Elements of mathematical analysis

Harald Hanche-Olsen

hanche@math.ntnu.no

Abstract. These notes were written as a supplement to a course on partial differential equations (PDEs), but have since been adapted for use in a course on linear analysis.

This material is covered in many books. The presentation in this note is quite terse, but I hope the motivated reader will not have any serious difficulty reading it.

If you find misprints or other mistakes or shortcomings of these notes, I would like to hear about it — preferably by email.

Introduction

in which the author tries to explain why studying this note is useful, and gives fatherly advice on how to do so.

In a sense, mathematical analysis can be said to be about continuity. The epsilon–delta arguments that you meet in a typical calculus course represent the beginnings of mathematical analysis. Unfortunately, too often these definitions are briefly presented, then hardly used at all and soon forgotten in the interest of not losing too many students and because, frankly, it is not *that* important in elementary calculus. As mathematics becomes more abstract, however, there is no way to proceed without a firm grounding in the basics. Most PDEs, for example, do not admit any solution by formulas. Therefore, emphasis is on different questions: Does a solution exist? If so, is it unique? And if so, does it depend on the data in a continuous manner? When you cannot write up a simple formula for the solution of a PDE, you must resort to other methods to prove existence. Quite commonly, some iterative method is used to construct a sequence which is then shown to converge to a solution. This requires careful estimates and a thorough understanding of the underlying issues. Similarly, the question of continuous dependence on the data is not a trivial task when all you have to work with is the existence of a solution and some of its properties.

How to read these notes. The way to read these notes is slowly. Because the presentation is so brief, you may be tempted to read too much at a time, and you get confused because you have not properly absorbed the previous material. If you get stuck, backtrack a bit and see if that helps.

The core material is contained in the first four sections — on metric spaces, completeness, compactness, and continuity. These sections should be read in sequence, more or less. The final two sections, one on ordinary differential equations and one on the implicit and inverse function theorems, are independent of each other.

The end of a proof is marked with ∎ in the right margin. Sometimes, you see the statement of a theorem, proposition etc. ended with such a box. If so, that means the proof is either contained in the previous text or left as an exercise (sometimes trivial, sometimes not — but always doable, I hope). If a proof is not complete, then this is probably intentional — the idea is for you to complete it yourself.

Metric spaces

in which the basic objects of study are introduced, and their elementary properties are established.

Most of mathematical analysis happens in some sort of metric space. This is a set in which we are given some way to measure the *distance* $d(x,y)$ between two points x and y. The distance function (metric) d has to satisfy some simple axioms in order to be useful for our purposes.

Later, we shall see that a metric space is the proper space on which to define continuity of functions (actually, there is a more general concept — that of a *topological space* — that is even more appropriate, but we shall not need that level of abstraction here).

1 Definition. *A* metric *on a set X is a real-valued function $d\colon X \times X \to \mathbb{R}$ satisfying:*

- $d(x,x) = 0$
- $d(x,y) > 0$ *if* $x \neq y$
- $d(x,y) = d(y,x)$ *(symmetry)*
- $d(x,z) \le d(x,y) + d(y,z)$ *(the triangle inequality)*

A metric space *is a pair (X,d) where X is a set and d a metric on X (however we often speak of the metric space X, where d is understood).*

Before moving on to the examples, we shall note that the triangle inequality can easily be generalised to more than three elements. For example, two applications of the triangle inequality yields the inequality

$$d(x,w) \le d(x,y) + d(y,w) \le d(x,y) + d(y,z) + d(z,w).$$

In fact, it is not difficult to prove the general inequality

$$d(x_0,x_n) \le \sum_{k=1}^{n} d(x_{k-1},x_k)$$

by induction on n. This is sometimes called the *generalised triangle inequality*, but we shall simply call this the triangle inequality as well. We shall resist the temptation to call it the polygonal inequality. While the original triangle inequality corresponds to the fact that the sum of two sides in a triangle is at least as large as the third, the above inequality corresponds to a similar statement about the sides of an $n+1$-gon.

2 Examples.

$\mathbb{R}$ or $\mathbb{C}$ with $d(x,y) = |x-y|$.

$\mathbb{R}^n$ or $\mathbb{C}^n$ with $d(x,y) = \|x-y\|$ (where $\|x\| = \sqrt{\sum_{j=1}^n |x_j|^2}$).

Any set X with $d(x,x) = 0$ and $d(x,y) = 1$ whenever $x \ne y$. This is called a *discrete* metric space.

3 Definition. *A norm on a real or complex vector space X is a map $\|\cdot\|\colon X \to \mathbb{R}$ satisfying:*

- $\|x\| > 0$ *if* $x \ne 0$
- $\|ax\| = |a| \cdot \|x\|$ *for every scalar a and $x \in X$*
- $\|x+y\| \le \|x\| + \|y\|$ *(the triangle inequality)*

A normed space is a vector space with a norm on it. Such a space is also a metric space with $d(x,y) = \|x-y\|$.

4 Examples.

$\mathbb{R}^n$ or $\mathbb{C}^n$ with $\|x\| = \sqrt{\sum_{j=1}^n |x_j|^2}$.

$\mathbb{R}^n$ or $\mathbb{C}^n$ with $\|x\| = \sum_{j=1}^n |x_j|$.

$\mathbb{R}^n$ or $\mathbb{C}^n$ with $\|x\| = \max\{|x_j|: 1 \le j \le n\}$.

The space l^∞ consisting of all bounded sequences $x = (x_1, x_2, \ldots)$ of real (or complex) numbers, with $\|x\| = \max\{|x_j|: 1 \le j \le \infty\}$.

We shall often need to consider subsets of a metric space as metric spaces in their own right. Thus, if (X, d) is a metric space and A is a subset of X, then $(A, d|_{A\times A})$ is a metric space. (The notation $f|S$ is often used to denote the *restriction* of a function f to the subset S, in the sense that $f|_S(x) = f(x)$ when $x \in S$ but $f|_S(x)$ is not defined when $x \notin S$.) We say that A is given the metric *inherited*, or *induced* from X. Later, we shall define what it means for a metric space to be compact. Then we shall say that a subset $A \subseteq X$ is compact if it is compact when given the inherited metric, and we can do similarly with any other concept relating to metric spaces.

5 Definition. *The open ε-neighbourhood, also called the open ε-ball (where $\varepsilon > 0$) of a point x in a metric space X is*

$$B_\varepsilon(x) = \{\xi \in X\colon d(x, \xi) < \varepsilon\}$$

The corresponding closed ε-neighbourhood (or –ball) is

$$\bar{B}_\varepsilon(x) = \{\xi \in X\colon d(x, \xi) \le \varepsilon\}$$

A neighbourhood of a point in a metric space is a subset containing an ε-ball for some $\varepsilon > 0$. A set is open if it is a neighbourhood of every one of its points. The interior of a subset A of X is the set of all points for which A is a neighbourhood.

Note that $x \in B_\varepsilon(x)$, and that $B_\varepsilon(x)$ may consist of no points other than x itself (if, for example, X is given the discrete metric and $\varepsilon \le 1$). It should also be noted that $\bar{B}_\varepsilon(x)$ is not necessarily the closure of $B_\varepsilon(x)$ (for example, with the discrete metric, $B_1(x) = \{x\}$ while $\bar{B}_1(x) = X$).

6 Exercise. Use the triangle inequality to show that an open ε-ball is in fact open. Prove that the interior of any set is open, and that the interior of A is in fact the largest open subset of A. Finally, show that the complement of a closed ε-ball, that is a set of the form $X \setminus \bar{B}_\varepsilon(x)$, is open.

7 Definition. *A sequence $(x_n)_{n=1}^{\infty}$ in a metric space X is said to converge to a limit $x \in X$ (and we write $x_n \to x$) if, for each $\varepsilon > 0$, there is an index N so that $d(x_n, x) < \varepsilon$ whenever $n \geq N$. A part $(x_n)_{n=N}^{\infty}$ is called a tail of the sequence. Thus, the sequence (x_n) converges to the limit x if and only if every neighbourhood of x contains some tail of the sequence. A sequence is called convergent if it converges to some limit.*

8 Exercise. Show that no sequence can have more than one limit.

As a result of the above exercise, we can talk about *the* limit of a convergent sequence, and write $\lim_{n\to\infty} x_n$ for the limit. Though a sequence can have only one limit, a non-convergent sequence can have many limit points:

9 Definition. *A limit point of a sequence (x_n) is a point $x \in X$ such that, for every $\varepsilon > 0$ and every N, there is some $n \geq N$ so that $d(x_n, x) < \varepsilon$. Equivalently, every neighbourhood of x contains at least one point (and therefore infinitely many points) from every tail of the sequence.*

10 Exercise. A sequence $(y_k)_{k=1}^{\infty}$ is called a *subsequence* of a sequence $(x_n)_{n=1}^{\infty}$ if it is possible to find $n_1 < n_2 < \cdots$ so that $y_k = x_{n_k}$ for all k. Show that x is a limit point of the sequence (x_n) if and only if some subsequence of (x_n) converges to x.

11 Definition. *A subset F of a metric space X is called closed if, whenever a sequence contained in F converges to a point in X, the limit is in F.*

12 Proposition. *For a subset F of a metric space X the following are equivalent:*

(*a*) *F is closed,*
(*b*) *the complement $X \setminus F$ is open,*
(*c*) *for every $x \in X$, if every neighbourhood of x has a nonempty intersection with F, then $x \in F$,*
(*d*) *F contains every limit point of every sequence contained in F.*

∎

13 Definition. *The closure of a subset A of the metric space X is the set $\bar{A}$ of limits of all convergent sequences in A.*

14 Proposition. *The closure of a subset A of a metric space X is a closed set containing A, and is in fact the smallest closed set containing A. It also consists of all limit points of sequences in A. Finally, a point x belongs to $\bar{A}$ if and only if every neighbourhood of x has a nonempty intersection with A.* ∎

15 Proposition. *The union of an arbitrary family of open sets is open, and the intersection of an arbitrary family of closed sets is closed. The intersection of a finite family of open sets is open, and the union of a finite family of closed sets is closed.* ∎

For the sake of completeness, I include the definition of a *topological space* here. This is a set X together with a family $\mathcal{O}$ of subsets of X satisfying the following requirements:

- $\emptyset \in \mathcal{O}$ and $X \in \mathcal{O}$,
- If $U \in \mathcal{O}$ and $V \in \mathcal{O}$ then $U \cap V \in \mathcal{O}$,
- The union of an arbitrary subfamily of $\mathcal{O}$ is in $\mathcal{O}$.

The members of $\mathcal{O}$ are called *open*, and their complements are called *closed*. A *neighbourhood* of a point is a set containing an open set containing the point.

Clearly, a metric space together with its open sets is a topological space.

In these notes you may notice that notions like continuity and compactness have equivalent formulations in terms of open or closed sets (or neighbourhoods). These notions, then, can be generalised to topological spaces.

Completeness

in which we already encounter our first Theorem.

16 Definition. *A sequence (x_n) is said to be a Cauchy sequence if, for each $\varepsilon > 0$ there is some N so that $d(x_m, x_n) < \varepsilon$ whenever $m, n \geq N$.*

It is easy to see that any convergent sequence is Cauchy (exercise), but the converse is not true, as can be seen from the metric space $\mathbb{Q}$ (all rational numbers, with the usual metric $d(x, y) = |x - y|$). A sequence in $\mathbb{Q}$ converging to $\sqrt{2}$ in $\mathbb{R}$ is convergent in $\mathbb{R}$ and hence Cauchy, but is not convergent in $\mathbb{Q}$.

17 Definition. *A metric space in which every Cauchy sequence converges is called complete. A complete normed space is called a Banach space.*

One reason to be interested in complete spaces is that one can often prove existence theorems by somehow constructing a Cauchy sequence and considering its limit: Proving that a sequence converges may be difficult, as you need to know the limit before you can use the definition of convergence; by comparison, showing that a sequence is Cauchy may be much easier. As an example, we state and prove the Banach fixed point theorem, also known as the *contraction principle*.

18 Definition. *A contraction on a metric space X is a function f from X to itself so that there is a constant $K < 1$, such that*

$$d(f(x), f(y)) \le K d(x, y)$$

for every $x, y \in X$. A fixed point of a map $f\colon X \to X$ is a point $x \in X$ so that $f(x) = x$.

19 Theorem. (Banach's fixed point theorem) *A contraction on a nonempty complete metric space has one, and only one, fixed point.*

Proof: Let $f\colon X \to X$ be a contraction, and $0 < K < 1$ such that $d(f(x), f(y)) \le K d(x, y)$ whenever $x, y \in X$. Let $x_0 \in X$ be any point, and define the sequence (x_n) by $x_{n+1} = f(x_n)$, $n = 0, 1, 2, \ldots$. It is easy to prove by induction that $d(x_{n+1}, x_n) \le K^n d(x_1, x_0)$ and hence, by repeated use of the triangle inequality, whenever $1 \le m < n$ we get

$$\begin{aligned} d(x_n, x_m) &\le \sum_{k=m}^{n-1} d(x_{k+1}, x_k) \le \sum_{k=m}^{n-1} K^k d(x_1, x_0) \\ &< d(x_1, x_0) \sum_{k=m}^{\infty} K^k = d(x_1, x_0) \frac{K^m}{1-K} \end{aligned}$$

and since $0 < K < 1$ it is then clear that (x_n) is a Cauchy sequence, and hence convergent since X is complete. Let x be the limit.

We need to prove that x is a fixed point. We know that x_n is an approximate fixed point when n is large, in the sense that $d(f(x_n), x_n) \to 0$

when $n \to \infty$ (because $d(f(x_n), x_n) = d(x_{n+1}, x_n) \le K^n d(x_1, x_0)$). We perform a standard gymnastic exercise using the triangle inequality: $f(x)$ is close to $f(x_n) = x_{n+1}$ which is close to x. More precisely:

$$\begin{aligned} d(f(x), x) &\le d(f(x), f(x_n)) + d(x_{n+1}, x) \\ &\le K d(x, x_n) + d(x_{n+1}, x) \to 0 \quad (n \to \infty) \end{aligned}$$

Thus, for any $\varepsilon > 0$ we can use the above inequality with a sufficiently large n to obtain $d(f(x), x) < \varepsilon$, and since $\varepsilon > 0$ was arbitrary, we must have $d(f(x), x) = 0$. Thus $f(x) = x$, and x is indeed a fixed point of f.

It remains to prove the uniqueness of the fixed point. So, assume x and y are fixed points, that is, $f(x) = x$ and $f(y) = y$. Then

$$d(x, y) = d(f(x), f(y)) \le K d(x, y)$$

and since $0 < K < 1$ while $d(x, y) \ge 0$, this is only possible if $d(x, y) = 0$. Thus $x = y$, and the proof is complete. ∎

In many applications of the fixed point theorem, we are given a function which is not a contraction on the entire space, but which is so locally. In this case, we need some other condition to ensure the existence of a fixed point. In the following very useful case, it turns out that the proof of the Banach fixed point theorem can be adapted.

20 Corollary. *Assume X is a complete metric space, that $x_0 \in X$, and that $f\colon \bar{B}_r(x_0) \to X$ is a continuous function. Assume further that $K < 1$ is so that*

$$\begin{gathered} d(f(x), f(y)) \le K d(x, y) \qquad (x, y \in \bar{B}_r(x_0)), \\ d(f(x_0), x_0) \le (1 - K) r. \end{gathered}$$

Then f has a unique fixed point in $\bar{B}_r(x_0)$.

Proof: The uniqueness proof is just as in the theorem. Next, let $x_1 = f(x_0)$, and more generally $x_{n+1} = f(x_n)$ whenever x_n is defined and $x_n \in \bar{B}_r(x_0)$. Just as in the proof of the theorem, we find

$$d(x_n, x_m) < d(x_1, x_0) \frac{K^m}{1 - K}$$

provided $x_0,\dots,x_n$ are defined. With $m=0$, this becomes

$$d(x_n,x_0) < d(f(x_0),x_0)\frac{1}{1-K} \le r$$

using the assumption. Thus $x_n \in \bar{B}_r(x_0)$, and therefore we can define x_{n+1}. By induction, then, x_n is defined and in $\bar{B}_r(x_0)$ for all n. The proof that this sequence converges to a limit which is a fixed point is just like before. ∎

21 Proposition. *A subset of a complete space is complete if and only if it is closed.*

Proof: Let X be a complete metric space, and A a subset of X.

First, assume that A is closed. To show that A is complete, consider a Cauchy sequence in A. Then this sequence is also a Cauchy sequence in X. But because X is complete, the sequence has a limit in X. Since A is closed and the original sequence was contained in A, the limit belongs to A. Thus the sequence converges in A, and we have proved that A is complete.

Second, assume that A is complete. To show that A is closed in X, consider a sequence in A converging to some point $x \in X$. Since A is complete, this sequence also has a limit in A. But no sequence can have more than one limit, so the latter limit must be x, which therefore must belong to A. Thus A is closed. ∎

22 Definition. *The diameter of any subset A of a metric space is*

$$\operatorname{diam} A = \sup\{d(x,y) : x,y \in A\}$$

23 Proposition. *A metric space X is complete if and only if whenever $F_1 \supseteq F_2 \supseteq F_3 \supseteq \cdots$ are closed nonempty subsets of X with $\operatorname{diam} F_n \to 0$, the intersection $\bigcap_{n=1}^{\infty} F_n$ is nonempty.*

Clearly $\bigcap_{n=1}^{\infty} F_n$, if nonempty, has diameter 0, and so contains only a single point.

Proof: First, assume X is complete, and let $F_1 \supseteq F_2 \supseteq F_3 \supseteq \cdots$ be closed nonempty subsets of X. Pick $x_n \in F_n$ for each n. If $\operatorname{diam} F_n \to 0$ the sequence (x_n) is Cauchy and hence convergent. The limit x belongs to each F_n because $x_j \in F_n$ whenever $j \geq n$, and because F_n is closed. Thus $x \in \bigcap_{n=1}^{\infty} F_n$.

To prove the converse, let (x_n) be a Cauchy sequence, and let $F_n = \overline{\{x_n, x_{n+1}, \ldots\}}$. Then $\bigcap_{n=1}^{\infty} F_n$ is the set of limit points of (x_n). Since $\bigcap_{n=1}^{\infty} F_n \neq \emptyset$, thus (x_n) has a limit point, which must be a limit of (x_n). ∎

24 Exercise. The above proof has several gaps and details left out. Identify these, and fill them in.

Compactness

in which we define a most useful property of metric spaces such as closed and bounded intervals.

25 Definition. *A metric space is called compact if every sequence in the space has at least one limit point (and hence a convergent subsequence).*

Note that $\mathbb{R}$ is not compact, since the sequence $x_n = n$ has no limit point.

26 Exercise. Prove that any closed subset of a compact metric space is compact. Also prove that every compact subset of any metric space is closed.

27 Definition. *A metric space X is called totally bounded if, for every $\varepsilon > 0$, there exist a finite number of points $x_1, \ldots, x_n \in X$ so that, for every $y \in X$, one of the points x_i satisfies $d(x_i, y) < \varepsilon$. In other words, X is a finite union of ε-balls for every $\varepsilon > 0$.*

28 Definition. *A cover of X is a set of subsets of X whose union is all of X. An open cover is a cover consisting of open sets.*

29 Definition. *A set $\mathscr{F}$ of subsets of X has the finite intersection property if every finite subset of $\mathscr{F}$ has nonempty intersection; i.e., if $F_1, \ldots, F_n \in \mathscr{F}$, then $F_1 \cap \cdots \cap F_n \neq \emptyset$.*

30 Theorem. *For a metric space X, the following are equivalent:*

(*a*) *X is compact,*

(b) *every open cover of X contains a finite cover of X,*
(c) *every set of closed subsets of X with the finite intersection property has nonempty intersection,*
(d) *X is totally bounded and complete.*

Proof: We prove $(d) \Rightarrow (c) \Rightarrow (a) \Rightarrow (d)$. The proof of the equivalence $(b) \Leftrightarrow (c)$ will be left as an exercise. (*Hint*: $\mathscr{F}$ is a set of closed sets with the finite intersection property but with empty intersection if and only if $\{X \setminus F \colon F \in \mathscr{F}\}$ is an open cover with no finite subset which is also a cover.)

$(d) \Rightarrow (c)$: Assume X is totally bounded and complete, and let $\mathscr{F}$ be a set of closed subsets of X with the finite intersection property. If $\varepsilon > 0$, by the total boundedness we may write $X = \bigcup_{k=1}^{n} \bar{B}_\varepsilon(x_k)$. Let $\mathscr{G}_k = \{F \cap \bar{B}_\varepsilon(x_k) \colon F \in \mathscr{F}\}$. At least one of the families $\mathscr{G}_1, \ldots, \mathscr{G}_n$ has the finite intersection property (exercise: prove this). Clearly each set in $\mathscr{G}_k$ has diameter at most 2ε.

So far we have proved: Every set $\mathscr{F}$ of closed sets with the finite intersection property has a *refinement* $\mathscr{G}$ (by which we mean a family of closed sets, also with the finite intersection property, so that for every $F \in \mathscr{F}$ there exists some $G \in \mathscr{G}$ with $G \subseteq F$), each of whose members has diameter no larger than some prescribed positive number.

Let now $\varepsilon_k \searrow 0$. Let $\mathscr{F}_0 = \mathscr{F}$ and, for $k = 1, 2, \ldots$ let $\mathscr{F}_k$ be a refinement of $\mathscr{F}_{k-1}$, each of whose members has diameter at most ε_k. Next let $F_0 = X$ and, for $k = 1, 2, \ldots$ let $F_k \in \mathscr{F}_k$ with $F_k \subseteq F_{k-1}$. Now apply Proposition 23 to see that, by the completeness of X, $\bigcap_{k=1}^{\infty} F_k = \{x\}$ for some $x \in X$.

Now let $G \in \mathscr{F}$. For each k, since $\mathscr{F}_k$ refines $\mathscr{F}$ there is some $G_k \in \mathscr{F}_k$ with $G_k \subseteq G$. Since $\mathscr{F}_k$ has the finite intersection property, $F_k \cap G_k \neq \emptyset$, so let $x_k \in F_k \cap G_k$. Since $x_k \in F_k$ for each k, $x = \lim_{k\to\infty} x_k$. But $x_k \in G$ and G is closed, and therefore $x \in G$. Since G was an arbitrary member of $\mathscr{F}$, we have shown $x \in \bigcap \mathscr{F}$, so that $\mathscr{F}$ does indeed have a nonempty intersection.

$(c) \Rightarrow (a)$: Let (x_n) be a sequence in X, and let $F_n = \overline{\{x_n, x_{n+1}, \ldots\}}$. Clearly, $\mathscr{F} = \{F_1, F_2, \ldots\}$ has the finite intersection property, and its intersection $\bigcap_{n=1}^{\infty} F_n$ consists of all limit points of (x_n). If (c) holds there is therefore at least one limit point.

$(a) \Rightarrow (d)$: Assume X is compact. Clearly it is complete, for if a Cauchy sequence has at least one limit point, that limit point is unique and the sequence converges to that point (exercise).

Assume X is not totally bounded. Then there is some $\varepsilon > 0$ so that no finite number of ε-balls covers X. Let $x_1 \in X$ be arbitrary, and for $n = 1,2,3,\dots$ pick $x_{n+1} \in X \setminus \bigcup_{k=1}^n B_\varepsilon(x_k)$. Then (x_n) is a sequence in X so that $d(x_m, x_n) \ge \varepsilon$ whenever $m \ne n$. Such a sequence can have no limit point, since no open $\varepsilon/2$-ball can contain more than one point from the sequence. This contradicts the compactness of X. ∎

The real numbers

in which we prove the Heine–Borel theorem and the completeness of the field of real numbers.

We shall take the following fundamental property of $\mathbb{R}$ for granted: Every nonempty subset $S \subseteq \mathbb{R}$ which has an upper bound has a *least upper bound* $a = \sup S$. More precisely, that a is an upper bound of S means that $x \le a$ for every $x \in S$. That a is a least upper bound means that it is an upper bound for S, such that $a \le b$ whenever b is an upper bound for S. Clearly, the least upper bound is unique. For completeness, we set $\sup \emptyset = -\infty$, and $\sup S = +\infty$ if S has no upper bound. The *greatest lower bound* $\inf S$ is defined similarly, but with all the inequalities reversed. The existence of the greatest lower bound can be deduced from the existence of the least upper bound by taking negatives; in fact $\inf S = -\sup\{-s\colon s \in S\}$. For completeness, we set $\inf \emptyset = +\infty$, and $\inf S = -\infty$ if S has no lower bound.

31 Lemma. *For any closed and bounded set $F \subseteq \mathbb{R}$, $\sup F \in F$.*

Proof: Let $a = \sup F$. If $a \notin F$ then, since F is closed, there is some $\varepsilon > 0$ so that $F \cap B_\varepsilon(a) = \emptyset$. But if $x \in F$ then $x \le a$ because a is an upper bound for F, and so $x \le a - \varepsilon$ (since otherwise $|x - a| < \varepsilon$). Thus $a - \varepsilon$ is an upper bound for F, which contradicts the definition of a as the *least* upper bound for F. This contradiction shows that $a \in F$. ∎

32 Theorem. (Heine–Borel) *Every bounded and closed set of real numbers is compact.*

Historically, the version of compactness to be proven below is called *Cantor's intersection theorem*, while it is the open covering version that is properly called the Heine–Borel theorem. The fact that any bounded sequence of real numbers has a limit point is known as the *Bolzano–Weierstrass theorem.*

Proof: Let $K \subseteq \mathbb{R}$ be a bounded and closed set, and let $\mathscr{F}$ be a family of closed subsets of K, with the finite intersection property. Define $\mathscr{F}'$ to be the set of all intersections of finite subsets of $\mathscr{F}$, and finally let

$$\omega = \inf\{\sup F \colon F \in \mathscr{F}'\}.$$

The claim is that $\omega \in \bigcap \mathscr{F}$. Thus let $F \in \mathscr{F}$. We need to prove that $\omega \in F$. Since F is closed, we only need to prove – for any $\varepsilon > 0$ – that $F \cap B_\varepsilon(\omega) \neq \emptyset$.

By the definition of ω, there is some $G \in \mathscr{F}'$ so that $\sup G < \omega + \varepsilon$. Since $F \cap G \in \mathscr{F}'$, we have $\sup(F \cap G) \geq \omega$, and of course $\sup(F \cap G) \leq \sup G < \omega + \varepsilon$. Hence $\sup(F \cap G)$ belongs to $B_\varepsilon(\omega)$. By Lemma 31, $\sup(F \cap G) \in F$, so $\sup(F \cap G) \in F \cap B_\varepsilon(\omega)$, which proves the claim. ∎

33 Corollary. *Every bounded sequence of real numbers has a limit point, and $\mathbb{R}$ is complete.*

Proof: The first statement follows from Theorem 30. To prove the second statement, note that a Cauchy sequence is certainly bounded. Thus it has a limit point. But any limit point of a Cauchy sequence is a limit of the sequence, which is therefore convergent. ∎

34 Proposition. *$\mathbb{R}^n$ is complete, and any closed and bounded subset of $\mathbb{R}^n$ is compact.*

Proof: It is enough to show that every bounded sequence in $\mathbb{R}^n$ has a limit point. So let (x_k) be such a sequence, and write $x_{k,j}$ for the j-th coordinate of x_k, so that $x_k = (x_{k,1}, \ldots, x_{k,n})$.

Since $(x_{k,1})$ is a bounded sequence of real numbers, some subsequence converges. By replacing the original sequence (x_k) by the corresponding subsequence, then, we conclude that $(x_{k,1})$ converges.

Next, by again replacing the just found subsequence with a further subsequence, we find that $(x_{k,2})$ converges. Repeating this procedure n times, we end up having replaced the original sequence with a subsequence such that $(x_{k,j})$ converges for $j = 1, 2, \ldots, n$. It is not hard to show that then the sequence (x_k) converges.

Indeed, let $y_j = \lim_{k\to\infty} x_{k,j}$. Given $\varepsilon > 0$ there is, for each j, some N_j so that $|x_{k,j} - y_j| < \varepsilon$ whenever $k > N_j$. Let $N = \max\{N_1, N_2, \ldots, N_n\}$. If $k > N$ then $|x_k - y| < \sqrt{n}\varepsilon$. Since $\sqrt{n}$ is a harmless constant in this context, the claimed convergence $x_k \to y$ follows. ∎

Continuity

in which we, at last, study the continuous functions, without which the study of metric spaces would be a fruitless and boring activity. As an application, we consider the problem of moving a differentiation operator under the integral sign.

35 Definition. *Let (X,d) and (Y,ρ) be metric spaces, and let $f\colon X \to Y$ be a function. f is said to be continuous at $x \in X$ if, for every $\varepsilon > 0$, there exists some $\delta > 0$ so that, whenever $\xi \in X$,*

$$d(\xi,x) < \delta \Rightarrow \rho(f(\xi), f(x)) < \varepsilon.$$

36 Exercise. For some fixed $y \in X$, let $f\colon X \to \mathbb{R}$ be the function defined by $f(x) = d(x,y)$. Show that f is continuous. Also, define a metric ρ on $X \times X$ by $\rho\big((x,y),(\xi,\eta)\big) = d(x,\xi) + d(y,\eta)$. Show that ρ is in fact a metric, and that $d\colon X \times X \to \mathbb{R}$ is continuous when $X \times X$ is given this metric.

37 Definition. *If $V \subseteq Y$ we write $f^{-1}[V] = \{x \in X\colon f(x) \in V\}$ (even if f has no inverse function f^{-1}).*

38 Proposition. *Let X and Y be metric spaces, $f\colon X \to Y$ a function, and $x \in X$. Then the following are equivalent:*

- (a) *f is continuous at x;*
- (b) *for each neighbourhood V of $f(x)$, $f^{-1}[V]$ is a neighbourhood of x;*
- (c) *$f(x_n) \to f(x)$ whenever $x_n \to x$.*

Proof: The equivalence of any two of these is easy to prove directly. We prove a cycle of implications.

$(a) \Rightarrow (b)$: If V is a neighbourhood of $f(x)$, then $B_\varepsilon(f(x)) \subseteq V$ for some $\varepsilon > 0$. Then, by continuity, for some $\delta > 0$ we get $B_\delta(x) \subseteq f^{-1}[B_\varepsilon(f(x))] \subseteq f^{-1}[V]$, so $f^{-1}[V]$ is a neighbourhood of x.

$(b) \Rightarrow (c)$: Let $x_n \to x$. Assume V is a neighbourhood of $f(x)$. Then, since $f^{-1}[V]$ is a neighbourhood of x, some tail of the sequence (x_n) is contained in $f^{-1}[V]$, and so the corresponding tail of the sequence $(f(x_n))$ is contained in V. Hence $f(x_n) \to f(x)$.

$(c) \Rightarrow (a)$: Assume f is not continuous at x. Then for some $\varepsilon > 0$ and every $\delta > 0$ there is some $\xi \in X$ with $d(\xi, x) < \delta$ but $\rho(f(\xi), f(x)) \geq \varepsilon$. Let $\delta_k \searrow 0$, and for each k, let $x_k \in X$ with $d(x_k, x) < \delta_k$ and $\rho(f(x_k), f(x)) \geq \varepsilon$. Then $x_k \to x$ but $f(x_k) \not\to f(x)$. ∎

39 Definition. *A function $f\colon X \to Y$ is said to be continuous if it is continuous at every point in X.*

40 Proposition. *Let X and Y be metric spaces and $f\colon X \to Y$ a function. Then the following are equivalent:*

- (a) *f is continuous;*
- (b) *for each open subset $V \subseteq Y$, $f^{-1}[V]$ is open;*
- (c) *for each closed subset $F \subseteq Y$, $f^{-1}[F]$ is closed.* ∎

41 Theorem. *Let X and Y be metric spaces and $f\colon X \to Y$ a continuous function. If X is compact then $f[X] = \{f(x)\colon x \in X\}$ is compact.*

Proof: It is instructive to give several proofs of this fact.

First, let (y_n) be a sequence in $f[X]$. Since $y_n \in f[X]$ we can write $y_n = f(x_n)$. By compactness the sequence (x_n) has a convergent subsequence (x_{n_k}). Since f is continuous, then the subsequence given by $y_{n_k} = f(x_{n_k})$ converges. Thus $f[X]$ is compact.

Second, let $\mathscr{U}$ be an open cover of $f[X]$. Then $\{f^{-1}(U)\colon U \in \mathscr{U}\}$ is an open cover of X, and so by the compactness of X there is a finite number of sets $U_1, \ldots, U_n \in \mathscr{U}$ so that $\bigcup_{k=1}^{n} f^{-1}[U_k] = X$. Then $\bigcup_{k=1}^{n} U_k = f[X]$, and hence $f[X]$ is compact.

Third, consider a family of closed subsets of $f[X]$, with the finite intersection property. The proof that the family has a nonempty intersection is left as an exercise for the reader. ∎

42 Corollary. *A continuous real-valued function on a compact space is bounded, and achieves its maximum as well as its minimum.*

Proof: A compact subset of $\mathbb{R}$ is bounded, and its infimum and supremum are members of the set because it is closed. ∎

43 Exercise. If A be a subset of some metric space X, the *distance* from a point $x \in X$ to A is the number

$$\operatorname{dist}(x, A) = \inf\{d(x, a) : a \in A\}.$$

Show that $x \mapsto \operatorname{dist}(x, A)$ is a continuous function. Assume now that K is a compact subset of X and that $K \cap \bar{A} = \emptyset$. Show that there is some $\varepsilon > 0$ so that $d(x, a) \geq \varepsilon$ whenever $x \in K$ and $a \in A$.

44 Definition. *Let (X, d) and (Y, ρ) be metric spaces. A function $f : X \to Y$ is called uniformly continuous if, whenever $\varepsilon > 0$, there exists some $\delta > 0$ so that $\rho(f(\xi), f(x)) < \varepsilon$ whenever $x, \xi \in X$ and $d(\xi, x) < \delta$.*

45 Exercise. Show that the real functions $x \mapsto x^2$ and $x \mapsto 1/x$ (with $x > 0$) are not uniformly continuous. Show that arctan is uniformly continuous.

46 Proposition. *A continuous function $f : X \to Y$ where X, Y are metric spaces and X is compact, is uniformly continuous.*

Proof: Let $\varepsilon > 0$. For every $x \in X$ there is some $\delta(x) > 0$ so that $\rho(f(\xi), f(x)) < \varepsilon$ whenever $d(\xi, x) < \delta(x)$. By the compactness of X there are $x_1, \ldots, x_n \in X$ so that $X = \bigcup_{j=1}^{n} B_{\delta(x_j)/2}(x_j)$. Let $\delta = \min\{\delta(x_1), \ldots, \delta(x_n)\}/2$.

Now, if $\xi, x \in X$, then $x \in B_{\delta(x_j)/2}(x_j)$ for some j. If furthermore $d(\xi, x) < \delta$ then $\xi \in B_{\delta(x_j)}(x_j)$ as well, and so

$$\rho(f(\xi), f(x)) \leq \rho(f(\xi), f(x_j)) + \rho(f(x_j), f(x)) < \varepsilon + \varepsilon = 2\varepsilon.$$

Hence f is uniformly continuous. ∎

The above theorem and the notion of uniform continuity has many uses. A simple application is the following result on differentiating under the integral.

47 Proposition. *Let f be a real function on some open subset $U \subset \mathbb{R}^2$. Let a, b, and x_0 be real numbers so that $(x_0, y) \in U$ whenever $a \leq y \leq$*

b. Assume that $\partial f(x,y)/\partial x$ exists and is continuous for each $(x,y)\in U$. Then the function $x\mapsto \int_a^b f(x,y)\,\mathrm{d}y$ is differentiable at x_0, with derivative

$$\frac{\mathrm{d}}{\mathrm{d}x}\bigg|_{x=x_0}\int_a^b f(x,y)\,\mathrm{d}y=\int_a^b \frac{\partial f}{\partial x}(x_0,y)\,\mathrm{d}y.$$

Proof: First, since the compact set $\{(x_0,y)\colon a\le y\le b\}$ is contained in the open set U, there is some $\delta_1>0$ so that the, likewise compact, set $\{(x,y)\colon |x-x_0|\le\delta_1, a\le y\le b\}$ is contained in U (exercise: prove this using exercise 43 with $A=X\setminus U$). Let $\varepsilon>0$. By uniform continuity of $\partial f/\partial x$ on this compact set, there is some $\delta>0$ so that

$$\left|\frac{\partial f}{\partial x}(x,y)-\frac{\partial f}{\partial x}(x_0,y)\right|<\varepsilon$$

whenever $|x-x_0|<\delta$ and $a\le y\le b$ (clearly, by picking $\delta\le\delta_1$ we make sure that $(x,y)\in U$ at the same time). Now

$$\int_a^b f(x,y)\,\mathrm{d}y-\int_a^b f(x_0,y)\,\mathrm{d}y=\int_a^b\int_{x_0}^x \frac{\partial f}{\partial x}(\xi,y)\,\mathrm{d}\xi\,\mathrm{d}y$$

and

$$\begin{aligned}
&\left|\frac{1}{x-x_0}\left(\int_a^b\int_{x_0}^x \frac{\partial f}{\partial x}(\xi,y)\,\mathrm{d}\xi\,\mathrm{d}y\right)-\int_a^b \frac{\partial f}{\partial x}(x_0,y)\,\mathrm{d}y\right|\\
&\qquad=\left|\frac{1}{x-x_0}\int_a^b\int_{x_0}^x\left(\frac{\partial f}{\partial x}(\xi,y)-\frac{\partial f}{\partial x}(x_0,y)\right)\mathrm{d}\xi\,\mathrm{d}y\right|\\
&\qquad\le\frac{1}{|x-x_0|}\int_a^b\int_{x_0}^x\left|\frac{\partial f}{\partial x}(\xi,y)-\frac{\partial f}{\partial x}(x_0,y)\right|\mathrm{d}\xi\,\mathrm{d}y\\
&\qquad<\frac{1}{|x-x_0|}\int_a^b\int_{x_0}^x \varepsilon\,\mathrm{d}\xi\,\mathrm{d}y\\
&\qquad=|b-a|\varepsilon
\end{aligned}$$

which completes the proof. ∎

The above result has many useful generalisations, in particular replacing the integral by multiple integrals or surface integrals. As long as we integrate over a compact set, essentially the same proof will work. However, for improper integrals uniform continuity will not work anymore.

For example, to prove a formula of the type

$$\frac{\mathrm{d}}{\mathrm{d}x}\bigg|_{x=x_0}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y = \int_{-\infty}^{\infty}\frac{\partial f}{\partial x}(x_0,y)\,\mathrm{d}y$$

uniform continuity is not enough, but if you can show, for every $\varepsilon > 0$, the existence of some $\delta > 0$ so that

$$\left|\frac{\partial f}{\partial x}(x,y) - \frac{\partial f}{\partial x}(x_0,y)\right| < \varepsilon g(y)$$

whenever $|x - x_0| < \delta$, and where the function g (independent of ε) satisfies $\int_{-\infty}^{\infty} g(x,y)\,\mathrm{d}y < \infty$, you can get the desired formula just as in the above proof.

48 Exercise. Complete the above argument.

The following result is sometimes called *Fubini's theorem*, but that is misleading – Fubini's theorem is much more general, and deals with Lebesgue integrable functions. The conclusion is the same, however. We include this simple special case because it is easy to prove with the tools at hand.

49 Proposition. *Assume that f is a continuous function on $[a,b]\times[c,d]$. Then*

$$\int_a^b\int_c^d f(x,y)\,dy\,dx = \int_c^d\int_a^b f(x,y)\,dx\,dy.$$

Proof: Replace b in both integrals by a variable ξ. Clearly, the integrals are equal when $\xi = a$. If we can show that both sides are differentiable with respect to ξ, with the same derivative, then they must have the same value for all ξ, including $\xi = b$.

First, note that f is uniformly continuous. A direct computation shows that $\int_c^d f(x,y)\,dy$ is a continuous function of x. Hence the fundamental theorem of calculus shows that

$$\frac{d}{d\xi}\int_a^\xi\int_c^d f(x,y)\,dy\,dx = \int_c^d f(\xi,y)\,dy.$$

Second, use the previous proposition to show that

$$\frac{d}{d\xi}\int_c^d\int_a^\xi f(x,y)\,dx\,dy = \int_c^d\frac{\partial}{\partial\xi}\int_a^\xi f(x,y)\,dx\,dy = \int_c^d f(\xi,y)\,dy.$$

∎

Ordinary differential equations

in which we state and prove the fundamental existence and uniqueness theorem.

In elementary calculus courses (the so-called “advanced calculus”), when discussing ordinary differential equations, the emphasis is on solutions by formula. Sure, the uniqueness of solutions for the initial value problem is often proved, but this is usually a result of the particular structure of the equation. In this section we will be concerned with the questions of existence and uniqueness in a more general setting.

The elementary theory can still throw some light on the general problem, and hint at what can and cannot be expected to hold true.

We shall be concerned with *initial value problems* — that is, problems of the form

$$\begin{aligned} \dot{x}(t) &= f(t, x(t)) \\ x(0) &= x_0 \end{aligned} \tag{1}$$

with given function f and initial value x_0. (We might, more generally, consider a given initial value $x(t_0) = x_0$ at some time t_0, but this generalisation is trivial. We shall always think of the independent variable t as time, though of course this would be misleading in many applications.)

Consider, for example, the equation $\dot{x} = x^2$. This separable equation is typically solved by formally rewriting it as $\mathrm{d}x/x^2 = \mathrm{d}t$ and integrating, with the result $x = 1/(\tau - t)$. (This method misses the trivial solution $x(t) = 0$, though.) If we are given an initial value $x(0) = x_0$, the integration constant must be given by $\tau = 1/x_0$. Thus, if $x_0 > 0$ then the solution goes to infinity (or “blows up”) at time $t = \tau = 1/x_0$.

Therefore, we cannot expect a general existence theorem to give *global* results, but we must settle for a *local* result instead: existence of a solution $x(t)$ for t in a neighbourhood of 0.

Another example is the equation $\dot{x} = x^{1/3}$, with the general solution $x = (2/3(t - \tau))^{3/2}$ in addition to the trivial solution $x(t) = 0$. We note that the general solution is only valid for $t > \tau$; however, we can make a solution valid everywhere by joining the general solution and the trivial one as follows:

$$x(t) = \begin{cases} 0 & t \le \tau, \\ (2/3(t-\tau))^{3/2} & t > \tau \end{cases}$$

However, this is an example of the breakdown of uniqueness, for the initial value problem with initial value $x(0) = 0$ has infinitely many solutions: Any solution like the one above with $\tau \ge 0$ will do. The problem lies with the right hand side $x^{1/3}$, which is too singular for uniqueness to hold.

The key to an existence and uniqueness result for the initial value problem (1) is to note that a function x defined on an interval surrounding 0 solves (1) if, and only if, it is continuous and satisfies the integral equation

$$x(t) = x_0 + \int_0^t f(\tau, x(\tau))\,\mathrm{d}\tau \tag{2}$$

for each t in the given interval.

We might imagine solving (2) by picking an arbitrary initial function $x_1(t)$ and proceeding to iterate:

$$x_{n+1}(t) = x_0 + \int_0^t f(\tau, x_n(\tau))\,\mathrm{d}\tau \tag{3}$$

and then hoping that x_n will converge to the desired solution function x as $n \to \infty$.

This is known as *Picard's method* and it does indeed work. We shall use Banach's fixed point theorem to show this. To carry out this program, then, we must first define a suitable complete metric space to be populated by functions $x(t)$.

To this end, we may replace the interval around 0 by an arbitrary metric space X. First, let $l^\infty(X)$ consist of all bounded, real-valued functions on X. We use the norm

$$\|f\|_\infty = \sup\{|f(x)| : x \in X\}$$

on this space. A sequence in $l^\infty(X)$ which converges in this norm is called *uniformly convergent*. In contrast, a sequence for which $f(x)$ converges for every $x \in X$ is called *pointwise convergent*.

50 Exercise. Show that a uniformly convergent sequence is pointwise convergent. Show that, given x, the map $f \mapsto f(x)$ is a continuous map from $l^\infty(X)$ to $\mathbb{R}$. What is the connection between these two statements?

Finally, show that a pointwise convergent sequence need not be uniformly convergent (let $f(x) = x/(1+x^2)$ and consider the sequence f_n, where $f_n(x) = f(nx)$).

51 Proposition. *$l^\infty(X)$ is a Banach space.*

Proof: Clearly, it is a normed space. We must show it is complete. Let (f_n) be a Cauchy sequence in $l^\infty(X)$. For each $x \in X$, $(f_n(x))$ is a Cauchy sequence in $\mathbb{R}$, and so has a limit which we will denote $f(x)$.

To show that $f_n \to f$ uniformly, let $\varepsilon > 0$. There is some N so that $\|f_n - f_m\| < \varepsilon$ whenever $m, n \geq N$. By the definition of the norm, that translates into $|f_n(x) - f_m(x)| < \varepsilon$ whenever $m, n \geq N$ and $x \in X$. Letting $m \to \infty$ in this inequality, we conclude $|f_n(x) - f(x)| \leq \varepsilon$ whenever $n \geq N$ and $x \in X$. But, again by the definition of the norm, this means $\|f_n - f\| \leq \varepsilon$ whenever $n \geq N$. In other words, $f_n \to f$ uniformly. ∎

The *continuous* functions in $l^\infty(X)$ form a subspace which we shall call $C_b(X)$.

52 Proposition. *$C_b(X)$ is a Banach space.*

Proof: By Proposition 21, we only need to show that $C_b(X)$ is closed. In other words, we must show that *a uniform limit of continuous functions is continuous.*

So let now $f_1, f_2, \ldots$ be continuous real-valued functions on X, and assume that $f_n \to f$ uniformly. We must show that f is continuous.

So let $x \in X$, and $\varepsilon > 0$. By uniform convergence, there is some N so that $n > N$ implies $\|f_n - f\|_\infty < \varepsilon$. Pick any $n > N$. Since f_n is continuous, there exists some $\delta > 0$ so that $|f_n(y) - f_n(x)| < \varepsilon$ whenever $y \in B_\delta(x)$. Now, for any such y we find

$$|f(y) - f(x)| \leq |f(y) - f_n(y)| + |f_n(y) - f_n(x)| + |f_n(x) - f(x)| < 3\varepsilon$$

since $|f(y) - f_n(y)| \leq \|f - f_n\|_\infty < \varepsilon$ (and similarly, $|f_n(x) - f(x)| < \varepsilon$). ∎

53 Definition. *A real function f defined on $\mathbb{R}$ is said to be Lipschitz continuous with Lipschitz constant L if $|f(x) - f(y)| \le L|x - y|$ for all x, y (this definition has an immediate generalisation to functions between arbitrary metric spaces, of course). Similarly, a function f of two variables is said to be uniformly Lipschitz continuous in the second variable if it satisfies $|f(t,x) - f(t,y)| \le L|x - y|$ for all t, x, y.*

54 Theorem. (Picard–Lindelöf) *Consider the initial value problem (1) where the right hand side f is defined and uniformly Lipschitz continuous in the second variable on a neighbourhood of $(t,x) = (0,x_0)$. Then (1) has a unique solution on some neighbourhood of $t = 0$.*

Proof: The simple idea of the proof is to use the Banach fixed point theorem on the function

$$\Phi(x)(t) = x_0 + \int_0^t f(\tau, x(\tau))\,\mathrm{d}\tau$$

since a fixed point of this function is a solution of (1), because (1) is equivalent to (2). The proof is somewhat complicated by the fact that $\Phi(x)$ may be undefined for some functions x, namely those x for which $x(t)$ is sometimes outside the domain of definition of f. So we use Corollary 20 instead.

First note that, given two functions x and y, we find

$$\begin{aligned}
|\big(\Phi(x) - \Phi(y)\big)(t)| &= \left|\int_0^t \big(f(\tau, x(\tau)) - f(\tau, y(\tau))\big)\,\mathrm{d}\tau\right| \\
&\le \int_0^t |f(\tau, x(\tau)) - f(\tau, y(\tau))|\,\mathrm{d}\tau \\
&\le \int_0^t L|x(\tau) - y(\tau)|\,\mathrm{d}\tau \\
&\le L|t| \cdot \|x - y\|
\end{aligned}$$

where L is the Lipschitz constant of f (in the second variable). Thus, to make Φ a contraction, we might restrict it to functions on the interval $[-T, T]$ where $T < 1/L$. Further, to use Corollary 20 we must have

$$\|\Phi(x_0) - x_0\|_\infty \le (1 - LT)r.$$

This norm is easily estimated:

$$\|\Phi(x_0) - x_0\|_\infty = \sup_{|t|\le T} \left| \int_0^t f(\tau, x(\tau))\,\mathrm{d}\tau \right| \le MT$$

where M is the maximum value of f on $[-T, T] \times \bar{B}_\varepsilon(x_0)$. By making T smaller if necessary (which does not increase M) we can achieve the inequality $MT \le (1 - LT)r$ which is exactly what we need to complete the existence proof.

Uniqueness also follows, at least locally, from which one can patch together a global uniqueness proof. It is better, however to use Grönwall's lemma to show uniqueness (but we shall not do so here). ∎

We have stated and proved the Picard–Lindelöf theorem for scalar initial value problems only. However the same proof works, without modification, for *systems* of first order equations: Just think of the unknown function as mapping an interval around 0 into $\mathbb{R}^n$, and f as defined on a suitable open subset of $\mathbb{R} \times \mathbb{R}^n$. It can in fact be seen to work in a yet more general setting, replacing $\mathbb{R}^n$ by a Banach space.

Furthermore, higher order equations are handled by reducing them to first order equations: For example, given a second order equation of the form $\ddot{x} = f(t, x, \dot{x})$ we put $y = \dot{x}$ and so arrive at the equivalent system $\dot{y} = f(t, x, y)$, $\dot{x} = y$.

A first order system of equations $\dot{x} = f(t, x)$ where f is Lipschitz defines a function ϕ by writing the solution to the system satisfying the initial condition $x(s) = \xi$ as $\phi(\xi, s, t)$; thus $\phi(\xi, s, t) = \xi$ and $\partial\phi/\partial t = f(\phi)$. It turns out (but we shall not prove it) that *if f is a C^1 function then ϕ is also C^1*. In other words, the solution of the system is a continuously differentiable function, not only of its parameter, but also of the initial conditions. We have now arrived at the beginnings of the theory of dynamical systems, and this is where we leave that theory.

The implicit and inverse function theorems

in which we state some conditions under which equations may be solved, and some properties of the solution.

The implicit function theorem concerns situations in which one can guarantee that an equation of the form $F(x, y) = 0$ defines y as a function of x; that is, when we can find a function g so that $F(x, g(x)) = 0$ for all x in some neighbourhood of a given point x_0 (and, moreover, we want some form of uniqueness, so that $F(x,y)=0$ has no solution $y\neq g(x)$, at least not locally).

The inverse function theorem concerns the existence of a local inverse of a given function f. Since the defining equation of the inverse, $f(g(x))=x$, can be written as $F(x,g(x))=0$ where $F(x,y)=f(y)-x$, it should be clear that the inverse function theorem will be a special case of the implicit function theorem. Thus we concentrate on the latter.

We shall need some definitions before we start.

The space of linear maps from $\mathbb{R}^n$ to $\mathbb{R}^m$ will be written $L(\mathbb{R}^n,\mathbb{R}^m)$. Of course we know that this can be identified with the space of all $m\times n$-matrices $\mathbb{R}^{m\times n}$, but mostly, we shall prefer the more abstract approach. On this space we use the operator norm defined by

$$\|A\| = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|}{\|x\|}.$$

(Much of what follows works just as well if the Euclidean spaces $\mathbb{R}^n$ are replaced by Banach spaces, and $L(X,Y)$ is the space of *bounded* linear operators from a Banach space X to a Banach space Y.)

55 Definition. *Let f be a function defined on a neighbourhood of a point $x\in\mathbb{R}^n$ and with values in $\mathbb{R}^m$. f is said to be (Fréchet) differentiable at x if there is some $A\in L(\mathbb{R}^n,\mathbb{R}^m)$ so that*

$$\lim_{\xi\to 0} \frac{\|f(x+\xi)-f(x)-A\xi\|}{\|\xi\|} = 0.$$

The operator A is uniquely determined (exercise!), and is called the derivative of f at x, written $A=Df(x)$.

This is more conveniently written as the first order Taylor's formula:

$$f(x+\xi) = f(x) + Df(x)\xi + o(\|\xi\|) \qquad (\xi\to 0)$$

where $o(\|\xi\|)$ is taken to mean some function $r(\xi)$ so that $r(\xi)/\|\xi\|\to 0$ as $\xi\to 0$.

56 Exercise. Show that, if f is differentiable at x with $Df(x)=A$, then all first order partial derivatives of f exist at x, and when A is interpreted as a matrix, we have

$$A_{ij} = \frac{\partial f_i}{\partial x_j}(x).$$

Show that the converse does not hold, for example by considering the function

$$f(x) = \begin{cases} \dfrac{x_1 x_2}{\sqrt{x_1^2 + x_2^2}} & x \neq 0, \\ 0 & x = 0. \end{cases}$$

Now, consider a function f defined in an open subset U of $\mathbb{R}^n$. The function is said to be C^1 if it is differentiable at each point of U, and the function $Df\colon U \to L(\mathbb{R}^n, \mathbb{R}^m)$ is continuous.

57 Proposition. *A function $f\colon U \to \mathbb{R}^m$ is C^1 in the open set $U \subseteq \mathbb{R}^n$ if and only if it has first order partial derivatives at each point of U, and those partial derivatives are continuous in U.*

Proof: We prove only the hard part, leaving the rest as an exercise. What we shall prove is the following: If f has continuous first order partial derivatives in a neighbourhood of 0, then f is differentiable there.

For brevity, we write $g_j = \partial f/\partial x_j$. Write e_j for the vector whose j component is 1, while all the others are 0 (so $e_1, \ldots, e_n$ is the standard basis of $\mathbb{R}^n$). Then $x = \sum_{j=1}^n x_j e_j$. We estimate $f(x) - f(0)$ by integrating the appropriate partial derivatives of f along the path consisting of straight line segments from 0 via $x_1 e_1$, $x_1 e_1 + x_2 e_2$, and so on to x: Let $\gamma_k(t) = \sum_{j=1}^{k-1} x_j e_j + t x_k e_k$ so that $\gamma_k(0) = \sum_{j=1}^{k-1} x_j e_j$, $\gamma_k(1) = \sum_{j=1}^{k} x_j e_j$, and $\gamma'_k(t) = x_k e_k$. Then

$$\begin{aligned} f(x) - f(0) &= \sum_{k=1}^n \Bigl[f\Bigl(\sum_{j=1}^k x_j e_j\Bigr) - f\Bigl(\sum_{j=1}^{k-1} x_j e_j\Bigr)\Bigr] \\ &= \sum_{k=1}^n \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma_k(t))\,\mathrm{d}t \\ &= \sum_{k=1}^n x_k \int_0^1 g_k(\gamma_k(t))\,\mathrm{d}t \\ &= \sum_{k=1}^n x_k g_k(0) + \sum_{k=1}^n x_k \int_0^1 [g_k(\gamma_k(t)) - g_k(0)]\,\mathrm{d}t \end{aligned}$$

The first sum on the last line is the desired linear function of x; it is $Df(0)x$. To show that this is really the Fréchet derivative, we must show the second sum is $o(\|x\|)$. But if $\varepsilon > 0$ is given we can find $\delta > 0$ so that $\|g_j(x) - g_j(0)\| < \varepsilon$ whenever $\delta > 0$. Clearly, for such an x and $0 \le t \le 1$ we have $\|\gamma_k(t)\| \le \|x\| < \delta$, and so each of the integrals in the second sum has norm $< \varepsilon$; hence their sum has norm $< \varepsilon \sum |x_j|$ and so it is $o(\|x\|)$ as $\|x\| \to 0$.

We leave the rest of the proof as an exercise. To show that f is C^1, note that once we know that f is differentiable, the matrix elements of Df are the partial derivatives of the components of f (in fact the columns of Df are the functions g_j). ∎

58 Proposition. (The chain rule) *Let g map a neighbourhood of $x \in \mathbb{R}^p$ into $\mathbb{R}^n$ and f map a neighbourhood of $g(x)$ into $\mathbb{R}^m$. If g is differentiable at x and f is differentiable at $g(x)$, then the composition $f \circ g$ is differentiable at x, with derivative*

$$D(f \circ g)(x) = Df\bigl(g(x)\bigr) \circ Dg(x)$$

When A and B are composable linear maps, it is more common to write their composition as BA rather than $B \circ A$. Thus the above formula is more commonly written as $D(f \circ g)(x) = Df(g(x))Dg(x)$.

Proof: Simply write

$$\begin{aligned} f \circ g(\xi) - f \circ g(x) &= Df\bigl(g(x)\bigr)\bigl(g(\xi) - g(x)\bigr) + o(\|g(\xi) - g(x)\|) \\ &= Df\bigl(g(x)\bigr)\bigl(Dg(x)(\xi - x) + o(\|\xi - x\|)\bigr) \\ &\quad + o\bigl(Dg(x)(\xi - x) + o(\|\xi - x\|)\bigr) \\ &= Df\bigl(g(x)\bigr)\bigl(Dg(x)(\xi - x)\bigr) + o(\|\xi - x\|) \end{aligned}$$

and the proof is complete. (Exercise: Write the argument out more carefully, dealing properly with all the epsilons and deltas.) ∎

We can now state and prove the implicit function theorem. First, however, let us consider the simple case of single variables. Clearly, the curve in Figure 1 is not the graph of a function. Nevertheless, some part surrounding the point (x_0, y_0) is the graph of a function $y = g(x)$. This illustrates the fact that we can only expect to be able to show the existence of a function g with $F(x, g(x)) = 0$ locally, that is, in some neighbourhood

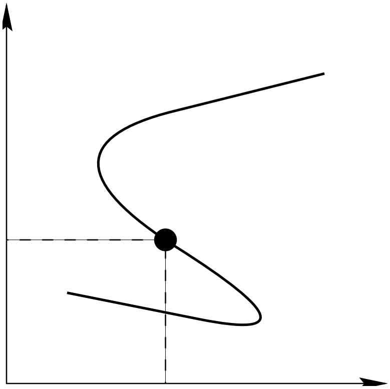

Figure 1: The curve $F(x,y)=0$ with a point (x_0,y_0) on it.

of x_0. The trouble spots seem to be where the curve has a vertical tangent or, equivalently, a horizontal normal. A normal vector is given by $\nabla F=(\partial F/\partial x,\partial F/\partial y)$, so the trouble spots are recognised by $\partial F/\partial y=0$.

We return to the general case of functions of several variables. If $x\in\mathbb{R}^m$ and $y\in\mathbb{R}^n$, we may write the vector $(x_1,\dots,x_m,y_1,\dots,y_n)\in\mathbb{R}^{m+n}$ as (x,y). If F is a function of (x,y), we write $D_yF(x,y)\in L(\mathbb{R}^n,\mathbb{R}^n)$ as $D_yF(x,y)\eta=DF(x,y)(0,\eta)$. We then note that the chain rule applied to the equation $F(x,g(x))=0$ yields $D_xF(x,y)+D_yF(x,g(x))Dg(x)=0$, so if $D_yF(x,g(x))$ is invertible, we find

$$Dg(x)=-D_yF(x,g(x))^{-1}D_xF(x,y).$$

It turns out that this invertibility condition, which gives us a unique value of $Dg(x)$, is sufficient to define the function g in a neighbourhood of x.

59 Theorem. (Implicit function theorem) *Assume given an $\mathbb{R}^n$-valued C^1 function F on a neighbourhood of $(x_0,y_0)\in\mathbb{R}^m\times\mathbb{R}^n$. Assume that $F(x_0,y_0)=0$, and that $D_yF(x_0,y_0)$ is invertible. Then there is a neighbourhood U of x_0 and a C^1 function $g\colon U\to\mathbb{R}^n$ with $g(x_0)=y_0$ and $F(x,g(x))=0$ for all $x\in U$.*

Proof: By replacing $F(x,y)$ by $F(x-x_0,y-y_0)$ we may assume $x_0=0$ and $y_0=0$. If A is an invertible $n\times n$ matrix, $F(x,y)=0$ is equivalent to

$AF(x,y) = 0$; hence we may replace F by AF. If we let $A = D_yF(0,0)^{-1}$ this means we may, and indeed shall, assume $D_yF(0,0) = I$.

Write $F(x,y) = y - H(x,y)$; thus $F(x,y) = 0 \Leftrightarrow y = H(x,y)$, and moreover $D_yH(0,0) = 0$. Given x, we propose to solve the equation $y = H(x,y)$ by the iteration $y_{n+1} = H(x,y_n)$ with $y_0 = 0$. We must show that $H(x,y)$, as a function of y, is a contraction. But

$$\begin{aligned} H(x,y) - H(x,z) &= \int_0^1 \frac{\partial}{\partial t} H(x, ty + (1-t)z)\,\mathrm{d}t \\ &= \int_0^1 D_yH(x, ty + (1-t)z)\,\mathrm{d}t \cdot (y - z) \end{aligned}$$

and since $D_yH(0,0) = 0$ and H is C^1, there is some $\varepsilon > 0$ so that, whenever $\|x\| \le \varepsilon$ and $\|y\| \le \varepsilon$, we have $\|D_yH(x,y)\| \le 1/2$. For such x and y, then, the above equality implies

$$\|H(x,y) - H(x,z)\| \le \frac{1}{2}\|y - z\|.$$

Clearly, then, for fixed x with $\|x\| < \varepsilon$, the map $y \mapsto H(x,y)$ is a contraction of $B_\varepsilon(0)$ into $\mathbb{R}^n$. This ball may however not be invariant for all x (that is, the map $y \mapsto H(x,y)$ may not map the ball into itself). However, if x is small enough, we shall see that Corollary 20 comes to the rescue. In fact, all that remains is to satisfy the second inequality in that Corollary, with $K = \frac{1}{2}$ and $r = \varepsilon$. In our current setting, that inequality becomes simply

$$|H(x,0)| \le \frac{1}{2}\varepsilon.$$

Since H is continuous and $H(0,0) = 0$, we can find a $\delta > 0$ so that the above inequality holds whenever $|x| < \delta$. By Corollary 20, then, there is therefore a unique solution $y = g(x) \in \bar{B}_\varepsilon(0)$ of the equation $y = H(x,y)$ wehenever $x < \delta$, and thus of the equation $F(x,y) = 0$.

It only remains to establish the C^1 nature of g. To this end, consider

$$\begin{aligned}
0 &= F\bigl(\xi, g(\xi)\bigr) - F\bigl(x, g(x)\bigr) \\
&= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} F\bigl(t\xi + (1-t)x, tg(\xi) + (1-t)g(x)\bigr)\,\mathrm{d}t \\
&= \int_0^1 DF\bigl(t\xi + (1-t)x, tg(\xi) + (1-t)g(x)\bigr)\,\mathrm{d}t \cdot \bigl(\xi - x, g(\xi) - g(x)\bigr) \\
&= \underbrace{\int_0^1 D_xF\,\mathrm{d}t}_{A} \cdot (\xi - x) + \underbrace{\int_0^1 D_yF\,\mathrm{d}t}_{B} \cdot \bigl(g(\xi) - g(x)\bigr)
\end{aligned}$$

where the matrices A and B satisfy inequalities of the form $\|A\| < C$ and $\|I - B\| < 1/2$ (the latter comes from the inequality $\|D_yH(x,y)\| < 1/2$, and the former is just the boundedness of D_xF in a neighbourhood of 0). But then B is invertible with $\|B^{-1}\| < 2$ (because $B^{-1} = \sum_{k=0}^{\infty}(I-B)^k$), and we have

$$g(\xi) - g(x) = -B^{-1}A(\xi - x).$$

In the limit $\xi \to x$, we find $A \to D_xF(x, g(x))$ while $B \to D_yF(x, g(x))$ from which the differentiability of g is easily shown. Furthermore,

$$Dg(x) = -D_yF(x, g(x))^{-1} D_xF(x, g(x))$$

which is continuous, so that g is C^1. ∎

60 Corollary. (Inverse function theorem) *Assume there is given an $\mathbb{R}^n$-valued C^1 function f on a neighbourhood of $y_0 \in \mathbb{R}^n$, and that $Df(y_0)$ is invertible. Then there is a neighbourhood U of $f(y_0)$ and a neighbourhood V of y_0 so that f maps V invertibly onto U, and the inverse map $g\colon U \to V$ is C^1.*

Proof: Write $F(x, y) = f(y) - x$ and apply the implicit function theorem to F at the point (x_0, y_0) where $x_0 = f(y_0)$. Thus there is a neighbourhood U_0 of x_0 and a C^1 function g on U_0 so that $f(g(x)) = x$ whenever $x \in U_0$.

By the chain rule, $Df(y_0)Dg(x_0) = I$ so that $Dg(x_0)$ is invertible, and by what we just proved (applied to g instead of to f) there is a neighbourhood V of y_0 and a C^1 function $h\colon V \to \mathbb{R}^n$ so that $g(h(y)) = y$ whenever $y \in V$.

Let $U = g^{-1}[V]$. It is not hard to show that, in fact, $h(y) = f(y)$ whenever $y \in V$, and the restrictions $f|_V$ and $g|_U$ are each other's inverses.

∎

Index