u}+\mathbf{v}) = c\mathbf{u}+c\mathbf{v}$ for all $c, d \in \mathbb{R}$ and $\mathbf{u}, \mathbf{v} \in V$ (*distributive laws*)
(5) $(cd)\mathbf{u} = c(d\mathbf{u})$ for all $c, d \in \mathbb{R}$ and $\mathbf{u} \in V$
(6) $1\mathbf{u} = \mathbf{u}$ for all $\mathbf{u} \in V$

Examples

(1) Recall that $\mathbb{R}^n$ is the set of all n-tuples $(a^1, \ldots, a^n)$ of real numbers. The *sum* of two n-tuples is defined by

$$(a^1, \ldots, a^n) + (b^1, \ldots, b^n) = (a^1+b^1, \ldots, a^n+b^n) \quad [3]. \tag{62}$$

The *product* of a scalar and an n-tuple is defined by

$$c(a^1, \ldots, a^n) = (ca^1, \ldots, ca^n). \tag{63}$$

The *zero* n-vector is defined to be

$$(0, \ldots, 0). \tag{64}$$

With these definitions it is easily checked that $\mathbb{R}^n$ becomes a vector space.

(2) Other very important examples of vector spaces are various spaces of functions. For example $\mathcal{C}[a, b]$, *the set of continuous*[4] *real-valued functions defined on the interval* $[a, b]$, with the usual addition of functions and multiplication of a scalar by a function, is a vector space (what is the zero vector?).

Remarks You should review the following concepts for a general vector space (see [**Fl**, Appendix 1] or [**An**]):

[1] One can define a vector space over the complex numbers in an analogous manner.

[2] It is common to denote vectors in boldface type.

[3] This is not a circular definition; we are defining addition of n-tuples in terms of addition of real numbers.

[4] We will discuss continuity in a later chapter. Meanwhile we will just use $\mathcal{C}[a, b]$ as a source of examples.

- linearly independent set of vectors, linearly dependent set of vectors,
- basis for a vector space, dimension of a vector space,
- linear operator between vector spaces.

The *standard basis* for $\mathbb{R}^n$ is defined by

$$\begin{aligned} \mathbf{e}_1 &= (1,0,\ldots,0) \\ \mathbf{e}_2 &= (0,1,\ldots,0) \\ &\vdots \\ \mathbf{e}_n &= (0,0,\ldots,1) \end{aligned} \tag{65}$$

Geometric Representation of $\mathbb{R}^2$ and $\mathbb{R}^3$ The vector $\mathbf{x} = (x^1, x^2) \in \mathbb{R}^2$ is represented geometrically in the plane either by the arrow from the origin $(0,0)$ to the point P with coordinates (x^1, x^2), or by any parallel arrow of the same length, or by the point P itself. Similar remarks apply to vectors in $\mathbb{R}^3$.

5.2. Normed Vector Spaces

A normed vector space is a vector space together with a notion of magnitude or length of its members, which satisfies certain axioms. More precisely:

DEFINITION 5.2.1. A *normed vector space* is a vector space V together with a real-valued function on V called a *norm*. The norm of $\mathbf{u}$ is denoted by $||\mathbf{u}||$ (sometimes $|\mathbf{u}|$). The following axioms are satisfied for all $\mathbf{u} \in V$ and all $\alpha \in \mathbb{R}$:

(1) $||\mathbf{u}|| \geq 0$ and $||\mathbf{u}|| = 0$ iff $\mathbf{u} = \mathbf{0}$ (positivity),
(2) $||\alpha\mathbf{u}|| = |\alpha|\,||\mathbf{u}||$ (homogeneity),
(3) $||\mathbf{u} + \mathbf{v}|| \leq ||\mathbf{u}|| + ||\mathbf{v}||$ (triangle inequality).

We usually abbreviate *normed vector space* to *normed space*.

Easy and important consequences (*exercise*) of the triangle inequality are

$$||\mathbf{u}|| \leq ||\mathbf{u} - \mathbf{v}|| + ||\mathbf{v}||, \tag{66}$$

$$\Big|\,||\mathbf{u}|| - ||\mathbf{v}||\,\Big| \leq ||\mathbf{u} - \mathbf{v}||. \tag{67}$$

Examples

(1) The vector space $\mathbb{R}^n$ is a normed space if we define $||(x^1,\ldots,x^n)||_2 = \left((x^1)^2 + \cdots + (x^n)^2\right)^{1/2}$. The only non-obvious part to prove is the triangle inequality. In the next section we will see that $\mathbb{R}^n$ is in fact an *inner product space*, that the norm we just defined is the norm corresponding to this inner product, and we will see that the triangle inequality is true for the norm in *any* inner product space.
(2) There are other norms which we can define on $\mathbb{R}^n$. For $1 \leq p < \infty$,

$$||(x^1,\ldots,x^n)||_p = \left(\sum_{i=1}^{n} |x^i|^p\right)^{1/p} \tag{68}$$

defines a norm on $\mathbb{R}^n$, called the *p-norm*. It is also easy to check that

$$||(x^1,\ldots,x^n)||_\infty = \max\{|x^1|,\ldots,|x^n|\} \tag{69}$$

defines a norm on $\mathbb{R}^n$, called the *sup norm*. *Exercise:* Show this notation is consistent, in the sense that

$$\lim_{p\to\infty} ||\mathbf{x}||_p = ||\mathbf{x}||_\infty \tag{70}$$

(3) Similarly, it is easy to check (*exercise*) that the *sup norm* on $\mathcal{C}[a,b]$ defined by

$$||f||_\infty = \sup|f| = \sup\{|f(x)| : a \le x \le b\} \tag{71}$$

is indeed a norm. (Note, incidentally, that since f is continuous, it follows that the sup on the right side of the previous equality is achieved at some $x \in [a,b]$, and so we could replace *sup* by *max*.)

(4) A norm on $\mathcal{C}[a,b]$ is defined by

$$||f||_1 = \int_a^b |f|. \tag{72}$$

Exercise: Check that this is a norm.

$\mathcal{C}[a,b]$ is also a normed space[5] with

$$||f|| = ||f||_2 = \left(\int_a^b f^2\right)^{1/2}. \tag{73}$$

Once again the triangle inequality is not obvious. We will establish it in the next section.

(5) Other examples are the set of all bounded sequences on $\mathbb{N}$:

$$\ell^\infty(\mathbb{N}) = \{(x_n) : ||(x_n)||_\infty = \sup|x_n| < \infty\}. \tag{74}$$

and its subset $c_0(\mathbb{N})$ of those sequences which converge to 0.

(6) On the other hand, for $\mathbb{R}^\mathbb{N}$, which is clearly a vector space under pointwise operations, has no natural norm. Why?

5.3. Inner Product Spaces

A (real) inner product space is a vector space in which there is a notion of magnitude and of orthogonality, see Definition 5.3.2. More precisely:

DEFINITION 5.3.1. An *inner product space* is a vector space V together with an operation called *inner product*. The inner product of $\mathbf{u}, \mathbf{v} \in V$ is a real number denoted by $\mathbf{u} \cdot \mathbf{v}$ or $(\mathbf{u}, \mathbf{v})$[6]. The following axioms are satisfied for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$:

(1) $\mathbf{u} \cdot \mathbf{u} \ge 0$, $\mathbf{u} \cdot \mathbf{u} = 0$ iff $\mathbf{u} = \mathbf{0}$ (*positivity*)
(2) $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$ (*symmetry*)
(3) $(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w} = \mathbf{u} \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w}$, $(c\mathbf{u}) \cdot \mathbf{v} = c(\mathbf{u} \cdot \mathbf{v})$ (*bilinearity*)[7]

Remark In the complex case $\mathbf{v} \cdot \mathbf{u} = \overline{\mathbf{u} \cdot \mathbf{v}}$. Thus from 2. and 3. The inner product is linear in the first variable and conjugate linear in the second variable, that is, it is *sesquilinear*.

Examples

(1) The *Euclidean inner product* (or *dot product* or *standard inner product*) of two vectors in $\mathbb{R}^n$ is defined by

$$(a^1, \ldots, a^n) \cdot (b^1, \ldots, b^n) = a^1 b^1 + \cdots + a^n b^n \tag{75}$$

It is easily checked that this does indeed satisfy the axioms for an inner product. The corresponding inner product space is denoted by E^n in [Fl], but we will abuse notation and use $\mathbb{R}^n$ for the set of n-tuples, for the corresponding vector space, and for the inner product space just defined.

---

[5] We will see the reason for the $||\cdot||_2$ notation when we discuss the L^p norm.

[6] Other notations are $\langle\cdot,\cdot\rangle$ and $(\cdot|\cdot)$.

[7] Thus an inner product is linear in the first argument. Linearity in the second argument then follows from 2.

(2) One can define *other* inner products on $\mathbb{R}^n$, these will be considered in the algebra part of the course. One simple class of examples is given by defining

$$(a^1,\dots,a^n)\cdot(b^1,\dots,b^n)=\alpha^1a^1b^1+\cdots+\alpha^na^nb^n, \tag{76}$$

where $\alpha^1,\dots,\alpha^n$ is any sequence of *positive* real numbers. *Exercise* Check that this defines an inner product.

(3) Another important example of an inner product space is $\mathcal{C}[a,b]$ with the inner product defined by $f\cdot g=\int_a^b fg$. *Exercise*: check that this defines an inner product.

DEFINITION 5.3.2. In an inner product space we define the *length* (or *norm*) of a vector by

$$|\mathbf{u}|=(\mathbf{u}\cdot\mathbf{u})^{1/2}, \tag{77}$$

and the notion of *orthogonality* between two vectors by

$$\mathbf{u} \text{ is orthogonal to } \mathbf{v} \text{ (written } \mathbf{u}\perp\mathbf{v}) \text{ iff } \mathbf{u}\cdot\mathbf{v}=0. \tag{78}$$

Example The functions

$$1,\cos x,\sin x,\cos 2x,\sin 2x,\dots \tag{79}$$

form an important (infinite) set of pairwise orthogonal functions in the inner product space $\mathcal{C}[0,2\pi]$, as is easily checked. This is the basic fact in the theory of Fourier series (you will study this theory at some later stage).

THEOREM 5.3.3. *An inner product on V has the following properties: for any $\mathbf{u},\mathbf{v}\in V$,*

$$|\mathbf{u}\cdot\mathbf{v}|\le|\mathbf{u}|\,|\mathbf{v}| \quad \textit{(}\text{Cauchy-Schwarz-Bunyakovsky Inequality}\textit{)}, \tag{80}$$

and if $\mathbf{v}\neq\mathbf{0}$ then equality holds iff $\mathbf{u}$ is a multiple of $\mathbf{v}$.

Moreover, $|\cdot|$ is a norm, and in particular

$$|\mathbf{u}+\mathbf{v}|\le|\mathbf{u}|+|\mathbf{v}| \quad \textit{(}\text{Triangle Inequality}\textit{)}. \tag{81}$$

If $\mathbf{v}\neq\mathbf{0}$ then equality holds iff $\mathbf{u}$ is a nonnegative *multiple of $\mathbf{v}$.*

The proof of the inequality is in [Fl, p. 6]. Although the proof given there is for the standard inner product in $\mathbb{R}^n$, the same proof applies to *any* inner product space. A similar remark applies to the proof of the triangle inequality in [Fl, p. 7]. The other two properties of a norm are easy to show.

An *orthonormal basis* for a finite dimensional inner product space is a basis $\{\mathbf{v}_1,\dots,\mathbf{v}_n\}$ such that

$$\mathbf{v}_i\cdot\mathbf{v}_j=\begin{cases}0 & \text{if } i\neq j\\ 1 & \text{if } i=j\end{cases} \tag{82}$$

Beginning from any basis $\{\mathbf{x_1},\dots,\mathbf{x_n}\}$ for an inner product space, one can construct an orthonormal basis $\{\mathbf{v}_1,\dots,\mathbf{v}_n\}$ by the *Gram-Schmidt process* described in [F, p.10 Question 10]. See Figure 1.

If $\mathbf{x}$ is a *unit vector* (i.e. $|\mathbf{x}|=1$) in an inner product space then the *component* of $\mathbf{v}$ in the direction of $\mathbf{x}$ is $\mathbf{v}\cdot\mathbf{x}$. In particular, in $\mathbb{R}^n$ the component of $(a^1,\dots,a^n)$ in the direction of $\mathbf{e}_i$ is a^i.

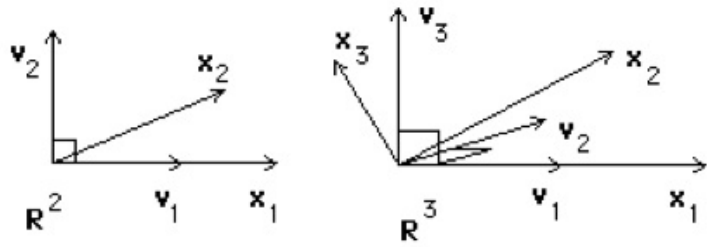

FIGURE 1. Gram-Schmidt process: Construct $\mathbf{v}_1$ of unit length in the subspace generated by $\mathbf{x}_1$; then construct $\mathbf{v}_2$ of unit length, orthogonal to $\mathbf{v}_1$, and in the subspace generated by $\mathbf{x}_1$ and $\mathbf{x}_2$; then construct $\mathbf{v}_3$ of unit length, orthogonal to $\mathbf{v}_1$ and $\mathbf{v}_2$, and in the subspace generated by $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$; etc.

CHAPTER 6

Metric Spaces

Metric spaces play a fundamental role in Analysis. In this chapter we will see that $\mathbb{R}^n$ is a particular example of a *metric space*. We will also study and use other examples of metric spaces.

6.1. Basic Metric Notions in $\mathbb{R}^n$

DEFINITION 6.1.1. The *distance* between two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is given by

$$d(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}| = \left((x^1 - y^1)^2 + \cdots + (x^n - y^n)^2\right)^{1/2}.$$

THEOREM 6.1.2. *For all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ the following hold:*

(1) $d(\mathbf{x}, \mathbf{y}) \geq 0$, $d(\mathbf{x}, \mathbf{y}) = 0$ *iff* $x = y$ (positivity),
(2) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (symmetry),
(3) $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$ (triangle inequality).

PROOF. The first two are immediate. For the third we have $d(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}| = |\mathbf{x} - \mathbf{z} + \mathbf{z} - \mathbf{y}| \leq |\mathbf{x} - \mathbf{z}| + |\mathbf{z} - \mathbf{y}| = d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$, where the inequality comes from version (81) of the triangle inequality in Section 5.3. □

6.2. General Metric Spaces

We now generalise these ideas as follows:

DEFINITION 6.2.1. A *metric space* (X, d) is a set X together with a distance function $d : X \times X \to \mathbb{R}$ such that for all $x, y, z \in X$ the following hold:

(1) $d(x, y) \geq 0$, $d(x, y) = 0$ iff $x = y$ (*positivity*),
(2) $d(x, y) = d(y, x)$ (*symmetry*),
(3) $d(x, y) \leq d(x, z) + d(z, y)$ (*triangle inequality*).

We often denote the corresponding metric space by (X, d), to indicate that a metric space is determined by *both* the set X and the metric d.

Examples

(1) We saw in the previous section that $\mathbb{R}^n$ together with the distance function defined by $d(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}|$ is a metric space. This is called the *standard* or *Euclidean* metric on $\mathbb{R}^n$.

    Unless we say otherwise, when referring to the metric space $\mathbb{R}^n$, we will always intend the Euclidean metric.

(2) More generally, *any normed space* is also a metric space, if we define

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

    The proof is the same as that for Theorem 6.1.2. As examples, the sup norm on $\mathbb{R}^n$, and both the inner product norm and the sup norm on $\mathcal{C}[a, b]$ (c.f. Section 5.2), induce corresponding metric spaces.

(3) An example of a metric space which is *not* a vector space is a smooth surface S in $\mathbb{R}^3$, where the distance between two points $\mathbf{x}, \mathbf{y} \in S$ is defined to be the length of the shortest curve joining $\mathbf{x}$ and $\mathbf{y}$ and lying entirely in S. Of course to make this precise we first need to define *smooth*, *surface*,

curve, and *length*, as well as consider whether there will exist a curve of shortest length (and is this necessary anyway?)

(4) *French metro, Post Office* Let $X = \{\mathbf{x} \in \mathbb{R}^2 : |\mathbf{x}| \leq 1\}$ and define

$$d(\mathbf{x}, \mathbf{y}) = \begin{cases} |\mathbf{x} - \mathbf{y}| & \text{if } \mathbf{x} = t\mathbf{y} \text{ for some scalar } t \\ |\mathbf{x}| + |\mathbf{y}| & \text{otherwise} \end{cases}$$

One can check that this defines a metric—the French metro with Paris at the centre. The distance between two stations on different lines is measured by travelling in to Paris and then out again.

(5) *p-adic metric.* Let $X = \mathbb{Z}$, and let $p \in \mathbb{N}$ be a fixed prime. For $x, y \in \mathbb{Z}, x \neq y$, we have $x - y = p^k n$ uniquely for some $k \in \mathbb{N}$, and some $n \in \mathbb{Z}$ not divisible by p. Define

$$d(x, y) = \begin{cases} (k+1)^{-1} & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}$$

One can check that this defines a metric which in fact satisfies the *strong triangle inequality* (which implies the usual one):

$$d(x, y) \leq \max\{d(x, z), d(z, y)\}.$$

Members of a general metric space are often called *points*, although they may be functions (as in the case of $\mathcal{C}[a, b]$), sets (as in 14.5) or other mathematical objects.

DEFINITION 6.2.2. Let (X, d) be a metric space. The *open ball* of radius $r > 0$ centred at x, is defined by

$$B_r(x) = \{y \in X : d(x, y) < r\}. \tag{83}$$

Note that the open balls in $\mathbb{R}$ are precisely the intervals of the form (a, b) (the centre is $(a + b)/2$ and the radius is $(b - a)/2$).

Exercise: Draw the open ball of radius 1 about the point $(1, 2) \in \mathbb{R}^2$, with respect to the Euclidean (L^2), sup (L^∞) and L^1 metrics. What about the French metro?

It is often convenient to have the following notion [1].

DEFINITION 6.2.3. Let (X, d) be a metric space. The subset $Y \subset X$ is a *neighbourhood* of $x \in X$ if there is $R > 0$ such that $B_r(x) \subset Y$.

DEFINITION 6.2.4. A subset S of a metric space X is *bounded* if $S \subset B_r(x)$ for some $x \in X$ and some $r > 0$.

PROPOSITION 6.2.5. *If S is a bounded subset of a metric space X, then for* every $y \in X$ *there exists $\rho > 0$ (ρ depending on y) such that $S \subset B_\rho(y)$.*

In particular, a subset S of a normed space is bounded iff $S \subset B_r(0)$ for some r, i.e. iff for some real number r, $||x|| < r$ for all $x \in S$.

PROOF. Assume $S \subset B_r(x)$ and $y \in X$. Then $B_r(x) \subset B_\rho(y)$ where $\rho = r + d(x, y)$; since if $z \in B_r(x)$ then $d(z, y) \leq d(z, x) + d(x, y) < r + d(x, y) = \rho$, and so $z \in B_\rho(y)$. Since $S \subset B_r(x) \subset B_\rho(y)$, it follows $S \subset B_\rho(y)$ as required. □

The previous proof is a typical application of the triangle inequality in a metric space.

[1] Some definitions of neighbourhood require the set to be open (see Section 6.4 below).

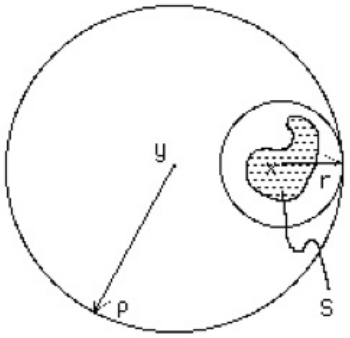

FIGURE 1. $S \subset B_r(x)$, and $B_r(x) \subset B_\rho(y)$ if $\rho = r + d(x, y)$.

6.3. Interior, Exterior, Boundary and Closure

Everything from this section, including proofs, unless indicated otherwise and apart from specific examples, applies with $\mathbb{R}^n$ replaced by an arbitrary metric space (X, d).

The following ideas make precise the notions of a point being strictly inside, strictly outside, on the boundary of, or having arbitrarily close-by points from, a set A.

DEFINITION 6.3.1. Suppose that $A \subset \mathbb{R}^n$. A point $\mathbf{x} \in \mathbb{R}^n$ is an *interior* (respectively *exterior*) point of A if some open ball centred at $\mathbf{x}$ is a subset of A (respectively A^c). If *every* open ball centred at $\mathbf{x}$ contains at least one point of A and at least one point of A^c, then $\mathbf{x}$ is a *boundary* point of A.

The set of interior (exterior) points of A is called the *interior* (*exterior*) of A and is denoted by A^0 or int A (ext A). The set of boundary points of A is called the *boundary* of A and is denoted by ∂A.

PROPOSITION 6.3.2. *Suppose that $A \subset \mathbb{R}^n$.*

$$\mathbb{R}^n = int\ A \cup \partial A \cup ext\ A, \tag{84}$$

$$ext\ A = int\ (A^c), \quad int\ A = ext\ (A^c), \tag{85}$$

$$int\ A \subset A, \quad ext\ A \subset A^c. \tag{86}$$

The three sets on the right side of (84) are mutually disjoint.

PROOF. These all follow immediately from the previous definition, *why?* □

We next make precise the notion of a point for which there are members of A which are arbitrarily close to that point.

DEFINITION 6.3.3. Suppose that $A \subset \mathbb{R}^n$. A point $\mathbf{x} \in \mathbb{R}^n$ is a *limit point* of A if every open ball centred at $\mathbf{x}$ contains at least one member of A *other than* $\mathbf{x}$. A point $\mathbf{x} \in A \subset \mathbb{R}^n$ is an *isolated point* of A if some open ball centred at $\mathbf{x}$ contains no members of A other than $\mathbf{x}$ itself.

NB The terms *cluster point* and *accumulation point* are also used here. However, the usage of these three terms is not universally the same throughout the literature.

DEFINITION 6.3.4. The *closure* of $A \subset \mathbb{R}^n$ is the union of A and the set of limit points of A, and is denoted by $\overline{A}$.

The following proposition follows directly from the previous definitions.

PROPOSITION 6.3.5. *Suppose $A \subset \mathbb{R}^n$.*

(1) *A limit point of A need not be a member of A.*
(2) *If $\mathbf{x}$ is a limit point of A, then every open ball centred at $\mathbf{x}$ contains an* **infinite** *number of points from A.*
(3) $A \subset \overline{A}$.
(4) *Every point in $\overline{A}$ is either a limit point of A or an isolated point of A, but not both.*
(5) $\mathbf{x} \in \overline{A}$ *iff every $B_r(\mathbf{x})$ $(r > 0)$ contains a point of A.*

PROOF. *Exercise.* □

Example 1 If $A = \{1, 1/2, 1/3, \ldots, 1/n, \ldots\} \subset \mathbb{R}$, then every point in A is an isolated point. The only limit point is 0.

If $A = (0, 1] \subset \mathbb{R}$ then there are no isolated points, and the set of limit points is $[0, 1]$.

Defining $f_n(t) = t^n$, set $A = \{m^{-1} f_n : m, n \in \mathbb{N}\}$. Then A has only limit point 0 in $(\mathcal{C}[0, 1], \|\cdot\|_\infty)$.

THEOREM 6.3.6. *If $A \subset \mathbb{R}^n$ then*

$$\overline{A} = (ext\ A)^c. \tag{87}$$

$$\overline{A} = int\ A \cup \partial A. \tag{88}$$

$$\overline{A} = A \cup \partial A. \tag{89}$$

PROOF. For (87) first note that $\mathbf{x} \in \overline{A}$ iff every $B_r(\mathbf{x})$ $(r > 0)$ contains at least one member of A. On the other hand, $\mathbf{x} \in$ ext A iff some $B_r(\mathbf{x})$ is a subset of A^c, and so $\mathbf{x} \in (\text{ext } A)^c$ iff it is *not* the case that some $B_r(\mathbf{x})$ is a subset of A^c, i.e. iff every $B_r(\mathbf{x})$ contains at least one member of A.

Equality (88) follows from (87), (84) and the fact that the sets on the right side of (84) are mutually disjoint.

For 89 it is sufficient from (87) to show $A \cup \partial A = (\text{int } A) \cup \partial A$. But clearly $(\text{int } A) \cup \partial A \subset A \cup \partial A$.

On the other hand suppose $\mathbf{x} \in A \cup \partial A$. If $\mathbf{x} \in \partial A$ then $\mathbf{x} \in (\text{int } A) \cup \partial A$, while if $\mathbf{x} \in A$ then $\mathbf{x} \notin$ ext A from the definition of exterior, and so $\mathbf{x} \in (\text{int } A) \cup \partial A$ from (84). Thus $A \cup \partial A \subset (\text{int } A) \cup \partial A$. □

Example 2

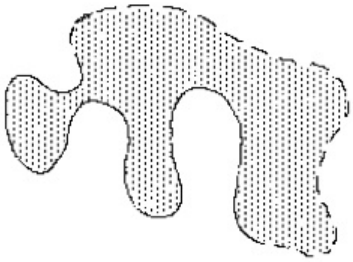

FIGURE 2. A is the shaded region together with the unbroken line, int A is the shaded region, ∂A is the unbroken line together with the broken line, $\overline{A}$ is the shaded region together with the unbroken line together with the broken line.

The following proposition shows that we need to be careful in relying too much on our intuition for $\mathbb{R}^n$ when dealing with an arbitrary metric space.

PROPOSITION 6.3.7. *Let $A = B_r(\mathbf{x}) \subset \mathbb{R}^n$. Then we have int $A = A$, ext $A = \{\mathbf{y} : d(\mathbf{y}, \mathbf{x}) > r\}$, $\partial A = \{\mathbf{y} : d(\mathbf{y}, \mathbf{x}) = r\}$ and $\overline{A} = \{\mathbf{y} : d(\mathbf{y}, \mathbf{x}) \leq r\}$.*

If $A = B_r(x) \subset X$ where (X, d) is an arbitrary metric space, then int $A = A$, ext $A \supset \{y : d(y, x) > r\}$, $\partial A \subset \{y : d(y, x) = r\}$ and $\overline{A} \subset \{y : d(y, x) \leq r\}$. Equality need not hold in the last three cases.

PROOF. We begin with the counterexample to equality. Let $X = \{0, 1\}$ with the metric $d(0, 1) = 1$, $d(0, 0) = d(1, 1) = 0$. Let $A = B_1(0) = \{0\}$. Then *(check)* int $A = A$, ext $A = \{1\}$, $\partial A = \emptyset$ and $\overline{A} = A$.

- **(*int* $A = A$):** Since int $A \subset A$, we need to show every point in A is an interior point. But if $y \in A$, then $d(y, x) = s$(say) $< r$ and $B_{r-s}(y) \subset A$ by the triangle inequality[2],
- **(*ext*$A \supset \{y : d(y, x) > r\}$):** If $d(y, x) > r$, let $d(y, x) = s$. Then $B_{s-r}(y) \subset A^c$ by the triangle inequality (*exercise*), i.e. y is an exterior point of A.
- **(*ext* $A = \{\mathbf{y} : d(\mathbf{y}, \mathbf{x}) > r\}$ *in* $\mathbb{R}^n$):** We have ext $A \supset \{\mathbf{y} : d(\mathbf{y}, \mathbf{x}) > r\}$ from the previous result. If $d(\mathbf{y}, \mathbf{x}) \leq r$ then every $B_s(\mathbf{y})$, where $s > 0$, contains points in A[3]. Hence $\mathbf{y} \notin$ ext A. The result follows.
- **($\partial A \subset \{\mathbf{y} : d(y, x) = r\}$, *with equality for* $\mathbb{R}^n$):** This follows from the previous results and the fact that $\partial A = X \setminus ((\text{int } A) \cup \text{ext } A)$.
- **($\overline{A} \subset \{\mathbf{y} : d(y, x) \leq r\}$, *with equality for* $\mathbb{R}^n$):** This follows from $\overline{A} = A \cup \partial A$ and the previous results.

□

Example If $\mathbb{Q}$ is the set of rationals in $\mathbb{R}$, then int $\mathbb{Q} = \emptyset$, $\partial\mathbb{Q} = \mathbb{R}$, $\overline{\mathbb{Q}} = \mathbb{R}$ and ext $\mathbb{Q} = \emptyset$ (*exercise*).

6.4. Open and Closed Sets

Everything in this section apart from specific examples, applies with $\mathbb{R}^n$ replaced by an arbitrary metric space (X, d).

The concept of an open set is very important in $\mathbb{R}^n$ and more generally is basic to the study of *topology*[4]. We will see later that notions such as connectedness of a set and continuity of a function can be expressed in terms of open sets.

DEFINITION 6.4.1. A set $A \subset \mathbb{R}^n$ is *open* iff $A \subset \text{int}A$.

Remark Thus a set is open iff all its members are interior points. Note that since always int$A \subset A$, it follows that

$$A \text{ is open iff } A = \text{int}A.$$

We usually show a set A is open by proving that *for every $\mathbf{x} \in A$ there exists $r > 0$ such that $B_r(\mathbf{x}) \subset A$* (which is the same as showing that every $\mathbf{x} \in A$ is an interior point of A). Of course, the value of r will depend on x in general.

Note that $\emptyset$ and $\mathbb{R}^n$ are both open sets (for a set to be open, *every* member must be an interior point—since the $\emptyset$ has *no* members it is trivially true that every member of $\emptyset$ is an interior point!). Proposition 6.3.7 shows that $B_r(\mathbf{x})$ is open, thus justifying the terminology of *open* ball.

The following result gives many examples of open sets.

THEOREM 6.4.2. *If $A \subset \mathbb{R}^n$ then int A is open, as is ext A.*

---

[2] If $z \in B_s(y)$ then $d(z, y) < r - s$. But $d(y, x) = s$ and so $d(z, x) \leq d(z, y) + d(y, x) < (r - s) + s = r$, i.e. $d(z, x) < r$ and so $z \in B_r(x)$ as required. *Draw a diagram in* $\mathbb{R}^2$.

[3] Why is this true in $\mathbb{R}^n$? It is not true in an arbitrary metric space, as we see from the counterexample.

[4] There will be courses on elementary topology and algebraic topology in later years. Topological notions are important in much of contemporary mathematics.

PROOF. See Figure 3.

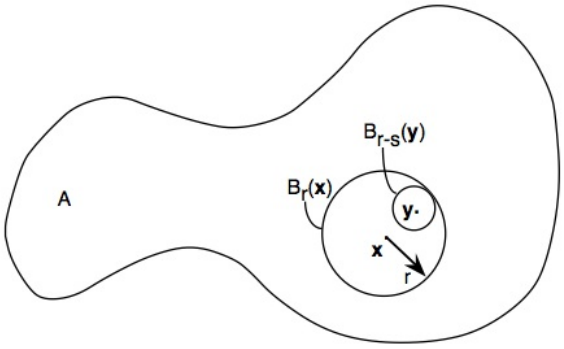

FIGURE 3. Diagram for the proof of Theorem 6.4.2.

Let $A \subset \mathbb{R}^n$ and consider any $\mathbf{x} \in \text{int } A$. Then $B_r(\mathbf{x}) \subset A$ for some $r > 0$. We claim that $B_r(\mathbf{x}) \subset \text{int } A$, thus proving int A is open.

To establish the claim consider any $\mathbf{y} \in B_r(\mathbf{x})$; we need to show that $\mathbf{y}$ is an interior point of A. Suppose $d(\mathbf{y}, \mathbf{x}) = s$ $(< r)$. From the triangle inequality (*exercise*), $B_{r-s}(\mathbf{y}) \subset B_r(\mathbf{x})$, and so $B_{r-s}(\mathbf{y}) \subset A$, thus showing $\mathbf{y}$ is an interior point.

The fact ext A is open now follows from (85). □

Exercise. Suppose $A \subset \mathbb{R}^n$. Prove that the interior of A with respect to the Euclidean metric, and with respect to the sup metric, are the same. *Hint:* First show that each open ball about $\mathbf{x} \in \mathbb{R}^n$ with respect to the Euclidean metric contains an open ball with respect to the sup metric, and conversely.

Deduce that the open sets corresponding to either metric are the same.

The next result shows that finite intersections and arbitrary unions of open sets are open. It is *not* true that an arbitrary intersection of open sets is open. For example, the intervals $(-1/n, 1/n)$ are open for each positive integer n, but $\bigcap_{n=1}^{\infty}(-1/n, 1/n) = \{0\}$ which is not open.

THEOREM 6.4.3. *If $A_1, \ldots, A_k$ are finitely many open sets then $A_1 \cap \cdots \cap A_k$ is also open. If $\{A_\lambda\}_{\lambda \in S}$ is a collection of open sets, then $\bigcup_{\lambda \in S} A_\lambda$ is also open.*

PROOF. Let $A = A_1 \cap \cdots \cap A_k$ and suppose $\mathbf{x} \in A$. Then $\mathbf{x} \in A_i$ for $i = 1, \ldots, k$, and for each i there exists $r_i > 0$ such that $B_{r_i}(\mathbf{x}) \subset A_i$. Let $r = \min\{r_1, \ldots, r_n\}$. Then $r > 0$ and $B_r(\mathbf{x}) \subset A$, implying A is open.

Next let $B = \bigcup_{\lambda \in S} A_\lambda$ and suppose $\mathbf{x} \in B$. Then $\mathbf{x} \in A_\lambda$ for some λ. For some such λ choose $r > 0$ such that $B_r(\mathbf{x}) \subset A_\lambda$. Then certainly $B_r(\mathbf{x}) \subset B$, and so B is open. □

We next define the notion of a *closed* set.

DEFINITION 6.4.4. A set $A \subset \mathbb{R}^n$ is *closed* iff its complement is open.

PROPOSITION 6.4.5. *A set is open iff its complement is closed.*

PROOF. *Exercise.* □

We saw before that a set is open iff it is contained in, and hence equals, its interior. Analogously we have the following result.

THEOREM 6.4.6. *A set A is closed iff $A = \overline{A}$.*

PROOF. A is closed iff A^c is open iff $A^c = \text{int } (A^c)$ iff $A^c = \text{ext } A$ (from (85)) iff $A = \overline{A}$ (taking complements and using (87)). □

Remark Since $A \subset \overline{A}$ it follows from the previous theorem that A is closed iff $\overline{A} \subset A$, i.e. iff A contains all its limit points.

The following result gives many examples of closed sets, analogous to Theorem (6.4.2).

THEOREM 6.4.7. *The sets $\overline{A}$ and ∂A are closed.*

PROOF. Since $\overline{A} = (\text{ext } A)^c$ it follows $\overline{A}$ is closed, $\partial A = (\text{int } A \cup \text{ext } A)^c$, so that ∂A is closed. □

Examples We saw in Proposition 6.3.7 that the set $C = \{\mathbf{y} : |\mathbf{y} - \mathbf{x}| \leq r\}$ in $\mathbb{R}^n$ is the closure of $B_r(\mathbf{x}) = \{\mathbf{y} : |\mathbf{y} - \mathbf{x}| < r\}$, and hence is closed. We also saw that in an arbitrary metric space we only know that $\overline{B_r(\mathbf{x})} \subseteq C$. But it is *always* true that C is closed.

To see this, note that the complement of C is $\{y : d(y, x) > r\}$. This is open since if y is a member of the complement and $d(x, y) = s$ $(> r)$, then $B_{s-r}(y) \subset C^c$ by the triangle inequality (*exercise*).

Similarly, $\{y : d(y, x) = r\}$ is always closed; it contains but need not equal $\partial B_r(\mathbf{x})$.

In particular, the interval $[a, b]$ is closed in $\mathbb{R}$.

Also $\emptyset$ and $\mathbb{R}^n$ are both closed, showing that a set can be both open and closed (these are the only such examples in $\mathbb{R}^n$, why?).

Remark "Most" sets are neither open nor closed. In particular, $\mathbb{Q}$ and $(a, b]$ are neither open nor closed in $\mathbb{R}$.

An analogue of Theorem 6.4.3 holds:

THEOREM 6.4.8. *If $A_1, \ldots, A_n$ are closed sets then $A_1 \cup \cdots \cup A_n$ is also closed. If $A_\lambda (\lambda \in S)$ is a collection of closed sets, then $\bigcap_{\lambda \in S} A_\lambda$ is also closed.*

PROOF. This follows from the previous theorem by DeMorgan's rules. More precisely, if $A = A_1 \cup \cdots \cup A_n$ then $A^c = A_1^c \cap \cdots \cap A_n^c$ and so A^c is open and hence A is closed. A similar proof applies in the case of arbitrary intersections. □

Remark The example $(0, 1) = \bigcup_{n=1}^{\infty} [1/n, 1 - 1/n]$ shows that a non-finite union of closed sets need not be closed.

In $\mathbb{R}$ we have the following description of open sets. A similar result is *not* true in $\mathbb{R}^n$ for $n > 1$ (with intervals replaced by open balls or open n-cubes).

THEOREM 6.4.9. *A set $U \subset \mathbb{R}$ is open iff $U = \bigcup_{i \geq 1} I_i$, where $\{I_i\}$ is a countable (finite or denumerable) family of disjoint open intervals.*

PROOF. *Suppose U is open in $\mathbb{R}$. Let $a \in U$. Since U is open, there exists an open interval I with $a \in I \subset U$. Let I_a be the union of all such open intervals. Since the union of a family of open intervals with a point in common is itself an open interval (*exercise*), it follows that I_a is an open interval. Clearly $I_a \subset U$.

We next claim that any two such intervals I_a and I_b with $a, b \in U$ are either disjoint or equal. For if they have some element in common, then $I_a \cup I_b$ is itself an open interval which is a subset of U and which contains both a and b, and so $I_a \cup I_b \subset I_a$ and $I_a \cup I_b \subset I_b$. Thus $I_a = I_b$.

Thus U is a union of a family $\mathcal{F}$ of disjoint open intervals. To see that $\mathcal{F}$ is countable, for each $I \in \mathcal{F}$ select a rational number in I (this is possible, as there is a rational number between any two real numbers, but does it require the axiom

of choice?). Different intervals correspond to different rational numbers, and so the set of intervals in $\mathcal{F}$ is in one-one correspondence with a subset of the rationals. Thus $\mathcal{F}$ is countable. □

***A Surprising Result** Suppose ϵ is a small positive number (e.g. 10^{-23}). Then there exist disjoint open intervals $I_1, I_2, \ldots$ such that $\mathbb{Q} \subset \bigcup_{i=1}^{\infty} I_i$ and such that $\sum_{i=1}^{\infty} |I_i| \leq \epsilon$ (where $|I_i|$ is the length of I_i)!

To see this, let $r_1, r_2, \ldots$ be an enumeration of the rationals. About each r_i choose an interval J_i of length $\epsilon/2^i$. Then $\mathbb{Q} \subset \bigcup_{i\geq 1} J_i$ and $\sum_{i\geq 1} |J_i| = \epsilon$. However, the J_i are not necessarily mutually disjoint.

We say two intervals J_i and J_j are "connectable" if there is a sequence of intervals $J_{i_1}, \ldots, J_{i_n}$ such that $i_1 = i$, $i_n = j$ and any two consecutive intervals $J_{i_p}, J_{i_{p+1}}$ have non-zero intersection.

Define I_1 to be the union of all intervals connectable to J_1.
Next take the first interval J_i after J_1 which is *not* connectable to J_1 and define I_2 to be the union of all intervals connectable to this J_i.
Next take the first interval J_k after J_i which is not connectable to J_1 or J_i and define I_3 to be the union of all intervals connectable to this J_k.
And so on.

Then one can show that the I_i are mutually disjoint intervals and that $\sum_{j=1}^{\infty} |I_i| \leq \sum_{i=1}^{\infty} |J_i| = \epsilon$.

6.5. Metric Subspaces

DEFINITION 6.5.1. Suppose (X, d) is a metric space and $S \subset X$. Then the *metric subspace* corresponding to S is the metric space (S, d_S), where

$$d_S(x, y) = d(x, y). \tag{90}$$

The metric d_S (often just denoted d) is called the *induced metric* on S [5].

It is easy (*exercise*) to see that the axioms for a metric space do indeed hold for (S, d_S).

Examples

(1) The sets $[a, b]$, $(a, b]$ and $\mathbb{Q}$ all define metric subspaces of $\mathbb{R}$.
(2) Consider $\mathbb{R}^2$ with the usual Euclidean metric. We can identify $\mathbb{R}$ with the "x-axis" in $\mathbb{R}^2$, more precisely with the subset $\{(x, 0) : x \in \mathbb{R}\}$, via the map $x \mapsto (x, 0)$. The Euclidean metric on $\mathbb{R}$ then corresponds to the induced metric on the x-axis.

Since a metric subspace (S, d_S) is a metric space, the definitions of open ball; of interior, exterior, boundary and closure of a set; and of open set and closed set; all apply to (S, d_S).

There is a simple relationship between an open ball about a point in a metric subspace and the corresponding open ball in the original metric space.

PROPOSITION 6.5.2. *Suppose (X, d) is a metric space and (S, d) is a metric subspace. Let $a \in S$. Let the open ball in S of radius r about a be denoted by $B_r^S(a)$. Then*

$$B_r^S(a) = S \cap B_r(a).$$

[5] There is no connection between the notions of a metric subspace and that of a vector subspace! For example, *every* subset of $\mathbb{R}^n$ defines a metric subspace, but this is certainly not true for vector subspaces.

PROOF.

$$\begin{aligned} B_r^S(a) &:= \{x \in S : d_S(x,a) < r\} = \{x \in S : d(x,a) < r\} \\ &= S \cap \{x \in X : d(x,a) < r\} = S \cap B_r(a). \end{aligned}$$

□

The symbol ":=" means "by definition, is equal to".

There is also a simple relationship between the open (closed) sets in a metric subspace and the open (closed) sets in the original space.

THEOREM 6.5.3. *Suppose (X,d) is a metric space and (S,d) is a metric subspace. Then for any $A \subset S$:*

(1) *A is open in S iff $A = S \cap U$ for some set $U\,(\subset X)$ which is open in X.*
(2) *A is closed in S iff $A = S \cap C$ for some set $C\,(\subset X)$ which is closed in X.*

PROOF. (i) Suppose that $A = S \cap U$, where $U\,(\subset X)$ is open in X. Then for each $a \in A$ (since $a \in U$ and U is open in X) there exists $r > 0$ such that $B_r(a) \subset U$. Hence $S \cap B_r(a) \subset S \cap U$, i.e. $B_r^S(a) \subset A$ as required.

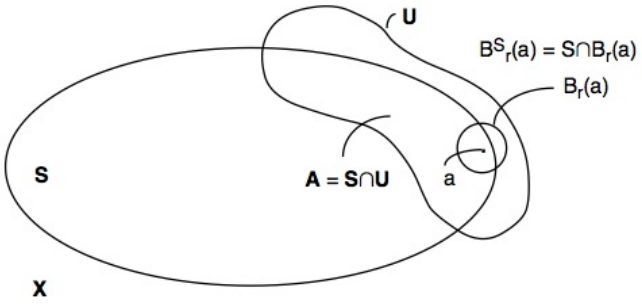

FIGURE 4. Diagram for proof of Theorem 6.5.3.

(ii) Next suppose A is open in S. Then for each $a \in A$ there exists $r = r_a > 0$[6] such that $B_{r_a}^S(a) \subset A$, i.e. $S \cap B_{r_a}(a) \subset A$. Let $U = \bigcup_{a \in A} B_{r_a}(a)$. Then U is open in X, being a union of open sets.

We claim that $A = S \cap U$. Now

$$S \cap U = S \cap \bigcup_{a \in A} B_{r_a}(a) = \bigcup_{a \in A} (S \cap B_{r_a}(a)) = \bigcup_{a \in A} B_{r_a}^S(a).$$

But $B_{r_a}^S(a) \subset A$, and for each $a \in A$ we trivially have that $a \in B_{r_a}^S(a)$. Hence $S \cap U = A$ as required.

The result for closed sets follow from the results for open sets together with DeMorgan's rules.

(iii) First suppose $A = S \cap C$, where $C\,(\subset X)$ is closed in X. Then $S \setminus A = S \cap C^c$ from elementary properties of sets. Since C^c is open in X, it follows from (1) that $S \setminus A$ is open in S, and so A is closed in S.

(iv) Finally suppose A is closed in S. Then $S \setminus A$ is open in S, and so from (1), $S \setminus A = S \cap U$ where $U \subset X$ is open in X. From elementary properties of sets it follows that $A = S \cap U^c$. But U^c is closed in X, and so the required result follows. □

Examples

[6] We use the notation $r = r_a$ to indicate that r depends on a.

(1) Let $S = (0, 2]$. Then $(0, 1)$ and $(1, 2]$ are both open *in* S (*why?*), but $(1, 2]$ is *not* open *in* $\mathbb{R}$. Similarly, $(0, 1]$ and $[1, 2]$ are both closed *in* S (*why?*), but $(0, 1]$ is *not* closed *in* $\mathbb{R}$.
(2) Consider $\mathbb{R}$ as a subset of $\mathbb{R}^2$ by identifying $x \in \mathbb{R}$ with $(x, 0) \in \mathbb{R}^2$. Then $\mathbb{R}$ is open and closed as a subset of itself, but is closed (and not open) as a subset of $\mathbb{R}^2$.
(3) Note that $[-\sqrt{2}, \sqrt{2}] \cap \mathbb{Q} = (-\sqrt{2}, \sqrt{2}) \cap \mathbb{Q}$. It follows that $\mathbb{Q}$ has many clopen sets.

CHAPTER 7

Sequences and Convergence

In this chapter you should initially think of the cases $X = \mathbb{R}$ and $X = \mathbb{R}^n$.

7.1. Notation

If X is a set and $x_n \in X$ for $n = 1, 2, \ldots$, then $(x_1, x_2, \ldots)$ is called a *sequence in* X and x_n is called the *nth term* of the sequence. We also write $x_1, x_2, \ldots$, or $(x_n)_{n=1}^{\infty}$, or just (x_n), for the sequence.

NB Note the difference between (x_n) and $\{x_n\}$.

More precisely, a sequence in X is a function $f : \mathbb{N} \to X$, where $f(n) = x_n$ with x_n as in the previous notation.

We write $(x_n)_{n=1}^{\infty} \subset X$ or $(x_n) \subset X$ to indicate that all terms of the sequence are members of X. Sometimes it is convenient to write a sequence in the form $(x_p, x_{p+1}, \ldots)$ for some (possible negative) integer $p \neq 1$.

Given a sequence (x_n), a *subsequence* is a sequence (x_{n_i}) where (n_i) is a strictly increasing sequence in $\mathbb{N}$.

7.2. Convergence of Sequences

DEFINITION 7.2.1. Suppose (X, d) is a metric space, $(x_n) \subset X$ and $x \in X$. Then we say the sequence (x_n) *converges* to x, written $x_n \to x$, if for every $r > 0$ [1] there exists an integer N such that

$$n \geq N \;\Rightarrow\; d(x_n, x) < r. \tag{91}$$

Thus $x_n \to x$ if for every open ball $B_r(x)$ centred at x the sequence (x_n) is *eventually* contained in $B_r(x)$. The "smaller" the ball, i.e. the smaller the value of r, the larger the value of N required for (91) to be true, as we see in the following diagram for three different balls centred at x. Although for each $r > 0$ there will be a *least* value of N such that (91) is true, this *particular* value of N is rarely of any significance.

[1] It is sometimes convenient to replace r by ϵ, to remind us that we are interested in *small* values of r (or ϵ).

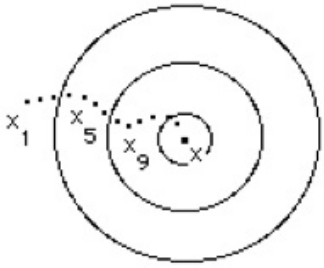

FIGURE 1. The sequence x_n converges to x.

Remark The notion of convergence in a metric space can be reduced to the notion of convergence in $\mathbb{R}$, since the above definition says $x_n \to x$ iff $d(x_n, x) \to 0$, and the latter is just convergence of a sequence of real numbers.

Examples

(1) Let $\theta \in \mathbb{R}$ be fixed, and set

$$x_n = \left(a + \frac{1}{n}\cos n\theta,\ b + \frac{1}{n}\sin n\theta\right) \in \mathbb{R}^2.$$

Then $x_n \to (a, b)$ as $n \to \infty$. The sequence (x_n) "spirals" around the point (a, b), with $d\,(x_n,\,(a, b)) = 1/n$, and with a rotation by the angle θ in passing from x_n to x_{n+1}.

(2) Let

$$(x_1, x_2, \ldots) = (1, 1, \ldots).$$

Then $x_n \to 1$ as $n \to \infty$.

(3) Let

$$(x_1, x_2, \ldots) = (1, \frac{1}{2}, 1, \frac{1}{3}, 1, \frac{1}{4}, \ldots).$$

Then it is *not* the case that $x_n \to 0$ and it is *not* the case that $x_n \to 1$. The sequence (x_n) does not converge.

(4) Let $A \subset \mathbb{R}$ be bounded above and suppose $a = \operatorname{lub} A$. Then there exists $(x_n) \subset A$ such that $x_n \to a$.

PROOF. Suppose n is a natural number. By the definition of least upper bound, $a - 1/n$ is *not* an upper bound for A. Thus there exists an $x \in A$ such that $a - 1/n < x \leq a$. Choose some such x and denote it by x_n. Then $x_n \to a$ since $d(x_n, a) < 1/n$ [2]. □

(5) As an indication of "strange" behaviour, for the p-adic metric on $\mathbb{Z}$ we have $p^n \to 0$.

Series An infinite series $\sum_{n=1}^{\infty} x_n$ of terms from $\mathbb{R}$ (more generally, from $\mathbb{R}^n$ or from a normed space) is just a certain type of sequence. More precisely, for each i we define the nth *partial sum* by

$$s_n = x_1 + \cdots + x_n.$$

Then we say the series $\sum_{n=1}^{\infty} x_n$ *converges* iff the sequence (of partial sums) (s_n) converges, and in this case the limit of (s_n) is called the *sum* of the series.

NB Note that changing the order (re-indexing) of the (x_n) gives rise to a possibly different sequence of partial sums (s_n).

[2] Note the implicit use of the axiom of choice to form the sequence (x_n).

Example

If $0 < r < 1$ then the geometric series $\sum_{n=0}^{\infty} r^n$ converges to $(1-r)^{-1}$.

7.3. Elementary Properties

THEOREM 7.3.1. *A sequence in a metric space can have at most one limit.*

PROOF. Suppose (X, d) is a metric space, $(x_n) \subset X$, $x, y \in X$, $x_n \to x$ as $n \to \infty$, and $x_n \to y$ as $n \to \infty$.

Supposing $x \neq y$, let $d(x, y) = r > 0$. From the definition of convergence there exist integers N_1 and N_2 such that

$$\begin{aligned} n \geq N_1 &\Rightarrow d(x_n, x) < r/4, \\ n \geq N_2 &\Rightarrow d(x_n, y) < r/4. \end{aligned}$$

Let $N = \max\{N_1, N_2\}$. Then

$$\begin{aligned} d(x, y) &\leq d(x, x_N) + d(x_N, y) \\ &< r/4 + r/4 = r/2, \end{aligned}$$

i.e. $d(x, y) < r/2$, which is a contradiction. □

DEFINITION 7.3.2. A sequence is *bounded* if the set of terms from the sequence is bounded.

THEOREM 7.3.3. *A convergent sequence in a metric space is bounded.*

PROOF. Suppose that (X, d) is a metric space, $(x_n)_{n=1}^{\infty} \subset X$, $x \in X$ and $x_n \to x$. Let N be an integer such that $n \geq N$ implies $d(x_n, x) \leq 1$.

Let $r = \max\{d(x_1, x), \ldots, d(x_{N-1}, x), 1\}$ (this is finite since r is the maximum of a *finite* set of numbers). Then $d(x_n, x) \leq r$ for *all* n and so $(x_n)_{n=1}^{\infty} \subset B_{r+1/10}(x)$. Thus (x_n) is bounded, as required. □

Remark This method, of using convergence to handle the 'tail' of the sequence, and some separate argument for the *finitely* many terms not in the tail, is of fundamental importance.

The following result on the distance function is useful. As we will see in Chapter 11, it says that the distance function is continuous.

THEOREM 7.3.4. *Let $x_n \to x$ and $y_n \to y$ in a metric space (X, d). Then $d(x_n, y_n) \to d(x, y)$.*

PROOF. Two applications of the triangle inequality show that

$$d(x, y) \leq d(x, x_n) + d(x_n, y_n) + d(y_n, y),$$

and so

$$d(x, y) - d(x_n, y_n) \leq d(x, x_n) + d(y_n, y). \tag{92}$$

Similarly

$$d(x_n, y_n) \leq d(x_n, x) + d(x, y) + d(y, y_n),$$

and so

$$d(x_n, y_n) - d(x, y) \leq d(x, x_n) + d(y_n, y). \tag{93}$$

It follows from (92) and (93) that

$$|d(x, y) - d(x_n, y_n)| \leq d(x, x_n) + d(y_n, y).$$

Since $d(x, x_n) \to 0$ and $d(y, y_n) \to 0$, the result follows immediately from properties of sequences of real numbers (or see the Comparison Test in the next section). □

7.4. Sequences in $\mathbb{R}$

The results in this section are particular to sequences in $\mathbb{R}$. They do not even make sense in a general metric space.

DEFINITION 7.4.1. A sequence $(x_n)_{n=1}^{\infty} \subset \mathbb{R}$ is

(1) *increasing* (or *non-decreasing*) if $x_n \leq x_{n+1}$ for all n,
(2) *decreasing* (or *non-increasing*) if $x_{n+1} \geq x_n$ for all n,
(3) *strictly increasing* if $x_n < x_{n+1}$ for all n,
(4) *strictly decreasing* if $x_{n+1} < x_n$ for all n.

A sequence is *monotone* if it is either increasing or decreasing.

The following theorem uses the Completeness Axiom in an essential way. It is not true if we replace $\mathbb{R}$ by $\mathbb{Q}$. For example, consider the sequence of rational numbers obtained by taking the decimal expansion of $\sqrt{2}$; i.e. x_n is the decimal approximation to $\sqrt{2}$ to n decimal places.

THEOREM 7.4.2. *Every bounded monotone sequence in $\mathbb{R}$ has a limit in $\mathbb{R}$.*

PROOF. Suppose $(x_n)_{n=1}^{\infty} \subset \mathbb{R}$ and (x_n) is increasing (if (x_n) is decreasing, the argument is analogous). Since the *set* of terms $\{x_1, x_2, \ldots\}$ is bounded above, it has a least upper bound x, say. We claim that $x_n \to x$ as $n \to \infty$.

To see this, note that $x_n \leq x$ for all n; but if $\epsilon > 0$ then $x_k > x - \epsilon$ for some k, as otherwise $x - \epsilon$ would be an upper bound. Choose such $k = k(\epsilon)$. Since $x_k > x - \epsilon$, then $x_n > x - \epsilon$ for all $n \geq k$ as the sequence is increasing. Hence

$$x - \epsilon < x_n \leq x$$

for all $n \geq k$. Thus $|x - x_n| < \epsilon$ for $n \geq k$, and so $x_n \to x$ (since $\epsilon > 0$ is arbitrary). □

It follows that a bounded closed set in $\mathbb{R}$ contains its infimum and supremum, which are thus the minimum and maximum respectively.

For sequences $(x_n) \subset \mathbb{R}$ it is also convenient to define the notions $x_n \to \infty$ and $x_n \to -\infty$ as $n \to \infty$.

DEFINITION 7.4.3. If $(x_n) \subset \mathbb{R}$ then $x_n \to \infty$ $(-\infty)$ as $n \to \infty$ if for every positive real number M there exists an integer N such that

$$n \geq N \text{ implies } x_n > M \ (x_n < -M).$$

We say (x_n) has limit ∞ $(-\infty)$ and write $\lim_{n\to\infty} x_n = \infty(-\infty)$.

Note When we say a sequence (x_n) converges, we usually mean that $x_n \to x$ for some $x \in \mathbb{R}$; i.e. we do *not* allow $x_n \to \infty$ or $x_n \to -\infty$.

The following *Comparison Test* is easy to prove (*exercise*). Notice that the assumptions $x_n < y_n$ for all n, $x_n \to x$ and $y_n \to y$, do *not* imply $x < y$. For example, let $x_n = 0$ and $y_n = 1/n$ for all n.

THEOREM 7.4.4 (Comparison Test).

(1) *If $0 \leq x_n \leq y_n$ for all $n \geq N$, and $y_n \to 0$ as $n \to \infty$, then $x_n \to 0$ as $n \to \infty$.*
(2) *If $x_n \leq y_n$ for all $n \geq N$, $x_n \to x$ as $n \to \infty$ and $y_n \to y$ as $n \to \infty$, then $x \leq y$.*
(3) *In particular, if $x_n \leq a$ for all $n \geq N$ and $x_n \to x$ as $n \to \infty$, then $x \leq a$.*

∎

Example Let $x_m = (1+1/m)^m$ and $y_m = 1+1+1/2!+\cdots+1/m!$

The sequence (y_m) is increasing and for each $m \in \mathbb{N}$,

$$y_m \leq 1+1+\frac{1}{2}+\frac{1}{2^2}+\cdots\frac{1}{2^m} < 3.$$

Thus $y_m \to y_0(\text{say}) \leq 3$, from Theorem 7.4.2.

From the binomial theorem,

$$x_m = 1+m\frac{1}{m}+\frac{m(m-1)}{2!}\frac{1}{m^2}+\frac{m(m-1)(m-2)}{3!}\frac{1}{m^3}+\cdots+\frac{m!}{m!}\frac{1}{m^m}.$$

This can be written as

$$\begin{aligned} x_m &= 1+1+\frac{1}{2!}\left(1-\frac{1}{m}\right)+\frac{1}{3!}\left(1-\frac{1}{m}\right)\left(1-\frac{2}{m}\right) \\ &\quad +\cdots+\frac{1}{m!}\left(1-\frac{1}{m}\right)\left(1-\frac{2}{m}\right)\cdots\left(1-\frac{m-1}{m}\right). \end{aligned}$$

It follows that $x_m \leq x_{m+1}$, since there is one extra term in the expression for x_{m+1} and the other terms (after the first two) are larger for x_{m+1} than for x_m. Clearly $x_m \leq y_m$ $(\leq y_0 \leq 3)$. Thus the sequence (x_m) has a limit x_0 (say) by Theorem 7.4.2. Moreover, $x_0 \leq y_0$ from the Comparison test.

In fact $x_0 = y_0$ and is usually denoted by $e(=2.71828...)$[3]. It is the base of the natural logarithms.

Example Let $z_n = \sum_{k=1}^{n}\frac{1}{k} - \log n$. Then (z_n) is monotonically decreasing and $0 < z_n < 1$ for all n. Thus (z_n) has a limit γ say. This is Euler's constant, and $\gamma = 0.577\ldots$. It arises when considering the Gamma function:

$$\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}dt = \frac{e^{\gamma z}}{z}\prod_1^\infty(1+\frac{1}{n})^{-1}e^{z/n}$$

For $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$.

7.5. Sequences and Components in $\mathbb{R}^k$

The result in this section is particular to sequences in $\mathbb{R}^n$ and does not apply (or even make sense) in a general metric space.

THEOREM 7.5.1. *A sequence $(\mathbf{x}_n)_{n=1}^\infty$ in $\mathbb{R}^n$ converges iff the corresponding sequences of components (x_n^i) converge for $i = 1,\ldots,k$. Moreover,*

$$\lim_{n\to\infty}\mathbf{x}_n = (\lim_{n\to\infty}x_n^1,\ldots,\lim_{n\to\infty}x_n^k).$$

PROOF. Suppose $(\mathbf{x}_n) \subset \mathbb{R}^n$ and $\mathbf{x}_n \to \mathbf{x}$. Then $|\mathbf{x}_n - \mathbf{x}| \to 0$, and since $|x_n^i - x^i| \leq |\mathbf{x}_n - \mathbf{x}|$ it follows from the Comparison Test that $x_n^i \to x^i$ as $n \to \infty$, for $i = 1,\ldots,k$.

Conversely, suppose that $x_n^i \to x^i$ for $i = 1,\ldots,k$. Then for any $\epsilon > 0$ there exist integers $N_1,\ldots,N_k$ such that

$$n \geq N_i \;\Rightarrow\; |x_n^i - x^i| < \epsilon$$

for $i = 1,\ldots,k$.

Since

$$|\mathbf{x}_n - \mathbf{x}| = \left(\sum_{i=1}^k |x_n^i - x^i|^2\right)^{\frac{1}{2}},$$

it follows that if $N = \max\{N_1,\ldots,N_k\}$ then

$$n \geq N \;\Rightarrow\; |\mathbf{x}_n - \mathbf{x}| < \sqrt{k\epsilon^2} = \sqrt{k}\epsilon.$$

[3] See the Problems.

Since $\epsilon > 0$ is otherwise arbitrary[4], the result is proved. □

7.6. Sequences and the Closure of a Set

The following gives a useful characterisation of the closure of a set in terms of convergent sequences.

THEOREM 7.6.1. *Let X be a metric space and let $A \subset X$. Then $x \in \overline{A}$ iff there is a sequence $(x_n)_{n=1}^{\infty} \subset A$ such that $x_n \to x$.*

PROOF. If $(x_n) \subset A$ and $x_n \to x$, then for every $r > 0$, $B_r(x)$ must contain some term from the sequence. Thus $x \in \overline{A}$ from Definition (6.3.3) of a limit point.

Conversely, if $x \in \overline{A}$ then (again from Definition (6.3.3)), $B_{1/n}(x) \cap A \neq \emptyset$ for each $n \in N$. Choose $x_n \in B_{1/n}(x) \cap A$ for $n = 1, 2, \ldots$. Then $(x_n)_{n=1}^{\infty} \subset A$. Since $d(x_n, x) \leq 1/n$ it follows $x_n \to x$ as $n \to \infty$. □

COROLLARY 7.6.2. *Let X be a metric space and let $A \subset X$. Then A is closed in X iff*

$$(x_n)_{n=1}^{\infty} \subset A \text{ and } x_n \to x \text{ implies } x_n \in A. \tag{94}$$

PROOF. From the theorem, (94) is true iff $\overline{A} = A$, i.e. iff A is closed. □

Remark Thus in a metric space X the closure of a set A equals the set of all limits of sequences whose members come from A. And this is generally *not* the same as the set of limit points of A (which points will be missed?). The set A is closed iff it is "closed" under the operation of taking limits of convergent sequences of elements from A[5].

Exercise Let $A = \{(\frac{n}{m}, \frac{1}{n}) : m, n \in \mathbb{N}\}$. Determine $\overline{A}$.

Exercise Use Corollary 7.6.2 to show directly that the closure of a set is indeed closed.

7.7. Algebraic Properties of Limits

The important cases in this section are $X = \mathbb{R}$, $X = \mathbb{R}^n$ and X is a function space such as $\mathcal{C}[a, b]$. The proofs are essentially the same as in the case $X = \mathbb{R}$. We need X to be a normed space (or an inner product space for the third result in the next theorem) so that the algebraic operations make sense.

THEOREM 7.7.1. *Let $(x_n)_{n=1}^{\infty}$ and $(y_n)_{n=1}^{\infty}$ be convergent sequences in a normed space X, and let α be a scalar. Let $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty} y_n = y$. Then the following limits exist with the values stated:*

$$\lim_{n\to\infty} (x_n + y_n) = x + y, \tag{95}$$

$$\lim_{n\to\infty} \alpha x_n = \alpha x. \tag{96}$$

More generally, if also $\alpha_n \to \alpha$, then

$$\lim_{n\to\infty} \alpha_n x_n = \alpha x. \tag{97}$$

---

[4]More precisely, to be consistent with the definition of convergence, we could replace ϵ throughout the proof by $\epsilon/\sqrt{k}$ and so replace $\epsilon\sqrt{k}$ on the last line of the proof by ϵ. We would not normally bother doing this.

[5]There is a possible inconsistency of terminology here. The *sequence* $(1, 1+1, 1/2, 1+1/2, 1/3, 1+1/3, \ldots, 1/n, 1+1/n, \ldots)$ has no limit; the *set* $A = \{1, 1+1, 1/2, 1+1/2, 1/3, 1+1/3, \ldots, 1/n, 1+1/n, \ldots\}$ has the two limit points 0 and 1; and the closure of A, i.e. the set of limit points of sequences whose members belong to A, is $A \cup \{0, 1\}$.

Exercise: What is a sequence with members from A converging to 0, to 1, to 1/3?

If X is an inner product space then also

$$\lim_{n\to\infty} x_n \cdot y_n = x \cdot y. \tag{98}$$

PROOF. Suppose $\epsilon > 0$. Choose N_1, N_2 so that

$$\begin{aligned} ||x_n - x|| &< \epsilon \quad \text{if } n \geq N_1, \\ ||y_n - y|| &< \epsilon \quad \text{if } n \geq N_2. \end{aligned}$$

Let $N = \max\{N_1, N_2\}$.

(i) If $n \geq N$ then

$$\begin{aligned} ||(x_n + y_n) - (x + y)|| &= ||(x_n - x) + (y_n - y)|| \\ &\leq ||x_n - x|| + ||y_n - y|| \\ &< 2\epsilon. \end{aligned}$$

Since $\epsilon > 0$ is arbitrary, the first result is proved.

(ii) If $n \geq N_1$ then

$$\begin{aligned} ||\alpha x_n - \alpha x|| &= |\alpha|\,||x_n - x|| \\ &\leq |\alpha|\epsilon. \end{aligned}$$

Since $\epsilon > 0$ is arbitrary, this proves the second result.

(iii) For the fourth claim, we have for $n \geq N$

$$\begin{aligned} |x_n \cdot y_n - x \cdot y| &= |(x_n - x) \cdot y_n + x \cdot (y_n - y)| \\ &\leq |(x_n - x) \cdot y_n| + |x \cdot (y_n - y)| \\ &\leq |x_n - x|\,|y_n| + |x|\,|y_n - y| \quad \text{by Cauchy-Schwarz} \\ &\leq \epsilon|y_n| + |x|\epsilon, \end{aligned}$$

Since (y_n) is convergent, it follows from Theorem 7.3.3 that $|y_n| \leq M_1$ (say) for all n. Setting $M = \max\{M_1, |x|\}$ it follows that

$$|x_n \cdot y_n - x \cdot y| \leq 2M\epsilon$$

for $n \geq N$. Again, since $\epsilon > 0$ is arbitrary, we are done.

(iv) The third claim is proved like the fourth (*exercise*). □

CHAPTER 8

Cauchy Sequences

8.1. Cauchy Sequences

Our definition of convergence of a sequence $(x_n)_{n=1}^{\infty}$ refers not only to the sequence but also to the limit. But it is often not possible to know the limit *a priori*, (see example in Section 7.4). We would like, if possible, a criterion for convergence which does not depend on the limit itself. We have already seen that a bounded *monotone* sequence in $\mathbb{R}$ converges, but this is a very special case.

Theorem 8.1.3 below gives a necessary and sufficient criterion for convergence of a sequence in $\mathbb{R}^n$, due to Cauchy (1789–1857), which does not refer to the actual limit. We discuss the generalisation to sequences in other metric spaces in the next section.

DEFINITION 8.1.1. Let $(x_n)_{n=1}^{\infty} \subset X$ where (X, d) is a metric space. Then (x_n) is a *Cauchy sequence* if for every $\epsilon > 0$ there exists an integer N such that

$$m, n \geq N \;\Rightarrow\; d(x_m, x_n) < \epsilon.$$

We sometimes write this as $d(x_m, x_n) \to 0$ as $m, n \to \infty$.

Thus a sequence is Cauchy if, for each $\epsilon > 0$, beyond a certain point in the sequence *all* the terms are within distance ϵ of one another.

Warning This is stronger than claiming that, beyond a certain point in the sequence, *consecutive* terms are within distance ϵ of one another.

For example, consider the sequence $x_n = \sqrt{n}$. Then

$$\begin{aligned} |x_{n+1} - x_n| &= \left(\sqrt{n+1} - \sqrt{n}\right) \frac{\sqrt{n+1} + \sqrt{n}}{\sqrt{n+1} + \sqrt{n}} \\ &= \frac{1}{\sqrt{n+1} + \sqrt{n}}. \end{aligned}$$

(99)

Hence $|x_{n+1} - x_n| \to 0$ as $n \to \infty$.

But $|x_m - x_n| = \sqrt{m} - \sqrt{n}$ if $m > n$, and so for any N we can choose $n = N$ and $m > N$ such that $|x_m - x_n|$ is as large as we wish. Thus the sequence (x_n) is *not* Cauchy.

THEOREM 8.1.2. *In a metric space, every convergent sequence is a Cauchy sequence.*

PROOF. Let (x_n) be a convergent sequence in the metric space (X, d), and suppose $x = \lim x_n$.

Given $\epsilon > 0$, choose N such that

$$n \geq N \Rightarrow d(x_n, x) < \epsilon.$$

It follows that for any $m, n \geq N$

$$\begin{aligned} d(x_m, x_n) &\leq d(x_m, x) + d(x, x_n) \\ &\leq 2\epsilon. \end{aligned}$$

Thus (x_n) is Cauchy. □

Remark The converse of the theorem is true in $\mathbb{R}^n$, as we see in the next theorem, but is not true in general. For example, a Cauchy sequence from $\mathbb{Q}$ (with the usual metric) will not necessarily converge to a limit *in* $\mathbb{Q}$ (take the usual example of the sequence whose nth term is the n place decimal approximation to $\sqrt{2}$). We discuss this further in the next section.

Cauchy implies Bounded A Cauchy sequence in any metric space is bounded. This simple result is proved in a similar manner to the corresponding result for convergent sequences (*exercise*).

THEOREM 8.1.3 (Cauchy). *A sequence in $\mathbb{R}^n$ converges (to a limit in $\mathbb{R}^n$) iff it is Cauchy.*

PROOF. We have already seen that a convergent sequence in *any* metric space is Cauchy.

Assume for the converse that the sequence $(\mathbf{x}_n)_{n=1}^{\infty} \subset \mathbb{R}^n$ is Cauchy.

The Case $k = 1$: We will show that if $(x_n) \subset \mathbb{R}$ is Cauchy then it is convergent by showing that it converges to the same limit as an associated monotone increasing sequence (y_n).

Define

$$y_n = \inf\{x_n, x_{n+1}, \ldots\}$$

for each $n \in \mathbb{N}$ [1]. It follows that $y_{n+1} \geq y_n$ since y_{n+1} is the infimum over a subset of the set corresponding to y_n. Moreover the sequence (y_n) is bounded since the sequence (x_n) is Cauchy and hence bounded (if $|x_n| \leq M$ for all n then also $|y_n| \leq M$ for all n).

From Theorem 7.4.2 on monotone sequences, $y_n \to a$ (say) as $n \to \infty$. We will prove that also $x_n \to a$.

Suppose $\epsilon > 0$. Since (x_n) is Cauchy there exists $N = N(\epsilon)$ [2] such that

$$x_n - \epsilon \leq x_m \leq x_\ell + \epsilon \tag{100}$$

for all $\ell, m, n \geq N$.

Claim:

$$y_n - \epsilon \leq x_m \leq y_n + \epsilon \tag{101}$$

for all $m, n \geq N$.

To establish the first inequality in (101), note that from the first inequality in (100) we immediately have for $m, n \geq N$ that

$$y_n - \epsilon \leq x_m,$$

since $y_n \leq x_n$.

To establish the second inequality in (101), note that from the second inequality in (100) we have for $\ell, m \geq N$ that

$$x_m \leq x_\ell + \epsilon\,.$$

But $y_n + \epsilon = \inf\{x_\ell + \epsilon : \ell \geq n\}$ as is easily seen[3]. Thus $y_n + \epsilon$ is greater or equal to any other lower bound for $\{x_n + \epsilon, x_{n+1} + \epsilon, \ldots\}$, whence

$$x_m \leq y_n + \epsilon.$$

It now follows from the Claim, by fixing m and letting $n \to \infty$, and from the Comparison Test (Theorem 7.4.4) that

$$a - \epsilon \leq x_m \leq a + \epsilon$$

[1] You may find it useful to think of an example such as $x_n = (-1)^n/n$.

[2] The notation $N(\epsilon)$ is just a way of noting that N depends on ϵ.

[3] If $y = \inf S$, then $y + \alpha = \inf\{x + \alpha : x \in S\}$.

for all $m \geq N = N(\epsilon)$.

Since $\epsilon > 0$ is arbitrary, it follows that $x_m \to a$ as $m \to \infty$. This finishes the proof in the case $k = 1$.

The Case $k > 1$: If $(\mathbf{x}_n)_{n=1}^{\infty} \subset \mathbb{R}^n$ is Cauchy it follows easily that each sequence of components is also Cauchy, since

$$|\mathbf{x}_n^i - \mathbf{y}_n^i| \leq |\mathbf{x}_n - \mathbf{y}_n|$$

for $i = 1, \ldots, k$. From the case $k = 1$ it follows $\mathbf{x}_n^i \to a^i$ (say) for $i = 1, \ldots, k$. Then from Theorem 7.5.1 it follows $\mathbf{x}_n \to \mathbf{a}$ where $\mathbf{a} = (a_1, \ldots, a_n)$. □

Remark In the case $k = 1$, and for any bounded sequence, the number a constructed above,

$$a = \sup_n \inf\{x_i : i \geq n\}$$

is denoted $\liminf x_n$ or $\underline{\lim}\, x_n$. It is the least of the limits of the subsequences of $(x_n)_{n=1}^{\infty}$ (why?).[4] One can analogously define $\limsup x_n$ or $\overline{\lim}\, x_n$ (*exercise*).

8.2. Complete Metric Spaces

DEFINITION 8.2.1. A metric space (X, d) is *complete* if every Cauchy sequence in X has a limit *in* X.

If a normed space is complete with respect to the associated metric, it is called a *complete normed space* or a *Banach space*.

We have seen that $\mathbb{R}^n$ is complete, but that $\mathbb{Q}$ is not complete. The next theorem gives a simple criterion for a subset of $\mathbb{R}^n$ (with the standard metric) to be complete.

Examples

(1) We will see in Corollary 12.3.5 that $\mathcal{C}[a, b]$ (see Section 5.1 for the definition) is a Banach space with respect to the sup metric. The same argument works for $\ell^{\infty}(\mathbb{N})$.

On the other hand, $\mathcal{C}[a, b]$ with respect to the L^1 norm (see 72) is *not* complete. For example, let $f_n \in \mathcal{C}[-1, 1]$ be defined by

$$f_n(x) = \begin{cases} 0 & -1 \leq x \leq 0 \\ nx & 0 \leq x \leq \frac{1}{n} \\ 1 & \frac{1}{n} \leq x \leq 1 \end{cases}$$

Then there is no $f \in \mathcal{C}[-1, 1]$ such that $||f_n - f||_{L^1} \to 0$, i.e. such that $\int_{-1}^{1} |f_n - f| \to 0$. (If there were such an f, then we would have to have $f(x) = 0$ if $-1 \leq x < 0$ and $f(x) = 1$ if $0 < x \leq 1$ (why?). But such an f cannot be continuous on $[-1, 1]$.)

The same example shows that $\mathcal{C}[a, b]$ with respect to the L^2 norm (see 73) is not complete.

(2) Take $X = \mathbb{R}$ with metric

$$d(x, y) = \left| \frac{x}{1 + |x|} - \frac{y}{1 + |y|} \right|.$$

(*Exercise*) show that this is indeed a metric.

If $(x_n) \subset \mathbb{R}$ with $|x_n - x| \to 0$, then certainly $d(x_n, x) \to 0$. Not so obviously, the converse is also true. But whereas $(\mathbb{R}, |\cdot|)$ is complete,

[4] Note that the limit of a subsequence of (x_n) may not be a limit point of the *set* $\{x_n\}$.

(X, d) is not, as consideration of the sequence $x_n = n$ easily shows. Define $Y = \mathbb{R} \cup \{-\infty, \infty\}$, and extend d by setting

$$d(\pm\infty, x) = \left| \frac{x}{1 + |x|} - (\pm 1) \right|, \quad d(-\infty, \infty) = 2$$

for $x \in \mathbb{R}$. Then (Y, d) is complete (*exercise*).

Remark The second example here utilizes the following simple fact. Given a set X, a metric space (Y, d_Y) and a suitable function $f : X \to Y$, the function $d_X(x, x') = d_Y(f(x), f(x'))$ is a metric on X. (*Exercise* : what does suitable mean here?) The cases $X = (0, 1)$, $Y = \mathbb{R}^2$, $f(x) = (x, \sin x^{-1})$ and $f(x) = (\cos 2\pi x, \sin 2\pi x)$ are of interest.

THEOREM 8.2.2. *If $S \subset \mathbb{R}^n$ and S has the induced Euclidean metric, then S is a complete metric space iff S is closed in $\mathbb{R}^n$.*

PROOF. Assume that S is complete. From Corollary 7.6.2, in order to show that S is closed in $\mathbb{R}^n$ it is sufficient to show that whenever $(x_n)_{n=1}^{\infty}$ is a sequence in S and $x_n \to x \in \mathbb{R}^n$, then $x \in S$. But (x_n) is Cauchy by Theorem 8.1.2, and so it converges to a limit *in* S, which must be x by the uniqueness of limits in a metric space[5].

Assume S is closed in $\mathbb{R}^n$. Let (x_n) be a Cauchy sequence in S. Then (x_n) is also a Cauchy sequence in $\mathbb{R}^n$ and so $x_n \to x$ for some $x \in \mathbb{R}^n$ by Theorem 8.1.3. But $x \in S$ from Corollary 7.6.2. Hence any Cauchy sequence from S has a limit *in* S, and so S with the Euclidean metric is complete. □

Generalisation If S is a closed subset of a *complete* metric space (X, d), then S with the induced metric is also a complete metric space. The proof is the same.

***Remark** A metric space (X, d) fails to be complete because there are Cauchy sequences from X which do not have any limit *in* X. It is always possible to enlarge (X, d) to a *complete* metric space (X^*, d^*), where $X \subset X^*$, d is the restriction of d^* to X, and every element in X^* is the limit of a sequence from X. We call (X^*, d^*) the *completion* of (X, d).

For example, the completion of $\mathbb{Q}$ is $\mathbb{R}$, and more generally the completion of any $S \subset \mathbb{R}^n$ is the closure of S in $\mathbb{R}^n$.

In outline, the proof of the existence of the completion of (X, d) is as follows[6]:

> Let $\mathcal{S}$ be the set of *all* Cauchy sequences from X. We say two such sequences (x_n) and (y_n) are *equivalent* if $|x_n - y_n| \to 0$ as $n \to \infty$. The idea is that the two sequences are "trying" to converge to the *same* element. Let X^* be the set of all equivalence classes from $\mathcal{S}$ (i.e. elements of X^* are sets of Cauchy sequences, the Cauchy sequences in any element of X^* are equivalent to one another, and any two equivalent Cauchy sequences are in the same element of X^*).
>
> Each $x \in X$ is "identified" with the set of Cauchy sequences equivalent to the Cauchy sequence $(x, x, \ldots)$ (more precisely one shows this defines a one-one map from X into X^*). The distance d^* between two elements (i.e. equivalence classes) of X^* is defined by
>
> $$d^*((x_n), (y_n)) = \lim_{n \to \infty} |x_n - y_n|,$$

[5]To be more precise, let $x_n \to y$ in S (and so in $\mathbb{R}^k$) for some $y \in S$. But we also know that $x_n \to x$ in $\mathbb{R}^k$). Thus by uniqeness of limits *in the metric space* $\mathbb{R}^k$) , it follows that $x = y$.

[6]Note how this proof also gives a construction of the reals from the rationals.

where (x_n) is a Cauchy sequence from the first eqivalence class and (y_n) is a Cauchy sequence from the second eqivalence class. It is straightforward to check that the limit exists, is independent of the choice of representatives of the equivalence classes, and agrees with d when restricted to elements of X^* which correspond to elements of X. Similarly one checks that every element of X^* is the limit of a sequence "from" X in the appropriate sense.

Finally it is necessary to check that (X^*, d^*) is complete. So suppose that we have a Cauchy sequence from X^*. Each member of this sequence is itself equivalent to a Cauchy sequence from X. Let the nth member x_n of the sequence correspond to a Cauchy sequence $(x_{n1}, x_{n2}, x_{n3}, \ldots)$. Let x be the (equivalence class corresponding to the) *diagonal* sequence $(x_{11}, x_{22}, x_{33}, \ldots)$ (of course, this sequence must be shown to be Cauchy in X). Using the fact that (x_n) is a Cauchy sequence (of equivalence classes of Cauchy sequences), one can check that $x_n \to x$ (with respect to d^*). Thus (X^*, d^*) is complete. ∎

It is important to reiterate that the completion depends crucially on the metric d, see Example 2 above.

8.3. Contraction Mapping Theorem

Let (X, d) be a metric space and let $F : X \to X$. We say F is a *contraction* if there exists λ where $0 \leq \lambda < 1$ such that

$$d(F(x), F(y)) \leq \lambda d(x, y) \tag{102}$$

for all $x, y \in X$.

Remark It is essential that there is a *fixed* λ, $0 \leq \lambda < 1$ in (102). The function $f(x) = x^2$ is a contraction on each interval on $[0, a], 0 < a < 0.5$, but is *not* a contraction on $[0, 0.5]$.

A simple example of a contraction map on $\mathbb{R}^n$ is the map

$$\mathbf{x} \mapsto \mathbf{a} + r(\mathbf{x} - \mathbf{b}), \tag{103}$$

where $0 \leq r < 1$ and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. In this case $\lambda = r$, as is easily checked.

If $\mathbf{b} = \mathbf{0}$, then (103) is just dilation by the factor r followed by translation by the vector $\mathbf{a}$. More generally, since

$$\mathbf{a} + r(\mathbf{x} - \mathbf{b}) = \mathbf{b} + r(\mathbf{x} - \mathbf{b}) + \mathbf{a} - \mathbf{b},$$

we see (103) is dilation about $\mathbf{b}$ by the factor r, followed by translation by the vector $\mathbf{a} - \mathbf{b}$.

We say z is a *fixed point* of a map $F : A(\subset X) \to X$ if $F(z) = z$. In the preceding example, it is easily checked that the unique fixed point for *any* $r \neq 1$ is $(\mathbf{a} - r\mathbf{b})/(1 - r)$.

The following result is known as the *Contraction Mapping Theorem* or as the *Banach Fixed Point Theorem*. It has many important applications; we will use it to show the existence of solutions of *differential equations* and of *integral equations*, the existence of certain types of *fractals*, and to prove the Inverse Function Theorem 19.1.1.

You should first follow the proof in the case $X = \mathbb{R}^n$.

THEOREM 8.3.1 (Contraction Mapping Theorem). *Let (X, d) be a complete metric space and let $F : X \to X$ be a contraction map. Then F has a unique fixed point*[7].

[7] In other words, F has *exactly one* fixed point.

PROOF. We will find the fixed point as the limit of a Cauchy sequence.
Let x be any point in X and define a sequence $(x_n)_{n=1}^{\infty}$ by

$$x_1 = F(x),\ x_2 = F(x_1),\ x_3 = F(x_2), \ldots,\ x_n = F(x_{n-1}), \ldots.$$

Let λ be the contraction ratio.

1. Claim: (x_n) is Cauchy.

We have

$$\begin{aligned} d(x_n, x_{n+1}) &= d(F(x_{n-1}), F(x_n) \\ &\leq \lambda d(x_{n-1}, x_n) \\ &= \lambda d(F(x_{n-2}, F(x_{n-1}) \\ &\leq \lambda^2 d(x_{n-2}, x_{n-1}) \\ &\vdots \\ &\leq \lambda^{n-1} d(x_1, x_2). \end{aligned}$$

Thus if $m > n$ then

$$\begin{aligned} d(x_m, x_n) &\leq d(x_n, x_{n+1}) + \cdots + d(x_{m-1}, x_m) \\ &\leq (\lambda^{n-1} + \cdots + \lambda^{m-2}) d(x_1, x_2). \end{aligned}$$

But

$$\begin{aligned} \lambda^{n-1} + \cdots + \lambda^{m-2} &\leq \lambda^{n-1}(1 + \lambda + \lambda^2 + \cdots) \\ &= \lambda^{n-1} \frac{1}{1-\lambda} \\ &\to 0 \text{ as } n \to \infty. \end{aligned}$$

It follows (why?) that (x_n) is Cauchy.

Since X is complete, (x_n) has a limit in X, which we denote by x.

2. Claim: x is a fixed point of F.

We claim that $F(x) = x$, i.e. $d(x, F(x)) = 0$. Indeed, for any n

$$\begin{aligned} d(x, F(x)) &\leq d(x, x_n) + d(x_n, F(x)) \\ &= d(x, x_n) + d(F(x_{n-1}), F(x)) \\ &\leq d(x, x_n) + \lambda d(x_{n-1}, x) \\ &\to 0 \end{aligned}$$

as $n \to \infty$. This establishes the claim.

3. Claim: The fixed point is unique.

If x and y are fixed points, then $F(x) = x$ and $F(y) = y$ and so

$$d(x, y) = d(F(x), F(y)) \leq \lambda d(x, y).$$

Since $0 \leq \lambda < 1$ this implies $d(x, y) = 0$, i.e. $x = y$. □

Remark Fixed point theorems are of great importance for proving *existence* results. The one above is perhaps the simplest, but has the advantage of giving an algorithm for determining the fixed point. In fact, it also gives an estimate of how close an iterate is to the fixed point (how?).

In applications, the following Corollary is often used.

COROLLARY 8.3.2. *Let S be a closed subset of a complete metric space (X, d) and let $F : S \to S$ be a contraction map on S. Then F has a unique fixed point in S.*

PROOF. We saw following Theorem 8.2.2 that S is a complete metric space with the induced metric. The result now follows from the preceding Theorem. □

Example Take $\mathbb{R}$ with the standard metric, and let $a > 1$. Then the map

$$f(x) = \frac{1}{2}(x + \frac{a}{x})$$

takes $[1, \infty)$ into itself, and is contractive with $\lambda = \frac{1}{2}$. What is the fixed point? (This was known to the Babylonians, nearly 4000 years ago.)

Somewhat more generally, consider Newton's method for finding a simple root of the equation $f(x) = 0$, given some interval containing the root. Assume that f'' is bounded on the interval. Newton's method is an iteration of the function

$$g(x) = x - \frac{f(x)}{f'(x)}\,.$$

To see why this could possibly work, suppose that there is a root ξ in the interval $[a, b]$, and that $f' > 0$ on $[a, b]$. Since

$$g'(x) = \frac{f(x)}{(f'(x)^2} f''(x)\,,$$

we have $|g'| < 0.5$ on some interval $[\alpha, \beta]$ containing ξ. Since $g(\xi) = \xi$, it follows that g is a contraction on $[\alpha, \beta]$.

Exercise When $S \subseteq \mathbb{R}$ is an interval it is often easy to verify that a function $f : S \to S$ is a contraction by use of the mean value theorem. What about the following function on $[0, 1]$?

$$f(x) = \begin{cases} \frac{\sin(x)}{x} & x \neq 0 \\ 1 & x = 0 \end{cases}$$

CHAPTER 9

Sequences and Compactness

9.1. Subsequences

Recall that if $(x_n)_{n=1}^{\infty}$ is a sequence in some set X and $n_1 < n_2 < n_3 < \ldots$, then the sequence $(x_{n_i})_{i=1}^{\infty}$ is called a *subsequence* of $(x_n)_{n=1}^{\infty}$.

The following result is easy.

THEOREM 9.1.1. *If a sequence in a metric space converges then every subsequence converges to the same limit as the original sequence.*

PROOF. Let $x_n \to x$ in the metric space (X, d) and let (x_{n_i}) be a subsequence. Let $\epsilon > 0$ and choose N so that

$$d(x_n, x) < \epsilon$$

for $n \geq N$. Since $n_i \geq i$ for all i, it follows

$$d(x_{n_i}, x) < \epsilon$$

for $i \geq N$. □

Another useful fact is the following.

THEOREM 9.1.2. *If a Cauchy sequence in a metric space has a convergent subsequence, then the sequence itself converges.*

PROOF. Suppose that (x_n) is cauchy in the metric space (X, d). So given $\epsilon > 0$ there is N such that $d(x_n, x_m) < \epsilon$ provided $m, n > N$. If (x_{n_j}) is a subsequence convergent to $x \in X$, then there is N' such that $d(x_{n_j}, x) < \epsilon$ for $j > N'$. Thus for $m > N$, take any $j > \max\{N, N'\}$ (so that $n_j > N$ certainly), to see that

$$d(x, x_m) \leq d(x, x_{n_j}) + d(x_{n_j}, x_m) < 2\epsilon\,.$$

□

9.2. Existence of Convergent Subsequences

It is often very important to be able to show that a sequence, even though it may not be convergent, has a convergent subsequence. The following is a simple criterion for sequences in $\mathbb{R}^n$ to have a convergent subsequence. The same result is not true in an arbitrary metic space, as we see in Remark 2 following the Theorem.

THEOREM 9.2.1 (Bolzano-Weierstrass). [1] *Every bounded sequence in $\mathbb{R}^n$ has a convergent subsequence.*

Let us give two proofs of this significant result.

PROOF. By the remark following the proof of Theorem 8.1.3, a bounded sequence in $\mathbb{R}$ has a limit point, and necessarily there is a subsequence which converges to this limit point.

Suppose then, that the result is true in $\mathbb{R}^n$ for $1 \leq k < m$, and take a bounded sequence $(\mathbf{x}_n) \subset \mathbb{R}^m$. For each n, write $\mathbf{x}_n = (\mathbf{y}_n, x_n^m)$ in the obvious way. Then

[1] Other texts may have different forms of the Bolzano-Weierstrass Theorem.

$(\mathbf{y}_n) \subset \mathbb{R}^{m-1}$ is a bounded sequence, and so has a convergent subsequence $(\mathbf{y}_{n_j})$ by the inductive hypothesis. But then $(x^m_{n_j})$ is a bounded sequence in $\mathbb{R}$, so has a subsequence $(x^m_{n'_j})$ which converges. It follows from Theorem 7.5.1 that the subsequence $(\mathbf{x}_{n_j})$of the original sequence $(\mathbf{x}_n)$ is convergent. Thus the result is true for $k = m$. □

Note the "diagonal" argument in the second proof.

PROOF. (See Figure 1.)

Let $(\mathbf{x}_n)_{n=1}^{\infty}$ be a bounded sequence of points in $\mathbb{R}^n$, which for convenience we rewrite as $(\mathbf{x}_n{}^{(1)})_{n=1}^{\infty}$. All terms $(\mathbf{x}_n{}^{(1)})$ are contained in a closed cube

$$I_1 = \left\{\mathbf{y} : |y^i| \leq r,\ i = 1, \ldots, k\right\}$$

for some $r > 0$.

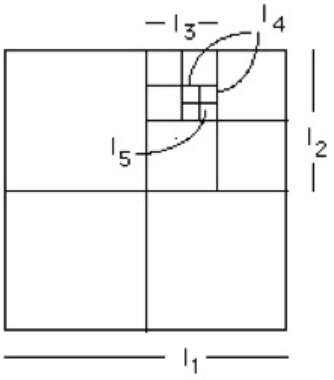

FIGURE 1. Diagram for the second proof of Theorem 9.2.1.

Divide I_1 into 2^k closed subcubes as indicated in the diagram. At least one of these cubes must contain an *infinite* number of terms from the sequence $(\mathbf{x}_n{}^{(1)})_{n=1}^{\infty}$. Choose one such cube and call it I_2. Let the corresponding subsequence in I_2 be denoted by $(\mathbf{x}_n{}^{(2)})_{n=1}^{\infty}$.

Repeating the argument, divide I_2 into 2^k closed subcubes. Once again, at least one of these cubes must contain an *infinite* number of terms from the sequence $(\mathbf{x}_n{}^{(2)})_{n=1}^{\infty}$. Choose one such cube and call it I_3. Let the corresponding subsequence be denoted by $(\mathbf{x}_n{}^{(3)})_{n=1}^{\infty}$.

Continuing in this way we find a decreasing sequence of closed cubes

$$I_1 \supset I_2 \supset I_3 \supset \cdots$$

and sequences

$$\begin{gathered}(\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \mathbf{x}_3^{(1)}, \ldots) \\ (\mathbf{x}_1^{(2)}, \mathbf{x}_2^{(2)}, \mathbf{x}_3^{(2)}, \ldots) \\ (\mathbf{x}_1^{(3)}, \mathbf{x}_2^{(3)}, \mathbf{x}_3^{(3)}, \ldots) \\ \vdots\end{gathered}$$

where each sequence is a *subsequence* of the preceding sequence and the terms of the ith sequence are all members of the cube I_i.

We now define the sequence $(\mathbf{y}_i)$ by $\mathbf{y}_i = \mathbf{x}_i^{(i)}$ for $i = 1, 2, \ldots$. This is a subsequence of the original sequence.

Notice that for each N, the terms $\mathbf{y}_N, \mathbf{y}_{N+1}, \mathbf{y}_{N+2}, \ldots$ are all members of I_N. Since the distance between any two points in I_N is[2] $\leq \sqrt{k}r/2^{N-2} \to 0$ as $N \to \infty$, it follows that $(\mathbf{y}_i)$ is a Cauchy sequence. Since $\mathbb{R}^n$ is complete, it follows $(\mathbf{y}_i)$ converges in $\mathbb{R}^n$. This proves the theorem. □

Remark 1 If a sequence in $\mathbb{R}^n$ is not bounded, then it need not contain a convergent subsequence. For example, the sequence $(1, 2, 3, \ldots)$ in $\mathbb{R}$ does not contain any convergent subsequence (since the nth term of any subsequence is $\geq n$ and so any subsequence is not even bounded).

Remark 2 The Theorem is not true if $\mathbb{R}^n$ is replaced by $\mathcal{C}[0, 1]$. For example, consider the sequence of functions (f_n) whose graphs are as shown in Figure 2.

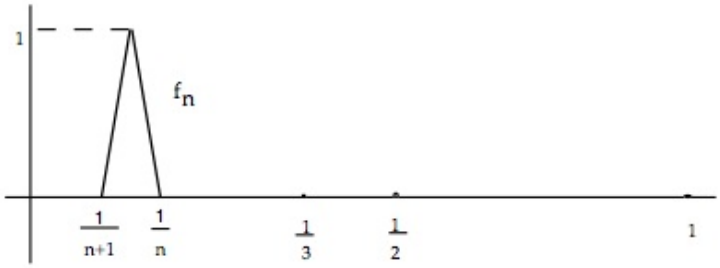

FIGURE 2. A sequence of uniformly bounded functions without a convergent subsequence in the sup norm.

The sequence is bounded since $||f_n||_\infty = 1$, where we take the sup norm (and corresponding metric) on $\mathcal{C}[0, 1]$.

But if $n \neq m$ then

$$||f_n - f_m||_\infty = \sup\{|f_n(x) - f_m(x)| : x \in [0, 1]\} = 1,$$

as is seen by choosing appropriate $x \in [0, 1]$. Thus no subsequence of (f_n) can converge in the sup norm[3].

We often use the previous theorem in the following form.

COROLLARY 9.2.2. *If $S \subset \mathbb{R}^n$, then S is closed and bounded iff every sequence from S has a subsequence which converges to a limit* in *S.*

PROOF. Let $S \subset \mathbb{R}^n$, be closed and bounded. Then any sequence from S is bounded and so has a convergent subsequence by the previous Theorem. The limit is in S as S is closed.

Conversely, first suppose S is not bounded. Then for every natural number n there exists $x_n \in S$ such that $|x_n| \geq n$. Any subsequence of (x_n) is unbounded and so cannot converge.

Next suppose S is not closed. Then there exists a sequence (x_n) from S which converges to $x \notin S$. Any subsequence also converges to x, and so does not have its limit in S. □

Remark The second half of this corollary holds in a general metric space, the first half does not.

[2]The distance between any two points in I_1 is $\leq 2\sqrt{k}r$, between any two points in I_2 is thus $\leq \sqrt{k}r$, between any two points in I_3 is thus $\leq \sqrt{k}r/2$, etc.

[3]We will consider the sup norm on functions in detail in a later chapter. Notice that $f_n(x) \to 0$ as $n \to \infty$ for *every* $x \in [0, 1]$—we say that (f_n) converges *pointwise* to the *zero function.* Thus here we have convergence pointwise but not in the sup metric. This notion of pointwise convergence *cannot* be described by a metric.

9.3. Compact Sets

DEFINITION 9.3.1. A subset S of a metric space (X, d) is *compact* if every sequence from S has a subsequence which converges to an element of S. If X is compact, we say the metric space itself is compact.

Remark

(1) This notion is also called *sequential compactness*. There is another definition of compactness in terms of coverings by open sets which applies to any topological space[4] and agrees with the definition here for metric spaces. We will investigate this more general notion in Chapter 15.
(2) Compactness turns out to be a stronger condition than completeness, though in some arguments one notion can be used in place of the other.

Examples

(1) From Corollary 9.2.2 the compact subsets of $\mathbb{R}^n$ are precisely the closed bounded subsets. Any such compact subset, with the induced metric, is a compact metric space. For example, $[a, b]$ with the usual metric is a compact metric space.
(2) The Remarks on $\mathcal{C}[0, 1]$ in the previous section show that the closed[5] bounded set $S = \{f \in \mathcal{C}[0, 1] : ||f||_\infty = 1\}$ is *not* compact. The set S is just the "closed unit sphere" in $\mathcal{C}[0, 1]$. (You will find later that $\mathcal{C}[0, 1]$ is not unusual in this regard, the closed unit ball in *any* infinite dimensional normed space fails to be compact.)

Relative and Absolute Notions Recall from the Note at the end of Section (6.4) that if X is a metric space the notion of a set $S \subset X$ being open or closed is a *relative* one, in that it depends also on X and not just on the induced metric on S.

However, whether or not S is compact depends *only* on S itself and the induced metric, and so we say compactess is an *absolute* notion. Similarly, completeness is an absolute notion.

9.4. Nearest Points

We now give a simple application in $\mathbb{R}^n$ of the preceding ideas.

DEFINITION 9.4.1. Suppose $A \subset X$ and $x \in X$ where (X, d) is a metric space. The *distance from x to A* is defined by

$$d(x, A) = \inf_{y \in A} d(x, y). \tag{104}$$

It is not necessarily the case that there exists $y \in A$ such that $d(x, A) = d(x, y)$. For example if $A = [0, 1) \subset \mathbb{R}$ and $x = 2$ then $d(x, A) = 1$, but $d(x, y) > 1$ for all $y \in A$.

Moreover, even if $d(x, A) = d(x, y)$ for some $y \in A$, this y may not be unique. For example, let $S = \{y \in \mathbb{R}^2 : ||y|| = 1\}$ and let $x = (0, 0)$. Then $d(x, S) = 1$ and $d(x, y) = 1$ for every $y \in S$.

Notice also that if $x \in A$ then $d(x, A) = 0$. But $d(x, A) = 0$ does not imply $x \in A$. For example, take $A = [0, 1)$ and $x = 1$.

However, we do have the following theorem. Note that in the result we need S to be closed but not necessarily bounded.

The technique used in the proof, of taking a "minimising" sequence and extracting a convergent subsequence, is a fundamental idea.

[4] You will study general topological spaces in a later course.

[5] S is the boundary of the unit ball $B_1(0)$ in the metric space $\mathcal{C}[0, 1]$ and is closed as noted in the Examples following Theorem 6.4.7.

THEOREM 9.4.2. *Let S be a closed subset of $\mathbb{R}^n$, and let $x \in \mathbb{R}^n$. Then there exists $y \in S$ such that $d(x, y) = d(x, S)$.*

PROOF. Let

$$\gamma = d(x, S)$$

and choose a sequence (y_n) in S such that

$$d(x, y_n) \to \gamma \text{ as } n \to \infty.$$

(*Draw a diagram.*) This is possible from (104) by the definition of *inf*.

The sequence (y_n) is bounded[6] and so has a convergent subsequence which we also denote by (y_n)[7].

Let y be the limit of the convergent subsequence (y_n). Then $d(x, y_n) \to d(x, y)$ by Theorem 7.3.4, but $d(x, y_n) \to \gamma$ since this is also true for the original sequence. It follows $d(x, y) = \gamma$ as required. □

[6]This is fairly obvious and the actual argument is similar to showing that convergent sequences are bounded, c.f. Theorem 7.3.3. More precisely, we have there exists an integer N such that $d(x, y_n) \leq \gamma + 1$ for all $n \geq N$, by the fact $d(x, y_n) \to \gamma$. Let

$$M = \max\{\gamma + 1, d(x, y_1), \ldots, d(x, y_{N-1})\}.$$

Then $d(x, y_n) \leq M$ for *all* n, and so (y_n) is bounded.

[7]This *abuse of notation* in which we use the same notation for the subsequence as for the original sequence is a common one. It saves using subscripts $y_{n_{i_j}}$ — which are particularly messy when we take subsequences of subsequences — and will lead to no confusion provided we are careful.

CHAPTER 10

Limits of Functions

10.1. Diagrammatic Representation of Functions

In this Chapter we will consider functions $f : A\,(\subset X) \to Y$ where X and Y are metric spaces.

Important cases to keep in mind are $f : A\,(\subset \mathbb{R}) \to \mathbb{R}$, $f : A\,(\subset \mathbb{R}^n) \to \mathbb{R}$ and $f : A\,(\subset \mathbb{R}^n) \to \mathbb{R}^m$.

Sometimes we can *represent* a function by its graph. Of course functions can be quite complicated, and we should be careful not to be misled by the simple features of the particular graphs which we are able to sketch.

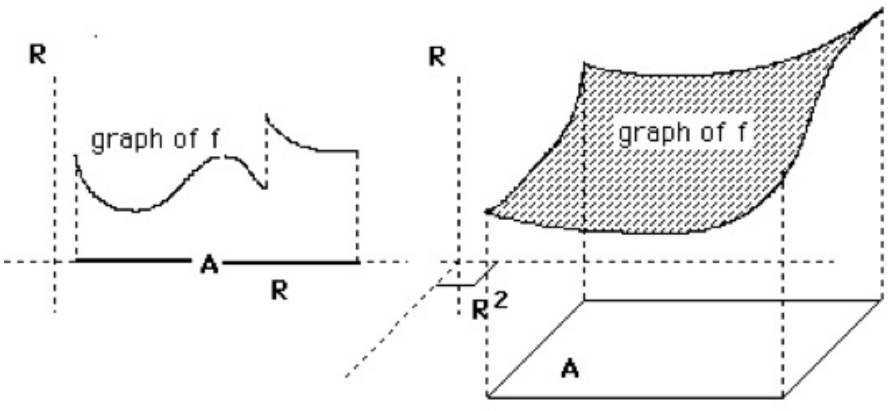

FIGURE 1. Graphs of a function $f : A\ (\subset \mathbb{R}) \to \mathbb{R}$ and a function $f : A\ (\subset \mathbb{R}^2) \to \mathbb{R}$.

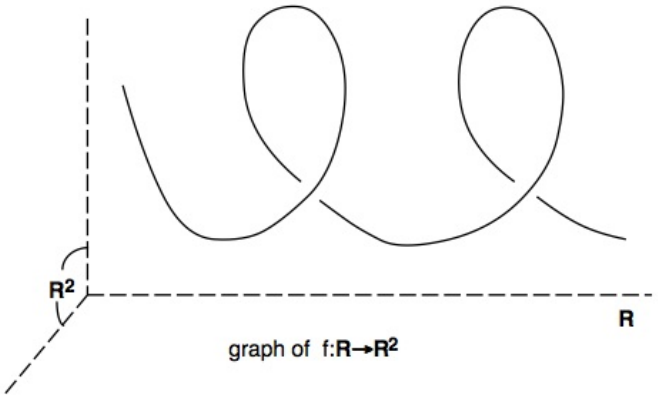

FIGURE 2. Graph of a function $f : U\ (\subset \mathbb{R}) \to \mathbb{R}^2$.

Sometimes we can sketch the domain and the range of the function, perhaps also with a coordinate grid and its image. See the following diagram.

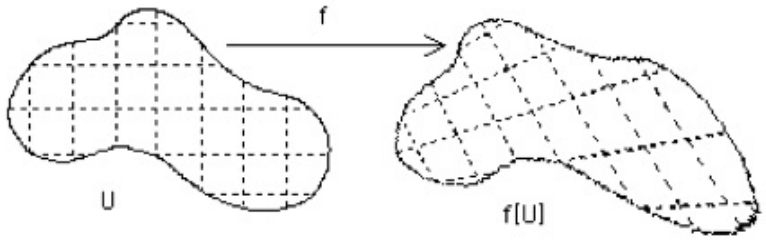

FIGURE 3. Domain with grid, range with image of grid, for a function $f : \mathbb{R}^2 \to \mathbb{R}^2$.

Sometimes we can represent a function by drawing a vector at various points in its domain to represent $f(\mathbf{x})$.

Sometimes we can represent a real-valued function by drawing the level sets (contours) of the function. See Section 17.6.2.

In other cases the best we can do is to represent the graph, or the domain and range of the function, in a highly idealised manner. See the following diagrams.

10.2. Definition of Limit

Suppose $f : A\,(\subset X) \to Y$ where X and Y are metric spaces. In considering the limit of $f(x)$ as $x \to a$ we are interested in the behaviour of $f(x)$ for x near a. We are *not* concerned with the value of f at a and indeed f need not even be defined at a, nor is it necessary that $a \in A$. See Example 1 following the definition below.

For the above reasons we assume a is a *limit point* of A, which is equivalent to the existence of some sequence $(x_n)_{n=1}^{\infty} \subset A \setminus \{a\}$ with $x_n \to a$ as $n \to \infty$. In particular, a is *not* an isolated point of A.

The following definition of limit in terms of sequences is equivalent to the usual ϵ–δ definition as we see in the next section. The definition here is perhaps closer

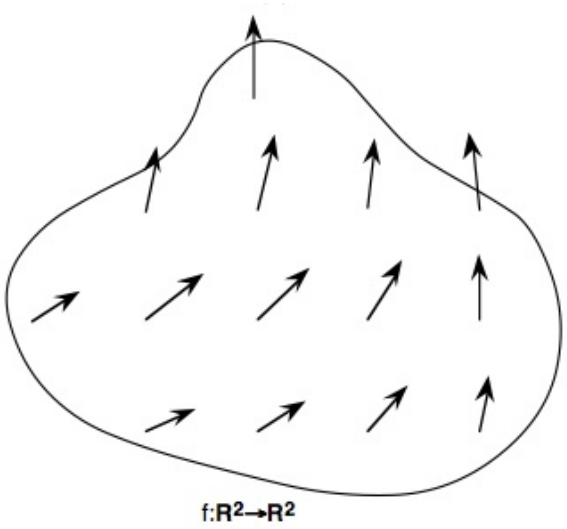

FIGURE 4. Vector field representation of a function $f : \mathbb{R}^2 \to \mathbb{R}^2$.

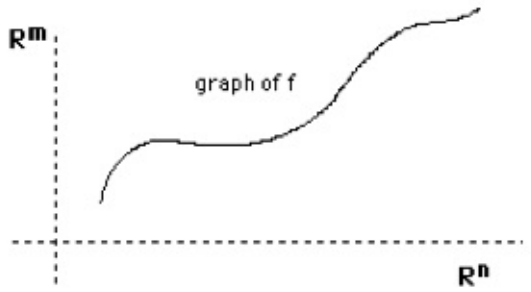

FIGURE 5. Schematic/Idealised representation of a function $f : \mathbb{R}^n \to \mathbb{R}^m$.

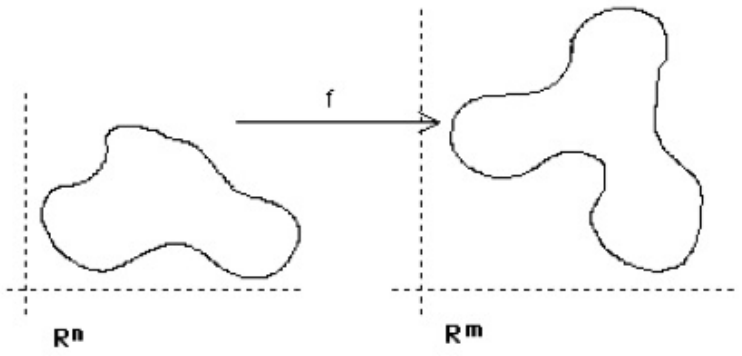

FIGURE 6. Another schematic/Idealised representation of a function $f : \mathbb{R}^n \to \mathbb{R}^m$.

to the usual intuitive notion of a limit. Moreover, with this definition we can deduce the basic properties of limits directly from the corresponding properties of sequences, as we will see in Section 10.4.

DEFINITION 10.2.1. Let $f : A\,(\subset X) \to Y$ where X and Y are metric spaces, and let a be a limit point of A. Suppose

$$(x_n)_{n=1}^{\infty} \subset A \setminus \{a\} \text{ and } x_n \to a \text{ together imply } f(x_n) \to b.$$

Then we say $f(x)$ *approaches* b *as* x *approaches* a, or f *has limit* b *at* a and write

$$f(x) \to b \text{ as } x \to a \ (x \in A),$$

or

$$\lim_{\substack{x \to a \\ x \in A}} f(x) = b,$$

or

$$\lim_{x \to a} f(x) = b$$

(where in the last notation the intended domain A is understood from the context).

DEFINITION 10.2.2. [*One-sided Limits*] If in the previous definition $X = \mathbb{R}$ and A is an interval with a as a left or right endpoint, we write

$$\lim_{x \to a^+} f(x) \text{ or } \lim_{x \to a^-} f(x)$$

and say *the limit as* x *approaches* a *from the right* or *the limit as* x *approaches* a *from the left*, respectively.

Example 1 (a) Let $A = (-\infty, 0) \cup (0, \infty)$. Let $f : A \to \mathbb{R}$ be given by $f(x) = 1$ if $x \in A$. Then $\lim_{x \to 0} f(x) = 1$, even though f is not defined at 0.

(b) Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = 1$ if $x \neq 0$ and $f(0) = 0$. Then again $\lim_{x \to 0} f(x) = 1$.

This example illustrates why in the Definition 10.2.1 we require $x_n \neq a$, even if $a \in A$.

Example 2 If $g : \mathbb{R} \to \mathbb{R}$ then

$$\lim_{x \to a} \frac{g(x) - g(a)}{x - a}$$

is (if the limit exists) called the *derivative* of g at a. Note that in Definition 10.2.1 we are taking $f(x) = \Big(g(x) - g(a)\Big)/(x - a)$ and that $f(x)$ is *not* defined at $x = a$. We take $A = \mathbb{R} \setminus \{a\}$, or $A = (a - \delta, a) \cup (a, a + \delta)$ for some $\delta > 0$.

Example 3 (*Exercise:* Draw a sketch.)

$$\lim_{x \to 0} x \sin\left(\frac{1}{x}\right) = 0. \tag{105}$$

To see this take any sequence (x_n) where $x_n \to 0$ and $x_n \neq 0$. Now

$$-|x_n| \leq x_n \sin\left(\frac{1}{x_n}\right) \leq |x_n|.$$

Since $-|x_n| \to 0$ and $|x_n| \to 0$ it follows from the Comparison Test that $x_n \sin(1/x_n) \to 0$, and so (105) follows.

We can also use the definition to show in some cases that $\lim_{x \to a} f(x)$ does *not* exist. For this it is sufficient to show that for *some* sequence $(x_n) \subset A \setminus \{a\}$ with $x_n \to a$ the corresponding sequence $(f(x_n))$ does *not* have a limit. Alternatively, it is sufficient to give two sequences $(x_n) \subset A \setminus \{a\}$ and $(y_n) \subset A \setminus \{a\}$ with $x_n \to a$ and $y_n \to a$ but $\lim x_n \neq \lim y_n$.

Example 4 (*Draw a sketch*). $\lim_{x\to 0}\sin(1/x)$ does not exist. To see this consider, for example, the sequences $x_n = 1/(n\pi)$ and $y_n = 1/((2n+1/2)\pi)$. Then $\sin(1/x_n) = 0 \to 0$ and $\sin(1/y_n) = 1 \to 1$.

Example 5

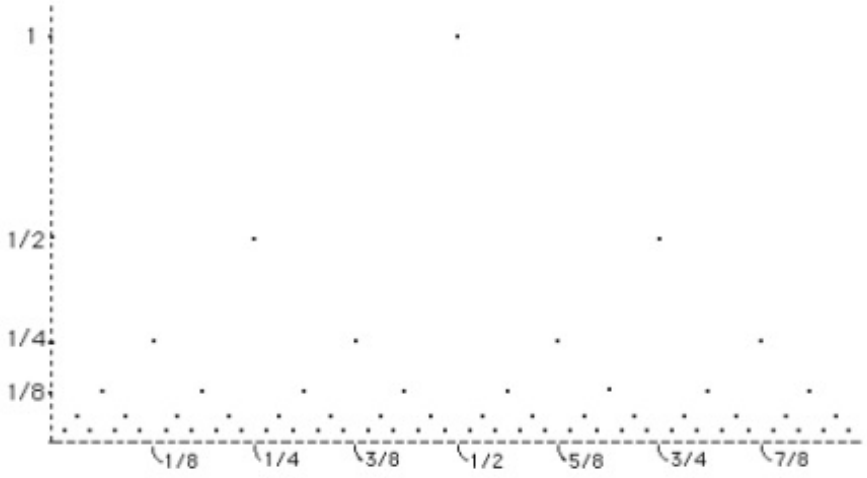

FIGURE 7. An attempt to sketch the graph of the function $g : [0,1] \to [0,1]$ in Example 5.

Consider the function $g\colon[0,1] \to [0,1]$ defined by

$$g(x) = \begin{cases} 1/2 & \text{if } x = 1/2 \\ 1/4 & \text{if } x = 1/4, 3/4 \\ 1/8 & \text{if } x = 1/8, 3/8, 5/8, 7/8 \\ \vdots & \\ 1/2^k & \text{if } x = 1/2^k, 3/2^k, 5/2^k, \ldots, (2^k-1)/2^k \\ \vdots & \end{cases}$$

$$g(x) = 0 \quad \text{otherwise.}$$

In fact, in simpler terms,

$$g(x) = \begin{cases} 1/2^k & x = p/2^k \text{ for } p \text{ odd}, 1 \le p < 2^k \\ 0 & \text{otherwise} \end{cases}$$

Then we claim $lim_{x\to a} g(x) = 0$ for *all* $a \in [0,1]$.

First define $S_k = \{1/2^k, 2/2^k, 3/2^k, \ldots, (2^k-1)/2^k\}$ for $k = 1, 2, \ldots$. Notice that $g(x)$ will take values $1/2, 1/4, \ldots, 1/2^k$ for $x \in S_k$, and

$$g(x) < 1/2^k \text{ if } x \notin S_k. \tag{106}$$

For each $a \in [0,1]$ and $k = 1, 2, \ldots$ define the *distance* from a to $S_k \setminus \{a\}$ (whether or not $a \in S_k$) by

$$d_{a,k} = \min\Big\{|x-a| : x \in S_k \setminus \{a\}\Big\}. \tag{107}$$

Then $d_{a,k} > 0$, even if $a \in S_k$, since it is the minimum of a *finite* set of *strictly positive* (i.e. > 0) numbers.

Now let (x_n) be any sequence with $x_n \to a$ and $x_n \neq a$ for all n. We need to show $g(x_n) \to 0$.

Suppose $\epsilon > 0$ and choose k so $1/2^k \leq \epsilon$. Then from (106), $g(x_n) < \epsilon$ if $x_n \notin S_k$.

On the other hand, $0 < |x_n - a| < d_{a,k}$ for all $n \geq N$ (say), since $x_n \neq a$ for all n and $x_n \to a$. It follows from (107) that $x_n \notin S_k$ for $n \geq N$. Hence $g(x_n) < \epsilon$ for $n \geq N$.

Since also $g(x) \geq 0$ for all x it follows that $g(x_n) \to 0$ as $n \to \infty$. Hence $\lim_{x\to a} g(x) = 0$ as claimed.

Example 6 Define $h : \mathbb{R} \to \mathbb{R}$ by

$$h(x) = \lim_{m\to\infty} \lim_{n\to\infty} (\cos(m!\pi x))^n$$

Then h fails to have a limit at every point of $\mathbb{R}$.

Example 7 Let

$$f(x, y) = \frac{xy}{x^2 + y^2}$$

for $(x, y) \neq (0, 0)$.

If $y = ax$ then $f(x, y) = a(1 + a^2)^{-1}$ for $x \neq 0$. Hence

$$\lim_{\substack{(x,y)\to a \\ y=ax}} f(x, y) = \frac{a}{1 + a^2}.$$

Thus we obtain a different limit of f as $(x, y) \to (0, 0)$ along different lines. It follows that

$$\lim_{(x,y)\to(0,0)} f(x, y)$$

does not exist.

A partial diagram of the graph of f is shown

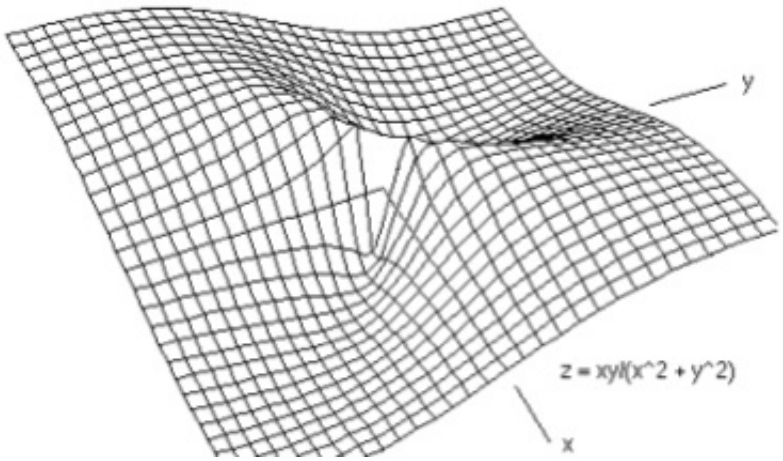

FIGURE 8. The function $f : \mathbb{R}^2 \to \mathbb{R}$ is continuous along every line through the origin, but is not continuous at the origin.

One can also visualize f by sketching *level sets*[1] of f as shown in the next diagram. Then you can visualise the graph of f as being swept out by a straight line rotating around the origin at a height as indicated by the level sets.

[1] A level set of f is a set on which f is constant.

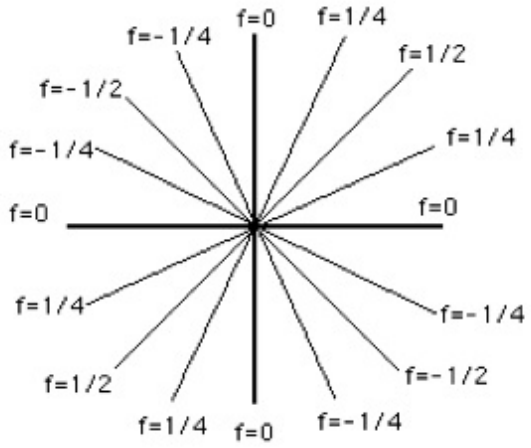

FIGURE 9. Level sets of the function in Figure 8.

Example 8 Let

$$f(x,y) = \frac{x^2 y}{x^4 + y^2}$$

for $(x,y) \neq (0,0)$.

Then

$$\begin{aligned} \lim_{\substack{(x,y)\to a \\ y=ax}} f(x,y) &= \lim_{x\to 0} \frac{ax^3}{x^4 + a^2x^2} \\ &= \lim_{x\to 0} \frac{ax}{x^2 + a^2} \\ &= 0. \end{aligned}$$

Thus the limit of f as $(x,y) \to (0,0)$ along *any* line $y = ax$ is 0. The limit along the y-axis $x = 0$ is also easily seen to be 0.

But it is still not true that $\lim_{(x,y)\to(0,0)} f(x,y)$ exists. For if we consider the limit of f as $(x,y) \to (0,0)$ along the parabola $y = bx^2$ we see that $f = b(1+b^2)^{-1}$ on this curve and so the limit is $b(1+b^2)^{-1}$.

You might like to draw level curves (corresponding to parabolas $y = bx^2$).

This example reappears in Chapter 17. Clearly we can make such examples as complicated as we please.

10.3. Equivalent Definition

In the following theorem, (2) is the usual ϵ–δ definition of a limit.

THEOREM 10.3.1. *Suppose (X,d) and (Y,ρ) are metric spaces, $A \subset X$, $f: A \to Y$, and a is a limit point of A. Then the following are equivalent:*

(1) $\lim_{x\to a} f(x) = b$;

(2) *For every $\epsilon > 0$ there is a $\delta > 0$ such that*

$$x \in A \setminus \{a\} \text{ and } d(x,a) < \delta \text{ implies } \rho(f(x),b) < \epsilon;$$
$$\text{i.e. } x \in \Big(A \cap B_\delta(a)\Big) \setminus \{a\} \;\Rightarrow\; f(x) \in B_\epsilon(b).$$

PROOF. (1) $\Rightarrow$ (2): Assume (1), so that whenever $(x_n) \subset A \setminus \{a\}$ and $x_n \to a$ then $f(x_n) \to b$.

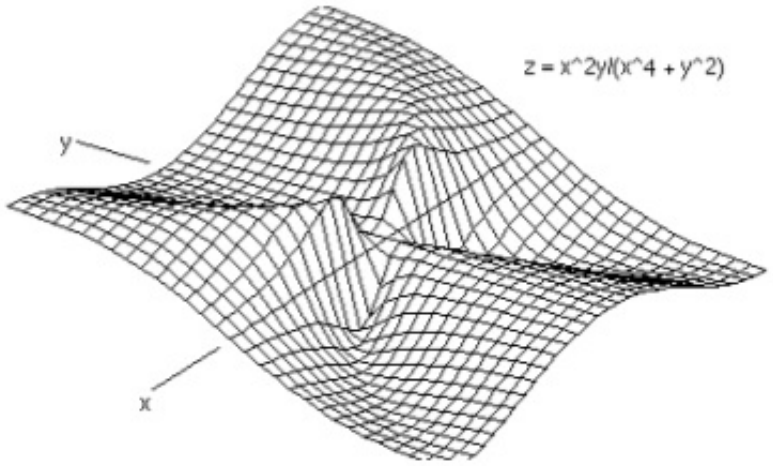

FIGURE 10. Graph of the function f in Example 8. The limit at the origin along every straight line through the origin exists and equals 0. Yet the limit at the origin does not exist.

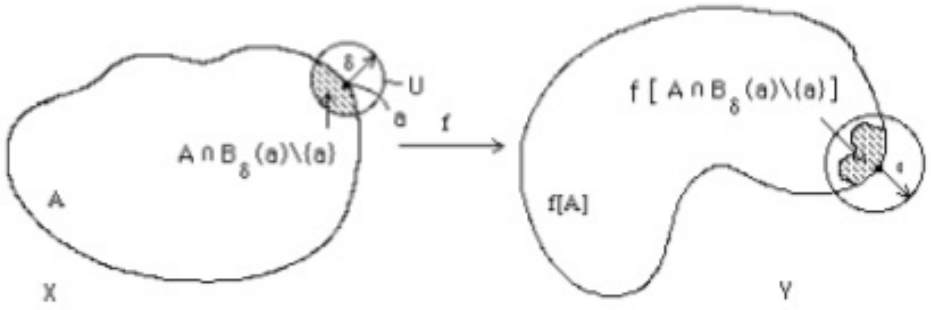

FIGURE 11. Illustration of condition (2) in Theorem 10.3.1.

Suppose (by way of obtaining a contradiction) that (2) is *not* true. Then for some $\epsilon > 0$ there is no $\delta > 0$ such that

$$x \in A \setminus \{a\} \text{ and } d(x,a) < \delta \text{ implies } \rho(f(x),b) < \epsilon.$$

In other words, for some $\epsilon > 0$ and every $\delta > 0$, there exists an x depending on δ, with

$$x \in A \setminus \{a\} \text{ and } d(x,a) < \delta \text{ and } \rho(f(x),b) \geq \epsilon. \tag{108}$$

Choose such an ϵ, and for $\delta = 1/n$, $n = 1, 2, \ldots$, choose $x = x_n$ satisfying (108). It follows $x_n \to a$ and $(x_n) \subset A \setminus \{a\}$ but $f(x_n) \not\to b$. This contradicts (1) and so (2) *is* true.

(2) $\Rightarrow$ (1): Assume (2).

In order to prove (1) suppose $(x_n) \subset A \setminus \{a\}$ and $x_n \to a$. We have to show $f(x_n) \to b$.

In order to do this take $\epsilon > 0$. By (2) there is a $\delta > 0$ (depending on ϵ) such that

$$\overbrace{x_n \in A \setminus \{a\} \text{ and } d(x_n, a) < \delta}^{*} \text{ implies } \rho(f(x_n), b) < \epsilon.$$

But * is true for all $n \geq N$ (say, where N depends on δ and hence on ϵ), and so $\rho(f(x_n), b) < \epsilon$ for all $n \geq N$. Thus $f(x_n) \to b$ as required and so (1) is true. □

10.4. Elementary Properties of Limits

Assumption *In this section we let $f, g : A\,(\subset X) \to Y$ where (X, d) and (Y, ρ) are metric spaces.*

The next definition is not surprising.

DEFINITION 10.4.1. The function f is *bounded* on the set $E \subset A$ if $f[E]$ is a bounded set in Y.

Thus the function $f : (0, \infty) \to \mathbb{R}$ given by $f(x) = x^{-1}$ is bounded on $[a, \infty)$ for any $a > 0$ but is not bounded on $(0, \infty)$.

PROPOSITION 10.4.2. *Assume $\lim_{x\to a} f(x)$ exists. Then for some $r > 0$, f is bounded on the set $A \cap B_r(a)$.*

PROOF. Let $\lim_{x\to a} f(x) = b$ and let $V = B_1(b)$. V is certainly a bounded set. For some $r > 0$ we have $f\left[(A \setminus \{a\}) \cap B_r(a)\right] \subset V$ from Theorem 10.3.1(2). Since a subset of a bounded set is bounded, it follows that $f\left[(A \setminus \{a\}) \cap B_r(a)\right]$ is bounded, and so $f[A \cap B_r(a)]$ is bounded if $a \notin A$. If $a \in A$ then $f[A \cap B_r(a)] = f\left[(A \setminus \{a\}) \cap B_r(a)\right] \cup \{f(a)\}$, and so again $f[A \cap B_r(a)]$ is bounded. □

Most of the basic properties of limits of functions follow directly from the corresponding properties of limits of sequences without the necessity for any ϵ–δ arguments.

THEOREM 10.4.3. *Limits are unique; in the sense that if $\lim_{x\to a} f(x) = b_1$ and $\lim_{x\to a} f(x) = b_2$ then $b_1 = b_2$.*

PROOF. Suppose $\lim_{x\to a} f(x) = b_1$ and $\lim_{x\to a} f(x) = b_2$. If $b_1 \neq b_2$, then $\epsilon = \frac{|b_1 - b_2|}{2} > 0$. By the defienition of the limit, there are $\delta_1, \delta_2 > 0$ such that

$$0 < d(x, a) < \delta_1 \Rightarrow \rho(f(x) - b_1) < \epsilon\,,$$
$$0 < d(x, a) < \delta_2 \Rightarrow \rho(f(x) - b_2) < \epsilon\,.$$

Taking $0 < d(x, a) < \min\{\delta_1, \delta - 2\}$ gives a contradiction. □

Notation Assume $f : A\,(\subset X) \to \mathbb{R}^n$. (In applications it is often the case that $X = \mathbb{R}^m$ for some m). We write

$$f(x) = (f^1(x), \ldots, f^k(x)).$$

Thus each of the f^i is just a real-valued function defined on A.

For example, the linear transformation $f : \mathbb{R}^2 \to \mathbb{R}^2$ described by the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is given by

$$f(\mathbf{x}) = f(x^1, x^2) = (ax^1 + bx^2, cx^1 + dx^2).$$

Thus $f^1(\mathbf{x}) = ax^1 + bx^2$ and $f^2(\mathbf{x}) = cx^1 + dx^2$.

THEOREM 10.4.4. *Let $f : A\,(\subset X) \to \mathbb{R}^n$. Then $\lim_{x\to a} f(x)$ exists iff the component limits, $\lim_{x\to a} f^i(x)$, exist for all $i = 1, \ldots, k$. In this case*

$$\lim_{x\to a} f(x) = (\lim_{x\to a} f^1(x), \ldots, \lim_{x\to a} f^k(x)). \tag{109}$$

PROOF. Suppose $\lim_{x\to a} f(x)$ exists and equals $\mathbf{b} = (b^1, \ldots, b^k)$. We want to show that $\lim_{x\to a} f^i(x)$ exists and equals b^i for $i = 1, \ldots, k$.

Let $(x_n)_{n=1}^{\infty} \subset A \setminus \{a\}$ and $x_n \to a$. From Definition (10.2.1) we have that $\lim f(x_n) = \mathbf{b}$ and it is sufficient to prove that $\lim f^i(x_n) = b^i$. But this is immediate from Theorem (7.5.1) on sequences.

Conversely, if $\lim_{x\to a} f^i(x)$ exists and equals b^i for $i = 1, \ldots, k$, then a similar argument shows $\lim_{x\to a} f(x)$ exists and equals $(b^1, \ldots, b^k)$. □

More Notation Let $f, g : S \to V$ where S is *any* set (not necessarily a subset of a metric space) and V is *any* vector space. In particular, $V = \mathbb{R}$ is an important case. Let $\alpha \in \mathbb{R}$. Then we define *addition* and *scalar multiplication* of functions as follows:

$$\begin{aligned} (f+g)(x) &= f(x) + g(x), \\ (\alpha f)(x) &= \alpha f(x), \end{aligned}$$

for all $x \in S$. That is, $f+g$ is the function defined on S whose value at each $x \in S$ is $f(x)+g(x)$, and similarly for αf. Thus addition of functions is defined by addition of the values of the functions, and similarly for multiplication of a function and a scalar. The *zero function* is the function whose value is everywhere 0. (It is easy to check that the set $\mathcal{F}$ of all functions $f : S \to V$ is a vector space whose "zero vector" is the zero function.)

If $V = \mathbb{R}$ then we define the *product* and *quotient* of functions by

$$\begin{aligned} (fg)(x) &= f(x)g(x), \\ \left(\frac{f}{g}\right)(x) &= \frac{f(x)}{g(x)}. \end{aligned}$$

The domain of f/g is defined to be $S \setminus \{x : g(x) = 0\}$.

If $V = X$ is an inner product space, then we define the inner product of the functions f and g to be the function $f \cdot g : S \to \mathbb{R}$ given by

$$(f \cdot g)(x) = f(x) \cdot g(x).$$

The following algebraic properties of limits follow easily from the corresponding properties of sequences. As usual you should think of the case $X = \mathbb{R}^n$ and $V = \mathbb{R}^n$ (in particular, $m = 1$).

THEOREM 10.4.5. *Let $f, g : A\,(\subset X) \to V$ where X is a metric space and V is a normed space. Let $\lim_{x\to a} f(x)$ and $\lim_{x\to a} g(x)$ exist. Let $\alpha \in \mathbb{R}$. Then the following limits exist and have the stated values:*

$$\begin{aligned} \lim_{x\to a}(f+g)(x) &= \lim_{x\to a} f(x) + \lim_{x\to a} g(x), \\ \lim_{x\to a}(\alpha f)(x) &= \alpha \lim_{x\to a} f(x). \end{aligned}$$

If $V = \mathbb{R}$ then

$$\begin{aligned} \lim_{x\to a}(fg)(x) &= \lim_{x\to a} f(x) \lim_{x\to a} g(x), \\ \lim_{x\to a}\left(\frac{f}{g}\right)(x) &= \frac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)}, \end{aligned}$$

provided in the last case that $g(x) \neq 0$ for all $x \in A \setminus \{a\}$[2] and $\lim_{x\to a} g(x) \neq 0$.

If $X = V$ is an inner product space, then

$$\lim_{x\to a}(f \cdot g)(x) = \lim_{x\to a} f(x) \cdot \lim_{x\to a} g(x).$$

[2]It is often convenient to instead just require that $g(x) \neq 0$ for all $x \in B_r(a) \cap (A \setminus \{a\})$ and some $r > 0$. In this case the function f/g will be defined everywhere in $B_r(a) \cap (A \setminus \{a\})$ and the conclusion still holds.

PROOF. Let $\lim_{x\to a} f(x) = b$ and $\lim_{x\to a} g(x) = c$.

We prove the result for addition of functions.

Let $(x_n) \subset A \setminus \{a\}$ and $x_n \to a$. From Definition 10.2.1 we have that

$$f(x_n) \to b, \quad g(x_n) \to c, \tag{110}$$

and it is sufficient to prove that

$$(f+g)(x_n) \to b + c.$$

But

$$\begin{aligned} (f+g)(x_n) &= f(x_n) + g(x_n) \\ &\to b + c \end{aligned}$$

from (110) and the algebraic properties of limits of sequences, Theorem 7.7.1. This proves the result.

The others are proved similarly. For the second last we also need the Problem in Chapter 7 about the ratio of corresponding terms of two convergent sequences. □

One usually uses the previous Theorem, rather than going back to the original definition, in order to compute limits.

Example If P and Q are polynomials then

$$\lim_{x\to a} \frac{P(x)}{Q(x)} = \frac{P(a)}{Q(a)}$$

if $Q(a) \neq 0$.

To see this, let $P(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$. It follows (*exercise*) from the definition of limit that $\lim_{x\to a} c = c$ (for any real number c) and $\lim_{x\to a} x = a$. It then follows by repeated applications of the previous theorem that $\lim_{x\to a} P(x) = P(a)$. Similarly $\lim_{x\to a} Q(x) = Q(a)$ and so the result follows by again using the theorem.

CHAPTER 11

Continuity

As usual, unless otherwise clear from context, we consider functions $f : A\,(\subset X) \to Y$*, where* X *and* Y *are metric spaces.*

You should think of the case $X = \mathbb{R}^n$ and $Y = \mathbb{R}^n$, and in particular $Y = \mathbb{R}$.

11.1. Continuity at a Point

We first define the notion of continuity at a point in terms of limits, and then we give a few useful equivalent definitions.

The idea is that $f : A \to Y$ is *continuous at* $a \in A$ if $f(x)$ is arbitrarily close to $f(a)$ for all x sufficiently close to a. Thus the value of f does not have a "jump" at a. However, one's intuition can be misleading, as we see in the following examples.

If a is a *limit* point of A, continuity of f at a means $\lim_{x\to a,\ x\in A} f(x) = f(a)$. If a is an *isolated* point of A then $\lim_{x\to a,\ x\in A} f(x)$ is not defined and we *always* define f to be continuous at a in this (uninteresting) case.

DEFINITION 11.1.1. Let $f : A\,(\subset X) \to Y$ where X and Y are metric spaces and let $a \in A$. Then f is *continuous at* a if a is an isolated point of A, or if a is a limit point of A and $\lim_{x\to a,\ x\in A} f(x) = f(a)$.

If f is continuous at *every* $a \in A$ then we say f is *continuous*. The set of all such continuous functions is denoted by

$$\mathcal{C}(A; Y),$$

or by

$$\mathcal{C}(A)$$

if $Y = \mathbb{R}$.

Example 1 Define

$$f(x) = \begin{cases} x \sin(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

From Example 3 of Section 10.2, it follows f is continuous at 0. From the rules about products and compositions of continuous functions, see Example 1 Section 11.2, it follows that f is continuous everywhere on $\mathbb{R}$.

Example 2 From the Example in Section 10.4 it follows that any rational function P/Q, and in particular any polynomial, is continuous everywhere it is defined, i.e. everywhere $Q(x) \neq 0$.

Example 3 Define

$$f(x) = \begin{cases} x & \text{if } x \in \mathbb{Q} \\ -x & \text{if } x \notin \mathbb{Q} \end{cases}$$

Then f is continuous at 0, and only at 0.

Example 4 If g is the function from Example 5 of Section 10.2 then it follows that g is not continuous at any x of the form $k/2^m$ but is continuous everywhere else.

The similar function

$$h(x) = \begin{cases} 0 & \text{if } x \text{ is irrational} \\ 1/q & x = p/q \text{ in simplest terms} \end{cases}$$

is continuous at every irrational and discontinuous at every rational. (*It is possible to prove that there is no function $f : [0,1] \to [0,1]$ such that f is continuous at every rational and discontinuous at every irrational.)

Example 5 Let $g : \mathbb{Q} \to \mathbb{R}$ be given by $g(x) = x$. Then g is continuous at every $x \in \mathbb{Q} = \operatorname{dom} g$ and hence is continuous. On the other hand, if f is the function defined in Example 3, then g agrees with f everywhere in $\mathbb{Q}$, but f is continuous only at 0. The point is that f and g have different domains.

The following equivalent definitions are often useful. They also have the advantage that it is not necessary to consider the case of isolated points and limit points separately. Note that, unlike in Theorem 10.3.1, we allow $x_n = a$ in (2), and we allow $x = a$ in (3).

THEOREM 11.1.2. *Let $f : A (\subset X) \to Y$ where (X, d) and (Y, ρ) are metric spaces. Let $a \in A$. Then the following are equivalent.*

(1) *f is continuous at a;*
(2) *whenever $(x_n)_{n=1}^{\infty} \subset A$ and $x_n \to a$ then $f(x_n) \to f(a)$;*
(3) *for each $\epsilon > 0$ there exists $\delta > 0$ such that*

$$x \in A \text{ and } d(x, a) < \delta \text{ implies } \rho(f(x), f(a)) < \epsilon;$$
$$\text{i.e. } f\Big(B_\delta(a)\Big) \subseteq B_\epsilon\Big(f(a)\Big).$$

PROOF. (1) $\Rightarrow$ (2): Assume (1). Then in particular for any sequence $(x_n) \subset A \setminus \{a\}$ with $x_n \to a$, it follows $f(x_n) \to f(a)$.

In order to prove (2) suppose we have a sequence $(x_n) \subset A$ with $x_n \to a$ (where we allow $x_n = a$). If $x_n = a$ for all $n \geq$ some N then $f(x_n) = f(a)$ for $n \geq N$ and so trivially $f(x_n) \to f(a)$. If this case does not occur then by deleting any x_n with $x_n = a$ we obtain a new (infinite) sequence $x'_n \to a$ with $(x'_n) \subset A \setminus \{a\}$. Since f is continuous at a it follows $f(x'_n) \to f(a)$. As also $f(x_n) = f(a)$ for all terms from (x_n) *not* in the sequence (x'_n), it follows that $f(x_n) \to f(a)$. This proves (2).

(2) $\Rightarrow$ (1): This is immediate, since if $f(x_n) \to f(a)$ whenever $(x_n) \subset A$ and $x_n \to a$, then certainly $f(x_n) \to f(a)$ whenever $(x_n) \subset A \setminus \{a\}$ and $x_n \to a$, i.e. f is continuous at a.

The equivalence of (2) and (3) is proved almost exactly as is the equivalence of the two corresponding conditions (1)and (2) in Theorem 10.3.1. The only essential difference is that we replace $A \setminus \{a\}$ everywhere in the proof there by A. □

Remark Property (3) here is perhaps the simplest to visualize, try giving a diagram which shows this property.

11.2. Basic Consequences of Continuity

Remark (See Figure 1.)

Note that if $f : A \to \mathbb{R}$, f is continuous at a, and $f(a) = r > 0$, then $f(x) > r/2$ for all x sufficiently near a. In particular, f is strictly positive for all x sufficiently near a. This is an immediate consequence of Theorem 11.1.2 (3), since $r/2 < f(x) < 3r/2$ if $d(x, a) < \delta$, say. Similar remarks apply if $f(a) < 0$.

A useful consequence of this observation is that if $f : [a, b] \to \mathbb{R}$ is continuous, and $\int_a^b |f| = 0$, then $f = 0$. (This fact has already been used in Section 5.2.)

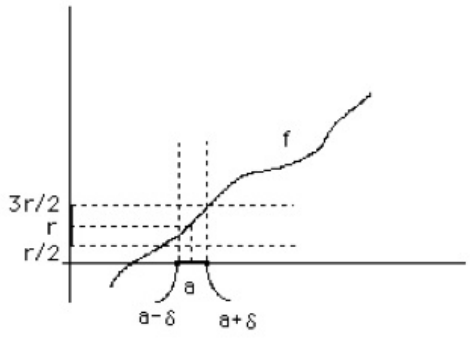

FIGURE 1. Since f is continuous at a and $f(a) = r > 0$, it follows that $f(x) > r/2$ for all x sufficiently near a.

The following two Theorems are proved using Theorem 11.1.2 (2) in the same way as are the corresponding properties for limits. The only difference is that we no longer require sequences $x_n \to a$ with $(x_n) \subset A$ to also satisfy $x_n \neq a$.

THEOREM 11.2.1. *Let $f : A\,(\subset X) \to \mathbb{R}^n$, where X is a metric space. Then f is continuous at a iff f^i is continuous at a for every $i = 1, \ldots, k$.*

PROOF. As for Theorem 10.4.4 □

THEOREM 11.2.2. *Let $f, g : A\,(\subset X) \to V$ where X is a metric space and V is a normed space. Let f and g be continuous at $a \in A$. Let $\alpha \in \mathbb{R}$. Then $f + g$ and αf are continuous at a.*

If $V = \mathbb{R}$ then fg is continuous at a, and moreover f/g is continuous at a if $g(a) \neq 0$.

If $X = V$ is an inner product space then $f \cdot g$ is continuous at a.

PROOF. Using Theorem 11.1.2 (2), the proofs are as for Theorem 10.4.5, except that we take sequences $(x_n) \subset A$ with possibly $x_n = a$. The only extra point is that because $g(a) \neq 0$ and g is continuous at a then from the remark at the beginning of this section, $g(x) \neq 0$ for all x sufficiently near a. □

The following Theorem implies that the composition of continuous functions is continuous.

THEOREM 11.2.3. *Let $f : A\,(\subset X) \to B\,(\subset Y)$ and $g : B \to Z$ where X, Y and Z are metric spaces. If f is continuous at a and g is continuous at $f(a)$, then $g \circ f$ is continuous at a.*

PROOF. Let $(x_n) \to a$ with $(x_n) \subset A$. Then $f(x_n) \to f(a)$ since f is continuous at a. Hence $g(f(x_n)) \to g(f(a))$ since g is continuous at $f(a)$. □

Remark Use of property (3) again gives a simple picture of this result.

Example 1 Recall the function

$$f(x) = \begin{cases} x \sin(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

from Section 11.1. We saw there that f is continuous at 0. Assuming the function $x \mapsto \sin x$ is everywhere continuous[1] and recalling from Example 2 of Section 11.1 that the function $x \mapsto 1/x$ is continuous for $x \neq 0$, it follows from the previous

[1] To prove this we need to first give a proper definition of $\sin x$. This can be done by means of an infinite series expansion.

theorem that $x \mapsto \sin(1/x)$ is continuous if $x \neq 0$. Since the function x is also everywhere continuous and the product of continuous functions is continuous, it follows that f is continuous at every $x \neq 0$, and hence everywhere.

11.3. Lipschitz and Hölder Functions

We now define some classes of functions which, among other things, are very important in the study of partial differential equations. An important case to keep in mind is $A = [a, b]$ and $Y = \mathbb{R}$.

DEFINITION 11.3.1. A function $f : A\,(\subset X) \to Y$, where (X, d) and (Y, ρ) are metric spaces, is a *Lipschitz continuous function* if there is a constant M with

$$\rho(f(x), f(x')) \leq M d(x, x')$$

for all $x, y \in A$. The least such M is the *Lipschitz constant* M of f.

More generally:

DEFINITION 11.3.2. A function $f : A\,(\subset X) \to Y$, where (X, d) and (Y, ρ) are metric spaces, is *Hölder continuous* with exponent $\alpha \in (0, 1]$ if

$$\rho(f(x), f(x')) \leq M d(x, x')^{\alpha}$$

for all $x, x' \in A$ and some fixed constant M.

Remarks

(1) Hölder continuity with exponent $\alpha = 1$ is just Lipschitz continuity.
(2) Hölder continuous functions are continuous. Just choose $\delta = \left(\frac{\epsilon}{M}\right)^{1/\alpha}$ in Theorem 11.1.2(3).
(3) A *contraction map* (recall Section 8.3) has Lipschitz constant $M < 1$, and conversely.

Examples

(1) Let $f : [a, b] \to \mathbb{R}$ be a *differentiable* function and suppose $|f'(x)| \leq M$ for all $x \in [a, b]$. If $x \neq y$ are in $[a, b]$ then from the Mean Value Theorem,

$$\frac{f(y) - f(x)}{y - x} = f'(\xi)$$

for some ξ between x and y. It follows $|f(y) - f(x)| \leq M|y - x|$ and so f is Lipschitz with Lipschitz constant at most M.

(2) An example of a Hölder continuous function defined on $[0, 1]$, which is not Lipschitz continuous, is $f(x) = \sqrt{x}$. This is Hölder continuous with exponent $1/2$ since

$$\begin{aligned} \left|\sqrt{x} - \sqrt{x'}\right| &= \frac{|x - x'|}{\sqrt{x} + \sqrt{x'}} \\ &= \sqrt{|x - x'|}\,\frac{\sqrt{|x - x'|}}{\sqrt{x} + \sqrt{x'}} \\ &\leq \sqrt{|x - x'|}. \end{aligned}$$

This function is *not* Lipschitz continuous since

$$\frac{|f(x) - f(0)|}{|x - 0|} = \frac{1}{\sqrt{x}},$$

and the right side is not bounded by any constant independent of x for $x \in (0, 1]$.

11.4. Another Definition of Continuity

The following theorem gives a definition of "continuous function" in terms only of open (or closed) sets. It does *not* deal with continuity of a function at a point.

THEOREM 11.4.1. *Let $f : X \to Y$, where (X, d) and (Y, ρ) are metric spaces. Then the following are equivalent:*

(1) *f is continuous;*
(2) *$f^{-1}[E]$ is open in X whenever E is open in Y;*
(3) *$f^{-1}[C]$ is closed in X whenever C is closed Y.*

PROOF. (See Figure 2.)
(1) $\Rightarrow$ (2)*:* Assume (1). Let E be open in Y. We wish to show that $f^{-1}[E]$ is open (in X).

Let $x \in f^{-1}[E]$. Then $f(x) \in E$, and since E is open there exists $r > 0$ such that $B_r\Big(f(x)\Big) \subset E$. From Theorem 11.1.2(3) there exists $\delta > 0$ such that $f\Big[B_\delta(x)\Big] \subset B_r\Big(f(x)\Big)$. This implies $B_\delta(x) \subset f^{-1}\Big[B_r\Big(f(x)\Big)\Big]$. But $f^{-1}\Big[B_r\Big(f(x)\Big)\Big] \subset f^{-1}[E]$ and so $B_\delta(x) \subset f^{-1}[E]$.

Thus every point $x \in f^{-1}[E]$ is an interior point and so $f^{-1}[E]$ is open.

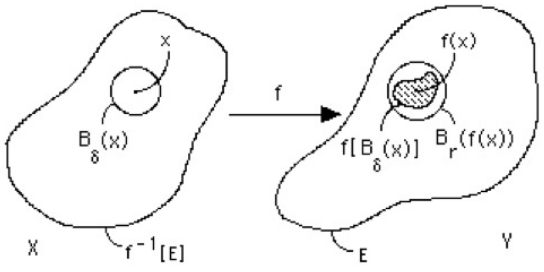

FIGURE 2. Diagram for (1) $\Longrightarrow$ (2) in proof of Theorem 11.4.1.

(2) $\Leftrightarrow$ (3)*:* Assume (2), i.e. $f^{-1}[E]$ is open in X whenever E is open in Y. If C is closed in Y then C^c is open and so $f^{-1}[C^c]$ is open. But $\Big(f^{-1}[C]\Big)^c = f^{-1}[C^c]$. Hence $f^{-1}[C]$ is closed.

We can similarly show (3) $\Rightarrow$ (2).

(2) $\Rightarrow$ (1)*:* Assume (2). We will use Theorem 11.1.2(3) to prove (1).

Let $x \in X$. In order to prove f is continuous at x take any $B_r\Big(f(x)\Big) \subset Y$. Since $B_r\Big(f(x)\Big)$ is open it follows that $f^{-1}\Big[B_r\Big(f(x)\Big)\Big]$ is open. Since $x \in f^{-1}[B_r\Big(f(x)\Big)]$ it follows there exists $\delta > 0$ such that $B_\delta(x) \subset f^{-1}[B_r\Big(f(x)\Big)]$. Hence $f\Big[B_\delta(x)\Big] \subset f\Big[f^{-1}\Big[B_r\Big(f(x)\Big)\Big]\Big]$; but $f\Big[f^{-1}\Big[B_r\Big(f(x)\Big)\Big]\Big] \subset B_r\Big(f(x)\Big)$ (*exercise*) and so $f\Big[B_\delta(x)\Big] \subset B_r\Big(f(x)\Big)$.

It follows from Theorem 11.1.2(3) that f is continuous at x. Since $x \in X$ was arbitrary, it follows that f is continuous on X. □

COROLLARY 11.4.2. *Let $f : S\ (\subset X) \to Y$, where (X, d) and (Y, ρ) are metric spaces. Then the following are equivalent:*

(1) *f is continuous;*

(2) $f^{-1}[E]$ *is open in* S *whenever* E *is open (in* Y*);*
(3) $f^{-1}[C]$ *is closed in* S *whenever* C *is closed (in* Y*).*

PROOF. Since (S, d) is a metric space, this follows immediately from the preceding theorem. □

Note The function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = 0$ if x is irrational and $f(x) = 1$ if x is rational is not continuous anywhere, (this is Example 10.2). However, the function g obtained by restricting f to $\mathbb{Q}$ is continuous everywhere on $\mathbb{Q}$.

Applications The previous theorem can be used to show that, loosely speaking, sets defined by means of continuous functions and the inequalities $<$ and $>$ are open; while sets defined by means of continuous functions, $=$, $\leq$ and $\geq$, are closed.

(1) The half space

$$H\,(\subset \mathbb{R}^n) = \{\mathbf{x} : \mathbf{z} \cdot \mathbf{x} < c\}\,,$$

where $\mathbf{z} \in \mathbb{R}^n$ and c is a scalar, is open (c.f. the Problems on Chapter 6.) To see this, fix z and define $f : \mathbb{R}^n \to \mathbb{R}$ by

$$f(\mathbf{x}) = \mathbf{z} \cdot \mathbf{x}.$$

Then $H = f^{-1}(-\infty, c)$. Since f is continuous[2] and $(-\infty, c)$ is open, it follows H is open.

(2) The set

$$S\,(\subset \mathbb{R}^2) = \left\{(x, y) : x \geq 0 \text{ and } x^2 + y^2 \leq 1\right\}$$

is closed. To see this let $S = S_1 \cap S_2$ where

$$S_1 = \{(x, y) : x \geq 0\}\,, \quad S_2 = \left\{(x, y) : x^2 + y^2 \leq 1\right\}.$$

Then $S_1 = g^{-1}[0, \infty)$ where $g(x, y) = x$. Since g is continuous and $[0, \infty)$ is closed, it follows that S_1 is closed. Similarly $S_2 = f^{-1}[0, 1]$ where $f(x, y) = x^2 + y^2$, and so S_2 is closed. Hence S is closed being the intersection of closed sets.

Remark It is *not* always true that a continuous image[3] of an open set is open; nor is a continuous image of a closed set always closed. But see Theorem 11.5.1 below.

For example, if $f : \mathbb{R} \to \mathbb{R}$ is given by $f(x) = x^2$ then $f[(-1, 1)] = [0, 1)$, so that a continuous image of an open set need not be open. Also, if $f(x) = e^x$ then $f[\mathbb{R}] = (0, \infty)$, so that a continuous image of a closed set need not be closed.

11.5. Continuous Functions on Compact Sets

We saw at the end of the previous section that a continuous image of a closed set need not be closed. However, the continuous image of a closed *bounded* subset of $\mathbb{R}^n$ is a closed bounded set.

More generally, for arbitrary metric spaces the continuous image of a compact set is compact.

THEOREM 11.5.1. *Let* $f : K\,(\subset X) \to Y$ *be a continuous function, where* (X, d) *and* (Y, ρ) *are metric spaces, and* K *is compact. Then* $f[K]$ *is compact.*

PROOF. Let (y_n) be any sequence from $f[K]$. We want to show that some subsequence has a limit in $f[K]$.

Let $y_n = f(x_n)$ for some $x_n \in K$. Since K is compact there is a subsequence (x_{n_i}) such that $x_{n_i} \to x$ (say) as $i \to \infty$, where $x \in K$. Hence $y_{n_i} = f(x_{n_i}) \to f(x)$

[2] If $\mathbf{x}_n \to \mathbf{x}$ then $f\mathbf{x}_n) = \mathbf{z} \cdot \mathbf{x}_n \to \mathbf{z} \cdot \mathbf{x} = f(\mathbf{x})$ from Theorem 7.7.1.

[3] By a *continuous image* we just mean the image under a continuous function.

since f is continuous, and moreover $f(x) \in f[K]$ since $x \in K$. It follows that $f[K]$ is compact. □

You know from your earlier courses on Calculus that a continuous function defined on a closed bounded interval is bounded above and below and has a maximum value and a minimum value. This is generalised in the next theorem.

THEOREM 11.5.2. *Let $f : K\,(\subset X) \to \mathbb{R}$ be a continuous function, where (X, d) is a metric space and K is compact. Then f is bounded (above and below) and has a maximum and a minimum value.*

PROOF. From the previous theorem $f[K]$ is a closed and bounded subset of $\mathbb{R}$. Since $f[K]$ is bounded it has a least upper bound b (say), i.e. $b \geq f(x)$ for all $x \in K$. Since $f[K]$ is closed it follows that $b \in f[K]$[4]. Hence $b = f(x_0)$ for some $x_0 \in K$, and so $f(x_0)$ is the maximum value of f on K.

Similarly, f has a minimum value taken at some point in K. □

Remarks The need for K to be compact in the previous theorem is illustrated by the following examples:

(1) Let $f(x) = 1/x$ for $x \in (0, 1]$. Then f is continuous and $(0, 1]$ is bounded, but f is not bounded above on the set $(0, 1]$.
(2) Let $f(x) = x$ for $x \in [0, 1)$. Then f is continuous and is even bounded above on $[0, 1)$, but does not have a maximum on $[0, 1)$.
(3) Let $f(x) = 1/x$ for $x \in [1, \infty)$. Then f is continuous and is bounded below on $[1, \infty)$ but does not have a minimum on $[1, \infty)$.

11.6. Uniform Continuity

In this Section, you should first think of the case X is an interval in $\mathbb{R}$ and $Y = \mathbb{R}$.

DEFINITION 11.6.1. Let (X, d) and (Y, ρ) be metric spaces. The function $f : X \to Y$ is *uniformly continuous* on X if for each $\epsilon > 0$ there exists $\delta > 0$ such that

$$d(x, x') < \delta \Rightarrow \rho(f(x), f(x')) < \epsilon,$$

for all $x, x' \in X$.

Remark The point is that δ may depend on ϵ, but does *not* depend on x or x'.

Examples

(1) *Hölder continuous (and in particular Lipschitz continuous) functions are uniformly continuous.* To see this, just choose $\delta = \left(\frac{\epsilon}{M}\right)^{1/\alpha}$ in Definition 11.6.1.
(2) The function $f(x) = 1/x$ is continuous at every point in $(0, 1)$ and hence is continuous on $(0, 1)$. But f is *not* uniformly continuous on $(0, 1)$.

    For example, choose $\epsilon = 1$ in the definition of uniform continuity. Suppose $\delta > 0$. By choosing x sufficiently close to 0 (e.g. if $|x| < \delta$) it is clear that there exist x' with $|x - x'| < \delta$ but $|1/x - 1/x'| \geq 1$. This contradicts uniform continuity.
(3) The function $f(x) = \sin(1/x)$ is continuous and bounded, but not uniformly continuous, on $(0, 1)$.

[4] To see this take a sequence from $f[K]$ which converges to b (the existence of such a sequence (y_n) follows from the definition of *least upper bound* by choosing $y_n \in f[K]$, $y_n \geq b - 1/n$.) It follows that $b \in f[K]$ since $f[K]$ is closed.

(4) Also, $f(x) = x^2$ is continuous, but not uniformly continuous, on $\mathbb{R}$(*why?*). On the other hand, f *is* uniformly continuous on any bounded interval from the next Theorem. *Exercise:* Prove this fact directly.

You should think of the previous examples in relation to the following theorem.

THEOREM 11.6.2. *Let $f : S \to Y$ be continuous, where X and Y are metric spaces, $S \subset X$ and S is compact. Then f is uniformly continuous.*

PROOF. If f is *not* uniformly continuous, then there exists $\epsilon > 0$ such that for every $\delta > 0$ there exist $x, y \in S$ with

$$d(x, y) < \delta \text{ and } \rho(f(x), f(y)) \geq \epsilon.$$

Fix some such ϵ and using $\delta = 1/n$, choose two sequences (x_n) and (y_n) such that for all n

$$x_n, y_n \in S,\ d(x_n, y_n) < 1/n,\ \rho(f(x_n), f(y_n)) \geq \epsilon. \tag{111}$$

See Figure 3. Since S is compact, by going to a subsequence if necessary we can suppose that $x_n \to x$ for some $x \in S$. Since

$$d(y_n, x) \leq d((y_n, x_n) + d(x_n, x),$$

and both terms on the right side approach 0, it follows that also $y_n \to x$.

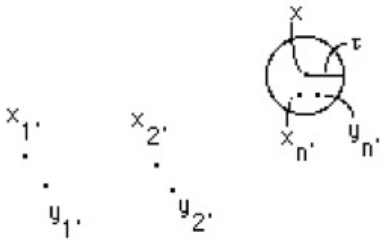

FIGURE 3. Here $x_n \to x$ and $d(x_n, y_n) \to 0$, so $y_n \to x$.

Since f is continuous at x, there exists $\tau > 0$ such that

$$z \in S,\ d(z, x) < \tau \Rightarrow \rho(f(z), f(x)) < \epsilon/2. \tag{112}$$

Since $x_n \to x$ and $y_n \to x$, we can choose k so $d(x_k, x) < \tau$ and $d(y_k, x) < \tau$. It follows from (112) that for this k

$$\begin{aligned} \rho(f(x_k), f(y_k)) &\leq \rho(f(x_k), f(x)) + \rho(f(x), f(y_k)) \\ &< \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

But this contradicts (111). Hence f *is* uniformly continuous. □

COROLLARY 11.6.3. *A continuous real-valued function defined on a closed bounded subset of $\mathbb{R}^n$ is uniformly continuous.*

PROOF. This is immediate from the theorem, as closed bounded subsets of $\mathbb{R}^n$ are compact. □

COROLLARY 11.6.4. *Let K be a continuous real-valued function on the square $[0, 1]^2$. Then the function*

$$f(x) = \int_0^1 K(x, t)dt$$

is (uniformly) continuous.

PROOF. We have

$$|f(x) - f(y)| \leq \int_0^1 |K(x,t) - K(y,t)|dt\,.$$

Uniform continuity of K on $[0,1]^2$ means that given $\epsilon > 0$ there is $\delta > 0$ such that $|K(x,s) - K(y,t)| < \epsilon$ provided $d((x,s),(y,t)) < \delta$. So if $|x-y| < \delta$ $|f(x) - f(y)| < \epsilon$. □

Exercise There is a converse to Corollary 11.6.3: if every function continuous on a subset of $\mathbb{R}$ is uniformly continuous, then the set is closed. Must it also be bounded?

CHAPTER 12

Uniform Convergence of Functions

12.1. Discussion and Definitions

Consider the following examples of sequences of functions $(f_n)_{n=1}^{\infty}$, with graphs as shown. In each case f is in some sense the limit function, as we discuss subsequently.

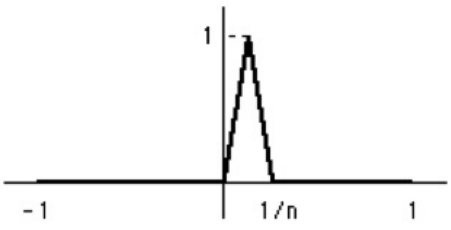

FIGURE 1. The graph of the continuous f_n is shown. Here $f_n \to \mathbf{0}$ pointwise where $\mathbf{0}$ is the zero function.

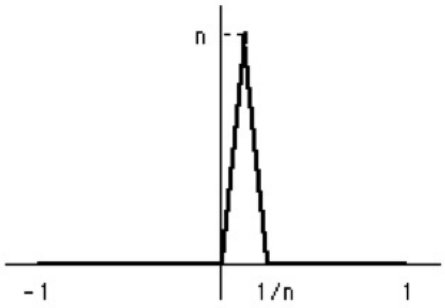

FIGURE 2. The graph of the continuous f_n is shown. Then $f_n \to \mathbf{0}$ pointwise, where $\mathbf{0}$ is the zero function. Note that $\int f_n = 1/2$ for all n, but $\int f = 0$.

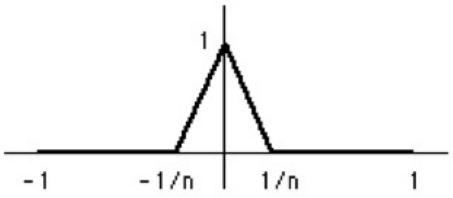

FIGURE 3. The graph of the continuous f_n is shown. Then $f_n \to \mathbf{0}$ pointwise, where $f(x) = 0$ if $x \neq 0$, and $f(0) = 1$.

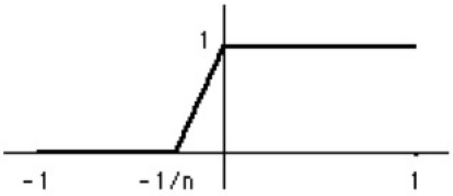

FIGURE 4. The graph of the continuous f_n is shown. Then $f_n \to \mathbf{0}$ pointwise, where $f(x) = 0$ if $x < 0$ and $f(x) = 1$ if $x \geq 0$.

In *every* one of the preceding cases, $f_n(x) \to f(x)$ as $n \to \infty$ for each $x \in \mathrm{dom} f$, where $dom f$ is the domain of f.

For example, in 1, consider the cases $x \leq 0$ and $x > 0$ separately. If $x \leq 0$ then $f_n(x) = 0$ for all n, and so it is certainly true that $f_n(x) \to 0$ as $n \to \infty$. On the

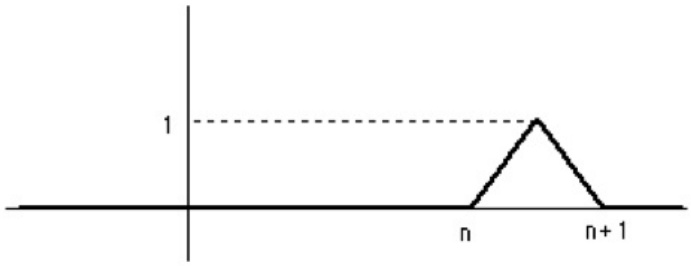

FIGURE 5. The graph of the continuous f_n is shown. Then $f_n \to \mathbf{0}$ pointwise, where $\mathbf{0}$ is the zero function.

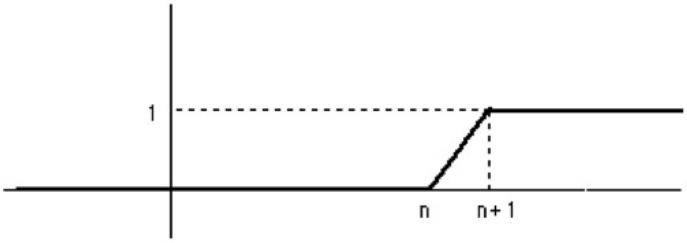

FIGURE 6. The graph of the continuous f_n is shown. Then $f_n \to \mathbf{0}$ pointwise, where $\mathbf{0}$ is the zero function.

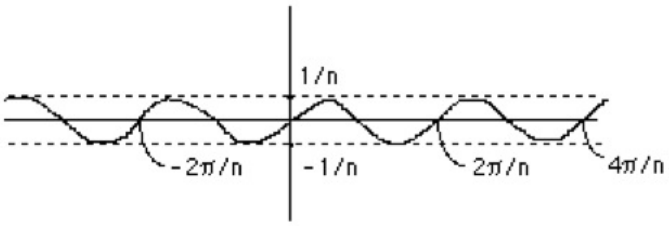

FIGURE 7. The graph of the continuous f_n is shown, where $f_n(x) = \frac{1}{n}\sin nx$. Then $f_n \to \mathbf{0}$ pointwise, where $\mathbf{0}$ is the zero function.

other hand, if $x > 0$, then $f_n(x) = 0$ for all $n > 1/x$ [1], and in particular $f_n(x) \to 0$ as $n \to \infty$.

In *all* cases we say that $f_n \to f$ in the *pointwise sense*. That is, for each $\epsilon > 0$ and each x there exists N such that

$$n \geq N \Rightarrow |f_n(x) - f(x)| < \epsilon.$$

In cases 1–6, N depends on x as well as ϵ: there is no N which works for all x.

[1] Notice that how large n needs to be depends on x.

We can see this by imagining the "ϵ-strip" about the graph of f, which we define to be the set of all points $(x, y) \in \mathbb{R}^2$ such that

$$f(x) - \epsilon < y < f(x) + \epsilon.$$

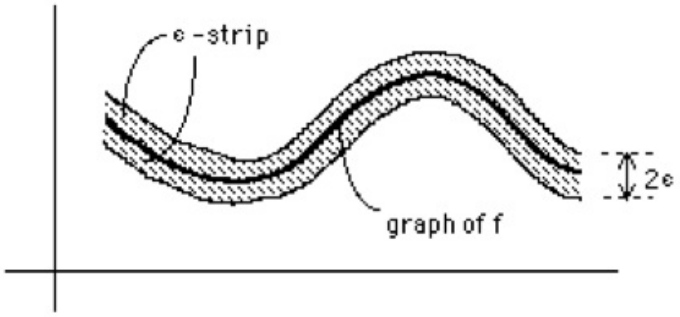

FIGURE 8. The ϵ-strip around the graph of f.

Then it is not the case for Examples 1–6 that the graph of f_n is a subset of the ϵ-strip about the graph of f for all sufficiently large n.

However, in Example 7, since

$$|f_n(x) - f(x)| = |\frac{1}{n} \sin nx - 0| \leq \frac{1}{n},$$

it follows that $|f_n(x) - f(x)| < \epsilon$ for all $n > 1/\epsilon$. In other words, the graph of f_n *is* a subset of the ϵ-strip about the graph of f for all sufficiently large n. In this case we say that $f_n \to f$ *uniformly*.

Finally we remark that in Examples 5 and 6, if we consider f_n and f restricted to any *fixed* bounded set B, then it *is* the case that $f_n \to f$ uniformly *on* B.

Motivated by the preceding examples we now make the following definitions.

In the following think of the case $S = [a, b]$ *and* $Y = \mathbb{R}$.

DEFINITION 12.1.1. Let $f, f_n : S \to Y$ for $n = 1, 2, \ldots$, where S is any set and (Y, ρ) is a metric space.

If $f_n(x) \to f(x)$ for all $x \in S$ then $f_n \to f$ *pointwise*.

If for every $\epsilon > 0$ there exists N such that

$$n \geq N \Rightarrow \rho\Big(f_n(x), f(x)\Big) < \epsilon$$

for all $x \in S$, then $f_n \to f$ *uniformly (on* S*)*.

Remarks (i) Informally, $f_n \to f$ uniformly means $\rho\Big(f_n(x), f(x)\Big) \to 0$ "uniformly" in x. See also Proposition 12.2.3.

(ii) If $f_n \to f$ uniformly then clearly $f_n \to f$ pointwise, but not necessarily conversely, as we have seen. However, see the next theorem for a partial converse.

(iii) Note that $f_n \to f$ does not converge uniformly iff there exists $\epsilon > 0$ and a sequence $(x_n) \subset S$ such that $|f(x) - f_n(x_n)| \geq \epsilon$ for all n.

It is also convenient to define the notion of a *uniformly Cauchy* sequence of functions.

DEFINITION 12.1.2. Let $f_n : S \to Y$ for $n = 1, 2, \ldots$, where S is any set and (Y, ρ) is a metric space. Then the sequence (f_n) is *uniformly Cauchy* if for every

$\epsilon > 0$ there exists N such that

$$m, n \geq N \Rightarrow \rho\Big(f_n(x), f_m(x)\Big) < \epsilon$$

for all $x \in S$.

Remarks (i) Thus (f_n) is uniformly Cauchy iff the following is true: for each $\epsilon > 0$, any two functions from the sequence after a certain point (which will depend on ϵ) lie within the ϵ-strip of each other.
(ii) Informally, (f_n) is uniformly Cauchy if $\rho\Big(f_n(x), f_m(x)\Big) \to 0$ "uniformly" in x as $m, n \to \infty$.
(iii) We will see in the next section that if Y is complete (e.g. $Y = \mathbb{R}^n$) then a sequence (f_n) (where $f_n : S \to Y$) is uniformly Cauchy iff $f_n \to f$ uniformly for some function $f : S \to Y$.

THEOREM 12.1.3 (Dini's Theorem). *Suppose (f_n) is an increasing sequence (i.e. $f_1(x) \leq f_2(x) \leq \ldots$ for all $x \in S$) of real-valued continuous functions defined on the compact subset S of some metric space (X, d). Suppose $f_n \to f$ pointwise and f is continuous. Then $f_n \to f$ uniformly.*

PROOF. Suppose $\epsilon > 0$. For each n let

$$A_n^\epsilon = \{x \in S : f(x) - f_n(x) < \epsilon\}.$$

Since (f_n) is an increasing sequence,

$$A_1^\epsilon \subset A_2^\epsilon \subset \ldots. \tag{113}$$

Since $f_n(x) \to f(x)$ for all $x \in S$,

$$S = \bigcup_{n=1}^{\infty} A_n^\epsilon. \tag{114}$$

Since f_n and f are continuous,

$$A_n^\epsilon \text{ is open in } S \tag{115}$$

for all n (see Corollary 11.4.2).

In order to prove uniform convergence, it is sufficient (*why*?) to show there exists n (depending on ϵ) such that

$$S = A_n^\epsilon \tag{116}$$

(note that then $S = A_m^\epsilon$ for all $m > n$ from (113)). If no such n exists then for each n there exists x_n such that

$$x_n \in S \setminus A_n^\epsilon \tag{117}$$

for all n. By compactness of S,there is a subsequence x_{n_k} with $x_{n_k} \to x_0$(say) $\in S$.

From (114) it follows $x_0 \in A_N^\epsilon$ for some N. From (115) it follows $x_{n_k} \in A_N^\epsilon$ for all $n_k \geq M$(say), where we may take $M \geq N$. But then from (113) we have

$$x_{n_k} \in A_{n_k}^\epsilon$$

for all $n_k \geq M$. This contradicts (117). Hence (116) *is* true for some n, and so the theorem is proved. □

Remarks (i) The same result hold if we replace "increasing" by "decreasing". The proof is similar; or one can deduce the result directly from the theorem by replacing f_n and f by $-f_n$ and $-f$ respectively.
(ii) The proof of the Theorem can be simplified by using the equivalent definition of compactness given in Definition 15.1.2. See the *Exercise* after Theorem 15.2.1.
(iii) *Exercise*: Give examples to show why it is necessary that the sequence be increasing (or decreasing) and that f be continuous.

12.2. The Uniform Metric

In the following think of the case $S = [a, b]$ and $Y = \mathbb{R}$.

In order to understand uniform convergence it is useful to introduce the *uniform "metric"* d_u. This is not quite a metric, only because it may take the value $+\infty$, but it otherwise satisfies the three axioms for a metric.

DEFINITION 12.2.1. Let $\mathcal{F}(S, Y)$ be the set of all functions $f : S \to Y$, where S is a set and (Y, ρ) is a metric space. Then the *uniform "metric"* d_u on $\mathcal{F}(S, Y)$ is defined by

$$d_u(f, g) = \sup_{x \in S} \rho\Big(f(x), g(x)\Big).$$

If Y is a normed space (e.g. $\mathbb{R}$), then we define the *uniform "norm"* by

$$||f||_u = \sup_{x \in S} ||f(x)||.$$

(Thus in this case the uniform "metric" is the metric corresponding to the uniform "norm", as in the examples following Definition 6.2.1)

In the case $S = [a, b]$ and $Y = \mathbb{R}$ it is clear that the ϵ-strip about f is precisely the set of functions g such that $d_u(f, g) < \epsilon$. A similar remark applies for general S and Y if the "ϵ-strip" is appropriately defined.

The *uniform metric (norm)* is also known as the *sup metric (norm)*. We have in fact already defined the sup metric and norm on the set $\mathcal{C}[a, b]$ of continuous real-valued functions; c.f. the examples of Section 5.2 and Section 6.2. The present definition just generalises this to other classes of functions.

The distance d_u between two functions can easily be $+\infty$. For example, let $S = [0, 1]$, $Y = \mathbb{R}$. Let $f(x) = 1/x$ if $x \neq 0$ and $f(0) = 0$, and let g be the zero function. Then clearly $d_u(f, g) = \infty$. In applications we will only be interested in the case $d_u(f, g)$ is finite, and in fact small.

We now show the three axioms for a metric are satisfied by d_u, provided we *define*

$$\infty + \infty = \infty, \quad c\,\infty = \infty \text{ if } c > 0, \quad c\,\infty = 0 \text{ if } c = 0. \tag{118}$$

THEOREM 12.2.2. *The uniform "metric" ("norm") satisfies the axioms for a metric space (Definition 6.2.1) (Definition 5.2.1) provided we interpret arithmetic operations on ∞ as in (118).*

PROOF. It is easy to see that d_u satisfies positivity and symmetry (*Exercise*). For the triangle inequality let $f, g, h : S \to Y$. Then[2]

$$\begin{aligned} d_u(f, g) &= \sup_{x \in S} \rho\Big(f(x), g(x)\Big) \\ &\leq \sup_{x \in S} \Big[\rho\Big(f(x), h(x)\Big) + \rho\Big(h(x), g(x)\Big)\Big] \\ &\leq \sup_{x \in S} \rho\Big(f(x), h(x)\Big) + \sup_{x \in S} \rho\Big(h(x), g(x)\Big) \\ &= d_u(f, h) + d_u(h, g). \end{aligned}$$

This completes the proof in this case. □

PROPOSITION 12.2.3. *A sequence of functions is uniformly convergent iff it is convergent in the uniform metric.*

[2]The third line uses uses the fact *(Exercise)* that if u and v are two real-valued functions defined on the same domain S, then $\sup_{x \in S}(u(x) + v(x)) \leq \sup_{x \in S} u(x) + \sup_{x \in S} v(x)$.

PROOF. Let S be a set and (Y, ρ) be a metric space.

Let $f, f_n \in \mathcal{F}(S, Y)$ for $n = 1, 2, \ldots$. From the definition we have that $f_n \to f$ uniformly iff for every $\epsilon > 0$ there exists N such that

$$n \geq N \Rightarrow \rho\Big(f_n(x), f(x)\Big) < \epsilon$$

for all $x \in S$. This is equivalent to

$$n \geq N \Rightarrow d_u(f_n, f) < \epsilon.$$

The result follows. □

PROPOSITION 12.2.4. *A sequence of functions is uniformly Cauchy iff it is Cauchy in the uniform metric.*

PROOF. As in previous result. □

We next establish the relationship between uniformly convergent, and uniformly Cauchy, sequences of functions.

THEOREM 12.2.5. *Let S be a set and (Y, ρ) be a metric space.*

If $f, f_n \in \mathcal{F}(S, Y)$ and $f_n \to f$ uniformly, then $(f_n)_{n=1}^{\infty}$ is uniformly Cauchy.

Conversely, if (Y, ρ) is a **complete** *metric space, and $(f_n)_{n=1}^{\infty} \subset \mathcal{F}(S, Y)$ is uniformly Cauchy, then $f_n \to f$ uniformly for some $f \in \mathcal{F}(S, Y)$.*

PROOF. First suppose $f_n \to f$ uniformly.

Let $\epsilon > 0$. Then there exists N such that

$$n \geq N \Rightarrow \rho(f_n(x), f(x)) < \epsilon$$

for all $x \in S$. Since

$$\rho\Big(f_n(x), f_m(x)\Big) \leq \rho\Big(f_n(x), f(x)\Big) + \rho\Big(f(x), f_m(x)\Big),$$

it follows

$$m, n \geq N \Rightarrow \rho(f_n(x), f_m(x)) < 2\epsilon$$

for all $x \in S$. Thus (f_n) is uniformly Cauchy.

Next assume (Y, ρ) is *complete* and suppose (f_n) is a uniformly Cauchy sequence.

It follows from the definition of *uniformly Cauchy* that $\Big(f_n(x)\Big)$ is a Cauchy sequence for each $x \in S$, and so has a limit in Y since Y is complete. Define the function $f : S \to Y$ by

$$f(x) = \lim_{n \to \infty} f_n(x)$$

for each $x \in S$.

We know that $f_n \to f$ in the pointwise sense, but we need to show that $f_n \to f$ uniformly.

So suppose that $\epsilon > 0$ and, using the fact that (f_n) is uniformly Cauchy, choose N such that

$$m, n \geq N \Rightarrow \rho\Big(f_n(x), f_m(x)\Big) < \epsilon$$

for all $x \in S$. Fixing $m \geq N$ and letting $n \to \infty$[3], it follows from the Comparison Test that

$$\rho\Big(f(x), f_m(x)\Big) \leq \epsilon$$

for all $x \in S$[4].

[3] This is a commonly used technique; it will probably seem strange at first.

[4] In more detail, we argue as follows: Every term in the sequence of real numbers

$$\rho\Big(f_N(x), f_m(x)\Big),\ \rho\Big(f_{N+1}(x), f_m(x)\Big),\ \rho\Big(f_{N+2}(x), f_m(x)\Big), \ldots$$

Since this applies to *every* $m \geq N$, we have that

$$m \geq N \Rightarrow \rho\Big(f(x), f_m(x)\Big) \leq \epsilon.$$

for all $x \in S$. Hence $f_n \to f$ uniformly. □

12.3. Uniform Convergence and Continuity

In the following think of the case $X = [a, b]$ *and* $Y = \mathbb{R}$.

We saw in Examples 3 and 4 of Section 12.1 that a *pointwise* limit of continuous functions need not be continuous. The next theorem shows however that a *uniform* limit of continuous functions *is* continuous.

THEOREM 12.3.1. *Let* (X, d) *and* (Y, ρ) *be metric spaces. Let* $f_n : X \to Y$ *for* $n = 1, 2, \ldots$ *be a sequence of continuous functions such that* $f_n \to f$ *uniformly. Then* f *is continuous.*

PROOF. (See Figure 9.) Consider any $x_0 \in X$; we will show f is continuous at x_0.

Suppose $\epsilon > 0$. Using the fact that $f_n \to f$ uniformly, *first* choose N so that

$$\rho\Big(f_N(x), f(x)\Big) < \epsilon \tag{119}$$

for *all* $x \in X$. *Next*, using the fact that f_N is continuous, choose $\delta > 0$ so that

$$d(x, x_0) < \delta \Rightarrow \rho\Big(f_N(x), f_N(x_0)\Big) < \epsilon. \tag{120}$$

It follows from (119) and (120) that if $d(x, x_0) < \delta$ then

$$\begin{aligned} \rho\Big(f(x), f(x_0)\Big) &\leq \rho\Big(f(x), f_N(x)\Big) + \rho\Big(f_N(x), f_N(x_0)\Big) + \rho\Big(f_N(x_0), f(x_0)\Big) \\ &< 3\epsilon. \end{aligned}$$

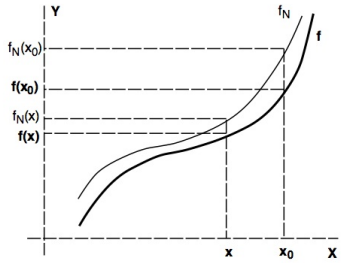

FIGURE 9. Diagram for the proof of Thorem 12.3.1.

Hence f is continuous at x_0, and hence continuous as x_0 was an arbitrary point in X. □

The next result is very important. We will use it in establishing the existence of solutions to systems of differential equations.

Recall from Definition 11.1.1 that if X and Y are metric spaces and $A \subset X$, then the set of all continuous functions $f : A \to Y$ is denoted by $\mathcal{C}(A; Y)$. If A

---

is $< \epsilon$. Since $f_{N+p}(x) \to f(x)$ as $p \to \infty$, it follows that $\rho\Big(f_{N+p}(x), f_m(x)\Big) \to \rho\Big(f(x), f_m(x)\Big)$ as $p \to \infty$ (this is clear if Y is $\mathbb{R}$ or $\mathbb{R}^k$, and follows in general from Theorem 7.3.4). By the Comparison Test it follows that $\rho\Big(f(x), f_m(x)\Big) \leq \epsilon$.

is compact, then we have seen that $f[A]$ is compact and hence bounded[5], i.e. f is bounded. If A is not compact, then continuous functions need not be bounded[6].

DEFINITION 12.3.2. The set of *bounded* continuous functions $f : A \to Y$ is denoted by

$$\mathcal{BC}(A;Y).$$

THEOREM 12.3.3. *Suppose $A \subset X$, (X,d) is a metric space and (Y,ρ) is a* **complete** *metric spaces. Then $\mathcal{BC}(A;Y)$ is a complete metric space with the uniform metric d_u.*

PROOF. It has already been verified in Theorem 12.2.2 that the three axioms for a metric are satisfied. We need only check that $d_u(f,g)$ is always finite for $f,g \in \mathcal{BC}(A;Y)$.

But this is immediate. For suppose $b \in Y$. Then since f and g are bounded on A, it follows there exist K_1 and K_2 such that $\rho(f(x),b) \le K_1$ and $\rho(g(x),b) \le K_2$ for all $x \in A$. But then $d_u(f,g) \le K_1 + K_2$ from the definition of d_u and the triangle inequality. Hence $\mathcal{BC}(A;Y)$ is a metric space with the uniform metric.

In order to verify completeness, let $(f_n)_{n=1}^{\infty}$ be a Cauchy sequence from $\mathcal{BC}(A;Y)$. Then (f_n) is uniformly Cauchy, as noted in Proposition 12.2.4. From Theorem 12.2.5 it follows that $f_n \to f$ uniformly, for some function $f : A \to Y$. From Proposition 12.2.3 it follows that $f_n \to f$ in the uniform metric.

From Theorem 12.3.1 it follows that f is continuous. It is also clear that f is bounded[7]. Hence $f \in \mathcal{BC}(A;Y)$.

We have shown that $f_n \to f$ in the sense of the uniform metric d_u, where $f \in \mathcal{BC}(A;Y)$. Hence $(\mathcal{BC}(A;Y), d_u)$ is a complete metric space. □

COROLLARY 12.3.4. *Let (X,d) be a metric space and (Y,ρ) be a complete metric space. Let $A \subset X$ be compact. Then $\mathcal{C}(A;Y)$ is a complete metric space with the uniform metric d_u.*

PROOF. Since A is compact, every continuous function defined on A is bounded. The result now follows from the Theorem. □

COROLLARY 12.3.5. *The set $\mathcal{C}[a,b]$ of continuous real-valued functions defined on the interval $[a,b]$, and more generally the set $\mathcal{C}([a,b] : \mathbb{R}^n)$ of continuous maps into $\mathbb{R}^n$, are complete metric spaces with the sup metric.*

We will use the previous corollary to find solutions to (systems of) differential equations.

12.4. Uniform Convergence and Integration

It is not necessarily true that if $f_n \to f$ *pointwise*, where $f, f_n : [a,b] \to \mathbb{R}$ are continuous, then $\int_a^b f_n \to \int_a^b f$. In particular, in Example 2 from Section 12.1, $\int_{-1}^{1} f_n = 1/2$ for all n but $\int_{-1}^{1} f = 0$. However, integration is better behaved under *uniform convergence.*

THEOREM 12.4.1. *Suppose that $f, f_n : [a,b] \to \mathbb{R}$ for $n = 1, 2, \ldots$ are continuous functions and $f_n \to f$ uniformly. Then*

$$\int_a^b f_n \to \int_a^b f.$$

---

[5] In $\mathbb{R}^n$, *compactness* is the same as *closed and bounded*. This is not true in general, but it *is* always true that *compact* implies *closed and bounded*. The proof is the same as in Corollary 9.2.2.

[6] Let $A = (0,1)$ and $f(x) = 1/x$, or $A = \mathbb{R}$ and $f(x) = x$.

[7] Choose N so $d_u(f_N, f) \le 1$. Choose any $b \in Y$. Since f_N is bounded, $\rho(f_N(x), b) \le K$ say, for all $x \in A$. It follows $\rho(f(x), b) \le K + 1$ for all $x \in A$.

Moreover, $\int_a^x f_n \to \int_a^x f$ *uniformly for* $x \in [a,b]$.

PROOF. Suppose $\epsilon > 0$. By uniform convergence, choose N so that

$$n \geq N \Rightarrow |f_n(x) - f(x)| < \epsilon \tag{121}$$

for all $x \in [a,b]$

Since f_n and f are continuous, they are Riemann integrable. Moreover,[8] for $n \geq N$

$$\begin{array}{rcl} \left|\int_a^b f_n - \int_a^b f\right| & = & \left|\int_a^b \Big(f_n - f\Big)\right| \\ & \leq & \int_a^b |f_n - f| \\ & \leq & \int_a^b \epsilon \qquad \text{from (121)} \\ & = & (b-a)\epsilon. \end{array}$$

It follows that $\int_a^b f_n \to \int_a^b f$, as required.

For uniform convergence just note that the same proof gives $\left|\int_a^x f_n - \int_a^x f\right| \leq (x-a)\epsilon \leq (b-a)\epsilon$. □

***Remarks**

(1) More generally, it is not hard to show that the uniform limit of a sequence of Riemann integrable functions is also Riemann integrable, and that the corresponding integrals converge. See [**Sm**, Theorem 4.4, page 101].

(2) There is a much more important notion of integration, called *Lebesgue integration*. Lebesgue integration has much nicer properties with respect to convergence than does Riemann integration. See, for example, [**Fl**], [**St**] and [**Sm**].

12.5. Uniform Convergence and Differentiation

Suppose that $f_n : [a,b] \to \mathbb{R}$, for $n = 1, 2, \ldots$, is a sequence of differentiable functions, and that $f_n \to f$ uniformly. It is not true that $f_n'(x) \to f'(x)$ for all $x \in [a,b]$, in fact it need not even be true that f is differentiable.

For example, let $f(x) = |x|$ for $x \in [0,1]$. Then f is not differentiable at 0. But, as indicated in the following diagram, it is easy to find a sequence (f_n) of differentiable functions such that $f_n \to f$ uniformly.

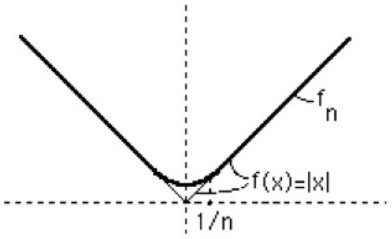

FIGURE 10. The uniform limit of a sequence of differentiable functions may not be differentiable.

[8] For the following, recall

$$\left|\int_a^b g\right| \leq \int_a^b |g|,$$

$$f(x) \leq g(x) \text{ for all } x \in [a,b] \Rightarrow \int_a^b f \leq \int_a^b g.$$

In particular, let

$$f_n(x) = \begin{cases} \frac{n}{2}x^2 + \frac{1}{2n} & 0 \le |x| \le \frac{1}{n} \\ |x| & \frac{1}{n} \le |x| \le 1 \end{cases}$$

Then the f_n are differentiable on $[-1, 1]$ (the only points to check are $x = \pm 1/n$), and $f_n \to f$ uniformly since $d_u(f_n, f) \le 1/n$.

Example 7 from Section 12.1 gives an example where $f_n \to f$ uniformly and f is differentiable, but f_n' does not converge for most x. In fact, $f_n'(x) = \cos nx$ which does not converge (unless $x = 2k\pi$ for some $k \in \mathbb{Z}$ (*exercise*)).

However, if the derivatives themselves converge uniformly to some limit, then we have the following theorem.

THEOREM 12.5.1. *Suppose that $f_n : [a, b] \to \mathbb{R}$ for $n = 1, 2, \ldots$ and that the f_n' exist and are continuous. Suppose $f_n \to f$ pointwise on $[a, b]$ and (f_n') converges uniformly on $[a, b]$.*

Then f' exists and is continuous on $[a, b]$ and $f_n' \to f'$ uniformly on $[a, b]$. Moreover, $f_n \to f$ uniformly on $[a, b]$.

PROOF. By the Fundamental Theorem of Calculus,

$$\int_a^x f_n' = f_n(x) - f_n(a) \tag{122}$$

for every $x \in [a, b]$.

Let $f_n' \to g$(say) uniformly. Then from (122), Theorem 12.4.1 and the hypotheses of the theorem,

$$\int_a^x g = f(x) - f(a). \tag{123}$$

Since g is continuous, the left side of (123) is differentiable on $[a, b]$ and the derivative equals g[9]. Hence the right side is also differentiable and moreover

$$g(x) = f'(x)$$

on $[a, b]$.

Thus f' exists and is continuous and $f_n' \to f'$ uniformly on $[a, b]$.

Since $f_n(x) = \int_a^x f_n'$ and $f(x) = \int_a^x f'$, uniform convergence of f_n to f now follows from Theorem 12.4.1. □

[9]Recall that the integral of a continuous function is differentiable, and the derivative is just the original function.

CHAPTER 13

First Order Systems of Differential Equations

The main result in this Chapter is the *Existence and Uniqueness Theorem* for *first order systems* of (ordinary) differential equations. Essentially any differential equation or system of differential equations can be reduced to a first-order system, so the result is very general. The Contraction Mapping Principle is the main ingredient in the proof.

The local Existence and Uniqueness Theorem for a single equation, together with the necessary preliminaries, is in Sections 13.3, 13.7–13.9. See Sections 13.10 and 13.11 for the global result and the extension to systems. These sections are independent of the remaining sections.

In Section 13.1 we give two interesting examples of systems of differential equations.

In Section 13.2 we show how higher order differential equations (and more generally higher order systems) can be reduced to first order systems.

In Sections 13.4 and 13.5 we discuss "geometric" ways of analysing and understanding the solutions to systems of differential equations.

In Section 13.6 we give two examples to show the necessity of the conditions assumed in the Existence and Uniqueness Theorem.

13.1. Examples

13.1.1. Predator-Prey Problem. Suppose there are two species of animals, and let the populations at time t be $x(t)$ and $y(t)$ respectively. We assume we can approximate $x(t)$ and $y(t)$ by differentiable functions. Species x is eaten by species y . The rates of increase of the species are given by

$$\begin{aligned} \frac{dx}{dt} &= ax - bxy - ex^2, \\ \frac{dy}{dt} &= -cy + dxy - fy^2. \end{aligned} \tag{124}$$

The quantities a, b, c, d, e, f are constants and depend on the environment and the particular species.

A quick justification of this model is as follows:

> The term ax represents the usual rate of growth of x in the case of an unlimited food supply and no predators. The term bxy comes from the number of contacts per unit time between predator and prey, it is proportional to the populations x and y, and represents the rate of decrease in species x due to species y eating it. The term ex^2 is similarly due to competition between members of species x for the limited food supply.
>
> The term $-cy$ represents the natural rate of decrease of species y if its food supply, species x, were removed. The term dxy is proportional to the number of contacts per unit time between predator and prey, and accounts for the growth rate of y in the absence

of other effects. The term fy^2 accounts for competition between members of species y for the limited food supply (species x).

We will return to this system later. It is *first order*, since only first derivatives occur in the equation, and *nonlinear*, since some of the terms involving the *unknowns* (or *dependent variables*) x and y occur in a nonlinear way (namely the terms xy, x^2 and y^2). It is a system of *ordinary* differential equations since there is only *one independent* variable t, and so we only form *ordinary* derivatives; as opposed to differential equations where there are two or more independent variables, in which case the differential equation(s) will involve *partial* derivatives.

13.1.2. A Simple Spring System. Consider a body of mass m connected to a wall by a spring and sliding over a surface which applies a frictional force, as shown in the following diagram.

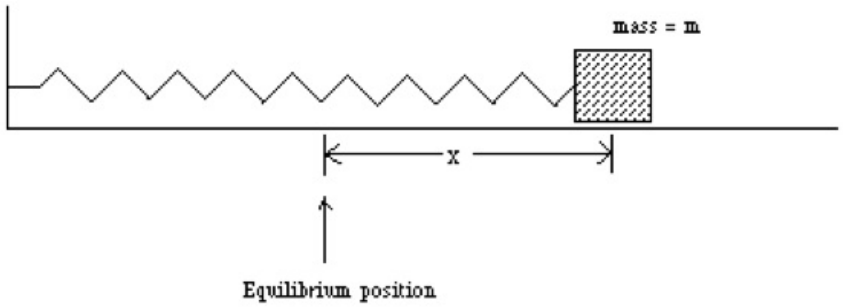

FIGURE 1. A mass on a spring displaced from its equilibrium position.

Let $x(t)$ be the displacement at time t from the equilibrium position. From Newton's second law, the force acting on the mass is given by

$$\textbf{Force} = mx''(t).$$

If the spring obeys Hooke's law, then the force is proportional to the displacement, but acts in the opposite direction, and so

$$\textbf{Force} = -kx(t),$$

for some constant $k > 0$ which depends on the spring. Thus

$$mx''(t) = -kx(t),$$

i.e.

$$mx''(t) + kx(t) = 0.$$

If there is also a force term, due to friction, and proportional to the velocity but acting in the opposite direction, then

$$\textbf{Force} = -kx - cx',$$

for some constant $c > 0$, and so

$$mx''(t) + cx'(t) + kx(t) = 0. \tag{125}$$

This is a *second order* ordinary differential equation, since it contains *second* derivatives of the "unknown" x, and is *linear* since the unknown and its derivatives occur in a *linear* manner.

13.2. Reduction to a First Order System

It is usually possible to reduce a *higher order* ordinary differential equation or system of ordinary differential equations to a *first order* system.

For example, in the case of the differential equation (125) for the spring system in Section 13.1.2, we introduce a new variable y corresponding to the velocity x', and so obtain the following first order system for the "unknowns" x and y:

$$\begin{aligned} x' &= y \\ y' &= -m^{-1}cy - m^{-1}kx \end{aligned} \tag{126}$$

This is a *first order* system (linear in this case).

If x, y is a solution of (126) then it is clear that x is a solution of (125). Conversely, if x is a solution of (125) and we define $y(t) = x'(t)$, then x, y is a solution of (126).

An *nth order* differential equation is a relation between a function x and its first n derivatives. We can write this in the form

$$F\Big(x^{(n)}(t), x^{(n-1)}(t), \ldots, x'(t), x(t), t\Big) = 0,$$

or

$$F\Big(x^{(n)}, x^{(n-1)}, \ldots, x', x, t\Big) = 0.$$

Here $t \in I$ for some interval $I \subset \mathbb{R}$, where I may be open, closed, or infinite, at either end. If $I = [a, b]$, say, then we take one-sided derivatives at a and b.

One can usually, in principle, solve this for x^n, and so write

$$x^{(n)} = G\Big(x^{(n-1)}, \ldots, x', x, t\Big). \tag{127}$$

In order to reduce this to a first order system, introduce new functions $x^1, x^2, \ldots, x^n$, where

$$\begin{aligned} x^1(t) &= x(t) \\ x^2(t) &= x'(t) \\ x^3(t) &= x''(t) \\ &\vdots \\ x^n(t) &= x^{(n-1)}(t). \end{aligned}$$

Then from (127) we see (*exercise*) that $x^1, x^2, \ldots, x^n$ satisfy the *first order* system

$$\begin{aligned} \frac{dx^1}{dt} &= x^2(t) \\ \frac{dx^2}{dt} &= x^3(t) \\ \frac{dx^3}{dt} &= x^4(t) \\ &\vdots \\ \frac{dx^{n-1}}{dt} &= x^n(t) \\ \frac{dx^n}{dt} &= G(x^n, \ldots, x^2, x^1, t) \end{aligned} \tag{128}$$

Conversely, if $x^1, x^2, \ldots, x^n$ satisfy (128) and we let $x(t) = x^1(t)$, then we can check (*exercise*) that x satisfies (127).

13.3. Initial Value Problems

Notation If x is a real-valued function defined in some interval I, we say x is *continuously differentiable* (or C^1) if x is differentiable on I and its derivative is continuous on I. Note that since x is differentiable, it is in particular continuous. Let

$$C^1(I)$$

denote the set of real-valued continuously differentiable functions defined on I. Usually $I = [a, b]$ for some a and b, and in this case the derivatives at a and b are one-sided.

More generally, let $\mathbf{x}(t) = (x^1(t), \dots, x^n(t))$ be a *vector-valued* function with values in $\mathbb{R}^n$.[1] Then we say $\mathbf{x}$ is *continuously differentiable* (or C^1) if each $x^i(t)$ is C^1. Let

$$C^1(I; \mathbb{R}^n)$$

denote the set of all such continuously differentiable functions.

In an analogous manner, if a function is *continuous*, we sometimes say it is C^0.

We will consider the general first order system of differential equations in the form

$$\begin{aligned} \frac{dx^1}{dt} &= f^1(t, x^1, x^2, \dots, x^n) \\ \frac{dx^2}{dt} &= f^2(t, x^1, x^2, \dots, x^n) \\ &\vdots \\ \frac{dx^n}{dt} &= f^n(t, x^1, x^2, \dots, x^n), \end{aligned}$$

which we write for short as

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(t, \mathbf{x}).$$

Here

$$\begin{aligned} \mathbf{x} &= (x^1, \dots, x^n) \\ \frac{d\mathbf{x}}{dt} &= \left(\frac{dx^1}{dt}, \dots, \frac{dx^n}{dt}\right) \\ \mathbf{f}(t, \mathbf{x}) &= \mathbf{f}(t, x^1, \dots, x^n) \\ &= \Big(f^1(t, x^1, \dots, x^n), \dots, f^n(t, x^1, \dots, x^n)\Big). \end{aligned}$$

It is usually convenient to think of t as representing *time*, but this is not necessary.

We will always assume $\mathbf{f}$ is *continuous* for all $(t, \mathbf{x}) \in U$, where $U \subset \mathbb{R} \times \mathbb{R}^n = \mathbb{R}^{n+1}$.

By an *initial condition* is meant a condition of the form

$$x^1(t_0) = x_0^1,\ x^2(t_0) = x_0^2, \dots,\ x^n(t_0) = x_0^n$$

for some given t_0 and some given $\mathbf{x}_0 = (x_0^1, \dots, x_0^n)$. That is,

$$\mathbf{x}(t_0) = \mathbf{x}_0.$$

Here, $(t_0, \mathbf{x}_0) \in U$.

The following diagram sketches the situation (schematically in the case $n > 1$).

In case $n = 2$, we have the following diagram.

[1] You can think of $\mathbf{x}(t)$ as tracing out a curve in $\mathbb{R}^n$.

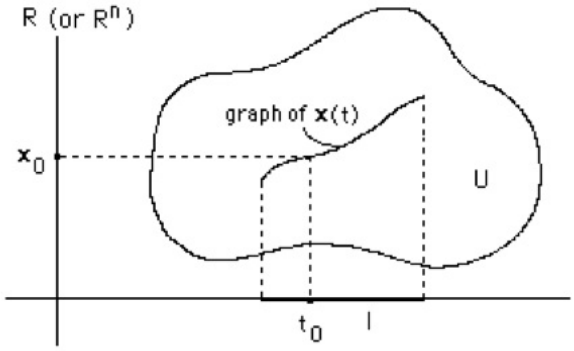

FIGURE 2. Notation and partial graph of solution for an ODE as discussed in Section 13.3.

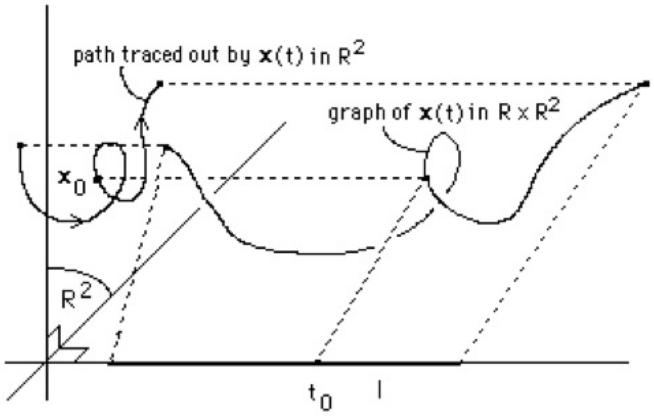

FIGURE 3. Notation and graph of solution for an ODE in case $n = 2$ as discussed in Section 13.3.

As an example, in the case of the *predator-prey problem* (124), it is reasonable to restrict (x, y) to $U = \{(x, y) : x > 0,\ y > 0\}$[2]. We might think of restricting t to $t \geq 0$, but since the right side of (124) is independent of t, and since in any case the choice of what instant in time should correspond to $t = 0$ is arbitrary, it is more reasonable not to make any restrictions on t. Thus we might take $U = \mathbb{R} \times \{(x, y) : x > 0,\ y > 0\}$ in this example. We also usually assume for this problem that we know the values of x and y at some "initial" time t_0.

DEFINITION 13.3.1. [**Initial Value Problem**] Assume $U \subset \mathbb{R} \times \mathbb{R}^n = \mathbb{R}^{n+1}$, U is open[3] and $(t_0, \mathbf{x}_0) \in U$. Assume $\mathbf{f}\,(= \mathbf{f}(t, \mathbf{x})) : U \to \mathbb{R}$ is *continuous*. Then the following is called an *initial value problem*, with *initial condition* (130):

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(t, \mathbf{x}), \tag{129}$$

$$\mathbf{x}(t_0) = \mathbf{x}_0. \tag{130}$$

We say $\mathbf{x}(t) = (x^1(t), \ldots, x^n(t))$ is a *solution* of this initial value problem for t in the interval I if:

[2]What happens if one of x or y vanishes at some point in time ?

[3]Sometimes it is convenient to allow U to be the closure of an open set.

(1) $t_0 \in I$,
(2) $\mathbf{x}(t_0) = \mathbf{x}_0$,
(3) $(t, \mathbf{x}(t)) \in U$ and $\mathbf{x}(t)$ is C^1 for $t \in I$,
(4) the system of equations (129) is satisfied by $\mathbf{x}(t)$ for all $t \in I$.

13.4. Heuristic Justification for the Existence of Solutions

To simplify notation, we consider the case $n = 1$ in this section. Thus we consider a single differential equation and an initial value problem of the form

$$x'(t) = f(t, x(t)), \tag{131}$$

$$x(t_0) = x_0. \tag{132}$$

As usual, assume f is *continuous* on U, where U is an open set containing (t_0, x_0).

It is reasonable to expect that there should exist a (unique) solution $x = x(t)$ to (131) satisfying the initial condition (132) and defined for all t in some time interval I containing t_0. We make this plausible as follows (See Figure 4.)

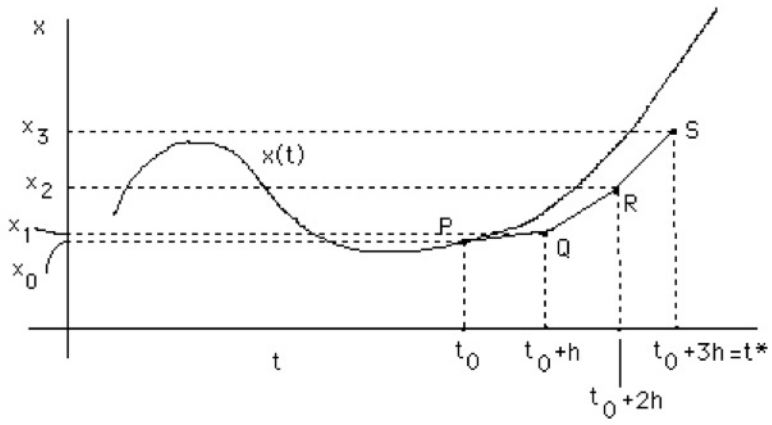

FIGURE 4. Diagram for the discussion in Section 13.4.

From (131) and (132) we know $x'(t_0) = f(t_0, x_0)$. It follows that for small $h > 0$

$$x(t_0 + h) \approx x_0 + hf(t_0, x_0) =: x_1 \text{[4]}$$

Similarly

$$\begin{aligned} x(t_0 + 2h) &\approx x_1 + hf(t_0 + h, x_1) =: x_2 \\ x(t_0 + 3h) &\approx x_2 + hf(t_0 + 2h, x_2) =: x_3 \\ &\vdots \end{aligned}$$

Suppose $t^* > t_0$. By taking sufficiently many steps, we thus obtain an approximation to $x(t^*)$ (in the diagram we have shown the case where h is such that $t^* = t_0 + 3h$). By taking $h < 0$ we can also find an approximation to $x(t^*)$ if $t^* < t_0$. By taking h

[4] $a := b$ means that a, by definition, is equal to b. And $a =: b$ means that b, by definition, is equal to a.

very small we expect to find an approximation to $x(t^*)$ to any desired degree of accuracy.

In the previous diagram

$$\begin{aligned} P &= (t_0, x_0) \\ Q &= (t_0 + h, x_1) \\ R &= (t_0 + 2h, x_2) \\ S &= (t_0 + 3h, x_3) \end{aligned}$$

The slope of PQ is $f(t_0, x_0) = f(P)$, of QR is $f(t_0 + h, x_1) = f(Q)$, and of RS is $f(t_0 + 2h, x_2) = f(R)$.

The method outlined is called the method of *Euler polygons*. It can be used to solve differential equations numerically, but there are refinements of the method which are much more accurate. Euler's method can also be made the basis of a rigorous proof of the existence of a solution to the initial value problem (131), (132). We will take a different approach, however, and use the Contraction Mapping Theorem.

Direction field

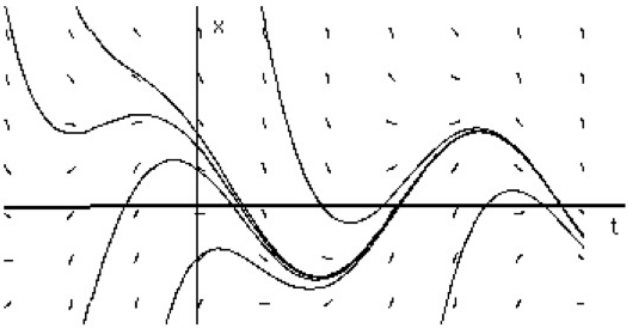

FIGURE 5. Direction field and solutions of $x'(t) = -x - \sin t$.

Consider again the differential equation (131). At each point in the (t, x) plane, one can draw a line segment with slope $f(t, x)$. The set of all line segments constructed in this way is called the *direction field* for the differential equation. The graph of any solution to (131) must have slope given by the line segment at each point through which it passes. The direction field thus gives a good idea of the behaviour of the set of solutions to the differential equation.

13.5. Phase Space Diagrams

A useful way of visualising the behaviour of solutions to a system of differential equations (129) is by means of a *phase space diagram*. This is nothing more than a set of paths (solution curves) in $\mathbb{R}^n$ (here called *phase space*) traced out by various solutions to the system. It is particularly usefu in the case $n = 2$ (i.e. two *unknowns*) and in case the system is *autonomous* (i.e. the right side of (129) is independent of time).

Note carefully the difference between the graph of a solution, and the path traced out by a solution in phase space. In particular, see the second diagram in Section 13.3, where $\mathbb{R}^2$ is phase space.

We now discuss some general considerations in the context of the following example.

Competing Species

Consider the case of two species whose populations at time t are $x(t)$ and $y(t)$. Suppose they have a good food supply but fight each other whenever they come into contact. By a discussion similar to that in Section 13.1.1, their populations may be modelled by the equations

$$\begin{aligned} \frac{dx}{dt} &= ax - bxy \ \Big(= f^1(x,y)\Big), \\ \frac{dy}{dt} &= cy - dxy \ \Big(= f^2(x,y)\Big), \end{aligned} \tag{133}$$

for suitable $a, b, c, d > 0$. Consider as an example the case $a = 1000$, $b = 1$, $c = 2000$ and $d = 1$.

If a solution $x(t), y(t)$ passes through a point (x, y) in phase space at some time t, then the "velocity" of the path at this point is $(f^1(x,y), f^2(x,y)) = (x(1000 - y), y(2000-x))$. In particular, the path is tangent to the vector $(x(1000-y), y(2000-x))$ at the point (x, y). The set of all such velocity vectors $(f^1(x,y), f^2(x,y))$ at the points $(x, y) \in \mathbb{R}^2$ is called the *velocity field* associated to the system of differential equations. Notice that as the example we are discussing is autonomous, the velocity field is *independent of time*.

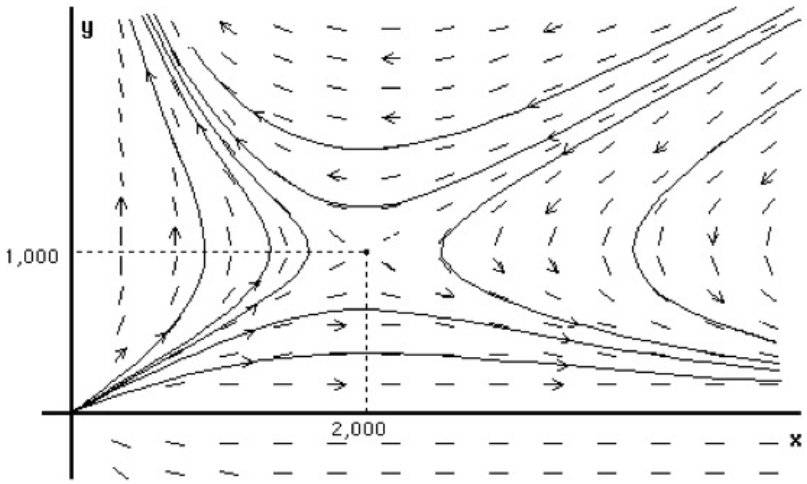

FIGURE 6. Direction field (only some arrows shown) and some solutions for the system (133).

In the previous diagram we have shown some vectors from the velocity field for the present system of equations. For simplicity, we have only shown their directions in a few cases, and we have *normalised* each vector to have the same length; we sometimes call the resulting vector field a *direction field*[5].

Once we have drawn the velocity field (or direction field), we have a good idea of the structure of the set of solutions, since each solution curve must be tangent to the velocity field at each point through which it passes.

[5] *Note* the distinction between the direction field in phase space and the direction field for the graphs of solutions as discussed in the last section.

Next note that $(f^1(x,y), f^2(x,y)) = (0,0)$ if $(x,y) = (0,0)$ or $(2000, 1000)$. Thus the "velocity" (or rate of change) of a solution passing through either of these pairs of points is zero. The pair of constant functions given by $x(t) = 2000$ and $y(t) = 1000$ for all t is a solution of the system, and from Theorem 13.10.1 is the only solution passing through $(2000, 1000)$. Such a constant solution is called a *stationary solution* or *stationary point*. In this example the other stationary point is $(0,0)$ (this is not surprising!).

The stationary point (2000, 1000) is *unstable* in the sense that if we change either population by a small amount away from these values, then the populations do not converge back to these values. In this example, one population will always die out. This is all clear from the diagram.

13.6. Examples of Non-Uniqueness and Non-Existence

Example 1 (Non-Uniqueness) Consider the initial value problem

$$\frac{dx}{dt} = \sqrt{|x|}, \tag{134}$$

$$x(0) = 0. \tag{135}$$

We use the method of separation of variables and formally compute from (134) that

$$\frac{dx}{\sqrt{|x|}} = dt.$$

If $x > 0$, integration gives

$$\frac{x^{1/2}}{1/2} = t - a,$$

for some a. That is, for $x > 0$,

$$x(t) = (t-a)^2/4. \tag{136}$$

We need to justify these formal computations. By differentiating, we check that (136) is indeed a solution of (134) provided $t \geq a$.

Note also that $x(t) = 0$ is a solution of (134) for all t.

Moreover, we can check that for each $a \geq 0$ there is a solution of (134) and (135) given by

$$x(t) = \begin{cases} 0 & t \leq a \\ (t-a)^2/4 & t > a. \end{cases}$$

See the following diagram.

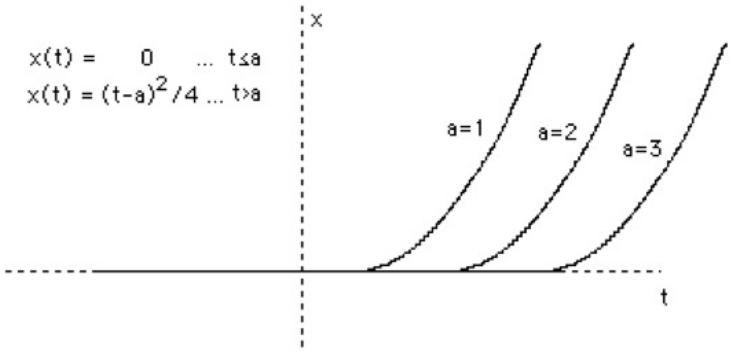

FIGURE 7. Different solutions of (134) and (135), one for each real a.

Thus we do *not* have uniqueness for solutions of (134), (135). There are even more solutions to (134), (135), what are they? (*exercise*).

We will later prove uniqueness of solutions of the initial value problem (131), (132) provided the function $f(t,x)$ is *locally Lipschitz with respect to* x, as defined in the next section.

Example 2 (Non-Existence) Let $f(t,x) = 1$ if $x \leq 1$, and $f(t,x) = 2$ if $x > 1$. Notice that f is *not* continuous. Consider the initial value problem

$$x'(t) = f(t, x(t)), \tag{137}$$

$$x(0) = 0. \tag{138}$$

Then it is natural to take the solution to be

$$x(t) = \begin{cases} t & t \leq 1 \\ 2t-1 & t > 1 \end{cases}$$

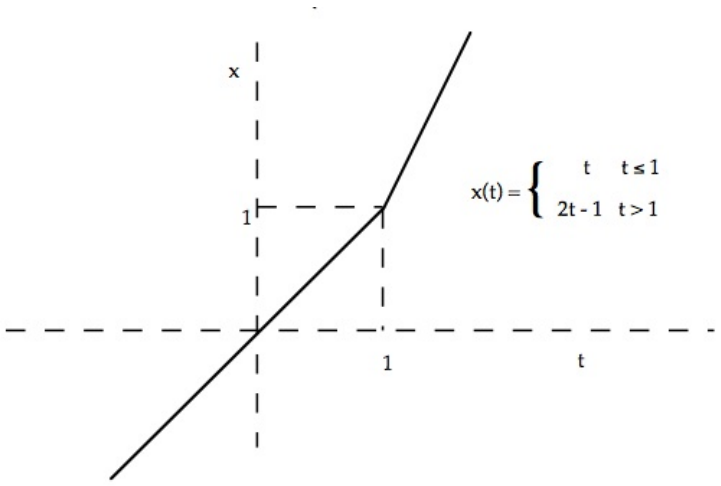

FIGURE 8. "Solution" of (137) and (138).

Notice that $x(t)$ satisfies the initial condition and also satisfies the differential equation provided $t \neq 1$. But $x(t)$ is not differentiable at $t = 1$. There is *no* solution of this initial value problem, in the usual sense of a *solution*. It is possible to generalise the notion of a solution, and in this case the "solution" given is the correct one.

13.7. A Lipschitz Condition

As we saw in Example 1 of Section 13.6, we need to impose a further condition on f, apart from continuity, if we are to have a unique solution to the Initial Value Problem (131), (132). We do this by generalising slightly the notion of a Lipschitz function as defined in Section 11.3.

DEFINITION 13.7.1. The function $\mathbf{f} = \mathbf{f}(t, \mathbf{x}) : A\,(\subset \mathbb{R} \times \mathbb{R}^n) \to \mathbb{R}$ is *Lipschitz with respect to* $\mathbf{x}$ (in A) if there exists a constant K such that

$$(t, \mathbf{x}_1), (t, \mathbf{x}_2) \in A \Rightarrow |\mathbf{f}(t, \mathbf{x}_1) - \mathbf{f}(t, \mathbf{x}_2)| \leq K|\mathbf{x}_1 - \mathbf{x}_2|.$$

If $\mathbf{f}$ is Lipschitz with respect to $\mathbf{x}$ in $A_{h,k}(t_0, \mathbf{x}_0)$, for every set $A_{h,k}(t_0, \mathbf{x}_0) \subset A$ of the form

$$A_{h,k}(t_0, \mathbf{x}_0) := \{(t, \mathbf{x}) : |t - t_0| \leq h,\ |\mathbf{x} - \mathbf{x}_0| \leq k\}, \tag{139}$$

then we say f is *locally Lipschitz with respect to* $\mathbf{x}$. (See Figure 9.)

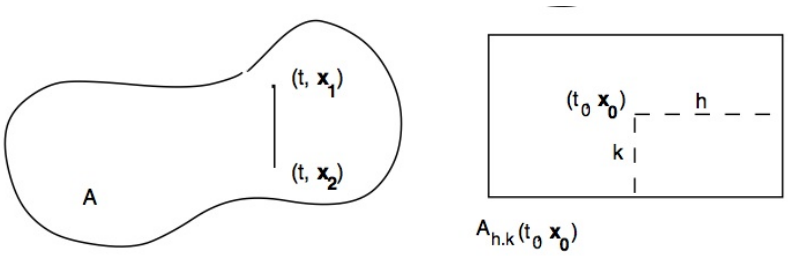

FIGURE 9. Diagram for Definition 13.7.1.

We could have replaced the sets $A_{h,k}(t_0, \mathbf{x}_0)$ by closed balls centred at $(t, \mathbf{x}_0)$ without affecting the definition, since each such ball contains a set $A_{h,k}(t_0, \mathbf{x}_0)$ for some $h, k > 0$, and conversely. We choose sets of the form $A_{h,k}(t_0, \mathbf{x}_0)$ for later convenience.

The difference between being *Lipschitz* with respect to $\mathbf{x}$ and being *locally Lipschitz* with respect to $\mathbf{x}$ is clear from the following Examples.

Example 1 Let $n = 1$ and $A = \mathbb{R} \times \mathbb{R}$. Let $f(t, x) = t^2 + 2\sin x$. Then

$$\begin{aligned} |f(t, x_1) - f(t, x_2)| &= |2\sin x_1 - 2\sin x_2| \\ &= |2\cos\xi|\,|x_1 - x_2| \\ &\leq 2|x_1 - x_2|, \end{aligned}$$

for some ξ between x_1 and x_2, using the Mean Value Theorem.

Thus f is Lipschitz with respect to x (it is also Lipschitz in the usual sense).

Example 2 Let $n = 1$ and $A = \mathbb{R} \times \mathbb{R}$. Let $f(t, x) = t^2 + x^2$. Then

$$\begin{aligned} |f(t, x_1) - f(t, x_2)| &= |x_1^2 - x_2^2| \\ &= |2\xi|\,|x_1 - x_2|, \end{aligned}$$

for some ξ between x_1 and x_2, again using the Mean Value Theorem. If $x_1, x_2 \in B$ for some bounded set B, in particular if B is of the form $\{(t, x) : |t - t_0| \leq h,\ |x - x_0| \leq k\}$, then ξ is also bounded, and so f is *locally* Lipschitz in A. But f is *not* Lipschitz in A.

We now give an analogue of the result from Example 1 in Section 11.3.

THEOREM 13.7.2. *Let $U \subset \mathbb{R} \times \mathbb{R}^n$ be open and let $\mathbf{f} = \mathbf{f}(t, \mathbf{x}) : U \to \mathbb{R}$. If the partial derivatives $\dfrac{\partial \mathbf{f}}{\partial x_i}(t, \mathbf{x})$ all exist and are continuous in U, then $\mathbf{f}$ is locally Lipschitz in U with respect to $\mathbf{x}$.*

PROOF. Let $(t_0, \mathbf{x}_0) \in U$. Since U is open, there exist $h, k > 0$ such that

$$A_{h,k}(t_0, \mathbf{x}_0) := \{(t, \mathbf{x}) : |t - t_0| \leq h,\ |\mathbf{x} - \mathbf{x}_0| \leq k\} \subset U.$$

Since the partial derivatives $\dfrac{\partial \mathbf{f}}{\partial x_i}(t, \mathbf{x})$ are continuous on the compact set $A_{h,k}$, they are also bounded on $A_{h,k}$ from Theorem 11.5.2. Suppose

$$\left| \frac{\partial \mathbf{f}}{\partial x_i}(t, \mathbf{x}) \right| \leq K, \tag{140}$$

for $i = 1, \ldots, n$ and $(t, \mathbf{x}) \in A_{h,k}$.

Let $(t, \mathbf{x}_1), (t, \mathbf{x}_2) \in A_{h,k}$. To simplify notation, let $n = 2$ and let $\mathbf{x}_1 = (x_1^1, x_1^2)$, $\mathbf{x}_2 = (x_2^1, x_2^2)$. Then
(see Figure 10–note that it is in $\mathbb{R}^n$, *not* in $\mathbb{R} \times \mathbb{R}^n$; here $n = 2$.)

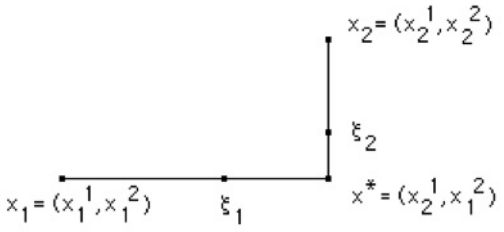

FIGURE 10. Diagram for the proof of Thorem 13.7.2.

$$\begin{aligned}
|\mathbf{f}(t,\mathbf{x}_1) - \mathbf{f}(t,\mathbf{x}_2)| &= |\mathbf{f}(t,x_1^1,x_1^2) - \mathbf{f}(t,x_2^1,x_2^2)| \\
&\le |\mathbf{f}(t,x_1^1,x_1^2) - \mathbf{f}(t,x_2^1,x_1^2)| + |\mathbf{f}(t,x_2^1,x_1^2) - \mathbf{f}(t,x_2^1,x_2^2)| \\
&= |\frac{\partial \mathbf{f}}{\partial x_1}(\xi_1)|\,|x_2^1 - x_1^1| + |\frac{\partial \mathbf{f}}{\partial x_2}(\xi_2)|\,|x_2^2 - x_1^2| \\
&\le K|x_2^1 - x_1^1| + K|x_2^2 - x_1^2| \quad \text{from (140)} \\
&\le 2K|\mathbf{x}_1 - \mathbf{x}_2|,
\end{aligned}$$

In the third line, $\xi_\mathbf{1}$ is between $\mathbf{x}_1$ and $\mathbf{x}^* = (x_2^1, x_1^2)$, and ξ_2 is between $\mathbf{x}^* = (x_2^1, x_1^2)$ and $\mathbf{x}_2$. This uses the usual Mean Value Theorem for a function of *one* variable, applied on the interval $[x_1^1, x_2^1]$, and on the interval $[x_1^2, x_2^2]$.

This completes the proof if $n = 2$. For $n > 2$ the proof is similar. □

13.8. Reduction to an Integral Equation

We again consider the case $n = 1$ in this section.

Thus we again consider the Initial Value Problem

$$x'(t) = f(t, x(t)), \tag{141}$$

$$x(t_0) = x_0. \tag{142}$$

As usual, assume f is *continuous* in U, where U is an open set containing (t_0, x_0).

The first step in proving the existence of a solution to (141), (142) is to show that the problem is equivalent to solving a certain *integral equation*. This follows easily by integrating both sides of (141) from t_0 to t. More precisely:

THEOREM 13.8.1. *Assume the function x satifies $(t, x(t)) \in U$ for all $t \in I$, where I is some closed bounded interval. Assume $t_0 \in I$. Then x is a C^1 solution to (141), (142) in I iff x is a C^0 solution to the integral equation*

$$x(t) = x_0 + \int_{t_0}^{t} f\Big(s, x(s)\Big)\, ds \tag{143}$$

in I.

PROOF. First let x be a C^1 solution to (141), (142) in I. Then the left side, and hence both sides, of (141) are continuous and in particular integrable. Hence

for any $t \in I$ we have by integrating (141) from t_0 to t that

$$x(t) - x(t_0) = \int_{t_0}^{t} f\Big(s, x(s)\Big)\, ds.$$

Since $x(t_0) = x_0$, this shows x is a C^1 (and in particular a C^0) solution to (143) for $t \in I$.

Conversely, assume x is a C^0 solution to (143) for $t \in I$. Since the functions $t \mapsto x(t)$ and $t \mapsto t$ are continuous, it follows that the function $s \mapsto (s, x(s))$ is continuous from Theorem 11.2.1. Hence $s \mapsto f(s, x(s))$ is continuous from Theorem 11.2.3. It follows from (143), using properties of indefinite integrals of continuous functions[6], that $x'(t)$ exists and

$$x'(t) = f(t, x(t))$$

for all $t \in I$. In particular, x is C^1 on I. Finally, it follows immediately from 143 that $x(t_0) = x_0$. Thus x is a C^1 solution to (141), (142) in I. □

Remark A bootstrap argument shows that the solution x is in fact C^∞ provided f is C^∞.

13.9. Local Existence

We again consider the case $n = 1$ in this section.

We first use the Contraction Mapping Theorem to show that the integral equation (143) has a solution on some interval containing t_0.

THEOREM 13.9.1. *Assume f is continuous, and locally Lipschitz with respect to the second variable, on the open set $U \subset \mathbb{R} \times \mathbb{R}$. Let $(t_0, x_0) \in U$. Then there exists $h > 0$ such that the integral equation*

$$x(t) = x_0 + \int_{t_0}^{t} f\Big(t, x(t)\Big)\, dt \tag{144}$$

has a unique C^0 solution for $t \in [t_0 - h, t_0 + h]$.

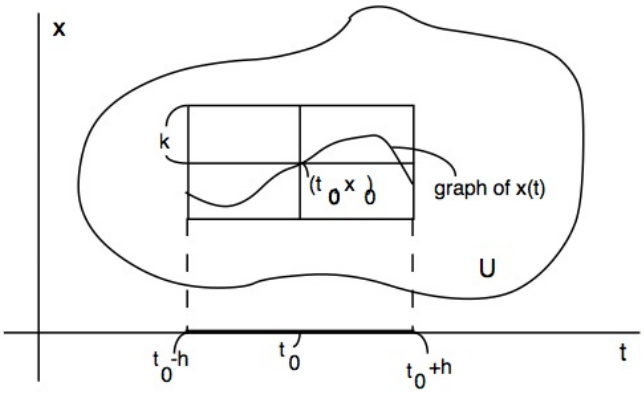

FIGURE 11. Diagram for the proof of Theorem 13.9.1.

[6] If h exists and is continuous on I, $t_0 \in I$ and $g(t) = \int_{t_0}^{t} h(s)\, ds$ for all $t \in I$, then g' exists and $g' = h$ on I. In particular, g is C^1.

PROOF. Choose $h, k > 0$ so that

$$A_{h,k}(t_0, \mathbf{x}_0) := \{(t, \mathbf{x}) : |t - t_0| \leq h,\ |\mathbf{x} - \mathbf{x}_0| \leq k\} \subset U.$$

Since f is continuous, it is bounded on the compact set $A_{h,k}(t_0, \mathbf{x}_0)$ by Theorem 11.5.2. Choose M such that

$$|f(t, x)| \leq M \quad \text{if } (t, x) \in A_{h,k}(t_0, \mathbf{x}_0). \tag{145}$$

Since f is locally Lipschitz with respect to x, there exists K such that

$$|f(t, x_1) - f(t, x_2)| \leq K|x_1 - x_2| \quad \text{if } (t, x_1), (t, x_2) \in A_{h,k}(t_0, \mathbf{x}_0). \tag{146}$$

By decreasing h if necessary, we will require

$$h \leq \min\left\{\frac{k}{M}, \frac{1}{2K}\right\}. \tag{147}$$

Let $\mathcal{C}^*[t_0 - h, t_0 + h]$ be the set of continuous functions defined on $[t_0 - h, t_0 + h]$ whose graphs lie in $A_{h,k}(t_0, \mathbf{x}_0)$. That is,

$$\begin{aligned}\mathcal{C}^*[t_0 - h, t_0 + h] &= \mathcal{C}[t_0 - h, t_0 + h] \bigcap \\ &\quad \Big\{x(t) : |x(t) - x_0| \leq k \text{ for all } t \in [t_0 - h, t_0 + h]\Big\}.\end{aligned}$$

Now $\mathcal{C}[t_0 - h, t_0 + h]$ is a complete metric space with the uniform metric, as noted in Example 1 of Section 12.3. Since $\mathcal{C}^*[t_0 - h, t_0 + h]$ is a closed subset[7], it follows from the "generalisation" following Theorem 8.2.2 that $\mathcal{C}^*[t_0 - h, t_0 + h]$ is also a complete metric space with the uniform metric.

We want to solve the integral equation (144).

To do this consider the map

$$T : \mathcal{C}^*[t_0 - h, t_0 + h] \to \mathcal{C}^*[t_0 - h, t_0 + h]$$

defined by

$$(Tx)(t) = x_0 + \int_{t_0}^{t} f\Big(t, x(t)\Big)\, dt \quad \text{for } t \in [t_0 - h, t_0 + h]. \tag{148}$$

Notice that the fixed points of T are precisely the solutions of (144).

We check that T is indeed a map into $\mathcal{C}^*[t_0 - h, t_0 + h]$ as follows:

(i): Since in (148) we are taking the definite integral of a continuous function, Corollary 11.6.4 shows that Tx is a continuous function.

(ii): Using (145) and (147) we have

$$\begin{aligned}|(Tx)(t) - x_0| &= \left|\int_{t_0}^{t} f\Big(t, x(t)\Big)\, dt\right| \\ &\leq \int_{t_0}^{t} \left|f\Big(t, x(t)\Big)\right| dt \\ &\leq hM \\ &\leq k.\end{aligned}$$

It follows from the definition of $\mathcal{C}^*[t_0 - h, t_0 + h]$ that $Tx \in \mathcal{C}^*[t_0 - h, t_0 + h]$.

[7] If $x_n \to x$ uniformly and $|x_n(t)| \leq k$ for all t, then $|x(t)| \leq k$ for all t.

We next check that T is a contraction map. To do this we compute for $x_1, x_2 \in \mathcal{C}^*[t_0 - h, t_0 + h]$, using (146) and (147), that

$$\begin{aligned} |(Tx_1)(t) - (Tx_2)(t)| &= \left| \int_{t_0}^{t} \Big(f\Big(t, x_1(t)\Big) - f\Big(t, x_2(t)\Big) \Big) dt \right| \\ &\leq \int_{t_0}^{t} \left| f\Big(t, x_1(t)\Big) - f\Big(t, x_2(t)\Big) \right| dt \\ &\leq \int_{t_0}^{t} K|x_1(t) - x_2(t)|\, dt \\ &\leq Kh \sup_{t \in [t_0 - h, t_0 + h]} |x_1(t) - x_2(t)| \\ &\leq \frac{1}{2} d_u(x_1, x_2). \end{aligned}$$

Hence

$$d_u(Tx_1, Tx_2) \leq \frac{1}{2} d_u(x_1, x_2).$$

Thus we have shown that T is a contraction map on the complete metric space $\mathcal{C}^*[t_0 - h, t_0 + h]$, and so has a unique fixed point. This completes the proof of the theorem, since as noted before the fixed points of T are precisely the solutions of (144). □

Since the contraction mapping theorem gives an algorithm for finding the fixed point, this can be used to obtain approximates to the solution of the differential equation. In fact the argument can be sharpened considerably. At the step (149)

$$\begin{aligned} |(Tx_1)(t) - (Tx_2)(t)| &\leq \int_{t_0}^{t} K|x_1(t) - x_2(t)|\, dt \\ &\leq K|t - t_0| d_u(x_1, x_2). \end{aligned}$$

Thus applying the next step of the iteration,

$$\begin{aligned} |(T^2x_1)(t) - (T^2x_2)(t)| &\leq \int_{t_0}^{t} K|(Tx_1)(t) - (Tx_2)(t)|\, dt \\ &\leq K^2 \int_{t_0}^{t} |t - t_0|\, dt\, d_u(x_1, x_2) \\ &\leq K^2 \frac{|t - t_0|^2}{2} d_u(x_1, x_2). \end{aligned}$$

Induction gives

$$|(T^r x_1)(t) - (T^r x_2)(t) \leq K^r \frac{|t - t_0|^r}{r!} d_u(x_1, x_2).$$

Without using the fact that $Kh < 1/2$, it follows that some power of T is a contraction. Thus by one of the problems T itself has a unique fixed point. This observation generally facilitates the obtaining of a larger domain for the solution.

Example The simple equation $x'(t) = x(t)$, $x(0) = 1$ is well known to have solution the exponential function. Applying the above algorithm, with $x_1(t) = 1$

we would have

$$\begin{aligned}(Tx_1)(t) &= 1+\int_{t_0}^{t} f\Big(t,x_1(t)\Big)\,dt = 1+t,\\ (T^2x_1)(t) &= 1+\int_{t_0}^{t}(1+t)dt = 1+t+\frac{t^2}{2}\\ &\vdots\\ (T^kx_1)(t) &= \sum_{i=0}^{k}\frac{t^i}{i!},\end{aligned}$$

giving the exponential series, which in fact converges to the solution uniformly on any bounded interval.

THEOREM 13.9.2 (Local Existence and Uniqueness). *Assume that $f(t,x)$ is continuous, and locally Lipschitz with respect to x, in the open set $U \subset \mathbb{R}\times\mathbb{R}$. Let $(t_0,x_0)\in U$. Then there exists $h>0$ such that the initial value problem*

$$\begin{aligned}x'(t) &= f(t,x(t)),\\ x(t_0) &= x_0,\end{aligned}$$

has a unique C^1 solution for $t\in[t_0-h,t_0+h]$.

PROOF. By Theorem 13.8.1, x is a C^1 solution to this initial value problem iff it is a C^0 solution to the integral equation (144). But the integral equation has a unique solution in some $[t_0-h,t_0+h]$, by the previous theorem. □

13.10. Global Existence

Theorem 13.9.2 shows the existence of a (unique) solution to the initial value problem in some (possibly small) time interval containing t_0. Even if $U=\mathbb{R}\times\mathbb{R}$ it is *not* necessarily true that a solution exists for all t.

Example 1 Consider the initial value problem

$$\begin{aligned}x'(t) &= x^2,\\ x(0) &= a,\end{aligned}$$

where for this discussion we take $a\geq 0$.

If $a=0$, this has the solution

$$x(t)=0,\quad (\text{all } t).$$

If $a>0$ we use separation of variables to show that the solution is

$$x(t)=\frac{1}{a^{-1}-t},\quad (t<a^{-1}).$$

It follows from the Existence and Uniqueness Theorem that for each a this gives the *only* solution.

Notice that if $a>0$, then the solution $x(t)\to\infty$ as $t\to a^{-1}$ from the left, and $x(t)$ is undefined for $t=a^{-1}$. Of course this x also satisfies $x'(t)=x$ for $t>a^{-1}$, as do all the functions

$$x_b(t)=\frac{1}{b^{-1}-t},\quad (0<b<a).$$

Thus the presciption of $x(0)=a$ gives zero information about the solution for $t>a^{-1}$.

The following diagram shows the solution for various a.

The following theorem more completely analyses the situation.

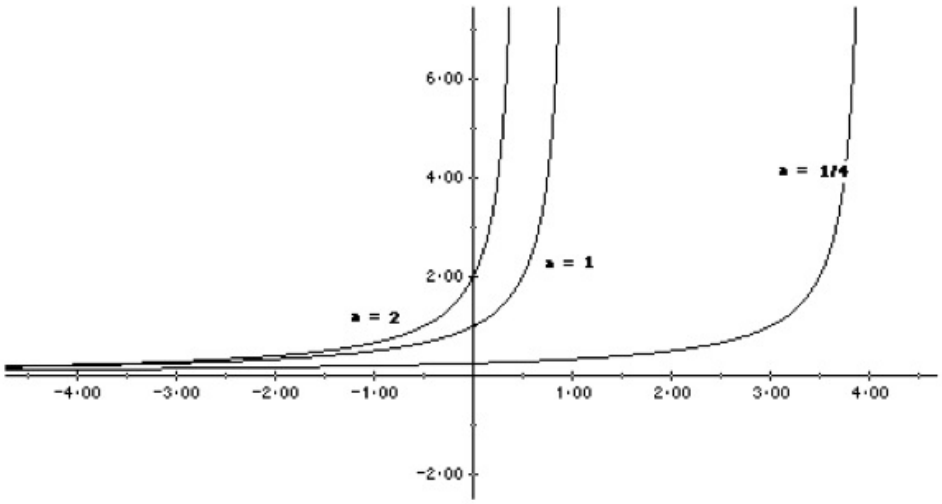

FIGURE 12. Solutions to the initial value problem in Example 13.10 for different a.

THEOREM 13.10.1 (Global Existence and Uniqueness). *There is a unique solution x to the initial value problem (141), (142) and*

(1) *either the solution exists for all $t \geq t_0$,*
(2) *or the solution exists for all $t_0 \leq t < T$, for some (finite) $T > t_0$; in which case for any closed bounded subset $A \subset U$ we have $(t, x(t)) \notin A$ for all $t < T$ sufficiently close to T.*

A similar result applies to $t \leq t_0$.

Remark* The second alternative in the Theorem just says that the graph of the solution eventually leaves any closed bounded $A \subset U$. We can think of it as saying that the graph of the solution either *escapes to infinity* or approaches the boundary of U as $t \to T$.

PROOF. * (*Outline*) Let T be the supremum of all t^* such that a solution exists for $t \in [t_0, t^*]$. If $T = \infty$, then we are done.

If T is finite, let $A \subset U$ where A is compact. If $(t, \mathbf{x}(t))$ does not eventually leave A, then there exists a sequence $t_n \to T$ such that $(t_n, \mathbf{x}(t_n)) \in A$. From the definition of compactness, a subsequence of $(t_n, \mathbf{x}(t_n))$ must have a limit in A. Let $(t_{n_i}, \mathbf{x}(t_{n_i})) \to (T, \overline{\mathbf{x}}) \in A$ (note that $t_{n_i} \to T$ since $t_n \to T$). In particular, $\mathbf{x}(t_{n_i}) \to \overline{\mathbf{x}}$.

The proof of the Existence and Uniqueness Theorem shows that a solution beginning at $(T, \overline{\mathbf{x}})$ exists for some time $h > 0$, and *moreover*, that for $t' = T - h/4$, say, the solution beginning at $(t', \mathbf{x}(t'))$ exists for time $h/2$. But this then extends the original solution past time T, contradicting the definition of T.

Hence $(t, \mathbf{x}(t))$ does eventually leave A. □

13.11. Extension of Results to Systems

The discussion, proofs and results in Sections 13.4, 13.8, 13.9 and 13.10 generalise to systems, with essentially only notational changes, as we now sketch.

Thus we consider the following initial value problem for systems:

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= \mathbf{f}(t,\mathbf{x}), \\ \mathbf{x}(t_0) &= \mathbf{x}_0. \end{aligned}$$

This is equivalent to the integral equation

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}\Big(s,\mathbf{x}(s)\Big)\,ds.$$

The integral of the vector function on the right side is defined componentwise in the natural way, i.e.

$$\int_{t_0}^{t} \mathbf{f}\Big(s,\mathbf{x}(s)\Big)\,ds := \left(\int_{t_0}^{t} f^1\Big(s,\mathbf{x}(s)\Big)\,ds, \ldots, \int_{t_0}^{t} f^2\Big(s,\mathbf{x}(s)\Big)\,ds\right).$$

The proof of equivalence is essentially the proof in Section 13.8 for the single equation case, applied to each component separately.

Solutions of the integral equation are precisely the fixed points of the operator T, where

$$(T\mathbf{x})(t) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}\Big(s,\mathbf{x}(s)\Big)\,ds \quad t \in [t_0-h, t_0+h].$$

As is the proof of Theorem 13.9.1, T is a contraction map on

$$\begin{aligned} \mathcal{C}^*([t_0-h,t_0+h];\mathbb{R}^n) &= \mathcal{C}([t_0-h,t_0+h];\mathbb{R}^n) \bigcap \\ &\quad \Big\{\mathbf{x}(t) : |\mathbf{x}(t)-\mathbf{x}_0| \le k \text{ for all } t \in [t_0-h,t_0+h]\Big\} \end{aligned}$$

for some I and some $k > 0$, provided $\mathbf{f}$ is locally Lipschitz in $\mathbf{x}$. This is proved exactly as in Theorem 13.9.1. Thus the integral equation, and hence the initial value problem has a unique solution in some time interval containing t_0.

The analogue of Theorem 13.10.1 for global (or long-time) existence is also valid, with the same proof.

CHAPTER 14

Fractals

So, naturalists observe, a flea
Hath smaller fleas that on him prey;
And these have smaller still to bite 'em;
And so proceed ad infinitum.

Jonathan Swift *On Poetry. A Rhapsody* [1733]

Big whorls have little whorls
which feed on their velocity;
And little whorls have lesser whorls,
and so on to viscosity.

Lewis Fry Richardson

Fractals are, loosely speaking, sets which

- have a *fractional dimension*;
- have certain *self-similarity* or *scale invariance* properties.

There is also a notion of a *random* (or *probabilistic*) fractal.

Until recently, fractals were considered to be only of mathematical interest. But in the last few years they have been used to model a wide range of mathematical phenomena—coastline patterns, river tributary patterns, and other geological structures; leaf structure, error distribution in electronic transmissions, galactic clustering, etc. etc. The theory has been used to achieve a very high magnitude of data compression in the storage and generation of computer graphics.

References include the book [**Ma**], which is written in a somewhat informal style but has many

and [**Ba**] provide an accessible discussion of many aspects of fractals and are quite readable. The book [**BD**] has a number of good articles.

14.1. Examples

14.1.1. Koch Curve. A sequence of approximations $A = A^{(0)}, A^{(1)}, A^{(2)}, \ldots, A^{(n)}, \ldots$ to the *Koch Curve* (or *Snowflake Curve*) is sketched in the following diagrams.

The actual Koch curve $K \subset \mathbb{R}^2$ is the limit of these approximations in a sense which we later make precise.

Notice that

$$A^{(1)} = S_1[A] \cup S_2[A] \cup S_3[A] \cup S_4[A],$$

where each $S_i : \mathbb{R}^2 \to \mathbb{R}^2$, and S_i equals a dilation with dilation ratio $1/3$, followed by a translation and a rotation. For example, S_1 is the map given by dilating with dilation ratio $1/3$ about the fixed point P, see the diagram. S_2 is obtained by

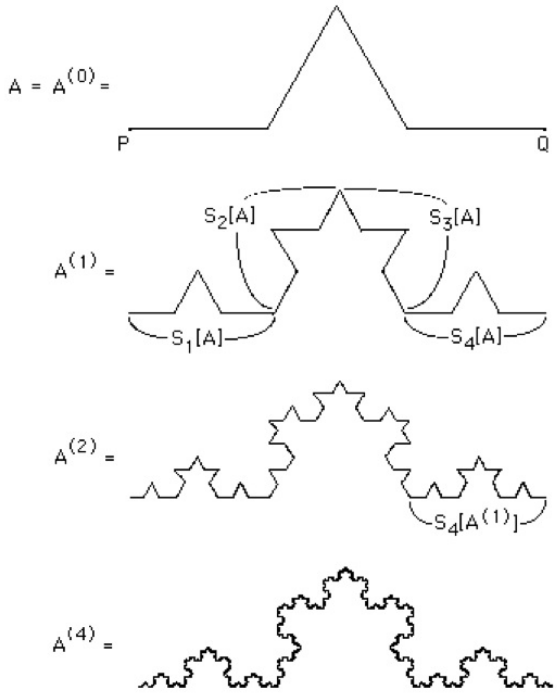

FIGURE 1. Approximations to the Koch curve.

composing this map with a suitable translation and then a rotation through 60^0 in the anti-clockwise direction. Similarly for S_3 and S_4.

Likewise,

$$A^{(2)} = S_1[A^{(1)}] \cup S_2[A^{(1)}] \cup S_3[A^{(1)}] \cup S_4[A^{(1)}].$$

In general,

$$A^{(n+1)} = S_1[A^{(n)}] \cup S_2[A^{(n)}] \cup S_3[A^{(n)}] \cup S_4[A^{(n)}].$$

Moreover, the Koch curve K itself has the property that

$$K = S_1[K] \cup S_2[K] \cup S_3[K] \cup S_4[K].$$

This is quite plausible, and will easily follow after we make precise the limiting process used to define K.

14.1.2. Cantor Set. We next sketch a sequence of approximations $A = A^{(0)}$, $A^{(1)}$, $A^{(2)}$,..., $A^{(n)}$, ... to the *Cantor Set* C.

We can think of C as obtained by first removing the *open middle third* $(1/3, 2/3)$ from $[0, 1]$; then removing the open middle third from each of the two closed intervals which remain; then removing the open middle third from each of the four closed interval which remain; etc.

More precisely, let

$$\begin{aligned} A = A^{(0)} &= [0,1] \\ A^{(1)} &= [0,1/3] \cup [2/3,1] \\ A^{(2)} &= [0,1/9] \cup [2/9,1/3] \cup [2/3,7/9] \cup [8/9,1] \\ &\vdots \end{aligned}$$

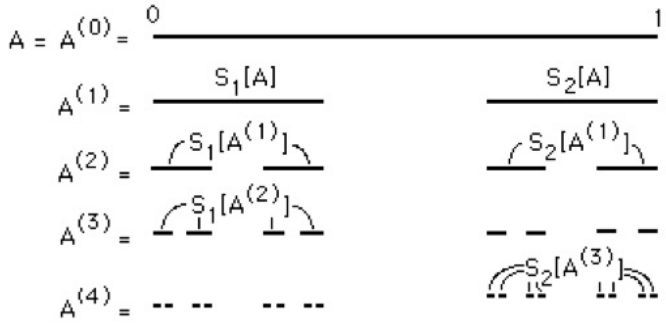

FIGURE 2. Approximations to the Cantor set.

Let $C = \bigcap_{n=0}^{\infty} A^{(n)}$. Since C is the intersection of a family of closed sets, C is closed.

Note that $A^{(n+1)} \subset A^{(n)}$ for all n and so the $A^{(n)}$ form a *decreasing* family of sets.

Consider the ternary expansion of numbers $x \in [0,1]$, i.e. write each $x \in [0,1]$ in the form

$$x = .a_1a_2 \dots a_n \dots = \frac{a_1}{3} + \frac{a_2}{3^2} + \dots + \frac{a_n}{3^n} + \dots \tag{149}$$

where $a_n = 0, 1$ or 2. Each number has either one or two such representations, and the only way x can have two representations is if

$$x = .a_1a_2 \dots a_n222 \dots = .a_1a_2 \dots a_{n-1}(a_n+1)000 \dots$$

for some $a_n = 0$ or 1. For example, $.210222 \dots = .211000 \dots$.

Note the following:

(1) $x \in A^{(n)}$ iff x has an expansion of the form (149) with each of $a_1, \dots, a_n$ taking the values 0 or 2.
(2) It follows that $x \in C$ iff x has an expansion of the form (149) with *every* a_n taking the values 0 or 2.
(3) Each endpoint of any of the 2^n intervals associated with $A^{(n)}$ has an expansion of the form (149) with each of $a_1, \dots, a_n$ taking the values 0 or 2 and the remaining a_i either *all* taking the value 0 or *all* taking the value 2.

Next let

$$S_1(x) = \frac{1}{3}x, \quad S_2(x) = 1 + \frac{1}{3}(x-1).$$

Notice that S_1 is a dilation with dilation ratio $1/3$ and fixed point 0. Similarly, S_2 is a dilation with dilation ratio $1/3$ and fixed point 1.

Then

$$A^{(n+1)} = S_1[A^{(n)}] \cup S_2[A^{(n)}].$$

Moreover,

$$C = S_1[C] \cup S_2[C].$$

14.1.3. Sierpinski Sponge. The following diagrams show two approximations to the *Sierpinski Sponge*.

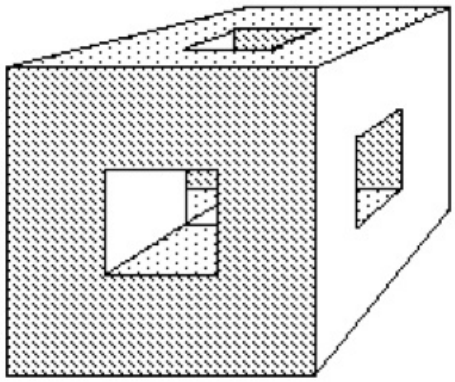

FIGURE 3. First approximation to the Sierpinski Sponge.

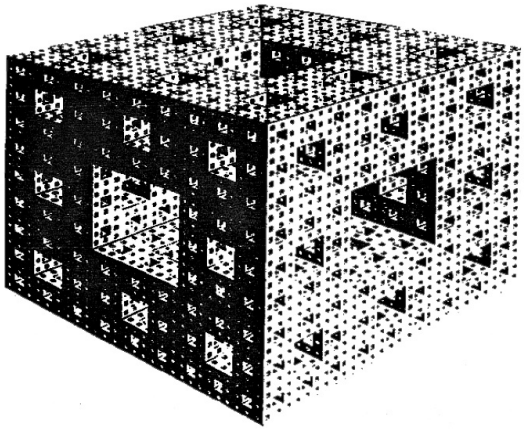

FIGURE 4. A subsequent approximation to the Sierpinski Sponge.

The Sierpinski Sponge P is obtained by first drilling out from the closed unit cube $A = A^{(0)} = [0,1] \times [0,1] \times [0,1]$, the three open, square cross-section, tubes

$$(1/3, 2/3) \times (1/3, 2/3) \times \mathbb{R},$$
$$(1/3, 2/3) \times \mathbb{R} \times (1/3, 2/3),$$
$$\mathbb{R} \times (1/3, 2/3) \times (1/3, 2/3).$$

The remaining (closed) set $A = A^{(1)}$ is the union of 20 small cubes (8 at the top, 8 at the bottom, and 4 legs).

From each of these 20 cubes, we again remove three tubes, each of cross-section equal to one-third that of the cube. The remaining (closed) set is denoted by $A = A^{(2)}$.

Repeating this process, we obtain $A = A^{(0)}, A^{(1)}, A^{(2)}, \ldots, A^{(n)}, \ldots$; a sequence of closed sets such that

$$A^{(n+1)} \subset A^{(n)},$$

for all n. We define

$$P = \bigcap_{n \geq 1} A^{(n)}.$$

Notice that P is also closed, being the intersection of closed sets.

14.2. Fractals and Similitudes

Motivated by the three previous examples we make the following:

DEFINITION 14.2.1. A *fractal*[1] in $\mathbb{R}^n$ is a compact set K such that

$$K = \bigcup_{i=1}^{N} S_i[K] \tag{150}$$

for some finite family

$$\mathcal{S} = \{S_1, \ldots, S_N\}$$

of *similitudes* $S_i : \mathbb{R}^n \to \mathbb{R}^n$.

Similitudes A *similitude* is any map $S : \mathbb{R}^n \to \mathbb{R}^n$ which is a composition of *dilations*[2], *orthonormal transformations*[3], and *translations*[4].

Note that translations and orthonormal transformations preserve distances, i.e. $|F(\mathbf{x}) - F(\mathbf{y})| = |\mathbf{x} - \mathbf{y}|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ if F is such a map. On the other hand, $|D(\mathbf{x}) - D(\mathbf{y})| = r|\mathbf{x} - \mathbf{y}|$ if D is a dilation with dilation ratio $r \geq 0$[5]. It follows that every similitude S has a well-defined dilation ratio $r \geq 0$, i.e.

$$|S(\mathbf{x}) - S(\mathbf{y})| = r|\mathbf{x} - \mathbf{y}|,$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

THEOREM 14.2.2. *Every similitude S can be expressed in the form*

$$S = D \circ T \circ O,$$

with D a dilation about $\mathbf{0}$, T a translation, and O an orthonormal transformation. In other words,

$$S(\mathbf{x}) = r(O\mathbf{x} + \mathbf{a}), \tag{151}$$

for some $r \geq 0$, some $\mathbf{a} \in \mathbb{R}^n$ and some orthonormal transformation O.

Moreover, the dilation ratio of the composition of two similitudes is the product of their dilation ratios.

PROOF. Every map of type (151) is a similitude.

On the other hand, any dilation, translation or orthonormal transformation is clearly of type (151). To show that a composition of such maps is also of type (151), it is thus sufficient to show that a composition of maps of type (151) is itself of type (151). But

$$r_1\Big(O_1\Big(r_2(O_2\mathbf{x} + \mathbf{a}_2)\Big) + \mathbf{a}_1\Big) = r_1 r_2\Big(O_1 O_2 \mathbf{x} + \Big(O_1\mathbf{a}_2 + r_2^{-1}\mathbf{a}_1\Big)\Big).$$

This proves the result, including the last statement of the Theorem. □

---

[1] The word *fractal* is often used to denote a wider class of sets, but with analogous properties to those here.

[2] A *dilation* with *fixed point* $\mathbf{a}$ and *dilation ratio* $r \geq 0$ is a map $D : \mathbb{R}^n \to \mathbb{R}^n$ of the form $D(\mathbf{x}) = \mathbf{a} + r(\mathbf{x} - \mathbf{a})$.

[3] An *orthonormal transformation* is a linear transformation $O : \mathbb{R}^n \to \mathbb{R}^n$ such that $O^{-1} = O^t$. In $\mathbb{R}^2$ and $\mathbb{R}^3$, such maps consist of a *rotation*, possibly followed by a *reflection*.

[4] A *translation* is a map $T : \mathbb{R}^n \to \mathbb{R}^n$ of the form $T(\mathbf{x}) = \mathbf{x} + \mathbf{a}$.

[5] There is no need to consider dilation ratios $r < 0$. Such maps are obtained by composing a positive dilation with the orthonormal transformation $-I$, where I is the identity map on $\mathbb{R}^n$.

14.3. Dimension of Fractals

A curve has dimension 1, a surface has dimension 2, and a "solid" object has dimension 3. By the k-volume of a "nice" k-dimensional set we mean its length if $k = 1$, area if $k = 2$, and usual volume if $k = 3$.

One can in fact define in a rigorous way the so-called *Hausdorff dimension* of an arbitrary subset of $\mathbb{R}^n$. The Hausdorff dimension is a real number h with $0 \leq h \leq n$. We will not do this here, but you will see it in a later course in the context of *Hausdorff measure*. Here, we will give a simple definition of dimension for fractals, which agrees with the Hausdorff dimension in many important cases.

Suppose by way of motivation that a k-dimensional set K has the property

$$K = K_1 \cup \cdots \cup K_N,$$

where the sets K_i are "almost"[6] disjoint. Suppose moreover, that

$$K_i = S_i[K]$$

where each S_i is a similitude with dilation ratio $r_i > 0$. See Figure 5 for a few examples.

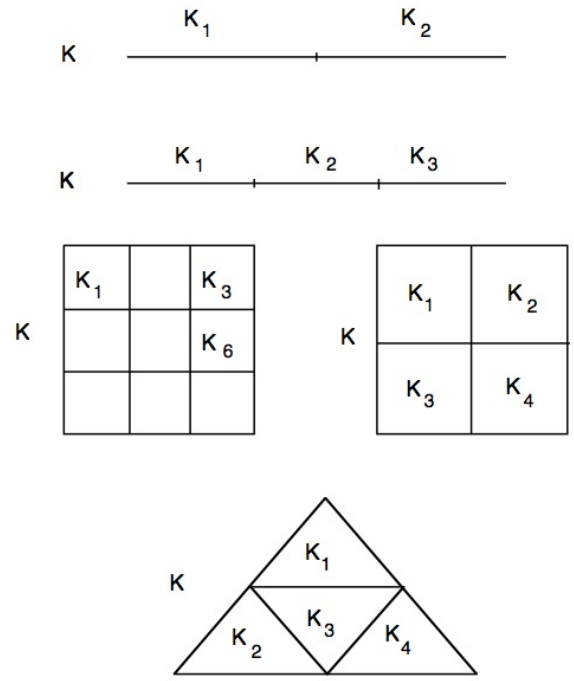

FIGURE 5. The set K is, in each case, the union of rescale copies of itself.

Suppose K is one of the previous examples and K is k-dimensional. Since dilating a k-dimensional set by the ratio r will multiply the k-volume by r^k, it follows that

$$V = r_1^k V + \cdots + r_N^k V,$$

where V is the k-volume of K. Assume $V \neq 0, \infty$, which is reasonable if V is k-dimensional and is certainly the case for the examples in the previous diagram.

[6] In the sense that the Hausdorff dimension of the intersection is less than k.

It follows

$$\sum_{i=1}^{N} r_i^k = 1. \tag{152}$$

In particular, if $r_1 = \ldots = r_N = r$, say, then

$$Nr^k = 1,$$

and so

$$k = \frac{\log N}{\log 1/r}.$$

Thus we have a *formula for the dimension* k in terms of the number N of "almost disjoint" sets K_i whose union is K, and the dilation ratio r used to obtain each K_i from K.

More generally, if the r_i are not all equal, the dimension k can be determined from N and the r_i as follows. Define

$$g(p) = \sum_{i=1}^{N} r_i^p.$$

Then $g(0) = N\,(> 1)$, g is a strictly decreasing function (assuming $0 < r_i < 1$), and $g(p) \to 0$ as $p \to \infty$. It follows there is a unique value of p such that $g(p) = 1$, and from (152) this value of p must be the dimension k.

The preceding considerations lead to the following definition:

DEFINITION 14.3.1. Assume $K \subset \mathbb{R}^n$ is a compact set and

$$K = S_1[K] \cup \cdots \cup S_n[K],$$

where the S_i are similitudes with dilation ratios $0 < r_i < 1$. Then the *similarity dimension* of K is the unique real number D such that

$$1 = \sum_{i=1}^{N} r_i^D.$$

Remarks This is only a good definition if the sets $S_i[K]$ are "almost" disjoint in some sense (otherwise different decompositions may lead to different values of D). In this case one can prove that the similarity dimension and the Hausdorff dimension are equal. The advantage of the similarity dimension is that it is easy to calculate.

Examples For the Koch curve,

$$N = 4,\ r = \frac{1}{3},$$

and so

$$D = \frac{\log 4}{\log 3} \approx 1.2619\,.$$

For the Cantor set,

$$N = 2,\ r = \frac{1}{3},$$

and so

$$D = \frac{\log 2}{\log 3} \approx 0.6309\,.$$

And for the Sierpinski Sponge,

$$N = 20,\ r = \frac{1}{3},$$

and so

$$D = \frac{\log 20}{\log 3} \approx 2.7268\,.$$

14.4. Fractals as Fixed Points

We defined a *fractal* in (150) to be a compact non-empty set $K \subset \mathbb{R}^n$ such that

$$K = \bigcup_{i=1}^{N} S_i[K], \tag{153}$$

for some finite family

$$\mathcal{S} = \{S_1, \ldots, S_N\}$$

of *similitudes* $S_i : \mathbb{R}^n \to \mathbb{R}^n$.

The surprising result is that given *any* finite family $\mathcal{S} = \{S_1, \ldots, S_N\}$ of similitudes with contraction ratios less than 1, there always exists a compact non-empty set K such that (153) is true. Moreover, K is unique.

We can replace the similitudes S_i by *any* contraction map (i.e. Lipschitz map with Lipschitz constant less than 1)[7]. The following Theorem gives the result.

THEOREM 14.4.1 (Existence and Uniqueness of Fractals). *Let $\mathcal{S} = \{S_1, \ldots, S_N\}$ be a family of contraction maps on $\mathbb{R}^n$. Then there is a unique compact non-empty set K such that*

$$K = S_1[K] \cup \cdots \cup S_N[K]. \tag{154}$$

PROOF. For any compact set $A \subset \mathbb{R}^n$, define

$$\mathcal{S}(A) = S_1[A] \cup \cdots \cup S_N[A].$$

Then $\mathcal{S}(A)$ is also a compact subset of $\mathbb{R}^n$ [8].

Let

$$\mathcal{K} = \{A : A \subset \mathbb{R}^n, A \neq \emptyset, \ A \text{ compact}\}$$

denote the family of all compact non-empty subsets of $\mathbb{R}^n$. Then $\mathcal{S} : \mathcal{K} \to \mathcal{K}$, and K satisfies (154) iff K is fixed point of $\mathcal{S}$.

In the next section we will define the *Hausdorff metric* d_H on $\mathcal{K}$, and show that $(\mathcal{K}, d_H)$ is a *complete metric space*. Moreover, we will show that $\mathcal{S}$ is a *contraction mapping on* $\mathcal{K}$ [9], and hence has a unique fixed point K, say. Thus there is a unique compact set K such that (154) is true. □

A Computational Algorithm From the proof of the Contraction Mapping Theorem, we know that if A is *any* compact subset of $\mathbb{R}^n$, then the sequence[10]

$$A, \mathcal{S}(A), \mathcal{S}^2(A), \ldots, \mathcal{S}^k(A), \ldots$$

converges to the fractal K (in the Hausdorff metric).

The approximations to the Koch curve which are shown in Section (14.1.1) were obtained by taking $A = A^{(0)}$ as shown there. We could instead have taken $A = [P, Q]$, in which case the A shown in the first approximation is obtained after just one iteration.

The approximations to the Cantor set were obtained by taking $A = [0, 1]$, and to the Sierpinski sponge by taking A to be the unit cube.

Another convenient choice of A is the set consisting of the N fixed points of the contraction maps $\{S_1, \ldots, S_N\}$. The advantage of this choice is that the sets $\mathcal{S}^k(A)$ are then *subsets* of the fractal K (*exercise*).

---

[7] The restriction to similitudes is only to ensure that the similarity and Hausdorff dimensions agree under suitable extra hypotheses.

[8] Each $S_i[A]$ is compact as it is the continuous image of a compact set. Hence $\mathcal{S}(A)$ is compact as it is a finite union of compact sets.

[9] Do not confuse this with the fact that the S_i are contraction mappings on $\mathbb{R}^n$.

[10] Define $\mathcal{S}^2(A) := \mathcal{S}(\mathcal{S}(A))$, $\mathcal{S}^3(A) := \mathcal{S}(\mathcal{S}^2(A))$, etc.

Variants on the Koch Curve Let K be the Koch curve. We have seen how we can write

$$K = S_1[K] \cup \cdots \cup S_4[K].$$

It is also clear that we can write

$$K = S_1[K] \cup S_2[K]$$

for suitable other choices of similitudes S_1, S_2. Here $S_1[K]$ is the left side of the Koch curve, as shown in the next diagram, and $S_2[K]$ is the right side.

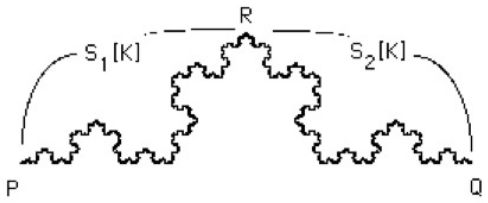

FIGURE 6. The Koch curve as the union of two scaled copies of itself.

The map S_1 consists of a reflection in the PQ axis, followed by a dilation about P with the appropriate dilation factor (which a simple calculation shows to be $1/\sqrt{3}$), followed by a rotation about P such that the final image of Q is R. Similarly, S_2 is a reflection in the PQ axis, followed by a dilation about Q, followed by a rotation about Q such that the final image of P is R.

The previous diagram was generated with a simple Fortran program by the previous computational algorithm, using $A = [P, Q]$, and taking 6 iterations.

Simple variations on $\mathcal{S} = \{S_1, S_2\}$ give quite different fractals. If S_2 is as before, and S_1 is also as before *except* that no reflection is performed, then the following *Dragon* fractal is obtained:

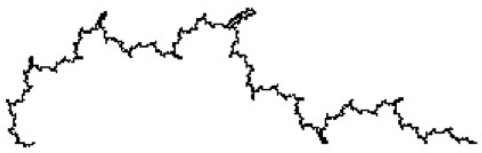

FIGURE 7. Dragon fractal.

If S_1, S_2 are as for the Koch curve, except that no reflection is performed in either case, then the following *Brain* fractal is obtained:

If S_1, S_2 are as for the previous case except that now S_1 maps Q to $(-.15, .6)$ instead of to $R = (0, 1/\sqrt{3}) \approx (0, .6)$, and S_2 maps P to $(.15, .6)$, then the following *Clouds* are obtained:

An important point worth emphasising is that despite the apparent complexity in the fractals we have just sketched, all the relevant information is already *encoded* in the family of generating similitudes. And any such similitude, as in (151), is determined by $r \in (0, 1)$, $\mathbf{a} \in \mathbb{R}^n$, and the $n \times n$ orthogonal matrix O. If $n = 2$,

FIGURE 8. Brain fractal.

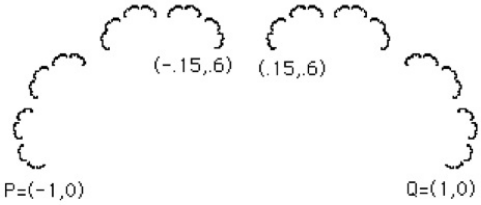

FIGURE 9. Clouds fractal.

then $O = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$, i.e. O is a rotation by θ in an anticlockwise direction, or $O = \begin{bmatrix} \cos\theta & -\sin\theta \\ -\sin\theta & -\cos\theta \end{bmatrix}$, i.e. O is a rotation by θ in an anticlockwise direction followed by reflection in the x-axis.

For a given fractal it is often a simple matter to work "backwards" and find a corresponding family of similitudes. One needs to find $S_1, \ldots, S_N$ such that

$$K = S_1[K] \cup \cdots \cup S_N[K].$$

If equality is only *approximately* true, then it is not hard to show that the fractal generated by $S_1, \ldots, S_N$ will be *approximately* equal to K[11].

In this way, complicated structures can often be encoded in very efficient ways. The point is to find appropriate $S_1, \ldots, S_N$. There is much applied and commercial work (and venture capital!) going into this problem.

14.5. *The Metric Space of Compact Subsets of $\mathbb{R}^n$

Let $\mathcal{K}$ is the family of compact non-empty subsets of $\mathbb{R}^n$.

In this Section we will define the Hausdorff metric $d_{\mathcal{H}}$ on $\mathcal{K}$, show that $(d_{\mathcal{H}}, \mathcal{K})$, is a complete metric space, and prove that the map $\mathcal{S}:\mathcal{K} \to \mathcal{K}$ is a contraction map with respect to $d_{\mathcal{H}}$. This completes the proof of Theorem (14.4.1).

Recall that the distance from $x \in \mathbb{R}^n$ to $A \subset \mathbb{R}^n$ was defined (c.f. (104)) by

$$d(x, A) = \inf_{a \in A} d(x, a). \tag{155}$$

If $A \in \mathcal{K}$, it follows from Theorem 9.4.2 that the sup is realised, i.e.

$$d(x, A) = d(x, a) \tag{156}$$

for some $a \in A$. Thus we could replace *inf* by *min* in (155).

DEFINITION 14.5.1. Let $A \subset \mathbb{R}^n$ and $\epsilon \geq 0$. Then for any $\epsilon > 0$ the *ϵ-enlargement* of A is defined by

$$A_\epsilon = \{x \in \mathbb{R}^n : d(x, A) \leq \epsilon\}.$$

[11] In the Hausdorff distance sense, as we will discuss in Section 14.5.

Hence from (156), $x \in A_\epsilon$ iff $d(x,a) \leq \epsilon$ for some $a \in A$.

The following diagram shows the ϵ-enlargement A_ϵ of a set A.

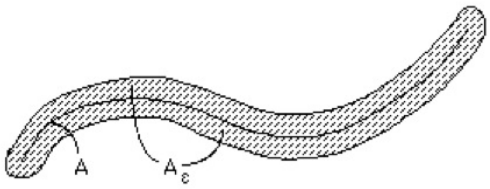

FIGURE 10. The set A_ϵ is the ϵ-enlargement of the curve A.

Properties of the ϵ-enlargement

(1) $A \subset B \Rightarrow A_\epsilon \subset B_\epsilon$.
(2) A_ϵ is closed. (*Exercise:* check that the complement is open)
(3) A_0 is the closure of A. (*Exercise*)
(4) $A \subset A_\epsilon$ for any $\epsilon \geq 0$, and $A_\epsilon \subset A_\gamma$ if $\epsilon \leq \gamma$. (*Exercise*)
(5)

$$\bigcap_{\epsilon > \delta} A_\epsilon = A_\delta. \tag{157}$$

To see this[12] first note that $A_\delta \subset \bigcap_{\epsilon>\delta} A_\epsilon$, since $A_\delta \subset A_\epsilon$ whenever $\epsilon > \delta$. On the other hand, if $x \in \bigcap_{\epsilon>\delta} A_\epsilon$ then $x \in A_\epsilon$ for all $\epsilon > \delta$. Hence $d(x,A) \leq \epsilon$ for all $\epsilon > \delta$, and so $d(x,A) \leq \delta$. That is, $x \in A_\delta$.

We regard two sets A and B as being *close* to each other if $A \subset B_\epsilon$ and $B \subset A_\epsilon$ for some small ϵ. This leads to the following definition.

DEFINITION 14.5.2. Let $A, B \subset \mathbb{R}^n$. Then the *(Hausdorff) distance* between A and B is defined by

$$d_{\mathcal{H}}(A,B) = d(A,B) = \inf\{\epsilon : A \subset B_\epsilon,\ B \subset A_\epsilon\}. \tag{158}$$

We call $d_{\mathcal{H}}$ (or just d), the *Hausdorff metric* on $\mathcal{K}$.

We give some examples in the following diagrams.

FIGURE 11. There is a small ϵ such that both $B \subset A_\epsilon$ and $A \subset B_\epsilon$. Thus $d(A,B) \leq \epsilon$ for this particular ϵ.

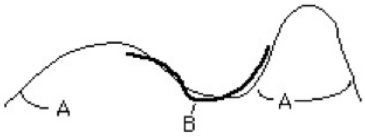

FIGURE 12. $B \subset A_\epsilon$ for some small ϵ, but there is no small ϵ such that $A \subset B_\epsilon$. Thus $d(A, B)$ is *not* small.

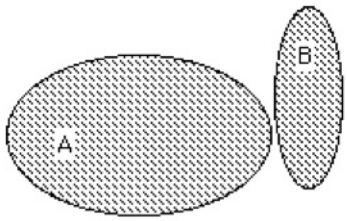

FIGURE 13. A and B are not close, although there are members from A and B respectively which are close to one another.

Remark 1 It is easy to see that the three notions[13] of d are consistent, in the sense that $d(x, y) = d(x, \{y\})$ and $d(x, y) = d(\{x\}, \{y\})$.

Remark 2 Let $\delta = d(A, B)$. Then $A \subset B_\epsilon$ for all $\epsilon > \delta$, and so $A \subset B_\delta$ from (157). Similarly, $B \subset A_\delta$. It follows that the inf in Definition 14.5.2 is realised, and so we could there replace *inf* by *min*.

Notice that if $d(A, B) = \epsilon$, then $d(a, B) \leq \epsilon$ for every $a \in A$. Similarly, $d(b, A) \leq \epsilon$ for every $b \in B$.

Elementary Properties of $d_{\mathcal{H}}$

(1) *(Exercise)* If $E, F, G, H \subset \mathbb{R}^n$ then

$$d(E \cup F, G \cup H) \leq \max\{d(E, G), d(F, H)\}.$$

(2) *(Exercise)* If $A, B \subset \mathbb{R}^n$ and $F : \mathbb{R}^n \to \mathbb{R}^n$ is a Lipschitz map with Lipschitz constant λ, then

$$d(F[A], F[B]) \leq \lambda d(A, B).$$

The Hausdorff metric is not a metric on the set of all subsets of $\mathbb{R}^n$. For example, in $\mathbb{R}$ we have

$$d\Big((a, b), [a, b]\Big) = 0.$$

Thus the distance between two non-equal sets is 0. But if we restrict to compact sets, the d *is* indeed a metric, and moreover it makes $\mathcal{K}$ into a *complete* metric space.

THEOREM 14.5.3. *$(\mathcal{K}, d)$ is a complete metric space.*

---

[12]The result is not completely obvious. Suppose we had *instead* defined $A_\epsilon = \{x \in \mathbb{R}^n : d(x, A) < \epsilon\}$. Let $A = [0, 1] \subset \mathbb{R}$. With this *changed* definition we would have $A_\epsilon = (-\epsilon, 1 + \epsilon)$, and so

$$\bigcap_{\epsilon > \delta} A_\epsilon = \bigcap_{\epsilon > \delta} (-\epsilon, 1 + \epsilon) = [-\delta, 1 + \delta] \neq A_\delta.$$

[13]The distance between two points, the distance between a point and a set, and the Hausdorff distance between two sets.

PROOF. (a) We first prove the three properties of a metric from Definition 6.2.1. In the following, all sets are compact and non-empty.

(1) Clearly $d(A,B) \geq 0$. If $d(A,B) = 0$, then $A \subset B_0$ and $B \subset A_0$. But $A_0 = A$ and $B_0 = B$ since A and B are closed. This implies $A = B$.
(2) Clearly $d(A,B) = d(B,A)$, i.e. symmetry holds.
(3) Finally, suppose $d(A,C) = \delta_1$ and $d(C,B) = \delta_2$. We want to show $d(A,B) \leq \delta_1 + \delta_2$, i.e. that the triangle inequality holds.

    We first claim $A \subset B_{\delta_1+\delta_2}$. To see this consider any $a \in A$. Then $d(a,C) \leq \delta_1$ and so $d(a,c) \leq d_1$ for some $c \in C$ (by (156)). Similarly, $d(c,b) \leq \delta_2$ for some $b \in B$. Hence $d(a,b) \leq \delta_1 + \delta_2$, and so $a \in B_{\delta_1+\delta_2}$, and so $A \subset B_{\delta_1+\delta_2}$, as claimed.

    Similarly, $B \subset A_{\delta_1+\delta_2}$. Thus $d(A,B) \leq \delta_1 + \delta_2$, as required.

(b) Assume $(A^i)_{i\geq 1}$ is a Cauchy sequence (of compact non-empty sets) from $\mathcal{K}$. Let

$$C^j = \overline{\bigcup_{i\geq j} A^i},$$

for $j = 1, 2, \ldots$. Then the C^j are closed and bounded[14], and hence compact. Moreover, the sequence (C^j) is decreasing, i.e.

$$C^j \subset C^k,$$

if $j \geq k$.

Let

$$C = \bigcap_{j\geq 1} C^j.$$

Then C is also closed and bounded, and hence compact.

Claim: $A^k \to C$ in the Hausdorff metric, i.e. $d(A^i, C) \to 0$ as $i \to \infty$.

Suppose that $\epsilon > 0$. Choose N such that

$$j,k \geq N \Rightarrow d(A^j, A^k) \leq \epsilon. \tag{159}$$

We *claim* that

$$j \geq N \Rightarrow d(A^j, C) \leq \epsilon,$$

i.e.

$$j \geq N \Rightarrow C \subset A^j_\epsilon. \tag{160}$$

and

$$j \geq N \Rightarrow A^j \subset C_\epsilon \tag{161}$$

To prove (160), note from (159) that if $j \geq N$ then

$$\bigcup_{i\geq j} A^i \subset A^j_\epsilon.$$

Since A^j_ϵ is closed, it follows

$$C^j = \overline{\bigcup_{i\geq j} A^i} \subset A^j_\epsilon.$$

Since $C \subset C^j$, this establishes (160).

To prove (161), assume $j \geq N$ and suppose

$$x \in A^j.$$

[14]This follows from the fact that (A^k) is a Cauchy sequence.

Then from (159), $x \in A^k_\epsilon$ if $k \geq j$, and so

$$k \geq j \Rightarrow x \in \bigcup_{i \geq k} A^i_\epsilon \subset \left(\bigcup_{i \geq k} A^i \right)_\epsilon \subset C^k_\epsilon,$$

where the first "$\subset$" follows from the fact $A^i_\epsilon \subset \left(\bigcup_{i \geq k} A^i\right)_\epsilon$ for each $i \geq k$. For each $k \geq j$, we can then choose $x^k \in C^k$ with

$$d(x, x^k) \leq \epsilon. \tag{162}$$

Since $(x^k)_{k \geq j}$ is a bounded sequence, there exists a subsequence converging to y, say. For each set C^k with $k \geq j$, all terms of the sequence $(x^i)_{i \geq j}$ beyond a certain term are members of C^k. Hence $y \in C^k$ as C^k is closed. But $C = \bigcap_{k \geq j} C^k$, and so $y \in C$.

Since $y \in C$ and $d(x, y) \leq \epsilon$ from (162), it follows that

$$x \in C_\epsilon$$

As x was an arbitrary member of A^j, this proves (161) □

Recall that if $\mathcal{S} = \{S_1, \ldots, S_N\}$, where the S_i are contractions on $\mathbb{R}^n$, then we defined $\mathcal{S}{:}\mathcal{K} \to \mathcal{K}$ by $\mathcal{S}(K) = S_1[K] \cup \ldots \cup S_N[K]$.

THEOREM 14.5.4. *If $\mathcal{S}$ is a finite family of contraction maps on $\mathbb{R}^n$, then the corresponding map $\mathcal{S}{:}\mathcal{K} \to \mathcal{K}$ is a contraction map (in the Hausdorff metric).*

PROOF. Let $\mathcal{S} = \{S_1, \ldots, S_N\}$, where the S_i are contractions on $\mathbb{R}^n$ with Lipschitz constants $r_1, \ldots, r_N < 1$.

Consider any $A, B \in \mathcal{K}$. From the earlier properties of the Hausdorff metric it follows

$$\begin{aligned} d(\mathcal{S}(A), \mathcal{S}(B)) &= d\left(\bigcup_{1 \leq i \leq N} S_i[A], \bigcup_{1 \leq i \leq N} S_i[B] \right) \\ &\leq \max_{1 \leq i \leq N} d(S_i[A], S_i[B]) \\ &\leq \max_{1 \leq i \leq N} r_i d(A, B), \end{aligned}$$

Thus $\mathcal{S}$ is a contraction map with Lipschitz constant given by $\max\{r_1, \ldots, r_n\}$. □

14.6. *Random Fractals

There is also a notion of a *random* fractal. A random fractal is not a particular compact set, but is a probability distribution on $\mathcal{K}$, the family of all compact subsets (of $\mathbb{R}^n$).

One method of obtaining random fractals is to *randomise* the Computational Algorithm in Section 14.4.

As an example, consider the Koch curve. In the discussion, "Variants on the Koch Curve", in Section 14.4, we saw how the Koch curve could be generated from two similitudes S_1 and S_2 applied to an initial compact set A. We there took A to be the closed interval $[P, Q]$, where P and Q were the fixed points of S_1 and S_2 respectively. The construction can be randomised by selecting $\mathcal{S} = \{S_1, S_2\}$ at each stage of the iteration according to some probability distribution.

For example, assume that S_1 is always a reflection in the PQ axis, followed by a dilation about P and then followed by a rotation, such that the image of Q is some point R. Assume that S_2 is a reflection in the PQ axis, followed by a dilation about Q and then followed by a rotation about Q, such that the image of P is the *same* point R. Finally, assume that R is chosen according to some

probability distribution over $\mathbb{R}^2$ (this then gives a probability distribution on the set of possible $\mathcal{S}$). We have chosen R to be normally distributed in $\Re^2$ with mean position $(0, \sqrt{3}/3)$ and variance $(.4, .5)$. Figure 14 shows three realisations.

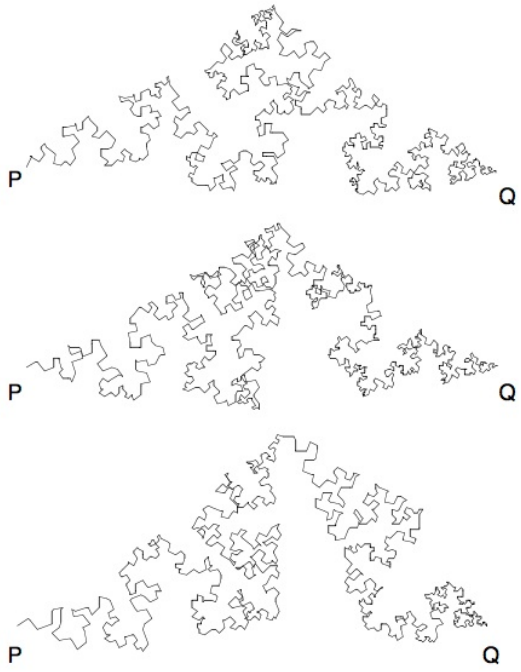

FIGURE 14. Three realisations of a random fractal.

CHAPTER 15

Compactness

15.1. Definitions

In Definition 9.3.1 we defined the notion of a compact subset of a *metric* space. As noted following that Definition, the notion defined there is usually called *sequential compactness*.

We now give another definition of compactness, in terms of coverings by open sets (which applies to any topological space)[1]. We will show that compactness and sequential compactness agree for *metric* spaces. (There are examples to show that neither implies the other in an arbitrary topological space.)

DEFINITION 15.1.1. A collection X_α) of subsets of a set X is a *cover* or *covering* of a subset Y of X if $\cup_\alpha X_\alpha supseteq Y$.

DEFINITION 15.1.2. A subset K of a metric space (X, d) is *compact* if whenever

$$K \subset \bigcup_{U \in \mathcal{F}} U$$

for some collection $\mathcal{F}$ of open sets from X, then

$$K \subset U_1 \cup \ldots \cup U_N$$

for some $U_1, \ldots, U_N \in \mathcal{F}$. That is, *every open cover has a finite subcover.*

If X is compact, we say the metric space itself is compact.

Remark *Exercise:* A subset K of a metric space (X, d) is compact in the sense of the previous definition iff the induced metric space (K, d) is compact.

The main point is that $U \subset K$ is open in (K, d) iff $U = K \cap V$ for some $V \subset X$ which is open in X.

Examples It is clear that if $A \subset \mathbb{R}^n$ and A is unbounded, then A is *not* compact according to the definition, since

$$A \subset \bigcup_{i=1}^{\infty} B_i(0),$$

but no finite number of such open balls covers A.

Also $B_1(0)$ is not compact, as

$$B_1(0) = \bigcup_{i=2}^{\infty} B_{1-1/i}(0),$$

and again no finite number of such open balls covers $B_1(0)$.

In Example 1 of Section 9.3 we saw that the sequentially compact subsets of $\mathbb{R}^n$ are precisely the closed bounded subsets of $\mathbb{R}^n$. It then follows from Theorem 15.2.1 below that the compact subsets of $\mathbb{R}^n$ are *precisely* the closed bounded subsets of $\mathbb{R}^n$.

In a general metric space, this is *not* true, as we will see.

[1] You will consider general topological spaces in later courses.

There is an equivalent definition of compactness in terms of closed sets, which is called the *finite intersection property*.

THEOREM 15.1.3. *A topological space X is compact iff for every family $\mathcal{F}$ of closed subsets of X,*

$$\bigcap_{C\in\mathcal{F}} C = \emptyset \Rightarrow C_1 \cap \cdots \cap C_N = \emptyset \text{ for some finite subfamily } \{C_1, \ldots, C_N\} \subset \mathcal{F}.$$

PROOF. The result follows from De Morgan's laws (*exercise*). □

15.2. Compactness and Sequential Compactness

We now show that these two notions agree in a metric space.

The following is an example of a *non-trivial* proof. Think of the case that $X = [0,1] \times [0,1]$ with the induced metric. Note that an open ball $B_r(a)$ in X is just the intersection of X with the usual ball $B_r(a)$ in $\mathbb{R}^2$.

THEOREM 15.2.1. *A metric space is compact iff it is sequentially compact.*

PROOF. *First suppose that the metric space (X, d) is compact.*

Let $(x_n)_{n=1}^{\infty}$ be a sequence from X. We want to show that some subsequence converges to a limit in X.

Let $A = \{x_n\}$. Note that A may be finite (in case there are only a finite number of distinct terms in the sequence).

(1) *Claim:* If A is finite, then some subsequence of (x_n) converges.

Proof: If A is finite there is only a finite number of distinct terms in the sequence. Thus there is a subsequence of (x_n) for which all terms are equal. This subsequence converges to the common value of all its terms.

(2) *Claim:* If A is infinite, then A has at least one limit point.

Proof: Assume A has *no* limit points.

It follows that $A = \overline{A}$ from Definition 6.3.4, and so A is *closed* by Theorem 6.4.6.

It also follows that each $a \in A$ is not a limit point of A, and so from Definition 6.3.3 there exists a neighbourhood V_a of a (take V_a to be some open ball centred at a) such that

$$V_a \cap A = \{a\}. \tag{163}$$

In particular,

$$X = A^c \cup \bigcup_{a\in A} V_a$$

gives an *open* cover of X. By compactness, there is a finite subcover. Say

$$X = A^c \cup V_{a_1} \cup \cdots \cup V_{a_N}, \tag{164}$$

for some $\{a_1, \ldots, a_N\} \subset A$. But this is impossible, as we see by choosing $a \in A$ with $a \neq a_1, \ldots, a_N$ (remember that A is infinite), and noting from (163) that a cannot be a member of the right side of (164). This establishes the *claim*.

(3) *Claim:* If x is a limit point of A, then some subsequence of (x_n) converges to x[2].

Proof: Any neighbourhood of x contains an *infinite* number of points from A (Proposition 6.3.5) and hence an infinite number of terms from the sequence (x_n). Construct a subsequence (x'_k) from (x_n) so that, for each k, $d(x'_k, x) < 1/k$ and x'_k is a term from the original sequence (x_n) which occurs later in that sequence

[2]From Theorem 7.5.1 there is a sequence (y_n) consisting of points from A (and hence of terms from the sequence (x_n)) converging to x; but this sequence may not be a subsequence of (x_n) because the terms may not occur in the right order. So we need to be a little more careful in order to prove the claim.

than any of the finite number of terms $x'_1, \ldots, x'_{k-1}$. Then (x'_k) is the required subsequence.

From (1), (2) and (3) we have established that compactness implies sequential compactness.

Next assume that (X, d) is sequentially compact.

(4) *Claim:* [3] For each integer k there is a finite set $\{x_1, \ldots, x_N\} \subset X$ such that

$$x \in X \Rightarrow d(x_i, x) < 1/k \text{ for some } i = 1, \ldots, N.$$

Proof: Choose x_1; choose x_2 so $d(x_1, x_2) \geq 1/k$; choose x_3 so $d(x_i, x_3) \geq 1/k$ for $i = 1, 2$; choose x_4 so $d(x_i, x_4) \geq 1/k$ for $i = 1, 2, 3$; etc. This procedure must terminate in a finite number of steps

> For if not, we have an *infinite* sequence (x_n). By sequential compactness, some subsequence converges and in particular is Cauchy. But this contradicts the fact that any two members of the subsequence must be distance at least $1/k$ apart.

Let $x_1, \ldots, x_N$ be some such (finite) sequence of maximum length.

It follows that any $x \in X$ satisfies $d(x_i, x) < 1/k$ for some $i = 1, \ldots, N$. For if not, we could enlarge the sequence $x_1, \ldots, x_N$ by adding x, thereby contradicting its maximality.

(5) *Claim:* There exists a countable dense[4] subset of X.

Proof: Let A_k be the finite set of points constructed in (4). Let $A = \bigcup_{k=1}^{\infty} A_k$. Then A is countable. It is also dense, since if $x \in X$ then there exist points in A arbitrarily close to x; i.e. x is in the closure of A.

(6) *Claim:* Every open cover of X has a countable subcover[5].

Proof: Let $\mathcal{F}$ be a cover of X by open sets. For each $x \in A$ (where A is the countable dense set from (5)) and each rational number $r > 0$, *if* $B_r(x) \subset U$ for some $U \in \mathcal{F}$, choose *one* such set U and denote it by $U_{x,r}$. The collection $\mathcal{F}^*$ of all such $U_{x,r}$ is a *countable* subcollection of $\mathcal{F}$. Moreover, we claim it is a cover of X.

To see this, suppose $y \in X$ and choose $U \in \mathcal{F}$ with $y \in U$. Choose $s > 0$ so $B_s(y) \subset U$. Choose $x \in A$ so $d(x, y) < s/4$ and choose a rational number r so $s/4 < r < s/2$. Then $y \in B_r(x) \subset B_s(y) \subset U$. In particular, $B_r(x) \subset U \in \mathcal{F}$ and so there *is* a set $U_{x,r} \in \mathcal{F}^*$ (by definition of $\mathcal{F}^*$). Moreover, $y \in B_r(x) \subset U_{x,r}$ and so y is a member of the union of all sets from $\mathcal{F}^*$. Since y was an arbitrary member of X, $\mathcal{F}^*$ is a countable cover of X.

(7) *Claim:* Every countable open cover of X has a finite subcover.

Proof: Let $\mathcal{G}$ be a countable cover of X by open sets, which we write as $\mathcal{G} = \{U_1, U_2, \ldots\}$. Let

$$V_n = U_1 \cup \cdots \cup U_n$$

for $n = 1, 2, \ldots$. Notice that the sequence $(V_n)_{n=1}^{\infty}$ is an increasing sequence of sets. We need to show that

$$X = V_n$$

for some n.

Suppose not. Then there exists a sequence (x_n) where $x_n \notin V_n$ for each n. By assumption of sequential compactness, some subsequence (x'_n) converges to x, say.

---

[3] The claim says that X is *totally bounded*, see Section 15.5.

[4] A subset of a topological space is *dense* if its closure is the entire space. A topological space is said to be *separable* if it has a countable dense subset. In particular, the reals are separable since the rationals form a countable dense subset. Similarly $\mathbb{R}^n$ is separable for any n.

[5] This is called the *Lindelöf* property.

Since $\mathcal{G}$ is a cover of X, it follows $x \in U_N$, say, and so $x \in V_N$. But V_N is open and so

$$x'_n \in V_N \text{ for } n > M, \tag{165}$$

for some M.

On the other hand, $x_k \notin V_k$ for all k, and so

$$x_k \notin V_N \tag{166}$$

for all $k \geq N$ since the (V_k) are an increasing sequence of sets.

From 165 and 166 we have a contradiction. This establishes the claim.

From (6) and (7) it follows that sequential compactness implies compactness. □

Exercise Use the definition of compactness in Definition 15.1.2 to simplify the proof of Dini's Theorem (Theorem 12.1.3).

15.3. *Lebesgue covering theorem

DEFINITION 15.3.1. The *diameter* of a subset Y of a metric space (X, d) is

$$d(Y) = \sup\{d(y, y') : y, y' \in Y\}.$$

Note this is not necessarily the same as the diameter of the smallest ball containing the set, however, $Y \subseteq \overline{B_{d(Y)}(y)}l$ for any $y \in Y$.

THEOREM 15.3.2. *Suppose (G_α) is a covering of a compact metric space (X, d) by open sets. Then there exists $\delta > 0$ such that any (non-empty) subset Y of X whose diameter is less than δ lies in some G_α.*

PROOF. Supposing the result fails, there are non-empty subsets $C_n \subseteq X$ with $d(C_n) < n^{-1}$ each of which fails to lie in any single G_α. Taking $x_n \in C_n$, (x_n) has a convergent subsequence, say, $x_{n_j} \to x$. Since (G_α) is a covering, there is some α such that $x \in G_\alpha$. Now G_α is open, so that $B_\epsilon(x) \subseteq G_\alpha$ for some $\epsilon > 0$. But $x_{n_j} \in B_\epsilon(x)$ for all j sufficiently large. Thus for j so large that $n_j > 2\epsilon^{-1}$, we have $C_{n_j} \subseteq B_{n_j^{-1}}(x_{n)j}) \subset B_\epsilon x$, contrary to the definition of (x_{n_j}). □

15.4. Consequences of Compactness

We review some facts:

1: As we saw in the previous section, *compactness and sequential compactness are the same* in a metric space.

2: In a metric space, *if a set is compact then it is closed and bounded.* The proof of this is similar to the proof of the corresponding fact for $\mathbb{R}^n$ given in the last two paragraphs of the proof of Corollary 9.2.2. As an *exercise* write out the proof.

3: *In $\mathbb{R}^n$ if a set is closed and bounded then it is compact.* This is proved in Corollary 9.2.2, using the Bolzano Weierstrass Theorem 9.2.1.

4: *It is not true in general that closed and bounded sets are compact.* In Remark 2 of Section 9.2 we see that the set

$$\mathcal{F} := C[0, 1] \cap \{f : ||f||_\infty \leq 1\}$$

is not compact. But it *is* closed and bounded (*exercise*).

5: *A subset of a compact metric space is compact iff it is closed. Exercise:* prove this directly from the definition of sequential compactness; and then give another proof directly from the definition of compactness.

We also have the following facts about continuous functions and compact sets:

6: *If f is continuous and K is compact, then $f[K]$ is compact* (Theorem 11.5.1).

7: *Suppose $f : K \to \mathbb{R}$, f is continuous and K is compact. Then f is bounded above and below and has a maximum and minimum value.* (Theorem 11.5.2)

8: *Suppose $f : K \to Y$, f is continuous and K is compact. Then f is uniformly continuous.* (Theorem 11.6.2)

It is not true in general that if $f : X \to Y$ is continuous, one-one and onto, then the inverse of f is continuous.

For example, define the function

$$f : [0, 2\pi) \to S^1 = \{(\cos\theta, \sin\theta) : [0, 2\pi)\} \subset \mathbb{R}^2$$

by

$$f(\theta) = (\cos\theta, \sin\theta).$$

Then f is clearly continuous (assuming the functions cos and sin are continuous), one-one and onto. But f^{-1} is not continuous, as we easily see by finding a sequence x_n $(\in S^1) \to (1, 0)$ $(\in S^1)$ such that $f^{-1}(x_n) \not\to f^{-1}((1, 0)) = 0$.

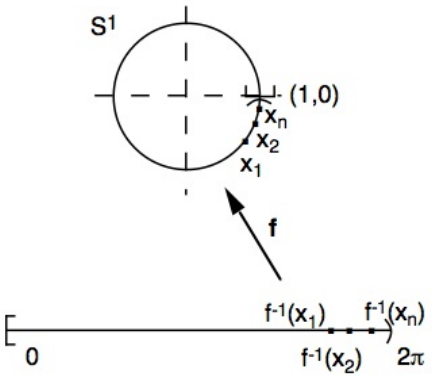

FIGURE 1. A one-one continuous map from the noncompact space $[0, 1)$ onto the compact space S^1. The inverse map is *not* continuous.

Note that $[0, 2\pi)$ is not compact (*exercise:* prove directly from the definition of sequential compactness that it is not sequentially compact, and directly from the definition of compactness that it is not compact).

THEOREM 15.4.1. *Let $f : X \to Y$ be continuous and bijective. If X is compact then f is a homeomorphism.*

PROOF. We need to show that the inverse function $f^{-1} : Y \to X$ [6] is continuous. To do this, we need to show that the *inverse* image under f^{-1} of a closed set $C \subset X$ is closed in Y; equivalently, that the *image* under f of C is closed.

But if C is closed then it follows C is compact from remark 5 at the beginning of this section; hence $f[C]$ is compact by remark 6; and hence $f[C]$ is closed by remark 2. This completes the proof. □

We could also give a proof using sequences (*Exercise*).

[6] The function f^{-1} exists as f is one-one and onto.

15.5. A Criterion for Compactness

We now give an important necessary and sufficient condition for a metric space to be compact. This will generalise the Bolzano-Weierstrass Theorem, Theorem 9.2.1. In fact, the proof of one direction of the present Theorem is very similar.

The most important application will be to finding compact subsets of $C[a, b]$.

DEFINITION 15.5.1. Let (X, d) be a metric space. A subset $A \subset X$ is *totally bounded* iff for every $\delta > 0$ there exist a finite set of points $x_1, \ldots, x_N \in X$ such that

$$A \subset \bigcup_{i=1}^{N} B_\delta(x_i).$$

Remark If necessary, we can assume that the centres of the balls belong to A.

To see this, first cover A by balls of radius $\delta/2$, as in the Definition. Let the centres be $x_1, \ldots, x_N$. If the ball $B_{\delta/2}(x_i)$ contains some point $a_i \in A$, then we replace the ball by the larger ball $B_\delta(a_i)$ which contains it. If $B_{\delta/2}(x_i)$ contains no point from A then we discard this ball. In this way we obtain a finite cover of A by balls of radius δ with centres in A.

Remark In any metric space, "totally bounded" implies "bounded". For if $A \subset \bigcup_{i=1}^{N} B_\delta(x_i)$, then $A \subset B_R(x_1)$ where $R = \max_i d(x_i, x_1) + \delta$.

In $\mathbb{R}^n$, we also have that "bounded" implies "totally bounded". To see this in $\mathbb{R}^2$, cover the bounded set A by a finite square lattice with grid size δ. Then A is covered by the finite number of open balls with centres at the vertices and radius $\delta\sqrt{2}$. In $\mathbb{R}^n$ take radius $\delta\sqrt{n}$.

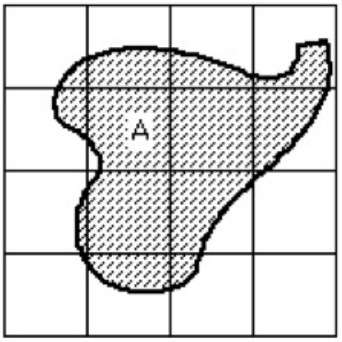

FIGURE 2. Covering of the set A by a finite square lattice.

Note that as the dimension n increases, the number of vertices in a grid of total side L is of the order $(L/\delta)^n$.

It is not true in a general metric space that "bounded" implies "totally bounded". The problem, as indicated roughly by the fact that $(L/\delta)^n \to \infty$ as $n \to \infty$, is that the number of balls of radius δ necessary to cover *may* be infinite if A is not a subset of a finite dimensional vector space.

In particular, the set of functions $A = \{f_n\}_{n \geq 1}$ in Remark 2 of Section 9.2 is clearly bounded. But it is not totally bounded, since the distance between any two functions in A is 1, and so no finite number of balls of radius less than $1/2$ can cover A as any such ball can contain at most one member of A.

In the following theorem, first think of the case $X = [a, b]$.

THEOREM 15.5.2. *A metric space X is compact iff it is complete and totally bounded.*

PROOF. (a) *First assume X is compact.*

In order to prove X is *complete*, let (x_n) be a Cauchy sequence from X. Since compactness in a metric space implies sequential compactness by Theorem 15.2.1, a subsequence (x'_k) converges to some $x \in X$. We claim the original sequence also converges to x.

This follows from the fact that

$$d(x_n, x) \leq d(x_n, x'_k) + d(x'_k, x).$$

Given $\epsilon > 0$, first use convergence of (x'_k) to choose N_1 so that $d(x'_k, x) < \epsilon/2$ if $k \geq N_1$. Next use the fact (x_n) is Cauchy to choose N_2 so $d(x_n, x'_k) < \epsilon/2$ if $k, n \geq N_2$. Hence $d(x_n, x) < \epsilon$ if $n \geq \max\{N_1, N_2\}$.

That X is *totally bounded* follows from the observation that the set of all balls $B_\delta(x)$, where $x \in X$, is an open cover of X, and so has a finite subcover by compactness of A.

(b) *Next assume X is complete and totally bounded.*

Let (x_n) be a sequence from X, which for convenience we rewrite as $(x_n^{(1)})$.

Using total boundedness, cover X by a *finite* number of balls of radius 1. Then at least one of these balls must contain an *(infinite)* subsequence of $(x_n^{(1)})$. Denote this subsequence by $(x_n^{(2)})$.

Repeating the argument, cover X by a *finite* number of balls of radius $1/2$. At least one of these balls must contain an *(infinite)* subsequence of $(x_n^{(2)})$. Denote this subsequence by $(x_n^{(3)})$.

Continuing in this way we find sequences

$$\begin{gathered}(x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \ldots) \\ (x_1^{(2)}, x_2^{(2)}, x_3^{(2)}, \ldots) \\ (x_1^{(3)}, x_2^{(3)}, x_3^{(3)}, \ldots) \\ \vdots\end{gathered}$$

where each sequence is a subsequence of the preceding sequence and the terms of the ith sequence are all members of some ball of radius $1/i$.

Define the (diagonal) sequence (y_i) by $y_i = x_i^{(i)}$ for $i = 1, 2, \ldots$. This is a subsequence of the original sequence.

Notice that for each i, the terms $y_i, y_{i+1}, y_{i+2}, \ldots$ are all members of the ith sequence and so lie in a ball of radius $1/i$. It follows that (y_i) is a Cauchy sequence. Since X is complete, it follows (y_i) converges to a limit in X.

This completes the proof of the theorem, since (y_i) is a subsequence of the original sequence (x_n). □

The following is a direct generalisation of the Bolzano-Weierstrass theorem.

COROLLARY 15.5.3. *A subset of a complete metric space is compact iff it is closed and totally bounded.*

PROOF. Let X be a complete metric space and A be a subset.

If A is closed (in X) then A (with the induced metric) is complete, by the generalisation following Theorem 8.2.2. Hence A is compact from the previous theorem.

If A is compact, then A is complete and totally bounded from the previous theorem. Since A is complete it must be closed[7] in X. □

15.6. Equicontinuous Families of Functions

Throughout this Section you should think of the case $X = [a, b]$ and $Y = \mathbb{R}$.

We will use the notion of equicontinuity in the next Section in order to give an important criterion for a family $\mathcal{F}$ of continuous functions to be compact (in the sup metric).

DEFINITION 15.6.1. Let (X, d) and (Y, ρ) be metric spaces. Let $\mathcal{F}$ be a family of functions from X to Y.

Then $\mathcal{F}$ is *equicontinuous* at the point $x \in X$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$d(x, x') < \delta \Rightarrow \rho(f(x), f(x')) < \epsilon$$

for all $f \in \mathcal{F}$. The family $\mathcal{F}$ is *equicontinuous* if it is equicontinuous at every $x \in X$.

$\mathcal{F}$ is *uniformly equicontinuous* on X if for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$d(x, x') < \delta \Rightarrow \rho(f(x), f(x')) < \epsilon$$

for all $x \in X$ and *all* $f \in \mathcal{F}$.

Remarks

(1) The members of an equicontinuous family of functions are clearly continuous; and the members of a uniformly equicontinuous family of functions are clearly uniformly continuous.
(2) In case of equicontinuity, δ may depend on ϵ and x but *not* on the particular function $f \in \mathcal{F}$. For uniform equicontinuity, δ may depend on ϵ, but *not* on x or x' (provided $d(x, x') < \delta$) and *not* on f.
(3) The most important of these concepts is the case of *uniform equicontinuity.*

Example 1 Let $\mathrm{Lip}_M(X; Y)$ be the set of Lipschitz functions $f : X \to Y$ with Lipschitz constant at most M. The family $\mathrm{Lip}_M(X; Y)$ is uniformly equicontinuous on the set X. This is easy to see since we can take $\delta = \epsilon/M$ in the Definition. Notice that δ does not depend on either x or on the particular function $f \in \mathrm{Lip}_M(X; Y)$.

Example 2 The family of functions $f_n(x) = x^n$ for $n = 1, 2, \ldots$ and $x \in [0, 1]$ is *not* equicontinuous at 1. To see this just note that if $x < 1$ then

$$|f_n(1) - f_n(x)| = 1 - x^n > 1/2, \text{ say}$$

for all sufficiently large n. So taking $\epsilon = 1/2$ there is no $\delta > 0$ such that $|1 - x| < \delta$ implies $|f_n(1) - f_n(x)| < 1/2$ for *all* n.

On the other hand, this family *is* equicontinuous at each $a \in [0, 1)$. In fact it is uniformly equicontinuous on any interval $[0, b]$ provided $b < 1$.

To see this, note

$$|f_n(a) - f_n(x)| = |a^n - x^n| = |f'(\xi)|\,|a - x| = n\xi^{n-1}|a - x|$$

for some ξ between x and a. If $a, x \leq b < 1$, then $\xi \leq b$, and so $n\xi^{n-1}$ is bounded by a constant $c(b)$ that depends on b but not on n (this is clear since $n\xi^{n-1} \leq nb^{n-1}$, and $nb^{n-1} \to 0$ as $n \to \infty$; so we can take $c(b) = \max_{n \geq 1} nb^{n-1}$). Hence

$$|f_n(1) - f_n(x)| < \epsilon$$

[7] To see this suppose $(x_n) \subset A$ and $x_n \to x \in X$. Then (x_n) is Cauchy, and so by completeness has a limit $x' \in A$. But then in X we have $x_n \to x'$ as well as $x_n \to x$. By uniqueness of limits in X it follows $x = x'$, and so $x \in A$.

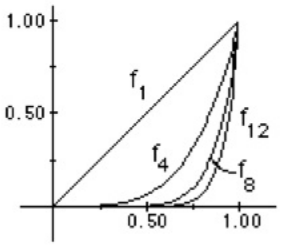

FIGURE 3. The family of continuous functions $f_n(x) = x^n$ for $n = 1, 2, 3, \ldots$ is *not* an equicontinuous family.

provided

$$|a - x| < \frac{\epsilon}{c(b)}.$$

Exercise: Prove that the family in Example 2 is uniformly equicontinuous on $[0, b]$ (if $b < 1$) by finding a Lipschitz constant independent of n and using the result in Example 1.

Example 3 In the first example, equicontinuity followed from the fact that the families of functions had a uniform Lipschitz bound.

More generally, families of Hölder continuous functions with a fixed exponent α and a fixed constant M (see Definition 11.3.2) are also uniformly equicontinuous. This follows from the fact that in the definition of uniform equicontinuity we can take $\delta = \left(\frac{\epsilon}{M}\right)^{1/\alpha}$.

We saw in Theorem 11.6.2 that a continuous function on a compact metric space is uniformly continuous. Almost exactly the same proof shows that an equicontinuous family of functions defined on a compact metric space is uniformly equicontinuous.

THEOREM 15.6.2. *Let $\mathcal{F}$ be an equicontinuous family of functions $f : X \to Y$, where (X, d) is a compact metric space and (Y, ρ) is a metric space. Then $\mathcal{F}$ is uniformly equicontinuous.*

PROOF. Suppose $\epsilon > 0$. For each $x \in X$ there exists $\delta_x > 0$ (where δ_x may depend on x as well as ϵ) such that

$$x' \in B_{\delta_x}(x) \Rightarrow \rho(f(x), f(x')) < \epsilon$$

for all $f \in \mathcal{F}$.

The family of all balls $B(x, \delta_x/2) = B_{\delta_x/2}(x)$ forms an open cover of X. By compactness there is a finite subcover $B_1, \ldots, B_N$ by open balls with centres $x_1, \ldots, x_n$ and radii $\delta_1/2 = \delta_{x_1}/2, \ldots, \delta_N/2 = \delta_{x_N}/2$, say.

Let

$$\delta = \min\{\delta_1, \ldots, \delta_N\}.$$

Take any $x, x' \in X$ with $d(x, x') < \delta/2$. See Figure 4.

Then $d(x_i, x) < \delta_i/2$ for some x_i since the balls $B_i = B(x_i, \delta_i/2)$ cover X. Moreover,

$$d(x_i, x') \leq d(x_i, x) + d(x, x') < \delta_i/2 + \delta/2 \leq \delta_i.$$

In particular, both $x, x' \in B(x_i, \delta_i)$.

It follows that for all $f \in \mathcal{F}$,

$$\begin{aligned} \rho(f(x), f(x')) &\leq \rho(f(x), f(x_i)) + \rho(f(x_i), f(x')) \\ &< \epsilon + \epsilon = 2\epsilon. \end{aligned}$$

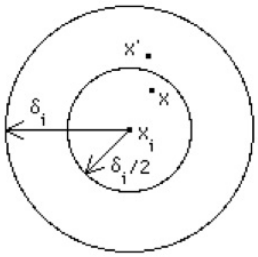

FIGURE 4. Diagram for the proof of Theorem 15.6.2.

Since ϵ is arbitrary, this proves $\mathcal{F}$ is a *uniformly* equicontinuous family of functions. □

15.7. Arzela-Ascoli Theorem

Throughout this Section you should think of the case $X = [a, b]$ and $Y = \mathbb{R}$.

THEOREM 15.7.1 (Arzela-Ascoli). *Let (X, d) be a compact metric space and let $\mathcal{C}(X; \mathbb{R}^n)$ be the family of continuous functions from X to $\mathbb{R}^n$. Let $\mathcal{F}$ be any subfamily of $\mathcal{C}(X; \mathbb{R}^n)$ which is closed, uniformly bounded*[8] *and uniformly equicontinuous. Then $\mathcal{F}$ is compact in the sup metric.*

Remarks

(1) Recall from Theorem 15.6.2 that since X is compact, we could replace *uniform equicontinuity* by *equicontinuity* in the statement of the Theorem.
(2) Although we do not prove it now, the converse of the theorem is also true. That is, $\mathcal{F}$ is compact *iff* it is closed, uniformly bounded, and uniformly equicontinuous.
(3) The Arzela-Ascoli Theorem is one of the most important theorems in Analysis. It is usually used to show that certain sequences of functions have a convergent subsequence (in the sup norm). See in particular the next section.

Example 1 Let $\mathcal{C}^\alpha_{M,K}(X; \mathbb{R}^n)$ denote the family of *Hölder continuous* functions $f : X \to \mathbb{R}^n$ with exponent α and constant M (as in Definition 11.3.2), which also satisfy the uniform bound

$$|f(x)| \leq K \quad \text{for all } x \in X.$$

Claim: $\mathcal{C}^\alpha_{M,K}(X; \mathbb{R}^n)$ is closed, uniformly bounded and uniformly equicontinuous, and hence compact by the Arzela-Ascoli Theorem.

We saw in Example 3 of the previous Section that $\mathcal{C}^\alpha_{M,K}(X; \mathbb{R}^n)$ is equicontinuous.

Boundedness is immediate, since the distance from any $f \in \mathcal{C}^\alpha_{M,K}(X; \mathbb{R}^n)$ to the zero function is at most K (in the sup metric).

In order to show *closure* in $\mathcal{C}(X; \mathbb{R}^n)$, suppose that

$$f_n \in \mathcal{C}^\alpha_{M,K}(X; \mathbb{R}^n)$$

for $n = 1, 2, \ldots$, and

$$f_n \to f \text{ uniformly as } n \to \infty,$$

(uniform convergence is just convergence in the sup metric). We know f is continuous by Theorem 12.3.1. We want to show $f \in \mathcal{C}^\alpha_{M,K}(X; \mathbb{R}^n)$.

[8] That is, bounded in the sup metric.

We first need to show

$$|f(x)| \leq K \text{ for each } x \in X.$$

But for any $x \in X$ we have $|f_n(x)| \leq K$, and so the result follows by letting $n \to \infty$.

We also need to show that

$$|f(x) - f(y)| \leq M|x - y|^\alpha \text{ for all } x, y \in X.$$

But for any fixed $x, y \in X$ this is true with f replaced by f_n, and so is true for f as we see by letting $n \to \infty$.

This completes the proof that $\mathcal{C}^\alpha_{M,K}(X; \mathbb{R}^n)$ is closed in $\mathcal{C}(X; \mathbb{R}^n)$.

Example 2 An important case of the previous example is $X = [a, b]$, $\mathbb{R}^n = \mathbb{R}$, and $\mathcal{F} = \text{Lip}_{M,K}[a, b]$ (the class of real-valued Lipschitz functions with Lipschitz constant at most M and uniform bound at most K).

You should keep this case in mind when reading the proof of the Theorem.

REMARK 15.7.2. The Arzela-Ascoli Theorem implies that any sequence from the class $\text{Lip}_{M,K}[a, b]$ has a convergent subsequence. This is not true for the set $\mathcal{C}_K[a, b]$ of all continuous functions f from $\mathcal{C}[a, b]$ merely satisfying $\sup |f| \leq K$. For example, consider the sequence of functions (f_n) defined by

$$f_n(x) = \sin nx \quad x \in [0, 2\pi].$$

See figure 5.

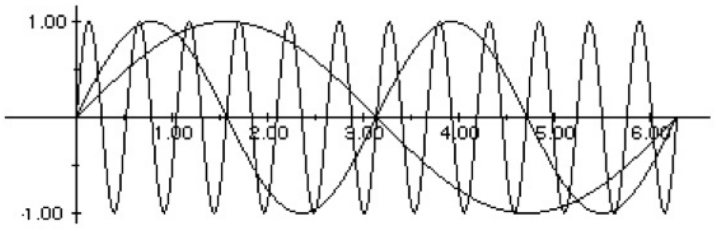

FIGURE 5. Graphs of $f_1(x) = \sin x$, $f_2(x) =\sim 2x$ and $f_{12}(x) = \sin 12x$, as in Remark 15.7.2.

It seems clear that there is no convergent subsequence. More precisely, one can show that for any $m \neq n$ there exists $x \in [0, 2\pi]$ such that $\sin mx > 1/2$, $\sin nx < -1/2$, and so $d_u(f_n, f_m) > 1$ (*exercise*). Thus there is no uniformly convergent subsequence as the distance (in the sup metric) between any two members of the sequence is at least 1.

If instead we consider the sequence of functions (g_n) defined by

$$g_n(x) = \frac{1}{n} \sin nx \quad x \in [0, 2\pi],$$

then the absolute value of the derivatives, and hence the Lipschitz constants, are uniformly bounded by 1. In this case the entire sequence converges uniformly to the zero function, as is easy to see.

Proof of Theorem We need to prove that $\mathcal{F}$ is complete and totally bounded.

(1) *Completeness of* $\mathcal{F}$.

We know that $\mathcal{C}(X; \mathbb{R}^n)$ is complete from Corollary 12.3.4. Since $\mathcal{F}$ is a closed subset, it follows $\mathcal{F}$ is complete as remarked in the generalisation following Theorem 8.2.2.

(2) *Total boundedness of* $\mathcal{F}$.

Suppose $\delta > 0$.

We need to find a *finite* set S of functions in $\mathcal{C}(X;\mathbb{R}^n)$ such that for any $f \in \mathcal{F}$,

$$\text{there exists some } g \in S \text{ satisfying } \max|f-g| < \delta. \tag{167}$$

From boundedness of $\mathcal{F}$ there is a finite K such that $|f(x)| \le K$ for all $x \in X$ and $f \in \mathcal{F}$.

By uniform equicontinuity choose $\delta_1 > 0$ so that

$$d(u,v) < \delta_1 \Rightarrow |f(u) - f(v)| < \delta/4$$

for all $u, v \in X$ and all $f \in \mathcal{F}$.

Next by total boundedness of X choose a finite set of points $x_1, \dots, x_p \in X$ such that for any $x \in X$ there exists at least one x_i for which

$$d(x, x_i) < \delta_1$$

and hence

$$|f(x) - f(x_i)| < \delta/4. \tag{168}$$

Also choose a finite set of points $y_1, \dots, y_q \in \mathbb{R}^n$ so that if $y \in \mathbb{R}^n$ and $|y| \le K$ then there exists at least one y_j for which

$$|y - y_j| < \delta/4. \tag{169}$$

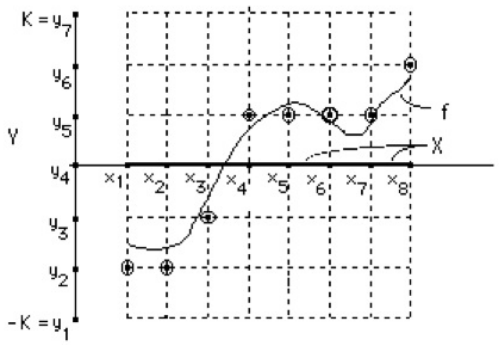

FIGURE 6. The graph of the discrete function α is indicated by the $\odot$. See (170).

Consider the set of all functions α where

$$\alpha : \{x_1, \dots, x_p\} \to \{y_1, \dots, y_q\}. \tag{170}$$

Thus α is a function assigning to each of $x_1, \dots, x_p$ one of the values $y_1, \dots, y_q$. There are only a finite number (in fact q^p) possible such α. For each α, *if* there exists a function $f \in \mathcal{F}$ satisfying

$$|f(x_i) - \alpha(x_i)| < \delta/4 \quad \text{for } i = 1, \dots, p,$$

then choose *one* such f and label it g_α. Let S be the set of all g_α. Thus

$$|g_\alpha(x_i) - \alpha(x_i)| < \delta/4 \quad \text{for } i = 1, \dots, p. \tag{171}$$

Note that S is a finite set (with at most q^p members).

Now consider any $f \in \mathcal{F}$. For each $i = 1, \dots, p$, by (169) choose one of the y_j so that

$$|f(x_i) - y_j| < \delta/4.$$

Let α be the function that assigns to each x_i this corresponding y_j. Thus

$$|f(x_i) - \alpha(x_i)| < \delta/4 \tag{172}$$

for $i = 1, \ldots, p$. Note that this implies the function g_α defined previously *does* exist.

We aim to show $d_u(f, g_\alpha) < \delta$.

Consider any $x \in X$. By (168) choose x_i so

$$d(x, x_i) < \delta_1. \tag{173}$$

Then

$$\begin{aligned} |f(x) - g_\alpha(x)| &\leq |f(x) - f(x_i)| + |f(x_i) - \alpha(x_i)| \\ &\quad + |\alpha(x_i) - g_\alpha(x_i)| + + |g_\alpha(x_i) - g_\alpha(x)| \\ &\leq 4 \times \frac{\delta}{4} \quad \text{from (168), (172), (173) and (171)} \\ &< \delta. \end{aligned}$$

This establishes (167) since x was an arbitrary member of X. Thus $\mathcal{F}$ is totally bounded. ∎

15.8. Peano's Existence Theorem

In this Section we consider the initial value problem

$$\begin{aligned} x'(t) &= f(t, x(t)), \\ x(t_0) &= x_0. \end{aligned}$$

For simplicity of notation we consider the case of a single equation, but everything generalises easily to a system of equations.

Suppose f is continuous, and locally Lipschitz with respect to the first variable, in some open set $U \subset \mathbb{R} \times \mathbb{R}$, where $(t_0, x_0) \in U$. Then we saw in the chapter on Differential Equations that there is a unique solution in some interval $[t_0 - h, t_0 + h]$ (and the solution is C^1).

If f is continuous, but not locally Lipschitz with respect to the first variable, then there need no longer be a unique solution, as we saw in Example 1 of the Differential Equations Chapter. The Example was

$$\begin{aligned} \frac{dx}{dt} &= \sqrt{|x|}, \\ x(0) &= 0. \end{aligned}$$

It turns out that this example is typical. Provided we at least assume that f is continuous, there *will* always be a solution. However, it may not be unique. Such examples are physically reasonable. In the example, $f(x) = \sqrt{x}$ may be determined by a one dimensional material whose properties do not change smoothly from the region $x < 0$ to the region $x > 0$.

We will prove the following Theorem.

THEOREM 15.8.1 (Peano). *Assume that f is continuous in the open set $U \subset \mathbb{R} \times \mathbb{R}$. Let $(t_0, x_0) \in U$. Then there exists $h > 0$ such that the initial value problem*

$$\begin{aligned} x'(t) &= f(t, x(t)), \\ x(t_0) &= x_0, \end{aligned}$$

has a C^1 solution for $t \in [t_0 - h, t_0 + h]$.

We saw in Theorem 13.8.1 that every (C^1) solution of the initial value problem is a solution of the integral equation

$$x(t) = x_0 + \int_{t_0}^{t} f\Big(s, x(s)\Big)\, ds$$

(obtained by integrating the differential equation). And conversely, we saw that every C^0 solution of the integral equation is a solution of the initial value problem (and the solution must in fact be $\mathcal{C}^1$). This Theorem only used the *continuity* of f.

Thus Theorem 15.8.1 follows from the following Theorem.

THEOREM 15.8.2. *Assume f is continuous in the open set $U \subset \mathbb{R} \times \mathbb{R}$. Let $(t_0, x_0) \in U$. Then there exists $h > 0$ such that the integral equation*

$$x(t) = x_0 + \int_{t_0}^{t} f\Big(s, x(s)\Big)\, ds \tag{174}$$

has a C^0 solution for $t \in [t_0 - h, t_0 + h]$.

Remark We saw in Theorem 13.9.1 that the integral equation does indeed have a solution, assuming f is also locally Lipschitz in x. The proof used the Contraction Mapping Principle. But if we merely assume continuity of f, then that proof no longer applies (if it did, it would also give uniqueness, which we have just remarked is not the case).

In the following proof, we show that some subsequence of the sequence of Euler polygons, first constructed in Section 13.4, converges to a solution of the integral equation.

Proof of Theorem

(1) Choose $h, k > 0$ so that

$$A_{h,k}(t_0, \mathbf{x}_0) := \{(t, \mathbf{x}) : |t - t_0| \le h,\ |\mathbf{x} - \mathbf{x}_0| \le k\} \subset U.$$

Since f is continuous, it is bounded on the *compact* set $A_{h,k}(t_0, \mathbf{x}_0)$. Choose M such that

$$|f(t, x)| \le M \quad \text{if } (t, x) \in A_{h,k}(t_0, \mathbf{x}_0). \tag{175}$$

By decreasing h if necessary, we will require

$$h \le \frac{k}{M}. \tag{176}$$

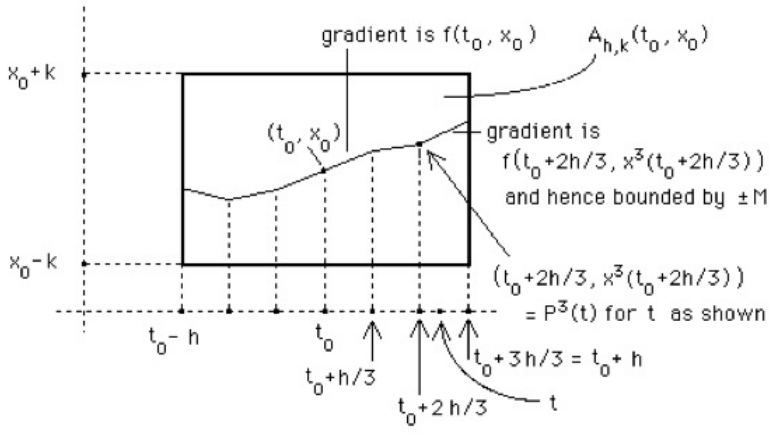

FIGURE 7. Graph of $x_3(t)$ as defined in Step (2) of the proof of Theorem 15.8.2.

(2) (See the diagram for $n = 3$) For each integer $n \geq 1$, let $x^n(t)$ be the *piecewise linear* function, defined as in Section 13.4, but with step-size h/n. More precisely, if $t \in [t_0, t_0 + h]$,

$$\begin{aligned} x^n(t) &= x_0 + (t - t_0)\, f(t_0, x_0) \\ &\qquad \text{for } t \in \left[t_0, t_0 + \frac{h}{n}\right] \\ x^n(t) &= x^n\Big(t_0 + \frac{h}{n}\Big) + \Big(t - \Big(t_0 + \frac{h}{n}\Big)\Big)\, f\Big(t_0 + \frac{h}{n},\, x^n\Big(t_0 + \frac{h}{n}\Big)\Big) \\ &\qquad \text{for } t \in \left[t_0 + \frac{h}{n}, t_0 + 2\frac{h}{n}\right] \\ x^n(t) &= x^n\Big(t_0 + 2\frac{h}{n}\Big) + \Big(t - \Big(t_0 + 2\frac{h}{n}\Big)\Big)\, f\Big(t_0 + 2\frac{h}{n},\, x^n\Big(t_0 + 2\frac{h}{n}\Big)\Big) \\ &\qquad \text{for } t \in \left[t_0 + 2\frac{h}{n}, t_0 + 3\frac{h}{n}\right] \\ &\vdots \end{aligned}$$

Similarly for $t \in [t_0 - h, t_0]$.

(3) From (175) and (176), and as is clear from the diagram, $|\frac{d}{dt}x^n(t)| \leq M$ (except at the points $t_0, t_0 \pm \frac{h}{n}, t_0 \pm 2\frac{h}{n}, \ldots$). It follows (*exercise*) that x^n is Lipschitz on $[t_0 - h, t_0 + h]$ with Lipschitz constant at most M.

In particular, since $k \geq Mh$, the graph of $t \mapsto x^n(t)$ remains in the closed rectangle $A_{h,k}(t_0, \mathbf{x}_0)$ for $t \in [t_0 - h, t_0 + h]$.

(4) From (3), the functions x^n belong to the family $\mathcal{F}$ of Lipschitz functions

$$f : [t_0 - h, t_0 + h] \to \mathbb{R}$$

such that

$$\mathrm{Lip} f \leq M$$

and

$$|f(t) - x_0| \leq k \quad \text{for all } t \in [t_0 - h, t_0 + h].$$

But $\mathcal{F}$ *is closed, uniformly bounded, and uniformly equicontinuous*, by the same argument as used in Example 1 of Section 15.7. It follows from the Arzela-Ascoli Theorem that *some subsequence* $(x^{n'})$ *of* (x^n) *converges uniformly to a function* $x \in \mathcal{F}$.

Our aim now is to show that x *is a solution of (174).*

(5) For each point $(t, x^n(t))$ on the graph of x^n, let $P^n(t) \in \mathbb{R}^2$ be the coordinates of the point at the left (right) endpoint of the corresponding line segment if $t \geq 0$ ($t \leq 0$). More precisely

$$P^n(t) = \left(t_0 + (i-1)\frac{h}{n},\; x^n\Big(t_0 + (i-1)\frac{h}{n}\Big)\right) \quad \text{if } t \in \left[t_0 + (i-1)\frac{h}{n}, t_0 + i\frac{h}{n}\right]$$

for $i = 1, \ldots, n$. A similar formula is true for $t \leq 0$.

Notice that $P^n(t)$ is constant for $t \in \big[t_0 + (i-1)\frac{h}{n}, t_0 + i\frac{h}{n}\big)$, (and in particular $P^n(t)$ is of course *not* continuous in $[t_0 - h, t_0 + h]$).

(6) Without loss of generality, suppose $t \in \left[t_0 + (i-1)\frac{h}{n}, t_0 + i\frac{h}{n}\right]$. Then from (5) and (3)

$$\begin{aligned} |P^n(t) - (t, x^n(t)| &\leq \sqrt{\left(t - \left(t_0 + i\frac{h}{n}\right)\right)^2 + \left(x^n(t) - x^n\left(t_0 + i\frac{h}{n}\right)\right)^2} \\ &\leq \sqrt{\left(\frac{h}{n}\right)^2 + \left(M\frac{h}{n}\right)^2} \\ &= \sqrt{1+M^2}\,\frac{h}{n}. \end{aligned}$$

Thus $|P^n(t) - (t, x^n(t))| \to 0$, uniformly in t, as $n \to \infty$.

(7) It follows from the definitions of x^n and P^n, and is clear from the diagram, that

$$x^n(t) = x_0 + \int_{t_0}^{t} f(P^n(s))\, ds \tag{177}$$

for $t \in [t_0 - h, t_0 + h]$. Although $P^n(s)$ is not continuous in s, the previous integral still exists (for example, we could define it by considering the integral over the various segments on which $P^n(s)$ and hence $f(P^n(s))$ is constant).

(8) Our intention now is to show (on passing to a suitable subsequence) that $x^n(t) \to x(t)$ uniformly, $P^n(t) \to (t, x(t))$ uniformly, and to use this and (177) to deduce (174).

(9) Since f is continuous on the compact set $A_{h,k}(t_0, \mathbf{x}_0)$, it is uniformly continuous there by (8) of Section 15.4.

Suppose $\epsilon > 0$. By uniform continuity of f choose $\delta > 0$ so that for any two points $P, Q \in A_{h,k}(t_0, \mathbf{x}_0)$, if $|P - Q| < \delta$ then $|f(P) - f(Q)| < \epsilon$.

In order to obtain (174) from (177), we compute

$$|f(s, x(s)) - f(P^n(s))| \leq |f(s, x(s)) - f(s, x^n(s))| + |f(s, x^n(s)) - f(P^n(s))|.$$

From (6), $|(s, x^n(s)) - P^n(s)| < \delta$ for all $n \geq N_1$ (say), independently of s. From uniform convergence (4), $|x(s) - x^{n'}(s)| < \delta$ for all $n' \geq N_2$ (say), independently of s. By the choice of δ it follows

$$|f(s, x(s)) - f(P^{n'}(s))| < 2\epsilon, \tag{178}$$

for all $n' \geq N := \max\{N_1, N_2\}$.

(10) From (4), the left side of (177) converges to the left side of (174), for the subsequence $(x^{n'})$.

From (178), the difference of the right sides of (177) and (174) is bounded by $2\epsilon h$ for members of the subsequence $(x^{n'})$ such that $n' \geq N(\epsilon)$. As ϵ is arbitrary, it follows that for this subsequence, the right side of (177) converges to the right side of (174).

This establishes (174), and hence the Theorem. ∎

CHAPTER 16

Connectedness

16.1. Introduction

One intuitive idea of what it means for a set S to be "connected" is that S *cannot* be written as the union of two sets which do not "touch" one another. We make this precise in Definition 16.2.1.

Another informal idea is that any two points in S can be connected by a "path" which joins the points and which lies entirely in S. We make this precise in definition 16.4.2.

These two notions are distinct, though they agree on open subsets of $\mathbb{R}^n$, see Theorem 16.4.4 below.

16.2. Connected Sets

DEFINITION 16.2.1. A metric space (X, d) is *connected* if there do *not* exist two non-empty disjoint *open* sets U and V such that $X = U \cup V$.

The metric space is *disconnected* if it is not connected, i.e. if there exist two non-empty disjoint open sets U and V such that $X = U \cup V$.

A set $S \subset X$ is connected (disconnected) if the metric subspace (S, d) is connected (disconnected).

Remarks and Examples

(1) In the following diagram, S is disconnected. On the other hand, T *is* connected; in particular although $T = T_1 \cup T_2$, any open set containing T_1 will have non-empty intersection with T_2. However, T is not pathwise connected — see Example 3 in Section 16.4.

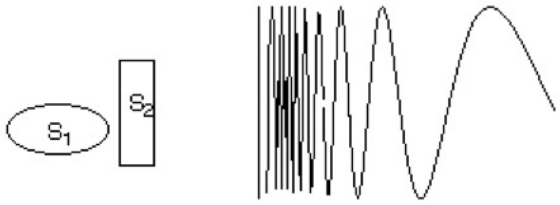

FIGURE 1. On the left side, $S = S - 1 \cup S_2$. On the right side $T = T_1 \cup T_2$, where T_1 is the interval $[-1, 1]$ on the y-axis and T_2 is the graph of $y = \sin(1/x)$ for $0 < x \leq 6\pi$. The set S is neither connected nor pathwise connected. The set T *is* connected, but it is not pathwise connected.

(2) The sets U and V in the previous definition are required to be open *in* X. For example, let

$$A = [0, 1] \cup (2, 3].$$

We claim that A is *disconnected*.

Let $U = [0, 1]$ and $V = (2, 3]$. Then both these sets are *open in* the metric subspace (A, d) (where d is the standard metric induced from $\mathbb{R}$). To see this, note that both U and V are the intersection of A with sets which are open in $\mathbb{R}$ (see Theorem 6.5.3). It follows from the definition that X is disconnected.

(3) In the definition, the sets U and V cannot be *arbitrary* disjoint sets. For example, we will see in Theorem 16.3.2 that $\mathbb{R}$ is connected. But $\mathbb{R} = U \cup V$ where U and V are the disjoint sets $(-\infty, 0]$ and $(0, \infty)$ respectively.

(4) $\mathbb{Q}$ is disconnected. To see this write

$$\mathbb{Q} = \Big(\mathbb{Q} \cap (-\infty, \sqrt{2})\Big) \cup \Big(\mathbb{Q} \cap (\sqrt{2}, \infty)\Big).$$

The following proposition gives two other definitions of connectedness.

PROPOSITION 16.2.2. *A metric space (X, d) is connected*

(1) *iff there do not exist two non-empty disjoint* **closed** *sets U and V such that $X = U \cup V$;*
(2) *iff the only non-empty subset of X which is both open and closed*[1] *is X itself.*

PROOF. (1) Suppose $X = U \cup V$ where $U \cap V = \emptyset$. Then $U = X \setminus V$ and $V = X \setminus U$. Thus U and V are both open iff they are both closed[2]. The first equivalence follows.

(2) In order to show the second condition is also equivalent to connectedness, first suppose that X is not connected and let U and V be the open sets given by Definition 16.2.1. Then $U = X \setminus V$ and so U is also closed. Since $U \neq \emptyset, X$, (2) in the statement of the theorem is not true.

Conversely, if (2) in the statement of the theorem is not true let $E \subset X$ be both open and closed and $E \neq \emptyset, X$. Let $U = E$, $V = X \setminus E$. Then U and V are non-empty disjoint open sets whose union is X, and so X is not connected. □

Example We saw before that if $A = [0, 1] \cup (2, 3]$ $(\subset \mathbb{R})$, then A is not connected. The sets $[0, 1]$ and $(2, 3]$ are both open and both closed in A.

[1] Such a set is called *clopen*.
[2] Of course, we mean open, or closed, in X.

16.3. Connectedness in $\mathbb{R}^n$

Not surprisingly, the connected sets in $\mathbb{R}$ are precisely the intervals in $\mathbb{R}$.
We first need a precise definition of *interval*.

DEFINITION 16.3.1. A set $S \subset \mathbb{R}$ is an *interval* if

$$a, b \in S \text{ and } a < x < b \Rightarrow x \in S.$$

THEOREM 16.3.2. *$S \subset \mathbb{R}$ is connected iff S is an interval.*

PROOF. (a) Suppose S is *not* an interval. Then there exist $a, b \in S$ and there exists $x \in (a, b)$ such that $x \notin S$.
Then

$$S = \Big(S \cap (-\infty, x)\Big) \cup \Big(S \cap (x, \infty)\Big).$$

Both sets on the right side are open *in* S, are disjoint, and are non-empty (the first contains a, the second contains b). Hence S is not connected.

(b) Suppose S is an interval.
Assume that S is not connected. Then there exist nonempty sets U and V which are open in S such that

$$S = U \cup V,\ U \cap V = \emptyset.$$

Choose $a \in U$ and $b \in V$. Without loss of generality we may assume $a < b$. Since S is an interval, $[a, b] \subset S$.
Let

$$c = \sup([a, b] \cap U).$$

Since $c \in [a, b] \subset S$ it follows $c \in S$, and so either $c \in U$ or $c \in V$.
Suppose $c \in U$. Then $c \neq b$ and so $a \leq c < b$. Since $c \in U$ and U is open, there exists $c' \in (c, b)$ such that $c' \in U$. This contradicts the definition of c as $\sup([a, b] \cap U)$.
Suppose $c \in V$. Then $c \neq a$ and so $a < c \leq b$. Since $c \in V$ and V is open, there exists $c'' \in (a, c)$ such that $[c'', c] \subset V$. But this implies that c is again not the *sup*. Thus we again have a contradiction.
Hence S *is* connected. □

Remark There is no such simple chararacterisation in $\mathbb{R}^n$ for $n > 1$.

16.4. Path Connected Sets

DEFINITION 16.4.1. A *path* connecting two points x and y in a metric space (X, d) is a continuous function $f : [0, 1]\ (\subset \mathbb{R}) \to X$ such that $f(0) = x$ and $f(1) = y$.

DEFINITION 16.4.2. A metric space (X, d) is *path connected* if any two points in X can be connected by a path in X.
A set $S \subset X$ is path connected if the metric subspace (S, d) is path connected.

The notion of *path connected* may seem more intuitive than that of *connected*. However, the latter is usually mathematically easier to work with.
Every path connected set is connected (Theorem 16.4.3). A connected set need not be path connected (Example (3) below), but for open subsets of $\mathbb{R}^n$ (an important case) the two notions of connectedness *are* equivalent (Theorem 16.4.4).

THEOREM 16.4.3. *If a metric space (X, d) is path connected then it is connected.*

PROOF. Assume X is *not* connected[3].
Thus there exist non-empty disjoint open sets U and V such that $X = U \cup V$.

[3]We want to show that X is not path connected.

Choose $x \in U$, $y \in V$ and suppose there is a path from x to y, i.e. suppose there is a continuous function $f : [0,1] (\subset \mathbb{R}) \to X$ such that $f(0) = x$ and $f(1) = y$.

Consider $f^{-1}[U], f^{-1}[V] \subset [0,1]$. They are open (continuous inverse images of open sets), disjoint (since U and V are), non-empty (since $0 \in f^{-1}[U]$, $1 \in f^{-1}[V]$), and $[0,1] = f^{-1}[U] \cup f^{-1}[V]$ (since $X = U \cup V$). But this contradicts the connectedness of $[0,1]$. Hence there is no such path and so X is not path connected. $\square$

Examples

(1) $B_r(\mathbf{x}) \subset \mathbb{R}^2$ is path connected and hence connected. Since for $\mathbf{u}, \mathbf{v} \in B_r(\mathbf{x})$ the path $f : [0,1] \to \mathbb{R}^2$ given by $f(t) = (1-t)\mathbf{u} + t\mathbf{v}$ is a path in $\mathbb{R}^2$ connecting $\mathbf{u}$ and $\mathbf{v}$. The fact that the path does lie in $\mathbb{R}^2$ is clear, and can be checked from the triangle inequality (*exercise*).

The same argument shows that in any *normed space* the open balls $B_r(x)$ are path connected, and hence connected. The *closed balls* $\{y : d(y,x) \leq r\}$ are similarly path connected and hence connected.

(2) $A = \mathbb{R}^2 \backslash \{(0,0), (1,0), (\frac{1}{2},0)\, (\frac{1}{3},0), \ldots, (\frac{1}{n},0), \ldots\}$ is path connected (take a semicircle joining points in A) and hence connected.

(3) Let

$$A = \{(x,y) : x > 0 \text{ and } y = \sin \tfrac{1}{x}, \text{ or } x = 0 \text{ and } y \in [0,1]\}.$$

Then A is connected but not path connected (**exercise*).

THEOREM 16.4.4. *Let $U \subset \mathbb{R}^n$ be an* **open** *set. Then U is connected iff it is path connected.*

PROOF. From Theorem 16.4.3 it is sufficient to prove that if U is connected then it is path connected.

Assume then that U is connected.

The result is trivial if $U = \emptyset$ (*why?*). So assume $U \neq \emptyset$ and choose some $\mathbf{a} \in \mathbf{U}$. Let

$$E = \{\mathbf{x} \in U : \text{there is a path in } U \text{ from } \mathbf{a} \text{ to } \mathbf{x}\}.$$

We want to show $E = U$. Clearly $E \neq \emptyset$ since $\mathbf{a} \in E$[4]. If we can show that E is both open and closed, it will follow from Proposition 16.2.2(2) that $E = U$[5].

To show that E is *open*, suppose $\mathbf{x} \in E$ and choose $r > 0$ such that $B_r(\mathbf{x}) \subset U$. From the previous Example(1), for each $\mathbf{y} \in B_r(\mathbf{x})$ there is a path in $B_r(\mathbf{x})$ from $\mathbf{x}$ to $\mathbf{y}$. If we "join" this to the path from $\mathbf{a}$ to $\mathbf{x}$, it is not difficult to obtain a path from $\mathbf{a}$ to $\mathbf{y}$[6]. Thus $\mathbf{y} \in E$ and so E is open.

To show that E is *closed in* U, suppose $(\mathbf{x}_n)_{n=1}^{\infty} \subset E$ and $\mathbf{x}_n \to \mathbf{x} \in U$. We want to show $\mathbf{x} \in E$. Choose $r > 0$ so $B_r(\mathbf{x}) \subset U$. Choose n so $\mathbf{x}_n \in B_r(\mathbf{x})$. There is a path in U joining $\mathbf{a}$ to $\mathbf{x}_n$ (since $\mathbf{x}_n \in E$) and a path joining $\mathbf{x}_n$ to $\mathbf{x}$ (as $B_r(\mathbf{x})$ is path connected). As before, it follows there is a path in U from $\mathbf{a}$ to $\mathbf{x}$. Hence $\mathbf{x} \in E$ and so E is closed.

Since E is open and closed, it follows as remarked before that $E = U$, and so we are done. □

16.5. Basic Results

THEOREM 16.5.1. *The continuous image of a connected set is connected.*

---

[4] A path joining $\mathbf{a}$ to $\mathbf{a}$ is given by $f(t) = \mathbf{a}$ for $t \in [0,1]$.

[5] This is a very important technique for showing that every point in a connected set has a given property.

[6] Suppose $f : [0,1] \to U$ is continuous with $f(0) = \mathbf{a}$ and $f(1) = \mathbf{x}$, while $g : [0,1] \to U$ is continuous with $g(0) = \mathbf{x}$ and $g(1) = \mathbf{y}$. Define

$$h(t) = \begin{cases} f(2t) & 0 \leq t \leq 1/2 \\ g(2t-1) & 1/2 \leq t \leq 1 \end{cases}$$

Then it is easy to see that h is a continuous path in U from $\mathbf{a}$ to $\mathbf{y}$ (the main point is to check what happens at $t = 1/2$).

PROOF. Let $f : X \to Y$, where X is connected.

Suppose $f[X]$ is *not* connected (we intend to obtain a contradiction).

Then there exists $E \subset f[X]$, $E \neq \emptyset, f[X]$, and E both open and closed in $f[X]$. It follows there exists an open $E' \subset Y$ and a closed $E'' \subset Y$ such that

$$E = f[X] \cap E' = f[X] \cap E'.$$

In particular,

$$f^{-1}[E] = f^{-1}[E'] = f^{-1}[E''],$$

and so $f^{-1}[E]$ is both open and closed in X. Since $E \neq \emptyset, f[X]$ it follows that $f^{-1}[E] \neq \emptyset, X$. Hence X is not connected, contradiction.

Thus $f[X]$ *is* connected. □

The next result generalises the usual Intermediate Value Theorem.

COROLLARY 16.5.2. *Suppose $f : X \to \mathbb{R}$ is continuous, X is connected, and f takes the values a and b where $a < b$. Then f takes all values between a and b.*

PROOF. By the previous theorem, $f[X]$ is a connected subset of $\mathbb{R}$. Then, by Theorem 16.3.2, $f[X]$ is an interval. Since $a, b \in f[X]$ it then follows $c \in f[X]$ for any $c \in [a, b]$. □

CHAPTER 17

Differentiation of Real-Valued Functions

17.1. Introduction

In this Chapter we discuss the notion of *derivative* (i.e. *differential*) for functions $f : D\,(\subset \mathbb{R}^n) \to \mathbb{R}$. In the next chapter we consider the case for functions $f : D\,(\subset \mathbb{R}^n) \to \mathbb{R}^n$.

We can represent such a function ($m = 1$) by drawing its graph, as is done in the first diagrams in Section 10.1 in case $n = 1$ or $n = 2$, or as is done "schematically" in the second last diagram in Section 10.1 for arbitrary n. In case $n = 2$ (or perhaps $n = 3$) we can draw the level sets, as is done in Section 17.6.

Convention Unless stated otherwise, we will always consider functions $f : D(\subset \mathbb{R}^n) \to \mathbb{R}$ where the domain D is *open*. This implies that for any $\mathbf{x} \in D$ there exists $r > 0$ such that $B_r(\mathbf{x}) \subset D$. See Figure 1.

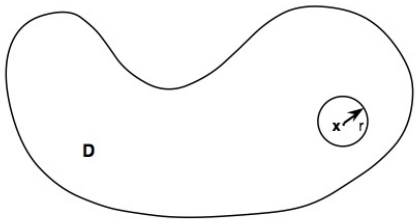

FIGURE 1. A domain $D \subset \mathbb{R}^2$.

Most of the following applies to more general domains D by taking one-sided, or otherwise restricted, limits. No essentially new ideas are involved.

17.2. Algebraic Preliminaries

The *inner product* in $\mathbb{R}^n$ is represented by

$$\mathbf{y} \cdot \mathbf{x} = y^1 x^1 + \ldots + y^n x^n$$

where $\mathbf{y} = (y^1, \ldots, y^n)$ and $\mathbf{x} = (x^1, \ldots, x^n)$.

For each *fixed* $\mathbf{y} \in \mathbb{R}^n$ the inner product enables us to define a *linear function*

$$L_{\mathbf{y}} = L : \mathbb{R}^n \to \mathbb{R}$$

given by

$$L(\mathbf{x}) = \mathbf{y} \cdot \mathbf{x}.$$

Conversely, we have the following.

PROPOSITION 17.2.1. *For any linear function*

$$L:\mathbb{R}^n \to \mathbb{R}$$

there exists a **unique** $\mathbf{y} \in \mathbb{R}^n$ *such that*

$$L(\mathbf{x}) = \mathbf{y}\cdot\mathbf{x} \quad \forall \mathbf{x} \in \mathbb{R}^n. \tag{179}$$

The components of $\mathbf{y}$ *are given by* $y^i = L(\mathbf{e_i})$.

PROOF. Suppose $L:\mathbb{R}^n \to \mathbb{R}$ is linear. *Define* $\mathbf{y} = (y^1, \ldots, y^n)$ by

$$y^i = L(\mathbf{e_i}) \quad i = 1, \ldots, n.$$

Then

$$\begin{aligned} L(\mathbf{x}) &= L(x^1\mathbf{e_1} + \cdots + x^n\mathbf{e_n}) \\ &= x^1 L(\mathbf{e_1}) + \cdots + x^n L(\mathbf{e_n}) \\ &= x^1 y^1 + \cdots + x^n y^n \\ &= \mathbf{y}\cdot\mathbf{x}. \end{aligned}$$

This proves the *existence* of $\mathbf{y}$ satisfying (179).

The *uniqueness* of $\mathbf{y}$ follows from the fact that if (179) is true for some $\mathbf{y}$, then on choosing $\mathbf{x} = \mathbf{e_i}$ it follows we *must* have

$$L(\mathbf{e_i}) = y^i \quad i = 1, \ldots, n.$$

□

Note that if L is the *zero operator*, i.e. if $L(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathbb{R}^n$, then the vector $\mathbf{y}$ corresponding to L is the zero vector.

17.3. Partial Derivatives

DEFINITION 17.3.1. The ith *partial derivative* of f at $\mathbf{x}$ is defined by

$$\begin{aligned} \frac{\partial f}{\partial x^i}(\mathbf{x}) &= \lim_{t\to 0} \frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x})}{t} \\ &= \lim_{t\to 0} \frac{f(x^1, \ldots, x^i + t, \ldots, x^n) - f(x^1, \ldots, x^i, \ldots, x^n)}{t}, \end{aligned} \tag{180}$$

provided the limit exists. The notation $\Delta_i f(\mathbf{x})$ is also used.

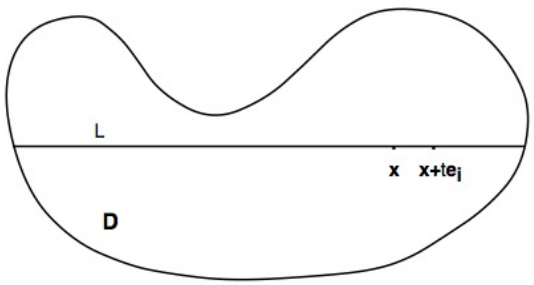

FIGURE 2. Diagram for Definition 17.3.1 and the discussion which follows it.

Thus $\dfrac{\partial f}{\partial x^i}(\mathbf{x})$ is just the usual derivative at $t = 0$ of the *real-valued* function g defined by $g(t) = f(x^1, \ldots, x^i + t, \ldots, x^n)$. Think of g as being defined along the line L in Figure 2, with $t = 0$ corresponding to the point $\mathbf{x}$.

17.4. Directional Derivatives

DEFINITION 17.4.1. The *directional derivative* of f at $\mathbf{x}$ in the direction $\mathbf{v} \neq \mathbf{0}$ is defined by

$$D_{\mathbf{v}} f(\mathbf{x}) = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t}, \tag{181}$$

provided the limit exists.

See Figure 3. It follows immediately from the definitions that

$$\frac{\partial f}{\partial x^i}(\mathbf{x}) = D_{\mathbf{e}_i} f(\mathbf{x}). \tag{182}$$

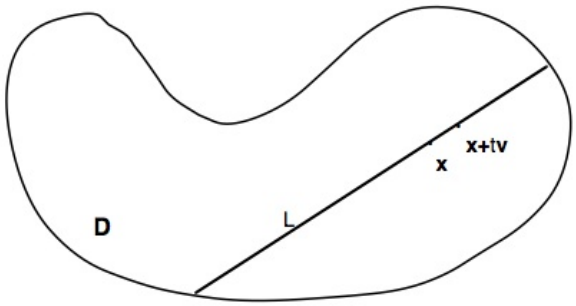

FIGURE 3. Diagram for Definition 17.4.1 and the discussion which follows it.

Note that $D_{\mathbf{v}} f(\mathbf{x})$ is just the usual derivative at $t = 0$ of the *real-valued* function g defined by $g(t) = f(\mathbf{x} + t\mathbf{v})$. As before, think of the function g as being defined along the line L in the previous diagram.

Thus we interpret $D_{\mathbf{v}} f(\mathbf{x})$ as the rate of change of f at $\mathbf{x}$ in the direction $\mathbf{v}$; at least in the case $\mathbf{v}$ is a unit vector.

Exercise: Show that $D_{\alpha\mathbf{v}} f(\mathbf{x}) = \alpha D_{\mathbf{v}} f(\mathbf{x})$ for any real number α.

17.5. The Differential (or Derivative)

Motivation Suppose $f : I\,(\subset \mathbb{R}) \to \mathbb{R}$ is differentiable at $a \in I$. Then $f'(a)$ can be used to define the *best linear approximation* to $f(x)$ for x near a. Namely:

$$f(x) \approx f(a) + f'(a)(x - a). \tag{183}$$

Note that the right-hand side of (183) is linear in x. (More precisely, the right side is a polynomial in x of degree one.)

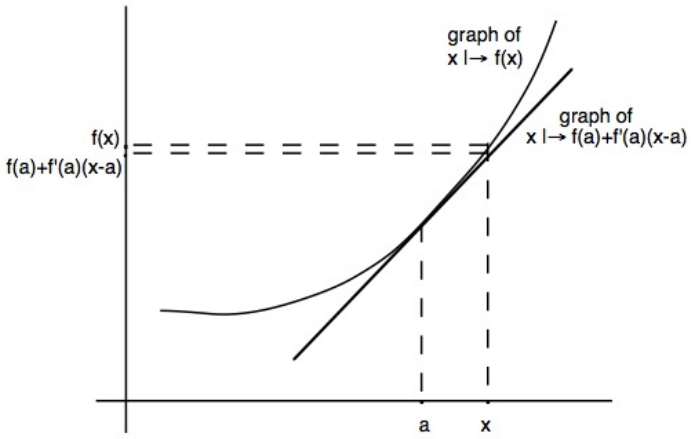

FIGURE 4. Graph of the best linear approximation to f near a.

The *error*, or difference between the two sides of (183), approaches zero as $x \to a$, *faster* than $|x-a| \to 0$. More precisely

$$\begin{aligned} \frac{\Big|f(x) - \Big(f(a) + f'(a)(x-a)\Big)\Big|}{|x-a|} &= \left|\frac{f(x) - \Big(f(a) + f'(a)(x-a)\Big)}{x-a}\right| \\ &= \left|\frac{f(x)-f(a)}{x-a} - f'(a)\right| \\ &\to 0 \quad \text{as } x \to a. \end{aligned} \tag{184}$$

We make this the basis for the next definition in the case $n > 1$.

DEFINITION 17.5.1. Suppose $f : D\,(\subset \mathbb{R}^n) \to \mathbb{R}$. Then f is *differentiable* at $\mathbf{a} \in D$ if there is a linear function $L : \mathbb{R}^n \to \mathbb{R}$ such that

$$\frac{\Big|f(\mathbf{x}) - \Big(f(\mathbf{a}) + L(\mathbf{x}-\mathbf{a})\Big)\Big|}{|\mathbf{x}-\mathbf{a}|} \to 0 \quad \text{as } \mathbf{x} \to \mathbf{a}. \tag{185}$$

The linear function L is denoted by $f'(\mathbf{a})$ or $df(\mathbf{a})$ and is called the *derivative* or *differential* of f at $\mathbf{a}$. (We will see in Proposition 17.5.2 that if L exists, it is uniquely determined by this definition.)

The idea is that the graph of $\mathbf{x} \mapsto f(\mathbf{a}) + L(\mathbf{x}-\mathbf{a})$ is "tangent" to the graph of $f(\mathbf{x})$ at the point $\Big(\mathbf{a}, f(\mathbf{a})\Big)$.

Notation: We write $\langle df(\mathbf{a}), \mathbf{x}-\mathbf{a}\rangle$ for $L(\mathbf{x}-\mathbf{a})$, and read this as "df at $\mathbf{a}$ applied to $\mathbf{x}-\mathbf{a}$". We think of $df(\mathbf{a})$ as a linear transformation (or function) which operates on vectors $\mathbf{x}-\mathbf{a}$ whose "base" is at $\mathbf{a}$.

The next proposition gives the connection between the differential operating on a vector $\mathbf{v}$, and the directional derivative in the direction corresponding to $\mathbf{v}$. In particular, it shows that the differential is *uniquely defined* by Definition 17.5.1.

Temporarily, we let $df(\mathbf{a})$ be *any* linear map satisfying the definition for the differential of f at $\mathbf{a}$.

PROPOSITION 17.5.2. *Let $\mathbf{v} \in \mathbb{R}^n$ and suppose f is differentiable at $\mathbf{a}$.*

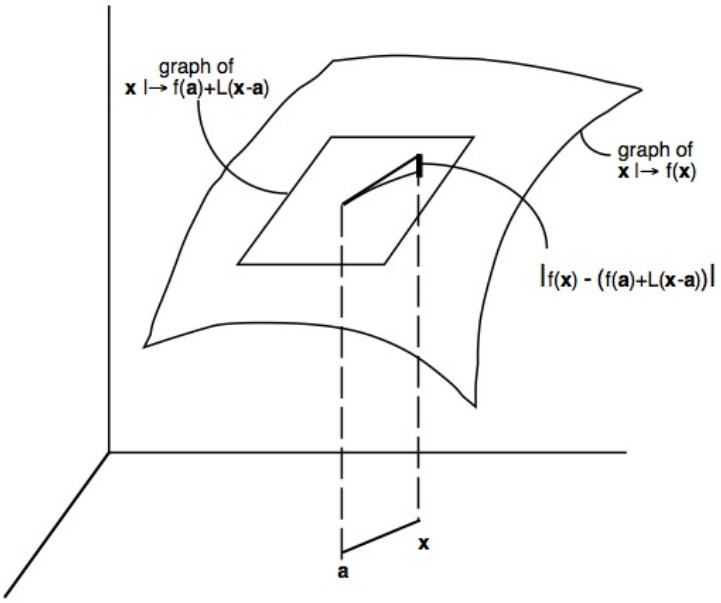

FIGURE 5. Graph of the best linear approximation to f near a, where $f : \mathbb{R}^2 \to \mathbb{R}$.

Then $D_{\mathbf{v}} f(\mathbf{a})$ *exists and*

$$\langle df(\mathbf{a}), \mathbf{v} \rangle = D_{\mathbf{v}} f(\mathbf{a}).$$

In particular, the differential is unique.

PROOF. Let $\mathbf{x} = \mathbf{a} + t\mathbf{v}$ in (185). Then

$$\lim_{t \to 0} \frac{\left| f(\mathbf{a} + t\mathbf{v}) - \Big(f(\mathbf{a}) + \langle df(\mathbf{a}), t\mathbf{v} \rangle \Big) \right|}{t} = 0.$$

Hence

$$\lim_{t \to 0} \frac{f(\mathbf{a} + t\mathbf{v}) - f(\mathbf{a})}{t} - \langle df(\mathbf{a}), \mathbf{v} \rangle = \mathbf{0}.$$

Thus

$$D_{\mathbf{v}} f(\mathbf{a}) = \langle df(\mathbf{a}), \mathbf{v} \rangle$$

as required. □

Thus $\langle df(\mathbf{a}), \mathbf{v} \rangle$ is just the directional derivative at $\mathbf{a}$ in the direction $\mathbf{v}$.

The next result shows $df(\mathbf{a})$ is the linear map given by the row vector of partial derivatives of f at $\mathbf{a}$.

COROLLARY 17.5.3. *Suppose* f *is differentiable at* $\mathbf{a}$. *Then for* any *vector* $\mathbf{v}$,

$$\langle df(\mathbf{a}), \mathbf{v} \rangle = \sum_{i=1}^{n} v^i \frac{\partial f}{\partial x^i}(\mathbf{a}).$$

PROOF.

$$\begin{aligned}\langle df(\mathbf{a}), \mathbf{v}\rangle &= \langle df(\mathbf{a}),\ v^1\mathbf{e}_1 + \cdots + v^n\mathbf{e}_n\rangle \\ &= v^1\langle df(\mathbf{a}), \mathbf{e}_1\rangle + \cdots + v^n\langle df(\mathbf{a}), \mathbf{e}_n\rangle \\ &= v^1 D_{\mathbf{e}_1} f(\mathbf{a}) + \cdots + v^n D_{\mathbf{e}_n} f(\mathbf{a}) \\ &= v^1 \frac{\partial f}{\partial x^1}(a) + \cdots + v^n \frac{\partial f}{\partial x^n}(\mathbf{a}).\end{aligned}$$

□

Example Let $f(x, y, z) = x^2 + 3xy^2 + y^3z + z$.

Then

$$\begin{aligned}\langle df(\mathbf{a}), \mathbf{v}\rangle &= v_1\frac{\partial f}{\partial x}(\mathbf{a}) + v_2\frac{\partial f}{\partial y}(\mathbf{a}) + v_3\frac{\partial f}{\partial z}(\mathbf{a}) \\ &= v_1(2a_1 + 3{a_2}^2) + v_2(6a_1a_2 + 3{a_2}^2a_3) + v_3({a_2}^3 + 1).\end{aligned}$$

Thus $df(\mathbf{a})$ is the linear map corresponding to the row vector $(2a_1 + 3{a_2}^2, 6a_1a_2 + 3{a_2}^2a_3, {a_2}^3 + 1)$.

If $\mathbf{a} = (1, 0, 1)$ then $\langle df(\mathbf{a}), \mathbf{v}\rangle = 2v_1 + v_3$. Thus $df(\mathbf{a})$ is the linear map corresponding to the row vector $(2, 0, 1)$.

If $\mathbf{a} = (1, 0, 1)$ and $\mathbf{v} = \mathbf{e}_1$ then $\langle df(1, 0, 1), \mathbf{e}_1\rangle = \dfrac{\partial f}{\partial x}(1, 0, 1) = 2$.

Rates of Convergence If a function $\psi(\mathbf{x})$ has the property that

$$\frac{|\psi(\mathbf{x})|}{|\mathbf{x} - \mathbf{a}|} \to 0 \text{ as } \mathbf{x} \to \mathbf{a},$$

then we say "$|\psi(\mathbf{x})| \to 0$ as $\mathbf{x} \to \mathbf{a}$, faster than $|\mathbf{x} - \mathbf{a}| \to 0$". We write $o(|\mathbf{x} - \mathbf{a}|)$ for $\psi(\mathbf{x})$, and read this as "little *oh* of $|\mathbf{x} - \mathbf{a}|$".

If

$$\frac{|\psi(\mathbf{x})|}{|\mathbf{x} - \mathbf{a}|} \leq M \quad \forall |\mathbf{x} - \mathbf{a}| < \epsilon,$$

for some M and some $\epsilon > 0$, i.e. if $\dfrac{|\psi(\mathbf{x})|}{|\mathbf{x} - \mathbf{a}|}$ is bounded as $\mathbf{x} \to \mathbf{a}$, then we say "$|\psi(\mathbf{x})| \to 0$ as $\mathbf{x} \to \mathbf{a}$, at least as fast as $|\mathbf{x} - \mathbf{a}| \to 0$". We write $O(|\mathbf{x} - \mathbf{a}|)$ for $\psi(\mathbf{x})$, and read this as "big *oh* of $|\mathbf{x} - \mathbf{a}|$".

For example, we can write

$$o(|x - a|) \text{ for } |x - a|^{3/2},$$

and

$$O(|x - a|) \text{ for } \sin(x - a).$$

Clearly, if $\psi(\mathbf{x})$ can be written as $o(|\mathbf{x} - \mathbf{a}|)$ then it can also be written as $O(|\mathbf{x} - \mathbf{a}|)$, but the converse may not be true as the above example shows.

The next proposition gives an equivalent definition for the differential of a function.

PROPOSITION 17.5.4. *If f is differentiable at $\mathbf{a}$ then*

$$f(\mathbf{x}) = f(\mathbf{a}) + \langle df(\mathbf{a}), \mathbf{x} - \mathbf{a}\rangle + \psi(\mathbf{x}),$$

where $\psi(\mathbf{x}) = o(|\mathbf{x} - \mathbf{a}|)$.

Conversely, suppose

$$f(\mathbf{x}) = f(\mathbf{a}) + L(\mathbf{x} - \mathbf{a}) + \psi(\mathbf{x}),$$

where $L{:}\mathbb{R}^n \to \mathbb{R}$ is linear and $\psi(\mathbf{x}) = o(|\mathbf{x} - \mathbf{a}|)$. Then f is differentiable at $\mathbf{a}$ and $df(\mathbf{a}) = L$.

PROOF. Suppose f is differentiable at $\mathbf{a}$. Let

$$\psi(\mathbf{x}) = f(\mathbf{x}) - \Big(f(\mathbf{a}) + \langle df(\mathbf{a}), \mathbf{x}-\mathbf{a}\rangle\Big).$$

Then

$$f(\mathbf{x}) = f(\mathbf{a}) + \langle df(\mathbf{a}), \mathbf{x}-\mathbf{a}\rangle + \psi(\mathbf{x}),$$

and $\psi(\mathbf{x}) = o(|\mathbf{x}-\mathbf{a}|)$ from Definition 17.5.1.

Conversely, suppose

$$f(\mathbf{x}) = f(\mathbf{a}) + L(\mathbf{x}-\mathbf{a}) + \psi(\mathbf{x}),$$

where $L:\mathbb{R}^n \to \mathbb{R}$ is linear and $\psi(\mathbf{x}) = o(|\mathbf{x}-\mathbf{a}|)$. Then

$$\frac{f(\mathbf{x}) - \Big(f(\mathbf{a}) + L(\mathbf{x}-\mathbf{a})\Big)}{|\mathbf{x}-\mathbf{a}|} = \frac{\psi(\mathbf{x})}{|\mathbf{x}-\mathbf{a}|} \to 0 \quad \text{as } \mathbf{x} \to \mathbf{a},$$

and so f is differentiable at $\mathbf{a}$ and $df(\mathbf{a}) = L$. □

Remark The word "differential" is used in [Sw] in an imprecise, and different, way from here.

Finally we have:

PROPOSITION 17.5.5. *If $f, g : D\,(\subset \mathbb{R}^n) \to \mathbb{R}$ are differentiable at $\mathbf{a} \in D$, then so are αf and $f+g$. Moreover,*

$$\begin{aligned} d(\alpha f)(\mathbf{a}) &= \alpha df(\mathbf{a}), \\ d(f+g)(\mathbf{a}) &= df(\mathbf{a}) + dg(\mathbf{a}). \end{aligned}$$

PROOF. This is straightforward (*exercise*) from Proposition 17.5.4. □

The previous proposition corresponds to the fact that the partial derivatives for $f+g$ are the sum of the partial derivatives corresponding to f and g respectively. Similarly for αf [1].

17.6. The Gradient

Strictly speaking, $df(\mathbf{a})$ is a *linear operator* on vectors in $\mathbb{R}^n$ (where, for convenience, we think of these vectors as having their "base at $\mathbf{a}$").

We saw in Section 17.2 that every linear operator from $\mathbb{R}^n$ to $\mathbb{R}$ corresponds to a unique vector in $\mathbb{R}^n$. In particular, the vector corresponding to the differential at $\mathbf{a}$ is called the *gradient* at $\mathbf{a}$.

DEFINITION 17.6.1. Suppose f is differentiable at $\mathbf{a}$. The *vector* $\nabla f(\mathbf{a}) \in \mathbb{R}^n$ (uniquely) determined by

$$\nabla f(\mathbf{a}) \cdot \mathbf{v} = \langle df(\mathbf{a}), \mathbf{v}\rangle \quad \forall \mathbf{v} \in \mathbb{R}^n,$$

is called the *gradient* of f at $\mathbf{a}$.

PROPOSITION 17.6.2. *If f is differentiable at $\mathbf{a}$, then*

$$\nabla f(\mathbf{a}) = \left(\frac{\partial f}{\partial x^1}(\mathbf{a}), \ldots, \frac{\partial f}{\partial x^n}(\mathbf{a})\right).$$

PROOF. It follows from Proposition 17.2.1 that the components of $\nabla f(\mathbf{a})$ are $\langle df(\mathbf{a}), \mathbf{e}_i\rangle$, i.e. $\dfrac{\partial f}{\partial x^i}(\mathbf{a})$. □

[1] We cannot establish the differentiability of $f+g$ (or αf) this way, since the existence of the partial derivatives does not imply differentiability.

Example For the example in Section 17.5 we have

$$\begin{aligned} \nabla f(\mathbf{a}) &= (2a_1 + 3{a_2}^2,\ 6a_1a_2 + 3{a_2}^2a_3,\ a_2^3 + 1), \\ \nabla f(1,0,1) &= (2,0,1). \end{aligned}$$

17.6.1. Geometric Interpretation of the Gradient.

PROPOSITION 17.6.3. *Suppose f is differentiable at $\mathbf{x}$. Then the directional derivatives at $\mathbf{x}$ are given by*

$$D_{\mathbf{v}} f(\mathbf{x}) = \mathbf{v} \cdot \nabla f(\mathbf{x}).$$

The unit vector $\mathbf{v}$ for which this is a maximum is $\mathbf{v} = \nabla f(\mathbf{x})/|\nabla f(\mathbf{x})|$ (assuming $|\nabla f(\mathbf{x})| \neq 0$), and the directional derivative in this direction is $|\nabla f(\mathbf{x})|$.

PROOF. From Definition 17.6.1 and Proposition 17.5.2 it follows that

$$\nabla f(\mathbf{x}) \cdot \mathbf{v} = \langle df(\mathbf{x}), \mathbf{v} \rangle = D_{\mathbf{v}} f(\mathbf{x})$$

This proves the first claim.

Now suppose $\mathbf{v}$ is a unit vector. From the Cauchy-Schwartz Inequality (80) we have

$$\nabla f(\mathbf{x}) \cdot \mathbf{v} \leq |\nabla f(\mathbf{x})|. \tag{186}$$

By the condition for equality in (80), equality holds in (186) iff $\mathbf{v}$ is a *positive* multiple of $\nabla f(\mathbf{x})$. Since $\mathbf{v}$ is a unit vector, this is equivalent to $\mathbf{v} = \nabla f(\mathbf{x})/|\nabla f(\mathbf{x})|$. The left side of (186) is then $|\nabla f(\mathbf{x})|$. □

17.6.2. Level Sets and the Gradient.

DEFINITION 17.6.4. If $f : \mathbb{R}^n \to \mathbb{R}$ then the *level set* through $\mathbf{x}$ is $\{\mathbf{y}\colon\ \mathrm{f}(\mathbf{y}) = \mathrm{f}(\mathbf{x})\ \}$.

For example, the contour lines on a map are the level sets of the height function.

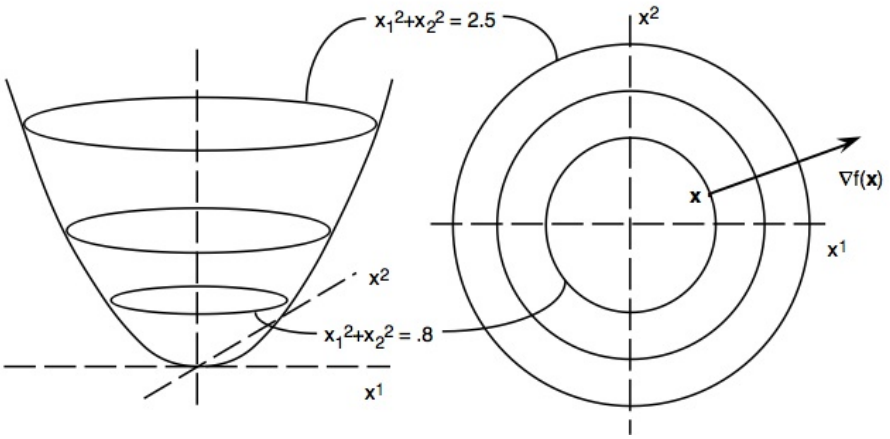

FIGURE 6. The graph of $f(\mathbf{x}) = x_1^2 + x_2^2$ is on the left. The level sets are on the right.

DEFINITION 17.6.5. A vector $\mathbf{v}$ is *tangent* at $\mathbf{x}$ to the level set S through $\mathbf{x}$ if

$$D_{\mathbf{v}} f(\mathbf{x}) = 0.$$

This is a reasonable definition, since f is *constant* on S, and so the rate of change of f in any direction tangent to S should be zero.

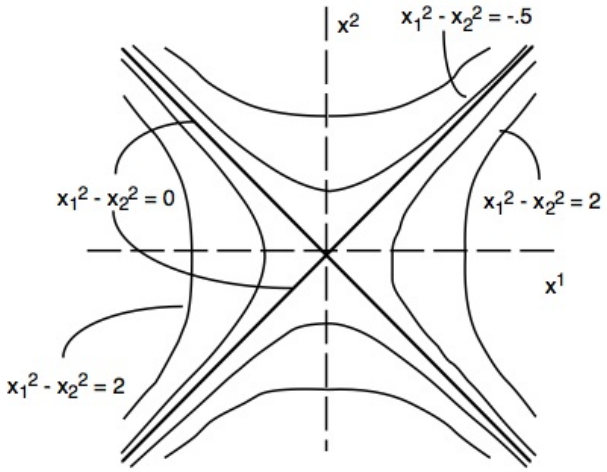

FIGURE 7. The level sets of $f(\mathbf{x}) = x_1^2 - x_2^2$. (The graph of f looks like a sadle.)

PROPOSITION 17.6.6. *Suppose f is differentiable at $\mathbf{x}$. Then $\nabla f(\mathbf{x})$ is orthogonal to all vectors which are tangent at $\mathbf{x}$ to the level set through $\mathbf{x}$.*

PROOF. Immediate by the previous Definition and Proposition 17.6.3. □

In the previous proposition, we say $\nabla f(\mathbf{x})$ *is orthogonal to the level set through* $\mathbf{x}$.

17.7. Some Interesting Examples

(1) An example where the partial derivatives exist but the other directional derivatives do not exist.

Let

$$f(x, y) = (xy)^{1/3}.$$

Then

(1) $\dfrac{\partial f}{\partial x}(0,0) = 0$ since $f = 0$ on the x-axis;
(2) $\dfrac{\partial f}{\partial y}(0,0) = 0$ since $f = 0$ on the y-axis;
(3) Let $\mathbf{v}$ be *any* vector. Then

$$\begin{aligned} D_{\mathbf{v}} f(0,0) &= \lim_{t \to 0} \frac{f(t\mathbf{v}) - f(0,0)}{t} \\ &= \lim_{t \to 0} \frac{t^{2/3}(v_1 v_2)^{1/3}}{t} \\ &= \lim_{t \to 0} \frac{(v_1 v_2)^{1/3}}{t^{1/3}}. \end{aligned}$$

This limit does *not* exist, unless $v_1 = 0$ or $v_2 = 0$.

(2) An example where the directional derivatives at some point all exist, but the function is not differentiable at the point.

Let

$$f(x, y) = \begin{cases} \dfrac{xy^2}{x^2 + y^4} & (x, y) \neq (0, 0) \\ 0 & (x, y) = (0, 0) \end{cases}$$

Let $\mathbf{v} = (v_1, v_2)$ be any non-zero vector. Then

$$\begin{aligned} D_{\mathbf{v}} f(0,0) &= \lim_{t\to 0} \frac{f(t\mathbf{v}) - f(0,0)}{t} \\ &= \lim_{t\to 0} \frac{\dfrac{t^3 v_1 {v_2}^2}{t^2 {v_1}^2 + t^4 {v_2}^4} - 0}{t} \\ &= \lim_{t\to 0} \frac{v_1 {v_2}^2}{{v_1}^2 + t^2 {v_2}^4} \\ &= \begin{cases} {v_2}^2/v_1 & v_1 \neq 0 \\ 0 & v_1 = 0 \end{cases} \end{aligned} \tag{187}$$

Thus the directional derivatives $D_{\mathbf{v}} f(0,0)$ exist for all $\mathbf{v}$, and are given by (187). In particular

$$\frac{\partial f}{\partial x}(0,0) = \frac{\partial f}{\partial y}(0,0) = 0. \tag{188}$$

But if f were differentiable at $(0,0)$, then we could compute any directional derivative from the partial drivatives. Thus for any vector $\mathbf{v}$ we would have

$$\begin{aligned} D_{\mathbf{v}} f(0,0) &= \langle df(0,0), \mathbf{v} \rangle \\ &= v_1 \frac{\partial f}{\partial x}(0,0) + v_2 \frac{\partial f}{\partial y}(0,0) \\ &= 0 \quad \text{from (188)} \end{aligned}$$

This contradicts (187).

(3) An Example where the directional derivatives at a point all exist, but the function is not continuous at the point

Take the same example as in (2). Approach the origin along the curve $x = \lambda^2$, $y = \lambda$. Then

$$\lim_{\lambda\to 0} f(\lambda^2, \lambda) = \lim_{\lambda\to 0} \frac{\lambda^4}{2\lambda^4} = \frac{1}{2}.$$

But if we approach the origin along any straight line of the form $(\lambda v_1, \lambda v_2)$, then we can check that the corresponding limit is 0.

Thus it is impossible to define f at $(0,0)$ in order to make f continuous there.

17.8. Differentiability Implies Continuity

Despite Example (3) in Section 17.7, we have the following result.

PROPOSITION 17.8.1. *If f is differentiable at $\mathbf{a}$, then it is continuous at $\mathbf{a}$.*

PROOF. Suppose f is differentiable at $\mathbf{a}$. Then

$$f(\mathbf{x}) = f(\mathbf{a}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x^i}(\mathbf{a})(x_i - a^i) + o(|\mathbf{x} - \mathbf{a}|).$$

Since $x^i - a^i \to 0$ and $o(|\mathbf{x} - \mathbf{a}|) \to 0$ as $\mathbf{x} \to \mathbf{a}$, it follows that $f(\mathbf{x}) \to f(\mathbf{a})$ as $\mathbf{x} \to \mathbf{a}$. That is, f is continuous at $\mathbf{a}$. □

17.9. Mean Value Theorem and Consequences

THEOREM 17.9.1. *Suppose f is continuous at all points on the line segment L joining $\mathbf{a}$ and $\mathbf{a} + \mathbf{h}$; and is differentiable at all points on L, except possibly at the end points.*

Then

$$f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) = \langle df(\mathbf{x}), \mathbf{h}\rangle \tag{189}$$

$$= \sum_{i=1}^{n} \frac{\partial f}{\partial x^i}(\mathbf{x})\, h^i \tag{190}$$

for some $\mathbf{x} \in L$, $\mathbf{x}$ *not an endpoint of* L.

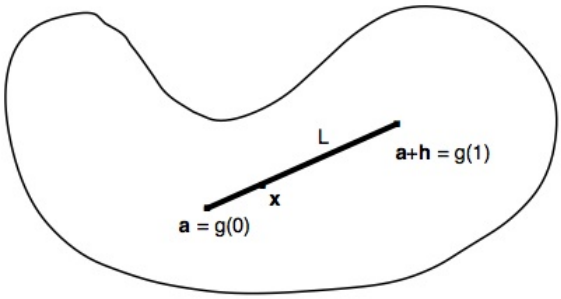

FIGURE 8. Diagram for Theorem 17.9.1 and its proof.

PROOF. Note that (190) follows immediately from (189) by Corollary 17.5.3.

Define the *one variable* function g by

$$g(t) = f(\mathbf{a}+t\mathbf{h}).$$

(See Figure 8.) Then g is continuous on $[0,1]$, being the composition of the continuous functions $t \mapsto \mathbf{a}+t\mathbf{h}$ and $\mathbf{x} \mapsto f(\mathbf{x})$. Moreover,

$$g(0) = f(\mathbf{a}),\ g(1) = f(\mathbf{a}+\mathbf{h}). \tag{191}$$

We next show that g is differentiable and compute its derivative.

If $0 < t < 1$, then f is differentiable at $\mathbf{a}+t\mathbf{h}$, and so

$$0 = \lim_{|\mathbf{w}|\to 0} \frac{f(\mathbf{a}+t\mathbf{h}+\mathbf{w}) - f(\mathbf{a}+t\mathbf{h}) - \langle df(\mathbf{a}+t\mathbf{h}), \mathbf{w}\rangle}{|\mathbf{w}|}. \tag{192}$$

Let $\mathbf{w} = s\mathbf{h}$ where s is a small real number, positive or negative. Since $|\mathbf{w}| = \pm s|\mathbf{h}|$, and since we may assume $\mathbf{h} \neq \mathbf{0}$ (as otherwise (189) is trivial), we see from (192) that

$$\begin{aligned} 0 &= \lim_{s\to 0} \frac{f\Big((\mathbf{a}+(t+s)\mathbf{h}\Big) - f(\mathbf{a}+t\mathbf{h}) - \langle df(\mathbf{a}+t\mathbf{h}), s\mathbf{h}\rangle}{s} \\ &= \lim_{s\to 0} \left(\frac{g(t+s)-g(t)}{s} - \langle df(\mathbf{a}+t\mathbf{h}), \mathbf{h}\rangle \right), \end{aligned}$$

using the linearity of $df(\mathbf{a}+t\mathbf{h})$.

Hence $g'(t)$ exists for $0 < t < 1$, and moreover

$$g'(t) = \langle df(\mathbf{a}+t\mathbf{h}), \mathbf{h}\rangle. \tag{193}$$

By the usual Mean Value Theorem for a function of one variable, applied to g, we have

$$g(1) - g(0) = g'(t) \tag{194}$$

for some $t \in (0,1)$.

Substituting (191) and (193) in (194), the required result (189) follows. □

If the norm of the gradient vector of f is bounded by M, then it is not surprising that the difference in value between $f(\mathbf{a})$ and $f(\mathbf{a}+\mathbf{h})$ is bounded by $M|\mathbf{h}|$. More precisely.

COROLLARY 17.9.2. *Assume the hypotheses of the previous theorem and suppose* $|\nabla f(\mathbf{x})| \leq M$ *for all* $\mathbf{x} \in L$. *Then*

$$|f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a})| \leq M|\mathbf{h}|$$

PROOF. From the previous theorem

$$\begin{aligned} |f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a})| &= |\langle df(\mathbf{x}), \mathbf{h}\rangle| \quad \text{for some } \mathbf{x} \in L \\ &= |\nabla f(\mathbf{x}) \cdot \mathbf{h}| \\ &\leq |\nabla f(\mathbf{x})|\,|\mathbf{h}| \\ &\leq M|\mathbf{h}|. \end{aligned}$$

□

COROLLARY 17.9.3. *Suppose* $\Omega \subset \mathbb{R}^n$ *is open and* **connected** *and* $f : \Omega \to \mathbb{R}$. *Suppose* f *is differentiable in* Ω *and* $df(\mathbf{x}) = 0$ *for all* $\mathbf{x} \in \Omega$[2].

Then f *is constant on* Ω.

PROOF. Choose any $\mathbf{a} \in \Omega$ and suppose $f(\mathbf{a}) = \alpha$. Let

$$E = \{\mathbf{x} \in \Omega : f(\mathbf{x}) = \alpha\}.$$

Then E is non-empty (as $\mathbf{a} \in E$). We will prove E is both open and closed *in* Ω. Since Ω is connected, this will imply that E is *all* of Ω[3]. This establishes the result.

To see E is *open*[4], suppose $\mathbf{x} \in E$ and choose $r > 0$ so that $B_r(\mathbf{x}) \subset \Omega$.

If $\mathbf{y} \in B_r(\mathbf{x})$, then from (189) for some $\mathbf{u}$ between $\mathbf{x}$ and $\mathbf{y}$,

$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{x}) &= \langle df(\mathbf{u}), \mathbf{y} - \mathbf{x}\rangle \\ &= 0, \quad \text{by hypothesis.} \end{aligned}$$

Thus $f(\mathbf{y}) = f(\mathbf{x})\,(= \alpha)$, and so $\mathbf{y} \in E$.

Hence $B_r(\mathbf{x}) \subset E$ and so E is open.

To show that E is *closed* in Ω, it is sufficient to show that $E^c = \{\mathbf{y} : f(\mathbf{x}) \neq \alpha\}$ is open in Ω.

From Proposition 17.8.1 we know that f is continuous. Since we have $E^c = f^{-1}[\mathbb{R} \setminus \{\alpha\}]$ and $\mathbb{R} \setminus \{\alpha\}$ is open, it follows that E^c is open in Ω. Hence E is closed in Ω, as required.

Since $E \neq \emptyset$, and E is both open and closed in Ω, it follows $E = \Omega$ (as Ω is connected).

In other words, f is constant $(= \alpha)$ on Ω. □

17.10. Continuously Differentiable Functions

We saw in Section 17.7, Example (2), that the partial derivatives (and even all the directional derivatives) of a function can exist without the function being differentiable.

However, we do have the following important theorem:

THEOREM 17.10.1. *Suppose* $f : \Omega\,(\subset \mathbb{R}^n) \to \mathbb{R}$ *where* Ω *is open. If the partial derivatives of* f *exist and are continuous at every point in* Ω, *then* f *is differentiable everywhere in* Ω.

---

[2] Equivalently, $\nabla f(\mathbf{x}) = \mathbf{0}$ in Ω.

[3] This is a standard technique for showing that all points in a connected set have a certain property, c.f. the proof of Theorem 16.4.4.

[4] Being open in Ω and being open in $\mathbb{R}^n$ is the same for subsets of Ω, since we are assuming Ω is itself open in $\mathbb{R}^n$.

Remark: If the partial derivatives of f exist in some neighbourhood of, and are continuous at, a *single point*, it does not necessarily follow that f is differentiable at that point. The hypotheses of the theorem need to hold at *all* points in some *open* set Ω.

PROOF. We prove the theorem in case $n = 2$ (the proof for $n > 2$ is only notationally more complicated). See Figure 9.

Suppose that the partial derivatives of f exist and are continuous in Ω. Then if $\mathbf{a} \in \Omega$ and $\mathbf{a} + \mathbf{h}$ is sufficiently close to $\mathbf{a}$,

$$\begin{aligned} f(a^1 + h^1, a^2 + h^2) &= f(a^1, a^2) \\ &\quad + f(a^1 + h^1, a^2) - f(a^1, a^2) \\ &\quad + f(a^1 + h^1, a^2 + h^2) - f(a^1 + h^1, a^2) \\ &= f(a^1, a^2) + \frac{\partial f}{\partial x^1}(\xi^1, a^2)h^1 + \frac{\partial f}{\partial x^2}(a^1 + h^1, \xi^2)h^2, \end{aligned}$$

for some ξ^1 between a^1 and $a^1 + h^1$, and some ξ^2 between a^2 and $a^2 + h^2$. The first partial derivative comes from applying the usual Mean Value Theorem, for a function of *one* variable, to the function $f(x^1, a^2)$ obtained by fixing a^2 and taking x^1 as a variable. The second partial derivative is similarly obtained by considering the function $f(a^1 + h^1, x^2)$, where $a^1 + h^1$ is fixed and x^2 is variable.

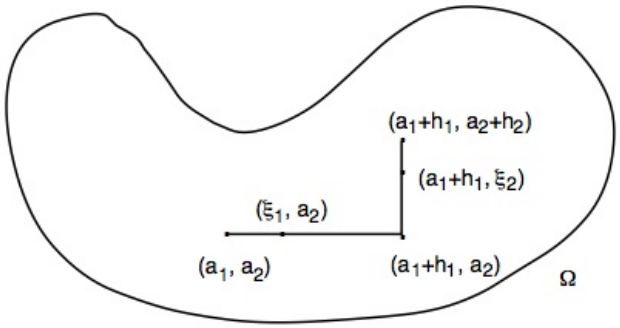

FIGURE 9. Diagram for the proof of Theorem 17.10.1.

Hence

$$\begin{aligned} f(a^1 + h^1, a^2 + h^2) &= f(a^1, a^2) + \frac{\partial f}{\partial x^1}(a^1, a^2)h^1 + \frac{\partial f}{\partial x^2}(a^1, a^2)h^2 \\ &\quad + \left(\frac{\partial f}{\partial x^1}(\xi^1, a^2) - \frac{\partial f}{\partial x^1}(a^1, a^2) \right) h^1 \\ &\quad + \left(\frac{\partial f}{\partial x^2}(a^1 + h^1, \xi^2) - \frac{\partial f}{\partial x^2}(a^1, a^2) \right) h^2 \\ &= f(a^1, a^2) + L(\mathbf{h}) + \psi(\mathbf{h}), \text{ say.} \end{aligned}$$

Here L is the linear map defined by

$$\begin{aligned} L(\mathbf{h}) &= \frac{\partial f}{\partial x^1}(a^1, a^2)h^1 + \frac{\partial f}{\partial x^2}(a^1, a^2)h^2 \\ &= \begin{bmatrix} \frac{\partial f}{\partial x^1}(a^1, a^2) & \frac{\partial f}{\partial x^2}(a^1, a^2) \end{bmatrix} \begin{bmatrix} h^1 \\ h^2 \end{bmatrix}. \end{aligned}$$

Thus L is represented by the previous 1×2 matrix.

We claim that the *error term*

$$\psi(\mathbf{h}) = \left(\frac{\partial f}{\partial x^1}(\xi^1, a^2) - \frac{\partial f}{\partial x^1}(a^1, a^2)\right) h^1 + \left(\frac{\partial f}{\partial x^2}(a^1 + h^1, \xi^2) - \frac{\partial f}{\partial x^2}(a^1, a^2)\right) h^2$$

can be written as $o(|\mathbf{h}|)$

This follows from the facts:

(1) $\dfrac{\partial f}{\partial x^1}(\xi^1, a^2) \to \dfrac{\partial f}{\partial x^1}(a^1, a^2)$ as $\mathbf{h} \to \mathbf{0}$ (by continuity of the partial derivatives),

(2) $\dfrac{\partial f}{\partial x^2}(a^1 + h^1, \xi^2) \to \dfrac{\partial f}{\partial x^2}(a^1, a^2)$ as $\mathbf{h} \to \mathbf{0}$ (again by continuity of the partial derivatives),

(3) $|h^1| \leq |\mathbf{h}|$, $|h^2| \leq |\mathbf{h}|$.

It now follows from Proposition 17.5.4 that f is differentiable at a, and the differential of f is given by the previous 1×2 matrix of partial derivatives.

Since $\mathbf{a} \in \Omega$ is arbitrary, this completes the proof. □

DEFINITION 17.10.2. If the partial derivatives of f exist and are continuous in the open set Ω, we say f is a C^1 (or *continuously differentiable*) function on Ω. One writes $f \in C^1(\Omega)$.

It follows from the previous Theorem that if $f \in C^1(\Omega)$ then f is indeed differentiable in Ω. *Exercise:* The converse may not be true, give a simple counterexample in $\mathbb{R}$.

17.11. Higher-Order Partial Derivatives

Suppose $f : \Omega\,(\subset \mathbb{R}^n) \to \mathbb{R}$. The partial derivatives $\dfrac{\partial f}{\partial x^1}, \ldots, \dfrac{\partial f}{\partial x^n}$, if they exist, are also functions from Ω to $\mathbb{R}$, and may themselves have partial derivatives.

The jth partial derivative of $\dfrac{\partial f}{\partial x_i}$ is denoted by

$$\frac{\partial^2 f}{\partial x_j \partial x_i} \text{ or } f_{ij} \text{ or } D_{ij} f.$$

If all first and second partial derivatives of f exist and are *continuous* in Ω [5] we write

$$f \in C^2(\Omega).$$

Similar remarks apply to higher order derivatives, and we similarly define $C^q(\Omega)$ for any integer $q \geq 0$.

Note that

$$C^0(\Omega) \supset C^1(\Omega) \supset C^2(\Omega) \supset \ldots$$

The usual rules for differentiating a sum, product or quotient of functions of a single variable apply to partial derivatives. It follows that $C^k(\Omega)$ is closed under addition, products and quotients (if the denominator is non-zero).

The next theorem shows that for higher order derivatives, the actual order of differentiation does not matter, only the number of derivatives with respect to each variable is important. Thus

$$\frac{\partial^2 f}{\partial x^i \partial x^j} = \frac{\partial^2 f}{\partial x^j \partial x^i},$$

[5] In fact, it is sufficient to assume just that the *second* partial derivatives are continuous. For under this assumption, each $\partial f/\partial x^i$ must be differentiable by Theorem 17.10.1 applied to $\partial f/\partial x^i$. From Proposition 17.8.1 applied to $\partial f/\partial x^i$ it then follows that $\partial f/\partial x^i$ is continuous.

and so

$$\frac{\partial^3 f}{\partial x^i \partial x^j \partial x^k} = \frac{\partial^3 f}{\partial x^j \partial x^i \partial x^k} = \frac{\partial^3 f}{\partial x^j \partial x^k \partial x^i}, \text{ etc.}$$

THEOREM 17.11.1. *If $f \in C^1(\Omega)$[6] and both f_{ij} and f_{ji} exist and are continuous (for some $i \neq j$) in Ω, then $f_{ij} = f_{ji}$ in Ω.*

In particular, if $f \in C^2(\Omega)$ then $f_{ij} = f_{ji}$ for all $i \neq j$.

PROOF. For notational simplicity we take $n = 2$. The proof for $n > 2$ is very similar.

Suppose $\mathbf{a} \in \Omega$ and suppose $h > 0$ is some sufficiently small real number.

Consider the *second difference quotient* defined by

$$\begin{aligned} A(h) &= \frac{1}{h^2}\Big(\big(f(a^1+h, a^2+h) - f(a^1, a^2+h)\big) \\ &\qquad -\big(f(a^1+h, a^2) - f(a^1, a^2)\big)\Big) \qquad (195) \\ &= \frac{1}{h^2}\big(g(a^2+h) - g(a^2)\big), \qquad (196) \end{aligned}$$

where

$$g(x^2) = f(a^1+h, x^2) - f(a^1, x^2).$$

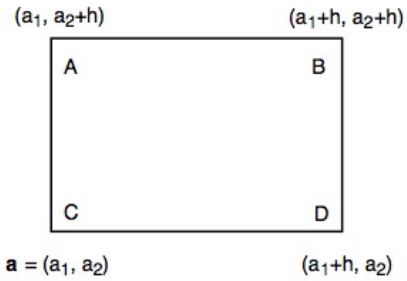

FIGURE 10. Diagram for the proof of Theorem 17.11.1. Note that

$$\begin{aligned} A(h) &= \Big(\big(f(B) - f(A)\big) - \big(f(D) - f(C)\big)\Big)/h^2 \\ &= \Big(\big(f(B) - f(D)\big) - \big(f(A) - f(C)\big)\Big)/h^2. \end{aligned}$$

From the definition of partial differentiation, $g'(x^2)$ exists and

$$g'(x^2) = \frac{\partial f}{\partial x^2}(a^1+h, x^2) - \frac{\partial f}{\partial x^2}(a^1, x^2) \qquad (197)$$

for $a^2 \leq x \leq a^2 + h$.

Applying the mean value theorem for a function of a single variable to (196), we see from (197) that

$$\begin{aligned} A(h) &= \frac{1}{h} g'(\xi^2) \quad \text{some } \xi^2 \in (a^2, a^2+h) \\ &= \frac{1}{h}\left(\frac{\partial f}{\partial x^2}(a^1+h, \xi^2) - \frac{\partial f}{\partial x^2}(a^1, \xi^2)\right). \qquad (198) \end{aligned}$$

[6] As usual, Ω is assumed to be open.

Applying the mean value theorem again to the function $\dfrac{\partial f}{\partial x^2}(x^1, \xi^2)$, with ξ^2 fixed, we see

$$A(h) = \frac{\partial^2 f}{\partial x^1 \partial x^2}(\xi^1, \xi^2) \quad \text{some } \xi^1 \in (a^1, a^1 + h). \tag{199}$$

If we now rewrite (195) as

$$\begin{aligned} A(h) \quad = \quad & \frac{1}{h^2}\Big(\Big(f(a^1 + h, a^2 + h) - f(a^1 + h, a^2)\Big) \\ & -\Big(f(a^1, a^2 + h) - f(a^1 + a^2)\Big)\Big) \end{aligned} \tag{200}$$

and interchange the roles of x^1 and x^2 in the previous argument, we obtain

$$A(h) = \frac{\partial^2 f}{\partial x^2 \partial x^1}(\eta^1, \eta^2) \tag{201}$$

for some $\eta^1 \in (a^1, a^1 + h)$, $\eta^2 \in (a^2, a^2 + h)$.

If we let $h \to 0$ then (ξ^1, ξ^2) and $(\eta^1, \eta^2) \to (a^1, a^2)$, and so from (199), (201) and the continuity of f_{12} and f_{21} at $\mathbf{a}$, it follows that

$$f_{12}(\mathbf{a}) = f_{21}(\mathbf{a}).$$

This completes the proof. □

17.12. Taylor's Theorem

If $g \in C^1[a, b]$, then we know

$$g(b) = g(a) + \int_a^b g'(t)\, dt$$

This is the case $k = 1$ of the following version of Taylor's Theorem for a function of *one* variable.

THEOREM 17.12.1 (Taylor's Formula; Single Variable, First Version). *Suppose $g \in C^k[a, b]$. Then*

$$\begin{aligned} g(b) \quad = \quad & g(a) + g'(a)(b - a) + \frac{1}{2!}g''(a)(b - a)^2 + \cdots \\ & + \frac{1}{(k-1)!}g^{(k-1)}(a)(b - a)^{k-1} + \int_a^b \frac{(b - t)^{k-1}}{(k - 1)!}g^{(k)}(t)\, dt. \end{aligned} \tag{202}$$

PROOF. An elegant (but not obvious) proof is to begin by computing:

$$\begin{aligned} & \frac{d}{dt}\left(g\varphi^{(k-1)} - g'\varphi^{(k-2)} + g''\varphi^{(k-3)} - \cdots + (-1)^{k-1}g^{(k-1)}\varphi\right) \\ & \quad = \left(g\varphi^{(k)} + g'\varphi^{(k-1)}\right) - \left(g'\varphi^{(k-1)} + g''\varphi^{(k-2)}\right) + \left(g''\varphi^{(k-2)} + g'''\varphi^{(k-3)}\right) - \\ & \qquad \cdots + (-1)^{k-1}\left(g^{(k-1)}\varphi' + g^{(k)}\varphi\right) \\ & \quad = g\varphi^{(k)} + (-1)^{k-1}g^{(k)}\varphi. \end{aligned} \tag{203}$$

Now choose

$$\varphi(t) = \frac{(b - t)^{k-1}}{(k - 1)!}.$$

Then

$$\begin{aligned}
\varphi'(t) &= (-1)\frac{(b-t)^{k-2}}{(k-2)!} \\
\varphi''(t) &= (-1)^2\frac{(b-t)^{k-3}}{(k-3)!} \\
&\vdots \\
\varphi^{(k-3)}(t) &= (-1)^{k-3}\frac{(b-t)^2}{2!} \\
\varphi^{(k-2)}(t) &= (-1)^{k-2}(b-t) \\
\varphi^{(k-1)}(t) &= (-1)^{k-1} \\
\varphi^{k}(t) &= 0. \qquad (204)
\end{aligned}$$

Hence from (203) we have

$$\begin{aligned}
&(-1)^{k-1}\frac{d}{dt}\left(g(t)+g'(t)(b-t)+g''(t)\frac{(b-t)^2}{2!}+\cdots+g^{k-1}(t)\frac{(b-t)^{k-1}}{(k-1)!}\right) \\
&\quad = (-1)^{k-1}g^{(k)}(t)\frac{(b-t)^{k-1}}{(k-1)!}.
\end{aligned}$$

Dividing by $(-1)^{k-1}$ and integrating both sides from a to b, we get

$$\begin{aligned}
&g(b)-\left(g(a)+g'(a)(b-a)+g''(a)\frac{(b-a)^2}{2!}+\cdots+g^{(k-1)}(a)\frac{(b-a)^{k-1}}{(k-1)!}\right) \\
&\quad = \int_a^b g^{(k)}(t)\frac{(b-t)^{k-1}}{(k-1)!}\,dt.
\end{aligned}$$

This gives formula (202). □

THEOREM 17.12.2 (Taylor's Formula; Single Variable, Second Version). *Suppose $g \in C^k[a,b]$. Then*

$$\begin{aligned}
g(b) &= g(a)+g'(a)(b-a)+\frac{1}{2!}g''(a)(b-a)^2+\cdots \\
&\quad +\frac{1}{(k-1)!}g^{(k-1)}(a)(b-a)^{k-1}+\frac{1}{k!}g^{(k)}(\xi)(b-a)^k \qquad (205)
\end{aligned}$$

for some $\xi \in (a,b)$.

PROOF. We establish (205) from (202).

Since $g^{(k)}$ is continuous in $[a,b]$, it has a minimum value m, and a maximum value M, say.

By elementary properties of integrals, it follows that

$$\int_a^b m\frac{(b-t)^{k-1}}{(k-1)!}\,dt \le \int_a^b g^{(k)}(t)\frac{(b-t)^{k-1}}{(k-1)!}\,dt \le \int_a^b M\frac{(b-t)^{k-1}}{(k-1)!}\,dt,$$

i.e.

$$m \le \frac{\displaystyle\int_a^b g^{(k)}(t)\frac{(b-t)^{k-1}}{(k-1)!}dt}{\displaystyle\int_a^b \frac{(b-t)^{k-1}}{(k-1)!}dt} \le M.$$

By the Intermediate Value Theorem, $g^{(k)}$ takes all values in the range $[m,M]$, and so the middle term in the previous inequality must equal $g^{(k)}(\xi)$ for some

$\xi \in (a,b)$. Since

$$\int_a^b \frac{(b-t)^{k-1}}{(k-1)!}dt = \frac{(b-a)^k}{k!},$$

it follows

$$\int_a^b g^{(k)}(t)\frac{(b-t)^{k-1}}{(k-1)!}dt = \frac{(b-a)^k}{k!}g^{(k)}(\xi).$$

Formula (205) now follows from (202). □

Remark For a direct proof of (205), which does not involve any integration, see [**Sw**, pp 582–3] or [**Fl**, Appendix A2].

Taylor's Theorem generalises easily to functions of more than one variable.

THEOREM 17.12.3 (Taylor's Formula; Several Variables).
Suppose $f \in C^k(\Omega)$ where $\Omega \subset \mathbb{R}^n$, and the line segment joining $\mathbf{a}$ and $\mathbf{a}+\mathbf{h}$ is a subset of Ω.

Then

$$\begin{aligned} f(\mathbf{a}+\mathbf{h}) &= f(\mathbf{a}) + \sum_{i-1}^{n} D_i f(\mathbf{a})\, h^i + \frac{1}{2!}\sum_{i,j=1}^{n} D_{ij} f(\mathbf{a})\, h^i h^j + \cdots \\ &\quad + \frac{1}{(k-1)!}\sum_{i_1,\cdots,i_{k-1}=1}^{n} D_{i_1\ldots i_{k-1}} f(\mathbf{a})\, h^{i_1}\cdot\ldots\cdot h^{i_{k-1}} + R_k(\mathbf{a},\mathbf{h}) \end{aligned}$$

where

$$\begin{aligned} R_k(\mathbf{a},\mathbf{h}) &= \frac{1}{(k-1)!}\sum_{i_1,\ldots,i_k=1}^{n}\int_0^1 (1-t)^{k-1} D_{i_1\ldots i_k} f(\mathbf{a}+t\mathbf{h})\,dt \\ &= \frac{1}{k!}\sum_{i_1,\ldots,i_k=1}^{n} D_{i_1,\ldots,i_k} f(\mathbf{a}+s\mathbf{h})\, h^{i_1}\cdot\ldots\cdot h^{i_k} \quad \textit{for some } s\in(0,1). \end{aligned}$$

PROOF. First note that for any differentiable function $F : D\,(\subset \mathbb{R}^n) \to \mathbb{R}$ we have

$$\frac{d}{dt}F(\mathbf{a}+t\mathbf{h}) = \sum_{i=1}^{n} D_i F(\mathbf{a}+t\mathbf{h})\, h^i. \tag{206}$$

This is just a particular case of the *chain rule*, which we will discuss later. This particular version follows from (193) and Corollary 17.5.3 (with f there replaced by F).

Let

$$g(t) = f(a+th).$$

Then $g : [0,1] \to \mathbb{R}$. We will apply Taylor's Theorem for a function of one variable to g.

From (206) we have

$$g'(t) = \sum_{i-1}^{n} D_i f(\mathbf{a}+t\mathbf{h})\, h^i. \tag{207}$$

Differentiating again, and applying (206) to $D_i F$, we obtain

$$\begin{aligned} g''(t) &= \sum_{i=1}^{n}\left(\sum_{j=1}^{n} D_{ij} f(\mathbf{a}+t\mathbf{h})\, h^j\right) h^i \\ &= \sum_{i,j=1}^{n} D_{ij} f(\mathbf{a}+t\mathbf{h})\, h^i h^j. \end{aligned} \tag{208}$$

Similarly

$$g'''(t) = \sum_{i,j,k=1}^{n} D_{ijk} f(\mathbf{a} + t\mathbf{h})\, h^i h^j h^k, \tag{209}$$

etc. In this way, we see $g \in C^k[0,1]$ and obtain formulae for the derivatives of g.

But from (202) and (205) we have

$$\begin{aligned} g(1) &= g(0) + g'(0) + \frac{1}{2!}g''(0) + \cdots + \frac{1}{(k-1)!}g^{(k-1)}(0) \\ &+ \begin{cases} \dfrac{1}{(k-1)!}\displaystyle\int_0^1 (1-t)^{k-1} g^{(k)}(t)\, dt \\ \qquad\qquad or \\ \dfrac{1}{k!} g^{(k)}(s) \quad \text{some } s \in (0,1). \end{cases} \end{aligned}$$

If we substitute (207), (208), (209) etc. into this, we obtain the required results. □

Remark The first two terms of Taylor's Formula give the best *first order approximation*[7] in $\mathbf{h}$ to $f(\mathbf{a}+\mathbf{h})$ for $\mathbf{h}$ near $\mathbf{0}$. The first three terms give the best *second order approximation*[8] in $\mathbf{h}$, the first four terms give the best *third order approximation*, etc.

Note that the *remainder term* $R_k(\mathbf{a}, \mathbf{h})$ in Theorem 17.12.3 can be written as $O(|\mathbf{h}|^k)$ (see the Remarks on rates of convergence in Section 17.5), i.e.

$$\frac{R_k(\mathbf{a}, \mathbf{h})}{|\mathbf{h}|^k} \text{ is bounded as } \mathbf{h} \to 0.$$

This follows from the second version for the remainder in Theorem 17.12.3 and the facts:

(1) $D_{i_1 \ldots i_k} f(\mathbf{x})$ is continuous, and hence bounded on compact sets,
(2) $|h^{i_1} \cdot \ldots \cdot h^{i_k}| \leq |\mathbf{h}|^k$.

Example Let

$$f(x,y) = (1+y^2)^{1/2} \cos x.$$

One finds the best second order approximation to f for (x,y) near $(0,1)$ as follows.

First note that

$$f(0,1) = 2^{1/2}.$$

Moreover,

$$\begin{array}{lllll} f_1 &= -(1+y^2)^{1/2} \sin x; &= 0 & \text{at } (0,1) \\ f_2 &= y(1+y^2)^{-1/2} \cos x; &= 2^{-1/2} & \text{at } (0,1) \\ f_{11} &= -(1+y^2)^{1/2} \cos x; &= -2^{1/2} & \text{at } (0,1) \\ f_{12} &= -y(1+y^2)^{-1/2} \sin x; &= 0 & \text{at } (0,1) \\ f_{22} &= (1+y^2)^{-3/2} \cos x; &= 2^{-3/2} & \text{at } (0,1). \end{array}$$

Hence

$$f(x,y) = 2^{1/2} + 2^{-1/2}(y-1) - 2^{1/2}x^2 + 2^{-3/2}(y-1)^2 + R_3\Big((0,1),(x,y)\Big),$$

where

$$R_3\Big((0,1),(x,y)\Big) = O\left(|(x,y)-(0,1)|^3\right) = O\left(\left(x^2+(y-1)^2\right)^{3/2}\right).$$

[7] I.e. constant plus linear term.
[8] I.e. constant plus linear term plus quadratic term.

CHAPTER 18

Differentiation of Vector-Valued Functions

18.1. Introduction

In this chapter we consider functions

$$\mathbf{f}: D\,(\subset \mathbb{R}^n) \to \mathbb{R}^n,$$

with $m \geq 1$. You should have a look back at Section 10.1.

We write

$$\mathbf{f}(x^1, \ldots, x^n) = \Big(f^1(x^1, \ldots, x^n), \ldots, f^m(x^1, \ldots, x^n)\Big)$$

where

$$f^i: D \to \mathbb{R}, \quad i = 1, \ldots, m,$$

are *real*-valued functions.

Example Let

$$\mathbf{f}(x, y, z) = (x^2 - y^2, 2xz + 1).$$

Then $f^1(x, y, z) = x^2 - y^2$ and $f^2(x, y, z) = 2xz + 1$.

Reduction to Component Functions For many purposes we can reduce the study of functions $\mathbf{f}$, as above, to the study of the corresponding *real*-valued functions $f^1, \ldots, f^m$. However, this is not always a good idea, since studying the f^i involves a choice of coordinates in $\mathbb{R}^n$, and this can obscure the geometry involved.

In Definitions 18.2.1, 18.3.1 and 18.4.1 we define the notion of partial derivative, directional derivative, and differential of $\mathbf{f}$ without reference to the component functions. In Propositions 18.2.2, 18.3.2 and 18.4.2 we show these definitions are equivalent to definitions in terms of the component functions.

18.2. Paths in $\mathbb{R}^m$

In this section we consider the case corresponding to $n = 1$ in the notation of the previous section. This is an important case in its own right and also helps motivates the case $n > 1$.

Definition 18.2.1. Let I be an interval in $\mathbb{R}$. If $\mathbf{f}: I \to \mathbb{R}^n$ then the *derivative* or *tangent vector* at t is the vector

$$\mathbf{f}'(t) = \lim_{s \to 0} \frac{\mathbf{f}(t+s) - \mathbf{f}(t)}{s},$$

provided the limit exists[1]. In this case we say $\mathbf{f}$ is *differentiable* at t. If, moreover, $\mathbf{f}'(t) \neq 0$ then $\mathbf{f}'(t)/|\mathbf{f}'(t)|$ is called the *unit tangent* at t.

Remark Although we say $\mathbf{f}'(t)$ is the tangent vector at t, we should really think of $\mathbf{f}'(t)$ as a vector with its "base" at $\mathbf{f}(t)$. See the next diagram.

[1] If t is an endpoint of I then one takes the corresponding one-sided limits.

PROPOSITION 18.2.2. *Let* $\mathbf{f}(t) = \left(f^1(t), \ldots, f^m(t)\right)$. *Then* $\mathbf{f}$ *is differentiable at* t *iff* $f^1, \ldots, f^m$ *are differentiable at* t. *In this case*

$$\mathbf{f}'(t) = \left(f^{1\prime}(t), \ldots, f^{m\prime}(t)\right).$$

PROOF. Since

$$\frac{\mathbf{f}(t+s) - \mathbf{f}(t)}{s} = \left(\frac{f^1(t+s) - f^1(t)}{s}, \ldots, \frac{f^m(t+s) - f^m(t)}{s}\right),$$

The theorem follows by applying Theorem 10.4.4. □

DEFINITION 18.2.3. If $\mathbf{f}(t) = \left(f^1(t), \ldots, f^m(t)\right)$ then $\mathbf{f}$ is C^1 if each f^i is C^1.

We have the usual rules for differentiating the sum of two functions from I to $\Re^m$, and the product of such a function with a real valued function (*exercise*: formulate and prove such a result). The following rule for differentiating the inner product of two functions is useful.

PROPOSITION 18.2.4. *If* $\mathbf{f}_1, \mathbf{f}_2 : I \to \mathbb{R}^n$ *are differentable at* t *then*

$$\frac{d}{dt}\Big(\mathbf{f}_1(t), \mathbf{f}_2(t)\Big) = \Big(\mathbf{f}_1'(t), \mathbf{f}_2(t)\Big) + \Big(\mathbf{f}_1(t), \mathbf{f}_2'(t)\Big).$$

PROOF. Since

$$\Big(\mathbf{f}_1(t), \mathbf{f}_2(t)\Big) = \sum_{i=1}^{m} f_1^i(t) f_2^i(t),$$

the result follows from the usual rule for differentiating sums and products. □

If $\mathbf{f} : I \to \mathbb{R}^n$, we can think of $\mathbf{f}$ as tracing out a "curve" in $\mathbb{R}^n$ (we will make this precise later). The terminology *tangent vector* is reasonable, as we see from the following diagram. Sometimes we speak of the tangent vector *at* $\mathbf{f}(t)$ rather than *at* t, but we need to be careful if $\mathbf{f}$ is not one-one, as in the second diagram in figure 1.

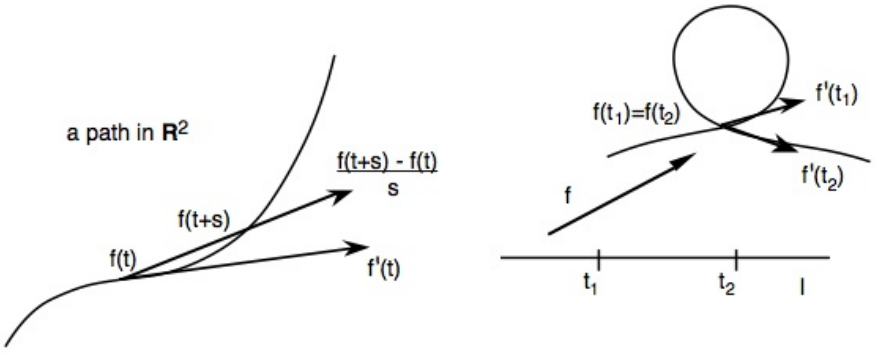

FIGURE 1. Two examples of a curve in $\mathbb{R}^2$. That is, of the path traced out by a function $f : I \to \mathbb{R}^2$, where I is an interval in $\mathbb{R}$.

Examples (See Figure 2.)

(1) Let

$$\mathbf{f}(t) = (\cos t, \sin t) \quad t \in [0, 2\pi).$$

This traces out a circle in $\mathbb{R}^2$ and

$$\mathbf{f}'(t) = (-\sin t, \cos t).$$

(2) Let

$$\mathbf{f}(t) = (t, t^2).$$

This traces out a parabola in $\mathbb{R}^2$ and

$$\mathbf{f}'(t) = (1, 2t).$$

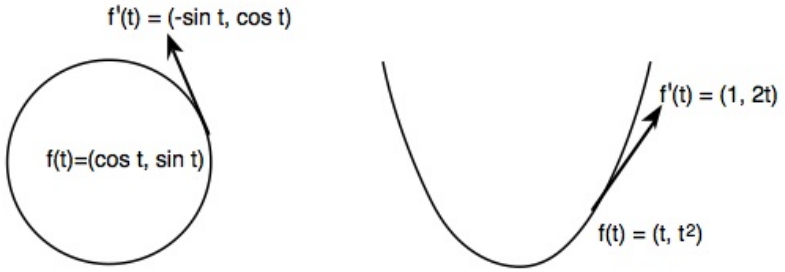

FIGURE 2. A circle and a parabola in $\mathbb{R}^2$, each being the image of a function $f : I \to \mathbb{R}^2$ for some interval $I \subset \mathbb{R}$.

Example Consider the functions

(1) $\mathbf{f}_1(t) = (t, t^3) \quad t \in \mathbb{R}$,
(2) $\mathbf{f}_2(t) = (t^3, t^9) \quad t \in \mathbb{R}$,
(3) $\mathbf{f}_3(t) = (\sqrt[3]{t}, t) \quad t \in \mathbb{R}$.

Then each function $\mathbf{f}_i$ traces out the same "cubic" curve in $\mathbb{R}^2$, (i.e., the image is the same set of points), and

$$\mathbf{f}_1(0) = \mathbf{f}_2(0) = \mathbf{f}_3(0) = (0, 0).$$

However,

$$\mathbf{f}_1'(0) = (1, 0),\ \mathbf{f}_2'(0) = (0, 0),\ \mathbf{f}_3'(0) \text{ is undefined.}$$

Intuitively, we will think of a *path* in $\mathbb{R}^n$ as a function $\mathbf{f}$ which neither stops nor reverses direction. It is often convenient to consider the variable t as representing "time". We will think of the corresponding *curve* as the set of points traced out by $\mathbf{f}$. Many different paths (i.e. functions) will give the same curve; they correspond to tracing out the curve at different times and velocities. We make this precise as follows:

DEFINITION 18.2.5. We say $\mathbf{f} : I \to \mathbb{R}^n$ is a *path*[2] in $\mathbb{R}^n$ if $\mathbf{f}$ is C^1 and $\mathbf{f}'(t) \neq 0$ for $t \in I$. We say the two paths $\mathbf{f}_1 : I_1 \to \mathbb{R}^n$ and $\mathbf{f}_2 : I_2 \to \mathbb{R}^n$ are *equivalent* if there exists a function $\phi : I_1 \to I_2$ such that $\mathbf{f}_1 = \mathbf{f}_2 \circ \phi$, where ϕ is C^1 and $\phi'(t) > 0$ for $t \in I_1$.

A *curve* is an equivalence class of paths. Any path in the equivalence class is called a *parametrisation* of the curve.

We can think of ϕ as giving another way of measuring "time".

We expect that the *unit* tangent vector to a curve should depend only on the curve itself, and not on the particular parametrisation. This is indeed the case, as is shown by the following Proposition.

[2]Other texts may have different terminology.

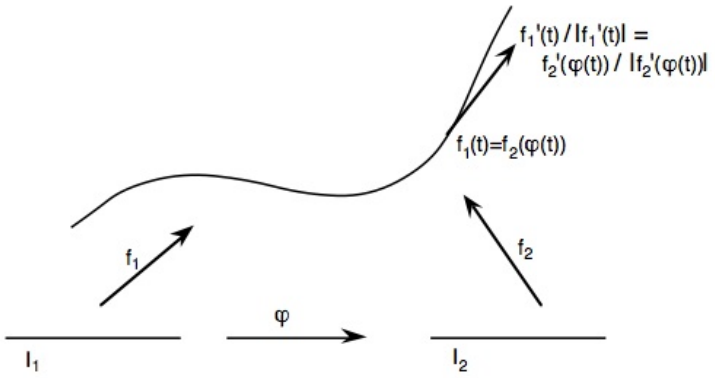

FIGURE 3. A curve with two parametrisations $f_1 : I_1 \to \mathbb{R}^2$ and $f_2 : I_1 \to \mathbb{R}^2$, where $\phi : I_1 \to I_2$ and $f_1(t) = f_2(\phi(t))$.

PROPOSITION 18.2.6. *Suppose* $\mathbf{f}_1 : I_1 \to \mathbb{R}^n$ *and* $\mathbf{f}_2 : I_2 \to \mathbb{R}^n$ *are equivalent parametrisations; and in particular* $\mathbf{f}_1 = \mathbf{f}_2 \circ \phi$ *where* $\phi : I_1 \to I_2$, ϕ *is* C^1 *and* $\phi'(t) > 0$ *for* $t \in I_1$. *Then* $\mathbf{f}_1$ *and* $\mathbf{f}_2$ *have the same unit tangent vector at* t *and* $\phi(t)$ *respectively.*

PROOF. From the chain rule for a function of one variable, we have

$$\begin{aligned} \mathbf{f}_1'(t) &= \Big(f_1^{1\prime}(t), \ldots, f_1^{m\prime}(t)\Big) \\ &= \Big(f_2^{1\prime}(\phi(t))\,\phi'(t), \ldots, f_2^{m\prime}(\phi(t))\,\phi'(t)\Big) \\ &= \mathbf{f}_2'(\phi(t))\,\phi'(t). \end{aligned}$$

Hence, since $\phi'(t) > 0$,

$$\frac{\mathbf{f}_1'(t)}{|\mathbf{f}_1'(t)|} = \frac{\mathbf{f}_2'(t)}{|\mathbf{f}_2'(t)|}.$$

□

DEFINITION 18.2.7. If $\mathbf{f}$ is a path in $\mathbb{R}^n$, then the *acceleration* at t is $\mathbf{f}''(t)$.

Example If $|\mathbf{f}'(t)|$ is constant (i.e. the "speed" is constant) then the velocity and the acceleration are orthogonal.

PROOF. Since $|\mathbf{f}(t)|^2 = \Big(\mathbf{f}'(t), \mathbf{f}'(y)\Big)$ is constant, we have from Proposition 18.2.4 that

$$\begin{aligned} 0 &= \frac{d}{dt}\Big(\mathbf{f}'(t), \mathbf{f}'(y)\Big) \\ &= 2\Big(\mathbf{f}''(t), \mathbf{f}'(y)\Big). \end{aligned}$$

This gives the result. □

18.2.1. Arc length. Suppose $\mathbf{f} : [a, b] \to \mathbb{R}^n$ is a path in $\mathbb{R}^n$. Let $a = t_1 < t_2 < \ldots < t_n = b$ be a partition of $[a, b]$, where $t_i - t_{i-1} = \delta t$ for all i.

We think of the length of the curve corresponding to $\mathbf{f}$ as being

$$\approx \sum_{i=2}^{N} |f(t_i) - f(t_{i-1})| = \sum_{i=2}^{N} \frac{|f(t_i) - f(t_{i-1})|}{\delta t}\delta t \approx \int_a^b \big|\mathbf{f}'(t)\big|\, dt. \tag{210}$$

See Figure 4.

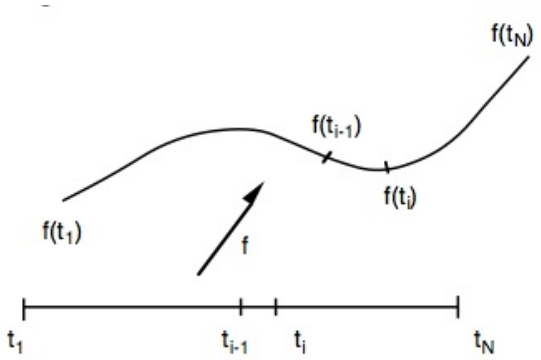

FIGURE 4. Diagram for (210).

Motivated by this we make the following definition.

DEFINITION 18.2.8. Let $\mathbf{f}:[a,b]\to\mathbb{R}^n$ be a path in $\mathbb{R}^n$. Then the *length* of the curve corresponding to $\mathbf{f}$ is given by

$$\int_a^b \left|\mathbf{f}'(t)\right| dt.$$

The next result shows that this definition is independent of the particular parametrisation chosen for the curve.

PROPOSITION 18.2.9. *Suppose* $\mathbf{f}_1:[a_1,b_1]\to\mathbb{R}^n$ *and* $\mathbf{f}_2:[a_2,b_2]\to\mathbb{R}^n$ *are equivalent parametrisations; and in particular* $\mathbf{f}_1=\mathbf{f}_2\circ\phi$ *where* $\phi:[a_1,b_1]\to[a_2,b_2]$, ϕ *is* C^1 *and* $\phi'(t)>0$ *for* $t\in I_1$. *Then*

$$\int_{a_1}^{b_1}\left|\mathbf{f}_1'(t)\right| dt=\int_{a_2}^{b_2}\left|\mathbf{f}_2'(s)\right| ds.$$

PROOF. From the chain rule and then the rule for change of variable of integration,

$$\int_{a_1}^{b_1}\left|\mathbf{f}_1'(t)\right| dt=\int_{a_1}^{b_1}\left|\mathbf{f}_2'(\phi(t))\right|\phi'(t)dt=\int_{a_2}^{b_2}\left|\mathbf{f}_2'(s)\right| ds.$$

□

18.3. Partial and Directional Derivatives

Analogous to Definitions 17.3.1 and 17.4.1 we have:

DEFINITION 18.3.1. The ith *partial derivative* of $\mathbf{f}$ at $\mathbf{x}$ is defined by

$$\frac{\partial\mathbf{f}}{\partial x^i}(\mathbf{x})\ \Big(\text{or } D_i\mathbf{f}(\mathbf{x})\Big)=\lim_{t\to0}\frac{\mathbf{f}(\mathbf{x}+t\mathbf{e}_i)-\mathbf{f}(\mathbf{x})}{t},$$

provided the limit exists. More generally, the *directional derivative* of $\mathbf{f}$ at $\mathbf{x}$ in the direction $\mathbf{v}$ is defined by

$$D_{\mathbf{v}}\mathbf{f}(\mathbf{x})=\lim_{t\to0}\frac{\mathbf{f}(\mathbf{x}+t\mathbf{v})-\mathbf{f}(\mathbf{x})}{t},$$

provided the limit exists.

Remarks See Figure 5.

(1) It follows immediately from the Definitions that

$$\frac{\partial\mathbf{f}}{\partial x^i}(\mathbf{x})=D_{\mathbf{e}_i}\mathbf{f}(\mathbf{x}).$$

(2) The partial and directional derivatives are *vectors in* $\mathbb{R}^n$. In the terminology of the previous section, $\dfrac{\partial \mathbf{f}}{\partial x^i}(\mathbf{x})$ is tangent to the path $t \mapsto \mathbf{f}(\mathbf{x} + t\mathbf{e}_i)$ and $D_{\mathbf{v}}\mathbf{f}(\mathbf{x})$ is tangent to the path $t \mapsto \mathbf{f}(\mathbf{x} + t\mathbf{v})$. Note that the curves corresponding to these paths are subsets of the image of $\mathbf{f}$.

(3) As we will discuss later, we may regard the partial derivatives at $\mathbf{x}$ as a basis for the tangent space to the image of $\mathbf{f}$ at $\mathbf{f}(\mathbf{x})$[3].

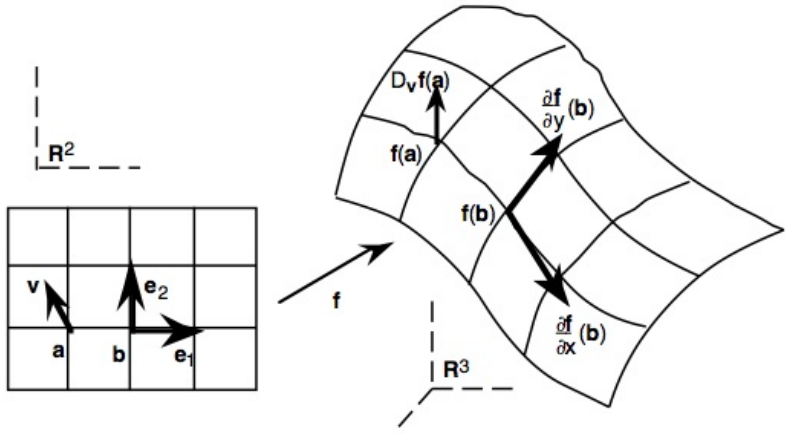

FIGURE 5. (Diagram is poor quality!) A rectangular grid in $\mathbb{R}^2$, its image under $f : \mathbb{R}^2 \to \mathbb{R}^3$, and a geometric representation of one directional derivative and two partial derivatives.

PROPOSITION 18.3.2. *If $f^1, \ldots, f^m$ are the component functions of $\mathbf{f}$ then*

$$\begin{aligned} \frac{\partial \mathbf{f}}{\partial x^i}(\mathbf{a}) &= \left(\frac{\partial f^1}{\partial x^i}(\mathbf{a}), \ldots, \frac{\partial f^m}{\partial x^i}(\mathbf{a}) \right) \quad \text{for } i = 1, \ldots, n \\ D_{\mathbf{v}}\mathbf{f}(\mathbf{a}) &= \left(D_{\mathbf{v}} f^1(\mathbf{a}), \ldots, D_{\mathbf{v}} f^m(\mathbf{a}) \right) \end{aligned}$$

in the sense that if one side of either equality exists, then so does the other, and both sides are then equal.

PROOF. Essentially the same as for the proof of Proposition 18.2.2. □

Example Let $f : \mathbb{R}^2 \to \mathbb{R}^3$ be given by

$$\mathbf{f}(x, y) = (x^2 - 2xy,\ x^2 + y^3,\ \sin x).$$

Then

$$\begin{aligned} \frac{\partial \mathbf{f}}{\partial x}(x, y) &= \left(\frac{\partial f^1}{\partial x}, \frac{\partial f^2}{\partial x}, \frac{\partial f^3}{\partial x} \right) = (2x - 2y,\ 2x,\ \cos x), \\ \frac{\partial \mathbf{f}}{\partial y}(x, y) &= \left(\frac{\partial f^1}{\partial y}, \frac{\partial f^2}{\partial y}, \frac{\partial f^3}{\partial y} \right) = (-2x,\ 3y^2,\ 0), \end{aligned}$$

are vectors in $\mathbb{R}^3$.

[3]More precisely, if $n \leq m$ and the differential $d\mathbf{f}(\mathbf{x})$ has rank n. See later.

18.4. The Differential

Analogous to Definition 17.5.1 we have:

DEFINITION 18.4.1. Suppose $\mathbf{f}: D\,(\subset \mathbb{R}^n) \to \mathbb{R}^n$. Then $\mathbf{f}$ is *differentiable* at $\mathbf{a} \in D$ if there is a linear transformation $\mathbf{L}:\mathbb{R}^n \to \mathbb{R}^n$ such that

$$\frac{\Big|\mathbf{f}(\mathbf{x}) - \Big(\mathbf{f}(\mathbf{a}) + \mathbf{L}(\mathbf{x}-\mathbf{a})\Big)\Big|}{|\mathbf{x}-\mathbf{a}|} \to 0 \quad \text{as } \mathbf{x} \to \mathbf{a}. \tag{211}$$

The linear transformation $\mathbf{L}$ is denoted by $\mathbf{f}'(\mathbf{a})$ or $d\mathbf{f}(\mathbf{a})$ and is called the *derivative* or *differential* of $\mathbf{f}$ at $\mathbf{a}$[4].

A vector-valued function is differentiable iff the corresponding component functions are differentiable. More precisely:

PROPOSITION 18.4.2. *$\mathbf{f}$ is differentiable at $\mathbf{a}$ iff $f^1, \dots, f^m$ are differentiable at $\mathbf{a}$. In this case the differential is given by*

$$\langle d\mathbf{f}(\mathbf{a}), \mathbf{v}\rangle = \Big(\langle df^1(\mathbf{a}), \mathbf{v}\rangle, \dots, \langle df^m(\mathbf{a}), \mathbf{v}\rangle\Big). \tag{212}$$

In particular, the differential is unique.

PROOF. For any linear map $\mathbf{L}: \mathbb{R}^n \to \mathbb{R}^n$, and for each $i = 1, \dots, m$, let $L^i:\mathbb{R}^n \to \mathbb{R}$ be the linear map defined by $L^i(\mathbf{v}) = \Big(\mathbf{L}(\mathbf{v})\Big)^i$.

From Theorem 10.4.4 it follows

$$\frac{\Big|\mathbf{f}(\mathbf{x}) - \Big(\mathbf{f}(\mathbf{a}) + \mathbf{L}(\mathbf{x}-\mathbf{a})\Big)\Big|}{|\mathbf{x}-\mathbf{a}|} \to 0 \quad \text{as } \mathbf{x} \to \mathbf{a}$$

iff

$$\frac{\Big|f^i(\mathbf{x}) - \Big(f^i(\mathbf{a}) + L^i(\mathbf{x}-\mathbf{a})\Big)\Big|}{|\mathbf{x}-\mathbf{a}|} \to 0 \quad \text{as } \mathbf{x} \to \mathbf{a} \quad \text{for } i = 1, \dots, m.$$

Thus $\mathbf{f}$ is differentiable at $\mathbf{a}$ iff $f^1, \dots, f^m$ are differentiable at $\mathbf{a}$.

In this case we must have

$$L^i = df^i(\mathbf{a}) \quad i = 1, \dots, m$$

(by uniqueness of the differential for *real*-valued functions), and so

$$\mathbf{L}(\mathbf{v}) = \Big(\langle df^1(\mathbf{a}), \mathbf{v}\rangle, \dots, \langle df^m(\mathbf{a}), \mathbf{v}\rangle\Big).$$

But this says that the differential $d\mathbf{f}(\mathbf{a})$ is unique and is given by (212). □

COROLLARY 18.4.3. *If $\mathbf{f}$ is differentiable at $\mathbf{a}$ then the linear transformation $d\mathbf{f}(\mathbf{a})$ is represented by the matrix*

$$\begin{bmatrix} \dfrac{\partial f^1}{\partial x^1}(\mathbf{a}) & \cdots & \dfrac{\partial f^1}{\partial x^n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f^m}{\partial x^1}(\mathbf{a}) & \cdots & \dfrac{\partial f^m}{\partial x^n}(\mathbf{a}) \end{bmatrix} : \mathbb{R}^n \to \mathbb{R}^n \tag{213}$$

PROOF. The ith column of the matrix corresponding to $d\mathbf{f}(\mathbf{a})$ is the vector $\langle d\mathbf{f}(\mathbf{a}), \mathbf{e}_i\rangle$[5]. From Proposition 18.4.2 this is the *column* vector corresponding to

$$\Big(\langle df^1(\mathbf{a}), \mathbf{e}_i\rangle, \dots, \langle df^m(\mathbf{a}), \mathbf{e}_i\rangle\Big),$$

[4]It follows from Proposition 18.4.2 that if $\mathbf{L}$ exists then it is unique and is given by the right side of (212).

[5]For any linear transformation $\mathbf{L}:\mathbb{R}^n \to \mathbb{R}^m$, the ith column of the corresponding matrix is $\mathbf{L}(\mathbf{e}_i)$.

i.e. to

$$\Big(\frac{\partial f^1}{\partial x^i}(\mathbf{a}),\dots,\frac{\partial f^m}{\partial x^i}(\mathbf{a})\Big).$$

This proves the result. □

Remark The jth *column* is the vector in $\mathbb{R}^n$ corresponding to the partial derivative $\dfrac{\partial \mathbf{f}}{\partial x^j}(\mathbf{a})$. The ith *row* represents $df^i(\mathbf{a})$.

The following proposition is immediate.

PROPOSITION 18.4.4. *If* $\mathbf{f}$ *is differentiable at* $\mathbf{a}$ *then*

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{a}) + \langle d\mathbf{f}(\mathbf{a}), \mathbf{x}-\mathbf{a}\rangle + \psi(\mathbf{x}),$$

where $\psi(\mathbf{x}) = o(|\mathbf{x}-\mathbf{a}|)$.

Conversely, suppose

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{a}) + L(\mathbf{x}-\mathbf{a}) + \psi(\mathbf{x}),$$

where $L:\mathbb{R}^n \to \mathbb{R}^n$ *is linear and* $\psi(\mathbf{x}) = o(|\mathbf{x}-\mathbf{a}|)$. *Then* $\mathbf{f}$ *is differentiable at* $\mathbf{a}$ *and* $d\mathbf{f}(\mathbf{a}) = L$.

PROOF. As for Proposition 17.5.4. □

Thus as is the case for real-valued functions, the previous proposition implies $\mathbf{f}(\mathbf{a}) + \langle d\mathbf{f}(\mathbf{a}), \mathbf{x}-\mathbf{a}\rangle$ gives the best first order approximation to $\mathbf{f}(\mathbf{x})$ for $\mathbf{x}$ near $\mathbf{a}$.

Example Let $\mathbf{f}:\mathbb{R}^2 \to \mathbb{R}^2$ be given by

$$\mathbf{f}(x,y) = (x^2 - 2xy, x^2 + y^3).$$

Find the best first order approximation to $\mathbf{f}(\mathbf{x})$ for $\mathbf{x}$ near $(1,2)$.

Solution:

$$\begin{aligned}
\mathbf{f}(1,2) &= \begin{bmatrix} -3 \\ 9 \end{bmatrix}, \\
d\mathbf{f}(x,y) &= \begin{bmatrix} 2x-2y & -2x \\ 2x & 3y^2 \end{bmatrix}, \\
d\mathbf{f}(1,2) &= \begin{bmatrix} -2 & -2 \\ 2 & 12 \end{bmatrix}.
\end{aligned}$$

So the best first order approximation near $(1,2)$ is

$$\begin{aligned}
&\mathbf{f}(1,2) + \langle d\mathbf{f}(1,2), (x-1, y-2)\rangle \\
&\quad = \begin{bmatrix} -3 \\ 9 \end{bmatrix} + \begin{bmatrix} -2 & -2 \\ 2 & 12 \end{bmatrix} \begin{bmatrix} x-1 \\ y-2 \end{bmatrix} \\
&\quad = \begin{bmatrix} -3 - 2(x-1) - 4(y-2) \\ 9 + 2(x-1) + 12(y-2) \end{bmatrix} \\
&\quad = \begin{bmatrix} 7 - 2x - 4y \\ -17 + 2x + 12y \end{bmatrix}.
\end{aligned}$$

Alternatively, working with each component separately, the best first order approximation is

$$\begin{aligned}
&\Big(f^1(1,2) + \frac{\partial f^1}{\partial x}(1,2)(x-1) + \frac{\partial f^1}{\partial y}(1,2)(y-2), \\
&\qquad f^2(1,2) + \frac{\partial f^2}{\partial x}(1,2)(x-1) + \frac{\partial f^2}{\partial y}(y-2)\Big) \\
&= \Big(-3 - 2(x-1) - 4(y-2),\ 9 + 2(x-1) + 12(y-2)\Big) \\
&= \Big(7 - 2x - 4y,\ -17 + 2x + 12y\Big).
\end{aligned}$$

Remark One similarly obtains second and higher order approximations by using Taylor's formula for each component function.

PROPOSITION 18.4.5. *If* $\mathbf{f}, \mathbf{g} : D\,(\subset \mathbb{R}^n) \to \mathbb{R}^n$ *are differentiable at* $\mathbf{a} \in D$*, then so are* $\alpha\mathbf{f}$ *and* $\mathbf{f}+\mathbf{g}$*. Moreover,*

$$\begin{aligned} d(\alpha\mathbf{f})(\mathbf{a}) &= \alpha d\mathbf{f}(\mathbf{a}), \\ d(\mathbf{f}+\mathbf{g})(\mathbf{a}) &= df(\mathbf{a}) + dg(\mathbf{a}). \end{aligned}$$

PROOF. This is straightforward (*exercise*) from Proposition 18.4.4. □

The previous proposition corresponds to the fact that the partial derivatives for $\mathbf{f}+\mathbf{g}$ are the sum of the partial derivatives corresponding to $\mathbf{f}$ and $\mathbf{g}$ respectively. Similarly for $\alpha\mathbf{f}$.

Higher Derivatives We say $\mathbf{f} \in C^k(D)$ iff $f^1, \ldots, f^m \in C^k(D)$.

It follows from the corresponding results for the component functions that

(1) $\mathbf{f} \in C^1(D) \Rightarrow \mathbf{f}$ is differentiable in D;
(2) $C^0(D) \supset C^1(D) \supset C^2(D) \supset \ldots$.

18.5. The Chain Rule

Motivation The chain rule for the composition of functions of one variable says that

$$\frac{d}{dx} g\Big(f(x)\Big) = g'\Big(f(x)\Big)\, f'(x).$$

Or to use a more informal notation, if $g = g(f)$ and $f = f(x)$, then

$$\frac{dg}{dx} = \frac{dg}{df}\frac{df}{dx}.$$

This is generalised in the following theorem. The theorem says that the linear approximation to $\mathbf{g} \circ \mathbf{f}$ (computed at $\mathbf{x}$) is the composition of the linear approximation to $\mathbf{f}$ (computed at $\mathbf{x}$) followed by the linear approximation to $\mathbf{g}$ (computed at $\mathbf{f}(\mathbf{x})$).

A Little Linear Algebra Suppose $L : \mathbb{R}^n \to \mathbb{R}^n$ is a *linear* map. Then we define the *norm* of L by

$$||L|| = \max\{|L(x)| : |x| \le 1\}^{6}.$$

A simple result *(exercise)* is that

$$|L(x)| \le ||L||\,|x| \tag{214}$$

for *any* $x \in \mathbb{R}^n$.

It is also easy to check *(exercise)* that $||\cdot||$ *does* define a norm on the vector space of linear maps from $\mathbb{R}^n$ into $\mathbb{R}^n$.

THEOREM 18.5.1 (Chain Rule). *Suppose* $\mathbf{f} : D\,(\subset \mathbb{R}^n) \to \Omega\,(\subset \mathbb{R}^n)$ *and* $g : \Omega\,(\subset \mathbb{R}^n) \to \mathbb{R}^r$*. Suppose* $\mathbf{f}$ *is differentiable at* $\mathbf{x}$ *and* $\mathbf{g}$ *is differentiable at* $\mathbf{f}(\mathbf{x})$*. Then* $\mathbf{g} \circ \mathbf{f}$ *is differentiable at* $\mathbf{x}$ *and*

$$d(\mathbf{g} \circ \mathbf{f})(\mathbf{x}) = d\mathbf{g}(\mathbf{f}(\mathbf{x})) \circ d\mathbf{f}(\mathbf{x}). \tag{215}$$

Schematically:

$$\begin{array}{ccccc} & & \xrightarrow{\quad g\circ f \quad} & & \\ D\,(\subset \mathbb{R}^n) & \xrightarrow{f} & \Omega\,(\subset \mathbb{R}^n) & \xrightarrow{g} & \mathbb{R}^r \\ & & \xrightarrow{\quad d(g\circ f)(x) = dg(f(x)) \circ df(x) \quad} & & \\ \mathbb{R}^n & \xrightarrow{df(x)} & \mathbb{R}^n & \xrightarrow{dg(f(x))} & \mathbb{R}^r \end{array}$$

[6] Here $|x|$, $|L(x)|$ are the usual Euclidean norms on $\mathbb{R}^n$ and $\mathbb{R}^m$. Thus $||L||$ corresponds to the maximum value taken by L on the unit ball. The maximum value *is* achieved, as L is continuous and $\{x : |x| \le 1\}$ is compact.

Example To see how all this corresponds to other formulations of the chain rule, suppose we have the following:

$$\begin{array}{ccccc} \mathbb{R}^3 & \xrightarrow{f} & \mathbb{R}^2 & \xrightarrow{g} & \mathbb{R}^2 \\ (x,y,z) & & (u,v) & & (p,q) \end{array}$$

Thus coordinates in $\mathbb{R}^3$ are denoted by (x,y,z), coordinates in the first copy of $\mathbb{R}^2$ are denoted by (u,v) and coordinates in the second copy of $\mathbb{R}^2$ are denoted by (p,q).

The functions f and g can be written as follows:

$$\begin{aligned} f &: \quad u = u(x,y,z),\ v = v(x,y,z), \\ g &: \quad p = p(u,v),\ q = q(u,v). \end{aligned}$$

Thus we think of u and v as functions of x, y and z; and p and q as functions of u and v.

We can also represent p and q as functions of x, y and z via

$$p = p\Big(u(x,y,z),\, v(x,y,z)\Big),\ q = q\Big(u(x,y,z),\, v(x,y,z)\Big).$$

The usual version of the chain rule in terms of partial derivatives is:

$$\begin{aligned} \frac{\partial p}{\partial x} &= \frac{\partial p}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial p}{\partial v}\frac{\partial v}{\partial x} \\ \frac{\partial p}{\partial x} &= \frac{\partial p}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial p}{\partial v}\frac{\partial v}{\partial x} \\ &\vdots \\ \frac{\partial q}{\partial z} &= \frac{\partial q}{\partial u}\frac{\partial u}{\partial z} + \frac{\partial q}{\partial v}\frac{\partial v}{\partial z}. \end{aligned}$$

In the first equality, $\frac{\partial p}{\partial x}$ is evaluated at (x,y,z), $\frac{\partial p}{\partial u}$ and $\frac{\partial p}{\partial v}$ are evaluated at $\Big(u(x,y,z), v(x,y,z)\Big)$, and $\frac{\partial u}{\partial x}$ and $\frac{\partial v}{\partial x}$ are evaluated at (x,y,z). Similarly for the other equalities.

In terms of the matrices of partial derivatives:

$$\underbrace{\begin{bmatrix} \frac{\partial p}{\partial x} & \frac{\partial p}{\partial y} & \frac{\partial p}{\partial z} \\ \frac{\partial q}{\partial x} & \frac{\partial q}{\partial y} & \frac{\partial q}{\partial z} \end{bmatrix}}_{d(g\circ f)(\mathbf{x})} = \underbrace{\begin{bmatrix} \frac{\partial p}{\partial u} & \frac{\partial p}{\partial v} \\ \frac{\partial q}{\partial u} & \frac{\partial q}{\partial v} \end{bmatrix}}_{dg(f(\mathbf{x}))} \underbrace{\begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} & \frac{\partial u}{\partial z} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} & \frac{\partial v}{\partial z} \end{bmatrix}}_{df(\mathbf{x})},$$

where $\mathbf{x} = (x,y,z)$.

Proof of Chain Rule: We want to show

$$(\mathbf{f}\circ\mathbf{g})(\mathbf{a}+\mathbf{h}) = (\mathbf{f}\circ\mathbf{g})(\mathbf{a}) + L(\mathbf{h}) + o(|\mathbf{h}|), \tag{216}$$

where $L = d\mathbf{f}(\mathbf{g}(\mathbf{a}))\circ d\mathbf{g}(\mathbf{a})$.

Now

$$\begin{aligned}
(\mathbf{f}\circ\mathbf{g})(\mathbf{a}+\mathbf{h}) &= \mathbf{f}\Big(\mathbf{g}(\mathbf{a}+\mathbf{h})\Big)\\
&= \mathbf{f}\Big(\mathbf{g}(\mathbf{a})+\mathbf{g}(\mathbf{a}+\mathbf{h})-\mathbf{g}(\mathbf{a})\Big)\\
&= \mathbf{f}\Big(\mathbf{g}(\mathbf{a})\Big)+\Big\langle d\mathbf{f}\Big(\mathbf{g}(\mathbf{a})\Big),\, \mathbf{g}(\mathbf{a}+\mathbf{h})-\mathbf{g}(\mathbf{a})\Big\rangle\\
&\qquad +o\Big(|\mathbf{g}(\mathbf{a}+\mathbf{h})-\mathbf{g}(\mathbf{a})|\Big)\\
&\qquad\qquad \dots \text{by the differentiability of } \mathbf{f}\\
&= \mathbf{f}\Big(\mathbf{g}(\mathbf{a})\Big)+\Big\langle d\mathbf{f}\Big(\mathbf{g}(\mathbf{a})\Big),\, \langle d\mathbf{g}(\mathbf{a}),\mathbf{h}\rangle + o(|\mathbf{h}|)\Big\rangle\\
&\qquad +o\Big(|\mathbf{g}(\mathbf{a}+\mathbf{h})-\mathbf{g}(\mathbf{a})|\Big)\\
&\qquad\qquad \dots \text{by the differentiability of } \mathbf{g}\\
&= \mathbf{f}\Big(\mathbf{g}(\mathbf{a})\Big)+\Big\langle d\mathbf{f}\Big(\mathbf{g}(\mathbf{a})\Big),\, \langle d\mathbf{g}(\mathbf{a}),\mathbf{h}\rangle\Big\rangle\\
&\qquad +\Big\langle d\mathbf{f}\Big(\mathbf{g}(\mathbf{a})\Big),\, o(|\mathbf{h}|)\Big\rangle + o\Big(|\mathbf{g}(\mathbf{a}+\mathbf{h})-\mathbf{g}(\mathbf{a})|\Big)\\
&= A+B+C+D
\end{aligned}$$

But $B = \Big\langle d\mathbf{f}\Big(\mathbf{g}(\mathbf{a})\Big)\circ d\mathbf{g}(\mathbf{a}),\, \mathbf{h}\Big\rangle$, by definition of the "composition" of two maps. Also $C = o(|\mathbf{h}|)$ from (214) (*exercise*). Finally, for D we have

$$\begin{aligned}
\Big|\mathbf{g}(\mathbf{a}+\mathbf{h})-\mathbf{g}(\mathbf{a})\Big| &= \Big|\langle d\mathbf{g}(\mathbf{a}),\mathbf{h}\rangle + o(|\mathbf{h}|)\Big| \dots \text{by differentiability of } \mathbf{g}\\
&\le ||d\mathbf{g}(\mathbf{a})||\,|\mathbf{h}| + o(|\mathbf{h}|) \dots \text{from (214)}\\
&= O(|\mathbf{h}|) \dots \textit{why?}
\end{aligned}$$

Substituting the above expressions into $A+B+C+D$, we get

$$(\mathbf{f}\circ\mathbf{g})(\mathbf{a}+\mathbf{h}) = \mathbf{f}\Big(\mathbf{g}(\mathbf{a})\Big)+\Big\langle d\mathbf{f}\Big(\mathbf{g}(\mathbf{a})\Big)\Big)\circ d\mathbf{g}(\mathbf{a}),\, \mathbf{h}\Big\rangle + o(|\mathbf{h}|). \tag{217}$$

If follows that $\mathbf{f}\circ\mathbf{g}$ *is* differentiable at $\mathbf{a}$, *and* moreover the differential equals $d\mathbf{f}(\mathbf{g}(\mathbf{a}))\circ d\mathbf{g}(\mathbf{a})$. This proves the theorem. ∎

CHAPTER 19

The Inverse Function Theorem and its Applications

19.1. Inverse Function Theorem

Motivation

(1) Suppose

$$f : \Omega\,(\subset \mathbb{R}^n) \to \mathbb{R}^n$$

and f is C^1. Note that the dimension of the domain and the range are the same. Suppose $f(x_0) = y_0$. Then a good approximation to $f(x)$ for x near x_0 is gven by

$$x \mapsto f(x_0) + \langle f'(x_0), x - x_0 \rangle. \tag{218}$$

See Figure 1.

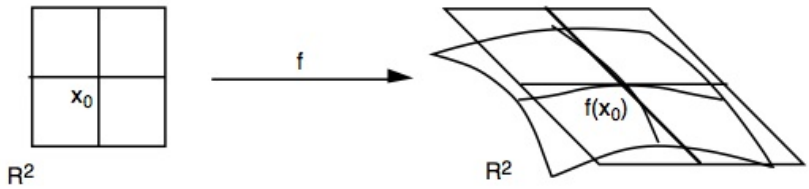

FIGURE 1. The right curved grid is the image of the left grid by the function$f$. The right flat grid is the image of the left grid by the linear linear approximation $x \mapsto f(x_0) + \langle f'(x_0), x - x_0 \rangle$.

We expect that if $f'(x_0)$ is a one-one and onto linear map, (which is the same as $\det f'(x_0) \neq 0$ and which implies the map in (218) is one-one and onto), then f should be one-one and onto near x_0. This is true, and is called the Inverse Function Theorem.

(2) Consider the set of equations

$$\begin{aligned} f^1(x^1, \ldots, x^n) &= y^1 \\ f^2(x^1, \ldots, x^n) &= y^2 \\ &\vdots \\ f^n(x^1, \ldots, x^n) &= y^n, \end{aligned}$$

where $f^1, \dots, f^n$ are certain real-valued functions. Suppose that these equations are satisfied if $(x^1, \dots, x^n) = (x_0^1, \dots, x_0^n)$ and $(y^1, \dots, y^n) = (y_0^1, \dots, y_0^n)$, and that $\det f'(x_0) \neq 0$. Then it follows from the Inverse Function Theorem that for all $(y^1, \dots, y^n)$ in some ball centred at $(y_0^1, \dots, y_0^n)$ the equations have a unique solution $(x^1, \dots, x^n)$ in some ball centred at $(x_0^1, \dots, x_0^n)$.

THEOREM 19.1.1 (Inverse Function Theorem). *Suppose $f : \Omega\,(\subset \mathbb{R}^n) \to \mathbb{R}^n$ is C^1 and Ω is open*[1]*. Suppose $f'(x_0)$ is invertible*[2] *for some $x_0 \in \Omega$.*

Then there exists an open set $U \ni x_0$ and an open set $V \ni f(x_0)$ such that

(1) *$f'(x)$ is invertible at every $x \in U$,*
(2) *$f : U \to V$ is one-one and onto, and hence has an inverse $g : V \to U$,*
(3) *g is C^1 and $g'(f(x)) = [f'(x)]^{-1}$ for every $x \in U$.*

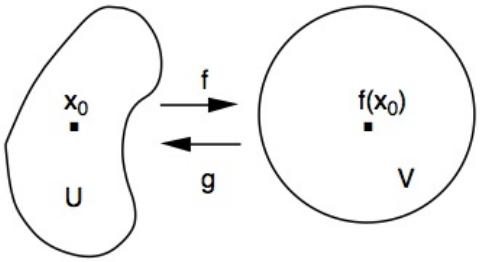

FIGURE 2. Diagram for Theorem 19.1.1.

PROOF. See Figure 3 for Steps 1 and 2, and Figure —4 for the latter part of Step 2.

Step 1 Suppose

$$y^* \in B_\delta(f(x_0)).$$

We will choose δ later. (We will take the set V in the theorem to be the open set $B_\delta(f(x_0))$)

For each such y, we want to prove the existence of x $(= x^*$, say) such that

$$f(x) = y^*. \tag{219}$$

We write $f(x)$ as a first order function plus an error term.

Look at Figure 3. Reformulating (219), we want to solve (for x)

$$f(x_0) + \langle f'(x_0), x - x_0 \rangle + R(x) = y^*, \tag{220}$$

where

$$R(x) := f(x) - f(x_0) - \langle f'(x_0), x - x_0 \rangle. \tag{221}$$

In other words, we want to find x such that

$$\langle f'(x_0), x - x_0 \rangle = y^* - f(x_0) - R(x),$$

i.e. such that

$$x = x_0 + \left\langle [f'(x_0)]^{-1}, y^* - f(x_0) \right\rangle - \left\langle [f'(x_0)]^{-1}, R(x) \right\rangle \tag{222}$$

[1]Note that the dimensions of the domain and range are equal.

[2]That is, the matrix $f'(x_0)$ is one-one and onto, or equivalently, $\det f'(x_0) \neq 0$.

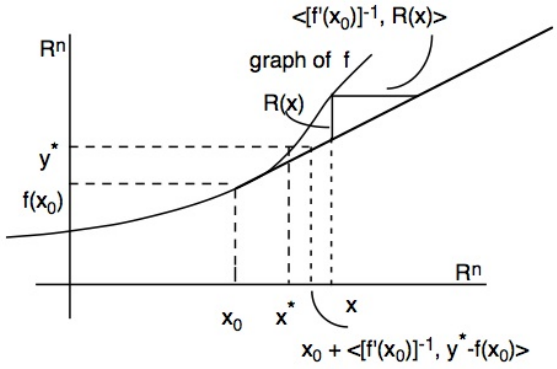

FIGURE 3. Diagram for Steps 1 and 2.

(*why*?).

The right side of (222) is the sum of two terms. The first term, that is $x_0 + \left\langle [f'(x_0)]^{-1}, y^* - f(x_0) \right\rangle$, is the solution of the linear equation $y^* = f(x_0) + \langle f'(x_0), x - x_0 \rangle$. The second term is the error term $-\left\langle [f'(x_0)]^{-1}, R(x) \right\rangle$, which is $o(|x - x_0|)$ because $R(x)$ is $o(|x - x_0|)$ and $[f'(x_0)]^{-1}$ is a fixed linear map.

Step 2 Because of (222) define

$$A_{y^*}(x) := x_0 + \left\langle [f'(x_0)]^{-1}, y^* - f(x_0) \right\rangle - \left\langle [f'(x_0)]^{-1}, R(x) \right\rangle . \tag{223}$$

Note that x is a fixed point of A_{y^*} iff x satisfies (222) and hence solves (219).

We *claim* that

$$A_{y^*} : \overline{B}_\epsilon(x_0) \to B_\epsilon(x_0), \tag{224}$$

and that A_{y^*} *is a contraction map*, provided $\epsilon > 0$ is sufficiently small (ϵ will depend only on x_0 and f) and provided $y^* \in B_\delta(y_0)$ (where $\delta > 0$ also depends only on x_0 and f).

To *prove* the claim, we compute

$$A_{y^*}(x_1) - A_{y^*}(x_2) = \left\langle [f'(x_0)]^{-1}, R(x_2) - R(x_1) \right\rangle ,$$

and so

$$|A_{y^*}(x_1) - A_{y^*}(x_2)| \leq K \, |R(x_1) - R(x_2)|, \tag{225}$$

where

$$K := \left\| [f'(x_0)]^{-1} \right\| . \tag{226}$$

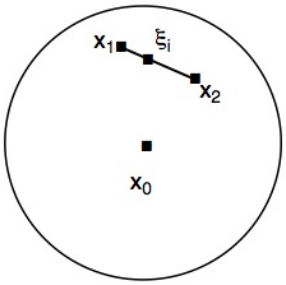

FIGURE 4. A few of the items in Step 2.

From (221), and perhaps looking at Figure 4,

$$R(x_2) - R(x_1) = f(x_2) - f(x_1) - \langle f'(x_0), x_2 - x_1 \rangle .$$

We apply the mean value theorem (17.9.1) to each of the *components* of this equation to obtain

$$\begin{aligned} \left|R^i(x_2) - R^i(x_1)\right| &= \left|\left\langle f^{i\prime}(\xi_i), x_2 - x_1 \right\rangle - \left\langle f^{i\prime}(x_0), x_2 - x_1 \right\rangle\right| \\ &\quad \text{for } i = 1, \ldots, n \text{ and some } \xi_i \in \mathbb{R}^n \text{ between } x_1 \text{ and } x_2 \\ &= \left|\left\langle f^{i\prime}(\xi_i) - f^{i\prime}(x_0), x_2 - x_1 \right\rangle\right| \\ &\leq \left|f^{i\prime}(\xi_i) - f^{i\prime}(x_0)\right| \, |x_2 - x_1|, \end{aligned}$$

by Cauchy-Schwartz, treating $f^{i\prime}$ as a "row vector".

By the continuity of the derivatives of f, it follows

$$|R(x_2) - R(x_1)| \leq \frac{1}{2K}|x_2 - x_1|, \tag{227}$$

provided $x_1, x_2 \in \overline{B}_\epsilon(x_0)$ for some $\epsilon > 0$ depending only on f and x_0. Hence from (225)

$$|A_{y^*}(x_1) - A_{y^*}(x_2)| \leq \frac{1}{2}|x_1 - x_2|. \tag{228}$$

This proves

$$A_{y^*} : \overline{B}_\epsilon(x_0) \to \mathbb{R}^n$$

is a contraction map, but we still need to prove (224).

For this we compute

$$\begin{aligned} |A_{y^*}(x) - x_0| &\leq \left|\left\langle [f'(x_0)]^{-1}, y^* - f(x_0) \right\rangle\right| + \left|\left\langle [f'(x_0)]^{-1}, R(x) \right\rangle\right| \text{ from (223)} \\ &\leq K|y^* - f(x_0)| + K|R(x)| \\ &= K|y^* - f(x_0)| + K|R(x) - R(x_0)| \qquad \text{as } R(x_0) = 0 \\ &\leq K|y^* - f(x_0)| + \frac{1}{2}|x - x_0| \qquad \text{from (227)} \\ &< \epsilon/2 + \epsilon/2 = \epsilon, \end{aligned}$$

provided $x \in \overline{B}_\epsilon(x_0)$ and $y^* \in B_\delta(f(x_0))$ (if $K\delta < \epsilon$). This establishes (224) and completes the proof of the claim.

Step 3 We now know that for each $y \in B_\delta(f(x_0))$ there is a unique $x \in B_\epsilon(x_0)$ such that $f(x) = y$. Denote this x by $g(y)$. Thus

$$g : B_\delta(f(x_0)) \to B_\epsilon(x_0).$$

We *claim* that this inverse function g is continuous.

To see this let $x_i = g(y_i)$ for $i = 1, 2$. That is, $f(x_i) = y_i$, or equivalently $x_i = A_{y_i}(x_i)$ (recall the remark after (223)). Then

$$\begin{aligned}
|g(y_1) - g(y_2)| &= |x_1 - x_2| \\
&= |A_{y_1}(x_1) - A_{y_2 1}(x_2)| \\
&\leq \left|\left\langle [f'(x_0)]^{-1}, y_1 - y_2 \right\rangle\right| + \left|\left\langle [f'(x_0)]^{-1}, R(x_1) - R(x_2) \right\rangle\right| \text{ by (223)} \\
&\leq K\,|y_1 - y_2| + K\,|R(x_1) - R(x_2)| \qquad \text{from (225)} \\
&\leq K\,|y_1 - y_2| + K\frac{1}{2K}|x_1 - x_2| \qquad \text{from (227)} \\
&= K\,|y_1 - y_2| + \frac{1}{2}|g(y_1) - g(y_2)|.
\end{aligned}$$

Thus

$$\frac{1}{2}|g(y_1) - g(y_2)| \leq K\,|y_1 - y_2|,$$

and so

$$|g(y_1) - g(y_2)| \leq 2K\,|y_1 - y_2|. \tag{229}$$

In particular, g is Lipschitz and hence continuous.

Step 4 Let

$$V = B_\delta(f(x_0)), \quad U = g\left[B_\delta(f(x_0))\right].$$

Since $U = B_\epsilon(x_0) \cap f^{-1}[V]$ (*why?*), it follows U is open. We have thus proved the second part of the theorem.

The first part of the theorem is easy. All we need do is first replace Ω by a smaller open set containing x_0 in which $f'(x)$ is invertible for all x. This is possible as $\det f'(x_0) \neq 0$ and the entries in the matrix $f'(x)$ are continuous.

Step 5 We *claim* g is C^1 on V and

$$g'(f(x)) = [f'(x)]^{-1}. \tag{230}$$

To see that g is differentiable at $y \in V$ and (230) is true, suppose $y, \overline{y} \in V$, and let $f(x) = y$, $f(\overline{x}) = \overline{y}$ where $x, \overline{x} \in U$. Then

$$\begin{aligned}
&\frac{\left|g(\overline{y}) - g(y) - \left\langle [f'(x)]^{-1}, \overline{y} - y \right\rangle\right|}{|\overline{y} - y|} \\
= &\frac{\left|\overline{x} - x - \left\langle [f'(x)]^{-1}, f(\overline{x}) - f(x) \right\rangle\right|}{|\overline{y} - y|} \\
= &\frac{\left|\left\langle [f'(x)]^{-1}, \langle f'(x), \overline{x} - x\rangle - f(\overline{x}) + f(x) \right\rangle\right|}{|\overline{y} - y|} \\
\leq &\left\|[f'(x)]^{-1}\right\| \frac{|f(\overline{x}) - f(x) - \langle f'(x), \overline{x} - x\rangle|}{|\overline{x} - x|} \frac{|\overline{x} - x|}{|\overline{y} - y|}.
\end{aligned}$$

If we fix y and let $\overline{y} \to y$, then x is fixed and $\overline{x} \to x$. Hence the last line in the previous series of inequalities $\to 0$, since f is differentiable at x and $|\overline{x} - x|/|\overline{y} - y| \leq K/2$ by (229). Hence g is differentiable at y and the derivative is given by (230).

The fact that g is C^1 follows from (230) and the expression for the inverse of a matrix. □

Remark We have

$$\begin{aligned}
g'(y) &= [f'(g(y))]^{-1} \\
&= \frac{\operatorname{Ad}[f'(g(y))]}{\det[f'(g(y))]},
\end{aligned} \tag{231}$$

where $\operatorname{Ad}[f'(g(y))]$ is the matrix of cofactors of the matrix $[f'(g(y))]$.

If f is C^2, then since we already know g is C^1, it follows that the terms in the matrix (231) are algebraic combinations of C^1 functions and so are C^1. Hence the terms in the matrix g' are C^1 and so g is C^2.

Similarly, if f is C^3 then since g is C^2 it follows the terms in the matrix (231) are C^2 and so g is C^3.

By induction we have the following Corollary.

COROLLARY 19.1.2. *If in the Inverse Function Theorem the function f is C^k then the local inverse function g is also C^k.*

Summary of Proof of Theorem

(1) Write the equation $f(x^*) = y$ as a perturbation of the first order equation obtained by linearising around x_0. See (220) and (221).

Write the solution x as the solution $T(y^*)$ of the linear equation plus an error term $E(x)$,

$$x = T(y^*) + E(x) =: A_{y^*}(x)$$

See (222).

(2) Show $A_{y^*}(x)$ is a contraction map on $B_\epsilon(x_0)$ (for ϵ sufficiently small and y^* near y_0) and hence has a fixed point. It follows that for all y^* near y_0 there exists a unique x^* near x_0 such that $f(x^*) = y^*$. Write $g(y^*) = x^*$.

(3) The local inverse function g is close to the inverse $T(y^*)$ of the linear function. Use this to prove that g is Lipschitz continuous.

(4) Wrap up the proof of parts 1 and 2 of the theorem.

(5) Write out the difference quotient for the derivative of g and use this and the differentiability of f to show g is differentiable.

19.2. Implicit Function Theorem

Motivation We can write the equations in the previous "Motivation" section as

$$f(x) = y,$$

where $x = (x^1, \ldots, x^n)$ and $y = (y^1, \ldots, y^n)$.

More generally we may have n equations

$$f(x, u) = y,$$

i.e.,

$$\begin{aligned} f^1(x^1, \ldots, x^n, u^1, \ldots, u^m) &= y^1 \\ f^2(x^1, \ldots, x^n, u^1, \ldots, u^m) &= y^2 \\ &\vdots \\ f^n(x^1, \ldots, x^n, u^1, \ldots, u^m) &= y^n, \end{aligned}$$

where we regard the $u = (u^1, \ldots, u^m)$ as *parameters*.

Write

$$\det\left[\frac{\partial f}{\partial x}\right] := \det\begin{bmatrix} \frac{\partial f^1}{\partial x^1} & \cdots & \frac{\partial f^1}{\partial x^n} \\ \vdots & & \vdots \\ \frac{\partial f^n}{\partial x^1} & \cdots & \frac{\partial f^n}{\partial x^n} \end{bmatrix}.$$

Thus $\det[\partial f/\partial x]$ is the determinant of the derivative of the map $f(x^1, \ldots, x^n)$, where $x^1, \ldots, x^m$ are taken as the variables and the $u^1, \ldots, u^m$ are taken to be *fixed*.

Now suppose that

$$f(x_0, u_0) = y_0, \qquad \det\left[\frac{\partial f}{\partial x}\right]_{(x_0, u_0)} \neq 0.$$

From the Inverse Function Theorem (still thinking of $u^1, \ldots, u^m$ as fixed), for y near y_0 there exists a unique x near x_0 such that

$$f(x, u_0) = y.$$

The *Implicit* Function Theorem says more generally that for y near y_0 *and* for u near u_0, there exists a unique x near x_0 such that

$$f(x, u) = y.$$

In *applications* we will usually take $y = y_0 = c$(say) to be *fixed*. Thus we consider an equation

$$f(x, u) = c \tag{232}$$

where

$$f(x_0, u_0) = c, \qquad \det \left[\frac{\partial f}{\partial x}\right]_{(x_0,u_0)} \neq 0.$$

Hence for u near u_0 there exists a unique $x = x(u)$ near x_0 such that

$$f(x(u), u) = c. \tag{233}$$

In words, suppose we have n equations involving n unknowns x and certain parameters u. Suppose the equations are satisfied at (x_0, u_0) and suppose that the determinant of the matrix of derivatives with respect to the x variables is non-zero at (x_0, u_0). Then the equations can be solved for $x = x(u)$ if u is near u_0.

Moreover, differentiating the ith equation in (233) with respect to u^j we obtain

$$\sum_k \frac{\partial f^i}{\partial x_k}\frac{\partial x^k}{\partial u^j} + \frac{\partial f^i}{\partial u_j} = 0.$$

That is

$$\left[\frac{\partial f}{\partial x}\right]\left[\frac{\partial x}{\partial u}\right] + \left[\frac{\partial f}{\partial u}\right] = [0],$$

where the first three matrices are $n \times n$, $n \times m$, and $n \times m$ respectively, and the last matrix is the $n \times m$ zero matrix. Since $\det [\partial f/\partial x]_{(x_0,u_0)} \neq 0$, it follows

$$\left[\frac{\partial x}{\partial u}\right]_{u_0} = -\left[\frac{\partial f}{\partial x}\right]^{-1}_{(x_0,u_0)}\left[\frac{\partial f}{\partial u}\right]_{(x_0,u_0)}. \tag{234}$$

Example 1 Consider the circle in $\mathbb{R}^2$ described by

$$x^2 + y^2 = 1.$$

Write

$$F(x, y) = 1. \tag{235}$$

Thus in (232), u is replaced by y and c is replaced by 1.

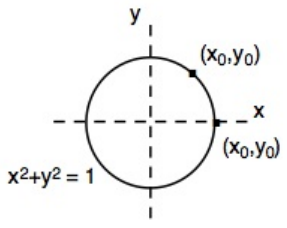

FIGURE 5. Diagram for Example 1.

Suppose $F(x_0, y_0) = 1$ and $\partial F/\partial x_0|_{(x_0,y_0)} \neq 0$ (i.e. $x_0 \neq 0$). Then for y near y_0 there is a unique x near x_0 satisfying (235). In fact $x = \pm\sqrt{1-y^2}$ according as $x_0 > 0$ or $x_0 < 0$. See the diagram for two examples of such points (x_0, y_0).

Similarly, if $\partial F/\partial y_0|_{(x_0,y_0)} \neq 0$, i.e. $y_0 \neq 0$, Then for x near x_0 there is a unique y near y_0 satisfying (235).

Example 2 Suppose a "surface" in $\mathbb{R}^3$ is described by

$$\Phi(x, y, z) = 0. \tag{236}$$

Suppose $\Phi(x_0, y_0, z_0) = 0$ and $\partial\Phi/\partial z\,(x_0, y_0, z_0) \neq 0$.

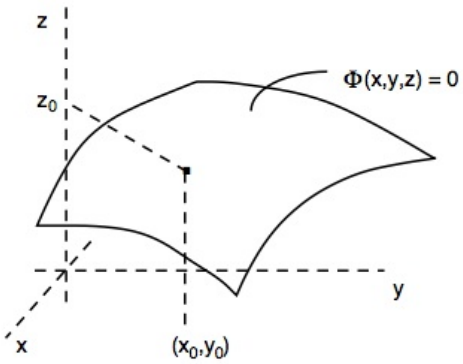

FIGURE 6. Diagram for Example 2.

Then by the Implicit Function Theorem, for (x, y) near (x_0, y_0) there is a unique z near z_0 such that $\Phi(x, y, z) = 0$. Thus the "surface" can locally[3] be written as a graph over the x-y plane

More generally, if $\nabla\Phi(x_0, y_0, z_0) \neq 0$ then at least one of the derivatives $\partial\Phi/\partial x\,(x_0, y_0, z_0)$, $\partial\Phi/\partial y\,(x_0, y_0, z_0)$ or $\partial\Phi/\partial z\,(x_0, y_0, z_0)$ does not equal 0. The corresponding variable x, y or z can then be solved for in terms of the other two variables and the surface is locally a graph over the plane corresponding to these two other variables.

Example 3 Suppose a "curve" in $\mathbb{R}^3$ is described by

$$\begin{aligned} \Phi(x, y, z) &= 0, \\ \Psi(x, y, z) &= 0. \end{aligned}$$

Suppose (x_0, y_0, z_0) lies on the curve, i.e. $\Phi(x_0, y_0, z_0) = \Psi(x_0, y_0, z_0) = 0$. Suppose moreover that the matrix

$$\left[\begin{array}{ccc} \frac{\partial\Phi}{\partial x} & \frac{\partial\Phi}{\partial y} & \frac{\partial\Phi}{\partial z} \\ \frac{\partial\Psi}{\partial x} & \frac{\partial\Psi}{\partial y} & \frac{\partial\Psi}{\partial z} \end{array}\right]_{(x_0,y_0,z_0)}$$

has rank 2. In other words, two of the three columns must be linearly independent. Suppose it is the first two. Then

$$\det \left|\begin{array}{cc} \frac{\partial\Phi}{\partial x} & \frac{\partial\Phi}{\partial y} \\ \frac{\partial\Psi}{\partial x} & \frac{\partial\Psi}{\partial y} \end{array}\right|_{(x_0,y_0,z_0)} \neq 0.$$

By the Implicit Function Theorem, we can solve for (x, y) near (x_0, y_0) in terms of z near z_0. In other words we can locally write the curve as a graph over the z axis.

[3]By "locally" we mean in some $B_r(a)$ for each point a in the surface.

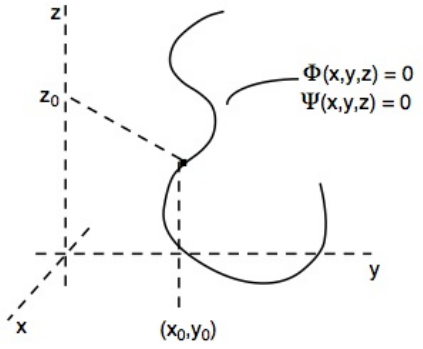

FIGURE 7. Diagram for Example 3.

Example 4 Consider the equations

$$\begin{aligned} f_1(x_1,x_2,y_1,y_2,y_3) &= 2e^{x_1}+x_2y_1-4y_2+3 \\ f_2(x_1,x_2,y_1,y_2,y_3) &= x_2\cos x_1-6x_1+2y_1-y_3. \end{aligned}$$

Consider the "three dimensional surface in $\mathbb{R}^5$ " given by $f_1(x_1,x_2,y_1,y_2,y_3)=0$, $f_2(x_1,x_2,y_1,y_2,y_3)=0$ [4]. We easily check that

$$f(0,1,3,2,7)=0$$

and

$$f'(0,1,3,2,7)=\begin{bmatrix} 2 & 3 & 1 & -4 & 0 \\ -6 & 1 & 2 & 0 & -1 \end{bmatrix}.$$

The first two columns are linearly independent and so we can solve for x_1,x_2 in terms of y_1,y_2,y_3 near $(3,2,7)$.

Moreover, from (234) we have

$$\begin{aligned} \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \frac{\partial x_1}{\partial y_3} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \frac{\partial x_2}{\partial y_3} \end{bmatrix}_{(3,2,7)} &= -\begin{bmatrix} 2 & 3 \\ -6 & 1 \end{bmatrix}^{-1}\begin{bmatrix} 1 & -4 & 0 \\ 2 & 0 & -1 \end{bmatrix} \\ &= -\frac{1}{20}\begin{bmatrix} 1 & -3 \\ 6 & 2 \end{bmatrix}\begin{bmatrix} 1 & -4 & 0 \\ 2 & 0 & -1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{4} & \frac{1}{5} & -\frac{3}{20} \\ -\frac{1}{2} & \frac{6}{5} & \frac{1}{10} \end{bmatrix} \end{aligned}$$

It folows that for (y_1,y_2,y_3) near $(3,2,7)$ we have

$$\begin{aligned} x_1 &\approx 0+\frac{1}{4}(y_1-3)+\frac{1}{5}(y_2-2)-\frac{3}{20}(y_3-7) \\ x_2 &\approx 1-\frac{1}{2}(y_1-3)+\frac{6}{5}(y_2-2)+\frac{1}{10}(y_3-7). \end{aligned}$$

We now give a precise statement and proof of the Implicit Function Theorem.

THEOREM 19.2.1 (Implicit function Theorem). *Suppose $f:D\,(\subset \mathbb{R}^n\times\mathbb{R}^k)\to\mathbb{R}^n$ is C^1 and D is open. Suppose $f(x_0,u_0)=y_0$ where $x_0\in\mathbb{R}^n$ and $u_0\in\mathbb{R}^m$. Suppose* $\det\left[\partial f/\partial x\right]|_{(x_0,u_0)}\neq 0$.

Then there exist $\epsilon,\delta>0$ such that for all $y\in B_\delta(y_0)$ and all $u\in B_\delta(u_0)$ there is a unique $x\in B_\epsilon(x_0)$ such that

$$f(x,u)=y.$$

[4] One constraint gives a four dimensional surface, two constraints give a three dimensional surface, etc. Each further constraint reduces the dimension by one.

If we denote this x by $g(u,y)$ then g is C^1. Moreover,

$$\left[\frac{\partial g}{\partial u}\right]_{(u_0,y_0)} = -\left[\frac{\partial f}{\partial x}\right]^{-1}_{(x_0,u_0)}\left[\frac{\partial f}{\partial u}\right]_{(x_0,u_0)}.$$

PROOF. Define

$$F : D \to \mathbb{R}^n \times \mathbb{R}^m$$

by

$$F(x,u) = \Big(f(x,u), u\Big).$$

Then clearly[5] F is C^1 and

$$\det F'|_{(x_0,u_0)} = \det\left[\frac{\partial f}{\partial x}\right]_{(x_0,u_0)}.$$

Also

$$F(x_0,u_0) = (y_0,u_0).$$

From the Inverse Function Theorem, for all (y,u) near (y_0,u_0) there exists a unique (x,w) near (x_0,u_0) such that

$$F(x,w) = (y,u). \tag{237}$$

Moreover, x and w are C^1 functions of (y,u). But from the definition of F it follows that (237) holds iff $w = u$ and $f(x,u) = y$. Hence for all (y,u) near (y_0,u_0) there exists a unique $x = g(u,y)$ near x_0 such that

$$f(x,u) = y. \tag{238}$$

Moreover, g is a C^1 function of (u,y).

The expression for $\left[\frac{\partial g}{\partial u}\right]_{(u_0,y_0)}$ follows from differentiating (238) precisely as in the derivation of (234). □

19.3. Manifolds

Discussion Loosely speaking, M is a k-dimensional manifold in $\mathbb{R}^n$ if M locally[6] looks like the graph of a function of k variables. Thus a 2-dimensional manifold is a surface and a 1-dimensional manifold is a curve.

We will give three different ways to define a manifold and show that they are equivalent.

We begin by considering manifolds of dimension $n-1$ in $\mathbb{R}^n$ (e.g. a curve in $\mathbb{R}^2$ or a surface in $\mathbb{R}^3$). Such a manifold is said to have *codimension one*.

Suppose

$$\Phi : \mathbb{R}^n \to \mathbb{R}$$

is C^1. Let

$$M = \{x : \Phi(x) = 0\}.$$

See Examples 1 and 2 in Section 19.2 (where $\Phi(x,y) = F(x,y) - 1$ in Example 1).

If $\nabla\Phi(a) \neq 0$ for some $a \in M$, then as in Examples 1 and 2 we can write M locally as the graph of a function of one of the variables x_i in terms of the remaining $n-1$ variables.

This leads to the following definition.

---

[5] Since

$$F' = \begin{bmatrix} \left[\frac{\partial f}{\partial x}\right] & \left[\frac{\partial f}{\partial u}\right] \\ O & I \end{bmatrix},$$

where O is the $m \times n$ zero matrix and I is the $m \times m$ identity matrix.

[6] "Locally" means in some neighbourhood for each $a \in M$.

DEFINITION 19.3.1. [**Manifolds as Level Sets**] Suppose $M \subset \mathbb{R}^n$ and for each $a \in M$ there exists $r > 0$ and a C^1 function $\Phi : B_r(a) \to \mathbb{R}$ such that

$$M \cap B_r(a) = \{x : \Phi(x) = 0\}.$$

Suppose also that $\nabla\Phi(x) \neq 0$ for each $x \in B_r(a)$.

Then M is an $n-1$ *dimensional manifold in* $\mathbb{R}^n$. We say M has *codimension one*.

The one dimensional space spanned by $\nabla\Phi(a)$ is called the *normal space to* M *at* a and is denoted by N_aM [7].

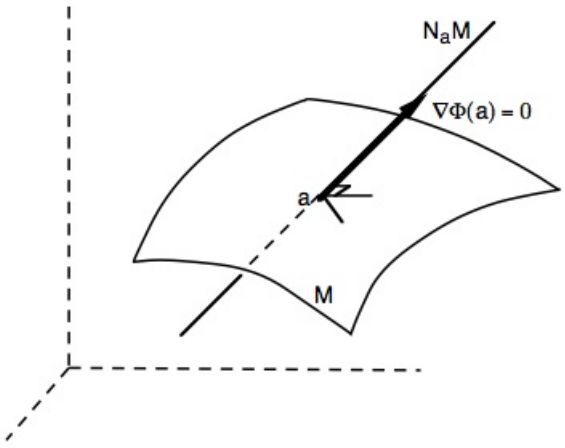

FIGURE 8. The manifold M is the level set $\{x : \Phi(x) = 0\}$. The equality in the diagram should read "$\nabla\Phi(a) \neq 0$".

Remarks

(1) Usually M is described by a *single* function Φ defined on $\mathbb{R}^n$
(2) See Section 17.6 for a discussion of $\nabla\Phi(a)$ which motivates the definition of N_aM.
(3) With the same proof as in Examples 1 and 2 from Section 19.2, we can locally write M as the graph of a function

$$x_i = f(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

for some $1 \leq i \leq n$.

Higher Codimension Manifolds Suppose more generally that

$$\Phi : \mathbb{R}^n \to \mathbb{R}^\ell$$

is C^1 and $\ell \geq 1$. See Example 3 in Section 19.2.

Now

$$M = M^1 \cap \cdots \cap M^\ell,$$

where

$$M^i = \{x : \Phi^i(x) = 0\}.$$

[7] The space N_aM does not depend on the particular Φ used to describe M. We show this in the next section.

Note that each Φ^i is real-valued. Thus we expect that, under reasonable conditions, M should have dimension $n-\ell$ in some sense. In fact, if

$$\nabla\Phi^1(x),\dots,\nabla\Phi^\ell(x)$$

are linearly independent for each $x\in M$, then the same argument as for Example 3 in the previous section shows that M is locally the graph of a function of ℓ of the variables $x_1,\dots,x_n$ in terms of the other $n-\ell$ variables.

This leads to the following definition which generalises the previous one.

DEFINITION 19.3.2. [**Manifolds as Level Sets**] Suppose $M\subset\mathbb{R}^n$ and for each $a\in M$ there exists $r>0$ and a C^1 function $\Phi\colon B_r(a)\to\mathbb{R}^\ell$ such that

$$M\cap B_r(a)=\{x:\Phi(x)=0\}.$$

Suppose also that $\nabla\Phi^1(x),\dots,\nabla\Phi^\ell(x)$ are linearly independent for each $x\in B_r(a)$.

Then M is an $n-\ell$ *dimensional manifold in* $\mathbb{R}^n$. We say M has *codimension* ℓ.

The ℓ dimensional space spanned by $\nabla\Phi^1(a),\dots,\nabla\Phi^\ell(a)$ is called the *normal space to* M *at* a and is denoted by N_aM [8].

Remarks With the same proof as in Examples 3 from the section on the Implicit Function Theorem, we can locally write M as the graph of a function of ℓ of the variables in terms of the remaining $n-\ell$ variables.

Equivalent Definitions There are two other ways to define a manifold. For simplicity of notation we consider the case M has codimension one, but the more general case is completely analogous.

DEFINITION 19.3.3. [**Manifolds as Graphs**] Suppose $M\subset\mathbb{R}^n$ and that for each $a\in M$ there exist $r>0$ and a C^1 function $f:\Omega\,(\subset\mathbb{R}^{n-1})\to\mathbb{R}$ such that for some $1\le i\le n$

$$M\cap B_r(a)=\{x\in B_r(a):x_i=f(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n)\}.$$

Then M is an $n-1$ *dimensional manifold in* $\mathbb{R}^n$.

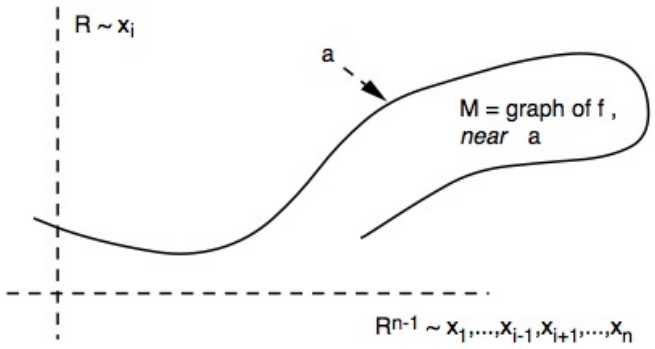

FIGURE 9. In the diagram $n=2$, the vertical axis corresponds to x_1 and the horizontal axis to x_2. The curve M is the level set of $f:\Omega\subset\mathbb{R}^2\to\mathbb{R}$ and is a one-dimensional manifold. Note that M can be written only locally as a graph, for example one part as a function of x_1 and two as functions of x_2. *Explain to someone.*

[8] The space N_aM does not depend on the particular Φ used to describe M. We show this in the next section.

Equivalence of the Level-Set and Graph Definitions Suppose M is a manifold as in the Graph Definition. Let

$$\Phi(x) = x_i - f(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).$$

Then

$$\nabla\Phi(x) = \left(-\frac{\partial f}{\partial x_1}, \ldots, -\frac{\partial f}{\partial x_{i-1}}, 1, -\frac{\partial f}{\partial x_{i+1}}, \ldots, -\frac{\partial f}{\partial x_n}\right)$$

In particular, $\nabla\Phi(x) \neq 0$ and so M is a manifold in the level-set sense.

Conversely, we have already seen (in the Remarks following Definitions 19.3.1 and 19.3.2) that if M is a manifold in the level-set sense then it is also a manifold in the graphical sense.

As an example of the next definition, see the diagram preceding Proposition 18.3.2.

DEFINITION 19.3.4. [**Manifolds as Parametrised Sets**] Suppose $M \subset \mathbb{R}^n$ and that for each $a \in M$ there exists $r > 0$ and a C^1 function

$$F : \Omega\,(\subset \mathbb{R}^{n-1}) \to \mathbb{R}^n$$

such that

$$M \cap B_r(a) = F[\Omega] \cap B_r(a).$$

Suppose moreover that the vectors

$$\frac{\partial F}{\partial u_1}(u), \ldots, \frac{\partial F}{\partial u_{n-1}}(u)$$

are *linearly independent* for each $u \in \Omega$.

Then M is an $n-1$ *dimensional manifold in* $\mathbb{R}^n$. We say that (F, Ω) is a parametrisation of (part of) M.

The $n-1$ dimensional space spanned by $\frac{\partial F}{\partial u_1}(u), \ldots, \frac{\partial F}{\partial u_{n-1}}(u)$ is called the *tangent space to* M *at* $a = F(u)$ and is denoted by T_aM [9].

Equivalence of the Graph and Parametrisation Definitions (See Figure 10.)

Suppose M is a manifold as in the Parametrisation Definition. We want to show that M is locally the graph of a C^1 function.

First note that the $n \times (n-1)$ matrix $\left[\frac{\partial F}{\partial u}(p)\right]$ has rank $n-1$ and so $n-1$ of the *rows* are linearly independent. Suppose the first $n-1$ rows are linearly independent.

It follows that the $(n-1) \times (n-1)$ matrix $\left[\frac{\partial F^i}{\partial u_j}(p)\right]_{1 \leq i,j \leq n-1}$ is invertible and hence by the Inverse Function Theorem there is locally a one-one correspondence between $u = (u_1, \ldots, u_{n-1})$ and points of the form

$$(x_1, \ldots, x_{n-1}) = (F^1(u), \ldots, F^{n-1}(u)) \in \mathbb{R}^{n-1} \simeq \mathbb{R}^{n-1} \times \{0\}\ (\subset \mathbb{R}^n),$$

with C^1 inverse G (so $u = G(x_1, \ldots, x_{n-1})$).

Thus points in M can be written in the form

$$\left(F^1(u), \ldots, F^{n-1}(u), F^n(u)\right) = \left(x_1, \ldots, x_{n-1}, (F^n \circ G)(x_1, \ldots, x_{n-1})\right).$$

Hence M is locally the graph of the C^1 function $F^n \circ G$.

Conversely, suppose M is a manifold in the graph sense. Then locally, after perhaps relabelling coordinates, for some C^1 function $f : \Omega\,(\subset \mathbb{R}^{n-1}) \to \mathbb{R}$,

$$M = \{(x_1, \ldots, x_n) : x_n = f(x_1, \ldots, x_{n-1})\}.$$

[9] The space T_aM does not depend on the particular Φ used to describe M. We show this in the next section.

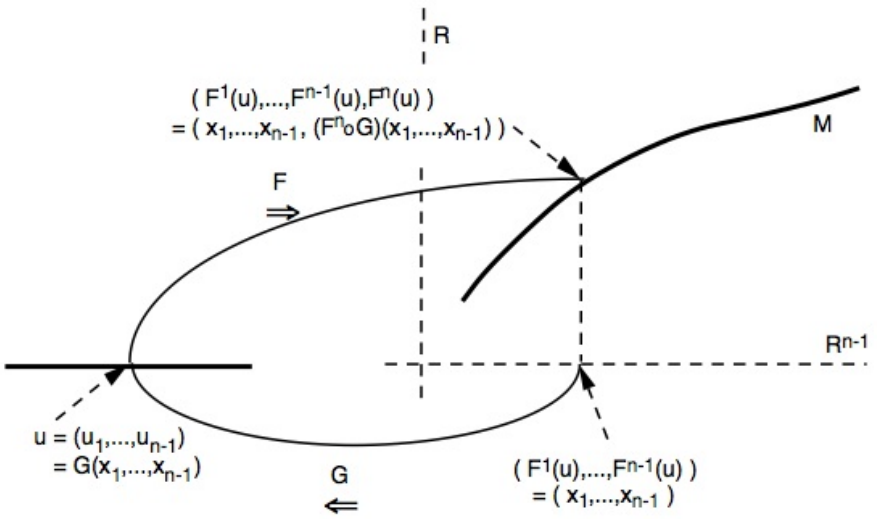

FIGURE 10. Diagram for the discussion: **Equivalence of the Graph and Parametrisation Definitions**.

It follows that M is also locally the *image* of the C^1 function $F:\Omega\,(\subset \mathbb{R}^{n-1}) \to \mathbb{R}^n$ defined by

$$F(x_1,\ldots,x_{n-1}) = (x_1,\ldots,x_{n-1}, f(x_1,\ldots,x_{n-1}))\,.$$

Moreover,

$$\frac{\partial F}{\partial x_i} = e_i + \frac{\partial f}{\partial x_i} e_n$$

for $i = 1,\ldots,n-1$, and so these vectors are linearly independent.

In conclusion, we have established the following theorem.

THEOREM 19.3.5. *The level-set, graph and parametrisation definitions of a manifold are equivalent.*

Remark If M is parametrised locally by a function $F:\Omega\,(\subset \mathbb{R}^k) \to \mathbb{R}^n$ and also given locally as the zero-level set of $\Phi:\mathbb{R}^n \to \mathbb{R}^\ell$ then it follows that

$$k + \ell = n.$$

To see this, note that previous arguments show that M is locally the graph of a function from $\mathbb{R}^k \to \mathbb{R}^{n-k}$ and also locally the graph of a function from $\mathbb{R}^{n-\ell} \to \mathbb{R}^\ell$. This makes it very plausible that $k = n-\ell$. A strict proof requires a little topology or measure theory.

19.4. Tangent and Normal vectors

If M is a manifold given as the zero-level set (locally) of $\Phi:\mathbb{R}^n \to \mathbb{R}^\ell$, then we defined the normal space N_aM to be the space spanned by $\nabla\Phi^1(a),\ldots,\nabla\Phi^\ell(a)$. If M is parametrised locally by $F:\mathbb{R}^k \to \mathbb{R}^n$ (where $k+\ell = n$), then we defined the tangent space T_aM to be the space spanned by $\frac{\partial F}{\partial u_1}(u),\ldots,\frac{\partial F}{\partial u_k(u)}$, where $F(u) = a$.

We next give a definition of T_aM which does not depend on the particular representation of M. We then show that N_aM is the orthogonal complement of T_aM, and so also N_aM does not depend on the particular representation of M.

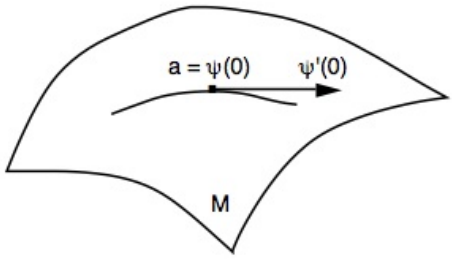

FIGURE 11. Diagram for Definition 19.4.1.

DEFINITION 19.4.1. Let M be a manifold in $\mathbb{R}^n$ and suppose $a \in M$. Suppose $\psi : I \to M$ is C^1 where $0 \in I \subset \mathbb{R}$, I is an interval and $\psi(0) = a$. Any vector h of the form

$$h = \psi'(0)$$

is said to be *tangent to M at A*. The set of all such vectors is denoted by T_aM.

THEOREM 19.4.2. *The set T_aM as defined above is indeed a vector space.*

If M is given locally by the parametrisation $F : \mathbb{R}^k \to \mathbb{R}^n$ and $F(u) = a$ then T_aM is spanned by

$$\frac{\partial F}{\partial u_1}(u), \ldots, \frac{\partial F}{\partial u_k}(u).^{10}$$

If M is given locally as the zero-level set of $\Phi : \mathbb{R}^n \to \mathbb{R}^\ell$ then T_aM is the orthogonal complement of the space spanned by

$$\nabla\Phi^1(a), \ldots, \nabla\Phi^\ell(a).$$

PROOF. *Step 1*: First suppose $h = \psi'(0)$ as in the Definition. Then

$$\Phi^i(\psi(t)) = 0$$

for $i = 1, \ldots, \ell$ and for t near 0. By the chain rule

$$\sum_{j=1}^{n} \frac{\partial \Phi^i}{\partial x^j}(a) \frac{d\psi^j}{dt}(0) \quad \text{for } i = 1, \ldots, \ell,$$

i.e.

$$\nabla\Phi^i(a) \perp \psi'(0) \quad \text{for } i = 1, \ldots, \ell.$$

This shows that T_aM (as in Definition 19.4.1) is orthogonal to the space spanned by $\nabla\Phi^1(a), \ldots, \nabla\Phi^\ell(a)$, and so is a *subset* of a space of dimension $n - \ell$.

Step 2: If M is parametrised by $F : \mathbb{R}^k \to \mathbb{R}^n$ with $F(u) = a$, then every vector

$$\sum_{i=1}^{k} \alpha_i \frac{\partial F}{\partial u_i}(u)$$

is a tangent vector as in Definition 19.4.1. To see this let

$$\psi(t) = F(u_1 + t\alpha_1, \ldots, u_n + t\alpha_n).$$

Then by the chain rule,

$$\psi'(0) = \sum_{i=1}^{k} \alpha_i \frac{\partial F}{\partial u_i}(u).$$

[10] As in Definition 19.3.4, these vectors are assumed to be linearly independent.

Hence T_aM contains the space spanned by $\frac{\partial F}{\partial u_1}(u), \ldots, \frac{\partial F}{\partial u_k}(u)$, and so *contains* a space of dimension $k(= n - \ell)$.

Step 3: From the last line in Steps 1 and 2, it follows that T_aM *is* a space of dimension $n - \ell$. It follows from Step 1 that T_aM is in fact the orthogonal complement of the space spanned by $\nabla\Phi^1(a), \ldots, \nabla\Phi^\ell(a)$, and from Step 2 that T_aM is in fact spanned by $\frac{\partial F}{\partial u_1}(u), \ldots, \frac{\partial F}{\partial u_k}(u)$. □

19.5. Maximum, Minimum, and Critical Points

In this section suppose $F : \Omega\,(\subset \mathbb{R}^n) \to \mathbb{R}$, where Ω is open.

DEFINITION 19.5.1. The point $a \in \Omega$ is a *local minimum point* for F if for some $r > 0$

$$F(a) \leq F(x)$$

for all $x \in B_r(a)$.

A similar definition applies for local maximum points.

THEOREM 19.5.2. *If F is C^1 and a is a local minimum or maximum point for F, then*

$$\frac{\partial F}{\partial x_1}(a) = \cdots = \frac{\partial F}{\partial x_n}(a) = 0.$$

Equivalently, $\nabla F(a) = 0$.

PROOF. Fix $1 \leq i \leq n$. Let

$$g(t) = F(a + te_i) = F(a_1, \ldots, a_{i-1}, a_i + t, a_{i+1}, \ldots, a_n).$$

Then $g : \mathbb{R} \to \mathbb{R}$ and g has a local minimum (or maximum) at 0. Hence $g'(0) = 0$.

But

$$g'(0) = \frac{\partial F}{\partial x_i}(a)$$

by the chain rule, and so the result follows. □

DEFINITION 19.5.3. If $\nabla F(a) = 0$ then a is a *critical point* for F.

Remark Every local maximum or minimum point is a critical point, but not conversely. In particular, a may correspond to a "saddle point" of the graph of F.

For example, if $F(x, y) = x^2 - y^2$, then $(0, 0)$ is a critical point. See the diagram before Definition 17.6.5.

19.6. Lagrange Multipliers

We are often interested in the problem of investigating the maximum and minimum points of a real-valued function F restricted to some manifold M in $\mathbb{R}^n$.

DEFINITION 19.6.1. Suppose M is a manifold in $\mathbb{R}^n$. The function $F : \mathbb{R}^n \to \mathbb{R}$ has a local minimum (maximum) at $a \in M$ *when F is constrained to M* if for some $r > 0$,

$$F(a) \leq (\geq)\, F(x)$$

for all $x \in B_r(a)$.

If F has a local (constrained) minimum at $a \in M$ then it is intuitively reasonable that the rate of change of F in any direction h in T_aM should be zero. Since

$$D_hF(a) = \nabla F(a) \cdot h,$$

this means $\nabla F(a)$ is orthogonal to *any* vector in T_aM and hence belongs to N_aM. We make this precise in the following Theorem.

THEOREM 19.6.2 (Method of Lagrange Multipliers). *Let M be a manifold in $\mathbb{R}^n$ given locally as the zero-level set of $\Phi:\mathbb{R}^n \to \mathbb{R}^\ell$* [11].

Suppose

$$F:\mathbb{R}^n \to \mathbb{R}$$

is C^1 and F has a constrained minimum (maximum) at $a \in M$. Then

$$\nabla F(a) = \sum_{j=1}^{\ell} \lambda_j \nabla \Phi^j(a)$$

for some $\lambda_1, \ldots, \lambda_\ell \in \mathbb{R}$ called **Lagrange Multipliers**.

Equivalently, let $H:\mathbb{R}^{n+\ell} \to \mathbb{R}$ be defined by

$$H(x_1, \ldots, x_n, \sigma_1, \ldots, \sigma_\ell) = F(x_1, \ldots, x_n) - \sigma_1 \Phi^1(x_1, \ldots, x_n) - \ldots - \sigma_\ell \Phi^\ell(x_1, \ldots, x_n).$$

Then H has a critical point at $a_1, \ldots, a_n, \lambda_1, \ldots, \lambda_\ell$ for some $\lambda_1, \ldots, \lambda_\ell$

PROOF. Suppose $\psi:I \to M$ where I is an open interval containing 0, $\psi(0) = a$ and ψ is C^1.

Then $F(\psi(t))$ has a local minimum at $t = 0$ and so by the chain rule

$$0 = \sum_{i=1}^{n} \frac{\partial F}{\partial x_i}(a) \frac{d\psi^i}{dt}(0),$$

i.e.

$$\nabla F(a) \perp \psi'(0).$$

Since $\psi'(0)$ can be any vector in $T_a M$, it follows $\nabla F(a) \in N_a M$. Hence

$$\nabla F(a) = \sum_{j=1}^{\ell} \lambda_j \nabla \Phi^j(a)$$

for some $\lambda_1, \ldots, \lambda_\ell$. This proves the first claim.

For the second claim just note that

$$\frac{\partial H}{\partial x_i} = \frac{\partial F}{\partial x_i} - \sum_j \sigma_j \frac{\partial \Phi^j}{\partial x_i}, \quad \frac{\partial H}{\partial \sigma_j} = -\Phi^j.$$

Since $\Phi^j(a) = 0$ it follows that H has a critical point at $a_1, \ldots, a_n, \lambda_1, \ldots, \lambda_\ell$ iff

$$\frac{\partial F}{\partial x_i}(a) = \sum_{j=1}^{\ell} \lambda_j \frac{\partial \Phi^j}{\partial x_i}(a)$$

for $i = 1, \ldots, n$. That is,

$$\nabla F(a) = \sum_{j=1}^{\ell} \lambda_j \nabla \Phi^j(a).$$

□

Example Find the maximum and minimum points of

$$F(x, y, z) = x + y + 2z$$

on the ellipsoid

$$M = \{(x, y, z) : x^2 + y^2 + 2z^2 = 2\}.$$

Solution: Let

$$\Phi(x, y, z) = x^2 + y^2 + 2z^2 - 2.$$

At a critical point there exists λ such that

$$\nabla F = \lambda \nabla \Phi.$$

[11] Thus Φ is C^1 and for each $x \in M$ the vectors $\nabla \Phi^1(x), \ldots, \nabla \Phi^\ell(x)$ are linearly independent.

That is

$$\begin{aligned} 1 &= \lambda(2x) \\ 1 &= \lambda(2y) \\ 2 &= \lambda(4z). \end{aligned}$$

Moreover

$$x^2 + y^2 + 2z^2 = 2.$$

These four equations give

$$x = \frac{1}{2\lambda},\ y = \frac{1}{2\lambda},\ z = \frac{1}{2\lambda},\ \frac{1}{\lambda} = \pm\sqrt{2}.$$

Hence

$$(x, y, z) = \pm\frac{1}{\sqrt{2}}(1, 1, 1).$$

Since F is continuous and M is compact, F must have a minimum and a maximum point. Thus one of $\pm(1, 1, 1)/\sqrt{2}$ must be the minimum point and the other the maximum point. A calculation gives

$$\begin{aligned} F\left(\frac{1}{\sqrt{2}}(1, 1, 1)\right) &= 2\sqrt{2} \\ F\left(\frac{1}{-\sqrt{2}}(1, 1, 1)\right) &= -2\sqrt{2}. \end{aligned}$$

Thus the minimum and maximum points are $-(1, 1, 1)/\sqrt{2}$ and $+(1, 1, 1)/\sqrt{2}$ respectively.

Bibliography

[An] H. Anton. *Elementary Linear Algebra.* 6th edn, 1991, Wiley.

[Ba] M. Barnsley. *Fractals Everywhere.* 1988, Academic Press.

[BM] G. Birkhoff and S. MacLane. *A Survey of Modern Algebra.* rev. edn, 1953, Macmillan.

[Br] V. Bryant. *Metric Spaces, iteration and application.* 1985, Cambridge University Press.

[Fl] W. Fleming. *Calculus of Several Variables.* 2nd ed., 1977, Undergraduate Texts in Mathematics, Springer-Verlag.

[La] S. R. Lay. *Analysis, an Introduction to Proof.* 1986, Prentice-Hall.

[Ma] B. Mandelbrot. *The Fractal Geometry of Nature.* 1982, Freeman.

[Me] E. Mendelson. *Number Systems and the Foundations of Analysis* 1973, Academic Press.

[Ms] E.E. Moise. *Introductory Problem Courses in Analysis and Toplogy.* 1982, Springer-Verlag.

[Mo] G. P. Monro. *Proofs and Problems in Calculus.* 2nd ed., 1991, Carslaw Press.

[PJS] H-O. Peitgen, H. Jürgens, D. Saupe. *Fractals for the Classroom.* 1992, Springer-Verlag.

[BD] J. Bélair, S. Dubuc. *Fractal Geometry and Analysis.* NATO Proceedings, 1989, Kluwer.

[Sm] K. T. Smith. *Primer of Modern Analysis.* 2nd ed., 1983, Undergraduate Texts in Mathematics, Springer-Verlag.

[Sp] M. Spivak. *Calculus.* 1967, Addison-Wesley.

[St] K. Stromberg. *An Introduction to Classical Real Analysis.* 1981, Wadsworth.

[Sw] E. W. Swokowski. *Calculus.* 5th ed., 1991, PWS-Kent.