"Fundamentals of analysis"

The course went from YBC7289 and the pyramids until Chapter 12, the notes were scrutinized by Bob, Harold and Thomas

Section 13.2 is the first return to YBC7289, the second return is in (16.29)

Chapter 14 is back to 1-variable calculus

Chapter 15 is the implicit function theorem from a 1-variable perspective

Chapter 16 is mainly about the Morse lemma, in the end avoiding the implicit function theorem

Chapter 18: multivariate calculus, analysis unpacked

Chapter 19 contains the basic statements about measures of parallelotopes

Chapter 20 is about Lagrange multipliers

Chapter 21 is about integration with more variables and a quick version of Green's Theorem

Chapter 22 is about Fourier series, to be adapted to the present context

Working on and rearranging Chapter 21 and Chapter 26: (line/surface) integrals and all that, also with forms

Section 25 relates to PDE's

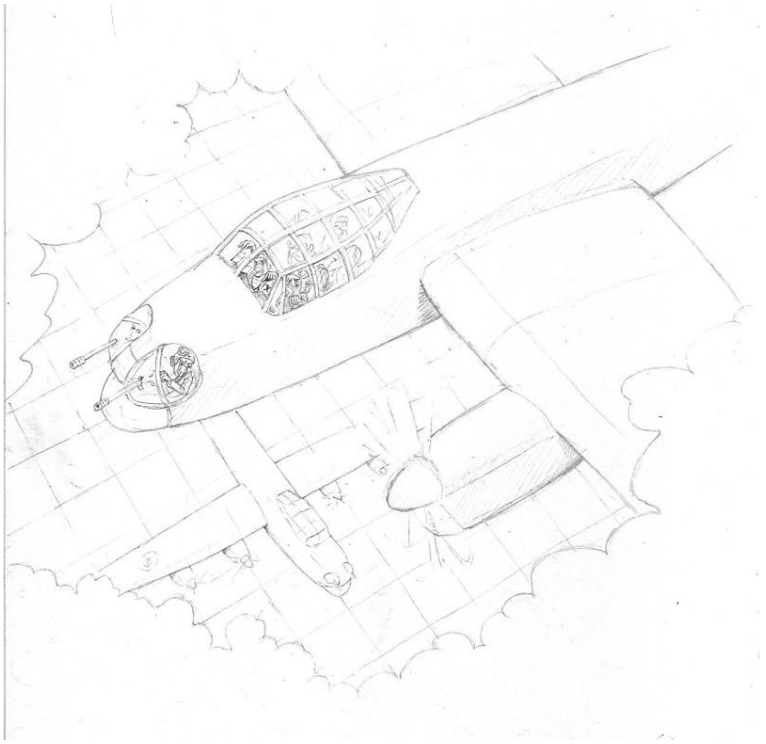
Chapter 28 is about the general Stokes Theorem and borrows a little from Chapter 27

jhf400@vu.nl, Oegstgeest, Amsterdam (2019)

Eventuele auteursinkomsten[1] komen ten goede aan de "geluksmachine":

`http://orchestra18c.com`

[1] Als het ervan komt...



The Hague, March 3, 1945

Contents

After the first 11 chapters, the student...

1. ... knows basic definitions concerning limits and continuity (convergence, Cauchy sequence, limit, completeness, continuity, uniform continuity) and is able to determine whether a sequence, series or function satisfies these definitions;

2. ... knows the definition of differentiability (i.e., that a function can be approximated by a linear one), can determine whether a function is differentiable, and is familiar with the more algebraic approach for power series);

3. ... knows the definition of Riemann integrability and can prove that certain functions (in particular, polynomials, monotone and uniformly continuous functions) are Riemann integrable, and knows the limit theorems about limits of integrals of uniformly convergent sequences of functions on [a,b], and of point wise convergent monotone functions;

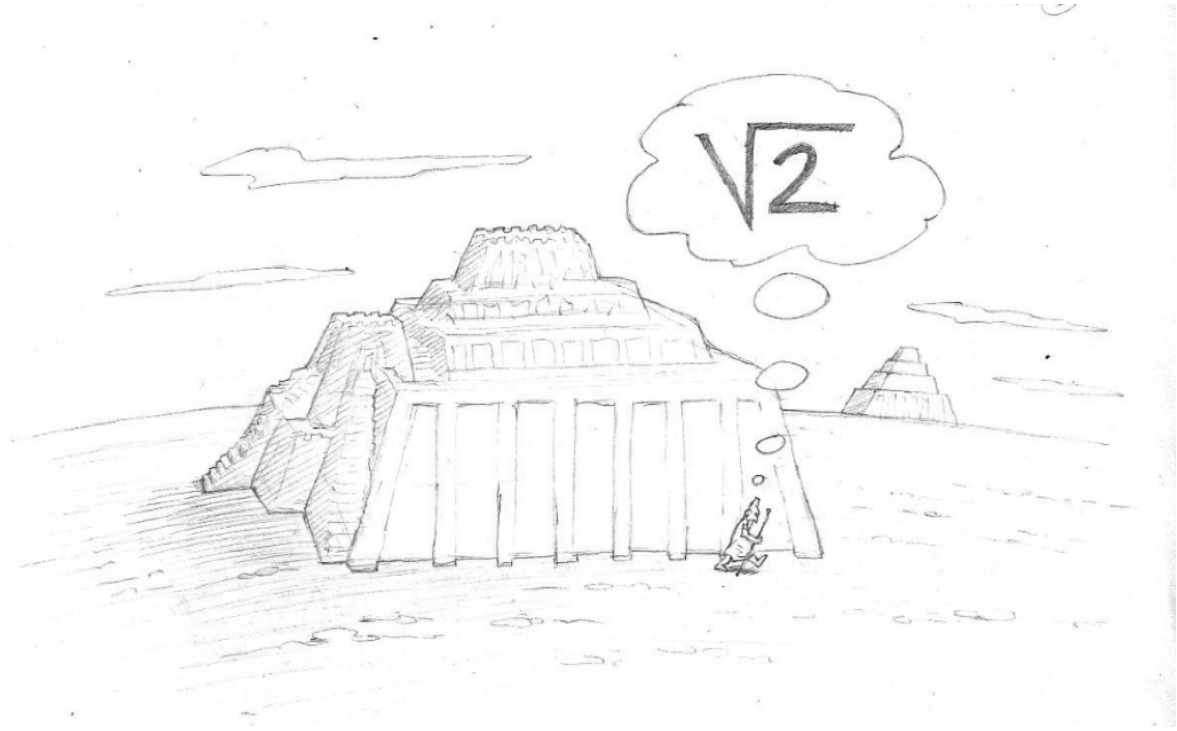
4. ... knows the definition of basic concepts from metric topology (metric, convergence, completeness, Banach space) and can prove that certain sets of functions satisfy these definitions, such as $C([a,b])$, the space of continuous functions $f : [a,b] \to \mathbb{R}$ with the uniform metric, and knows that convergence in this space corresponds to uniform convergence;

5. ... knows the statement of the Banach Fixed Point Theorem, and can apply this theorem to solve fixed point equations, in particular integral equations in $C([a,b])$ for solutions of differential equations.

Course content

This course treats the rigorous mathematical theory behind Calculus: limits, continuity, linear approximation, differentiability, integrability, and the mutual relation between these concepts. The mathematical tools that are necessary for formulating and proving the essential results of Calculus are first presented in the context of real valued sequences and real valued functions of a real variable, in such a way that everything can later be generalised (to Y-valued functions of variables in X, with X and Y Banach spaces). The space $C([a,b])$ of real valued continuous functions on an interval $[a,b]$ will appear as the first example of such a Banach space.

Starting point of the course are an ancient iterative scheme for solving equations, and the fundamental properties of (the set of) real numbers. Highlights: a fairly complete exposition of power series directly based on a systematic algebraic approach for monomials, and if time permits an early introduction of the Implicit Function Theorem via a contraction argument and the Banach Fixed Point Theorem.


√2

'I like fonctions of one variable'

Xavier Cabré adressing Abel prize winner Louis Nirenberg and a small analysis group at Tor Vergata in June 2015.

1 Introduction

These lecture notes for an analysis course for first year students of mathematics and what can follow later. First year topics covered are

1. Cauchy sequences, convergence, limits;
2. Completeness of the real numbers; theorem of Bolzano-Weierstrass;
3. Continuity and uniform continuity;
4. The concept of differentiability;
(including differentiability of power series);
5. The concept of Riemann integrability (including Riemann integrability of monotone and uniformly continuous functions);
6. The language of metric topology;
7. Completeness of the space $C([a,b])$, uniform convergence;
8. The Banach Fixed Point Theorem (with applications to integral and differential equations, and the implicit function theorem).

Some of these terms may mean nothing to you yet. This introduction is meant to give you a flavour of how and what we do in analysis, with some historical perspective. We introduce some of the notation along the way, as well as a few basic principles. Some familiarity with what used to be highschool calculus is assumed: limits, continuity, differentiability and integration, in the context of real valued functions $f(x)$ of a real variable x. In particular you have probably seen the integration formula

$$\int_a^b f(x)\,dx = F(b) - F(a),$$

in which F is a primitive function of f, meaning that the derivative of $F(x)$ is given by $F'(x) = f(x)$.

Perhaps you have also seen the Newton scheme

$$x_n = x_{n-1} - \frac{g(x_{n-1})}{g'(x_{n-1})} =: f(x_{n-1}) \tag{1.1}$$

for solving the equation $g(x) = 0$ numerically.

Exercise 1.1. Consider the graph defined by $y = g(x)$. Use your highschool maths to write down a formula for the line tangent to the graph of g in the point $(x, y) = (x_{n-1}, g(x_{n-1}))$. Intersect this line with the x-axis and denote the x-value in the intersection point by x_n. Show that it is given by (1.1). Hint: make a picture first, for instance if g is given by $g(x) = x^2 - 2$.

Starting with some x_0 this scheme produces a sequence x_1, x_2, ..., which typically converges to a solution of $g(x) = 0$ very fast, see Chapter 13.2.

Exercise 1.2. Show for $g(x) = x^2 - 2$ that (1.1) reduces to

$$x_n = f(x_{n-1}) := \frac{x_{n-1}}{2} + \frac{1}{x_{n-1}} \tag{1.2}$$

and experiment, with $x_0 = 1$ as starting value for instance.

1.1 The square root of two

The example in Exercises 1.1,1.2 takes us way back to Babylonian times, and the origins of differential calculus. It concerns $\sqrt{2}$, a geometric number which appears as the length of the diagonal in the unit square.

The first recorded attempt[1] to compute the positive number r defined by $r^2 = 2$ can be found on the Babylonian clay tablet YBC7289. Dating back around 37 centuries, it contains the picture of a square with its diagonals, and several number sequences written in cuneiform.

In decimal notation one of these number sequences is

$$1 \quad 24 \quad 51 \quad 10$$

and stands for[2]

$$1 + \frac{24}{60} + \frac{51}{3600} + \frac{10}{216000} = 1.41421\underline{296},$$

which is a remarkably good *hexagesimal* approximation of

$$\sqrt{2} = 1.4142135\ldots,$$

the irrational square root of 2.

[1] That I know of.

[2] The repeating part of the decimal expansion is underlined.

In our notation this approximation is believed to have resulted from rather clever calculations employing the approximation

$$\sqrt{1+x} \approx 1 + \frac{x}{2}$$

for small x. The clarifying formula would be that

$$\sqrt{2} \approx \frac{577}{408} \approx 1 + \frac{24}{60} + \frac{51}{3600} + \frac{10}{216000},$$

in which the Babylonian approximation is a truncated hexigesimal expansion for $\frac{577}{408}$. This works as follows.

Let $r > 0$ be a possibly not so very good approximation of $\sqrt{2}$. Then

$$\sqrt{2} = \sqrt{r^2 + 2 - r^2} = r\sqrt{1 + \frac{2-r^2}{r^2}} \approx r\left(1 + \frac{1}{2}\frac{2-r^2}{r^2}\right) = \frac{r}{2} + \frac{1}{r},$$

which is possibly a better approximation of $\sqrt{2}$. You should recognise the example of Newton's method in Exercises 1.1,1.2. Starting with the bad approximation $r = 1$ the new approximation of $\sqrt{2}$ is $\frac{3}{2}$, which is not that bad really. Redoing the approximation with $r = \frac{3}{2}$ gives $\frac{17}{12}$, much better, and $r = \frac{17}{12}$ in turn gives

$$\frac{17}{24} + \frac{12}{17} = \frac{289 + 288}{24 \times 17} = \frac{577}{408} \approx 1 + \frac{24}{60} + \frac{51}{60^2} + \frac{10}{60^3},$$

the approximation on YBC7289, which is where the Babylonians apparently stopped.

This method for approximating $\sqrt{2}$ is also known as Heron's method. In this course we will take these methods to the limit. In Chapter 2 we will give a proper formulation and proof of the statement that the sequence x_n defined in Exercise 1.2 has the property that

$$x_n \to \sqrt{2} \quad \text{as} \quad n \to \infty, \tag{1.3}$$

to be pronounced as "x_n goes to $\sqrt{2}$ as n goes to infinity". We shall show that it does so extremely fast.

1.2 One third of what?

Another geometric number is $\frac{1}{3}$. It appears as the volume V of a pyramid with unit square base and unit height. To see how and why we divide this pyramid into 10 horizontal layers of height $\frac{1}{10}$ and write n for 10. The maximal width of each layer varies from 1 at the bottom to $\frac{1}{10} = \frac{1}{n}$ at the top.

Exercise 1.3. Draw a picture and convince yourself that from top to bottom these maximal widths are

$$\frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \cdots, 1.$$

Thus the total volume V of the "unit" pyramid is certainly less than

$$\frac{1}{n}\left(\frac{1}{n^2} + \frac{4}{n^2} + \frac{9}{n^2} + \cdots + 1\right) = \frac{1}{n^3}\sum_{k=1}^{n} k^2.$$

Likewise the minimal widths of the layers are

$$\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \cdots, \frac{n-1}{n},$$

so V is larger than

$$\frac{1}{n^3}\sum_{k=0}^{n-1} k^2.$$

Combining the two bounds we have

$$\underline{S}_n := \frac{1}{n^3}\sum_{k=0}^{n-1} k^2 < V < \frac{1}{n^3}\sum_{k=1}^{n} k^2 =: \bar{S}_n, \quad \text{while} \quad \bar{S}_n - \underline{S}_n = \frac{n^2}{n^3} - \frac{0}{n^3} = \frac{1}{n},$$

in which we don't really have to exhaust ourselves to understand that this is also true for values of n different from 10 as large as we like.

How many numbers V can satisfy this inequality for all n? At most one according to Archimedes. Because for two such numbers, say $V < W$, we would have

$$0 < W - V < \bar{S}_n - \underline{S}_n = \frac{1}{n} \quad \text{for all} \quad n \in \mathbb{N}. \tag{1.4}$$

Archimedes took it for granted that therefore the difference of V and W must be zero, and who are we to dispute? As a consequence of what we now call the Archimedean Principle there is indeed at most one number that qualifies as the volume of the pyramid.

By the way, Archimedes also knew the identity

$$(C_n) \qquad \sum_{k=1}^{n} k^2 = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6},$$

so the inequalities become

$$\frac{1}{3} - \frac{1}{2n} + \frac{1}{6n^2} < V < \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}$$

and we see that $V = \frac{1}{3}$ fits. If we agree that the unit pyramid has a volume, then its volume must be $\frac{1}{3}$ because it is the *only* value that fits[3]. It's quite amusing that we actually found this value as the coefficient of n^3 in (C_n).

In modern language we say that V is the integral

$$\int_0^1 (1-z)^2\, dz = \frac{1}{3},$$

in which $(1-z)^2$ is the area of the intersection of the pyramid with a horizontal plane at height z. The integration variable z ranges from $z = 0$ at the bottom to $z = 1$ at the top of the pyramid.

Having guessed (C_n) one way or another you can prove it by induction: starting with $n = 1$ and (C_1) being a statement that is trivially true, the implication

$$(C_n) \implies (C_{n+1})$$

is easy to verify. Indeed, using (C_n) we have that

$$\sum_{k=1}^{n+1} k^2 = \sum_{k=1}^{n} k^2 + (n+1)^2 = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} + (n+1)^2,$$

which happens to be equal to

$$\frac{(n+1)^3}{3} + \frac{(n+1)^2}{2} + \frac{n+1}{6}.$$

So (C_{n+1}) holds if (C_n) holds. This is called the induction step, which here is valid for every $n \geq 1$. Verifying (C_1) via

$$\sum_{k=1}^{1} k^2 = 1^2 = 1 = \frac{1^3}{3} + \frac{1^2}{2} + \frac{1}{6}$$

we then conclude that for every natural number n the identity (C_n) holds because

$$C_1 \implies C_2 \implies C_3 \implies C_4 \implies \cdots$$

This trick to prove (C_n) for all positive integers n is also called proof by induction. Think of the n^{th} statement (C_n) as being written on the n^{th}

[3]There is no obvious way to think of this volume as one third of the unit cube!

domino. Put all dominos in a never ending queue. Kick the first domino ($n = 1$) over and watch. The statements still to be checked are the dominos still standing.

You may have noted that[4]

$$\int_0^1 (1-z)^2\,dz = \int_0^1 x^2\,dx.$$

This integral belongs to a family

$$J_1 = \int_0^1 x\,dx = \frac{1}{2}, \quad J_2 = \int_0^1 x^2\,dx = \frac{1}{3}, \quad J_3 = \int_0^1 x^3\,dx = \frac{1}{4}, \ldots,$$

expressions that you must have seen before for the area J_p of the set

$$A_p = \{(x, y) : 0 \le y \le x^p \le 1\}$$

in the xy-plane.

Archimedean type expressions for sums of powers can be used to show directly that the sequence $J_1, J_2, J_3, \ldots$ continues as suggested. Unfortunately the sum formulas for exponents p larger than 3 become a bit cumbersome. The inequalities

$$\sum_{k=0}^{n-1} k^p < \frac{n^{p+1}}{p+1} < \sum_{k=1}^{n} k^p$$

do a quicker job. They hold for all positive integers p, n and dividing by n^{p+1} it follows that

$$\frac{1}{n^{p+1}} \sum_{k=0}^{n-1} k^p < \frac{1}{p+1} < \frac{1}{n^{p+1}} \sum_{k=1}^{n} k^p$$

for lower and upper approximations of J_p. Since these approximations differ by $\frac{1}{n}$, Archimedes tells us that

$$J_p = \int_0^1 x^p\,dx = \frac{1}{p+1}. \tag{1.5}$$

holds for every positive integer p.

An important goal in this course will be to give a rigorous meaning to integrals such as (1.5). The above reasoning will guide us in Chapter 6.

[4]Via the substitution $z = 1 - x$.

1.3 The Archimedean Principle

We continue this introduction with an overview of the different number sets that we use in analysis, tied up with Archimedes' principle. You are of course familiar with

$$\mathbb{Z} = \{\dots, -4, -3, -2, -1, 0, 1, 2, 3, 4, \dots\} \subset \mathbb{Q} = \left\{\frac{p}{q} : p \in \mathbb{Z},\, q \in \mathbb{N}\right\},$$

the set of all integers and the set of all rationals. We think of $\mathbb{Z}$ as a bi-infinite sequence of marked points on a number line with no endpoints. The other numbers of $\mathbb{Q}$ lie in the intervals between. If $r \in \mathbb{Q}$ is not in $\mathbb{Z}$ then $r = m + q$ with $m \in \mathbb{Z}$, $q \in \mathbb{Q}$ and $0 < q < 1$.

Many geometrically defined numbers such as π and $\sqrt{2}$ are not rational and correspond to other points on the number line, which we think of as corresponding to the set $\mathbb{R}$ of all real numbers. Thus

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}.$$

Beginning with $\mathbb{N}$ all of these are sets with infinitely many elements, as they all contain the infinite set $\mathbb{N}$ enumerated by $1, 2, 3, \dots$. It is also easy to enumerate $\mathbb{Q}$, but you really should convince yourself that such a one-to-one correspondence between $\mathbb{N}$ and the set of all points on the real number line cannot exist.

To wit, assume

$$x_1, x_2, x_3, \dots$$

is an enumeration of $\mathbb{R}$. Then $\mathbb{R}$ is completely covered by the intervals[5]

$$\left(x_1 - \frac{1}{4}, x_1 + \frac{1}{4}\right), \left(x_2 - \frac{1}{8}, x_1 + \frac{1}{8}\right), \left(x_3 - \frac{1}{16}, x_3 + \frac{1}{16}\right),$$

$$\left(x_4 - \frac{1}{32}, x_4 + \frac{1}{32}\right), \left(x_5 - \frac{1}{64}, x_5 + \frac{1}{64}\right), \left(x_6 - \frac{1}{128}, x_6 + \frac{1}{128}\right),$$

etcetera. The total length of these covering intervals is at most

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} + \frac{1}{128} + \frac{1}{256} + \cdots,$$

which I hope you agree is 1. Similar reasoning would bound the total length by $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and so on. This is an absurdity that we are not willing to accept: the total length[6] of the real number line should be larger than any positive number. Have we proved the following theorem?

[5] For numbers $a < b$ we denote by (a, b) the set of all real numbers x with $a < x < b$.

[6] We touch upon measure theory here, see Section 8.4.

Theorem 1.4. *The set* $\mathbb{R}$ *of real numbers is not enumerable. In other words,* $\mathbb{R}$ *is not a sequence of numbers.*

A more direct proof of Theorem 1.4 is via never ending decimal expansions. Indeed: one possible and very natural definition of the set $\mathbb{R}$ of real numbers is by means of such expansions. Assume that the real numbers between 0 and 1 are enumerated by

$$x_n = \sum_{j=1}^{\infty} \frac{d_{nj}}{10^j} \quad \text{for} \quad n = 1, 2, 3, \ldots,$$

and put the digits[7] d_{nj} in a block

$$\begin{array}{cccccccc}
d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} & d_{17} & d_{18} & \ldots \\
d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} & d_{27} & d_{28} & \ldots \\
d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} & d_{37} & d_{38} & \ldots \\
d_{41} & d_{42} & d_{43} & d_{44} & d_{45} & d_{46} & d_{47} & d_{48} & \ldots \\
d_{51} & d_{52} & d_{53} & d_{54} & d_{55} & d_{56} & d_{57} & d_{58} & \ldots \\
d_{61} & d_{62} & d_{63} & d_{64} & d_{65} & d_{66} & d_{67} & d_{68} & \ldots \\
d_{71} & d_{72} & d_{73} & d_{74} & d_{75} & d_{76} & d_{77} & d_{78} & \ldots \\
d_{81} & d_{82} & d_{83} & d_{84} & d_{85} & d_{86} & d_{87} & d_{88} & \ldots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots &
\end{array}$$

Now choose d_n with $|d_n - d_{nn}| = 2$ and observe that the real number

$$\sum_{j=1}^{\infty} \frac{d_j}{10^j}$$

does not appear as any x_n in our enumeration, a contradiction.

To make decimal representations unique, we may choose to exclude expansions which only have finitely many nonzero digits. The number $1 \in \mathbb{N}$ is then represented in $\mathbb{R}$ as

$$1 = 0.9999999\ldots = \frac{9}{10} + \frac{9}{100} + \frac{9}{1000} + \cdots, \tag{1.6}$$

[7]Which can be any of $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$.

whence for example

$$\frac{1}{9} = 0.11111111\ldots = \frac{1}{10}+\frac{1}{100}+\frac{1}{1000}+\cdots = \frac{1}{10}+\frac{1}{10^2}+\frac{1}{10^3}+\cdots = \sum_{n=1}^{\infty}\frac{1}{10^n}.$$

This is just like

$$1 = \frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\cdots = \sum_{n=1}^{\infty}\frac{1}{2^n}, \tag{1.7}$$

which relates to binary representations of the real numbers.

The equalities in the above expressions relate to the Archimedean principle again. For instance, the absolute value of the difference between 1 and

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\cdots$$

is clearly smaller than every power of $\frac{1}{2}$, and thus smaller than every $\frac{1}{n}$. According to Archimedes it must thus be zero. We shall honour Archimedes by stating his principle as a theorem in which we use the modern symbols $\forall$ and $\exists$.

Theorem 1.5. *The Archimedean Principle:*

$$\forall_{\varepsilon>0}\,\exists_{N\in\mathbb{N}}:\ \frac{1}{N} < \varepsilon.$$

Theorem 1.5 implies that there is no strictly positive[8] real number smaller than every $\frac{1}{n}$, which is what we used to conclude from (1.4) that the only candidate for the volume of the pyramid is $\frac{1}{3}$. Our task will be to understand the mathematical proof of what was obvious to Archimedes[9].

Remark 1.6. *If we use the Archimedean Principle with $\varepsilon = \frac{1}{x}$, in which $x \in \mathbb{R}$ is positive, we obtain the equivalent statement*

$$\forall_{x>0}\,\exists_{N\in\mathbb{N}}:\ N > x.$$

Exercise 1.7. We used the symbol N to exhibit that the statements in Theorem 1.5 and Remark 1.6 concern the existence of a single N. For which n other than $n = N$ do these Archimedean statements also hold?

[8] By positive we mean strictly positive from now on!

[9] Don't we do important work?

1.4 The geometric series

See

`https://en.wikipedia.org/wiki/Geometric_series`

for the title of this subsection. We have seen in Section 1.3 that in the set $\mathbb{R}$ it must hold that

$$\frac{1}{10}+\frac{1}{10^2}+\frac{1}{10^3}+\frac{1}{10^4}+\frac{1}{10^5}+\cdots=\frac{1}{10-1}.$$

Substituting $10=n$ we "discover" that

$$\frac{1}{n}+\frac{1}{n^2}+\frac{1}{n^3}+\frac{1}{n^4}+\frac{1}{n^5}+\cdots=\frac{1}{n-1}. \tag{1.8}$$

It's easy to convince yourself why (1.8) should be true for every integer $n>1$: order one pizza for $n-1$ persons, slice it in n pieces, eat, slice, eat, and so. If you have been born with n fingers ($n>1$) you are likely to discover (1.8) as a fact of every day arithmetic life, long before you eat pizza's. Have a look at

`https://en.wikipedia.org/wiki/Zeno_of_Elea`

before we continue but don't spend too long there.

For $x\in\mathbb{R}$ the more general expression

$$\sum_{n=0}^{\infty}x^n=1+x+x^2+x^3+x^4+\cdots \tag{1.9}$$

is called a geometric series. The formula

$$\sum_{n=0}^{N}x^n=1+x+x^2+\cdots+x^N=\frac{1-x^{N+1}}{1-x} \tag{1.10}$$

for the finite sums leads to a remarkable conclusion.

Theorem 1.8. *For $x\in\mathbb{R}$ it holds that*

$$\sum_{n=0}^{\infty}x^n=\frac{1}{1-x}\quad \textit{if}\quad |x|<1. \tag{1.11}$$

Exercise 1.9. Sketch the graphs defined by

$$y=\frac{1}{1-x},\quad y=1+x,\quad y=1+x+x^2,\quad y=1+x+x^2+x^3,\ldots$$

to *see* what this actually means.

In particular it follows for all x with $|x| < 1$ that[10]

$$\sum_{n=1}^{\infty} x^n = x + x^2 + x^3 + \cdots = \frac{x}{1-x},$$

which reduces to (1.8) for $x = \frac{1}{n}$, but is a far more general statement[11].

A mathematical proof of Theorem 1.8 first of all requires an algebraic proof of (1.10), i.e. that

$$\sum_{n=0}^{N} x^n = \underbrace{\frac{1-x^{N+1}}{1-x} = \frac{1-x^{N+1}}{1-x}}_{\text{LaTe}\chi \text{ sucks}},$$

and then a limit argument for $N \to \infty$, which boils down to the statement that

$$x^N \to 0 \quad \text{as} \quad N \to \infty \quad \text{if} \quad |x| < 1. \tag{1.12}$$

You should contrast this with[12]

$$\sqrt[n]{x} \to 1 \quad \text{as} \quad n \to \infty \quad \text{if} \quad x > 0. \tag{1.13}$$

Making such and other limits statements mathematically sound is another important task for this course, but let's also not forget the algebraic beauty in (1.10).

Exercise 1.10. We turn (1.10) around. Show that for every $x \neq 1$ it holds that

$$\frac{x^n - 1}{x - 1} = 1 + x + x^2 + \cdots + x^{n-1},$$

and observe that the right hand side is equal to n if $x = 1$. Generalise to

$$\frac{x^n - a^n}{x - a}$$

with a and x in $\mathbb{R}$. What does this tell you about the line tangent to the graph defined by $y = x^n$ in $x = a$?

Your answer to Exercise 1.10 will be our starting point for the theory of differentiation and guide us in an approach that avoids the usual limits.

[10] Subtracting 1 on both sides, or multiplying by x.

[11] Play with this formula, for instance, replace x by $-x$ and draw some graphs.

[12] How would you use (1.12) to prove (1.13)?

1.5 Outlook: beyond the real numbers

A small detour: you will see elsewhere that (1.11) is true even in the form

$$\sum_{n=0}^{\infty} A^n = (I - A)^{-1}, \tag{1.14}$$

where A is a square matrix, and in which A^n is a *matrix product*[13], i.e.

$$A^2 = A\,A,\ A^3 = A\,A\,A,\ A^4 = A\,A\,A\,A,$$

and so on. To give a rigorous meaning to (1.14), we will in fact need a condition of the form $|A| < 1$. A possible "absolute value" of a matrix is

$$|A|_{\text{Frob}} = \sqrt{\sum_{i,j} A_{ij}^2},$$

the square root of the sum of all the squared entries of A. This norm is called the Frobenius[14] norm of A. The Frobenius norm has the remarkable properties that

$$|A\,B|_{\text{Frob}} \leq |A|_{\text{Frob}}\,|B|_{\text{Frob}} \quad \text{and} \quad |A + B|_{\text{Frob}} \leq |A|_{\text{Frob}} + |B|_{\text{Frob}} \tag{1.15}$$

for all square[15] matrices A and B of the same size.

In this course we will not so much study matrices and matrix norms. However, we will often work with the "absolute value" of functions $f : [a, b] \to \mathbb{R}$, many of which you have seen before. This absolute value or norm is defined as

$$|f|_{\text{max}} = \max_{a \leq x \leq b} |f(x)|, \tag{1.16}$$

if this maximum exists. For two functions[16] we will have that

$$|fg|_{\text{max}} \leq |f|_{\text{max}}\,|g|_{\text{max}} \quad \text{and} \quad |f + g|_{\text{max}} \leq |f|_{\text{max}} + |g|_{\text{max}}, \tag{1.17}$$

where in general $|fg|_{\text{max}} < |f|_{\text{max}}\,|g|_{\text{max}}$.

We will speak about $f_n \to f$ for sequences of such functions, just like we speak of convergent sequences of real numbers x_n. This concept of convergence of sequences of functions will be extremely useful for solving many problems in analysis, including integral and differential equations.

[13] Matrix products are explained in Section 19.1.

[14] To some dismay of Euclides and Pythagoras perhaps.

[15] In general $AB \neq BA$! In fact the estimates only require AB or $A + B$ to be defined.

[16] A function and a gunction, with clearly $fg = gf$.

1.6 Exercises

Exercise 1.11. Look at (1.1). Verify that

$$g(x) = \frac{x}{(1-x^7)^{\frac{1}{7}}} \quad \text{gives} \quad f(x) = x^8.$$

What does the scheme $x_n = f(x_{n-1})$ do in relation to g? Play with the obvious similar examples.

Exercise 1.12. Use long division to find the expansion

$$\frac{1}{7} = \sum_{j=1}^{\infty} \frac{d_j}{10^j}.$$

What's the periodic part in the expansion? Divide the sum by that periodic part to obtain (1.8) with n a power of 10 and check that your answer was right.

Exercise 1.13. Find the complete hexigesimal expansion[17] of

$$\frac{17}{24} + \frac{12}{17} = \frac{289+288}{24 \times 17} = \frac{577}{408}.$$

Hint: use hexigesimal long division to write

$$\frac{12}{17} = \sum_{j=1}^{\infty} \frac{h_j}{60^j},$$

which should come out periodic. Then add the finite hexigesimal expansion of $\frac{17}{24}$.

Exercise 1.14. Find a formula for

$$\sum_{k=1}^{n} k^3.$$

Hint: try $an^4 + bn^3 + cn^2 + dn$, find a, b, c, d from $n = 1, 2, 3, 4$, then use dominos.

[17] You need the multiplicative tables in base 60.

Exercise 1.15. You may enjoy proving that

$$\frac{1}{n} \geq 2^{1-n}$$

for all $n \in \mathbb{N}$. This tells us that

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}} :\ \frac{1}{2^{N-1}} < \varepsilon.$$

Hint: domino principle. When you're done do this next one:

Exercise 1.16. This exercise and (1.18) will be crucial in (3.8). Recall that we accept (1.7) as the obvious inequality below, supplemented with an Archimedean argument that the inequality cannot be strict. Let's examine the inequality more closely and cut it up in pieces, for instance

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \underbrace{\frac{1}{2} + \frac{1}{4}}_{\frac{3}{4}} + \frac{1}{8} + \underbrace{\frac{1}{16} + \frac{1}{32} + \frac{1}{64} + \frac{1}{128}}_{\frac{15}{128} = \frac{15}{16}\frac{1}{8} < \frac{1}{8} \leq \frac{1}{4}} + \cdots \leq 1,$$

to draw additional conclusions such as for example

$$\sum_{k=4}^{7} \frac{1}{2^k} < \frac{1}{2^2}.$$

Generalise and prove that

$$\forall_{m,n,N\in\mathbb{N}} :\quad m \geq n \geq N \implies \sum_{k=n}^{m} \frac{1}{2^k} < \frac{1}{2^{N-1}}.$$

Then take N as in Exercise 1.15 to conclude that

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{m,n\in\mathbb{N}} :\quad m \geq n \geq N \implies \sum_{k=n}^{m} \frac{1}{2^k} < \varepsilon. \tag{1.18}$$

Exercise 1.17. Let $p \in \mathbb{N}$. Complete the expression

$$a^{p+1} - b^{p+1} = (a-b) \sum_{j=0}^{p} \ldots$$

and show that

$$(p+1)b^p < \frac{a^{p+1}-b^{p+1}}{a-b} < (p+1)a^p$$

for $a > b > 0$. Then put $a = k+1$ and $b = k$ and take the sum over $k = 0, 1, \ldots, n-1$ to show that[18]

$$\sum_{k=0}^{n-1} k^p < \frac{n^{p+1}}{p+1} < \sum_{k=0}^{n} k^p$$

for $p, n \in \mathbb{N}$. NB *In Chapter 6 these inequalities lead to*

$$\int_0^1 x^p \, dx = \frac{1}{p+1}.$$

Exercise 1.18. Use (1.10) to show for $n \in \mathbb{N}$ that

$$nx^{n-1} < \frac{1}{1-x} \quad \text{if} \quad 0 < x < 1.$$

Exercise 1.19. Write

$$\frac{x}{1+x}$$

as a power series for x with $|x| < 1$, and as a power series in $\frac{1}{x}$ for x with $|x| > 1$. As in Exercise 1.9: draw graphs to examine how well the partial sums do as approximations.

Exercise 1.20. Referring to Exercise 1.10 take $n = 7$. Use long division to show that

$$\frac{x^7 - a^7}{x-a} = x^6 + ax^5 + a^2x^4 + a^3x^3 + a^4x^2 + a^5x + a^6,$$

and then whatever algebra you like to deduce that

$$x^7 = a^7 + 7a^6(x-a) + (x^5 + 2ax^4 + 3a^2x^3 + 4a^3x^2 + 5a^4x + 6a^5)(x-a)^2.$$

What's the formula for general $n \in \mathbb{N}$?

[18] Frits Beukers showed me this neat trick.

Exercise 1.21. This exercise relates to (1.12). Suppose that $0 < x < 1$. From (1.10) it follows that[19]

$$(N+1)x^N < \frac{1}{1-x}$$

for every $N \in \mathbb{N}$. Combine with Theorem 1.5 to show that

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N} : x^n < \varepsilon.$$

Hint: first show that the statement

$$\exists_{\varepsilon>0}\, \forall_{N\in\mathbb{N}} : x^N \geq \varepsilon$$

is false.

Exercise 1.22. This exercise relates to (1.13). Suppose that $x > 1$. For each $n \in \mathbb{N}$ let $y = \sqrt[n]{x}$ be defined by $y^n = x$. This implies that $\sqrt[n]{x} < \sqrt[m]{x}$ if $n > m$. To prove that

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N} : 0 < \sqrt[n]{x} - 1 < \varepsilon$$

it therefore suffices to prove that

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}} : 0 < \sqrt[N]{x} - 1 < \varepsilon.$$

Prove this latter statement. Hint: assume it is false; derive a contradiction with the hint in Exercise 1.21.

Exercise 1.23. This exercise also relates to (1.13). Suppose that $0 < x < 1$. Prove that

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N} : 0 < 1 - \sqrt[n]{x} < \varepsilon.$$

Exercise 1.24. This exercise introduces a happy couple for later. Let $p > 1$ and $q > 1$ be real numbers. Show that

$$\frac{1}{p} + \frac{1}{q} = 1 \iff (p-1)(q-1) = 1 \iff q = \frac{p}{p-1} \iff p = \frac{q}{q-1}$$

[19] Compare to Exercise 1.18.

2 What Heron tells us about sequences in $\mathbb{R}$

In Exercise 1.2 Heron's scheme (1.2) with $x_0 = 1$ produced the numbers

$$x_1 = \frac{3}{2} > x_2 = \frac{17}{12} > x_3 = \frac{577}{408} > x_4 = \frac{665857}{470832}$$

and so on. This sequence of numbers x_n indexed by $n \in \mathbb{N}$ was designed by Heron to solve the equation

$$x^2 = 2. \tag{2.1}$$

For $x > 0$ we now introduce the notation

$$\tilde{x} = f(x) = \frac{x}{2} + \frac{1}{x}, \tag{2.2}$$

which we think of as an input-output relation defined by the formula[1] $f(x)$. The input is some freely chosen x, and the output is some other $\tilde{x}$, defined by (2.2). With this notation every x_n in Heron's sequence is obtained as an $\tilde{x}$ from a previous $x = x_{n-1}$, starting from the fixed value $x_0 = 1$.

Note that

$$\tilde{x}^2 - 2 = \left(\frac{x}{2} + \frac{1}{x}\right)^2 - 2 = \left(\frac{x}{2} - \frac{1}{x}\right)^2 > 0$$

unless $x^2 = 2$, and that $\tilde{x}$ differs from x by

$$\tilde{x} - x = \frac{x}{2} + \frac{1}{x} - x = \frac{1}{x} - \frac{x}{2} = \frac{1}{2x}(2 - x^2). \tag{2.3}$$

Thus it follows that

$$x_n^2 > 2 \quad \text{and} \quad 0 < x_{n+1} < x_n \quad \text{for all} \quad n \in \mathbb{N}. \tag{2.4}$$

In particular Heron's sequence (of rational numbers x_n) has

$$\frac{3}{2} = x_1 > x_2 > x_3 > \cdots > \frac{4}{3},$$

in which $\frac{4}{3}$ is a rather arbitrarily chosen rational lower bound for the decreasing rational numbers in the sequence. Our goal is to show that this lower bound may be replaced by the larger lower bound $\sqrt{2}$, and that no larger lower bound is possible.

[1] Or the function f if you like.

Exercise 2.1. Prove that $\frac{4}{3}$ is indeed a lower bound for the sequence, but that there are larger rational lower bounds. Hint: maybe verify first that $\sqrt{2} > \frac{4}{3}$. Or maybe not. Simpler is to use that the squares are all larger than 2 and the reciprocals are all bounded from below by $\frac{2}{3}$. Write what x_{n+1} is and factor out the reciprocal of x_n.

We shall want to be able to conclude that

$$x_n \to \sqrt{2} \tag{2.5}$$

as n gets larger and larger. We therefore have an urgent need for (a meaning of) the statement

$$x_n \to \bar{x},$$

for some $\bar{x}$ we usually don't know yet a priori[2]. The reasoning should then be that

$$x_n = f(x_{n-1}) \to \bar{x} = f(\bar{x}), \tag{2.6}$$

and that therefore $\bar{x}$ is a solution of

$$x = f(x) = \frac{x}{2} + \frac{1}{x},$$

a purposefully perverted equivalent version of the equation $x^2 = 2$ we were hoping to solve. In particular this will involve the implication

$$x_n \to \bar{x} \implies f(x_n) \to f(\bar{x}), \tag{2.7}$$

which will be called *continuity* of f in $\bar{x}$.

Exercise 2.2. Verify that for $x \neq 0$ the equation

$$x = \frac{x}{2} + \frac{1}{x}$$

is equivalent to the equation $x^2 = 2$.

2.1 Bounded monotone sequences have limits!

We saw that Heron's sequence is strictly decreasing and bounded from below. Sequences of numbers[3] x_n with either

$$x_1 \leq x_2 \leq x_3 \leq \cdots \quad \text{or} \quad x_1 \geq x_2 \geq x_3 \geq \cdots,$$

[2] Although in this example we do have a hunch.

[3] For the moment rational numbers.

are called *monotone* sequences. There are two types of monotone sequences: nondecreasing and nonincreasing.

If such a sequence is bounded we think of it as approximating a number, be it rational or irrational. For instance, the sequence

$$\frac{1}{2}, \frac{1}{2}+\frac{1}{4} = \frac{3}{4}, \frac{1}{2}+\frac{1}{4}+\frac{1}{8} = \frac{7}{8}, \frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16} = \frac{15}{16}, \frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\frac{1}{32} = \frac{31}{32}, \cdots$$

is bound to approximate the rational number 1. Most nondecreasing bounded sequences however will define a number which is not rational, as you can infer from Theorem 1.4.

Exercise 2.3. Show that there exists a sequence

$$x_1 = 1 < x_2 = 1.4 < x_3 = 1.41 < x_4 = 1.414 < x_5 = 1.4142 < x_6 = 1.41421 < \cdots,$$

such that for every $n \in \mathbb{N}$ the number x_n is the largest number[4] with n digits that has the property that $x_n^2 < 2$.

The idea behind the construction of $\mathbb{R}$ is to add to $\mathbb{Q}$ all the *lowest upper bounds* of bounded nondecreasing sequences[5] which do not approximate a rational number. This is consistent with the decimal approach in the proof of Theorem 1.4 and in Exercise 2.3. The resulting[6] set $\mathbb{R}$ has the property that it contains $\mathbb{Q}$, and is just like $\mathbb{Q}$ as far as the algebraic operations addition and multiplication, and the ordering of the numbers are concerned.

Unlike $\mathbb{Q}$ the set $\mathbb{R}$ has the important property that every nondecreasing bounded sequence x_n in $\mathbb{R}$ has a smallest upper bound (*supremum*)

$$S = \sup_{n\in\mathbb{N}} x_n \in \mathbb{R}.$$

This number S turns out to be the unique limit of the sequence x_n (in terms of a definition that will follow shortly). Likewise for every nonincreasing bounded sequence x_n in $\mathbb{R}$, its largest lower bound (*infimum*)

$$L = \inf_{n\in\mathbb{N}} x_n \in \mathbb{R}$$

must be the limit of that sequence. Let's make these notions more precise.

[4]Counting 5 digits in 1.4142.

[5]And then also the real non-rational *largest lower bounds* of nonincreasing sequences.

[6]Details of this construction are omitted, we assume the existence of a such a set $\mathbb{R}$.

Definition 2.4. *Let x_n be a sequence of numbers in $\mathbb{R}$ indexed by $n \in \mathbb{N}$. Then the sequence is called*

- nondecreasing *if*
$$\forall_{n\in\mathbb{N}} : x_n \leq x_{n+1},$$
i.e. $x_n \leq x_{n+1}$ for every natural number n;

- strictly increasing *if*
$$\forall_{n\in\mathbb{N}} : x_n < x_{n+1};$$

- nonincreasing *if*
$$\forall_{n\in\mathbb{N}} : x_n \geq x_{n+1};$$

- strictly decreasing *if*
$$\forall_{n\in\mathbb{N}} : x_n > x_{n+1};$$

- bounded from above *if*
$$\exists_{M\in\mathbb{R}} \forall_{n\in\mathbb{N}} : x_n \leq M,$$
in which case the number M is called an upper bound*; a number $S \in \mathbb{R}$ is called a* lowest upper bound (supremum) *for the sequence x_n if it is an upper bound and if there are no upper bounds M with $M < S$, notation*
$$S = \sup_{n\in\mathbb{N}} x_n \in \mathbb{R};$$

- bounded from below *if*
$$\exists_{m\in\mathbb{R}} \forall_{n\in\mathbb{N}} : x_n \geq m,$$
in which case the number m is called a lower bound*; a number $L \in \mathbb{R}$ is called a* largest lower bound (infimum) *if it is a lower bound and if there are no lower bounds m with $m > L$, notation*
$$L = \inf_{n\in\mathbb{N}} x_n \in \mathbb{R};$$

- bounded *if it is bounded from below* and *bounded from above.*

For example, Heron's sequence is a strictly decreasing bounded sequence, bounded from above by $M = x_1 = \frac{3}{2}$, and bounded from below by $m = \frac{4}{3}$. In particular the following theorem applies to it.

Theorem 2.5. *Every nonincreasing bounded sequence in* $\mathbb{R}$ *has a unique infimum in* $\mathbb{R}$. *Equivalently: every nondecreasing bounded sequence in* $\mathbb{R}$ *has a unique supremum in* $\mathbb{R}$.

We will not prove this theorem. It follows from every proper construction of $\mathbb{R}$, for instance via decimal expansions as used in the proof of Theorem 1.4 and Exercise 2.3. Applied to Heron's sequence Theorem 2.5 gives us L_{Heron}, the largest lower bound of the Heron sequence. Our goal is to prove that

$$L_{\text{Heron}} = \sqrt{2},$$

and we need some definitions to get started with this proof.

2.2 The limit definition: epsilons

The defining property of the infimum L of a sequence x_n is that $x_n \geq L$ for all $n \in \mathbb{N}$, but that there is no larger number for which this is also the case. Thus, if $\varepsilon > 0$, the number $L + \varepsilon$ is not a lower bound, meaning there must exist $N \in \mathbb{N}$ such that $x_N < L + \varepsilon$. Since the sequence is nonincreasing it then also follows that

$$L \leq x_n \leq x_N < L + \varepsilon \quad \text{for all} \quad n \geq N.$$

We conclude that

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N} \,:\, |x_n - L| < \varepsilon, \tag{2.8}$$

a statement to be pronounced as: for all (real numbers) $\varepsilon > 0$ there exists a natural number N such that for all natural numbers n with $n \geq N$ it holds that

$$\underbrace{\text{the } \textit{distance} \text{ between } x_n \text{ and } L}_{d(x_n,L)=|x_n-L|}$$

is smaller than ε. For the moment $d(x, y)$ is only a short hand notation for the distance between *two real numbers* x *and* y, and therefore defined by

$$d(x, y) = |x - y|. \tag{2.9}$$

This formulation is based on *algebra*[7] with real[8] numbers.

By the way, the statement in (2.8) makes sense for every real L and every real sequence[9], not just for monotone sequences.

[7] For now: a human activity with the operations $+, -, \times, /$ and certain algebraic rules.
[8] So it's not really algebra....
[9] It does not matter that n runs from 1 upwards, any other starting integer is fine.

Definition 2.6. *A sequence of real numbers x_n indexed by $n \in \mathbb{N}$ is called* convergent *if there exists an $L \in \mathbb{R}$ such that*

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N}\, :\, |x_n - L| < \varepsilon.$$

We then write

$$x_n \to L \quad (as \quad n \to \infty),$$

or equivalently

$$\lim_{n\to\infty} x_n = L.$$

The number L is called the limit *of the sequence. We say that x_n converges to L (as n goes to infinity). We often don't explicitly write "as $n \to \infty$".*

Take note of the convention that Greek letters always stand for real numbers and the lower case letters in the middle of the alphabet are integers, unless otherwise specified.

Remark 2.7. *Convergence of the sequence x_n means that*

$$\exists_{\bar{x}\in\mathbb{R}}\, \forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N}\, :\, \underbrace{|x_n - \bar{x}|}_{d(x_n,\bar{x})} < \varepsilon. \tag{2.10}$$

The negation of (2.10) reads

$$\forall_{\bar{x}\in\mathbb{R}}\, \exists_{\varepsilon>0}\, \forall_{N\in\mathbb{N}}\, \exists_{n\geq N}\, :\, \underbrace{|x_n - \bar{x}|}_{d(x_n,\bar{x})} \geq \varepsilon. \tag{2.11}$$

The negation is obtained from (2.10) by negating the statement following the semi-colon, and changing every $\exists$ to $\forall$ and vice versa. Sequences that are not convergent, i.e. for which (2.11) holds, are called divergent.

Remark 2.8. *Is Definition 2.6 of any use? Heron's method requires to know that*

$$\lim_{n\to\infty} x_n = L \implies \lim_{n\to\infty} x_n^2 = L^2. \tag{2.12}$$

Proof of (2.12). We know that the left hand side of the implication in (2.12) says that $|x_n - L|$ is small for n large. To prove the right hand side we have to show that $|x_n^2 - L^2|$ is small for n large. Note moreover that

$$\underbrace{|x_n^2 - L^2|}_{\text{small for } n \text{ large?}} = \underbrace{|x_n + L|}_{\text{not too large?}} \cdot \underbrace{|x_n - L|}_{\text{small for } n \text{ large!}}, \tag{2.13}$$

in which the multiplicative dot is included for the purpose of clarification only. We first make the smallness of the second factor in (2.13) precise using

the definition of $x_n \to L$. So let $\varepsilon > 0$. Then according to the definition of $x_n \to L$ there exists $N \in \mathbb{N}$ such that

$$\forall_{n \geq N} : |x_n - L| < \varepsilon. \tag{2.14}$$

With the factor $|x_n - L|$ small there's neither need nor reason for the first factor in the right hand side of (2.13) to be small. We do want get rid of its n-dependence though, to make sure that the product of the two factors is also small. To this end we apply the definition of $x_n \to L$ with just one[10] convenient choice of $\varepsilon > 0$, say $\varepsilon = 1$, and we obtain[11]

$$\exists_{N_1 \in \mathbb{N}} \forall_{n \geq N_1} : |x_n - L| < 1.$$

The triangle inequality[12] then gives

$$|x_n + L| = |x_n - L + 2L| \leq |x_n - L| + |2L| < 1 + 2|L|, \tag{2.15}$$

for all $n \geq N_1$. Note very carefully how we bring $|x_n - L|$ into play in the first step of (2.15) by[13] subtracting and adding L before we use the triangle inequality.

Combining (2.14) and (2.15) it follows from (2.13) that

$$|x_n^2 - L^2| = |x_n + L||x_n - L| \leq (1 + 2|L|)\,|x_n - L| < (1 + 2|L|)\varepsilon \tag{2.16}$$

for all $n \geq \max(N, N_1)$. Writing $M = 1 + 2|L|$ we have thus established that

$$\forall_{\varepsilon > 0} \exists_{N \in \mathbb{N}} \forall_{n \geq N} : |x_n^2 - L^2| < M\varepsilon. \tag{2.17}$$

If it happens to be the case that $M \leq 1$ then the proof is complete with (2.17), but here this only occurs if $L = 0$. For $L \neq 0$ we have $M > 1$.

Now recall that to estimate the second factor in (2.13) we used (2.14) with the $\varepsilon > 0$ that was given at the start of the proof. But we can also use (2.14) with $\varepsilon > 0$ replaced by

$$\tilde{\varepsilon} = \frac{\varepsilon}{M}, \tag{2.18}$$

which is also positive[14]. This will give a different value of N, say $\tilde{N}$, such that

$$\forall_{n \geq N} : |x_n - L| < \frac{\varepsilon}{M}$$

[10] See also your proof Proposition 2.9 below.

[11] With a subscript 1 on N to distinguish from the N for the arbitrary choice of $\varepsilon > 0$.

[12] This triangle inequality will be reviewed in Exercise 2.13 below.

[13] The subtract and add the same term trick.

[14] Let's call this the M-trick.

holds. It then follows that

$$|x_n^2 - L^2| = |x_n + L||x_n - L| \leq M|x_n - L| < M\,\frac{\varepsilon}{M} = \varepsilon$$

for all n with[15]

$$n \geq \max(N_1, \tilde{N}).$$

Since $\varepsilon > 0$ was arbitrary this then completes the proof that $x_n^2 \to L^2$. Proposition 2.9 records one of the two[16] important items in this proof. □

Proposition 2.9. *Any convergent sequence is bounded, i.e. if x_n is convergent then there exists $M > 0$ such that $|x_n| \leq M$ for all n.*

Exercise 2.10. Prove Proposition 2.9. Hint: apply the definition of convergence with just one[17] convenient choice of ε and use the triangle inequality. Don't forget the n with $n < N$.

Proposition 2.11. *A convergent sequence can only have one limit.*

Exercise 2.12. Prove Proposition 2.11. Hint: if not then there are two limits, say L_1 and L_2, and you can apply the definition of convergence twice, with L_1 and with L_2; the subtract and add trick[18], the triangle inequality, and the specific choice[19]

$$\varepsilon = |L_1 - L_2| > 0$$

allow you to derive a contradiction.

Exercise 2.13. Note that (2.8) is the first occurrence of an absolute[20] value in a definition. We recall that $|x| = x$ for $x \geq 0$ and $|x| = -x$ for $x < 0$. In the proof of (2.12) we used the *triangle inequality*, which reads

$$|a + b| \leq |a| + |b|.$$

Prove that this inequality, as well as the *reverse triangle inequality*[21]

$$||a| - |b|| \leq |a - b|$$

[15] With a tilde on N to distinguish from the earlier (also arbitrary) choice of ε.
[16] The other one being the trick with M.
[17] So you don't use the full strength of the definition!
[18] as in the proof of (2.12).
[19] This requires the full strength of the definition!
[20] Recall $|x|$ is also called the norm of x.
[21] A nice statement about the map $x \to |x|$ from $\mathbb{R}$ to $[0, \infty)$.

hold for all $a, b \in \mathbb{R}$. Combined these statements are equivalent to

$$||a| - |b|| \leq |a + b| \leq |a| + |b|.$$

Exercise 2.14. Substitute $a = x - z$ and $b = z - y$ to obtain

$$\underbrace{|x - y|}_{d(x,y)} \leq \underbrace{|x - z|}_{d(x,z)} + \underbrace{|z - y|}_{d(z,y)},$$

in which we indicate what the triangle inequality looks like if we implement the notation introduced in the discussion of (2.9).

Theorem 2.15. *If x_n is a convergent sequence with limit L, then $|x_n|$ is also a convergent sequence, with limit $|L|$.*

Exercise 2.16. Prove Theorem 2.15. Hint: use the reverse triangle inequality.

Exercise 2.17. Let $N \in \mathbb{N}$. Prove that

$$|x_1 + \cdots + x_N| \leq |x_1| + \cdots + |x_N|$$

for all $x_1, \ldots, x_N \in \mathbb{R}$.

2.3 What about Heron's limit?

We note from (2.3) that Heron's sequence has

$$x_{n+1} - x_n = \frac{1}{x_n} - \frac{x_n}{2},$$

whence

$$2x_n(x_{n+1} - x_n) = 2 - x_n^2. \tag{2.19}$$

Exercise 2.18. Recall that Heron's sequence is convergent. Use this to prove[22] that $x_{n+1} - x_n \to 0$.

[22] And give an example of a divergent sequence for which $x_{n+1} - x_n \to 0$.

Exercise 2.19. Prove that it holds for Heron's sequence that $x_n^2 \to 2$. Hint: combine (2.19) and Exercise 2.18.

Exercise 2.20. Recall that

$$L_{\text{Heron}} = \inf_{n \in \mathbb{N}} x_n = \lim_{n \to \infty} x_n.$$

Prove that $L_{\text{Heron}}^2 = 2$. Hint: combine Exercise 2.19 with (2.12).

Exercise 2.21. Prove there is only one positive real number L such that $L^2 = 2$. No hint.

Exercise 2.22. By construction L_{Heron} is a positive number because $L_{\text{Heron}} \geq \frac{4}{3}$. Prove that L_{Heron} is the only positive real number which squares to 2. This then justifies the conclusion that $L_{\text{Heron}} = \sqrt{2}$.

Exercise 2.23. Exercise 2.3 produced a bounded nondecreasing[23] sequence which therefore has a supremum S. Prove that $S^2 = 2$. Thus $S = L_{\text{Heron}} = \sqrt{2}$.

2.4 Suprema and infima of sets

Every sequence $x_n \in \mathbb{R}$ indexed by $n \in \mathbb{N}$ defines a nonempty subset

$$\{x_n : n \in \mathbb{N}\} \subset \mathbb{R}.$$

Likewise every function $f : [a, b] \to \mathbb{R}$ defines a set

$$R_f = \{f(x) : a \leq x \leq b\},$$

called the range of f. This section will be a bit of an abstract project on the properties of subsets of $\mathbb{R}$, and is necessary for Theorem 4.4 in Chapter 4 and for the theory of integration in Chapter 6.

[23] Is that sequence strictly increasing?

Definition 2.24. *A nonempty subset A of $\mathbb{R}$ is called* bounded from above *if there exists $M_0 \in \mathbb{R}$ such that $a \leq M_0$ for all $a \in A$. Such an M_0 is called an* upper bound *for A. Likewise, A is called* bounded from below *if there exists $m_0 \in \mathbb{R}$ such that $a \geq m_0$ for all $a \in A$. Such an m_0 is called a* lower bound *for A.*

We want to show that every nonempty subset A of $\mathbb{R}$ which is bounded from above has a lowest upper bound. Suppose that A is such a set, and let M_0 be an upper bound for A. Take an $a_0 \in A$ and consider

$$m_0 = \frac{a_0 + M_0}{2}.$$

If m_0 is also an upper bound for A define $a_1 = a_0 \in A$ and $M_1 = m_0$. If m_0 is not an upper bound then there exists $a_1 > m_0$ with $a_1 \in A$ and therefore $a_0 < m_0 < a_1 \leq M_0$. In this case define $M_1 = M_0$. In both cases if follows that

$$a_1 \geq a_0, \quad M_1 \leq M_0, \quad 0 \leq M_1 - a_1 \leq \frac{M_0 - a_0}{2}.$$

Repeat the argument. This gives $a_2 \in A$ and an upper bound M_2, a_3 and M_3, and so on. We thus obtain two bounded monotone sequences. The nondecreasing sequence a_n has a supremum $\bar{a}$ and the nonincreasing sequence M_n has an infimum that we will call S.

Exercise 2.25. Prove that $S = \bar{a}$, and that S is the lowest upper bound of A.

It may or may not happen that $S \in A$, but in both cases the conclusion is the same:

Theorem 2.26. *Let A be a nonempty subset of $\mathbb{R}$ which is bounded from above. Then A has a lowest upper bound S in $\mathbb{R}$, notation*

$$S = \sup A.$$

Likewise, if A is bounded from below then A has a largest lower bound I in $\mathbb{R}$, denoted[24] *by*

$$I = \inf A.$$

Remark 2.27. *Thus S en I are real numbers, if they exist. If A is not bounded from above we say that $\sup A = \infty$. If A is not bounded from below we say that $\inf A = -\infty$. Neither ∞ nor $-\infty$ exists, for that matter.*

[24]Before we used L, for reasons of presentation.

2.5 Examples of convergent sequences

In Section 2.2 we tailored the definition of convergence so that the following theorem has already been proved.

Theorem 2.28. *Every bounded monotone sequence in* $\mathbb{R}$ *is convergent. If the sequence is nonincreasing then its limit is the infimum of the sequence, if the sequence is nondecreasing then its limit is the supremum of the sequence.*

This theorem in particular implies that the limit of the sequence

$$\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \dots$$

exists, but it does not yet tell us that this limit is 0.

Theorem 2.29. *(The Archimedean Principle in limit form)*

$$\lim_{n\to\infty} \frac{1}{n} = 0.$$

Exercise 2.30. Use Definition 2.6 to explain why Theorem 1.5 in Section 1.5 is equivalent to Theorem 2.29.

Proof. By Theorem 2.28 the limit exists as the largest lower bound of the sequence $\frac{1}{n}$. It is also clear that 0 is a lower bound. Could there be a larger lower bound? If so this would imply that there is a lower bound[25] $m > 0$ for the sequence, i.e.

$$\frac{1}{n} \geq m > 0 \quad \text{for all} \quad n \in \mathbb{N} \quad \text{and thus} \quad n \leq \frac{1}{m} = M \in \mathbb{R}$$

for all $n \in \mathbb{N}$. This looks absurd: how could the sequence

$$1, 2, 3, 4, 5, 6, 7, 8, 9, \dots$$

be bounded?

Actually it cannot, because then the sequence $x_n = n$ would have a supremum $S \in \mathbb{R}$. With this lowest upper bound S at our disposal[26], we then observe $S - \frac{1}{2}$ is not an upper bound. This means that there exists $n \in \mathbb{N}$ with $n > S - \frac{1}{2}$. Hence[27] the number $n+1 \in \mathbb{N}$ satisfies $n+1 > S + \frac{1}{2} > S$ and

[25] Here m is a real number.

[26] To dispose of in fact.

[27] We use that $n \in \mathbb{N} \implies n + 1 \in \mathbb{N}$.

disqualifies S as the supremum of the sequence $x_n = n$, since it is not even an upper bound. This completes the proof of Theorem 2.29. In particular we have

$$\inf_{n\in\mathbb{N}} \frac{1}{n} = 0, \tag{2.20}$$

and Theorem 1.5 is also proved. □

Exercise 2.31. Why does it now also follow that

$$\frac{1}{2^n} \to 0$$

as $n \to \infty$? Adapt the argument in the proof of (2.20) if that adapted proof wasn't already part of your answer.

It is highly unlikely that you will be impressed by Theorem 2.29 and the result in Exercise 2.31, but we had to make sure that what obviously must be true can indeed be proved within our framework for mathematical analysis. There are many more such obvious statements.

Example 2.32. *The sequence x_n defined by*

$$x_n = \frac{n-1}{n+1}$$

is convergent. You don't need to be knowledgable in mathematics to guess its limit: when n is large the numerator and denominator contain the same large term, so the limit is bound to be 1. To prove the obvious let $\varepsilon > 0$ be arbitrary. We need to establish that

$$\left|\frac{n-1}{n+1} - 1\right| < \varepsilon$$

for n sufficiently large, i.e. larger than some N which will depend[28] *on ε. Observe that*

$$\left|\frac{n-1}{n+1} - 1\right| = \left|\frac{-2}{n+1}\right| = \frac{2}{n+1},$$

and that

$$\frac{2}{n+1} < \varepsilon \iff n+1 > \frac{2}{\varepsilon} \iff n > \frac{2}{\varepsilon} - 1 = \frac{2-\varepsilon}{\varepsilon}.$$

[28] As before we prefer not to use a subscript on N when $\varepsilon > 0$ is not specified.

Thus the desired inequality is equivalent to

$$n > \frac{2-\varepsilon}{\varepsilon},$$

which certainly holds all $n \in \mathbb{N}$ *if* $\varepsilon \geq 2$.

For $\varepsilon < 2$ *we invoke the Archimedian Principle again. It is slightly more convenient to use the restated form in Remark 1.6. This gives the existence of an* $N \in \mathbb{N}$ *with*

$$N > \frac{2-\varepsilon}{\varepsilon},$$

whence[29] *also*

$$n > \frac{2-\varepsilon}{\varepsilon} \quad \text{for all} \quad n \geq N.$$

In both cases we have shown that there exists N *such that*

$$\left|\frac{n-1}{n+1} - 1\right| < \varepsilon \quad \text{for all} \quad n \geq N.$$

This proves the claim that

$$\lim_{n\to\infty} \frac{n-1}{n+1} = 1.$$

Do take note of the careful reasoning with inequalities. More of the same in Exercise 2.46. □

2.6 Basic theorems about convergent sequences

Proposition 2.33. *Assume that*

$$\lim_{n\to\infty} x_n = L.$$

If $x_n \geq a$ *for some number* $a \in \mathbb{R}$ *and all* $n \in \mathbb{N}$ *then also* $L \geq a$. *The same statement holds with* $\geq$ *replaced by* $\leq$.

Exercise 2.34. Prove Proposition 2.33. Hint: assume that $L < a$ and apply the Definition 2.6 with $\varepsilon = a - L$ to derive a contradiction. Can the conclusion of Proposition 2.33 be strenghtened if $x_n > a$ for all n?

[29] See the point made by Exercise 1.7.

Exercise 2.35. Here's a variant of the *subtract, add, then trangle inequality trick* that we will use for product sequences next. Let $a, b, c, d \in \mathbb{R}$. Prove that

$$|ab - cd| \leq |a - c|\,|b| + |c|\,|b - d|.$$

Theorem 2.36. *If x_n and y_n are convergent sequences, with limits $\bar{x}$ and $\bar{y}$, then so are the sequences $x_n + y_n$, $x_n - y_n$ and $x_n y_n$, with limits $\bar{x} + \bar{y}, \bar{x} - \bar{y}$, and $\bar{x}\bar{y}$ respectively.*

Proof of the sum statement. The limit of the sequence $x_n + y_n$ should be $\bar{x} + \bar{y}$, so we have to show that the distance between $x_n + y_n$ and $\bar{x} + \bar{y}$ is small for n large. We will try to estimate this distance in such a way that the distances $|x_n - \bar{x}|$ and $|y_n - \bar{y}|$ come into play. There is no general approach here, you have to figure out how to do it. If we use the triangle inequality with an intermediate step we obtain

$$\underbrace{|(x_n + y_n) - (\bar{x} + \bar{y})|}_{d(x_n+y_n,\bar{x}+\bar{y})} = \underbrace{|(x_n - \bar{x}) + (y_n - \bar{y})|}_{\textit{reshuffled}} \leq \underbrace{|x_n - \bar{x}|}_{d(x_n,\bar{x})} + \underbrace{|y_n - \bar{y}|}_{d(y_n,\bar{y})}. \quad (2.21)$$

The equality in (2.21) is a *reshuffle trick*. It uses the algebraic properties of addition and subtraction in $\mathbb{R}$.

With (2.21) we are in position to start up the proof with a default sentence.

Let $\varepsilon > 0$ be arbitrary.

Since $x_n \to \bar{x}$ we have

$$\exists_{N_x \in \mathbb{N}}\, \forall_{n \geq N_x} \,:\, \underbrace{|x_n - \bar{x}|}_{d(x_n,\bar{x})} < \varepsilon,$$

in which we use a subscript x on N to indicate that this is the statement for the sequence x_n to converge to $\bar{x}$.

We then do *copy-paste* followed by *search x replace by y*.

Indeed, since $y_n \to \bar{y}$ we have

$$\exists_{N_y \in \mathbb{N}}\, \forall_{n \geq N_y} \,:\, \underbrace{|y_n - \bar{y}|}_{d(y_n,\bar{y})} < \varepsilon,$$

in which we use a subscript y on N to indicate that this is the statement for the sequence y_n to converge to $\bar{y}$.

Next we set $N = \max(N_x, N_y)$ to let the ε-engine run.

Our initial estimate (2.21) and the two ε-statements establish that

$$\forall_{n\geq N} \ : \ \underbrace{|(x_n + y_n) - (x + y)|}_{d(x_n+y_n,\bar{x}+\bar{y})} < \varepsilon + \varepsilon = 2\varepsilon. \tag{2.22}$$

Now we are not completely happy with 2ε. Looking back at the proof of (2.12) we conclude that we must invoke a 2-trick rather than an M-trick, see (2.18). We replace the default choice $\varepsilon > 0$ above by

$$\tilde{\varepsilon} = \frac{\varepsilon}{2}, \tag{2.23}$$

which is also positive. This then gives two different values N_x and N_y, say $\tilde{N}_x$ and $\tilde{N}_x$, such that

$$\forall_{n\geq \tilde{N}_x} \ : \ \underbrace{|x_n - \bar{x}|}_{d(x_n,\bar{x})} < \tilde{\varepsilon},$$

and

$$\forall_{n\geq \tilde{N}_y} \ : \ \underbrace{|y_n - \bar{y}|}_{d(y_n,\bar{y})} < \tilde{\varepsilon}.$$

With

$$\tilde{N} = \max(\tilde{N}_x, \tilde{N}_y)$$

our initial estimate (2.21) and the two new $\tilde{\varepsilon}$-statements establish that

$$\forall_{n\geq \tilde{N}} \ : \ \underbrace{|(x_n + y_n) - (x + y)|}_{d(x_n+y_n,\bar{x}+\bar{y})} < \tilde{\varepsilon} + \tilde{\varepsilon} = \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary we have verified that

$$x_n + y_n \to \bar{x} + \bar{y} \quad \text{as} \quad n \to \infty.$$

□

Remark 2.37. *In hindsight we might just as well start with (2.21), jump to (2.23) and continue from there to finish the proof. Before we allow ourselves to think about such proof shortenings we do the proof for the product sequence. And then we shall reconsider our lack of happiness with (2.22), and maybe forget about (2.23) and what followed.*

Proof of the product statement. The limit of the sequence $x_n y_n$ should be $\bar{x}\bar{y}$, so we have to show that the distance between $x_n y_n$ and $\bar{x}\bar{y}$ is small for n large. Therefore we estimate this distance first, trying to get the distances $|x_n - \bar{x}|$ and $|y_n - \bar{y}|$ into play. Again there is no general approach. If we use the *subtract, add, then trangle inequality trick* from Exercise 2.35 and write

$$x_n y_n - \bar{x}\bar{y} = x_n y_n - \bar{x} y_n + \bar{x} y_n - \bar{x}\bar{y} = (x_n - \bar{x}) y_n + \bar{x}(y_n - \bar{y}),$$

it follows that

$$\underbrace{|x_n y_n - \bar{x}\bar{y}|}_{d(x_n y_n, \bar{x}\bar{y})} \leq \underbrace{|x_n - \bar{x}|}_{d(x_n,\bar{x})} |y_n| + |\bar{x}| \underbrace{|y_n - \bar{y}|}_{d(y_n,\bar{y})}. \tag{2.24}$$

Next we do copy-paste of what's between (2.21) and (2.22) but undo paste before we continue. Uuuuh, maybe not. Here's a partial paste.

Let $\varepsilon > 0$ be arbitrary.

Since $x_n \to \bar{x}$ *and* $y_n \to \bar{y}$ we have

$$\exists_{N\in\mathbb{N}} \forall_{n\geq N} : \underbrace{|x_n - \bar{x}|}_{d(x_n,\bar{x})} < \varepsilon \quad \textit{and} \quad \underbrace{|y_n - \bar{y}|}_{d(y_n,\bar{y})} < \varepsilon,$$

in which N is the maximum of the two subscripted N's we had from the definition of $x_n \to \bar{x}$ *and* the definition of $y_n \to \bar{y}$.

Now we use (2.24). We arrive, for *the same* $N \in \mathbb{N}$, at

$$\forall_{n\geq N} : |x_n y_n - \bar{x}\bar{y}| \leq \underbrace{|x_n - \bar{x}|}_{<\varepsilon} |y_n| + |\bar{x}| \underbrace{|y_n - \bar{y}|}_{<\varepsilon}. \tag{2.25}$$

If we are not happy with the prefactor $|\bar{x}|$, we are even more unhappy with the n-dependence in the postfactor $|y_n|$. Fortunately we have Proposition 2.9 at our disposal. Thus there exists $M > 0$ such that $|y_n| \leq M$ for all $n \in \mathbb{N}$ and it follows from (2.25) that

$$\forall_{n\geq N} : |x_n y_n - \bar{x}\bar{y}| < (M + |\bar{x}|)\varepsilon. \tag{2.26}$$

Now we are happy again, because with (2.26) we are at the same point as in the proof of (2.12) with (2.16). The M-trick with M replaced by $M + |\bar{x}|$ concludes the proof that $x_n y_n \to \bar{x}\bar{y}$ and well deserves a remark next. □

Remark 2.38. *A sequence x_n converges to $\bar{x}$ if and only if*

$$\exists_{M>0} \forall_{\varepsilon>0} \exists_{N\in\mathbb{N}} \forall_{n\geq N} : \underbrace{|x_n - \bar{x}|}_{d(x_n,\bar{x})} < M\varepsilon,$$

so from now on we will be happily content with $< M\varepsilon$ in the proofs of $\forall_{\varepsilon>0}$-statements ending with $< \varepsilon$, or $\leq \varepsilon$ for that matter.

Exercise 2.39. Prove the statement in Remark 2.38 as well as the statement in Theorem 2.36 for $x_n - y_n$.

Theorem 2.36 does not deal with quotients. Suppose $x_n \neq 0$ is a convergent sequence with limit $\bar{x} \neq 0$. We would like to prove that

$$\frac{1}{x_n} \to \frac{1}{\bar{x}} \quad \text{as} \quad n \to \infty.$$

We observe that

$$|x_n - \bar{x}| < \varepsilon \iff x_n \in (\bar{x} - \varepsilon, \bar{x} + \varepsilon) \tag{2.27}$$

so applied to $\varepsilon = \frac{1}{2}|\bar{x}|$ we have

$$x_n > \bar{x} - \varepsilon = \frac{1}{2}\bar{x} > 0 \quad \text{if} \quad \bar{x} > 0 \quad \text{and} \quad x_n < \bar{x} + \varepsilon = \frac{1}{2}\bar{x} < 0 \quad \text{if} \quad \bar{x} < 0,$$

for $n \in \mathbb{N}$ with $n \geq N$ as in (2.8). In both cases it follows that

$$|x_n| > \frac{1}{2}|\bar{x}| \quad \text{whence} \quad \left|\frac{1}{x_n}\right| < \frac{2}{|\bar{x}|} \tag{2.28}$$

and therefore also

$$\left|\frac{1}{x_n} - \frac{1}{\bar{x}}\right| = \frac{|x_n - \bar{x}|}{|\bar{x}|\,|\bar{x}_n|} \leq \frac{2}{|\bar{x}|^2}|x_n - \bar{x}|.$$

This basically proves the following theorem[30].

Theorem 2.40. *Let x_n be a sequence with $x_n \neq 0$ for all n. If x_n is convergent with limit $\bar{x} \neq 0$ then the sequence $\frac{1}{x_n}$ is convergent with limit $\frac{1}{\bar{x}}$.*

Exercise 2.41. Write out a complete proof of Theorem 2.40.

2.7 Exercises

Exercise 2.42. Let $a, b \in \mathbb{R}$ with $a < b$. Use the Archimedean principle to show that there exists $q \in \mathbb{Q}$ with $a < q < b$. Hint: $b - a > 0$. This is called the *density of $\mathbb{Q}$ in $\mathbb{R}$*.

[30] You may like to state and prove a theorem which only requires the limit to be nonzero.

Exercise 2.43. Let $a, b \in \mathbb{R}$ with $a < b$. Show that there exists $c \in \mathbb{R}$ with $a < c < b$ but $c \notin \mathbb{Q}$. Hint: consider $a - \sqrt{2}$ and $b - \sqrt{2}$ and use the result in Exercise 2.42.

Exercise 2.44. Define sequences s_n and S_n by

$$s_n = \sum_{k=1}^{n} \frac{1}{k(k+1)} \quad \text{and} \quad S_n = \sum_{k=1}^{n} \frac{1}{k^2}.$$

Use partial fractions to compute a formula for s_n and take the limit $n \to \infty$. Then prove that

$$\lim_{n\to\infty} S_n$$

exists. Hint: the conclusion would follow from $S_n \leq s_n$, but that's not the case. But if you look at s_{n+1} instead....

Exercise 2.45. Use monotonicity arguments to examine the convergence of the sequence x_n defined by $x_n = f(x_{n-1})$ if $x_0 = 1$ and f is given by

$$f(x) = \frac{1}{4-x}, \quad f(x) = \sqrt{2+x}, \quad f(x) = \sqrt{2x}.$$

Exercise 2.46. For each of the following sequences decide on convergence and prove your conclusion directly from Definition 2.6.

$$(-1)^n,\ \frac{1+n}{2+n},\ \frac{n^2}{n^2-\frac{1}{2}},\ \frac{\sqrt{1+n^2}}{n},\ \frac{\sqrt{n+1}-1}{\sqrt{n}},\ n\left(\sqrt{1+\frac{1}{n}}-1\right),\ \frac{1+n^2}{n}.$$

Exercise 2.47. Suppose that the sequence x_n is convergent with limit L. Referring to the proof of (2.12): give a proof in the same spirit that $x_n^3 \to L^3$ if $n \to \infty$.

Exercise 2.48. Let $k \in \mathbb{Z}$ and x_n be a sequence indexed by

$$n \in \mathbb{N}_k = \{n \in \mathbb{Z} : n \geq k\}.$$

Give the obvious definition of x_n being convergent.

Exercise 2.49. Give a definition of $x_n \to \infty$ which is equivalent to $x_n > 0$ for sufficiently large n and $\frac{1}{x_n} \to 0$ as $n \to \infty$.

Exercise 2.50. Referring to Theorem 2.36: assume that $\bar{y} \neq 0$ and prove that

$$\frac{x_n}{y_n} \to \frac{\bar{x}}{\bar{y}}$$

as $n \to \infty$.

Exercise 2.51. Let A and B be nonempty subsets of $\mathbb{R}$. We say that $A \leq B$ if $a \leq b$ for all $a \in A$ and all $b \in B$. Prove that $\sup A \leq \inf B$ if $A \leq B$.

Exercise 2.52. Let A and B be nonempty subsets of $\mathbb{R}$. What can you say about the supremum of $A \cup B$ in terms of $\sup A$ and $\sup B$? Prove your statement.

Exercise 2.53. Same question for

$$A + B = \{a + b : a \in A, b \in B\} \quad \text{and} \quad A - B = \{a - b : a \in A, b \in B\}$$

about the suprema and infima of $A + B$ and $A - B$ in terms of $\sup A$, $\sup B$, $\inf A$ and $\inf B$.

Exercise 2.54. Let $I_n = [a_n, b_n]$ be a sequence of closed intervals in $\mathbb{R}$ with the property that $I_{n+1} \subset I_n$ for all $n \in \mathbb{N}$. Prove that there exists $c \in \mathbb{R}$ such that $c \in I_n$ for every $n \in \mathbb{N}$.

Exercise 2.55. Suppose that X is a set and that $d : X \times X \to \mathbb{R}$ satisfies $d(x, y) = d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in X$. Prove that $d(x, y) \geq 0$ for all $x, y \in X$.

3 Contractions and non-monotone sequences

We now present another approach to conclude that Heron's sequence has a limit. In this approach we do not use the monotonicity of the sequence, but look at the size of the "increments"

$$\xi_n = x_n - x_{n-1}.$$

These increments can be used to reproduce x_n from x_0 because

$$x_n - x_0 = \underbrace{x_1 - x_0}_{\xi_1} + \cdots + \underbrace{x_n - x_{n-1}}_{\xi_n} = \underbrace{\sum_{k=1}^{n} \xi_k}_{S_n}. \tag{3.1}$$

The special case $x_0 = 1$ was dealt with in Chapter 2. In the exposition below we will take $x_0 > 0$ as a parameter that we can vary[1].

The strategy in Section 3.1 below will be to show that all these increments can not take the sequence x_n very far. To do so we look for estimates that guarantee that sums of the form

$$M_n = |\xi_1| + |\xi_2| + |\xi_3| + \cdots + |\xi_n|$$

remain bounded as $n \to \infty$, using the geometric series of Section 1.4 that Zeno never liked. In fact this will force the sequence x_n converge. The issue of general sums

$$S_n = \sum_{k=1}^{n} \xi_k$$

will be dropped for now, but will come back in Section 3.7, see Theorem 3.37.

3.1 Estimates for the increments

Exercise 3.1. For $x_0 > 0$ let the sequence $x_n > 0$ be defined by (1.2) in Exercise 1.2, i.e.

$$x_n = \frac{x_{n-1}}{2} + \frac{1}{x_{n-1}},$$

and let $\xi_n = x_n - x_{n-1}$. Show that

$$\xi_{n+1} = \xi_n \left(\frac{1}{2} - \frac{1}{x_{n-1} x_n} \right),$$

[1] By the way, variation of parameters helps in solving equations, see Exercise 3.44.

and that therefore

$$-\frac{1}{2} \leq \frac{\xi_{n+1}}{\xi_n} < \frac{1}{2} \tag{3.2}$$

for every $n \in \mathbb{N}$. Hint: you need $x_n x_{n-1} = \frac{1}{2}x_{n-1}^2 + 1$ for the inequalities.

From Exercise 3.2 it follows that

$$|x_{n+1} - x_n| = |\xi_{n+1}| \leq \frac{1}{2}\,|\xi_n| = \frac{1}{2}|x_n - x_{n-1}| \quad \text{for all} \quad n \in \mathbb{N}, \tag{3.3}$$

i.e. every consecutive increment is at least twice as small as the previous one. Now the first increment has norm $|\xi_1| = |x_1 - x_0|$, which may be large (depending on x_0). But every next increment is much smaller because[2]

$$|\xi_2| \leq \frac{1}{2}\,|\xi_1|, \quad |\xi_3| \leq \frac{1}{2}\,|\xi_2| \leq \frac{1}{4}\,|\xi_1|, \quad |\xi_4| \leq \frac{1}{8}\,|\xi_1|, \quad |\xi_5| \leq \frac{1}{16}\,|\xi_1| = \frac{1}{2^4}\,|\xi_1|,$$

and so on. It follows that

$$|\xi_n| \leq \frac{1}{2^{n-1}}\,|\xi_1| \tag{3.4}$$

for all $n \in \mathbb{N}$. Thus the increments get smaller and smaller exponentially fast.

Exercise 3.2. Let the map[3] f be defined by

$$f(x) = \frac{x}{2} + \frac{1}{x}$$

as in (2.2). Verify that f has the property that

$$\forall_{x\geq 1}\,\forall_{y\geq 1} : |f(x) - f(y)| \leq \frac{1}{2}\,|x - y|, \tag{3.5}$$

and that therefore the sequence x_n defined by $x_n = f(x_{n-1})$ has

$$|x_{n+1} - x_n| = |f(x_n) - f(x_{n-1})| \leq \frac{1}{2}\,|x_n - x_{n-1}|. \tag{3.6}$$

for all $n \in \mathbb{N}$ if $x_0 > 0$.

[2]The inequalities are strict unless the increments are zero.

[3]Or function, we shall prefer to use the word map for functions which are not $\mathbb{R}$-valued.

If f satisfies (3.5) then f is called *contractive* (with contraction factor $\frac{1}{2}$) on the set

$$A = [1, \infty) = \{x \in \mathbb{R} : \, x \geq 1\}.$$

This a special case of what is called Lipschitz continuity:

Definition 3.3. *Let $A \subset \mathbb{R}$. A function $f : A \to \mathbb{R}$ is called* Lipschitz continuous *with Lipschitz*[4] *constant $L > 0$ if for all $x, y \in A$ it holds that*

$$|f(x) - f(y)| \leq \, L|x - y|. \tag{3.7}$$

If $L < 1$ and $f(A) \subset A$ then f is called a contraction *with contraction factor L. If $f(A)$ is not necessarily a subset of A then f is called* contractive *if $L < 1$, and* nonexpanding *if $L = 1$.*

Exercise 3.4. Show that the map $x \to |x|$ is nonexpanding.

Exercise 3.5. Let $f : A \to \mathbb{R}$ be contractive. Prove that there can be at most one solution $x \in A$ to the equation $x = f(x)$. Hint: if there were two solutions you can use (3.7) with $L < 1$.

Exercise 3.6. This is a warming up exercise for what's to come. Let A be a subset of $\mathbb{R}$ and suppose that $f : A \to A$ is a contraction with contraction factor $\frac{1}{2}$. Suppose that the sequence x_n, defined by $x_n = f(x_{n-1})$ and some given $x_0 \in A$, converges to a limit $\bar{x}$ in A. Prove that $\bar{x}$ is *a solution* of $f(x) = x$. Hint:

$$|f(\bar{x}) - \bar{x}| \leq |f(\bar{x}) - f(x_n)| + |f(x_n) - \bar{x}| = \underbrace{|f(\bar{x}) - f(x_n)|}_{\leq \frac{1}{2}|\bar{x} - x_n|} + |x_{n+1} - \bar{x}|.$$

Recall from Exercise 3.5 that there is at most one solution to $f(x) = x$. What do you conclude about sequences starting from *other initial values* x_0 in A?

3.2 Properties of Heron's sequence due to contraction

Look at (3.1). What can happen to Heron's sequence x_n after say N steps? For $m > n$ the difference between x_m and x_n is equal to

$$x_m - x_n = x_{n+1} - x_n + \cdots + x_m - x_{m-1} = \xi_{n+1} + \cdots + \xi_m.$$

[4] We shall prefer another symbol when $L < 1$.

Using (3.4) it follows that

$$|x_m - x_n| \le |\xi_{n+1}| + \cdots + |\xi_m| \le \frac{|\xi_1|}{2^n} + \cdots + \frac{|\xi_1|}{2^{m-1}}.$$

Now go back to (1.18) and what we spelled out in Exercises 1.15 and 1.16 with the observation that

$$\forall_{m,n,N\in\mathbb{N}}: \quad m \ge n \ge N \implies \sum_{k=n}^{m} \frac{1}{2^k} < \frac{1}{2^{N-1}}.$$

It follows that

$$|x_m - x_n| \le |\xi_1| \sum_{k=n}^{m-1} \frac{1}{2^k} \le |\xi_1| \underbrace{\sum_{k=n}^{m} \frac{1}{2^k}}_{<\varepsilon}, \tag{3.8}$$

in which the ε-estimate holds for all m, n with $m > n \ge N$, provided N is as in Exercise 1.15. We conclude that for all $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that[5]

$$|x_n - x_m| < \varepsilon \quad \text{for all} \quad m, n \ge N,$$

which brings us to a crucial section next.

3.3 Cauchy sequences, monotone subsequences

We just concluded that the Heron sequence $x_1, x_2, x_3, \ldots$ has the property that

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{m,n\ge N}\, : \underbrace{|x_n - x_m|}_{d(x_n,x_m)} < \varepsilon, \tag{3.9}$$

a statement to be pronounced as: for all (real) $\varepsilon > 0$ there exists a natural number N such that for all natural numbers m, n with $m \ge N$ and $n \ge N$ the distance between x_n and x_m is smaller than ε.

Definition 3.7. *A sequence of real numbers x_n indexed by $n \in \mathbb{N}$ is called* Cauchy*, or a* Cauchy sequence*, if (3.9) holds.*

We already knew that Heron's sequence is convergent. Compare Definition 3.7 to Definition 2.6 in Section 2.2 for convergence of x_n. Unlike Definition 2.6 the new definition does not involve any number that candidates for being the limit of the sequence. Thus it may be verified without knowing the limit. Can it be used as an alternative definition of convergence?

[5] Also for $m = n$.

Exercise 3.8. Prove that every convergent sequence is a Cauchy sequence. Hint:

$$\underbrace{|x_n - x_m|}_{d(x_n,x_m)} \leq \underbrace{|x_n - \bar{x}|}_{d(x_n,\bar{x})} + \underbrace{|\bar{x} - x_m|}_{d(\bar{x},x_m)}.$$

Theorem 3.9. *A sequence is a convergent if and only if it is Cauchy.*

Proof of Theorem 3.9. Exercise 3.9 proves that every convergent sequence is Cauchy, so it remains to prove that every Cauchy sequence is convergent. We will do this in a number of steps, each of which by itself is not very hard, although Theorem 3.10 is rather clever.

Theorem 3.10. *Let x_n be a sequence of real numbers indexed by $n \in \mathbb{N}$. Then there exists a sequence of positive integers n_k, indexed by $k \in \mathbb{N}$, with the property that*

$$n_1 < n_2 < n_3 < \cdots,$$

and such that the subsequence x_{n_k}, indexed by k, is monotone[6].

Exercise 3.11. Prove Theorem 3.10. Hint: call an integer $m \in \mathbb{N}$ a *topindex* of the sequence x_n if $x_m > x_n$ for all $n > m$. A sequence may have no topindices at all. Show that it then has as a nondecreasing subsequence. A sequence may have only a finite number of topindices. Reduce this to the previous case. It remains to consider the case that the sequence has an infinite number of topindices. Conclude.

Exercise 3.12. Prove that every Cauchy sequence x_n is bounded, hence so is every subsequence x_{n_k} of x_n.

Exercise 3.13. Suppose that x_n is a Cauchy sequence of real numbers which has a convergent subsequence x_{n_k} with limit $\bar{x}$. Prove that the sequence x_n is itself convergent and that its limit is $\bar{x}$. That is to say

$$\lim_{n\to\infty} x_n = \lim_{k\to\infty} x_{n_k}.$$

[6]In particular this statement also holds for every sequence of rational numbers.

Once you know Theorem 3.10 you observe that every Cauchy sequence is bounded by Exercise 3.12. Thus so is the monotone subsequence provided by Theorem 3.10, which then has a limit in view of Theorem 2.28. By Exercise 3.13 this limit turns out to be the limit of the whole sequence as well. This completes the proof of Theorem 3.9. □

3.4 The Banach contraction theorem in $\mathbb{R}$

We have seen that if f is a contraction from a subset A of $\mathbb{R}$ to itself with contraction factor $\frac{1}{2}$, then every sequence defined by $x_n = f(x_{n-1})$ starting from any $x_0 \in A$ is convergent.

The reasoning started with estimate (3.8) and your answer to Exercise 2.31. We concluded that the sequence x_n had the property stated in (3.9), i.e. that it is a Cauchy sequence.

In Section 3.3 we then established a basic property of the real numbers with Theorem 3.9. It stated that every Cauchy sequence is convergent. In particular the sequence x_n defined by $x_n = f(x_{n-1})$ in Exercise 3.6 is convergent. Next we formulate a condition on A which implies that its limit is in A.

Definition 3.14. *A subset A of $\mathbb{R}$ is called* closed in $\mathbb{R}$ *if the convergence of a sequence $x_n \in A$ implies that its limit $\bar{x}$ is in A, i.e.*

$$A \ni x_n \to \bar{x} \quad \text{as} \quad n \to \infty \qquad \Longrightarrow \qquad \bar{x} \in A.$$

Let us now assume[7] that A is closed. Then $\bar{x} \in A$ if $\bar{x}$ is the limit of the sequence x_n in Exercise 3.6. By Exercise 3.6 it is the unique solution of the equation $f(x) = x$ in A.

This proves a special case of Theorem 3.16 below, namely for closed sets $A \subset \mathbb{R}$ and contractive maps f from A to A with contraction factor $\frac{1}{2}$. Here's the general theorem, which requires a definition first.

Definition 3.15. *Let A be a set and $f : A \to A$. Then $x \in A$ is called a fixed point of f if $x = f(x)$.*

Theorem 3.16. *(Banach contraction theorem for closed subsets of $\mathbb{R}$) Let A be a closed subset of $\mathbb{R}$ and let $f : A \to A$ be a contraction, i.e.*

$$\exists_{\theta \in (0,1)} \, \forall_{x,y \in A} : \, |f(x) - f(y)| \leq \theta \, |x - y|. \tag{3.10}$$

Then f has a unique fixed point $\bar{x} \in A$. For every $x_0 \in A$ this $\bar{x}$ is the limit of the sequence x_n defined by $x_n = f(x_{n-1})$ for all $n \in \mathbb{N}$.

[7] Or pray.

Proof of Theorem 3.16. We first formulate two essential ingredients for the proof as exercises.

Exercise 3.17. Assume that $\theta \in (0,1)$. Prove that $\theta^n \to 0$ as $n \to \infty$. This exercise generalises Exercise 2.31 and also establishes, somewhat overdue perhaps, (1.12) and (1.13). Hint: the sequence θ^n is decreasing[8].

Exercise 3.18. Prove for the sequence x_n defined in Theorem 3.16 that

$$|x_m - x_n| \le \theta^n|\xi_1| + \cdots + \theta^m|\xi_1| \le \frac{\theta^N}{1-\theta}\,|\xi_1|$$

for $m > n \ge N$.

These two exercises imply that x_n is a Cauchy sequence. Thus x_n is convergent and the limit $\bar{x}$ lies in A because A is closed.

We reason as in the hint for Exercise 3.6 to conclude. The *subtract, add, then triangle inequality trick* gives

$$|f(\bar{x})-\bar{x}| \le |f(\bar{x})-f(x_n)|+|f(x_n)-\bar{x}| = \underbrace{|f(\bar{x})-f(x_n)|}_{\le\theta|\bar{x}-x_n|}+|x_{n+1}-\bar{x}|, \quad (3.11)$$

in which the estimates depend on n, while what's being estimated clearly does not. To deal with the n-dependent final estimate in (3.11) we let $\varepsilon > 0$ and apply the definition of $x_n \to \bar{x}$, i.e. there is an $N \in \mathbb{N}$ such that $|\bar{x} - x_n| < \varepsilon$ for all $n \ge N$. We then conclude from (3.11) that[9]

$$|f(\bar{x}) - \bar{x}| \le \theta|\bar{x} - x_n| + |x_{n+1} - \bar{x}| < (\theta + 1)\varepsilon$$

for all $n \ge N$.

Since $\varepsilon > 0$ was arbitrary we conclude that $|f(\bar{x}) - \bar{x}| = 0$, so $f(\bar{x}) = \bar{x}$ is a fixed point of f. This limit $\bar{x}$ is in fact the *unique* solution of $x = f(x)$ in A, because (3.10) prevents the existence of two solutions. Indeed, for two solutions x and y with $x \ne y$ we would have that

$$0 < |x - y| = |f(x) - f(y)| \le \theta|x - y| < |x - y|$$

because $0 < \theta < 1$, a contradiction. This completes the proof of Theorem 3.16. □

[8]Incidentally, it is defined by $x_0 = 1$ and $x_n = \theta x_{n-1}$ for $n \in \mathbb{N}$.

[9]Note that with $n \ge N$ also $n + 1 \ge N$.

Remark 3.19. *You should carefully note that*

$$\text{we concluded that} \quad f(x_n) \to f(\bar{x}) \quad \text{because} \quad x_n \to \bar{x} \tag{3.12}$$

and f is contractive. The conclusion in (3.12) holds for a much larger class of functions than those satisfying (3.10) in fact. This will take us to the issue of continuity, but first we discuss a bit more about sequences and sets.

3.5 Convergent subsequences

We note that Theorems 2.28 and 3.10 also immediately imply Theorem 3.20 below, which will be essential for proving essential theorems[10] about continuous[11] functions later on.

Theorem 3.20. *(Bolzano-Weierstrass) Let x_n be a bounded sequence of real numbers indexed by $n \in \mathbb{N}$. Then x_n has a convergent subsequence.*

Proof. The standard proof of Theorem 3.20 is different. It is given in Section 3.9. In the proof here we simply observe that Theorem 3.10 states that every bounded sequence has a monotone (and also bounded) subsequence, and that Theorem 2.28 says this subsequence must be convergent. □

Definition 3.21. *A limit of a convergent subsequence of a sequence is called a* limit point *of the original sequence.*

Exercise 3.22. Prove that $\bar{x}$ is a limit point of the sequence x_n if and only if

$$\forall_{\varepsilon>0}\, \forall_{N\in\mathbb{N}}\, \exists_{n\geq N} \,:\, |x_n - \bar{x}| < \varepsilon.$$

Not easy, no hint. Test your abilities.

Remark 3.23. *Theorem 3.20 states for bounded sequences x_n of real numbers that*

$$\exists_{\bar{x}\in\mathbb{R}}\, \forall_{\varepsilon>0}\, \forall_{N\in\mathbb{N}}\, \exists_{n\geq N} \,:\, |x_n - \bar{x}| < \varepsilon,$$

a statement that looks deceptively similar to the statement (2.10) for convergence of a sequence x_n.

Theorem 3.24. *A bounded sequence of real numbers is convergent if and only if it has exactly one limit point.*

[10] Like having the integral $\int_a^b f$ to have a meaning for $f : [a, b] \to \mathbb{R}$ continuous.
[11] We used this term in relation to (2.7).

Exercise 3.25. Prove Theorem 3.24. Hint: by Theorem 3.20 the sequence has a limit point $\bar{x}$; to get one of the two implications in the statement of Theorem 3.24 assume the bounded sequence x_n does not converge and reason from (2.11); reapply Theorem 3.20 to obtain another limit point. For the other implication you are on your own.

3.6 Closed and open sets

This section is about points and sets in $\mathbb{R}$. We systematically use (2.9) and write $d(x, y)$ instead of $|x - y|$ to prepare this section to be carried[12] over to[13] Chapter 5. We recall that we defined in Definition 3.14 what a closed subset of $\mathbb{R}$ is.

Remark 3.26. *Informally Definition 3.14 says that a subset A of $\mathbb{R}$ is closed if you cannot get out of A by taking limits, which makes "closed" a natural adjective; "closed" and "bounded" are important adjectives for a set $A \subset \mathbb{R}$: bounded to have convergent subsequences of sequences in A by Theorem 3.20, closed to have their limits in A.*

Definition 3.27. *Let $A \subset \mathbb{R}$. Then $\xi \in \mathbb{R}$ is called an accumulation point of A if*

$$\forall_{\delta>0}\, \exists_{x\in A} \,: 0 < d(x, \xi) = |x - \xi| < \delta. \tag{3.13}$$

An accumulation point of A need not be in A. The name is explained by the following theorem.

Theorem 3.28. *Let $A \subset \mathbb{R}$. Then $\xi \in \mathbb{R}$ is an accumulation point of A if and only if there exists a sequence $x_n \in A$ with $x_n \neq \xi$ and $x_n \to \xi$.*

Proof. Let ξ be an accumulation point of A. We have to prove the existence of a sequence with the properties stated in Theorem 3.28. We use Definition 3.27. For each $n \in \mathbb{N}$ let $x_n \in A$ be provided by (3.13) with $\delta = \frac{1}{n}$. To prove that $x_n \to \xi$ let $\varepsilon > 0$ be arbitrary. Choose[14] $N \in \mathbb{N}$ with $\frac{1}{N} < \varepsilon$. Then

$$d(x_n, \xi) < \frac{1}{n} \leq \frac{1}{N} < \varepsilon$$

for all $n \geq N$, as desired for one of the two implications in the theorem. The other implication is left as Exercise 3.29. □

[12]Only Theorem 3.20 *will not generalise* to the metric space context in Chapter 5.
[13]With $\mathbb{R}$ replaced by X, X and d as in definition 5.1.
[14]This uses the Archimedean Principle in the form of Remark 1.6 again.

Exercise 3.29. Prove the opposite implication in Theorem 3.28: if such a sequence exists then its limit ξ is an accumulation point

Theorem 3.30. *Let $A \subset \mathbb{R}$. Then A is closed if and only if A contains all its accumulation points.*

Proof. Suppose ξ is an accumulation point of A. By Theorem 3.28 it is the limit of a sequence x_n in A and thereby in A if A is closed. So A contains all its accumulation points if A is closed.

Conversely, suppose A is not closed. Then there is a sequence x_n in A which converges to a limit $\bar{x}$ which is not in A. But then, by Theorem 3.28, $\bar{x}$ must be an accumulation point of A that is not in A. This completes the proof. □

Definition 3.31. *A point x_0 in a subset A of $\mathbb{R}$ is called an* interior point *of A if there exists $\delta > 0$ such that for all $x \in \mathbb{R}$ with $d(x, x_0) < \delta$ it holds that $x \in A$. That is to say*[15]

$$B_\delta(x_0) = \{x \in \mathbb{R} : \underbrace{|x - x_0|}_{d(x,x_0)} < \delta\} = (x_0 - \delta, x_0 + \delta) \subset A.$$

The set of all interior points of A is called the interior *of A, notation $int(A)$.*

Definition 3.32. *A subset $\mathcal{O}$ of $\mathbb{R}$ is called open if $int(\mathcal{O}) = \mathcal{O}$.*

Exercise 3.33. Prove that $B_\delta(x_0)$ in Exercise 3.31 is itself an open subset of $\mathbb{R}$. Hint: use the *triangle inequality* again.

Theorem 3.34. *A subset $A \subset \mathbb{R}$ is closed if and only if its complement*

$$A^c = \{x \in \mathbb{R} : x \notin A\}$$

in $\mathbb{R}$ is open.

Exercise 3.35. Prove Theorem 3.34. Hint: in the spirit of the proofs above.

Remark 3.36. *It is more common in the literature to first define what open sets are, and to then call a set closed if its complement is open.*

[15] In Chapter 5 the set $B_r(x_0)$ is called an open ball, but it's not. It's an open interval.

3.7 Absolute convergence of series

Theorem 3.37. *Let x_n be a sequence of real numbers indexed by $n \in \mathbb{N}$. Suppose that*

$$M_n = \sum_{k=1}^{n} |x_k| = |x_1| + \cdots + |x_n|$$

defines a bounded sequence M_n. Then M_n is convergent and the sequence defined by

$$S_n = \sum_{k=1}^{n} x_k = x_1 + \cdots + x_n$$

is also convergent. Its limit S satisfies

$$|S| \leq \bar{M} := \lim_{n\to\infty} M_n = \sup_{n\in\mathbb{N}} M_n \in \mathbb{R}.$$

Proof. Do the following two exercises. □

Exercise 3.38. Prove the convergence of both sequences M_n and S_n. Hint:

$$|S_n - S_m| = \left| \sum_{k=m+1}^{n} x_k \right| \leq \sum_{k=m+1}^{n} |x_k| = M_n - M_m \quad \text{for} \quad m, n \in \mathbb{N} \quad \text{with} \quad m < n.$$

Exercise 3.39. (continued) Show that

$$S := \lim_{n\to\infty} S_n = \lim_{n\to\infty} \sum_{k=1}^{n} x_k \tag{3.14}$$

satisfies

$$|S| \leq \bar{M} := \lim_{n\to\infty} M_n = \sup_{n\in\mathbb{N}} M_n \in \mathbb{R}.$$

See Exercise 5.39 for a general statement about absolutely convergent series.

Remark 3.40. *Informally we write*

$$\sum_{n=1}^{\infty} |x_n| < \infty \implies \left| \sum_{n=1}^{\infty} x_n \right| \leq \sum_{n=1}^{\infty} |x_n|, \tag{3.15}$$

to say that the series

$$\sum_{n=1}^{\infty} x_n$$

is absolutely convergent*, by which we merely mean that the monotone sequence M_n is bounded and thereby convergent. We then write*

$$S = \sum_{n=1}^{\infty} x_n. \tag{3.16}$$

It may of course happen that the sequence M_n is not bounded. Then (3.15) has no meaning but (3.14) may still hold for a number $S \in \mathbb{R}$. In that case we say that the series is convergent with sum S*, but* not absolutely convergent*.*

3.8 Unconditional convergence of series

Exercise 3.41. Think about

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots,$$

and show that it defines a real number. Hint: look at the partial sums with an even number of terms and the partial sums with an odd number of terms.

Can we manipulate with such sums like (3.16) as we do with finite sums? For instance,

$$x_0+x_1+x_2 = x_0+x_2+x_1 = x_1+x_0+x_2 = x_1+x_2+x_0 = x_2+x_0+x_1 = x_2+x_1+x_0$$

is 6 ways to write the same sum

$$\sum_{k=0}^{3} x_k.$$

We would similarly like to have that

$$S = \sum_{k=0}^{\infty} x_{\phi(n)} = \sum_{k=0}^{\infty} x_k \tag{3.17}$$

for every bijection $\phi : \mathbb{N}_0 \to \mathbb{N}_0$.

Proof of (3.17) if M_n is bounded. We wish to conclude for

$$S_n^\phi = \sum_{k=0}^{n} x_{\phi(n)} \quad \text{and} \quad \bar{M}_n^\phi = \sum_{k=0}^{n} |x_{\phi(n)}|,$$

that

$$S_n^\phi \to S \quad \text{and} \quad \bar{M}_n^\phi \to \bar{M} \tag{3.18}$$

as $n \to \infty$. Let's see how this can be done.

What we know is that

$$|S_n| \leq \bar{M}, \quad |S_n^\phi| \leq \bar{M}, \quad S_n \to S, \quad M_n \to \bar{M}, \quad |S| \leq \bar{M}.$$

So for all $\varepsilon > 0$ there exists an integer $N \in \mathbb{N}_0$ such

$$\bar{M} - \varepsilon < \sum_{k=0}^{N} |x_k| \leq \bar{M}, \tag{3.19}$$

for otherwise $\bar{M}$ is not the lowest upper bound. But then also

$$\bar{M} - \varepsilon < \sum_{k=0}^{n} |x_k| \leq \bar{M}$$

for all $n \geq N$. This is just the proof that

$$M_n \to \bar{M} = \sum_{k=0}^{\infty} |x_k|$$

redone.

Subtracting the partial sum in (3.19) from (3.19) we obtain in particular that

$$\sum_{k=N+1}^{\infty} |x_k| - \varepsilon < 0 \leq \sum_{k=N+1}^{\infty} |x_k|,$$

whence

$$\sum_{k=N+1}^{\infty} |x_k| < \varepsilon. \tag{3.20}$$

Now what about $\bar{M}_n^\phi$? The bijection $\phi : \mathbb{N}_0 \to \mathbb{N}_0$ is a permutation of $\mathbb{N}_0$. If we enumenerate

$$\mathbb{N}_0 = \{k = \phi(l) : l \in \mathbb{N}_0\}$$

via ϕ with $l \in \mathbb{N}_0$, then

$$\{0, 1, \ldots, N\} \subset \{\phi(0), \phi(1), \ldots, \phi(L)\}$$

for some $L \in \mathbb{N}_N$ Therefore

$$\bar{M} - \varepsilon < M_N \leq \bar{M}_L^\phi \leq \bar{M}_l^\phi \leq \bar{M}.$$

if $l \geq L$. We also have that

$$|S_l^\phi - S_N| \leq \sum_{k=N+1}^{\infty} |x_k| < \varepsilon,$$

because $S_l^\phi - S_N$ is a finite sum of terms x_k with $k > N$ if $m \geq L$. The proof of(3.17) is completed by the following exercise. □

Exercise 3.42. Show that S_n^ϕ converges to the same sum $S \in \mathbb{R}$ if M_n is bounded.

3.9 Extra: another diagonal argument

In this section we give the standard proof of Theorem 3.20. A similar argument will also be used and explained in the proof of the Arzéla-Ascoli Theorem.

Another proof of Theorem 3.20 (Bolzano-Weierstrass). Assume $x_n \in \mathbb{R}$ is a bounded sequence, say $x_n \in [0, 1]$. Then at least one of the intervals $[\frac{0}{2}, \frac{1}{2}]$, $[\frac{1}{2}, \frac{2}{2}]$ must contain x_n for infinitely many values of n. Call this interval

$$I_1 = \left[\frac{m_1}{2}, \frac{m_1 + 1}{2}\right].$$

So $m_1 = 0$ or $m_1 = 1$. Enumerate these n as $n_{1j} \in \mathbb{N}$. The first index 1 indicates that this is the first subsequence we choose.

Apply the same argument again. One of $[\frac{m_1}{2} + \frac{0}{4}, \frac{m_1}{2} + \frac{1}{4}]$ and $[\frac{m_1}{2} + \frac{1}{4}, \frac{m_1}{2} + \frac{2}{4}]$ must contain a further subsequence. Call this interval

$$I_2 = \left[\frac{m_1}{2} + \frac{m_2}{4}, \frac{m_1}{2} + \frac{m_2 + 1}{4}\right],$$

and enumerate this subsequence as $n_{2j} \in \mathbb{N}$. And so on. We obtain further and further subsequences

$$x_{n_{kj}} \in I_k = \left[\sum_{l=1}^{k} \frac{m_l}{2^l}, \sum_{l=1}^{k} \frac{m_l}{2^l} + \frac{1}{2^{k+1}}\right] = [a_k, b_k],$$

and the diagonal subsequence has

$$x_{n_{kk}} \in I_k = [a_k, b_k]$$

for every k. The proof will be completed in the following exercise. □

Exercise 3.43. Finish this proof of Theorem 3.20. Hint: $a_k \leq x_{n_{kk}} \leq b_k$ and the sequences a_k, b_k are monotone and have $b_k - a_k = 2^{-k}$.

3.10 Exercises

Exercise 3.44. Solve the equation $x^3 + x = q$ using Cardano's trick $x = y + z$ and an additional equation for y and z which gets rid of the terms $y^2 z$ and yz^2. Compare what you get to the obvious "solution" $q = x^3 + x$ for the parameter q.

Exercise 3.45. Referring to Definition 3.3, let $f : A \to \mathbb{R}$ be Lipschitz continuous and assume that x_n is a convergent sequence in A. Prove that the sequence $f(x_n)$ is convergent. Then, denoting the limit of x_n by L, assume that y_n is another convergent sequence in A with the same limit L. Prove that

$$\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} f(y_n).$$

Exercise 3.46. For each of the functions in Exercise 2.45 find a closed subset $A \subset \mathbb{R}$ such that $f : A \to A$ is a contraction.

Exercise 3.47. Determine all limit points of the sequences defined by $x_n = (-1)^n$, $x_n = (-1)^n + \frac{1}{n}$, $x_n = (-1)^n + (-1)^{2n}$.

Exercise 3.48. Let x_n be an enumeration of $\mathbb{Q}$. Prove that every element of $\mathbb{R}$ is a limit point of this sequence. Hint: use that every $\bar{x} \in \mathbb{R}$ appears as the limit of a sequence in $\mathbb{Q}$.

Exercise 3.49. For $a > 0$ let the sequence x_n be defined by

$$x_n = \frac{1}{2}\left(x_{n-1} + \frac{a}{x_{n-1}}\right) \quad \text{and} \quad x_0 = 1.$$

Does the sequence convergence? If so prove it and determine (the square of) the limit.

Exercise 3.50. For $a > 1$ let the sequence x_n be defined by

$$x_n = \frac{1}{2}\left(x_{n-1} + \frac{a}{x_{n-1}^2}\right) \quad \text{and} \quad x_0 = 1.$$

Does the sequence converge? If so prove it and determine (the cube of) the limit.

Exercise 3.51. Same question for $0 < a < 1$.

Exercise 3.52. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \frac{x}{1+x^2}.$$

Prove that f is Lipschitz continuous with Lipschitz constant $L = 1$. Hint: use your fractional abilities to write

$$f(x) - f(y) = (x-y)\,\frac{\cdots}{\cdots}$$

and rework the quotient as the difference of two terms, one of which is $f(x)f(y)$. Use this to first show that $|f(x) - f(y)| < |x - y|$ if $x, y \geq 0$ and $x \neq y$.

Exercise 3.53. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \frac{x}{2+x^2}.$$

Prove that f is a contraction. Hint: use your tricks from exercise 3.52.

Exercise 3.54. Let $a, b \in \mathbb{R}$ with $a < b$. Prove that the closed interval $[a, b]$ is a closed subset of $\mathbb{R}$.

Exercise 3.55. Let $a, b \in \mathbb{R}$ with $a < b$. Prove that the open interval (a, b) is an open subset of $\mathbb{R}$.

Exercise 3.56. Let $a, b \in \mathbb{R}$ with $a < b$. Prove that the intervals $(a, b]$ and $[a, b)$ are neither closed nor open in $\mathbb{R}$.

Exercise 3.57. Let A and B be closed subsets of $\mathbb{R}$. Prove that $A \cup B$ and $A \cap B$ are closed.

Exercise 3.58. Let I be any index set and let $A_i \subset \mathbb{R}$ be closed for every $i \in I$. Prove that the intersection

$$\cap_{i \in I} A_i = \{x \in \mathbb{R} : \, x \in A_i \text{ for all } i \in I\}$$

is closed.

Exercise 3.59. Formulate and prove similar statements for open subsets.

Exercise 3.60. Let G_n be a sequence of closed subsets of $\mathbb{R}$ with the property that $G_{n+1} \subset G_n$ for all $n \in \mathbb{N}$. Such sequences are called *nested*. Is it necessarily true that there exists $c \in \mathbb{R}$ such that $c \in G_n$ for every $n \in \mathbb{N}$? If not, which additional assumption is required?

Exercise 3.61. Consider the set C of numbers

$$\sum_{n=1}^{\infty} \frac{t_n}{3^n},$$

with $t_n \in \{0, 2\}$ for every $n \in \mathbb{N}$, but no further restrictions[16]. Prove that C is a closed uncountable set with empty interior, and that for two such numbers

$$\sum_{n=1}^{\infty} \frac{t_n}{3^n} = \sum_{n=1}^{\infty} \frac{\tilde{t}_n}{3^n} \iff \forall_{n \in \mathbb{N}} : \, t_n = \tilde{t}_n.$$

[16]Unlike in the context of (1.6), when expansions ending in only zero's were excluded.

Hint: construct C from a sequence of nested closed sets C_n of such numbers with $t_n \in \{0, 1, 2\}$, and $t_1, \ldots, t_n \neq 1$. The representation of numbers in C_n is not unique but in C it is. The set C is called *Cantor's discontinuum.*

Exercise 3.62. (continued) Describe $D = \{x \in [0, 1] : x \notin C\}$ as a countable disjoint union of open intervals indexed by a *binary tree.*

Exercise 3.63. Let x_n be a convergent sequence of real numbers indexed by $n \in \mathbb{N}$, and let

$$\xi_n = \frac{1}{n} \sum_{k=1}^{n} x_k = \frac{1}{n}(x_1 + \cdots + x_n).$$

Does

$$\lim_{n \to \infty} \xi_n$$

exist? Prove your answer. Can it happen that this limit exists if the sequence x_n is divergent?

Exercise 3.64. Same question for

$$S_n = \sum_{k=1}^{n} x_k \quad \text{and the "Cesàro" sums} \quad \sigma_n = \frac{1}{n}(S_1 + \cdots + S_n).$$

Exercise 3.65. Consider the series in Exercise 3.41. Its sum is $\ln 2$. Show that by carefully choosing the bijection $\phi : \mathbb{N} \to \mathbb{N}$ the sequence S_n^{ϕ} can actually be made to converge to zero[17].

[17] Or any other number you like.

4 Normed algebras of continuous functions

This chapter is mainly about the set $C([a,b])$ of functions $f : [a,b] \to \mathbb{R}$ which have the property that f is continuous in every point[1] of $[a,b]$. Here $[a,b]$ is a given bounded closed interval with $a < b$. Our tools will be

sequences of real numbers;
the equivalent[2] definitions of convergent and Cauchy sequences;
the elementary properties of convergent sequences;
the Bolzano-Weierstrass Theorem;
we also need the suprema and infima of bounded sets of real numbers[3].

In fact the existence of convergent subsequences of bounded sequences in $\mathbb{R}$ will be needed for the proof that (1.16) is indeed a proper definition of the "absolute value" of a function $f \in C([a,b])$.

Recall that we use the notation

$$x_n \to \bar{x}$$

to say that

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N} \,:\, \underbrace{|x_n - \bar{x}|}_{d(x_n,\bar{x})} < \varepsilon.$$

Definition 4.1. *Let $A \subset \mathbb{R}$ be nonempty, $f : A \to \mathbb{R}$ and $\xi \in A$. Then f is called* continuous *in ξ if*

$$f(x_n) \to f(\xi)$$

for every sequence x_n in A with the property that

$$x_n \to \xi.$$

If f is continuous in every $\xi \in A$ then $f : A \to \mathbb{R}$ is called continuous.

Remark 4.2. *If f fails to be continuous in ξ, then it is still possible that there exists $L \in \mathbb{R}$ such that*

$$f(x_n) \to L$$

for every sequence x_n in A with $x_n \neq \xi$ and $x_n \to \xi$. In that case we say that the limit

$$\lim_{x\to\xi} f(x)$$

exists and is equal to L. This terminology makes sense if and only if ξ is an accumulation point of A, and there's no need to assume that $\xi \in A$.

[1] First mentioned in (2.7), Definition 4.1 should not come unexpected.
[2] In hindsight.
[3] See Section 2.4.

4.1 Extrema and the maximum norm

One of the highlights of analysis is that a real valued continuous function that is defined on a closed and bounded subset A of $\mathbb{R}$, has a global maximum and global minimum on A. Here's a definition that is needed to formulate this result more precisely.

Definition 4.3. *Let A be a set and let $f : A \to \mathbb{R}$ a real valued function. If $\bar{x} \in A$ has the property that $f(x) \leq f(\bar{x})$ for every $x \in A$, then $M = f(\bar{x})$ is called a* global maximum *of f and $\bar{x}$ is called a* maximizer *of f. Likewise, if $\underline{x} \in A$ has the property that $f(x) \geq f(\underline{x})$ for every $x \in A$, then $m = f(\underline{x})$ is called a* global minimum *of f and $\underline{x}$ is called a* minimizer *of f.*

The question which functions $f : A \to \mathbb{R}$ have global *extrema*, i.e global maxima and minima, is a central issue in analysis.

Theorem 4.4. *Let $A \subset \mathbb{R}$ be a nonempty bounded closed subset, and let $f : A \to \mathbb{R}$ be continuous (in every point of A). Then f has a global maximum and a global minimum on A.*

Proof of Theorem 4.4. With Theorem 3.20 the hard work has already been done. Let

$$R_f = \{f(x) : x \in A\}$$

be the range of f.

Suppose R_f is bounded from above. Theorem 2.26 says that R_f has a smallest lower bound which we call M. By definition every $M - \frac{1}{n}$ with $n \in \mathbb{N}$ is not an upper bound then. Therefore there exist $x_n \in A$ with

$$M - \frac{1}{n} < f(x_n) \leq M.$$

It follows that $f(x_n) \to M$.

If R_f is not bounded then no $n \in \mathbb{N}$ is an upper bound. Then we know that for every $n \in \mathbb{N}$ there exist $x_n \in A$ with $f(x_n) > n$.

In both cases the sequence x_n is bounded because it is contained in the bounded set A. So in both cases it has a convergent subsequence x_{n_k} because of Theorem 3.20. The limit $\bar{x}$ is in A because A is closed[4]. Since f is continuous in $\bar{x}$ it follows from Definition 4.1 that $f(x_{n_k}) \to f(\bar{x})$. Proposition 2.9 then says that $f(x_{n_k})$ is a bounded sequence. This excludes the possibility $f(x_n) > n$ for every $n \in \mathbb{N}$.

Thus R_f is bounded. With both limits $f(x_n) \to M$ and $f(x_{n_k}) \to f(\bar{x})$ then established, it follows that $M = f(\bar{x})$. This is because Proposition 2.11

[4]You showed this for $A = [a, b]$ in Exercise 3.54.

says the limit of the convergent subsequence $f(x_{n_k})$ is unique. But then $M = f(\bar{x})$ is the global maximum of f, and $\bar{x}$ is a maximizer[5].

The argument for the global minimum is similar. This completes the proof of Theorem 4.4. □

Definition 4.5. *Let $[a,b] \subset \mathbb{R}$ be a closed interval. The set of all continuous functions $f : [a,b] \to \mathbb{R}$ is denoted by $C([a,b])$. Because $[a,b]$ is closed and bounded we can now define the number*

$$|f|_{max} = \max_{a \le x \le b} |f(x)| \in \mathbb{R}$$

for every $f \in C([a,b])$. This number is called the maximum norm *of f.*

Exercise 4.6. Let $f \in C([a,b])$ and $\varepsilon > 0$. Explain very carefully why

$$|f|_{\mathsf{max}} < \varepsilon \iff \forall_{x \in [a,b]} : |f(x)| < \varepsilon.$$

Hint: explain first that the function $|f|$, defined by $|f|(x) = |f(x)|$, is in $C([a,b])$, and that

$$\big|\,|f|\,\big|_{\mathsf{max}} = |f|_{\mathsf{max}}.$$

Theorem 4.7. *Let $f, g \in C([a,b])$. Define the functions $f+g$ and fg by*

$$(f+g)(x) = f(x) + g(x) \quad \text{and} \quad (fg)(x) = f(x)g(x).$$

Then $f+g \in C([a,b])$, $fg \in C([a,b])$, and

$$|f+g|_{max} \le |f|_{max} + |g|_{max} \quad \text{and} \quad |fg|_{max} \le |f|_{max}\,|g|_{max}.$$

Proof of Theorem 4.7. There is not much to prove. Thanks to Theorem 2.15, Theorem 2.36 and Definition 4.1, the functions $f+g$ and fg are in $C([a,b])$.

For example, let ξ be any point in $[a,b]$ and x_n a sequence in $[a,b]$ with $x_n \to \xi$. Then $f(x_n) \to f(\xi)$ and $g(x_n) \to g(\xi)$ by Definition 4.1, because f and g are continuous in ξ. By Theorem 2.36 we therefore have that the sequence $f(x_n)+g(x_n)$ converges to $f(\xi)+g(\xi)$ and the sequence $f(x_n)g(x_n)$ to $f(\xi)g(\xi)$. This holds for every sequence $x_n \to \xi$ with $x_n \in [a,b]$, definition 4.1 then says that $f+g$ and fg are continuous in ξ. Moreover, the argument is valid for every $\xi \in [a,b]$. Thus $f+g, fg \in C([a,b])$.

[5]Which need not be unique, see Exercise 4.35.

Finally, let $\bar{x}, \bar{y}, \bar{z}, \bar{w} \in [a,b]$ be the maximizers for the continuous[6] functions $|f|$, $|g|$, $|f+g|$, $|fg|$ respectively. Then

$$|f+g|_{\max} = |f(\bar{z}) + g(\bar{z})| \leq |f(\bar{z})| + |g(\bar{z})| \leq |f(\bar{x})| + |g(\bar{y})| = |f|_{\max} + |g|_{\max}$$

and

$$|fg|_{\max} = |f(\bar{w})|\,|g(\bar{w})| \leq |f(\bar{x})|\,|g(\bar{y})| = |f|_{\max}\,|g|_{\max}.$$

This completes the proof. □

4.2 Uniform convergence

Definition 4.8. *For $f, g \in C([a,b])$ the number*

$$d(f,g) = |f-g|_{max} = \max_{a\leq x\leq b} |f(x) - g(x)|, \tag{4.1}$$

is called the uniform distance *between f and g.*

A sequence of functions f_n in $C([a,b])$ is called uniformly convergent *if there exists $f \in C([a,b])$ such that*

$$d(f_n, f) = |f_n - f|_{max} \to 0 \quad as \quad n \to \infty,$$

i.e. if

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N} \,:\, d(f_n,f) = |f_n - f|_{max} < \varepsilon.$$

The sequence f_n is called a uniform Cauchy sequence *if*

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{m,n\geq N} \,:\, d(f_n,f_m) = |f_m - f_n|_{max} < \varepsilon, \tag{4.2}$$

Exercise 4.9. Take $a = 0, b = 1$, $f(x) = x^2$, $g(x) = x(1-x)$. Compute $d(f,g)$. Hint: sketch the graphs of $y = f(x)$ and $y = g(x)$ in the xy-plane and explain what $d(f,g)$ is before you actually compute it. Then draw the graphs of some other functions f for which $d(f,g)$ has the same value. Wat are the largest and smallest of such functions?

Exercise 4.10. Show that there are bounded sequences in $C([a,b])$ which do not have any uniformly convergent subsequence. Hint: $[a,b] = [0,1]$, $f_n(x) = x^n$. Do two arguments. One by contradiction: which function would the limit have to be? The other argument by an explicit calculation of $|f_n - f_m|_{\max}$ for which you use your calculus abilities.

[6] See the hint in Exercise 4.6.

Proposition 4.1. *For all $f, g, h \in C([a,b])$ it holds that*

$$d(f,f) = 0; \tag{4.3}$$

$$d(f,g) = d(g,f) > 0 \quad \text{if} \quad f \neq g; \tag{4.4}$$

$$d(f,g) \leq d(f,h) + d(h,g). \tag{4.5}$$

Exercise 4.11. Prove Proposition 4.1. Explain why (4.5) is called the *triangle inequality*. The property in (4.4) that $d(f,g) = d(g,f)$ is called the *symmetry* of f. The property in (4.4) that $d(f,g) > 0$ if $f \neq g$ is called the *positivity* of d. Note the similarity with the distance function (2.9) on $\mathbb{R}$.

The following theorem is the counterpart for sequences in $C([a,b])$ of one of the two implications in Theorem 3.9 for sequences in $\mathbb{R}$.

Theorem 4.12. *Let f_n be a uniform Cauchy sequence in $C([a,b])$. Then f_n is uniformly convergent. Its limit is defined by the (pointwise) limit*

$$f(x) = \lim_{n\to\infty} f_n(x)$$

for every $x \in [a,b]$. In particular, $f \in C([a,b])$.

Exercise 4.13. Formulate and prove the counterpart for sequences in $C([a,b])$ of the other implication in Theorem 3.9.

Proof of Theorem 4.12. Let f_n be a Cauchy sequence in $C([a,b])$ and let $\varepsilon > 0$. Then there exists $N \in \mathbb{N}$ such that

$$\underbrace{|f_n - f_m|_{\max}}_{d(f_n,f_m)} < \varepsilon \quad \text{for all} \quad m,n \geq N. \tag{4.6}$$

Note that N depends on $\varepsilon > 0$. By Exercise 4.6 the statement in (4.6) is equivalent to

$$\forall_{m,n\geq N}\, \forall_{\xi\in[a,b]} \quad |f_n(\xi) - f_m(\xi)| < \varepsilon, \tag{4.7}$$

with N depending only on ε. We say that f_n is a *uniform Cauchy sequence*. In particular it holds for every $\xi \in [a,b]$ that $f_n(\xi)$ is a Cauchy sequence in $\mathbb{R}$ and thereby convergent. We denote its limit by $f(\xi)$.

Since $\xi \in [a,b]$ was arbitrary this defines a function $f : [a,b] \to \mathbb{R}$. Moreover, for every fixed $\xi \in [a,b]$ and every fixed $n \geq N$ we can take the limit of the left hand side of (4.7) as $m \to \infty$. Exercise 2.33 then tells us that

$$|f_n(\xi) - f(\xi)| \leq \varepsilon \tag{4.8}$$

for all $n \geq N$. Recall that N depends on $\varepsilon > 0$, but not on ξ.

Suppose that $f \in C([a,b])$. We can then take the maximum of (4.8) over $\xi \in [a,b]$ and conclude that

$$d(f_n, f) = |f_n - f|_{\max} = \max_{a \leq \xi \leq b} |f_n(\xi) - f(\xi)| \leq \varepsilon$$

for all $n \geq N$, and this would complete the proof. In fact the continuity of f is consequence of the statement in Theorem 4.14 below. With a proof of Theorem 4.14 the proof of Theorem 4.12 will thus be complete.

Theorem 4.14. *Let f_n be a sequence in $C([a,b])$, and let f be another function from $[a,b]$ to $\mathbb{R}$. If*

$$\forall_{\varepsilon>0}\, \exists_{N \in \mathbb{N}}\, \forall_{n \geq N}\, \forall_{x \in [a,b]} \,:\, |f_n(x) - f(x)| \leq \varepsilon, \tag{4.9}$$

then f is in $C([a,b])$.

Proof of Theorem 4.14. Let $\xi \in [a,b]$. To prove that f is continuous in ξ let x_k be a sequence converging to ξ. We need to show that $f(x_k) \to f(\xi)$ as $k \to \infty$.

Let $\varepsilon > 0$. The splitting

$$f(x_k) - f(\xi) = f(x_k) - f_n(x_k) + f_n(x_k) - f_n(\xi) + f_n(\xi) - f(\xi)$$

implies that

$$|f(x_k) - f(\xi)| \leq \underbrace{|f(x_k) - f_n(x_k)|}_{\leq \varepsilon} + |f_n(x_k) - f_n(\xi)| + \underbrace{|f_n(\xi) - f(\xi)|}_{\leq \varepsilon}.$$

We indicated with underbraces that (4.9) can be applied to two of the terms. The inequalities hold for all $n \geq N$.

In particular it follows with $n = N$ that

$$|f(x_k) - f(\xi)| \leq 2\varepsilon + |f_N(x_k) - f_N(\xi)| \tag{4.10}$$

before we let $k \to \infty$. The second term on the right hand side of (4.10) goes to 0 as $k \to \infty$. Thus we can combine (4.10) with the continuity of f_N in ξ, and use that

$$f_N(x_k) \to f_N(\xi)$$

as $k \to \infty$ because $x_k \to \xi$. It follows that there must exist $K \in \mathbb{N}$ such that

$$|f(x_k) - f(\xi)| \leq 2\varepsilon + \underbrace{|f_N(x_k) - f_N(\xi)|}_{<\varepsilon} < 3\varepsilon$$

for all $k \geq K$.

Remark 2.38 with $M = 3$ now tells us that the proof of Theorem 4.14 is complete. Thus the proof of Theorem 4.12 is also complete. We don't forget to record the property of sequences formulated by Theorem 4.14 in a definition for functions that are not necessarily continuous. □

Definition 4.15. *A sequence of functions $f_n : [a,b] \to \mathbb{R}$ is called uniformly convergent on $[a,b]$ with limit $f : [a,b] \to \mathbb{R}$ if (4.9) holds, or equivalently*[7]*, if*

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N}\, \forall_{x\in[a,b]} \,:\, |f_n(x) - f(x)| < \varepsilon.$$

Theorem 4.14 says that the limit of a uniformly convergent sequence of functions f_n inherits the continuity properties of f_n. This is formulated a bit sharper in the following exercise.

Exercise 4.16. Let the sequence of functions $f_n : [a,b] \to \mathbb{R}$ be uniformly convergent on $[a,b]$ with limit $f : [a,b] \to \mathbb{R}$, and let $\xi \in [a,b]$. If the functions f_n are all continuous in ξ, then so is f. Prove this statement by adapting the proof of Theorem 4.14.

Remark 4.17. *Observe that $C([a,b])$ is a lot like $\mathbb{R}$ as far as multiplication, addition and norms are concerned. A minor difference is that in general*

$$|fg|_{max} \leq |f|_{max}\, |g|_{max}$$

does not hold with equality[8]*. A major difference is that there is no Theorem 3.20 for $C([a,b])$: bounded sequences do not have to have convergent subsequences, as Exercise 4.10 showed. Because of the properties in Theorem 4.7 and Theorem 4.12 we say that $C([a,b])$ is a* complete[9] normed algebra*. Such algebras are also called* Banach algebras*. In particular it is also a* complete normed (vector) space*. Such spaces are called* Banach spaces*. Exercise 4.36 gives another nice example.*

Here we will not bother you with definitions like "an algebra is a set on which two operations are defined denoted by ...", *and continue with* "such

[7] See Remark 2.38.

[8] Whereas $|xy| = |x|\,|y|$ for alle $x, y \in \mathbb{R}$.

[9] The word "complete" will be explained in Definition 5.4.

that the following rules hold ...". *But we do distinguish between algebras in which multiplication is commutative and algebras in which it's not*[10].

Remark 4.18. *The space $C([a,b])$ can be used for the construction of solutions of differential equations, via a transformation to so-called integral equations. Such integral equations will be solved via Theorem 5.7. We note that $C([a,b])$ is also a natural function space on which to consider the (linear) map*

$$f \to \int_a^b f(x)\,dx,$$

once this integral has been properly defined[11]*. It is is contained in the Banach algebra $B([a,b])$ of all bounded functions, which are normed by*

$$|f|_\infty = \sup_{a\le x\le b} |f(x)|. \tag{4.11}$$

The completeness of $B([a,b])$ follows (much easier) along the lines of the proof of Theorem 4.12. For $f \in C([a,b])$ the supremum norm in (4.11) is just the maximum norm announced in (1.16), see Definition 4.5.

4.3 Exercises

Exercise 4.19. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^3$. Prove directly from Definition 4.1 that f is continuous.

Exercise 4.20. Prove that in Definition 4.1 it is sufficient to verify the condition for monotone sequences $x_n \to \xi$. Hint: assume $x_n \to a$ but $f(x_n) \not\to f(a)$ as $n \to \infty$. Then for some $\varepsilon > 0$ no N exists with $|f(x_n) - f(a)| < \varepsilon$ for all $n \ge N$. Apply Theorem 3.10 to a suitably chosen[12] subsequence of x_n to derive a contradiction.

Exercise 4.21. Let $f : [0,1] \to [0,1]$ be defined by $f(x) = \sqrt{x}$. Prove that f is continuous. Hint: you may prefer to work with monotone sequences in Definition 4.1.

[10] See the footnotes in Section 1.5.

[11] The commonly used space in fact, but we'll have second thoughts in Section 7.6.

[12] See also Exercise 5.27.

Exercise 4.22. Let $g : \mathbb{R} \to \mathbb{R}$ be a function with $|g(x)| \leq 1$ for all $x \in \mathbb{R}$. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = xg(x)$. Prove directly from Definition 4.1 that f is continuous in $x = 0$.

Exercise 4.23. Let $g : \mathbb{R} \to \mathbb{R}$ be a function with $|g(x)| \leq 100 + x^{100}$ for all $x \in \mathbb{R}$. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = xg(x)$. Prove directly from Definition 4.1 that f is continuous in $x = 0$.

Exercise 4.24. Let $g : \mathbb{R} \to \mathbb{R}$ be a function with $|g(x)| \leq \frac{1}{x^2}$ for all $x \in \mathbb{R}$ with $x \neq 0$. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x^3 g(x)$. Prove directly from Definition 4.1 that f is continuous in $x = 0$.

Exercise 4.25. Let A be a subset of $\mathbb{R}$. Use Definition 3.27 to show there are sequences x_n in A with $x_n \neq \xi$ and $x_n \to \xi$ if and only if ξ is an accumulation point of A.

Exercise 4.26. Let A be a subset of $\mathbb{R}$, let $f : A \to \mathbb{R}$ and assume that $\xi \in A$ is an accumulation point of A. Explain why Remark 4.1 implies that f is continuous in ξ if and only if

$$\lim_{x \to \xi} f(x)$$

exists *and* is equal to $f(\xi)$.

Exercise 4.27. Let A be a subset of $\mathbb{R}$, let $f : A \to \mathbb{R}$ and assume that $\xi \in A$ is *not* an accumulation point of A. Explain why Definition 4.1 says that f is continuous in ξ.

Exercise 4.28. Let $I \subset \mathbb{R}$ be an nonempty open interval, let $f : I \to \mathbb{R}$, and ξ in I. Adapt Definition 4.1 to include a proper statement of what it means for

$$\lim_{x \downarrow \xi} f(x) \quad \text{and} \quad \lim_{x \uparrow \xi} f(x)$$

individually to exists.

Exercise 4.29. Let $I \subset \mathbb{R}$ be an nonempty open interval, and let $f : I \to \mathbb{R}$ be nonincreasing. Prove that

$$f(\xi^+) := \lim_{x \downarrow \xi} f(x) \quad \text{and} \quad f(\xi^-) := \lim_{x \uparrow \xi} f(x)$$

exist for every ξ in I, and that $f(\xi^-) \leq f(\xi^+)$.

Exercise 4.30. (continued) Prove that

$$\{\xi \in I : \, f(\xi^-) < f(\xi^+)\}$$

is finite or countable. Hint: consider open subintervals $(a, b) \subset I$ first.

Exercise 4.31. Construct a nondecreasing continuous function $f : [0, 1] \to [0, 1]$ with $f(0) = 0$, $f(1) = 1$ which is constant on every open interval in the disjoint union that describes the set D in Exercise 3.62. Hint: take the values on these intervals to be fractions with denominators equal to a power of 2.

Exercise 4.32. Construct a nondecreasing function $f : \mathbb{R} \to \mathbb{R}$ which is discontinuous in every $q \in \mathbb{Q}$ but continuous in every $\xi \notin \mathbb{Q}$. Hint: for every $q \in \mathbb{Q}$ let $H_q(x) = 0$ for $x < q$ and $H_q(x) = 1$ for $x \geq q$. Enumerate $\mathbb{Q}$ as a sequence q_n and consider

$$\sum_{n=1}^{\infty} \frac{1}{n^2} H_{q_n}(x).$$

Use Exercise 2.44.

Exercise 4.33. Let A be a subset of $\mathbb{R}$. Then A is called (sequentially) compact if every sequence in A has a convergent subsequence with its limit also in A. Prove that A is compact if and only if A is both bounded and closed.

Exercise 4.34. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \left((x+1)^2 + 3\right)^4 \left((x+5)^6 + 7\right)^8.$$

Prove that f has a global positive minimum. Hint: don't try to compute the minimizer but apply Theorem 4.3 with $A = [-R, R]$; specify a value of $R > 0$ for which the minimum m_R has $m_R < f(-5) \leq f(x)$ for all x with $|x| \geq R$.

Exercise 4.35. Let C be the Cantor set from Exercise 3.61 and let $f : C \to \mathbb{R}$ be defined by $f(x) = x(1-x)$. Explain why f has a global maximum on C, then find its maximizers.

Exercise 4.36. Let $f : \mathbb{R} \to \mathbb{R}$ be a function. We say that f vanishes at infinity if

$$\forall_{\varepsilon>0}\, \exists_{R>0}\, \forall_{x\in\mathbb{R}} : |x| \geq R \implies |f(x)| < \varepsilon. \tag{4.12}$$

Informally we write $f(\pm\infty) = 0$. Now let $C_0(\mathbb{R})$ be the set of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$ that vanish at infinity. In $C_0(\mathbb{R})$ we have the obvious definitions of addition and multiplication. Show that $C_0(\mathbb{R})$ is a complete normed algebra[13] with the (maximum-)norm well-defined by

$$|f|_{\max} = \max_{x\in\mathbb{R}} |f(x)|.$$

Hint: go through the programme for $C([a,b])$. The only new thing you have to do is show that the norm is well-defined.

Exercise 4.37. Let $F : \mathbb{R} \to \mathbb{R}$ be defined by

$$F(x) = \frac{x}{(1+x)^2},$$

and define $f : \mathbb{R} \to \mathbb{R}$ by $f_n(x) = F(nx)$. Show that

$$f(x) = \lim_{n\to\infty} f_n(x)$$

exists for every $x \in \mathbb{R}$. Is the convergence uniform on $\mathbb{R}$? And on $[0,\infty)$? And on $[0,1]$?

Exercise 4.38. Same question as in Exercise 4.37, but with

$$F(x) = \frac{|x|}{1+|x|}.$$

[13] A bit less like $\mathbb{R}$ since it does not contain a neutral element for multiplication.

Exercise 4.39. Same question as in Exercise 4.37 and Exercise 4.38, but with

$$f_n(x) = F\left(\frac{x}{n}\right).$$

Exercise 4.40. Suppose that $f_n : [0,1] \to [0,\infty)$ is a sequence of continuous functions nonincreasing in n with $f_n(x) \to 0$ for every $x \in [0,1]$ as $n \to \infty$, i.e.

$$\inf_{n\in\mathbb{N}} f_n(x) = 0. \tag{4.13}$$

Prove that

$$\max_{0\le x\le 1} f_n(x) \to 0$$

as $n \to \infty$. Hint: if not then there exists a sequence $x_n \in [0,1]$ such that $f_n(x_n) \not\to 0$. Let $\bar{x}$ be a limit point of this sequence and write

$$f_N(\bar{x}) = \underbrace{f_N(\bar{x}) - f_N(x_n)}_{\text{use continuity of } f_N} + \underbrace{f_N(x_n) - f_n(x_n)}_{\text{use } f_n \text{ nonincreasing}} + f_n(x_n)$$

to derive a contradiction with (4.13).

Exercise 4.41. Referring to Remark 4.18, prove that $B([a,b])$ is a complete metric space with the metric defined by $d(f,g) = |f-g|_\infty$.

5 Metric spaces and continuity

Recall that we wrote

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

for the distance between two number x and y in $\mathbb{R}$, and

$$d(f, g) = \max_{a \le x \le b} |f(x) - g(x)|$$

for the (uniform) distance between two functions f and g in $C([a, b])$. Henceforth we shall call such d, which assigns to every pair of elements of a set X (here $X = \mathbb{R}$ or $X = C([a, b])$) a number in $\mathbb{R}$, a *metric* if it has the following three properties:

$$d(x, x) = 0 \quad \text{for all} \quad x \in X; \tag{5.1}$$

$$d(x, y) = d(y, x) > 0 \quad \text{for all} \quad x, y \in X \quad \text{with} \quad x \neq y; \tag{5.2}$$

$$d(x, y) \le d(x, z) + d(z, y) \quad \text{for all} \quad x, y, z \in X. \tag{5.3}$$

It is time to introduce the abstract notion of a *metric space*[1].

Definition 5.1. *Let X be a nonempty set. A function*

$$d : X \times X \to \mathbb{R}$$

is called a metric *if the properties (5.1),(5.2),(5.3) hold. The set X is then called a* metric space *with metric d. The number $d(x, y)$ is commonly called the distance from x to y.*

In particular $X = \mathbb{R}$ is an example of a metric space, its metric defined by (2.9). Every nonempty subset A of a metric space X is also a metric space, with its metric inherited from the metric on X. And in Chapter 4 we encountered the example $C([a, b])$ with the uniform distance as metric. In Linear Algebra you must have seen normed vector spaces.

Exercise 5.2. Think about other examples. Subsets of $\mathbb{R}^2$ with the Pythagorean distance[2]. Point sets with a metric taking only the values 0 and 1. The unit sphere in $\mathbb{R}^3$ with the length of the shortest path connecting two points. Another example of a metric you have seen is the distance between nodes in a network or in a graph.

[1] Forgetting about $\mathbb{R}$ and its algebra for now.

[2] Illustrate the triangle inequality with a picture of a triangle in this case!

Have a look at Exercise 4.11 to extrapolate some terminology to the general case. The metric d is called a *strictly positive symmetric function*, because axiom[3] (5.2) says that $d(x,y) = d(y,x) > 0$ for $x \neq y$. Axiom (5.3), the *triangle inequality*, was already hinted at in Exercise 2.14, in the absence of triangles. The first axiom (5.1) stands by itself in its assignment that $d(x,x) = 0$ for all $x \in X$. Let's play with the axioms before we go on.

Remark 5.3. *The axioms (5.1,5.2,5.3) may be replaced by the axioms*

$$d(x,y) = 0 \iff x = y$$

and

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

for all $x, y, z \in X$. The nonnegativity follows when combining symmetry and the triangle inequality. See Exercise 2.55.

Many of the theorems we proved for $\mathbb{R}$ have counterparts in general metric spaces X, and also hold for $X = C([a,b])$ and $X = C_0(\mathbb{R})$ from Exercise 4.36 for instance. We simply replace absolute values $|x-y|$ by distances $d(x,y)$ in the definitions, theorems and proofs. The Banach Contraction Theorem is a nice example. The formulation and proof of Theorem 3.16 lead to the statement and proof of essentially the same theorem, for which we only have to adapt two basic definitions.

Definition 5.4. *A sequence x_n in a metric space X is a* Cauchy sequence *if*

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{m,n\geq N} \,:\, d(x_n, x_m) < \varepsilon,$$

and convergent *if*

$$\exists_{\bar{x}\in X}\, \forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N} \,:\, d(x_n, \bar{x}) < \varepsilon.$$

The metric space X is called complete *if every Cauchy sequence in X is convergent[4]. If such a complete metric space X happened to be a normed (vector) space and $d(x,y) = |x-y|$ then X is called a* Banach space. *In particular $\mathbb{R}$ is a Banach space[5].*

Exercise 5.5. Explain again why $\mathbb{R}$ is complete with $d(x,y) = |x-y|$, and that so is every closed subset of $\mathbb{R}$.

[3] An axiom is a property that we assume.

[4] With limit $\bar{x}$ in X, because there's nothing outside X here.

[5] The completeness assumption is in fact equivalent to the statement in Theorem 2.5.

Exercise 5.6. Explain why $C([a,b])$ is complete with the metric defined by

$$d(f,g) = |f-g|_{\max}.$$

5.1 The Banach Contraction Theorem

In view of Definition 5.4 it is now copy/paste from Theorem 3.16 with A and $\mathbb{R}$ both replaced by X to get the main result of this section.

Theorem 5.7. *Let X be a complete metric space and let $f : X \to X$ be a contraction, i.e.*

$$\exists_{\theta\in(0,1)}\, \forall_{x,y\in X} :\ d(f(x),f(y)) \le \theta\, d(x,y).$$

Then f has a unique fixed point, i.e a solution $\bar{x} \in X$ of $f(x) = x$. For every $x_0 \in X$, this $\bar{x}$ is the limit of the sequence x_n defined by $x_n = f(x_{n-1})$.

Proof. Could be an exercise now, but let's do it anyhow. Note that differences $x_n - x_m$ have meaning nor part in the formulation of Theorem 5.7, so the proof of Theorem 3.16 cannot be copy-pasted as it is. Still, the proof remains largely the same, with small changes making the proof more transparent perhaps. Here we go.

Consider a sequence as defined in the theorem by $x_n = f(x_{n-1})$ and let $m > n$. Before we bring in the arbitrary $\varepsilon > 0$ we observe that

$$d(x_1,x_2) \le \theta d(x_0,x_1), \quad d(x_2,x_3) \le \theta d(x_1,x_2) \le \theta^2 d(x_0,x_1),$$

$$d(x_3,x_4) \le \theta d(x_2,x_3) \le \theta^3 d(x_0,x_1), \quad d(x_4,x_{4+1}) \le \theta^4 d(x_0,x_1),$$

and so on. Replacing 4 by n in the last inequality we have

$$d(x_n,x_{n+1}) \le \theta^n d(x_0,x_1), \tag{5.4}$$

which holds[6] for all $n \in \mathbb{N}$. Now assume that x_0 is not a fixed point of f. By repeated use of the triangle inequality we then get[7] for $m > n \ge N$, N waiting for $\varepsilon > 0$ to show up. Here it is.

[6] By induction if you insist.

[7] As in Exercise 3.18.

$$d(x_n,x_m) \le d(x_n,x_{n+1}) + d(x_{n+1},x_m) \le d(x_n,x_{n+1}) + \cdots + d(x_{m-1},x_m)$$

$$\le (\theta^n + \cdots + \theta^{m-1})\, d(x_0,x_1) < \frac{\theta^n}{1-\theta}\, d(x_0,x_1) \le \frac{\theta^N}{1-\theta}\, d(x_0,x_1)$$

Let $\varepsilon > 0$. Choose N so large that

$$\frac{\theta^N}{1-\theta}\, d(x_0, x_1) < \varepsilon.$$

This is possible in view of Exercise 3.17. For all $m > n \geq N$ it then holds that

$$d(x_n, x_m) < \varepsilon.$$

We have thus proved that x_n is a Cauchy sequence because $\varepsilon > 0$ was arbitrary.

Since X is complete the sequence x_n is convergent[8]. Denote its limit by $\bar{x}$ and introduce x_n as before in (3.11) by means of the triangle inequality. This yields

$$d(\bar{x}, f(\bar{x})) \leq d(\bar{x}, x_{n+1}) + d(x_{n+1}, f(\bar{x})) = d(\bar{x}, x_{n+1}) + d(f(x_n), f(\bar{x}))$$

$$\leq d(\bar{x}, x_{n+1}) + \theta d(x_n, \bar{x}) < (1+\theta)\varepsilon$$

for all $n \geq N$, the N that comes with ε in the statement that $x_n \to \bar{x}$. As in the proof of Theorem 3.16 it follows that $d(\bar{x}, f(\bar{x})) = 0$ whence $\bar{x} = f(\bar{x})$. Another solution $\tilde{x}$ of $x = f(x)$ cannot exist, because we would then have

$$0 < d(\bar{x}, \tilde{x}) = d(f(\bar{x}), f(\tilde{x})) \leq \theta\, d(\bar{x}, \tilde{x}) < d(\bar{x}, \tilde{x}),$$

a contradiction. This completes a clean proof without algebra. □

Theorem 5.7 is often applied to subsets of complete metric spaces. This requires such a subset to be complete by itself. To characterise this property a version of Definition 3.14 with $\mathbb{R}$ replaced by X is needed.

Definition 5.8. *A subset A of a metric space X is called* closed *X if the limit $\bar{x}$ of a convergent sequence x_n is in A whenever all x_n are in A.*

This terminology was already best explained in Section 3.6, a section which can be copy-pasted here with $\mathbb{R}$ replaced by X, with in Remark 3.26: a subset A of a metric space X is closed if by taking limits of sequences contained in A you cannot get out of A. The repair for subsets A of X flawing this property was not yet formulated[9]:

Theorem 5.9. *Let A be a subset of a complete metric space X, and let $\bar{A}$ be the set of all limits of all convergent sequences[10] x_n with $x_n \in A$. Then $\bar{A}$ is the smallest closed subset of X which contains A, and $\bar{A}$ is called the* closure *of A.*

[8] Of course the same conclusion trivially holds if x_0 is a fixed point of f.

[9] Which you should compare to constructions of $\mathbb{R}$ out of the rational numbers.

[10] Including sequences $a, a, a, a, \ldots$ with $a \in A$.

Exercise 5.10. Prove Theorem 5.9. Hint: first show that $\bar{A}$ is closed, then show that there is no closed subset $\tilde{A}$ with $A \subset \tilde{A} \subset \bar{A}$ and $\tilde{A} \neq \bar{A}$.

Theorem 5.11. *Let X be a complete metric space and $A \subset X$. Then A is by itself a complete metric space if and only if A is closed.*

Exercise 5.12. Prove Theorem 5.11.

5.2 More of the same: continuity in metric spaces

Definition 4.1 used converging sequences to formulate the concept of continuity in a given point $\xi \in A \subset \mathbb{R}$ for a function $f : A \to \mathbb{R}$. We copy-paste it with the first and the second $\mathbb{R}$ replaced by X and Y.

Definition 5.13. *Let X, Y be a metric spaces, $A \subset X$ nonempty, $f : A \to Y$ and $\xi \in A$. Then f is called* continuous *in ξ if $f(x_n) \to f(\xi)$ for every sequence x_n in A with $x_n \to \xi$. If f is continuous in every $\xi \in A$ then $f : A \to Y$ is called* continuous.

Exercise 5.14. Let X, Y, Z be metric spaces, $f : X \to Y$ continuous in $a \in X$, $g : Y \to Z$ continuous in $f(a)$. Prove that $g \circ f$ is continuous in a. Conclude for $A = [0, \infty) \subset \mathbb{R}$, $f : A \to \mathbb{R}$ continuous, X a metric space, and $\xi \in X$ that $F : X \to \mathbb{R}$ defined by $F(x) = f(d(x, \xi))$ is continuous.

Remark 5.15. *Let X be a metric space, and let $f : X \to \mathbb{R}$ be continuous in every point of X. The proof of Theorem 4.4 can be copy-pasted with A replaced by X, provided X has the property that every sequence x_n in X has a limit point, i.e. if*[11]

$$\exists_{\bar{x} \in \mathbb{R}} \forall_{\varepsilon > 0} \forall_{N \in \mathbb{N}} \exists_{n \geq N} \, : \, d(x_n, \bar{x}) < \varepsilon. \tag{5.5}$$

Such metric spaces are called (sequentially) compact[12]*. This leads to:*

Theorem 5.16. *Let X be a sequentially compact metric space, i.e. every sequence in X has a convergent subsequence. If $f : X \to \mathbb{R}$ is continuous in*

[11] See the reformulation of the convergent subsequence property in Remark 3.23.
[12] See Exercise 4.33.

every point of X then f has a global maximum and a global minimum. The real number

$$|f|_{max} = \max_{x \in X} |f(x)| \tag{5.6}$$

is thus well-defined and called the maximum norm *of f.*

Exercise 5.17. Not done myself yet, but let X be a metric space which contains a sequence without limit points. Can you construct a continuous function on X which is unbounded? Hint: use Exercise 5.14 and the negation of (5.5) as a starting point for your imagination.

Remark 5.18. *We obtained $f(\bar{x}) = \bar{x}$ from $d(x_n, x) \to 0$ and the contraction property of f, which was a special stronger case of Lipschitz continuity, see Definition 3.3. For maps between metric spaces the definition is given below.*

Definition 5.19. *Let X and Y be metric spaces with metrics $d_{_X}$ for X and $d_{_Y}$ for Y. A map $f : X \to Y$ is called* Lipschitz continuous *with Lipschitz constant $L > 0$ if for all $x, y \in X$ it holds that*

$$d_{_Y}(f(x), f(y)) \leq L\, d_{_X}(x, y). \tag{5.7}$$

Examples[13] are $Y = X$, with

$$d(f(x), f(y)) \leq L\, d(x, y),$$

and $Y = \mathbb{R}$, with

$$|f(x) - f(y)| \leq L\, d(x, y).$$

Exercise 5.20. Prove that Lipschitz continuity implies pointwise continuity.

5.3 Outlook: topology

There's more to be copy-pasted from Section 3.6 with $\mathbb{R}$ replaced by X.

Definition 5.21. *Let X be metric space with metric d. A subset $\mathcal{O}$ of X is called* open in X *if*

$$\forall_{\xi \in \mathcal{O}}\, \exists_{r>0} : B_r(\xi) = \{x \in X :\, d(x, \xi) < r\} \subset \mathcal{O}.$$

The set $B_r(\xi)$ is called[14] *an* open ball *centered at ξ with radius $r > 0$.*

[13] Not to bore you with the example in Definition 3.3.

[14] Whatever meaning these words may have.

Exercise 5.22. Prove that the set $B_r(\xi)$ in Definition 5.21 is open.

Exercise 5.23. Prove that arbitrary unions of open subsets of a metric space X are open. Prove that the intersection of two open subsets of X is also open. Prove that X is open in itself. Prove that the empty subset $\emptyset$ of X is open.

Remark 5.24. *If we denote the collection of all open subsets of X by $\mathcal{T}$, then Exercise 5.23 says that*

$$\emptyset \in \mathcal{T},\, X \in \mathcal{T},$$

$$A, B \in \mathcal{T} \implies A \cap B \in \mathcal{T},$$

$$\forall_{i\in I} :\, A_i \in \mathcal{T} \implies \cup_{i\in I} A_i \in \mathcal{T}.$$

A collection $\mathcal{T}$ of subsets of a given set X with these properties is called a topology *on X. Thus every metric on X defines a topology on X, consisting of the open sets as defined in Definition 5.21.*

Theorem 5.25. *Let X, Y be metric spaces and $f : X \to Y$ a map. Then f is continuous in every point of X if and only if the inverse image*

$$f^{-1}(\mathcal{O}) = \{x \in X :\, f(x) \in \mathcal{O}\}$$

of $\mathcal{O}$ under f is open in X for every set $\mathcal{O} \subset Y$ that is open in Y.

Proof. Assume that f is continuous, i.e.

$$x_n \to \xi \implies f(x_n) \to f(\xi)$$

for every $\xi \in X$ and let $\mathcal{O} \subset Y$ be open in Y. To show that $f^{-1}(\mathcal{O})$ is open take $\xi \in X$ with $f(\xi) \in \mathcal{O}$. Suppose there is no $r > 0$ such that $B_r(\xi) \subset f^{-1}(\mathcal{O})$. Then we can choose[15] a sequence x_n in X such that $x_n \to \xi$ while $f(x_n) \notin \mathcal{O}$. By definition of continuity $f(x_n) \to f(\xi) \in \mathcal{O}$.

Choose $\varepsilon > 0$ such that

$$B_\varepsilon(f(\xi)) = \{y \in Y :\, d_{_Y}(y, f(\xi)) < \varepsilon\} \subset \mathcal{O}$$

and apply the definition of $f(x_n) \to f(\xi)$. Then there exists $N \in \mathbb{N}$ such that $f(x_n) \in B_\varepsilon(f(\xi)) \subset \mathcal{O}$ for all $n \geq N$, a contradiction. Thus there does

[15]The reasoning is similar to that in the proof of Theorem 3.28.

exist $r > 0$ such that $B_r(\xi) \subset f^{-1}(\mathcal{O})$. This holds for every $\xi \in f^{-1}(\mathcal{O})$. We have thus proved that $f^{-1}(\mathcal{O})$ is open.

For the opposite implication, assume that $f^{-1}(\mathcal{O})$ is open in X for every $\mathcal{O}$ open in Y, and let x_n be a convergent sequence with limit ξ. We have to prove that $f(x_n) \to f(\xi)$. We follow our nose. Let $\varepsilon > 0$ and consider the open ball $B_\varepsilon(f(\xi))$. By assumption its pre-image $f^{-1}(B_\varepsilon(f(\xi)))$ is open in X and contains ξ. Therefore there exists $r > 0$, but let's call it δ, such that

$$B_\delta(\xi)) \subset f^{-1}(B_\varepsilon(f(\xi))).$$

This is equivalent to

$$f(B_\delta(\xi) \subset B_\varepsilon(f(\xi)),$$

and says that

$$d_{_X}(x, \xi) < \delta \implies d_{_Y}(f(x), f(\xi)) < \varepsilon. \tag{5.8}$$

To finish we should not forget the sequence x_n we started with, and its limit ξ. Apply the definition of convergence in the form

$$\exists_{N \in \mathbb{N}} \forall_{n \geq N} \, : \, d(x_n, \xi) < \delta.$$

Then $d_{_Y}(f(x_n), f(\xi)) < \varepsilon$ for all $n \geq N$. This shows that $f(x_n) \to f(\xi)$ and completes the proof. □

Remark 5.26. *The reformulation of continuity in every point in terms of open sets given in Theorem 5.25 involved the first ε-δ-statement (5.8) in these lecture notes. Such statements reappear when we come to integrals next.*

5.4 Exercises

Exercise 5.27. Let x_n be a sequence in a metric space X and $\bar{x} \in X$. Prove that $x_n \to \bar{x}$ if and only if every subsequence of x_n has itself a subsequence that converges to $\bar{x}$. Hint: reason as in Exercise 4.20.

Exercise 5.28. Let X and Y be metric spaces, and $f : X \to Y$. Prove that f is continuous if and only if $f^{-1}(G) = \{x \in X : \, f(x) \in G\}$ is closed in X for every set G closed in Y.

Exercise 5.29. Let $X = C([0, 1])$ and let $g : \mathbb{R} \to \mathbb{R}$ be continuous. For $f \in X$ define

$$F(f) = g \circ f, \text{ i.e. } \quad (F(f))(x) = g(f(x)) \quad \forall x \in [0, 1].$$

Prove that $F(f) \in X$ and that $F : X \to X$ is continuous. What do you have to assume about g to ensure that F is Lipschitz continuous? Discuss the examples in which g is defined[16] by

$$g(y) = y^2 \quad \text{and} \quad g(y) = \frac{y}{1+y^2}.$$

Exercise 5.30. Let $X = C([0,1])$. Define $F : X \to \mathbb{R}$ by $F(f) = f(0) + f(1)^2$. Prove directly from Definition 4.1 and Definition 4.8 that F is continuous.

Exercise 5.31. Let $X = C([0,1])$. Define $F : X \to X$ by

$$(F(f))(x) = 1 + \frac{1}{2} f\left(\frac{x}{2}\right).$$

Prove that F is a contraction. What is the contraction factor of F? What is the unique fixed point of F?

Exercise 5.32. For $x = (x_1, x_2) \in \mathbb{R}^2$ define $|x|_{\max} = \max(|x_1, |x_2|)$. Prove that this defines a norm and thereby a metric. Show that the topology defined by this metric is the same as the topology defined by the Euclidean distance. Hint: roll in some balls first and draw them in the x_1x_2-plane.

Exercise 5.33. (continued) Same question for $|x|_1 = |x_1| + |x_2|$.

Exercise 5.34. (continued) The Euclidean distance derives from the norm defined by $|x|_2 = \sqrt{x_1^2 + x_2^2}$. Hint: prove the triangle inequality[17].

Exercise 5.35. An alternative way to say that $O \in \mathbb{R}^2$ is open is to demand that for every $\xi \in O$ it holds that[18]

$$\xi \in K_1 \cap K_2 \cap K_3 \subset O,$$

[16] See Exercise 3.52.

[17] No pictures allowed in the proof.

[18] The number of halfspaces needed is $3 = 2 + 1$, the dimension of $\mathbb{R}^2$ plus 1.

with K_1, K_2, K_3 open half planes. An open half plane is a set of the form

$$K = \{x \in \mathbb{R}^2 : a_1 x_1 + a_2 x_2 < b\}$$

with $a_1, a_2, b \in \mathbb{R}$ and a_1, a_2 not both equal to zero. Prove this statement.

Exercise 5.36. Referring to Exercise 2.55, let X be a set, let $d : X \times X \to \mathbb{R}$ satisfy

$$d(y,x) = d(x,y) \leq d(x,z) + d(z,y) \quad \text{for all} \quad x, y, z \in X.$$

Let $f : X \to X$ have the property that

$$\exists_{\theta \in (0,1)} \, \forall_{x,y \in X} : d(f(x), f(y)) \leq \theta \, d(x,y),$$

and define for $x_0 \in X$ the sequence x_n by $x_n = f(x_{n-1})$. Prove that

$$0 \leq d(x_n, x_{n+1}) \leq \theta^n d(x_0, x_1)$$

for every $n \in \mathbb{N}$. Then prove that

$$\forall_{\varepsilon > 0} \, \exists_{N \in \mathbb{N}} \, \forall_{m,n \geq N} : d(x_n, x_m) < \varepsilon,$$

which is *not* the definition of x being a Cauchy sequence: we have not assumed that d is a metric.

Exercise 5.37. Suppose that the sequence x_n defined in Exercise 5.36 has the property that $d(x_n, \bar{x}) \to 0$, for some $\bar{x} \in X$. Under the same conditions as in Exercise 5.36, prove that $d(\bar{x}, f(\bar{x})) = 0$.

Exercise 5.38. Suppose y_n is another sequence as in Exercise 5.37, defined by $y_n = f(y_{n-1})$ as in Exercise 5.36 for some $y_0 \in X$, with the property that $d(y_n, \bar{y}) \to 0$ and $d(\bar{y}, y_n) \to 0$, for some $\bar{y} \in X$. Prove that $d(\bar{x}, \bar{y}) = d(\bar{y}, \bar{x}) = 0$.

Exercise 5.39. Let X be a vector space over $\mathbb{R}$ with a norm, i.e. for every x in X there is defined a real number $|x|_X \geq 0$ such that

$$|x|_X = 0 \iff x = 0; \quad |\lambda x|_X = |\lambda| \, |x|_X; \quad |x + y|_X \leq |x|_X + |y|_X$$

for all $x, y \in X$ and all $\lambda \in \mathbb{R}$. Suppose that for some sequence x_n in X it holds that $S_n = x_1 + \cdots + x_n$ is a convergent sequence with limit $S \in X$, and also that

$$\sum_{n=1}^{\infty} |x_n|_X < \infty. \tag{5.9}$$

Prove that

$$|S|_X \leq \sum_{n=1}^{\infty} |x_n|_X.$$

Exercise 5.40. (continued) Prove that X is complete as a metric space with the norm defined by $d(x, y) = |x - y|_X$ if (5.9) implies that the sequence S_n is convergent. Also formulate and prove the converse of this stament: if x_n is a sequence in a Banach space X which satisfies (5.9), then the sequence S_n is convergent.

Exercise 5.41. Suppose that X is a compact metric space, $f_n \in C(X)$ for $n \in \mathbb{N}$, $f \in C(X)$, and $f_n(x) \to f(x)$ for every $x \in X$ as $n \to \infty$. Assume that $f_n(x)$ is a nonincreasing in n for every $x \in X$. Prove that $f_n \to f$ in $C(X)$, i.e. $f_n \to f$ uniformly on X (Dini's theorem). Hint: Exercise 4.40.

Exercise 5.42. Almost forgot: prove that every compact metric space is complete.

5.5 Compactness with open coverings

Definition 5.43. *Let X be a metric space and $A \subset X$. A collection*

$$\{O_i : \ i \in I\},$$

in which I is an index set and O_i is an open subset of X for every $i \in I$, is called an open covering of A if

$$A \subset \cup_{i \in I} O_i.$$

Theorem 5.44. *Let $A \subset X$ be sequentially compact, i.e. every sequence x_n in A has a limit point in A. Then for every open covering $\{O_i : \ i \in I\}$ of A there exist $i_1, \ldots, i_m \in I$ such that*

$$A \subset O_{i_1} \cup \cdots \cup O_{i_m},$$

and $\{O_{i_1}, \ldots, O_{i_m}\}$ is callled a finite subcovering.

Proof. We first assume that $I = \mathbb{N}$ and

$$A \subset \cup_{i\in\mathbb{N}} O_i.$$

If the statement were false then for every $n \in \mathbb{N}$ there would be a $p_n \in A$ with

$$p_n \notin O_1 \cup \cdots \cup O_n. \tag{5.10}$$

Since A is sequentially compact the sequence p_n has a limit point p in A, and p must be contained in some O_m. But O_m is open so there exists an open ball $B_\varepsilon(p) \subset O_m$. Then it must be that $p_n \in B_\varepsilon(p)$ for some $n \geq m$, otherwise p is not a limit point. This contradicts (5.10) because then

$$B_\varepsilon(p) \subset O_m \subset O_1 \cup \cdots \cup O_n.$$

So for general I we only have to show that there exists a sequence i_n such that

$$A \subset \cup_{n\in I} O_{i_n}.$$

We now first assume that A is *separable*, i.e. that there exists a sequence p_n in A such that every p in A is a limit point of this sequence. We claim[19] that thereby

$$p \in B_{\frac{1}{m}}(p_n) \subset O_i$$

for some $i \in I$ and some $m, n \in \mathbb{N}$. If so then the pairs (m, n) thus encountered by varying $p \in A$ form a countable set J and

$$A \subset \cup_{(m,n)\in J} B_{\frac{1}{m}}(p_n).$$

For each such (m, n) choose $i = i_{mn} \in I$ such that $B_{\frac{1}{m}}(p_n) \subset O_i$ as above. Then

$$\cup_{m,n\in\mathbb{N}} O_{i_{mn}}$$

a countable open cover of A.

It now remains to show that A is separable. For subsets of separable metric spaces X this is always true, but requires an argument we leave for now. Instead we show that sequentially compact sets are totally bounded, i.e. for every $\varepsilon > 0$ there are finitely many $p_1, \ldots, p_n$ in A such that

$$A \subset B_\varepsilon(p_1) \cup B_\varepsilon(p_2) \cup \cdots \cup B_\varepsilon(p_n).$$

Clearly this implies that A is separable.

[19]Prove this claim.

So suppose A is sequentially compact but not totally bounded. Then there exists $\varepsilon > 0$ for which no $p_1, \ldots, p_n$ as above exist. Choose $p_1 \in A$ and inductively for $n = 1, 2, \ldots$ a point $p_{n+1} \in A$ with

$$p_{n+1} \notin B_\varepsilon(p_1) \cup B_\varepsilon(p_2) \cup \cdots \cup B_\varepsilon(p_n).$$

Then $d(p_i, p_j) \geq \varepsilon$ for all $i \neq j$, so the sequence p_n can not have a convergent subsequence. This completes the proof. □

Theorem 5.45. *Let $A \subset X$ have the property that every open covering of A has a finite subcovering. Then A is sequentially compact.*

Proof. Let a_n be a sequence in A and suppose it has no convergent subsequence. Then for every $p \in A$ there must be and $\varepsilon_p > 0$ and $N_p \in \mathbb{N}$ such that $a_n \notin B_{\varepsilon_p}(p)$ for all $n > N_p$. Clearly $\{B_{\varepsilon_p}(p) : \ p \in A\}$ is an open covering of A, so there exists $p_1, p_2, \ldots, p_m$ in A such that

$$A \subset B_{\varepsilon_{p_1}}(p_1) \cup B_{\varepsilon_{p_2}}(p_2) \cup \cdots \cup B_{\varepsilon_{p_m}}(p_m).$$

Thus A contains at most finitely elements of the sequence a_n, so at least on element a_n of the sequence occurs infinitely many times in the sequence, say for $n = n_k$, with $n_1 < n_2 < \cdots$. This makes a_{n_k} a trivially convergent subsequence, a contradiction that completes the proof. □

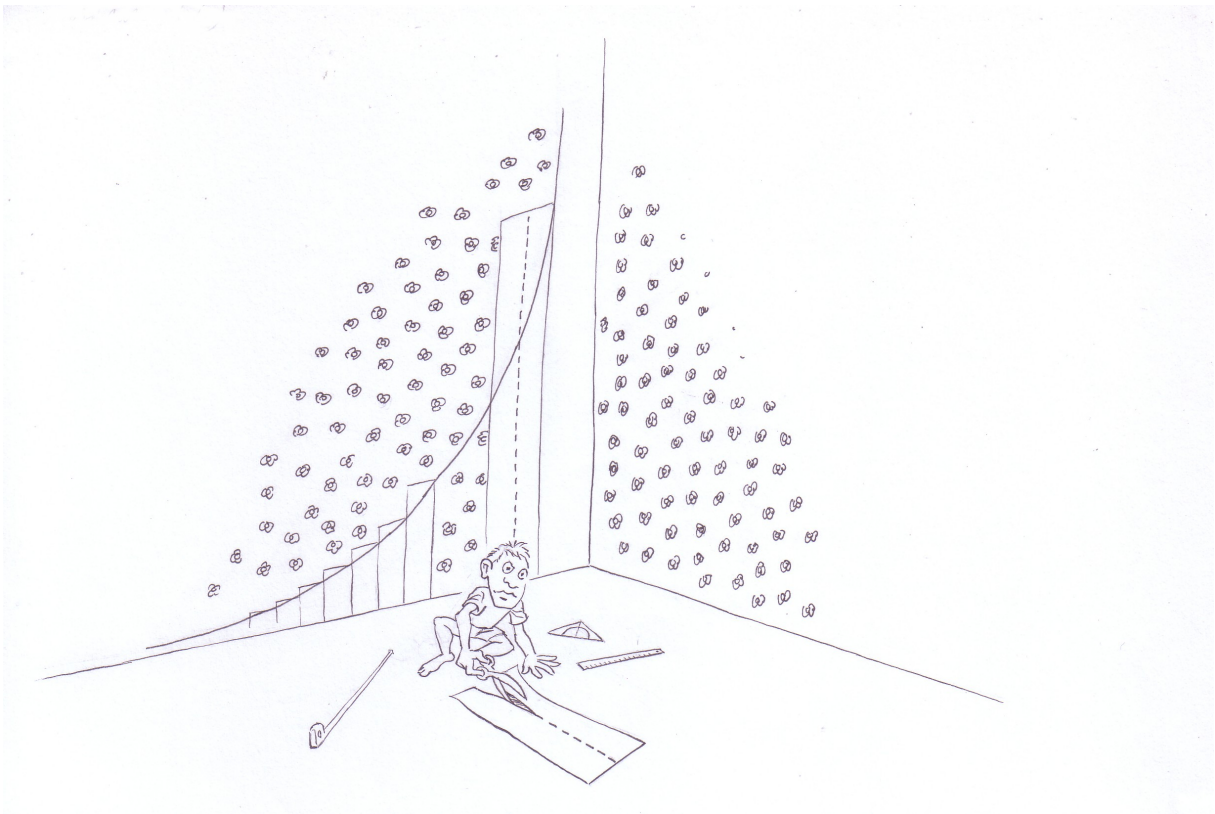
The story so far:

"In the beginning the Universe was created. This has made a lot of people very angry and been widely regarded as a bad move. Many races believe that it was created by some sort of God, though the Jatravartid people of Viltvodle VI believe that the entire Universe was in fact sneezed out of the nose of a being called the Great Green Arkleseizure. The Jatravartids, who live in perpetual fear of the time they call The Coming of The Great White Handkerchief, are small blue creatures with more than fifty arms each, who are therefore unique in being the only race in history to have invented the aerosol deodorant before the wheel. However, the Great Green Arkleseizure Theory is not widely accepted outside Viltvodle VI and so, the Universe being the puzzling place it is, other explanations are constantly being sought." (Douglas Adams)

Part 1 was about what we learned from YBC7289 and Archimedes: everything about limits and limit points of sequences, the Banach Contraction Theorem (BCT), $C([a,b])$ as a complete metric space in which BCT thereby holds, its metric defined by $d(f,g) = |f-g|_\infty$, the maximum norm well-defined for every $f \in C([a,b])$ by

$$|f|_{\max} = \max_{a \le x \le b} |f(x)|,$$

and showing off with the statement that $C([a,b])$ is in fact a Banach algebra. We continue with Part 2, revisit Archimedes and the pyramids, to first study integrals of functions $f : [a,b] \to \mathbb{R}$, ignoring functions in $C([a,b])$ while we can.



6 Integration of monotone functions

Let us slow down the pace. This chapter is meant to be largely independent of what we've done[1] since Archimedes and the pyramids in Sections 1.2 and 1.3. Let $a, b \in \mathbb{R}$ and let $f : [a, b] \to \mathbb{R}$ be a nice function, nice in a meaning to be made precise later. Consider the sets

$$A_+ = \{(x, y) \in \mathbb{R}^2 : 0 < y < f(x),\, a < x < b\}$$

and

$$A_- = \{(x, y) \in \mathbb{R}^2 : f(x) < y < 0,\, a < x < b\}.$$

If both these sets have a well-defined finite area, denoted by $|A_+|$ and $|A_-|$, then based on what you have seen in highschool you would expect that the integral of f from a to b is given by

$$\int_a^b f(x)\,dx = |A_+| - |A_-|.$$

Exercise 6.1. Sketch the graph of the function $f : [0, 1] \to \mathbb{R}$ defined by

$$f(x) = x(1 - x)(x - \frac{1}{3})$$

and indicate the two sets A_- and A_+.

Here we will *not bother to define the area of general subsets* of the plane, but we opt for a definition of the integral only. The definition should not make you uncomfortable in relation to what your intuition says that the area of the sets A_+ and A_- should be.

6.1 Integrals of monomials

Have a look at (1.5) in Section 1.2 and the work you did in Exercise 1.17. You probably convinced yourself that

$$J_p := \int_0^1 x^p\,dx = \frac{1}{p+1}$$

for every $p \geq 2$. But it is also instructive to look at the easy cases $p = 0$ and $p = 1$ first. Starting point for the definition of the integral is the consensus that the area of the open square

$$S = \{(x, y) \in \mathbb{R}^2 : 0 < x < 1,\, 0 < y < 1\}$$

[1] https://www.youtube.com/watch?v=2vcvh2K9wIk

is equal to 1, and that

$$\int_0^1 x^0\,dx = \int_0^1 1\,dx = |S| = 1.$$

So for $p = 0$ all is clear[2].

Next we consider $p = 1$. Let the function f be defined by $f(x) = x$. Again we have $A_- = \emptyset$, but now the set A_+ is an open triangle. The interior of $S\backslash A_+$ is also an open triangle, twinned to A_+ by reflection in the line $y = x$. We therefore conclude that the area of A_+ must be equal to half of the area of S, i.e.

$$\int_0^1 x\,dx = |A_+| = \frac{|S|}{2} = \frac{1}{2}$$

must be the outcome for any reasonable definition of the integral.

But for $p = 2, 3, 4, \ldots$ there is no such symmetry argument. Thus the example

$$f(x) = x^2$$

requires a new approach. We will first look for a sensible meaning of

$$J_2 = \int_0^1 x^2\,dx$$

that coincides with what we believe is the area of

$$A_2 = \{(x, y) \in \mathbb{R}^2 : 0 < y < x^2 < 1\}.$$

The idea now is to evaluate $y = x^2$ at values of x given by

$$0 = \frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n} = 1,$$

These particular x-values give you points (x, y) in the unit square S.

Exercise 6.2. Choose $n = 10$. Look at the set A_2 in S bounded by $y = 0$, $x = 1$ and $y = x^2$. Make a sketch in which S is large (so that there's not much outside of S) to convince yourself that the area $|A_2|$ of A_2 is less than the *upper sum*

$$\frac{1}{10}\left(\frac{1}{100} + \frac{4}{100} + \frac{9}{100} + \frac{16}{100} + \frac{25}{100} + \frac{36}{100} + \frac{49}{100} + \frac{64}{100} + \frac{81}{100} + \frac{100}{100}\right),$$

but more than the *lower sum*

$$\frac{1}{10}\left(\frac{0}{100} + \frac{1}{100} + \frac{4}{100} + \frac{9}{100} + \frac{16}{100} + \frac{25}{100} + \frac{36}{100} + \frac{49}{100} + \frac{64}{100} + \frac{81}{100}\right).$$

[2]Don't bother about 0^0 in $x = 0$ yet but note that we also agree that $|\bar{S}| = 1$.

Hint: look at the cartoon preceding this chapter. Every nonzero term in the two sums is the area of a rectangle with width $\frac{1}{10}$ in your sketch.

If this worked out, you will also convince yourself that

$$|A_2| < \frac{1}{n^3} \sum_{k=1}^{n} k^2 \tag{6.1}$$

for every natural number n. Now recall[3] from (C_n) in Section 1.2 that

$$\sum_{k=1}^{n} k^2 = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6},$$

and enjoy the cubic version

`https://twitter.com/i/status/1116738152935374853`

from another perspective if you like. Together with (6.1) the sum of the first n squares formula implies that

$$|A_2| < \frac{1}{n^3} \sum_{k=1}^{n} k^2 = \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2} < \frac{1}{3} + \frac{2}{3n}. \tag{6.2}$$

Likewise you will conclude that

$$|A_2| > \frac{1}{n^3} \sum_{k=0}^{n-1} k^2 = \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2} - \frac{1}{n} = \frac{1}{3} - \frac{1}{2n} + \frac{1}{6n^2} > \frac{1}{3} - \frac{1}{2n}. \tag{6.3}$$

Thus the area $|A_2|$ should satisfy

$$\frac{1}{3} - \frac{1}{2n} < |A_2| < \frac{1}{3} + \frac{2}{3n} \quad \text{for all} \quad n \in \mathbb{N}. \tag{6.4}$$

This squeezes the area in, and allows for no other conclusion than[4]

$$J_2 = |A_2| = \frac{1}{3},$$

the very same number that we found for the volume of the pyramid in Section 1.2.

[3] Proved (C_n) with the domino principle, click on the link for $n = 3$.

[4] Note the same reasoning applies to $\bar{A}_2$.

Exercise 6.3. In Exercise 1.17 of Chapter 1 we established that[5]

$$\sum_{k=0}^{n-1} k^p < \frac{n^{p+1}}{p+1} < \sum_{k=1}^{n} k^p$$

for all $p, n \in \mathbb{N}$. Convince yourself that for all $p \in \mathbb{N}$ it must therefore hold that

$$|A_p| = \frac{1}{p+1}.$$

Hint: use lower and upper sums.

Remark 6.4. *For $p = 1$ an approach with lower and upper sums may look a bit silly. But it does reproduce the right number for the area of the triangle A_1. Our new calculation for $J_2 = |A_2|$ is identical to the calculation of the volume of the pyramid in Section 1.2.*

6.2 Integrals of monotone functions via finite sums

In the previous section we have hopefully convinced you that a proper definition of the integral leads to

$$\int_0^1 x^p \, dx = \frac{1}{p+1}. \tag{6.5}$$

Now let $a, b \in \mathbb{R}$ with $a < b$. A definition of

$$J = \int_a^b f \; = \int_a^b f(x) \, dx \tag{6.6}$$

will now be designed for a large class of functions $f : [a, b] \to \mathbb{R}$ so as to describe the area $|A|$ of the set

$$A = \{(x, y) \in \mathbb{R}^2 : 0 < y < f(x),\, a < x < b\} \tag{6.7}$$

if f has the property that $f(x) \geq 0$ for all $a < x < b$. For a start we take f to be nondecreasing and nonnegative, just like in (6.5).

Definition 6.5. *Let $a, b \in \mathbb{R}$ with $a < b$ and $f : [a, b] \to \mathbb{R}$. Then f is called* nonnegative *if $f(x) \geq 0$ for all $x \in [a, b]$; f is called* nondecreasing *if the implication*

$$x_1 \leq x_2 \implies f(x_1) \leq f(x_2)$$

holds for all $x_1, x_2 \in [a, b]$.

[5]For reasons of consistency with what is to come we let the first sum start with $k = 0$.

Such nonnegative nondecreasing functions can be pretty wild[6], but for the indicated approach with lower and upper sums we will now show that there are no problems in defining an integral.

Definition 6.6. *A partition P of $[a,b]$ is a choice of real numbers $x_0,\dots,x_N$ with*

$$a = x_0 \le x_1 \le \cdots \le x_N = b \qquad (N \ge 2). \tag{6.8}$$

Given such a partition P and a nondecreasing nonnegative $f : [a,b] \to \mathbb{R}$ we define the *left endpoint sums*[7]

$$L := \sum_{k=1}^{N} \underbrace{f(x_{k-1})}_{m_k}(x_k - x_{k-1}) \tag{6.9}$$

$$= f(x_0)(x_1 - x_0) + \cdots + f(x_{N-1})(x_N - x_{N-1}).$$

Each nonzero term in (6.9) is the area of an open rectangle[8]

$$(x_{k-1}, x_k) \times (0, f(x_{k-1})) = \{(x,y) \in \mathbb{R}^2 : \, 0 < y < f(x_{k-1}),\ x_{k-1} < x < x_k\}$$

contained in A. This follows because f is nondecreasing, so that x_{k-1} is a *minimizer* for f on $[x_{k-1}, x_k]$, that is

$$m_k = \min_{x \in I_k} f(x) = f(x_{k-1}), \quad \text{where} \quad I_k = [x_{k-1}, x_k] \tag{6.10}$$

for $k = 1,\dots,N$. These rectangles are mutually disjoint. Therefore the sum of their areas must be a lower bound for the area of A. We therefore agree that the left endpoint sum L is a *lower Riemann sum* for the integral (6.6) that we want to define. In other words, the number J satisfies

$$L \le J$$

if J exists.

In the same fashion the closed rectangles[9]

$$[x_{k-1}, x_k] \times [0, f(x_k)] = \{(x,y) \in \mathbb{R}^2 : \, 0 \le y \le f(x_k),\ x_{k-1} \le x \le x_k\},$$

with k running from 1 to N, cover A completely, because we recognize x_k as *maximizer* for f on I_k:

$$M_k = \max_{x \in I_k} f(x) = f(x_k). \tag{6.11}$$

[6] See Exercises 4.31 and 4.32.

[7] Make a sketch in which you see what these sums are.

[8] Possibly empty, if $x_{k-1} = x_k$ or $f(x_{k-1}) = 0$.

[9] Possibly reducing to line segments or points with zero area.

We thus say that the *right endpoint sum*

$$R = \sum_{k=1}^{N} \underbrace{f(x_k)}_{M_k}(x_k - x_{k-1}) \tag{6.12}$$

is an *upper Riemann sum* for the integral (6.6) that we want to define. In particular,

$$J \leq R$$

if J exists. We are ready to give a definition of integrability for nondecreasing functions.

Definition 6.7. *A nondecreasing*[10] *function* $f : [a, b] \to \mathbb{R}$ *is called* Riemann integrable *if there is a unique number* J *such that*

$$L \leq J \leq R \tag{6.13}$$

for all possible choices of the partition P*. This number* J *is then called the* integral *of* f *over* $[a, b]$ *and we write*

$$J = \int_{[a,b]} f = \int_a^b f = \int_a^b f(x)\, dx.$$

In the above notation x is a *dummy* variable, which may be replaced by any other symbol[11].

But now observe that for *equidistant partitions*, i.e. partitions

$$x_0 < x_1 < \cdots < x_N \quad \text{with} \quad x_k - x_{k-1} = \frac{b-a}{N},$$

the corresponding (left endpoint) lower and (right endpoint) upper sums, denoted by L_N and R_N, satisfy[12]

$$0 \leq R_N - L_N = \sum_{k=1}^{N} \left(f(x_k) - f(x_{k-1})\right) \frac{b-a}{N} = \left(f(b) - f(a)\right) \frac{b-a}{N}. \tag{6.14}$$

But here $N \in \mathbb{N}$ arbitrary! Archimedes thus tells us that there is *at most one number* J that can reasonably qualify as the integral. It remains to find it. Here it is.

[10] Not necessarily nonnegative.

[11] Preferably not 1, 2, a, b, d or f.

[12] We say that this finite sum is telescoping.

Proposition 6.8. *Let $f : [a, b] \to \mathbb{R}$ be a nondecreasing function. Then*

$$\lim_{n\to\infty} L_{2^n} = \lim_{n\to\infty} R_{2^n}$$

exist. If f is integrable then both limits are equal to the integral $J = \int_a^b f$.

Proof. Restricting to $N = 2^n$ we obtain equidistant partitions with the property that

$$L_1 \le L_2 \le L_4 \le L_8 \le \cdots \le R_8 \le R_4 \le R_2 \le R_1. \tag{6.15}$$

You will prove this in Exercise 6.9 below. This by itself[13] implies that

$$\sup_{n\in\mathbb{N}} L_{2^n} \le \inf_{n\in\mathbb{N}} R_{2^n},$$

but strict inequality is impossible in view of (6.14). Thus we must have

$$\lim_{n\to\infty} L_{2^n} = \sup_{n\in\mathbb{N}} L_{2^n} = \inf_{n\in\mathbb{N}} R_{2^n} = \lim_{n\to\infty} R_{2^n}$$

because of Theorem 2.28. If f is integrable, then $J = \int_a^b f$ satisfies

$$L_{2^n} \le J \le R_{2^n}$$

and is therefore equal to both limits. □

Exercise 6.9. Prove (6.15). Hint: the equidistant partition with $N = 2^{n+1}$ is a refinement of the equidistant partition with $N = 2^n$.

Exercise 6.10. Verify that for nonincreasing functions the story is exactly the same, except for reversed roles of the left and right endpoint sums.

6.3 Non-equidistant partitions; common refinements

With Proposition 6.8 we have in fact established the existence of a unique number J which candidates for being called the integral of f from a to b. If (6.13) turns out to hold for all partitions, then it must be that[14] f is integrable. To show that (6.13) does indeed hold we need the following theorem.

[13] In particular every such lower sum is less than or equal to every such upper sum.

[14] We did not specify f so we cannot compute J like we did for $f(x) = x^p$ with $p \in \mathbb{N}$.

Theorem 6.11. *Let $f : [a, b] \to \mathbb{R}$ be nondecreasing, let P be a partition given by*

$$a = x_0 \leq x_1 \leq \cdots \leq x_N = b,$$

and let Q be another partition given by

$$a = y_0 \leq y_1 \leq \cdots \leq y_M = b.$$

Define the upper sum[15]

$$\bar{S} = \sum_{k=1}^{N} f(x_k)(x_k - x_{k-1})$$

and the lower sum

$$\underline{S} = \sum_{l=1}^{M} f(y_{l-1})(y_l - y_{l-1}).$$

Then $\underline{S} \leq \bar{S}$.

For the proof of Theorem 6.11 we need one more definition.

Definition 6.12. *For P and Q as in Theorem 6.11, the* common refinement

$$a = z_0 \leq z_1 \leq \cdots \leq z_K = b, \tag{6.16}$$

is the partition that is obtained by simultaneously putting the numbers

$$x_1 \leq \cdots \leq x_{N-1} \quad \textit{and} \quad y_1 \leq \cdots \leq y_{M-1}$$

in increasing order. So $K - 1 = M - 1 + N - 1$ and every z_i is either an x_k or a y_l.

Proof of Theorem 6.11. Let

$$m_l = \min_{[y_{l-1}, y_l]} f = f(l_{l-1}), \quad \tilde{m}_i = \min_{[z_{i-1}, z_i]} f = f(z_{i-1}),$$

$$\tilde{M}_i = \max_{[z_{i-1}, z_i]} f = f(z_i), \quad M_k = \max_{[x_{k-1}, x_k]} f = f(x_k).$$

Then

$$\sum_{l=1}^{M} m_l(y_l - y_{l-1}) \leq \sum_{i=1}^{K} \tilde{m}_i(z_i - z_{i-1}) \leq \sum_{i=1}^{K} \tilde{M}_l(z_i - z_{i-1}) \leq \sum_{k=1}^{N} M_k(x_k - x_{k-1})$$

for the lower sum obtained from Q and the upper sum obtained from P. It follows for every lower sum $\underline{S}$ and every upper sum $\bar{S}$ that $\underline{S} \leq \bar{S}$. □

[15]For future purposes we write $\bar{S}$ and $\underline{S}$ for R and L now.

Theorem 6.13. *Let $f : [a, b] \to \mathbb{R}$ be nondecreasing*[16]*. Then f is Riemann integrable. In other words, there is a unique $J \in \mathbb{R}$ such that*

$$\underline{S} \le J \le \bar{S}$$

for every lower Riemann sum $\underline{S}$ and every upper Riemann sum $\bar{S}$. This real number J is by Definition 6.7 the integral *of f from a to b, notation*

$$J = \int_a^b f(x)\,dx.$$

Proof of Theorem 6.13. Let $\underline{S}$ and $\bar{S}$ be lower and upper sums for some partitions. By Theorem 6.11 we have that $\underline{S} \le \bar{S}$. So every upper sum is an upper bound for the nonempty set

$$S_{\text{lower}} = \left\{ \sum_{k=1}^{N} f(x_{k-1})(x_k - x_{k-1}) : a = x_0 \le x_1 \le \cdots \le x_N = b \right\}$$

of all possible lower sums. Let J be the lowest upper bound of S_{lower}. Then $\underline{S} \le J$ for every $\underline{S}$ because J is an upper bound of S_{lower}. Since $\bar{S}$ is also an upper bound of S_{lower} it must then be that $J \le \bar{S}$ because J is the *lowest* upper bound of S_{lower}. Thus $\underline{S} \le J \le \bar{S}$ for all $\underline{S}$, $\bar{S}$. No other number $\tilde{J}$ can have this property in view of (6.14) and Archimedes' principle. □

Exercise 6.14. Explain once more how Theorem 1.5 is used in the conclusion of the proof of Theorem 6.13.

Remark 6.15. *For monotone functions the integral is the unique number squeezed in between all lower and all upper sums. In other words, monotone functions are integrable. This fundamental result is a straightforward consequence of Archimedes' Theorem 1.5 and Theorem 6.11. It could have been stated and proved in Section 1.3.*

Exercise 6.16. Let f and g be nondecreasing functions defined on $[a, b]$. Then also $f + g$ is nondecreasing and therefore the functions $f, g, f + g$ are integrable according to Theorem 6.13. Prove that

$$\int_a^b (f + g) = \int_a^b f + \int_a^b g.$$

[16]The statement for nondecreasing functions is similar.

6.4 A limit theorem

What about integrals of sequences of monotone functions f_n? The following theorem says that

$$\lim_{n\to\infty} \int_a^b f_n(x)\,dx = \int_a^b f(x)\,dx \quad \text{if} \quad f(x) = \lim_{n\to\infty} f_n(x)$$

exists for every $x \in [a,b]$.

Theorem 6.17. *Let $f_n : [a,b] \to \mathbb{R}$ be a sequence of nondecreasing functions indexed by $n \in \mathbb{N}$. Suppose that*

$$f(x) = \lim_{n\to\infty} f_n(x)$$

exists for every $x \in [a,b]$. Then the function f thus defined is nondecreasing and the integrals

$$J_n = \int_a^b f_n(x)\,dx$$

define a sequence J_n which converges to

$$J = \int_a^b f(x)\,dx$$

as $n \to \infty$, i.e.

$$\lim_{n\to\infty} \int_a^b f_n(x)\,dx = \int_a^b f(x)\,dx. \tag{6.17}$$

A similar (equivalent) statement holds for sequences of nonincreasing functions $f_n : [a,b] \to \mathbb{R}$.

Proof of Theorem 6.17. The monotonicity of

$$f(x) = \lim_{n\to\infty} f_n(x)$$

follows from Definition 6.5: we consider the sequence $f_n(x_2) - f_n(x_1) \geq 0$ for arbitrary $a \leq x_1 \leq x_2 \leq b$ and apply Proposition 2.33 to conclude that $f(x_2) - f(x_1) \geq 0$.

As many times before, let $\varepsilon > 0$. Consider a lower sum L and an upper sum R for the limit function f, with the partition P as in Definition 6.8 chosen such that

$$R - L < \varepsilon.$$

This is possible because f is monotone and therefore Theorem 6.13 applies. Denote the lower and upper sums for $\int_a^b f_n$ for that same partition by L_n and R_n. Then we have

$$L_n \le J_n \le R_n \quad \textit{and} \quad L \le J \le R.$$

It also holds that $L_n \to L$ and $R_n \to R$. This holds because $f_n(x_k) \to f(x_k)$ as $n \to \infty$ for every $k = 0, \dots, N$. In particular it follows that there is an $N \in \mathbb{N}$ such that for all $n \ge N$ we have

$$L - \varepsilon < L_n \le J_n \le R_n < R + \varepsilon.$$

But we also have that

$$L - \varepsilon < L \le J \le R < R + \varepsilon.$$

Thus[17]

$$|J_n - J| < R - L + 2\varepsilon < 3\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary this completes the proof. □

6.5 Scaling and shifting; logarithm and exponential

Exercise 6.18. Let $a, b, \xi, \lambda \in \mathbb{R}$, $a < b, \lambda > 0$. Let $f : [a, b] \to \mathbb{R}$ be a monotone function. Show directly from Theorem 6.13 that

$$\int_a^b f(x)\,dx = \int_{a+\xi}^{b+\xi} f(x-\xi)\,dx \quad \text{and} \quad \int_a^b f(x)\,dx = \frac{1}{\lambda}\int_{a\lambda}^{b\lambda} f\left(\frac{x}{\lambda}\right)\,dx.$$

Exercise 6.19. For $b > 0$ and $p \in \mathbb{N}$ the area of

$$\{(x, y) \in \mathbb{R}^2 : 0 \le y \le x^p \le b^p\}$$

equals the quotient of b^{p+1} and $p + 1$. In integral notation this means that

$$\int_0^b x^p\,dx = \frac{b^{p+1}}{p+1}.$$

Prove this statement from the known statement for $b = 1$ and relate it to scaling the units on the axes.

[17] With a bit more care we get $|J_n - J| < 2\varepsilon$ but so what?

Exercise 6.20. Likewise, for $0 \le a < b$ and $p \in \mathbb{N}$, the area of

$$\{(x, y) \in \mathbb{R}^2 : a \le x \le b,\, 0 \le y \le x^p\}$$

is

$$\int_a^b x^p \, dx = \left[\frac{x^p}{p+1}\right]_a^b = \frac{b^{p+1}}{p+1} - \frac{a^{p+1}}{p+1}.$$

Use Theorem 6.13 and whatever it takes to prove this formula.

Definition 6.21. *For $x > 0$ we define $\ln x$, the natural logarithm of x, somewhat unnaturally, by*

$$\ln x = \int_1^x \frac{1}{s}\, ds.$$

Exercise 6.22. Apply Exercise 6.18 to Definition 6.21 and rewrite the formula for $\ln y$ as an integral from x to xy if $x > 1$ and $y > 1$. Conclude that

$$\ln xy = \ln x + \ln y.$$

Then prove this identity for all $x, y \in \mathbb{R}^+$. Hint: show first that

$$\ln x + \ln \frac{1}{x} = 0$$

for all $x > 0$. Explain the meaning of all these identities in terms of areas.

Exercise 6.23. We define the functions $e_n : [0, \infty) \to [1, \infty)$ by

$$e_n(x) = 1 + \int_0^x e_{n-1} \quad \text{and} \quad e_0(x) = 1 \quad \text{for every} \quad x \ge 0.$$

Then $e_n(0) = 1$ for every $n \in \mathbb{N}$. Prove that $e_n(x)$ is a strictly increasing convergent sequence for every $x > 0$ and that

$$\exp(x) := \lim_{n\to\infty} e_n(x) = 1 + \int_0^x \exp$$

for every $x \ge 0$. Hint: use Exercise 6.19 and Exercise 6.16 in every iteration step. You need to establish that $e_n(x)$ is bounded from above for every fixed $x > 0$ to conclude.

Exercise 6.24. (continued) Also show that

$$\exp(\mu x) = 1 + \mu \int_0^x \exp(\mu s)\, ds$$

for every $\mu > 0$ and every $x \ge 0$. Hint: combine Exercise 6.23 with Exercise 6.18.

6.6 Exercises

Exercise 6.25. It follows from Definition 6.21 that $\ln$ is a strictly increasing function on $\mathbb{R}^+$. Prove and use

$$\ln n \geq \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{>\frac{1}{2}} + \underbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}_{>\frac{1}{2}} + \cdots + \frac{1}{n} = \sum_{k=2}^{n} \frac{1}{k}$$

to show[18] that $\ln x \to \infty$ as $x \to \infty$. What can you conclude for $x \to 0$?

Exercise 6.26. Use the definition of the integral and Definition 6.21 to show that

$$\ln 2 = \int_0^1 \frac{1}{1+x}\, dx.$$

Exercise 6.27. Let $g : [0,1] \to \mathbb{R}$ be defined by

$$g(x) = \frac{1}{1+x}.$$

Let

$$f(x) = \begin{cases} g(x) & \text{for } 0 \leq x < 1 \\ 0 & \text{for } x = 1 \end{cases} \qquad \text{and} \quad f_n(x) = \frac{1-x^{2n}}{1+x}.$$

Combine Exercise 6.26 and Theorem 6.17 with $[a,b] = [0,1]$ to prove that

$$\left(\frac{1}{1} - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{5} - \frac{1}{6}\right) + \left(\frac{1}{7} - \frac{1}{8}\right) + \cdots = \ln 2.$$

Hint: it follows from Theorem 1.8 that[19]

$$g(x) = \frac{1}{1+x} = \underbrace{1 - x + x^2 - x^3 + x^4 - x^5 + x^6 - x^7}_{f_4(x)} + \cdots = \lim_{n\to\infty} f_n(x)$$

for all x with $0 \leq x < 1$.

[18] Give a definition first, in the spirit of Exercise 2.49.

[19] Plot some graphs to see what's going on.

Exercise 6.28. Let $f_n : [a,b] \to \mathbb{R}$ be a sequence of functions with $f_n(x)$ nondecreasing in n and x. Then $J_n = \int_a^b f_n$ is a nondecreasing sequence. Suppose that J_n is bounded. Prove that

$$f(x) = \lim_{n\to\infty} f_n(x)$$

exists for every $x \in [a,b)$ and is nondecreasing in x, and that

$$\int_a^x f \to J = \lim_{n\to\infty} J_n \quad \text{as} \quad x \to b.$$

Exercise 6.29. (continued) If J_n is not bounded then a definition as in Exercise 2.49 applies to J_n. Formulate and prove a statement about

$$\int_a^x f \quad \text{for} \quad x \to \infty.$$

Exercise 6.30. Consider in $\mathbb{R}^2$ the points

$$P_1 = \left(\frac{1}{\sqrt{2}}, 0\right), \quad P_2 = \left(0, \frac{1}{\sqrt{2}}\right) \quad \text{and} \quad P_3 = (\lambda, \lambda),$$

where $\lambda > 0$ is chosen such that $d(P_i, P_j) = 1$ for all $i, j \in \{1,2,3\}$ with $i \neq j$. Then P_1, P_2, P_3 are the vertices of an equilateral triangle with all edges of unit length. Denote its area by V_2. Determine its area using the base times height formula with prefactor $\frac{1}{2}$. Hint: you have to solve a quadratic equation for λ. This gives two values of λ that you can both use to determine the height.

Exercise 6.31. In $\mathbb{R}^3$ the points

$$\left(\frac{1}{\sqrt{2}}, 0, 0\right), \quad \left(0, \frac{1}{\sqrt{2}}, 0\right) \quad \text{and} \quad \left(0, 0, \frac{1}{\sqrt{2}}\right)$$

are also the vertices of an equilateral triangle with all edges of unit length. Choose a fourth point with all coordinates positive and identical to one another to construct a tretrahedron with all edges of unit length. Determine its volume V_3 using the base times height rule with prefactor $\frac{1}{3}$.

Exercise 6.32. Then take the four points

$$\left(\frac{1}{\sqrt{2}},0,0,0\right), \quad \left(0,\frac{1}{\sqrt{2}},0,0\right), \quad \left(0,0,\frac{1}{\sqrt{2}},0\right) \quad \text{and} \quad \left(0,0,0,\frac{1}{\sqrt{2}}\right)$$

in $\mathbb{R}^4$ and a fifth point with all coordinates positive and identical to one another to construct a so-called simplex with all edges of unit length. Determine its 4-dimensional volume V_4. What's the prefactor in the base times height rule? And so on. What's the formula for general $n \in \mathbb{N}$? Hint: $V_1 = 1$, express V_n in V_{n-1}. Pay some attention to the other point you can choose.

7 Integration of bounded functions?

Let $a, b \in \mathbb{R}$ with $a < b$. We have seen that monotone functions $f : [a, b] \to \mathbb{R}$ are integrable. If f is nondecreasing then its range[1]

$$R_f = \{f(x) : a \leq f(x) \leq b\}, \tag{7.1}$$

is contained in the interval $[f(a), f(b)]$. A function f is called bounded if its range R_f is a bounded set. Clearly every nondecreasing function $f : [a, b] \to \mathbb{R}$ has this property. Monotone functions defined on *bounded closed* intervals are thus bounded. In this chapter we consider bounded but not necessarily monotone functions defined on intervals $[a, b]$ and ask the question: can we still integrate them?

7.1 Bounded integrable functions

Without a monotonicity assumption, the left and right endpoint sums (6.9) and (6.12) are no longer bounds for an integral that we would like to define. For some partitions we may have $R < L$, while $L < R$ for other partitions. In fact the maxima M_k and minima m_k used in these Riemann sums need not even exist. Instead we shall use, for $k = 1, \ldots, N$, the real numbers m_k, M_k defined by

$$\begin{aligned} m_k &= \inf\{f(x) : x \in I_k\} \\ M_k &= \sup\{f(x) : x \in I_k\} \end{aligned} \quad \text{in which} \quad I_k = [x_{k-1}, x_k]. \tag{7.2}$$

These numbers exist because[2] the range of f restricted to I_k is a bounded nonempty set contained in R_f. From Theorem 4.4 we do know for continuous $f : [a, b] \to \mathbb{R}$ that m_k and M_k are actually minima and maxima, but we will postpone the study of integrals of continuous functions for now.

Definition 7.1. *Let $a, b \in \mathbb{R}$ with $a < b$ and let $f : [a, b] \to \mathbb{R}$ be a bounded function, i.e. a function with bounded range. The function f is called* integrable *if there exists a unique number $J \in \mathbb{R}$ such that*

$$\underline{S} = \sum_{k=1}^{N} m_k(x_k - x_{k-1}) \leq J \leq \sum_{k=1}^{N} M_k(x_k - x_{k-1}) = \bar{S}$$

for all partitions (6.8) of $[a, b]$, where the numbers m_k, M_k are defined as in (7.2). The number J is called the integral *of f over $[a, b]$. We write*

$$J = \int_a^b f(x)\, dx.$$

[1] We have used this notation before in Section 2.4 and Theorem 7.5.

[2] See again Section 2.4.

The following theorem characterises the bounded integrable functions.

Theorem 7.2. *A bounded function $f : [a, b] \to \mathbb{R}$ is integrable if and only if for every $\varepsilon > 0$ there exists a partition P with $\bar{S} - \underline{S} < \varepsilon$. If so then in particular $J = \int_a^b f$ is contained in $[\underline{S}, \bar{S}]$, an interval of length less than ε.*

Proof. We copy the proof of Theorem 6.11, with min replaced by inf and max replaced by sup. That is we use (7.2) for the intervals of the partitions P, Q, and their common refinement R. It follows in exactly the same fashion that

$$\underline{S}_P \leq \underline{S}_R \leq \bar{S}_R \leq \bar{S}_Q.$$

□

Exercise 7.3. Take some to time reflect on this simple and effective "if and only if" criterion for the integrability of bounded functions.

Exercise 7.4. Prove that the function f defined by

$$f(x) = \begin{cases} 1 & \text{for } x \in \mathbb{Q} \\ 0 & \text{for } x \notin \mathbb{Q} \end{cases}$$

is not integrable on $[0, 1]$.

Exercise 7.4 shows that not every bounded function $f : [a, b] \to \mathbb{R}$ can be integrated. Too bad. In Chapter 8 we will show that every $f \in C([a, b])$ is integrable, but for now we are happy with the statement in the following theorem. It has the integrability of Lipschitz continuous functions as an obvious consequence[3].

Theorem 7.5. *Suppose the bounded function $f : [a, b] \to \mathbb{R}$ is integrable, and that $F : R_f \to \mathbb{R}$ is a Lipschitz continuous function defined on the (bounded) range*

$$R_f = \{f(x) : a \leq x \leq b\}$$

of f. Then the composition $F \circ f : [a, b] \to \mathbb{R}$ is also bounded and integrable on $[a, b]$. In particular every Lipschitz continuous function $F : [a, b] \to \mathbb{R}$ is integrable.

[3]And also the integrability of $F \circ f$ with F Lipschitz continuous and f monotone.

Proof of Theorem 7.5. The function

$$f^* := F \circ f$$

is bounded because F is Lipschitz continuous and f is bounded. Let M_k^* and m_k^* be the suprema and infima of f^* on the intervals I_k of a partition P, and let L be the Lipschitz constant of F. It should be clear from[4]

$$|F(y) - F(\tilde{y})| \leq L|y - \tilde{y}| \quad \text{for all} \quad y, \tilde{y} \in R_f,$$

that then also the estimate

$$M_k^* - m_k^* \leq L(M_k - m_k) \tag{7.3}$$

holds. You are asked to prove this claim in Exercise 7.6 below.

Now let $\varepsilon > 0$ and let P be a partition for which

$$0 \leq \bar{S} - \underline{S} = \sum_{k=1}^{N} (M_k - m_k)(x_k - x_{k-1}) < \varepsilon,$$

with m_k, M_k defined in (7.2). This P is provided by Theorem 7.2 because we assumed that f is integrable on $[a, b]$. We examine how P performs for f^*. As a consequence of (7.3) we have for the Riemann sums $\underline{S}^*$ and $\bar{S}^*$ of $F \circ f = f^*$ that

$$0 \leq \bar{S}^* - \underline{S}^* = \sum_{k=1}^{N} (M_k^* - m_k^*)(x_k - x_{k-1}) \leq$$

$$L \sum_{k=1}^{N} (M_k - m_k)(x_k - x_{k-1}) < L\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, Theorem 7.2 and an L-trick[5] complete the proof. The special case that $f(x) = x$ and the integrability of monotone functions imply that Lipschitz continuous functions $F : [a, b] \to \mathbb{R}$ are integrable. $\square$

Exercise 7.6. Prove (7.3). Hint: it suffices to consider the case that $N = 1$ and show for

$$m = \inf_{a \leq x \leq b} f(x), \quad m^* = \inf_{a \leq x \leq b} F(f(x)),$$

[4]See (3.7) in Definition 3.3.

[5]See (2.18).

$$M = \sup_{a \le x \le b} f(x), \quad M^* = \sup_{a \le x \le b} F(f(x))$$

that

$$M^* - m^* \le L(M - m).$$

My original hint was: explain why there are sequences $\bar{x}_n$ and $\underline{x}_n$ such that

$$F(f(\bar{x}_n)) \to M^* \quad \text{and} \quad F(f(\underline{x}_n)) \to m^*,$$

and estimate $F(f(\bar{x}_n)) - F(f(\underline{x}_n))$. But I prefer Harold's hint: show first that

$$\sup_{x \in I} F(f(x)) - \inf_{x \in I} F(f(x)) = \sup_{x, \tilde{x} \in I} (F(f(x)) - F(f(\tilde{x}))).$$

7.2 Variations and elementary properties

Here we collect some rather trivial properties of the integral without proof.

Exercise 7.7. Let $f : [a, b] \to \mathbb{R}$ be bounded and $c \in (a, b)$. Prove that f is integrable over $[a, b]$ if and only if f integrable over both $[a, c]$ and $[c, b]$. If so, it holds that

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

Definition 7.8. *Let $f : [a, b] \to \mathbb{R}$ be bounded and integrable. Then*[6]

$$\int_b^a f(x)\,dx := -\int_a^b f(x)\,dx.$$

Exercise 7.9. Prove for all $a, b, c \in \mathbb{R}$ that

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$

if all integrals exist[7].

[6] Consistent with the possible intuition that dx in $\int_a^b f(x)\,dx$ is negative if $a > b$.

[7] As integrals of bounded functions of course.

Exercise 7.10. A bounded integrable function $f : [a, b] \to \mathbb{R}$ can be modified in a point $x_0 \in [a, b]$ by introducing the function $g : [a, b] \to \mathbb{R}$ defined by $g(x_0) = c_0$ and $g(x) = f(x)$ for all $x \in [a, b]$ with $x \neq x_0$. Prove that $g : [a, b] \to \mathbb{R}$ is integrable and $\int_a^b f(x)\,dx = \int_a^b g(x)\,dx$, no matter what the number $c_0 \in \mathbb{R}$ actually is.

Exercise 7.11. Is the function f defined by

$$f(x) = \begin{cases} 1 & \text{if } \frac{1}{x} \in \mathbb{N} \\ 0 & \text{if not} \end{cases}$$

integrable on $[0, 1]$?

7.3 Improper integrals

Let $I = (a, b) \subset \mathbb{R}$ be an open nonempty interval, possibly unbounded, so

$$-\infty \leq a < b \leq \infty,$$

and suppose that $f : (a, b) \to \mathbb{R}$ is integrable on every closed bounded interval $[\alpha, \beta] \subset (a, b)$. Then we define the improper integral $\int_a^b f$ by

$$\int_a^b f = \int_a^b f(x)\,dx = \lim_{\alpha \downarrow a} \lim_{\beta \uparrow b} \int_\alpha^\beta f(x)\,dx = \lim_{\beta \uparrow b} \lim_{\alpha \downarrow a} \int_\alpha^\beta f(x)\,dx$$

if the double limits exist. It's not hard to show that if one the double limits exists then so does the other and the limit values coincide. In the case that (a, b) is a bounded interval and $f : (a, b) \to \mathbb{R}$ is bounded the existence of the improper integral is equivalent to the proper integral of $f : [a, b] \to \mathbb{R}$ with any choice of value for $f(a)$ and $f(b)$, and the values of the integrals coincide.

7.4 Another limit theorem

We already saw one theorem of the type

$$\text{if } f_n \to f \text{ then } \int_a^b f_n \to \int_a^b f, \tag{7.4}$$

namely Theorem 6.17 in which all f_n were monotone. Here is another and perhaps more important such theorem. More important because it can be interpreted as the continuity statement of the map that sends integrable functions to real numbers by taking their integrals.

Theorem 7.12. *Let $f_n : [a,b] \to \mathbb{R}$ be a sequence of bounded integrable functions indexed by $n \in \mathbb{N}$. Suppose that f_n converges uniformly on $[a,b]$ to some function $f : [a,b] \to \mathbb{R}$. Then f is also (bounded and) integrable, and*

$$\int_a^b f_n(x)\,dx \to \int_a^b f(x)\,dx \quad as \quad n \to \infty.$$

Proof of Theorem 7.12. In view of Exercise 6.18 it suffices to give the proof of the statements in the theorem for the special case that $[a,b] = [0,1]$. We first apply Definition 4.15 with $\varepsilon = 1$ to conclude that the limit function f is bounded. Next, let $\varepsilon > 0$ and take $N \in \mathbb{N}$ such that for all $n \geq N$ and all $x \in [0,1]$ it holds that

$$|f_n(x) - f(x)| < \varepsilon. \tag{7.5}$$

This is possible since f_n is uniformly convergent on $[0,1]$.

We then apply Theorem 7.2 to obtain a partition P with lower and upper sums $\underline{S}_N$ and $\bar{S}_N$ for $\int_0^1 f_N$ such that

$$\bar{S}_N - \underline{S}_N < \varepsilon.$$

Let us examine how P does for the limit function f.

Consider the suprema $M_k^{(N)}$ and infima $m_k^{(N)}$ used for f_N on the intervals I_k of the partition in the definition of $\bar{S}_N$ and $\underline{S}_N$. Then

$$m_k^{(N)} \leq f_N(x) \leq M_k^{(N)} \quad \text{for all} \quad x \in I_k.$$

Combined with (7.5) this yields

$$m_k^{(N)} - \varepsilon \leq f(x) \leq M_k^{(N)} + \varepsilon \quad \text{for all} \quad x \in I_k.$$

It follows for the suprema M_k and infima m_k of f on I_k that

$$M_k - m_k \leq (M_k^{(N)} + \varepsilon) - (m_k^{(N)} - \varepsilon).$$

Adding up we then find that

$$\bar{S} - \underline{S} \leq \bar{S}_N - \underline{S}_N + 2\varepsilon < 3\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary Theorem 7.2 and a 3-trick[8] prove that $J = \int_0^1 f$ exists. This is the first statement in the theorem, and in particular

$$J \in [\underline{S}, \bar{S}],$$

[8]See (2.18).

an interval of length less than 3ε.

Let us also examine how P does for the functions f_n. For $n \geq N$ we have from (7.5) that[9] $|\underline{S}_n - \underline{S}| \leq \varepsilon$ and $|\bar{S}_n - \bar{S}| \leq \varepsilon$. Therefore $J_n = \int_a^b f_n$ has the property that

$$\underline{S} - \varepsilon \leq \underline{S}_n \leq J_n \leq \bar{S}_n \leq \bar{S} + \varepsilon.$$

Thus

$$J_n \in [\underline{S} - \varepsilon, \bar{S} + \varepsilon], \text{ while also } J \in [\underline{S}, \bar{S}].$$

But then it follows that

$$|J_n - J| \leq \varepsilon + \bar{S} - \underline{S} < 4\varepsilon \quad \text{for all } n \geq N.$$

Since $\varepsilon > 0$ was arbitrary a 4-trick[10] completes the proof that $J_n \to J$ as $n \to \infty$, which is the second statement in the theorem. □

7.5 Integrals are continuous linear functionals

The title of this section is explained by the convention of calling maps from spaces of functions to $\mathbb{R}$ functionals[11].

Theorem 7.13. *Let*

$$\mathrm{RI}([a,b]) = \{f : [a,b] \to \mathbb{R} : \ f \text{ is bounded and integrable}\} \tag{7.6}$$

be the space of bounded integrable functions on $[a,b]$. Then $\mathrm{RI}([a,b])$ is a complete metric space with respect to the metric defined by

$$d(f,g) = \sup_{a \leq x \leq b} |f(x) - g(x)| \tag{7.7}$$

for all $f, g \in \mathrm{RI}([a,b])$.

Proof of Theorem 7.13. First we reformulate Exercise 4.41 as a separate result in Theorem 7.15 below. Recall that in Remark 4.18 we introduced the metric space

$$B([a,b]) = \{f : [a,b] \to \mathbb{R} : \ R_f \text{ is bounded}\}. \tag{7.8}$$

of bounded functions on $[a,b]$. The range R_f of $f : [a,b] \to \mathbb{R}$ was already defined in Section 2.4.

[9] Is it clear why?

[10] Not another footnote.

[11] So functionals are functions of functions.

Definition 7.14. *Let $B([a,b])$ be the space of all bounded functions from $[a,b]$ to $\mathbb{R}$ defined in (7.8) . The metric in $B([a,b])$ is defined by*

$$d(f,g) = \sup_{a\le x\le b} |f(x)-g(x)|$$

for all $f,g \in B([a,b])$, just as[12] *in (7.7).*

Theorem 7.15. *The space $B([a,b])$ is a complete metric space*[13].

Proof of Theorem 7.15. We only have to show that $B([a,b])$ is complete[14]. We note that for $\varepsilon > 0$ and[15] $f,g \in B([a,b])$

$$d(f,g) \le \varepsilon \iff \forall_{x\in[a,b]} : |f(x)-g(x)| \le \varepsilon \tag{7.9}$$

holds by the definition of supremum. Note that it does not matter whether we write $\le \varepsilon$ or $< \varepsilon$ in ε-statements for convergence.

Now let f_n be a Cauchy sequence in $B([a,b])$. This means that

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{m,n\ge N}\, \forall_{x\in[a,b]} : |f_n(x)-f_m(x)| < \varepsilon.$$

Just like in the proof of Theorem 4.12 it then follows that

$$f(x) = \lim_{m\to\infty} f_m(x)$$

exists for every $x \in [a,b]$, and that

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\ge N}\, \forall_{x\in[a,b]} : |f_n(x)-f(x)| \le \varepsilon.$$

In other words

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\ge N} : d(f_n,f) \le \varepsilon.$$

This statement implies on the one hand that $f \in B([a,b])$, and on the other hand that $f_n \to f$ in $B([a,b])$. This completes the proof of Theorem 7.15. □

We now complete the proof of Theorem 7.13. Recall that by (7.9) convergence in $B([a,b])$ is equivalent to uniform convergence. The first part of Theorem 7.12 says that the space $\mathrm{RI}([a,b])$ is a closed subset[16] of the complete metric space $B([a,b])$. Theorem 5.11 then implies that $\mathrm{RI}([a,b])$ is complete and so then is the proof of Theorem 7.13. □

[12] Check that this indeed defines a metric.

[13] A Banach algebra in fact, see Remark 4.18; $[a,b]$ may be replaced by any set $A \neq \emptyset$.

[14] Note that its metric "extends" the metric defined in the smaller metric space $\mathrm{RI}([a,b])$.

[15] In fact only $f-g \in B([a,b])$ is needed to define $d(f,g)$.

[16] See Definition 5.8.

Theorem 7.16. *The map or functional* $\phi : \mathrm{RI}([a,b]) \to \mathbb{R}$ *defined by*

$$\phi(f) = \int_a^b f, \tag{7.10}$$

is continuous.

Proof of Theorem 7.16. By the definition of continuity[17] this is now just a reformulation of the second part of Theorem 7.12. □

We finish with a theorem that says that the integral is in fact a linear Lipschitz continuous functional. The exercises below the theorem ask you to supply the proofs of the separate statements in the theorem.

Theorem 7.17. *If* $f, g \in \mathrm{RI}([a,b])$ *and* $\lambda \in \mathbb{R}$ *then also* $f + g \in \mathrm{RI}([a,b])$ *and* $\lambda f \in \mathrm{RI}([a,b])$. *Moreover,*

$$\int_a^b (f(x) + g(x))\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx;$$

$$\int_a^b \lambda f(x)\,dx = \lambda \int_a^b f(x)\,dx.$$

In other words, $\mathrm{RI}([a,b])$ *is a vector space, and the map* ϕ *defined by (7.10) is linear. Moreover, the function* $|f|$ *defined by* $|f|(x) = |f(x)|$ *is also in* $\mathrm{RI}([a,b])$, *and*

$$\left| \int_a^b f(x)\,dx \right| \leq \int_a^b |f(x)|\,dx. \tag{7.11}$$

Thus the functional defined by (7.10) in Theorem 7.16 satisfies

$$|\phi(f) - \phi(f)| = (b-a)\,d(f,g)$$

for all $f, g \in \mathrm{RI}([a,b])$ *and is thereby Lipschitz continuous.*

Remark 7.18. *Summing up, the space* $\mathrm{RI}([a,b])$ *is a complete metric* vector *space*[18], *and the map* $\phi : \mathrm{RI}([a,b]) \to \mathbb{R}$ *defined by* $\phi(f) = \int_a^b f$ *is linear and Lipschitz continuous with Lipschitz constant* $L = b - a$.

Exercise 7.19. Prove the statements about $f + g$ in Theorem 7.17. Hint: take a partition refining[19] two partitions chosen for f and g upon applying Theorem 7.2. Alternative hint: reason directly from Definition 7.1.

[17] See Definition 5.13.

[18] Such spaces are called Banach spaces.

[19] See Theorem 6.11 and Exercise 7.2.

Exercise 7.20. Easy: prove the statements about λf in Theorem 7.17.

Exercise 7.21. Give a proof that $f \in \mathrm{RI}([a,b])$ implies $|f| \in \mathrm{RI}([a,b])$ and prove (7.11) directly from Definition 7.1.

Exercise 7.22. Prove the Lipschitz continuity[20] of ϕ. Hint: use (7.11).

7.6 Integral equations

Exercise 6.23 provided us with a function[21] $f : [0,\infty) \to \mathbb{R}$ satisfying

$$f(x) = 1 + \int_0^x f = 1 + \int_0^x f(s)\,ds \tag{7.12}$$

for every $x > 0$. Now let $[a,b]$ be a closed bounded interval with

$$0 \in [a,b],$$

and consider (7.12) as an integral equation for $f \in \mathrm{RI}([a,b])$. Thus (7.12) must hold for all $x \in [a,b]$.

Exercise 7.23. An exercise for your calculus course. Assume that f is continuously differentiable on $[a,b]$ and satisfies (7.12) for all $x \in [a,b]$. Prove that $f'(x) = f(x)$.

The goal of this section is to establish that integral equations such as (7.12) have (unique) solutions in $\mathrm{RI}([a,b])$. In fact we will consider more general integral equations[22]. For a given $f_0 \in \mathbb{R}$ and $F : \mathbb{R} \to \mathbb{R}$ consider the problem of finding a function $f : [a,b] \to \mathbb{R}$ such that

$$f(x) = f_0 + \int_0^x F(f(s))\,ds \quad \text{for all} \quad x \in [a,b]. \tag{7.13}$$

We will solve this integral equation under the assumption that $F : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous, with Lipschitz constant L.

[20] Use that

$$\left|\int_a^b f\right| \le (b-a)\,|f|_\infty \quad \text{in which} \quad |f|_\infty = \sup_{a\le x\le b} |f(x)|.$$

[21] For good reasons denoted by exp.

[22] Designed to solve $f'(x) = F(f(x))$ with "initial" condition $f(0) = f_0$, Remark 7.26.

Theorem 7.24. *Let $f_0 \in \mathbb{R}$ and let $F : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous with Lipschitz constant L. Define*

$$\Phi(f)(x) = f_0 + \int_0^x F(f(s))\,ds \quad \textit{for} \quad x \in [a,b] \tag{7.14}$$

and $f \in \mathrm{RI}([a,b])$. Then (7.14) defines a Lipschitz continuous map

$$\Phi : \mathrm{RI}([a,b]) \to \mathrm{RI}([a,b])$$

with Lipschitz constant less or equal than $L(b-a)$.

Proof. The right hand side of (7.14) is well-defined for every $x \in [a,b]$ and every $f \in \mathrm{RI}([a,b])$ thanks to Theorem 7.5. Every $f \in \mathrm{RI}([a,b])$ is mapped by Φ to a function $\Phi(f) : [a,b] \to \mathbb{R}$ defined by (7.14). How well-behaved is this function $\Phi(f)$? For $a \le y \le x \le b$ we have[23]

$$|\Phi(f)(x) - \Phi(f)(y)| = \left| \int_y^x F(f(s))\,ds \right| \le \underbrace{\sup_{a \le s \le b} |F(f(s))|}_{= |F \circ f|_\infty < \infty} (x-y).$$

Thus $\Phi(f)$ is Lipschitz continuous and thereby in $\mathrm{RI}([a,b])$, according to (the special case in) Theorem 7.5. It follows that $\Phi : \mathrm{RI}([a,b]) \to \mathrm{RI}([a,b])$.

Next we consider the difference $\Phi(f_1) - \Phi(f_2)$ for $f_1, f_2 \in \mathrm{RI}([a,b])$. This difference is defined by

$$\left(\Phi(f_1) - \Phi(f_2)\right)(x) = \int_0^x \left(F(f_1(s)) - F(f_2(s))\right)\,ds \quad \text{for} \quad x \in [a,b].$$

Here the value of $\Phi(f_1) - \Phi(f_2)$ in x is denoted $\left(\Phi(f_1) - \Phi(f_2)\right)(x)$, with brackets around $\Phi(f_1) - \Phi(f_2)$. We estimate this value next. Taking absolute values we have[24]

$$\begin{aligned} |\left(\left(\Phi(f_1) - \Phi(f_2)\right)(x)\right| &= \left| \int_0^x F(f_1(s)) - F(f_2(s))\,ds \right| \\ \le \int_0^x |F(f_1(s)) - F(f_2(s))|\,ds &\le \int_0^x L\,|f_1(s) - f_2(s)|\,ds \\ = L \int_0^x |f_1(s) - f_2(s)|\,ds &\le L \underbrace{\sup_{a \le x \le b} |f_1(s) - f_2(s)|}_{d(f_1,f_2)}\,|x| \end{aligned}$$

[23] Recall (4.11).

[24] Using the inequality in (7.11).

for every $x \in [a,b]$. If we take the supremum over all such x it follows that[25]

$$d(\Phi(f_1), \Phi(f_2)) \leq L(b-a)\, d(f_1, f_2) \quad \text{for all} \quad f_1, f_2 \in \mathrm{RI}([a,b]).$$

This completes the proof. □

Theorem 7.25. *Let $F : \mathbb{R} \to \mathbb{R}$ be a Lipschitz continuous function with Lipschitz constant L, let $a \leq 0 \leq b$ with $a < b$ and let $f_0 \in \mathbb{R}$. Assume that $L(b-a) < 1$. Then there exists a unique $f \in \mathrm{RI}([a,b])$ such that*

$$f(x) = f_0 + \int_0^x F(f(s))\, ds \tag{7.15}$$

for all $x \in [a,b]$.

Proof. By Theorem 7.24 the Banach Contraction Theorem applies to the equation $f = \Phi(f)$ in $\mathrm{RI}([a,b])$. □

Remark 7.26. *It turns out that Theorem 10.10 implies that the unique solution f of the integral equation*

$$f(x) = f_0 + \int_0^x F(f(s))\, ds$$

is also the unique solution of the differential equation

$$f'(x) = F(f(x)) \quad \textit{with initial condition} \quad f(0) = f_0.$$

This is important in the theory of differential equations.

7.7 Exercises

Exercise 7.27. Let $f : [-1,1] \to \mathbb{R}$ be a bounded integrable function. Assume that f is odd, i.e. $f(x) = -f(-x)$ for all $x \in [-1,1]$. Prove that $\int_{-1}^{1} f = 0$.

Exercise 7.28. Let $f : [-1,1] \to \mathbb{R}$ be a bounded integrable function. Assume that f is even, i.e. $f(x) = f(-x)$ for all $x \in [-1,1]$. Prove that $\int_{-1}^{1} f = 2\int_{0}^{1} f$.

[25] The smallest possible Lipschitz constant is the maximum of $-a$ and b since $a \leq 0 \leq b$.

Exercise 7.29. Let $p > 1$ and $q > 1$ be as in Exercise 1.24, i.e.

$$\frac{1}{p} + \frac{1}{q} = 1,$$

and let $a > 0$ and $b > 0$ be real numbers. Use the integrals

$$\int_0^a x^{p-1}\,dx \quad \text{and} \quad \int_0^b y^{q-1}\,dy$$

and their interpretation as areas to explain why it must be that

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \quad \text{(Young's ineqality)}. \tag{7.16}$$

For amusement: give a direct proof using only algebra.

Exercise 7.30. Let $p > 1$ and $q > 1$ be as in Exercise 7.29, and let $a_1, \ldots, a_n \geq 0$, $b_1, \ldots, b_n \geq 0$ be real numbers, $n \in \mathbb{N}$. Prove that

$$\sum_{k=1}^{n} a_k b_k \leq \left(\sum_{k=1}^{n} a_k^p\right)^{\frac{1}{p}} \left(\sum_{k=1}^{n} b_k^q\right)^{\frac{1}{q}} \quad \text{(Hölder's inequality)}.$$

Hint: show first that it is sufficient to prove the inequality for the case that

$$\sum_{k=1}^{n} a_k^p = \sum_{k=1}^{n} b_k^q = 1$$

and then use Exercise 7.29.

Exercise 7.31. Exhibit a function $f : [a, b] \to \mathbb{R}$ not in $\mathrm{RI}([a, b])$ for which $|f|$ is.

Exercise 7.32. Show that

$$f, g \in \mathrm{RI}([a, b]) \implies fg \in \mathrm{RI}([a, b]).$$

Thus $\mathrm{RI}([a, b])$ is also a Banach algebra[26]. Hint:

$$\sup_I fg - \inf_I fg = \sup_{x,y \in I} |f(x)g(x) - f(y)g(y)|$$

[26] See Remark 4.18.

for intervals $I \subset [a,b]$; use

$$f(x)g(x) - f(y)g(y) = (f(x) - f(y))g(x) + f(y)(g(x) - g(y)),$$

estimate in terms of (bounds on) $|f|$ and $|g|$, $\sup_I f - \inf_I f$, $\sup_I g - \inf_I g$; take a partition refining two partitions chosen for f and g and conclude.

Exercise 7.33. Use the inequality[27]

$$|\sum_{i=1}^{n} a_i b_i| \leq \left(\sum_{i=1}^{n} |a_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |b_i|^q\right)^{\frac{1}{q}},$$

which holds for $p, q > 1$ with

$$\frac{1}{p} + \frac{1}{q},$$

to show that

$$|\int_a^b f(x)g(x)\,dx| \leq \left(\int_a^b |f(x)|^p\,dx\right)^{\frac{1}{p}} \left(\int_a^b |g(x)|^q\,dx\right)^{\frac{1}{q}}$$

for such p and q and $f, g \in \mathcal{RI}([a,b])$. Hint: use the conclusion of Exercise 7.32 and combine the inequality for the sums with the definition of integrability via finite sums.

Exercise 7.34. Recall from (4.11) that the supremum norm[28] in the vector space $B([a,b])$ was defined by

$$|f|_\infty = \sup\{|f(x)| : x \in [a,b]\}.$$

The importance of this concept and linear structures was scaled down in our approach to metric spaces of functions. But you should prove that for all $\lambda \in \mathbb{R}$ and for all $f, g \in B([a,b])$, with f not equal to the zero element in $B([a,b])$, the following norm axioms[29] hold:

$$|f|_\infty > 0, \quad |\lambda f|_\infty = |\lambda|\,|f|_\infty, \quad |f+g|_\infty \leq |f|_\infty + |g|_\infty. \tag{7.17}$$

The zero element in $B([a,b])$ is the zero function defined by $f(x) = 0$ for all $x \in [a,b]$).

[27] See Exercise 7.30 and also Section 20.3.

[28] The use of the subscript ∞ is related to the limit of $\left(\int_a^b |f|^p\right)^{\frac{1}{p}}$ as $p \to \infty$ for nice f.

[29] These axioms may have been mentioned in Linear Algebra, see also Exercise 5.39.

Exercise 7.35. Explain that $f \in B([a,b])$ is not equal to the zero element in $B([a,b])$ if and only if

$$\exists_{x\in[a,b]} : f(x) \neq 0.$$

Exercise 7.36. The norm of the zero element in $B([a,b])$ is the real number zero. Prove this from (7.17).

Exercise 7.37. Prove that the *triangle inequality* $|f+g|_\infty \leq |f|_\infty + |g|_\infty$ holds for all $f, g \in B([a,b])$.

Exercise 7.38. For $f \in \mathrm{RI}([0,1])$ define the function $\Phi(f) : [0,1] \to \mathbb{R}$ by

$$\Phi(f)(x) = \int_0^x (1 + f(s))\, ds$$

for all $x \in [0,1]$. Show that this defines a Lipschitz continuous map

$$\Phi : \mathrm{RI}([0,1]) \to \mathrm{RI}([0,1]).$$

Is Φ a contraction?

Exercise 7.39. Same questions as in Exercise 7.38 for Φ defined by

$$\Phi(f)(x) = \int_0^x \frac{1}{1+f(s)}\, ds,$$

but restricted to $\mathrm{RI}_+([0,1]) = \{f \in \mathrm{RI}([0,1]) : f(x) \geq 0 \text{ for all } x \in [0,1]\}$.

Exercise 7.40. For $\mu > 0$ let $B_\mu([0,\infty))$ be the space of functions $f : [0,\infty) \to \mathbb{R}$ for which the norm

$$|f|_\mu = \sup_{x\geq 0} \frac{|f(x)|}{\exp(\mu x)}$$

exists as a finite number. Show that

$$d_\mu(f,g) = |f-g|_\mu$$

defines a metric d_μ on $B_\mu([0,\infty))$. Show that this metric makes $B_\mu([0,\infty))$ a complete metric space, and that

$$\mathrm{RI}_\mu([0,\infty)) = \{f \in B_\mu([0,\infty)) : \ f \text{ is integrable over every } [0,T]\}$$

is a closed subspace.

Exercise 7.41. Consider the integral equation

$$f(x) = f_0 + \int_0^x F(f(s))\,ds = f_0 + \underbrace{\int_0^x F \circ f}_{\Phi(f)(x)}, \tag{7.18}$$

in which $F : \mathbb{R} \to \mathbb{R}$ is a Lipschitz continuous with Lipschitz constant $L > 0$ and $f_0 \in \mathbb{R}$ is given. Use Exercise 6.23 to show that

$$|\Phi(f)(x) - \Phi(g)(x)| \le L \int_0^x |f(s) - g(s)|\,ds = \frac{L}{\mu}\,\exp(\mu x)\,\underbrace{|f-g|_\mu}_{d_\mu(f,g)}$$

for all $f, g \in \mathrm{RI}_\mu([0,\infty))$ and conclude that for the metric d_μ we have

$$d_\mu(\Phi(f), \Phi(g)) \le \frac{L}{\mu}\, d_\mu(f,g).$$

Then prove that there exists a $\mu > 0$ such that (7.18) has a unique solution in $\mathrm{RI}_\mu([0,\infty))$ for every $f_0 \in \mathbb{R}$. Hint: use the Banach Contraction Theorem.

Exercise 7.42. Show that the integral equation (7.18) has a unique integrable solution $f : \mathbb{R} \to \mathbb{R}$, that is, f is integrable over every interval $[a,b] \subset \mathbb{R}$, and (7.18) holds for all $x \in \mathbb{R}$. Hint: put $x = -\xi$ to handle negative x.

Exercise 7.43. Let $F : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous. In view of Exercise 7.48 the integral equation (7.18), that is

$$f(x) = f_0 + \int_0^x F(f(s))\,ds,$$

has a unique solution f for every $f_0 \in \mathbb{R}$, defined for every $x \in \mathbb{R}$. We write $f(x; f_0)$ to indicate the dependence of the solution on f_0. We also write

$$S(x)(f_0) = f(x; f_0). \tag{7.19}$$

This defines a family of functions $S(x) : \mathbb{IR} \to \mathbb{IR}$. Prove that

$$S(x_1 + x_2) = S(x_2) \circ S(x_1) = S(x_1) \circ S(x_2)$$

for every $x_1, x_2 \in \mathbb{IR}$. Hint: write

$$f(x_1 + x_2) = f_0 + \int_0^{x_1} F(f(s))\, ds + \int_{x_1}^{x_1+x_2} F(f(s))\, ds,$$

rewrite the second integral as an integral from 0 to x_2, and recognise an integral equation for g defined by $g(s) = f(s + x_1)$.

Exercise 7.44. Prove that $\exp(x_1 + x_2) = \exp(x_1)\exp(x_2)$. Hint: Exercise 7.43.

Exercise 7.45. Consider the integral equation

$$f(x) = \int_0^x \int_0^t F(f(s))\, ds\, dt$$

for $x \in [0, T]$, $T > 0$. Assume that $F : \mathbb{IR} \to \mathbb{IR}$ is Lipschitz continuous with Lipschitz constant $L > 0$. Prove that this integral equation has a unique solution in $\mathrm{RI}([0, T])$ if $LT^2 < 2$. Hint: reason as for (7.15).

Exercise 7.46. Let f be the solution in Exercise 7.45. Use your calculus skills to find the differential equation that is satisfied by the solution f. What can you say about $f(0)$ and $f'(0)$? Write the integral equation for solving the differential equation that you found with initial data $f(0) = 1$ and $f'(0) = 2$.

Exercise 7.47. Prove that the integral equation in Exercise 7.45 has a unique solution in $\mathrm{RI}([0, T])$ for every $T > 0$. Hint: reason as in Exercise 7.41.

Exercise 7.48. Prove that the integral equation in Exercise 7.45 has a unique solution in $\mathrm{RI}([-T, T])$ for every $T > 0$.

Exercise 7.49. Consider the integral equation

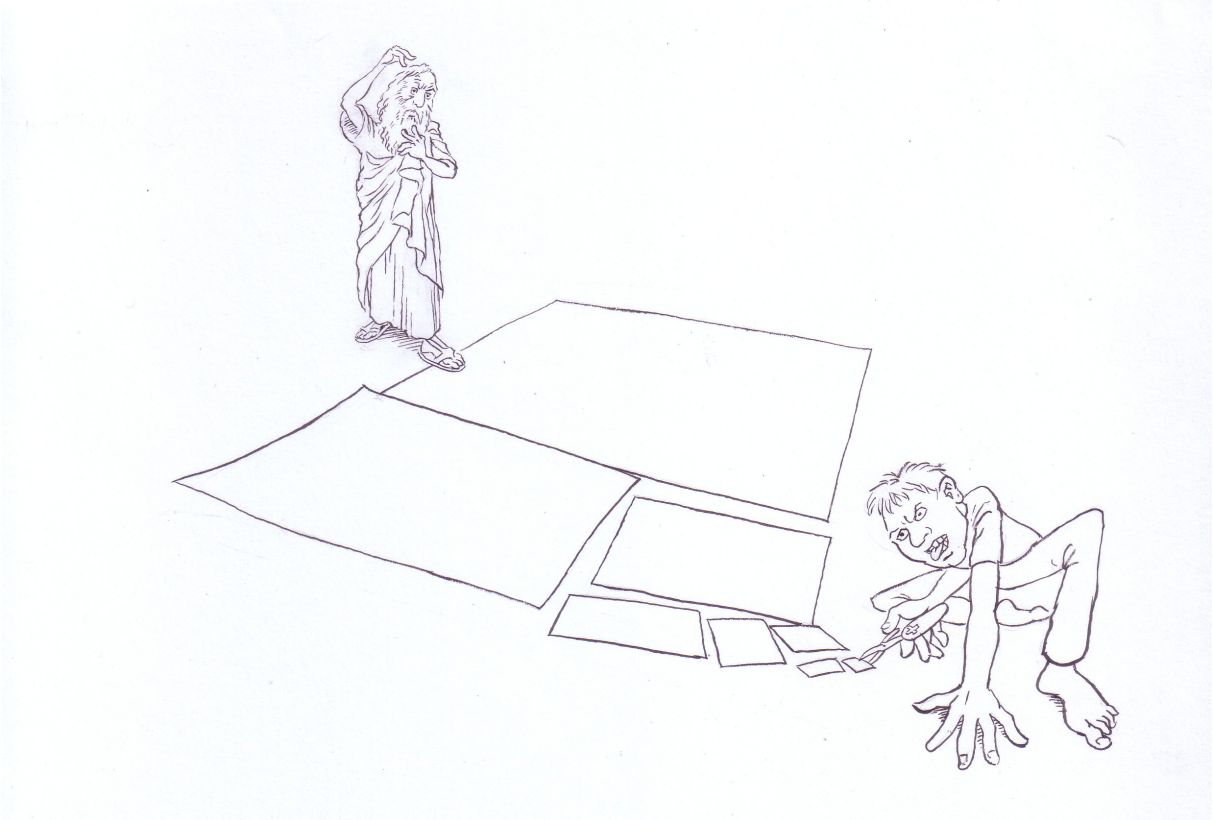
$$f(x) = \int_0^x \frac{1}{1+f(s)}\, ds.$$

Show that it has solution defined for all nonnegative $x \in \mathbb{R}$. Can you find a formula for $f(x)$? Examine what goes wrong for $x < 0$.

Exercise 7.50. Consider the integral equation

$$f(x) = f_0 + \int_0^x \frac{f(s)}{1+f(s)^2}\, ds.$$

Show for every $f_0 \in \mathbb{R}$ that it has solution defined for all $x \in \mathbb{R}$.



What about measure theory in $\mathbb{R}^2$?

8 Epsilons and deltas

In Definition 4.1 of Chapter 4 we called a function $f : A \to \mathbb{R}$, $A \subset \mathbb{R}$, *continuous* in $\xi \in A$ if

$$f(x_n) \to f(\xi)$$

for every sequence x_n in A with $x_n \to \xi$. Definition 5.13 in Chapter 5 copied Definition 4.1 for $A \subset X$, $f : A \to Y$, and X, Y abstract metric spaces. Theorem 8.1 below explains the title of this chapter and formulates *the other natural characterisation of continuity*[1]. We only state it for $X = A \subset \mathbb{R}$ and $Y = \mathbb{R}$.

Theorem 8.1. *Let $A \subset \mathbb{R}$ be nonempty, let $f : A \to \mathbb{R}$ be a function and let $\xi \in A$. Then f is continuous in ξ if and only if*

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{x\in A} : \underbrace{|x-\xi|}_{d(x,\xi)} < \delta \implies \underbrace{|f(x)-f(\xi)|}_{d(f(x),f(\xi))} < \varepsilon. \tag{8.1}$$

Proof of Theorem 8.1. To prove (8.1) from the statement in Definition 4.1 we argue by contraposition. Let $\xi \in A$ and suppose that (8.1) does not hold. Then

$$\exists_{\varepsilon>0}\, \forall_{\delta>0}\, \exists_{x\in A} : |x-\xi| < \delta \quad \text{and} \quad |f(x)-f(\xi)| \geq \varepsilon. \tag{8.2}$$

For every $n \in \mathbb{N}$ we use (8.2) with $\delta = \frac{1}{n}$. Denote the corresponding x by x_n. This defines a sequence x_n with $|x_n - \xi| < \frac{1}{n}$ whence $x_n \to \xi$ as $n \to \infty$. But $|f(x_n) - f(\xi)| \geq \varepsilon$ prevents $f(x_n) \to f(\xi)$ as $n \to \infty$. This is in contradiction with the continuity statement quoted in the first sentence of this chapter. We therefore conclude that (8.1) does indeed follow from the statement in Definition 4.1.

Conversely, assume that (8.1) holds. We have to show that $f(x_n) \to f(\xi)$ if x_n is a sequence in A with $x_n \to \xi$ as $n \to \infty$. So let $\varepsilon > 0$. Then (8.1) provides a $\delta > 0$ such that $|f(x_n) - f(\xi)| < \varepsilon$ if $|x_n - \xi| < \delta$. So we apply the definition of $x_n \to \xi$ with ε replaced by δ. This gives an N such that for all $n \geq N$ it holds that $|x_n - \xi| < \delta$ and thereby $|f(x_n) - f(\xi)| < \varepsilon$. This completes the proof of Theorem 8.1. □

[1] Actually the proof of Theorem 5.25 already contained this statement, namely

$$\forall_{\varepsilon>0}\, \exists_{\delta>0} : d_{_X}(x,\xi) < \delta \implies d_{_Y}(f(x), f(\xi)) < \varepsilon.$$

Exercise 8.2. Let $f : \mathbb{R} \to \mathbb{R}$, $\xi \in \mathbb{R}$, $\eta = f(\xi)$. For values $\varepsilon > 0$ and $\delta > 0$ draw the lines $x = \xi - \delta$, $x = \xi + \delta$, $y = \eta - \varepsilon$, $y = \eta + \varepsilon$, and explain geometrically what the implication in (8.1) says.

Exercise 8.3. Let $\xi = 2$, $f(x) = 2x + 1$, $f : \mathbb{R} \to \mathbb{R}$. Verify (8.1) by computing $\delta > 0$ in terms of $\varepsilon > 0$. Same question for $f(x) = x^2$.

Exercise 8.4. Let $A = [0, 1]$ and $f : A \to \mathbb{R}$ be defined by $f(x) = x^2$. Verify (8.1) for every $\xi \in A$. Is it possible to choose $\delta > 0$ depending on $\varepsilon > 0$ only? Same question for $A = \mathbb{R}$.

Exercise 8.5. Same question for $A = (0, 1)$ and $f(x) = \frac{1}{x}$.

In the above exercises we saw that sometimes δ depending on ε can be chosen independent of ξ for all ε, and sometimes it cannot. Such independence of δ on ξ is needed to prove a theorem that we have postponed so far, namely[2] that every continuous function $f : [a, b] \to \mathbb{R}$ is integrable, i.e.

$$C([a, b]) \subset \mathrm{RI}([a, b]). \tag{8.3}$$

8.1 Uniform continuity and integrability

Theorem 8.6. *Let $f \in C([a, b])$. Then f is integrable on $[a, b]$.*

To prove this theorem we recall that Theorem 7.2 decides on the integrability of bounded functions $f : [a, b] \to \mathbb{R}$. Given $\varepsilon > 0$ we have to show that

$$0 \leq \bar{S} - \underline{S} = \sum_{k=1}^{N} (M_k - m_k)(x_k - x_{k-1}) < \varepsilon \tag{8.4}$$

for at least one partition P of $[a, b]$. If this holds then the function is integrable. If not, then the function f is not integrable. It turns out that the following definition guarantees this property.

[2] See Remark 4.18, Definition 4.5 and Theorem 7.13.

Definition 8.7. *Let $A \subset \mathbb{R}$ and $f : A \to \mathbb{R}$. Then f is called* uniformly continuous on A *if*

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{x,\xi\in A} : \underbrace{|x-\xi|}_{d(x,\xi)} < \delta \implies \underbrace{|f(x)-f(\xi)|}_{d(f(x),f(\xi))} < \varepsilon.$$

You should carefully compare the statement in Definition 8.7 to the statement in Theorem 8.1. These are two different statements. One looks clearly stronger than the other but, according to Theorem 8.10 below, *both statements are equivalent if*[3] $A = [a, b]$.

Remark 8.8. *The statement that f is continuous in every $\xi \in A$ rewrites*[4] *as the non-uniform statement that*

$$\forall_{\varepsilon>0}\, \underbrace{\forall_{\xi\in A}\, \exists_{\delta>0}}_{pointwise}\, \forall_{x\in A} : |x-\xi| < \delta \implies |f(x)-f(\xi)| < \varepsilon,$$

and differs by one $\forall_{\xi\in A}$-$\exists_{\delta>0}$ swap from the uniform statement refrased from Definition 8.7 as

$$\forall_{\varepsilon>0}\, \underbrace{\exists_{\delta>0}\, \forall_{\xi\in A}}_{uniform}\, \forall_{x\in A} : |x-\xi| < \delta \implies |f(x)-f(\xi)| < \varepsilon.$$

Exercise 8.9. Let $A = \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$. For values $\xi \in \mathbb{R}$, $\varepsilon > 0$ and $\delta > 0$ the lines $x = \xi - \delta$, $x = \xi + \delta$, $y = f(\xi) - \varepsilon$, $y = f(\xi) + \varepsilon$, bound a rectangle centered in $(\xi, f(\xi))$, which we can now slide along the graph $y = f(x)$. Explain[5] geometrically what the implication in Definition 8.7 says, and compare to Exercise 8.2.

Theorem 8.10. *Let $f \in C([a, b])$. Then f is uniformly continuous on $[a, b]$.*

Proof of Theorem 8.10. As in the proof of Theorem 8.1 we argue by contradiction. So suppose that f is not uniformly continuous. Then the contraposition of the statement in Definition 8.10 holds, i.e.

$$\exists_{\varepsilon>0}\, \forall_{\delta>0}\, \exists_{x,\xi\in[a,b]} : |x-\xi| < \delta \quad \text{and} \quad |f(x)-f(\xi)| \geq \varepsilon.$$

Again this provides us with an $\varepsilon > 0$ and the possibility to choose $\delta > 0$ as we like. We choose $\delta = \frac{1}{n}$, with $n \in \mathbb{N}$ arbitrary, and conclude there exist sequences $x_n, \xi_n \in [a, b]$ for which it holds that

$$|x_n - \xi_n| < \frac{1}{n} \quad \text{and} \quad |f(x_n) - f(\xi_n)| \geq \varepsilon. \tag{8.5}$$

[3] Or any other closed bounded nonempty set in $\mathbb{R}$.

[4] No difference between $\forall_{\varepsilon>0}\, \forall_{\xi\in A}$ and $\forall_{\xi\in A}\, \forall_{\varepsilon>0}$!

[5] This nice explanation of *uniform* continuity is due to Thomas Rot.

Both sequences are bounded. As in the proof of Theorem 4.4 it is the Bolzano-Weierstrass Theorem[6] that gives the existence of a convergent subsequence x_{n_k} with limit $\bar{x} \in [a,b]$. The continuity of f then yields $f(x_{n_k}) \to f(\bar{x})$ as $k \to \infty$. But

$$|x_{n_k} - \xi_{n_k}| < \frac{1}{n_k} \leq \frac{1}{k}$$

implies that also $\xi_{n_k} \to \bar{x}$, so also $f(\xi_{n_k}) \to f(\bar{x})$ and therefore

$$f(x_{n_k}) - f(\xi_{n_k}) \to 0.$$

This happily contradicts (8.5) and completes the proof of Theorem 8.10. □

Proof of Theorem 8.6. Assume $f \in C([a,b])$. By Theorem 8.10 the function f is uniformly continuous. By now we are done with cosmetics, so let $\varepsilon > 0$ and apply Definition 8.7. Then $|f(x) - f(\xi)| < \varepsilon$ if $|x - \xi| < \delta$, $\delta > 0$ provided by the definition. Choose an equidistant[7] partition with

$$\frac{b-a}{N} < \delta,$$

it follows for M_k and m_k in (8.4) that $M_k - m_k < \varepsilon$ for all $k = 1, \ldots, N$. This is because m_k and M_k as defined in (7.2) are realised[8] as values of f in I_k, and I_k has length smaller than δ. But then it follows that

$$0 \leq \bar{S} - \underline{S} = \sum_{k=1}^{N} \underbrace{(M_k - m_k)}_{<\varepsilon}(x_k - x_{k-1}) < \varepsilon \sum_{k=1}^{N} (x_k - x_{k-1}) = \varepsilon(b-a).$$

Once again Theorem 7.2 completes a proof because $\varepsilon > 0$ was arbitrary[9]. □

8.2 Reflection: uniform epsilon statements

Definition 4.15 said that the sequence $f_n(x)$ converges to $f(x)$ as $n \to \infty$ with a choice of $N \in \mathbb{N}$ depending on $\varepsilon > 0$ but *independent of* x. This is why we speak of uniform convergence. We copy[10] the statement for $f_n, f : A \to \mathbb{R}$ and take

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N}\ \underbrace{\forall_{x\in A}}_{\text{uniform}}\ :\ |f_n(x) - f(x)| < \varepsilon \tag{8.6}$$

[6] Theorem 3.20.

[7] Or any other partition with $x_k - x_{k-1} < \delta$ for all $k = 1, \ldots, N$.

[8] Theorem 4.4 provides us with min- and maximizers.

[9] Should we mention the $(b-a)$-trick?

[10] Here A could be any non-empty set!

as the definition of $f_n \to f$ uniformly on A. Uniform convergence is stronger than pointwise convergence, which only says that

$$\underbrace{\forall_{x\in A}}_{\text{pointwise}} \ \forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N} \ : \ |f_n(x) - f(x)| < \varepsilon, \tag{8.7}$$

and allows N to *depend on both* $\varepsilon > 0$ *and* $x \in A$. Of course this can only weaken the statement made in (8.6) which has N *depending on* $\varepsilon > 0$ *only*.

Remark 8.11. *The uniform convergence statement (8.8) and the non-uniform pointwise convergence statement (8.9) differ by just one* $\forall$-$\exists$ *swap if we write them as*[11]

$$\forall_{\varepsilon>0}\ \underbrace{\exists_{N\in\mathbb{N}}\, \forall_{x\in A}}_{\textit{uniform}}\ \forall_{n\geq N} \ : \ |f_n(x) - f(x)| < \varepsilon, \tag{8.8}$$

and[12]

$$\forall_{\varepsilon>0}\ \underbrace{\forall_{x\in A}\, \exists_{N\in\mathbb{N}}}_{\textit{pointwise}}\ \forall_{n\geq N} \ : \ |f_n(x) - f(x)| < \varepsilon. \tag{8.9}$$

Indeed, $\forall_{x\in A}$ *and* $\exists_{N\in\mathbb{N}}$ *occur in different order in (8.8) and (8.9).*

You should compare Remark 8.11 to Remark 8.8. Recalling (7.9) we emphasise again that the stronger uniform statement (8.6) is equivalent to

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N} \ : \ \underbrace{d(f_n, f) = \sup_{x\in A} |f_n(x) - f(x)| \leq \varepsilon}_{\iff \forall_{x\in A}\,:\,|f_n(x)-f(x)|\leq\varepsilon}\,. \tag{8.10}$$

As indicated in (8.10), this is just (8.6) with $< \varepsilon$ replaced by $\leq \varepsilon$. *After all, the metric in* $B(A)$ *was chosen so as to make convergence of a sequence* f_n *in* $B(A)$ *equivalent*[13] *to uniform convergence on* A.

8.3 Uniform convergence and equicontinuity

We recall that we introduced the space $C([a, b])$ of continuous functions in Definition 4.5 and subsequently proved in Section 4.2 that it is a complete metric space with its metric defined in terms of the maximum norm. In Remark 4.17 we compared $C([a, b])$ to $\mathbb{R}$, and observed that the Bolzano-Weierstrass Theorem does not hold in $C([0, 1])$. A nice counter example is $f_n(x) = x^n$ in Exercise 4.10. The sequence f_n is bounded in $C([0, 1])$ but does not have a uniformly convergent subsequence.

[11] Compare (8.6) and (8.8): there is no difference between $\forall_{n\geq N}\, \forall_{x\in A}$ and $\forall_{x\in A}\, \forall_{n\geq N}$.
[12] Compare (8.7) and (8.9): also no difference between $\forall_{\varepsilon>0}\, \forall_{x\in A}$ and $\forall_{x\in A}\, \forall_{\varepsilon>0}$.
[13] Note again that only $f_n - f \in B(A)$ is needed to have $d(f_n, f)$ well defined.

We now re-adress this issue and formulate a condition on sequences in $C([a,b])$ that allows to prove that a bounded sequence satisfying this condition has a uniformly convergent subsequence. So let f_n be a sequence of functions defined on $[a,b]$ or any other nonempty subset A of $\mathbb{R}$. Then we can speak of continuity of f_n which is uniform is ξ, but also[14] of continuity which is simultaneously uniform is ξ and n. The following definition allows to formulate a Bolzano-Weierstrass type of statement in $C([a,b])$.

Definition 8.12. *Let $f_n : A \to \mathbb{R}$ be a sequence of functions. Then f_n is called* uniformly equicontinuous *on A if*

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\,\forall_{x,\xi\in A}\,\forall_{n\in\mathbb{N}} : \underbrace{|x-\xi|}_{d(x,\xi)} < \delta \implies \underbrace{|f_n(x)-f_n(\xi)|}_{d(f_n(x),f_n(\xi))} < \varepsilon.$$

Theorem 8.13. *(Arzéla-Ascoli) Let $f_n : [a,b] \to \mathbb{R}$ be a bounded sequence of uniformly equicontinuous functions. Then f_n has a convergent subsequence in $C([a,b])$ with limit $f \in C([a,b])$.*

Proof of Theorem 8.13. For (notational) convenience (only) we replace $[a,b]$ by $[0,1]$. A natural first step is try to define the limit function f. The sequence $f_n(0)$ is bounded in $\mathbb{R}$ and therefore has a convergent subsequence by the Bolzano-Weierstrass Theorem. Denote the limit by $f(0)$. Again by Theorem 3.20 this subsequence of f_n contains a further subsequence which converges in $x = 1$ as well. Denote the limit by $f(1)$. Along another further subsequence $f_n(\frac{1}{2})$ converges. The limit defines $f(\frac{1}{2})$. Repeating the argument we likewise define the values of a desired limit function f in $\frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8}$ and so on.

The indices[15] of all these subsequences are given by

$$\begin{array}{llllllll}
n_{11} & n_{12} & n_{13} & n_{14} & n_{15} & n_{16} & \dots & \text{for convergence in } 0,\\
n_{21} & n_{22} & n_{23} & n_{24} & n_{25} & n_{26} & \dots & \text{for convergence also in } 1,\\
n_{31} & n_{32} & n_{33} & n_{34} & n_{35} & n_{36} & \dots & \text{for convergence also in } \frac{1}{2},\\
n_{41} & n_{42} & n_{43} & n_{44} & n_{45} & n_{46} & \dots & \text{for convergence also in } \frac{1}{4},\\
n_{51} & n_{52} & n_{53} & n_{54} & n_{55} & n_{56} & \dots & \text{for convergence also in } \frac{3}{4},\\
n_{61} & n_{62} & n_{63} & n_{64} & n_{65} & n_{66} & \dots & \text{for convergence also in } \frac{1}{8},
\end{array}$$

[14] We won't consider pointwise equicontinuity.

[15] Have a look at the proof of Theorem 1.4.

a process we can continue *until every such dyadic*[16] *number in* $[0,1]$ *has occurred*. Each of these sequences is a subsequence of the previous sequence, and has the diagonal subsequence n_{kk} as a further subsequence.

It follows that the sequence F_k defined by $F_k = f_{n_{kk}}$ is a subsequence of f_n with the property that

$$F_k(a) = f_{n_{kk}}(a) \to f(a)$$

for every

$$a \in \mathcal{D} = \{0, 1, \frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8}, \dots\},$$

with the function $f : \mathcal{D} \to \mathbb{R}$ defined in the subsequence arguments above. In particular every $F_k(a)$ is a Cauchy sequence in $\mathbb{R}$.

In view of the completeness[17] of $C([0,1])$ it now suffices to show that the sequence F_k is a uniform Cauchy sequence. To do so we use that as a subsequence of f_n the sequence F_k is also equicontinuous. So let $\varepsilon > 0$ and apply Definition 8.12. Adapting the notation to the present context it says that

$$\exists_{\delta>0}\, \forall_{x,a\in A}\, \forall_{k\in\mathbb{N}} \,:\, |x-a| < \delta \implies |F_k(x) - F_k(a)| < \varepsilon.$$

We now choose $l \in \mathbb{N}$ with $2^l\delta > 1$ and estimate the difference of $F_k(x)$ and $F_m(x)$ for arbitrary $x \in [0,1]$ by

$$|F_k(x) - F_m(x)| \le \underbrace{|F_k(x) - F_k(a)|}_{<\varepsilon} + |F_k(a) - F_m(a)| + \underbrace{|F_m(a) - F_m(x)|}_{<\varepsilon},$$

in which for every $x \in [0,1]$ a number

$$a \in \mathcal{D}_l = \{0, \frac{1}{2^l}, \frac{2}{2^l}, \frac{3}{2^l}, \dots, 1\}$$

with

$$|x-a| < \frac{1}{2^l} < \delta \quad \text{is chosen to ensure} \quad |F_k(x) - F_k(a)| < \varepsilon.$$

We then choose $N \in \mathbb{N}$ such that $|F_k(a) - F_m(a)| < \varepsilon$ for all $k, m \ge N$, and for all $a \in \mathcal{D}_l$. This is possible because every $F_k(a)$ is a Cauchy sequence and $\mathcal{D}_l$ is a finite set. It follows that

$$|F_k(x) - F_m(x)| < 3\varepsilon$$

for all $k, m \ge N$. Since N is independent of x and $\varepsilon > 0$ was arbitrary, a usual 3-trick establishes that F_k has the property (4.7) stated in the proof of Theorem 4.12, namely that it is a uniform Cauchy sequence. Theorem 4.12, which stated the completeness of $C([a,b])$, then completes the proof. $\square$

[16] What's that? Wikipedia!

[17] See Theorem 4.12.

8.4 Extra: more on continuity and integration

Recall that Theorem 1.4 was preceded by a "proof" in which we argued by contradiction to conclude that $A = \mathbb{R}$ would have the property that, given any $\varepsilon > 0$, we can cover A with a countable union of say closed intervals, i.e.

$$A \subset \cup_{n\in\mathbb{N}}[a_n, b_n],$$

such that

$$\sum_{n\in\mathbb{N}}(b_n - a_n) < \varepsilon.$$

This conclusion is in fact the very definition of what it means for a subset A of $\mathbb{R}$ to have zero length, i.e. zero 1-dimensional (Lebesgue) measure. Without proof we state a fundamental theorem.

Theorem 8.14. *Let $f : [a, b] \to \mathbb{R}$ be a bounded function. Denote the set of points in which f is not continuous by A. Then $f \in \mathrm{RI}([a, b])$ if and only if A is set of measure zero.*

For functions $f : [a, b] \to \mathbb{R}$ we were able to avoid continuity issues for many of our integrational purposes by using the ordering of the real numbers. For X-valued functions continuity is more important.

Theorem 8.15. *Let X be a complete metric vector space and $f : [a, b] \to X$ be a continuous function. Denote the norm in X by the usual bars, i.e. $|x|$ is the norm of $x \in X$. Then there exists a unique $J \in X$ such that for every $\varepsilon > 0$ a $\delta > 0$ exists such that, for every partition P as in (6.8) and every choice of intermediate points ξ_k with*

$$a = x_0 \le \xi_1 \le x_1 \le \xi_2 \cdots \le \xi_N \le x_N = b,$$

it holds that

$$S = \sum_{k=1}^{N} f(\xi_k)(x_k - x_{k-1})$$

satisfies

$$|S - J| < \varepsilon$$

provided

$$\max_{k=1,\ldots,N}(x_k - x_{k-1}) < \delta.$$

We write

$$J = \int_a^b f$$

and we have

$$|J| \le \int_a^b |f(x)|\, dx.$$

Exercise 8.16. Not so easy. Give a proof of Theorem 8.15 for the case that $X = \mathbb{R}$ which does not rely on lower and upper sums. Hint: try a proof for the statement with only right endpoint sums for equidistant partitions as in the proof of Theorem 6.8 first. If all goes well you find the same[18] J as well as a proof for $f : [a, b] \to X$ continuous. Then think about such sums for other partitions and other choices of the points in the intervals of the partition.

8.5 Extra: global monotone inverse function theorem

The material in this section does not fit in with our overall philosophy that we discuss theory for $y = f(x)$ with $x, y \in \mathbb{R}$ that generalises to a context in which $x \in X$ and $y \in Y$. The result to remember from this section is that a continuous strictly monotone real valued function f defined on some interval I has a range $J = f(I)$ which is itself an interval, and that there exists a unique continuous strictly monotone real valued function g defined on J with range I such that

$$y = f(x) \iff x = g(y) \tag{8.11}$$

for all $x \in I$ and $y \in J$. Thus (8.11) defines a bijection between I and J. Formulated in Theorem 8.19 for open intervals I and J only, the proof relies crucially on Theorem 8.18 below, which has the simple[19] but important statement in Theorem 8.17 as a special case.

Theorem 8.17. *Let $a, b \in \mathbb{R}$ with $a < b$, and let $f : [a, b] \to \mathbb{R}$ be continuous. If $f(a)f(b) < 0$ then f has a zero in (a, b), i.e. there exists $x_0 \in (a, b)$ such that $f(x_0) = 0$.*

Theorem 8.17 can be restated as the *intermediate value theorem*:

Theorem 8.18. *Let I be an open interval in $\mathbb{R}$ and $f : I \to \mathbb{R}$ a continuous function. For $a, b \in I$ with $a < b$ let*

$$f([a, b]) = \{f(x) : a \leq x \leq b\}$$

be the image of $[a, b]$ under f. Then

$$f(a) < f(b) \implies [f(a), f(b)] \subset f([a, b]),$$

and

$$f(a) > f(b) \implies [f(b), f(a)] \subset f([a, b]).$$

[18] Why?

[19] By now obvious?

Proof. To prove this statement assume first that $f(a) < c < f(b)$. Then

$$\xi = \sup\{x \in [a,b] : \ f(x) < c\}$$

exists as the supremum of a bounded set which contains a. Can it be that $f(\xi) < c$? If so then $\xi < b$ because $f(b) > c$. Choose $\varepsilon > 0$ with $\varepsilon < c - f(\xi)$ and apply the ε-δ statement of continuity in (8.1). Then

$$f(x) - f(\xi) \leq |f(x) - f(\xi)| < \varepsilon < c - f(\xi)$$

for all $x \in I$ with $|x - \xi| < \delta$. But then $f(x) < c$ for all such x, which contradicts that ξ is an upper bound.

Can it be that $f(\xi) > c$? Choose $\varepsilon > 0$ with $\varepsilon < f(\xi) - c$ and apply (8.1). Then $f(\xi) - f(x) \leq |f(x) - f(\xi)| < \varepsilon < f(\xi) - c$ for all $x \in I$ with $|x - \xi| < \delta$. But then $f(x) > c$ for all such x. This makes $\xi - \delta$ an upper bound and contradicts that ξ is the lowest upper bound. Thereby the proof for $f(a) < f(b)$ is complete. For $f(a) > f(b)$ the proof is of course similar. □

Theorem 8.19. *Let I be an open interval in $\mathbb{R}$ and $f : I \to \mathbb{R}$ a continuous function with the property that*

$$\forall_{a,b \in I} \quad a < b \implies f(a) < f(b),$$

i.e. f is strictly increasing on I. Then

$$J = f(I) = \{f(x) : \ x \in I\}$$

is also an open interval and the equation $f(x) = y$ defines x as $g(y)$ for every $y \in J$, with the function $g : J \to \mathbb{R}$ continuous, strictly increasing i.e.

$$\forall_{c,d \in J} \quad c < d \implies g(c) < g(d),$$

and

$$I = g(J) = \{g(y) : \ y \in J\}.$$

Proof. By definition $f(x) = y$ has a solution in I for every $y \in f(I)$. The strict monotonicity of f makes that solution unique and thereby settles the existence of $g : J \to \mathbb{R}$ with the same strict monotonicity property. We next show that J is an open interval.

Let $c, d \in J$ with $c < d$. Then $c = f(a)$ and $d = f(b)$ for some a and b in I, and $[c,d] \subset J$ by Theorem 8.18. Thus J is an interval. Also, if $y_0 \in J$ then $y_0 = f(x_0)$, $x_0 \in I$ and $[x_0 - \delta_0, x_0 + \delta_0] \subset I$ for some

$\delta_0 > 0$. Thus $[f(x_0 - \delta_0), f(x_0 + \delta_0)] \subset J$ so y_0 is an interior point because $f(x_0 - \delta_0) < f(x_0) < f(x_0 + \delta_0)$. We conclude that J is an open interval.

It remains to prove the continuity of g, so let $y_0 = f(x_0)$ and $\varepsilon > 0$. It is no limitation to choose $\varepsilon < \delta_0$, δ_0 as just above. Then

$$(f(x_0 - \varepsilon), f(x_0 + \varepsilon)) \subset [f(x_0 - \delta_0), f(x_0 + \delta_0)] \subset J$$

and we can choose $\delta > 0$ such that

$$f(x_0-\delta_0) < \underbrace{f(x_0 - \varepsilon)}_{\substack{\downarrow g \\ x_0-\varepsilon}} < y_0-\delta < \underbrace{f(x_0) = y_0}_{\substack{\downarrow g \\ x_0=g(y_0)}} < y_0+\delta < \underbrace{f(x_0 + \varepsilon)}_{\substack{\downarrow g \\ x_0+\varepsilon}} < f(x_0+\delta_0),$$

whence

$$g((y_0 - \delta, y_0 + \delta)) \subset (g(y_0) - \varepsilon, g(y_0) + \varepsilon).$$

This completes the proof. □

8.6 Exercises

Exercise 8.20. Let $f(x) = 2x + 1$. Prove directly from the definition that f is uniformly continuous on $\mathbb{R}$.

Exercise 8.21. Let $f(x) = x^2$ and $A = (0, 1)$. Prove directly from the definition that f is uniformly continuous on A. Is f uniformly continuous on $\mathbb{R}$?

Exercise 8.22. Let $f(x) = \frac{1}{x}$ and $A = (1, \infty)$. Prove that f is uniformly continuous on A. Is f uniformly continuous on $(0, 1)$?

Exercise 8.23. Let $f : A \to \mathbb{R}$ be Lipschitz continous. Prove that f is uniformly continuous.

Exercise 8.24. Let $A \subset \mathbb{R}$ and let $f : A \to \mathbb{R}$ be uniformly continuous. Suppose that x_n is a Cauchy sequence in A. Prove that $f(x_n)$ is also a Cauchy sequence.

Exercise 8.25. Let $a, b \in \mathbb{R}$ with $a < b$, and let $f : (a, b) \to \mathbb{R}$ be uniformly continuous. Then there exists a unique $\bar{f} \in C([a, b])$ such that $f(x) = \bar{f}(x)$ for all $x \in (a, b)$. Hint: use Exercise 8.24 to define $\bar{f}(a)$ and $\bar{f}(b)$.

Exercise 8.26. Recall that Theorem 7.12 says that $\mathrm{RI}([a, b])$ is a complete metric vector space. Why is $C([a, b])$ a closed linear subspace of $\mathrm{RI}([a, b])$?

Exercise 8.27. Examine the function f defined by

$$f(x) = \frac{x}{1 + x}.$$

What is the largest open interval I containing 0 to which you can apply Theorem 8.19? Specify J and compute $g(y)$. What is J if $I = (0, \infty)$?

Exercise 8.28. Formulate Theorem 8.19 for strictly decreasing functions.

Exercise 8.29. Let $f : [a, b] \to \mathbb{R}$ be a bounded integrable function, assume that $R_f = \{f(x) : a \leq x \leq b\} \subset [c, d]$ and let $F : [c, d] \to \mathbb{R}$ be continuous. Prove that $F \circ f$ is integrable on $[a, b]$. Hint[20]: approximate F uniformly with a sequence of Lipschitz continuous functions and then use both Theorem 7.5 and Theorem 7.12.

Exercise 8.30. Let $f_n : [-1, 1] \to \mathbb{R}$ be a bounded sequence of integrable functions, and let $F : \mathbb{R} \to \mathbb{R}$ be continuous. Suppose that

$$f_n(x) = \int_0^x F(f_n(s))\, ds$$

holds for all $x \in [-1, 1]$ and all $n \in \mathbb{N}$. Prove that f_n has a uniformly convergent subsequence. Hint: Theorem 8.13. NB. The right hand side exists in view of Exercise 8.29.

[20] Harold drew my attention to a different rather clever proof due to Rudin which only uses Theorem 7.2. For $\varepsilon > 0$ choose $\delta > 0$ according to the definition of uniform continuity of F and then a partition for which the lower and upper sums for $\int_a^b f$ with $m_k(x_k - x_{k-1})$ and $M_k(x_k - x_{k-1})$ differ by at most δ^2. Then distinguish between the bad k for which $M_k - m_k \geq \delta$ and the good k for which $M_k - m_k < \delta$. Estimate the sum of $x_k - x_{k-1}$ over the bad k in terms of what you then know. Use the boundedness of f to get a final estimate for the sum of $(M_k - m_k)(x_k - x_{k-1})$ over all k. Then complete the proof.

Exercise 8.31. Let $F_n : \mathbb{R} \to \mathbb{R}$ be a a sequence of continuous functions which is bounded in the sense that there exists $M > 0$ such that $|F_n(y)| \leq M$ for all $n \in \mathbb{N}$ and all $y \in \mathbb{R}$. Suppose that $f_n : [-1, 1] \to \mathbb{R}$ is a sequence of integrable functions such that

$$f_n(x) = \int_0^x F_n(f_n(s))\, ds$$

holds for all $x \in [-1, 1]$ and all $n \in \mathbb{N}$. Prove that f_n has a uniformly convergent subsequence. Hint: Theorem 8.13.

Exercise 8.32. (continued) Suppose that $F_n \to F$ uniformly on $[-M, M]$ and let f be a limit function of a uniformly convergent subsequence as in Exercise 8.31. Prove that

$$f(x) = \int_0^x F(f(s))\, ds$$

holds for all $x \in [-1, 1]$. Hint: first show that $|f_n(x)| \leq M|x|$ and then use

$$|F_n(f_n(s)) - F(f(s))| \leq |F_n(f_n(s)) - F_n(f(s))| + |F_n(f(s)) - F(f(s))|$$

to apply Theorem 7.12.

Exercise 8.33. (continued) Let F_n and F be as in Exercise 8.32. Assume that every $F_n : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous, without further assumptions on the Lipschitz constants L_n. Prove that the integral equation

$$f(x) = \int_0^x F(f(s))\, ds$$

has an integrable solution $f : [-1, 1] \to \mathbb{R}$. Hint: recall Exercise 7.42 where you showed that the integral equation has a unique solution $f : \mathbb{R} \to \mathbb{R}$ in the class of integrable functions if $F : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous, which is not an assumption on F here.

Exercise 8.34. (continued) Assume that a bounded sequence of Lipschitz continuous functions $F_n : \mathbb{R} \to \mathbb{R}$ has the property that $F_n \to F$ uniformly on every bounded interval. Prove that the integral equation

$$f(x) = \int_0^x F(f(s))\, ds$$

has a solution $f : \mathbb{R} \to \mathbb{R}$.

Exercise 8.35. (continued) Let $F : \mathbb{R} \to \mathbb{R}$ be a bounded continuous function. Prove that the integral equation

$$f(x) = \int_0^x F(f(s))\, ds$$

has a solution $f : \mathbb{R} \to \mathbb{R}$. Hint: show that there exists a sequence F_n as in Exercise 8.34.

9 Differential calculus for power series

You will be familiar with the formula

$$f'(a) = \lim_{h\to 0} \frac{f(a+h)-f(a)}{h} = \lim_{x\to a} \frac{f(x)-f(a)}{x-a}. \tag{9.1}$$

This is the usual definition of the derivative $f'(a)$ of a real valued function f of a real variable x in a point $x = a$ on the real line. If the limit of the difference quotient in (9.1) exists it is called the *differential quotient* of f in $x = a$. Differential quotients are sometimes formally denoted as a fractions[1] with denominator df and numerator dx, just like difference quotients are denoted as

$$\frac{f(x+\Delta x)-f(x)}{\Delta x} = \frac{\Delta f}{\Delta x}$$

with $\Delta x = h \neq 0$. Notations to be handled with care or simply avoided.

For the simplest examples we consider first, monomials such as

$$f_{42}(x) = x^{42},$$

it turns out that the difference quotient is defined for $x = a$ as well[2]. For instance, for $x \neq a$ it holds that

$$\frac{x^{42}-a^{42}}{x-a} = x^{41} \underbrace{+\cdots+}_{\text{Exercise 1.10!}} a^{41},$$

but the right hand side is equal to $42a^{41}$ for $x = a$. Whatever the value of a, it must thus follow that f_{42} is differentiable in a, with

$$f'_{42}(a) = 42a^{41}.$$

In what follows we will avoid limits of difference quotients and think of *differentiation as a method to*[3] *approximate a given (nonlinear) function* f by a linear one, i.e. to write

$$f(x) \approx f(a) + A(x-a) = Ax + B.$$

We want to choose A and B so that

$$R_a(x) = f(x) - Ax - B$$

[1] This also suggests to write $df = f'(x)dx$.

[2] Parafrasing: as Jaap Murre stipulated, who would need a limit concept here?

[3] In some sense best.

is as small as possible near $x = a$. We write

$$f(x) = f(a) + A(x - a) + R_a(x)$$

and aim to identify $f'(a)$ as the *unique value* of A for which the remainder term $R_a(x)$ has a *smallness property that fails for other choices* of A. Below we derive this property by purely algebraic manipulations starting from the difference quotient

$$\frac{f(x) - f(a)}{x - a}$$

in (9.1). Hope you don't mind this long introduction, which was really written for highschool students and their teachers in my booklet with Ronald Meester.

9.1 Linear approximations of monomials

So consider such a *difference quotient* for the function f_7 defined by $f_7(x) = x^7$. A little algebra[4] in Chapter 1 told you that

$$\frac{x^7 - a^7}{x - a} = x^6 + ax^5 + a^2x^4 + a^3x^3 + a^4x^2 + a^5x + a^6,$$

which you rewrote as[5]

$$x^7 = a^7 + (x^6 + ax^5 + a^2x^4 + a^3x^3 + a^4x^2 + a^5x + a^6)(x - a) = \tag{9.2}$$

$$\underbrace{a^7 + 7a^6\,(x - a)}_{Ax + B} + \quad \underbrace{(x^5 + 2ax^4 + 3a^2x^3 + 4a^3x^2 + 5a^4x + 6a^5)\,(x - a)^2}_{\text{remainder term}}.$$

The *particular choice*

$$A = 7a^6, \quad B = -6a^7 \tag{9.3}$$

followed from putting $x = a$ in the 7 terms of the[6] prefactor in the second term on the right hand side of (9.2). Of course you already "knew" that $f_7'(x) = 7x^6$ so you recognise $7a^6$ as $f_7'(a)$ computed via (9.1).

The first two terms can be seen as the *best approximation* of the form

$$Ax + B = 7a^6x - 6a^7$$

[4] Long division for instance.

[5] See Exercise 1.20.

[6] Typographically large....

to $f_7(x) = x^7$ when x is close to a. This is because the above values of A and B appear as the *only choice*[7] which makes the *resulting remainder term*[8] contain a factor $(x-a)^2$.

Moreover, the prefactor in the remainder term under (9.2) is easily estimated if we assume that x and a are contained in a fixed interval $[-r, r]$. For example, if

$$|x| \leq r \quad \text{and} \quad |a| \leq r,$$

this prefactor is estimated by

$$(1+2+3+4+5+6)\, r^5 = \frac{7 \times 6}{2}\, r^5.$$

You will not be surprised that (9.2) and its splitting in a linear term and such a remainder term generalise to general $n \in \mathbb{N}$.

Theorem 9.1. *For $n \in \mathbb{N}$ and $x, a \in \mathbb{R}$ let $R_{an}(x)$ be defined by*

$$x^n = a^n + na^{n-1}(x-a) + R_{an}(x),$$

and let $r > 0$. Then

$$|R_{an}(x)| \leq \underbrace{\frac{n(n-1)}{2} r^{n-2}}_{r\text{-dependent constant}} (x-a)^2$$

for all $x, a \in [-r, r]$.

Exercise 9.2. You may guess a nice expression for $R_{an}(x)$ from (9.2). Guess right, prove what you guessed for all $n \in \mathbb{N}$, and then prove Theorem 9.1.

9.2 Linear approximations of polynomials

Let $\alpha_0, \alpha_1, \alpha_2, \ldots$ be a sequence of real coefficients. Then for the polynomials

$$p_k(x) = \sum_{n=0}^{k} \alpha_n x^n = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_k x^k$$

[7] Of course both A and B depend on a.

[8] A polynomial in x with coefficients depending on the choice of a, A, B.

of degree $k \geq 2$ the story is quite the same as in Section 9.1. Simply multiply both sides of the equality and inequality in Theorem 9.1 by α_n and take the sum over n. With some care for $n = 0, 1, 2$ it follows that

$$p_k(x) = p_k(a) + \underbrace{\sum_{n=1}^{k} n\alpha_n a^{n-1}\,(x-a)}_{\text{linear approximation}} + \underbrace{\sum_{n=2}^{k} \alpha_n R_{an}(x)}_{\text{remainder term}}, \tag{9.4}$$

in which for all $x, a \in [-r, r]$ the remainder term satisfies

$$|\underbrace{\sum_{n=2}^{k} \alpha_n R_{an}(x)}_{\text{remainder term}}| \leq \underbrace{\sum_{n=2}^{k} |\alpha_n| \frac{n(n-1)}{2} r^{n-2}}_{r\text{-dependent constant}} (x-a)^2. \tag{9.5}$$

As before

$$\underbrace{p_k(a) + \sum_{n=1}^{k} n\alpha_n a^{n-1}}_{p_k'(a)}\,(x-a)$$

is the *best linear approximation* of $p_k(x)$ near $x = a$, in which we recognise the value of derivative of p_k in a as the coefficient of $(x-a)$.

9.3 Power series: the fundamental theorem

The step from polynomials to power series like

$$p(x) = 1 + 2x + 3x^2 + \cdots \tag{9.6}$$

is a small step for the text editor if we use the illuminating dots notation. Recall from calculus that every power series

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots$$

has a critical radius R. For $x \in \mathbb{R}$ with $|x| < R$ the power series is absolutely convergent, for $|x| > R$ the individual terms are an unbounded sequence and therefore there is no way to give meaning to the sum. The behaviour for $|x| = R$ may be complicated but is for later worries.

Theorem 9.3. *Every power series*

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n$$

with $\alpha_n \in \mathbb{R}$ *for* $n \in \mathbb{N}_0$ *has a* radius of convergence $R \in [0, \infty]$ *such that the series is absolutely convergent for all* $x \in \mathbb{R}$ *with* $|x| < R$*. For such* x *it holds that*

$$p'(x) = \sum_{n=1}^{\infty} n\alpha_n x^{n-1} = \sum_{n=0}^{\infty} (n+1)\alpha_{n+1} x^n,$$

in which p' *is the* derivative *of* p *on* $\{x \in \mathbb{R} : |x| < R\}$ *in the usual sense of limits of difference quotients, namely*

$$p'(a) = \lim_{x \to a} \frac{p(x) - p(a)}{x - a}$$

for every a *with* $|a| < R$*. The power series for* $p'(x)$ *is also absolutely convergent for all* $x \in \mathbb{R}$ *with* $|x| < R$*, and the convergence of both series is uniform on every* $\{x \in \mathbb{R} : |x| \leq r\}$ *with* $0 < r < R$*. For* $x \in \mathbb{R}$ *with* $|x| > R$ *the terms in both series for* $p'(x)$ *and* $p(x)$ *are unbounded in* n *and none of the two series converge.*

Proof. We continue from (9.5). If for some $r > 0$ it holds that

$$C_r := \sum_{n=2}^{\infty} |\alpha_n| \frac{n(n-1)}{2} r^{n-2} < \infty, \tag{9.7}$$

we can let $k \to \infty$ in (9.4). Indeed, it then follows from Exercises 3.37 and 3.39 that the sums

$$\sum_{n=0}^{\infty} \alpha_n x^n, \quad \sum_{n=1}^{\infty} \alpha_n a^n, \quad \sum_{n=1}^{\infty} n\alpha_n a^{n-1}$$

exist for all $x, a \in [-r, r]$ because

$$1 \leq n \leq \frac{n(n-1)}{2}$$

for $n \geq 2$, and so does the sum

$$R_a(x) = \sum_{n=2}^{\infty} \alpha_n R_{an}(x).$$

Thus (9.7) allows to take the limit $k \to \infty$ in (9.4) and (9.5) to obtain[9]

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n = p(a) + \underbrace{\sum_{n=1}^{\infty} n\alpha_n a^{n-1}}_{A} (x - a) + R_a(x) \tag{9.8}$$

[9]The convergence is in fact uniform on $[-r, r]$, why?

with

$$|R_a(x)| \leq C_r(x-a)^2 \tag{9.9}$$

for all $x, a \in [-r, r]$. As before we observe that

$$A = \sum_{n=1}^{\infty} n\alpha_n a^{n-1} \tag{9.10}$$

is the only value of A for which

$$p(x) = p(a) + A(x-a) + R_a(x)$$

holds in combination with an estimate of the form (9.9) and a constant which depends only on r. The difference quotient in (9.1) with f replaced by p then evaluates as

$$\frac{p(x)-p(a)}{x-a} = A + \frac{R_a(x)}{x-a},$$

and (9.9) suffices to conclude from (9.7,9.8) that

$$\lim_{x\to a} \frac{p(x)-p(a)}{x-a} = A \tag{9.11}$$

as given by (9.10).

To conclude we note that the r-values for which (9.7) holds form an interval

$$\{r \geq 0 : \sum_{n=1}^{\infty} n^2|\alpha_n| r^n < \infty\}$$

which contains $r = 0$. The only possibilities for this interval are

$$\{0\},\ [0, R),\ [0, R],\ [0, \infty),$$

with $R > 0$ in the second and third case, and $R = \infty$ and $R = 0$ in the extreme fourth and first case. This completes the proof of Theorem 9.3, except for the statement about $|x| > R$, which follows from Exercise 9.4. □

Exercise 9.4. Suppose $R < \infty$ and let $x_0 \in \mathbb{R}$ with $|x_0| > R$. Assume the terms in $p(x_0)$ form a bounded sequence indexed by n. Derive a contradiction by showing that both $p(x)$ and $p'(x)$ are then absolutely convergent for every $x \in \mathbb{R}$ with $|x| < |x_0|$.

Exercise 9.5. Show that R is characterised by saying that $a_n x^n$ is an unbounded sequence if $|x| > R$ and a sequence converging to 0 if $|x| < R$.

Remark 9.6. *The limit statement (9.11) is equivalent to saying that*

$$\lim_{x\to a}\frac{R_a(x)}{x-a}=0. \tag{9.12}$$

This means that for every $\varepsilon>0$ there exists $\delta>0$ such that

$$|R_a(x)|<\varepsilon|x-a| \quad \text{if} \quad |x-a|<\delta, \tag{9.13}$$

a statement much weaker than the statement in (9.9). It will be used in Chapter 10 to define differentiability of functions not given by power series.

Exercise 9.7. The intervals

$$I_k:=\{r\geq 0:\sum_{n=1}^{\infty}n^k|\alpha_n|r^n \text{ exists}\}$$

don't change much if we vary $k\in\mathbb{N}$. It is clear that

$$I_1\supset I_2\supset I_3\supset\cdots,$$

but you should prove the existence of $R\in[0,\infty]$ such that for every $k\in\mathbb{N}$ either $I_k=[0,R)$ or $I_k=[0,R]$. Give examples of $R=0$, $R=1$ and $R=\infty$.

9.4 Extra: Taylor's formula for power series

We substitute $x=x_0+h$ in

$$p(x)=\sum_{n=0}^{\infty}\alpha_n x^n. \tag{9.14}$$

Changing the order of summation[10] we find

$$\begin{aligned}p(x_0+h)&=\sum_{n=0}^{\infty}\alpha_n(x_0+h)^n=\sum_{n=0}^{\infty}\alpha_n\sum_{k=0}^{n}\binom{n}{k}x_0^{n-k}h^k\\&=\sum_{n=0}^{\infty}\sum_{k=0}^{n}\alpha_n\frac{n(n-1)\dots(n-k+1)}{k!}x_0^{n-k}h^k\\&=\sum_{k=0}^{\infty}\sum_{n=k}^{\infty}\alpha_n\frac{n(n-1)\dots(n-k+1)}{k!}x_0^{n-k}h^k\end{aligned}$$

[10] This section relies on Section 3.8 but we will not expand on this here.

$$= \sum_{k=0}^{\infty} \frac{1}{k!} \underbrace{\sum_{n=k}^{\infty} \alpha_n n(n-1)\dots(n-k+1)\, x_0^{n-k}}_{p^{(k)}(x_0)} h^k$$

$$= \sum_{k=0}^{\infty} \frac{p^{(k)}(x_0)}{k!}\, h^k,$$

i.e.

$$p(x) = p(x_0 + h) = \sum_{n=0}^{\infty} \frac{p^{(n)}(x_0)}{n!}\, h^n = \sum_{n=0}^{\infty} \frac{p^{(n)}(x_0)}{n!}\, (x - x_0)^n. \tag{9.15}$$

In this form the power series is called a *Taylor series*. Do note the special case $x_0 = 0$ and $h = x$,

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n = \sum_{n=0}^{\infty} \frac{p^{(n)}(0)}{n!} x^n,$$

which is called a *Maclaurin series*.

Exercise 9.8. Let R be the radius of convergence of the power series $P(x)$. Show that (9.15) holds for all x_0 and h in $\mathbb{R}$ with $|x_0| + |h| < R$, as the sum of an absolutely convergent series. Hint: recall the concept of unconditional convergence, see Section 3.8.

Remark 9.9. *Everything we did for the differentiation of power series in (9.17) also works for (Laurent series)*

$$L(x) = \sum_{n=-\infty}^{\infty} \alpha_n x^n = \cdots + \frac{\alpha_{-2}}{x^2} + \frac{\alpha_{-1}}{x} + \alpha_0 + \alpha x + \alpha_2 x^2 + \cdots,$$

with $|x|$ not too large for the positive exponents and $|x|$ not too small for the negative exponents. Start with e.g.

$$\frac{1}{x^7} = \frac{1}{a^7} - \frac{7}{a^8}(x - a) + R_a(x),$$

figure out what $R_a(x)$ is, and you're in business.

9.5 Power series solutions of differential equations

We can solve linear differential equations for power series (9.14), for instance

$$p'(x) = p(x), \tag{9.16}$$

with boundary condition $p(0) = 1$. Let us try to find a solution of the form

$$p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \cdots ,$$

which may make sense for $|x| < R$, R hopefully positive. Provided $|x| < R$ it follows from Theorem 9.3 that

$$p'(x) = \alpha_1 + 2\alpha_2 x + 3\alpha_3 x^2 + 4\alpha_4 x^3 + \cdots ,$$

and so

$$p'(x) - p(x) = (\alpha_1 - \alpha_0) + (2\alpha_2 - \alpha_1)x + (3\alpha_3 - \alpha_2)x^2 + (4\alpha_4 - \alpha_3)x^3 + \cdots .$$

This can only be zero for all $x \in \mathbb{R}$ if

$$0 = \alpha_1 - \alpha_0 = 2\alpha_2 - \alpha_1 = 3\alpha_3 - \alpha_2 = 4\alpha_4 - \alpha_3 = \cdots ,$$

and from $\alpha_0 = p(0) = 1$ it then follows that

$$\alpha_1 = 1,\ \alpha_2 = \frac{1}{2},\ \alpha_3 = \frac{1}{2}\frac{1}{3},\ \alpha_4 = \frac{1}{2}\frac{1}{3}\frac{1}{4},\ \cdots ,\alpha_n = \frac{1}{n!}.$$

Thus we encounter a function we have seen before, namely in Exercise 6.23.

Theorem 9.10. *Let $r > 0$. The only possible power series that can satisfy $p'(x) = p(x)$ for all $x \in \mathbb{R}$ with $|x| < r$, and have $p(0) = 1$, is*

$$p(x) = \exp(x) := \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + \frac{x^6}{720} + \cdots .$$

In fact this power series converges for all $x \in \mathbb{R}$, and therefore satisfies $p'(x) = p(x)$ for all $x \in \mathbb{R}$, as well as $p(0) = 1$.

Exercise 9.11. Prove that $\exp(x)$ has $R = \infty$ and you have solved your first differential equation[11]. Hint: show for $N \in \mathbb{N}$ that

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^N}{N!} + R_N(x),$$

[11]No other functions can satisfy $f(0) = 1$ and $f'(x) = f(x)$, why is not clear yet.

in which

$$R_N(x) = \frac{x^N}{N!}\left(\frac{x}{N+1} + \frac{x^2}{(N+1)(N+2)} + \cdots\right)$$

is estimated by

$$|R_N(x)| \le \frac{|x|^N}{N!}\left(\frac{|x|}{N+1} + \left(\frac{|x|}{N+1}\right)^2 + \left(\frac{|x|}{N+1}\right)^3 + \cdots\right) = \frac{|x|^N}{N!}\frac{|x|}{N+1-|x|}$$

if $N+1 > |x|$.

Definition 9.12. *Let $a \in \mathbb{R}$. We say that $f(x) \to 0$ as $x \to \infty$ for a function $f : [a, \infty) \to \mathbb{R}$ if*

$$\forall_{\varepsilon>0}\, \exists_{\xi\in\mathbb{R}}\, \forall_{x\in\mathbb{R}} \quad x > \xi \implies |f(x)| < \varepsilon.$$

.

Exercise 9.13. Show for every fixed $n \in \mathbb{N}$ that

$$\frac{x^n}{\exp(x)} \to 0 \quad \text{as} \quad x \to \infty.$$

This is the standard limit that says that $\exp(x)$ beats every power of x as $x \to \infty$.

Theorem 9.14. *Let $r > 0$. The only possible power series that can satisfy $p''(x) + p(x) = 0$ for all $x \in \mathbb{R}$ with $|x| < r$, and have $p(0) = 0$ and $p'(0) = 1$, is*

$$p(x) = \sin x = x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \cdots.$$

In fact this power series converges for all $x \in \mathbb{R}$, so it satisfies

$$\begin{cases} p''(x) + p(x) = 0 \quad \textit{for all } x \in \mathbb{R}; \\ p(0) = 0 \ \textit{ and } \ p'(0) = 1 \ \textit{ in } x = 0. \end{cases}$$

Exercise 9.15. Write $p(x)$ using the sum notation and prove Theorem 9.14. Let $\cos x = p'(x)$. What is the derivative of cos?

At this point we don't know yet that $\exp(x)$, $\sin x$, $\cos x$ are what they should be. One way to verify what is and will ever be is to check all the formulas by brute force calculation. For instance:

Exercise 9.16. Show for all $x \in \mathbb{R}$ that

$$\cos^2 x + \sin^2 x = 1,$$

by substituting the power series for $\cos x$ and $\sin x$ and working out the squares.

Exercise 9.17. In Exercise 9.16 you may have realised that the square of a power series is also a power series with the same radius of convergence. Now let

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n \quad \text{and} \quad q(x) = \sum_{n=0}^{\infty} \beta_n x^n$$

be two power series. Theorem 9.3 gives R for $p(x)$ and S for $q(x)$.

a) Let a_n and b_n be sequences indexed by $n \in \mathbb{N}_0$, and

$$c_n = \sum_{k=0}^{n} a_k b_{n-k}.$$

If

$$\sum_{n=0}^{\infty} |a_n| < \infty \quad \text{and} \quad \sum_{n=0}^{\infty} |b_n| < \infty,$$

then

$$\sum_{n=0}^{\infty} |c_n| < \infty,$$

and

$$\left(\sum_{n=0}^{\infty} a_n\right)\left(\sum_{n=0}^{\infty} b_n\right) = \sum_{n=0}^{\infty} c_n = a_0b_0 + (a_0b_1 + a_1b_0) + (a_0b_2 + a_1b_1 + a_2b_0) + \cdots,$$

a statement we should have proved in Section 3.8 really. Apply this statement to show that

$$s(x) = p(x)q(x) = \alpha_0\beta_0 + (\alpha_1\beta_0 + \alpha_0\beta_1)x + (\alpha_2\beta_0 + \alpha_1\beta_1 + \alpha_0\beta_2)x^2 + \cdots$$

is also a power series, with radius of convergence at least equal to the minimum of R and S.

b) Then multiply (9.8) by the corresponding expression for $q(x)$ and prove that

$$s'(a) = p'(a)q(a) + p(a)q'(a)$$

for every a with $|a| < R$ and $|a| < S$.

c) Much easier, show that the same statement holds for the sum $t(x) = p(x)+q(x)$ with $t'(a) = p'(a) + q'(a)$.

d) Prove the equality in Exercise 9.16. Hint: you need Theorem 9.14 and Exercise 9.15 to conclude.

Exercise 9.18. Write down the power series solution of the differential equation

$$(1+x)f'_\alpha(x) = \alpha f_\alpha(x) \quad \text{with} \quad f_\alpha(0) = 1$$

and show that its radius of convergence is 1, unless $\alpha \in \mathbb{N}_0$. Hint: what you get should be consistent with what you know for $\alpha \in \mathbb{N}_0$.

Exercise 9.19. Prove that $\exp(x+a) = \exp(x)\exp(a)$ for all $x, a \in \mathbb{R}$.

9.6 Extra: integral calculus for power series

Consider

$$p(x) = \sum_{n=1}^{\infty} \alpha_n x^n. \tag{9.17}$$

In Exercise 6.20 we saw that

$$\int_a^b x^n \, dx = \left[\frac{x^n}{n+1}\right]_a^b = \frac{b^{n+1}}{n+1} - \frac{a^{n+1}}{n+1} \tag{9.18}$$

for $0 \leq a < b$. Via Theorem 6.13 and Definition 7.8 this restriction on a and b disappears:

Exercise 9.20. Verify that (9.18) holds for all $n \in \mathbb{N}$ and any $a, b \in \mathbb{R}$.

Theorem 7.17 then implies for

$$p_k(x) = \sum_{n=1}^{k} \alpha_n x^n = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_k x^k, \tag{9.19}$$

the partial polynomial sums of (9.17), that

$$\int_a^b p_k(x)\,dx = P_{k+1}(b) - P_{k+1}(a), \tag{9.20}$$

with P_{k+1} defined by

$$P_{k+1}(x) = \alpha_0 x + \frac{\alpha_1}{2}x^2 + \frac{\alpha_2}{3}x^3 + \cdots + \frac{\alpha_k}{k+1}x^{k+1}. \tag{9.21}$$

You recognise $p_k(x)$ as the derivative of $P_{k+1}(x)$ the way you computed it in highschool, and $P_{k+1}(x)$ as a primitive function for $p_k(x)$.

Now assume for some $r > 0$ that

$$\sum_{n=1}^{\infty} |\alpha_n| r^n < \infty. \tag{9.22}$$

Then

$$|p_k(x) - p(x)| = |\sum_{n=k+1}^{\infty} \alpha_n x^n| \le \sum_{n=k+1}^{\infty} |\alpha_n x^n| \le \sum_{n=k+1}^{\infty} |\alpha_n| r^n,$$

provided $[a,b] \subset [-r,r]$. It follows that $p_k \to p$ in $C([a,b])$ and thus by Theorem 7.12 that

$$\int_a^b p_k(x)\,dx \to \int_a^b p(x)\,dx \tag{9.23}$$

as $k \to \infty$. Combining (9.21) and (9.23) we arrive at the statements in the following theorem[12] for integration variable $x \in [a,b] \subset (-R,R)$.

Theorem 9.21. *If α_n is a sequence of real coefficients indexed by $n \in \mathbb{N}_0$, then there exists a maximal $R \in [0,\infty]$ such*

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots \tag{9.24}$$

exists for all $x \in \mathbb{R}$ with $|x| < R$. For those values

$$P(x) = \alpha_0 x + \frac{\alpha_1}{2}x^2 + \frac{\alpha_2}{3}x^3 + \cdots = \sum_{n=0}^{\infty} \frac{\alpha_n}{n+1}x^{n+1} = \sum_{n=1}^{\infty} \frac{\alpha_{n-1}}{n}x^n \tag{9.25}$$

also exists. Moreover

$$\int_a^b p(x)\,dx = P(b) - P(a)$$

whenever $[a,b] \subset (-R,R)$.

[12] This is really Theorem 9.3 if you think about it.

Exercise 9.22. Finish the proof of Theorem 7.12. Hint: consider the set of values $r > 0$ for which (9.22) holds. It is either empty, the whole of $\mathbb{R}_+$, or an interval of the form $(0, R)$ or $(0, R]$ with $R \in \mathbb{R}_+$.

10 Differentiability via linear approximation

In this chapter we formulate the linearisation approach to differentiation, first for a real valued function f defined on[1] a domain D_f in $\mathbb{R}$ around a point x_0 in the interior of D_f. Writing

$$x = x_0 + h$$

the considerations below concern $h = x - x_0$ sufficiently small. The main difference with Chapter 9 is that the functions under consideration are not specified by algebraic formulas. As a consequence there is no reason to have remainder terms which are quadratic, such as for instance the remainder term in Theorem 9.1. It's analysis again in this chapter.

Definition 10.1. *Let x_0 be an interior point of D_f, let $f : D_f \to \mathbb{R}$ and let $A_0 \in \mathbb{R}$. Then for some $\delta_0 > 0$ the equality*

$$f(x_0 + h) = f(x_0) + A_0 h + R_0(h) \tag{10.1}$$

defines a remainder term $R_0(h)$ *for all $h \in \mathbb{R}$ with $|h| < \delta_0$. It may happen that for every $\varepsilon > 0$ a $\delta > 0$ can be chosen such that*

$$|R_0(h)| < \varepsilon |h| \quad \text{if} \quad 0 < |h| < \delta. \tag{10.2}$$

If so then the function f is called differentiable *in x_0, and we say that $R_0(h)$ is* "small o of h" for h going to zero[2]. *Notation:*

$$R_0(h) = o(|h|) \quad \text{for} \quad h \to 0.$$

Theorem 10.2. *Let x_0 be an interior point of D_f, $f : D_f \to \mathbb{R}$, and suppose that f is differentiable in x_0. Then there is only one $A_0 \in \mathbb{R}$ for which the statement in Definition 10.1 holds, and $f'(x_0) = A_0$ is called the* derivative *of f in x_0.*

Proof. Suppose there is another A_0 that does the job, say B_0 instead of A_0 in (10.1), with remainder term $S(h)$, also satisfying $S(h) = o(|h|)$, just like $R(h)$. Subtraction then gives

$$(A_0 - B_0)h = S(h) - R(h) = o(|h|).$$

Divide by h and take the limit $h \to 0$ to conclude that $A_0 = B_0$. □

[1] For good reasons we kick the habit of writing A for D_f that started in Definition 3.3.
[2] Not to be confused with big O of h.

Exercise 10.3. Give the ε-δ argument that shows $A_0 - B_0 = 0$ in the above proof.

Exercise 10.4. Going back to Definition 10.1, let $g_0 \in \mathbb{R}$ and define the function $g : D_f \to \mathbb{R}$ by

$$g(x_0) = g_0 \quad \text{and} \quad g(x) = \frac{f(x) - f(x_0)}{x - x_0}$$

for all $x \in D_f$, $x \neq x_0$. Prove that f is differentiable in x_0 if and only if it is possible to choose g_0 such that g is continuous in x_0.

10.1 Critical points and the mean value theorem

A critical point[3] of a differentiable function $f : \mathcal{O} \to \mathbb{R}$ is by definition a point $\xi \in \mathcal{O}$ where $f'(\xi) = 0$. This statement makes sense for $\mathcal{O} \subset X$ open and X any real normed space. The following theorem is formulated for the case that $\mathcal{O} = (a, b) \subset \mathbb{R} = X$ and $f : (a, b) \to \mathbb{R}$ differentiable, but generalises to $f : \mathcal{O} \to \mathbb{R}$.

Theorem 10.5. *Let $f : (a, b) \to \mathbb{R}$ and assume that $\xi \in (a, b)$ is such that $f(x) \leq f(\xi)$ for all $x \in (a, b)$. Then $f'(\xi) = 0$ provided f is differentiable in ξ.*

Exercise 10.6. Prove Theorem 10.5. Hint: argue by contradiction.

Theorem 10.7. *The mean value theorem: if $f \in C([a, b])$ is differentiable on (a, b) then for at least one ξ in (a, b) it holds that*

$$\frac{f(b) - f(a)}{b - a} = f'(\xi).$$

Remember this theorem as stating that the difference quotient on the left is equal to a differential quotient *in some point ξ strictly between a and b.*

Proof. In the special case that $f(a) = f(b)$ the point ξ appears as maximizer or minimizer of f on $[a, b]$. Such a maximizer and minimizer must exist in $[a, b]$ in view of Theorem 4.4.

If that maximizer ξ lies in (a, b) then $f'(\xi) = 0$ in view of Theorem 10.5, which is exactly what Theorem 10.7 asserts in the case that $f(a) = f(b)$.

[3]Also: a stationary point.

The same conclusion holds if the minimizer lies in (a, b). One of these two possibilities must occur because otherwise the minimizer and maximizer can only be a or b, forcing the globale maximum and global minimum of f to both be equal to $f(a) = f(b)$, and thereby and $f(x) = f(a) = f(b)$ for all $x \in [a, b]$.

This contradicts the assumption that maximizers and minimizers do not occur in (a, b) and thus completes the proof in case $f(a) = f(b)$, which is also called Rolle's Theorem[4]. You will complete the proof of Theorem 10.7 in Exercise 10.8 by reduction of the general case to this special case. □

Exercise 10.8. Reduce the general case in Theorem 10.7 to the special case $f(a) = f(b)$ and prove Theorem 10.7. Hint: subtract a multiple of x to get equal function values in $x = a$ and $x = b$.

10.2 The fundamental theorem of calculus

Recall the example

$$\ln(x) = \int_1^x \frac{1}{s}\, ds$$

in Exercise 6.21, an integral that makes sense and defines $\ln(x)$ for every real x with $x > 0$. A trickier example you may enjoy to examine is the function from Exercise 4.32.

Exercise 10.9. Let f be the bounded nondecreasing function in Exercise 4.32 which is discontinuous in every point of $\mathbb{Q}$, and define $F : \mathbb{R} \to \mathbb{R}$ by

$$F(x) = \int_0^x f.$$

In which points is F differentiable? In which points is F continuous?

Theorem 10.10. *Let $a, b \in \mathbb{R}$ with $a < b$. Define for $f \in \mathrm{RI}([a, b])$ the function $F \in C([a, b])$ by*

$$F(x) = \int_a^x f(s)\, ds \tag{10.3}$$

Then F is differentiable in every $x_0 \in [a, b]$ where f is continuous, with derivative $F'(x_0) = f(x_0)$.

[4]Read about Rolle in wikipidia.

Note that x_0 is also allowed to be one of the boundary points, for which case the obvious one-sided statement[5] that F is differentiable was not given yet.

Proof. Take $x_0 \in [a,b]$ and write

$$F(x) = F(x_0) + \int_{x_0}^{x} f(s)\,ds = F(x_0) + \int_{x_0}^{x} f(x_0)\,ds + \int_{x_0}^{x} (f(s) - f(x_0))\,ds.$$

With $h = x - x_0$ it follows that

$$F(x) = F(x_0) + f(x_0)h + R_0(h),$$

in which

$$R_0(h) = \int_{x_0}^{x_0+h} (f(s) - f(x_0))\,ds.$$

To conclude that F is differentiable in x_0 with $F'(x_0) = f(x_0)$ we need to show that $R_0(h) = o(|h|)$ as $h \to 0$. Since the integral in the right hand side above is over an interval of length h, continuity of f in x_0 suffices to conclude that F is differentiable in x_0. Indeed, from

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\,\forall_{s\in[a,b]}: \quad 0 < |s - x_0| < \delta \implies |f(s) - f(x_0)| < \varepsilon,$$

we have

$$|R_0(h)| \le \Big|\int_{x_0}^{x_0+h} |f(s) - f(x_0)|\,ds\Big| < \varepsilon|h| \quad \text{if} \quad 0 < |h| \le \delta \tag{10.4}$$

and $x = x_0 + h \in [a,b]$. This completes the proof. □

Definition 10.11. *If $F : [a,b] \to \mathbb{R}$ is differentiable in every $x \in [a,b]$ with $F'(x) = f(x)$, then F is called a* primitive function[6] *of f.*

Theorem 10.10 thus says that every continuous function $f : [a,b] \to \mathbb{R}$ has a primitive on $[a,b]$. For this particular primitive we have that[7]

$$\int_a^b f(x)\,dx = F(b) - F(a), \tag{10.5}$$

because $F(a) = 0$. If we add a constant to F the equality in (10.5) does not change. But does (10.5) hold for every primitive of F of f? To put it differently, is every primitive of f of the form (10.3), up to an additive constant? Theorem 10.7 provides the positive answer. It is not possible for a function to have a zero derivative in every point of an interval without being constant.

[5] Formulate this statement for $x_0 = a$ and $x_0 = b$.

[6] Or anti-derivative.

[7] Have a look at Exercise 6.20 again.

Theorem 10.12. *The fundamental theorem of calculus: for every $f \in C([a,b])$ it holds that*

$$\int_a^b f(x)\,dx = F(b) - F(a),$$

in which F is any primitive of f. Such a primitive exists in view of (10.3). If G is any other primitive than the primitive defined by (10.3), then $F - G$ is constant on $[a,b]$.

Proof. Apply[8] the Mean Value Theorem 10.7 to $F - G$. □

Exercise 10.13. Let $F : \mathbb{R} \to \mathbb{R}$ be continuous, let $T > 0$ and suppose that $f : [0,T] \to \mathbb{R}$ is bounded. Prove that

$$f(t) = \int_0^t F(f(s))\,ds \quad \text{for all} \quad t \in [0,T]$$

if and only if

$$f(0) = 0 \quad \text{and} \quad f'(t) = F(f(t)) \quad \text{for all} \quad t \in [0,T].$$

NB. The first statement requires the assumption that $F \circ f \in \mathrm{RI}([0,T]$, the second statement requires the assumption that f is differentiable in every $t \in [0,T]$.

10.3 A word on notation for later

The formula in Theorem 10.12 is often written as

$$\int_{[a,b]} dF = F(x)\Big|_a^b \quad \text{with} \quad dF = F'(x)dx = f(x)dx, \tag{10.6}$$

and

$$F(x)\Big|_a^b = \big[F(x)\big]_a^b = F(b) - F(a).$$

This formal notation with the d of F will be also used in vector calculus with expressions like $dF = f(x,y)dx + g(x,y)dy$ and products of terms $f(x,y)dx$ en $g(x,y)dy$. The expression $f(x)dx$ is called a 1-form, $F = F(x)$ is called a 0-form, and thus a 1-form can be the d of a 0-form. The d of a 1-form in turn will be a 2-form, and $u(x,y)dxdy$ is an example of a two form[9], and so on.

[8] I don't know of a proof without.

[9] Usually witten as $u(x,y)dx \wedge dy$.

The algebra with forms will be defined later to mimic natural operations in multivariate integral calculus, and will be based on the formal rules[10] $dxdy + dydx = 0$, $ddx = 0$, and a Leibniz type rule, see Chapter 21 and further. We already note that in Theorem 10.12 the expression on the left can be seen as

$$\int_a^b \quad \text{acting on} \quad f(x)dx,$$

and the expression in the right as

$$|_a^b \quad \text{acting on} \quad F(x),$$

an interaction between "integrals" and differential forms.

10.4 Some strange examples

Exercise 10.14. If $g : \mathbb{R} \to \mathbb{R}$ is continuous in $x = 0$ with $g(0) = 0$, then $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = xg(x)$ is differentiable in $x = 0$ with $f'(0) = 0$. Show this directly from the definition of differentiability.

Remark 10.15. *For g in Exercise 10.14 you can take a strange function like for instance g defined by $g(x) = 0$ for $x \in \mathbb{Q}$ and $g(x) = x$ for $x \notin \mathbb{Q}$. Then $f : \mathbb{R} \to \mathbb{R}$ is discontinuous in every $x \neq 0$ while differentiable in $x = 0$.*

Exercise 10.16. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(0) = 0$ and

$$f(x) = x^2 \sin \frac{1}{x^2}$$

for $x \neq 0$. Show that f is differentiable in every $x \in \mathbb{R}$ but that $f'(x)$ is unbounded on $[0, 1]$.

Exercise 10.17. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(0) = 0$ and

$$f(x) = \exp(-\frac{1}{x^2})$$

for $x \neq 0$. Sketch the graph of f. Show that f is differentiable on the whole of $\mathbb{R}$, and that $f'(0) = 0$. Then show that the same is true for f', namely $(f')'(x) = f''(x)$ exists for all $x \in \mathbb{R}$ and $f''(0) = 0$. And so on for f''', f'''' and all higher order derivatives.

[10] Recall from Definition 7.8 that we think of dx and thus also dy as having a sign.

11 The rules for differentiation

In this chapter we formulate and prove the rules of differentiation that you have been using in calculus. In Chapter 14 these differentiation rules transform into the rules for integration, by means of Theorem 10.12 above, the fundamental theorem of calculus.

11.1 The sum and product rules

For real valued functions f and g of the same variable x we have the sum and product rules. We formulate them for real valued functions of a real variable first[1].

Theorem 11.1. *Let x_0 be an interior point of $D_f \cap D_g$, $f : D_f \to \mathbb{R}$ and $g : D_g \to \mathbb{R}$ differentiable in x_0. Then $f + g$ and fg are also differentiable in x_0 with the sum rule*

$$(f + g)'(x_0) = f'(x_0) + g'(x_0)$$

and the Leibniz product rule

$$(fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0).$$

Proof. Both proofs are straightforward. Writing expansions with $x - x_0$ instead of h, and the remainder term as $R_0(x)$, we expand $f(x)$ as

$$f(x) = f(x_0) + A_0(x - x_0) + R_0(x). \tag{11.1}$$

Here

$$A_0 = f'(x_0)$$

if

$$R_0(x) = o(|x - x_0|) \quad \text{as} \quad x \to x_0,$$

i.e. if

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{x\in D_f} \quad 0 < |x - x_0| < \delta \implies |R_0(x)| < \varepsilon|x - x_0|. \tag{11.2}$$

Note that we still write R_0 for the remainder term, but now choose to see it as a function of x. For g this becomes[2]

$$g(x) = g(x_0) + \underbrace{B_0}_{g'(x_0)}(x - x_0) + S_0(x), \quad S_0(x) = o(|x - x_0|). \tag{11.3}$$

[1] Again the results generalise.

[2] We use the alphabetic shift convention.

Adding (11.1) to (11.3) gives

$$(f+g)(x) = f(x) + g(x) =$$
$$f(x_0) + g(x_0) + A_0(x-x_0) + B_0(x-x_0) + R_0(x) + S_0(x) =$$
$$(f+g)(x_0) + \underbrace{(A_0+B_0)}_{(f+g)'(x_0)}(x-x_0) + \underbrace{R_0(x)+S_0(x)}_{\text{remainder term}}$$

for all $x \in D_f \cap D_g$. The remainder term clearly has the same properties as the individual remainder terms $R_0(x)$ and $S_0(x)$, warranting the conclusion that $f+g$ is differentiable in x_0 if f and g are, with

$$(f+g)'(x_0) = A_0 + B_0 = f'(x_0) + g'(x_0). \tag{11.4}$$

Carefully note that the argument sees no difference between $D_f \cap D_g \subset \mathbb{R}$ and $D_f \cap D_g \subset X$.

Next consider the product function fg defined by

$$(fg)(x) = f(x)g(x)$$

for all $x \in D_f \cap D_g$ and multiply (11.1) and (11.3) to get

$$(fg)(x) = f(x)g(x) = (f(x_0)+A_0(x-x_0)+R_0(x))(g(x_0)+B_0(x-x_0)+S_0(x))$$

$$= f(x_0)g(x_0) + \underbrace{A_0(x-x_0)g(x_0) + f(x_0)B_0(x-x_0)}_{(fg)'(x_0)(x-x_0)?} + T_0(x). \tag{11.5}$$

The remainder term $T_0(x)$ consists of the 6 other combinations of the 3 terms in (11.1) and (11.3). To conclude that fg is differentiable in x_0 you must check that each of these 6 terms is $o(|x-x_0|)$ as $x \to x_0$. Once it has been shown that

$$T_0(x) = o(|x-x_0|) \quad \text{as} \quad x \to x_0 \tag{11.6}$$

we read off from (11.5) that

$$(fg)'(x_0) = g(x_0)A_0 + f(x_0)B_0 = g(x_0)f'(x_0) + f(x_0)g'(x_0). \tag{11.7}$$

So do Exercise 11.2 below to complete the proof. □

Exercise 11.2. Prove that (11.6) holds. That is, use

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\,\forall_{x\in D_f} \quad 0 < |x-x_0| < \delta \implies |R_0(x)| < \varepsilon|x-x_0|$$

and the same statement for $S_0(x)$ to prove the same statement for each of the above 6 terms in $T_0(x)$ with x restricted to $D_f \cap D_g$.

Remark 11.3. *Both arguments see no difference between $D_f \cap D_g \subset \mathbb{R}$ and $D_f \cap D_g \subset X$. Note that $f(x_0) \in \mathbb{R}$ and $g(x_0) \in \mathbb{R}$ appear as scalars and are moved to the left in front of the linear map from X to $\mathbb{R}$ in each of the two terms in (11.7). In Chapter 12 we discuss the general case in which the other $\mathbb{R}$ is also replaced by Y. But then we must distinguish between the sum and the product rule.*

11.2 The chain rule

We now derive the chain rule, a rule which is in fact easier than (11.7), easier because it only needs *linear* algebra. So consider $g(f(x))$, with f defined on some domain D_f and g defined on some domain D_g. To be specific, we start with

$$x_0 \in D_f,$$

and assume that

$$y_0 = f(x_0) \in D_g.$$

Theorem 11.4. *Let x_0 be an interior point of the domain of f, assume f differentiable in x_0, let $y_0 = f(x_0)$ be an interior point of the domain of g, and assume that g differentiable in y_0. Let*

$$g(f(x) = (g \circ f)(x)$$

define the composition $g \circ f$ of g and f. Then x_0 is in the interior of the domain of $g \circ f$ and $g \circ f$ is differentiable in x_0 with

$$(g \circ f)'(x_0) = g'(y_0)f'(x_0). \tag{11.8}$$

Proof[3]. We want to linearise $g \circ f$ around x_0. To do so

$$g(y) = g(y_0) + \underbrace{B_0}_{g'(y_0)}(y - y_0) + S_0(y),$$

has to be combined with

$$f(x) = f(x_0) + \underbrace{A_0}_{f'(y_0)}(x - x_0) + R_0(x).$$

We assume both remainder terms $R_0(x)$ and $S_0(y)$ have the property needed for differentiability of f in x_0 and g in y_0, namely (12.7) for f,

$$\forall_{\varepsilon>0}\, \exists_{\delta>0} \quad 0 < |x - x_0| < \delta \implies |R_0(x)| < \varepsilon|x - x_0|,$$

[3]Simplify! Restrict to $x_0 = 0, y_0 = f(0) = 0, g(0) = 0$ and drop all subscripts.

and

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\,0<|y-y_0|<\delta \implies |S_0(y)|<\varepsilon|y-y_0| \tag{11.9}$$

for g. In particular these two statements provide us with $\delta>0$ for which

$$B_\delta(x_0)\subset D_f \quad \text{and} \quad B_\delta(y_0)\subset D_g.$$

Next we verify that the properties of the remainder terms $R_0(x)$ and $S_0(y)$ carry over to the remainder term $T_0(x)$ in

$$g(f(x))=g(y)=g(y_0)+B_0(\underbrace{f(x)-f(x_0)}_{y-y_0})+S_0(y)=$$

$$g(y_0)+B_0A_0(x-x_0)+\underbrace{B_0R_0(x)+S_0(y)}_{T_0(x)}.$$

The first term in $T_0(x)$ exists for all $x\in\mathbb{R}$ and is estimated via

$$|B_0R_0(x)|=|B_0|\,|R_0(x)|,$$

and therefore has the desired property that it is $o(|x-x_0|_X)$ as $x\to x_0$, simply because $R_0(x)$ does. For the second term we pick $\varepsilon>0$ and then know that

$$|S_0(y)|<\varepsilon|y-y_0| \quad \text{if} \quad 0<|y-y_0|<\delta,$$

with $\delta>0$ as in (11.9). What we want is an estimate in terms of a multiple of $\varepsilon|x-x_0|$ if $0<|x-x_0|<\tilde{\delta}$ for some other $\tilde{\delta}>0$ chosen depending on the positive ε we started with.

If by chance $y=y_0$ there's no work to be done. If not, then we need

$$0<|y-y_0|<\delta$$

if we want to conclude via (11.9). We actually have

$$|y-y_0|=|f(x)-f(x_0)|=|A_0(x-x_0)+R_0(x)|\le|A_0|\,|x-x_0|+|R_0(x)|$$
$$<(|A_0|+1)\,|x-x_0| \quad \text{if} \quad 0<|x-x_0|<\delta_R,$$

in which $\delta_R>0$ is provided by (10.2) applied with $\varepsilon=1$. So we indeed conclude via (11.9) if

$$0<|x-x_0|<\frac{\delta}{|A_0|+1}=\tilde{\delta},$$

which then implies that the second term in $T_0(x)$ exists so that x is actually in the domain of $g\circ f$. Moreover the second term is estimated by

$$|S_0(y)|<\varepsilon|y-y_0|<\underbrace{\varepsilon(|A_0|+1)}_{\tilde{\varepsilon}}\,|x-x_0|.$$

Leaving further cosmetics to the reader this concludes the proof that also the second term in $T_0(x)$ is $o(|x-x_0|_X)$ as $x\to x_0$. We have derived and proved the chain rule. □

11.3 Extra: differentiability of inverse functions

Consider the functions f and g in Theorem 8.19. We ask about the differentiability of g in some $y_0 = f(x_0)$ with $x_0 \in (a, b)$ and f differentiable in x_0 with $f'(x_0) > 0$. The positive answer to this question is that g is differentiable in y_0 and that

$$f'(x_0)g'(y_0) = 1, \tag{11.10}$$

a statement which is symmetric in f and g.

Proof of (11.10). To establish the positive answer we first make our lifes easy by noting that without loss of generality we may assume that $0 = x_0 = y_0 = 0 = f(0)$, and that $f'(x_0) = 1$. This means that

$$f(x) = x + o(x) \quad \text{as} \quad x \to 0, \tag{11.11}$$

i.e.

$$\forall_{\varepsilon>0}\, \exists_{\delta>0} \quad 0 < |x| < \delta \implies |f(x) - x| < \varepsilon|x|. \tag{11.12}$$

The inequality for $|f(x) - x|$ means that

$$(1-\varepsilon)x < y < (1+\varepsilon)x \quad \text{if} \quad 0 < x < \delta \quad \text{and} \quad y = f(x), \tag{11.13}$$

and the other way around for $-\delta < x < 0$. We want to replace this statement by an equivalent statement which is symmetric in x and y, and thereby also equivalent to

$$g(y) = y + o(y) \quad \text{as} \quad y \to 0. \tag{11.14}$$

How do we get the equivalent symmetric statement? Clearly the condition $y = f(x)$ already is symmetric because

$$y = f(x) \iff x = g(y),$$

but the inequalities with x and y are not. Note though that

$$(1-\varepsilon)x < y < (1+\varepsilon)x \implies (1-\varepsilon)x < y < \frac{1}{1-\varepsilon}x$$

if $x > 0$ and $0 < \varepsilon < 1$. In other words (11.12) implies that

$$\forall_{\varepsilon\in(0,1)}\, \exists_{\delta>0} \quad \begin{matrix} 0 < x < \delta \\ y = f(x) \end{matrix} \implies (1-\varepsilon)x < y < \frac{1}{1-\varepsilon}x, \tag{11.15}$$

and likewise[4] for $-\delta < x < 0$.

[4] With the same δ given $0 < \varepsilon < 1$, and with reversed inequalities for y.

Next observe that (11.15) and its version for $x < 0$ in turn imply

$$\forall_{\varepsilon\in(0,1)}\,\exists_{\delta>0}\quad 0<|x|<\delta \implies |f(x)-x| < \frac{\varepsilon}{1-\varepsilon}|x|, \tag{11.16}$$

since

$$\frac{1}{1-\varepsilon} = 1 + \frac{\varepsilon}{1-\varepsilon}.$$

But (11.16) and (11.12) are equivalent, by setting

$$\tilde{\varepsilon} = \frac{\varepsilon}{1-\varepsilon},$$

and thus (11.15) and its version for $x < 0$ make up for an equivalent definition of (11.11): we have

$$\forall_{\varepsilon\in(0,1)}\,\exists_{\delta>0}: \quad G_\delta = \{(x,y)\in\mathbb{R}^2 : 0<|x|<\delta,\, y=f(x)\} \subset S_\varepsilon, \tag{11.17}$$

in which

$$S_\varepsilon = \{(x,y)\neq(0,0) : \frac{1}{1-\varepsilon} < \frac{y}{x} < 1-\varepsilon\} \tag{11.18}$$

is clearly symmetric in x and y. Now choose $\tilde{\delta} > 0$ such that, for the same $\varepsilon \in (0,1)$, it holds that

$$F_{\tilde{\delta}} = \{(x,y)\in\mathbb{R}^2 : 0<|y|<\tilde{\delta},\, x=g(y)\} \subset S_\varepsilon.$$

How? Draw a picture to see that

$$\tilde{\delta} = (1-\varepsilon)\delta$$

does the job. This completes the proof. □

Exercise 11.5. In view of Section 11.3 and Theorem 8.19 the function $\ln$ has an inverse function $f : \mathbb{R} \to \mathbb{R}^+$. Show that $f(0) = 1$ and that $f'(y) = f(y)$ for all $y \in \mathbb{R}$. Look at Theorem 9.10 and explain why $f = \exp$.

Exercise 11.6. Show again that $\exp(x+y) = \exp(x)\exp(y)$ for all $x, y \in \mathbb{R}$, and that with $e = \exp(1)$ defined by

$$\ln e = \int_1^e \frac{1}{x}\,dx = 1,$$

it follows that

$$\exp(\frac{p}{q}) = e^{\frac{p}{q}} = \sqrt[q]{e^p}$$

for all $p \in \mathbb{Z}$ and all $q \in \mathbb{N}$. By general agreement we define $e^x = \exp(x)$ for all other $x \in \mathbb{R}$ as well.

Likewise for x^α with $x > 0$. Via

$$x^n = (e^{\ln x})^n = e^{n \ln x}$$

for $n \in \mathbb{N}$, but also with $n \in \mathbb{N}$ replaced by $r = \frac{p}{q} \in \mathbb{Q}$ and finally by general agreement:

$$x^\alpha = e^{\alpha \ln x} \quad \text{for} \quad x > 0 \quad \text{and} \quad \alpha \in \mathbb{R}. \tag{11.19}$$

Exercise 11.7. Show that

$$x \to \frac{\sin x}{\cos x} = \tan x$$

is strictly increasing on $(-\frac{\pi}{2}, \frac{\pi}{2})$ and has an inverse function

$$y \to \arctan y$$

on $\mathbb{R}$ with derivative

$$\frac{1}{1+y^2}.$$

Show that

$$\arctan y = y - \frac{1}{3}y^3 + \frac{1}{5}y^5 - \cdots$$

for $|y| < 1$.

Exercise 11.8. Show that

$$x \to \sin x$$

is strictly increasing on $(-\frac{\pi}{2}, \frac{\pi}{2})$ and has an inverse function

$$y \to \arcsin y$$

on $(-1, 1)$ with derivative

$$\frac{1}{\sqrt{1-y^2}}.$$

Derive a power series expression for $\arcsin y$ for $|y| < 1$.

Exercise 11.9. Consider

$$x \to \cos x$$

on $(0, \pi)$. Show for the inverse that $\arccos y + \arcsin y$ is constant on $(-1, 1)$. Which constant?

Exercise 11.10. Show that $\sin$ is a periodic function. Its period is by definition 2π. Show that $-\sin(-x) = \sin x = \sin(\pi - x) > 0$ for $0 < x < \pi$.

Exercise 11.11. Solve the differential equation in Exercise 9.18 via

$$\frac{f'_\alpha(x)}{f_\alpha(x)} = \frac{\alpha}{1+x}$$

and integration from 1 to x. Prove that

$$(1+x)^\alpha = 1 + \alpha x + \frac{\alpha(\alpha-1)}{2}x^2 + \frac{\alpha(\alpha-1)(\alpha-2)}{3\cdot 2}x^3 + \cdots = \sum_{k=0}^{\infty}\binom{\alpha}{k}x^k \quad (11.20)$$

for all $x \in \mathbb{R}$ with $|x| < 1$.

Exercise 11.12. Take $\alpha = \frac{1}{2}$ and square the series in (11.20). Prove that

$$\left(\sum_{k=0}^{\infty}\binom{\frac{1}{2}}{k}x^k\right)^2 = 1 + x,$$

for all $x \in \mathbb{R}$ with $|x| < 1$. To some extent this was perhaps known to the Babylonians.

Exercise 11.13. Write out the first few terms of

$$\sqrt[n]{1+x} = 1 + \frac{x}{n} + \cdots \quad \text{and} \quad \frac{1}{\sqrt[n]{1+x}} = 1 - \frac{x}{n} + \cdots$$

12 Extra: differentiation in normed spaces

In fact we may just as well speak about $D_f \subset X$, X a normed space, x_0 in the interior of D_f, $f : D_f \to \mathbb{R}$,

$$f(x_0 + h) = f(x_0) + \phi_0(h) + R_0(h),$$

in which $\phi_0 : X \to \mathbb{R}$ is linear and Lipschitz[1] continuous[2]. The ε-δ statement (10.2) then becomes

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{h\in X} : \quad 0 < |h|_X < \delta \implies |R(h)| < \varepsilon|h|.$$

It implies that such ϕ_0, if it exists, is unique, with $\phi_0(h) = A_0 h$ in the special case under consideration in Definition 10.1.

If you understand what's going on you see that everything also works for maps Φ from $D_\Phi \subset X$ to Y, X and Y normed spaces. We shall write

$$\Phi(x_0 + h) = \Phi(x_0) + A_0 h + R_0(h),$$

in which we write $A_0 h$ instead of $A_0(h)$ for $A_0 : X \to Y$ linear and Lipschitz[3] continuous. Definition 10.1 and Theorem 10.2 are just special cases of the following definition and theorem.

Definition 12.1. *Let X, Y be real normed spaces, $D_\phi \subset X$, $\Phi : D_\Phi \to Y$, x_0 an interior point of D_Φ and let $A_0 : X \to Y$ be* Lipschitz continuous. *Then*

$$\Phi(x_0 + h) = \Phi(x_0) + A_0 h + R_0(h) \tag{12.1}$$

defines a remainder term *$R_0(h)$ for $h \in X$ with $|h|_X < \delta_0$ for some $\delta_0 > 0$. It may happen that for every $\varepsilon > 0$ a $\delta > 0$ can be chosen such that*

$$|R_0(h)|_Y < \varepsilon|h|_X \quad \text{if} \quad 0 < |h|_X < \delta. \tag{12.2}$$

If so then the map Φ is called differentiable *in x_0.*

Theorem 12.2. *Let X, Y be real normed spaces, $D_\Phi \subset X$, $\Phi : D_\Phi \to Y$, x_0 an interior point of D_Φ, and suppose that f is* differentiable *in x_0. Then there is* precisely one linear Lipschitz continuous map *$A_0 : X \to Y$ for which the statement in Definition 12.1 holds, and $\Phi'(x_0) = A_0$ is called* the derivative *of Φ in x_0.*

[1] If $\phi : X \to \mathbb{R}$ is linear an continuous in 0 then it is Lipschitz continuous.
[2] In order to have f differentiable in x_0 imply that f is continuous in x_0, explain!
[3] Again: if $A_0 : X \to Y$ is linear an continuous in 0 then it is Lipschitz continuous.

Remark 12.3. *The space of all Lipschitz continuous linear maps A from X to Y that qualify to be used in Definition 12.1 is denoted by $L(X,Y)$. We shall write*

$$|A|_{L(X,Y)} \tag{12.3}$$

for the best (smallest) Lipschitz constant of such an A.

Theorem 12.4. *Let $x, y \in X$, X a normed space, $\mathcal{O} \subset X$ open,*

$$[x,y] = \{\xi(t) = (1-t)x + ty;\ 0 \le t \le 1\} \subset \mathcal{O},$$

and $f : \mathcal{O} \to \mathbb{R}$ differentiable. Then there exists

$$\xi \in (x,y) = \{(1-t)x + ty;\ 0 < t < 1\}$$

such that

$$f(y) - f(x) = f'(\xi)(y-x).$$

Exercise 12.5. Give a direct proof that[4]

$$t \to f(\xi(t)) \tag{12.4}$$

is differentiable on $[0,1]$. Then use Theorem 10.7 to prove Theorem 12.4. Can the assumption $[x,y] \subset \mathcal{O}$ be weakened?

The argument in Theorem 11.1 for the sum function immediately generalises to $\Phi : D_\Phi \to Y$ and $\Psi : D_\Psi \to Y$ as in Definition 12.1 and Theorem 12.2. For the general Leibniz rule we suppose Φ and Ψ map to a normed algebra Y and are as in Definition 12.1 and Theorem 12.2. If the multiplication is commutative we have

$$\underbrace{A_0(x-x_0)}_{\text{in } Y}\underbrace{\Psi(x_0)}_{\text{in } Y} = \underbrace{\Psi(x_0)A_0}_{\text{in } L(X,Y)}(x-x_0) \in Y$$

and (11.7) remains unaltered. Only the notation changes when we write

$$(\Phi\Psi)'(x_0) = \Psi(x_0)A_0 + \Phi(x_0)B_0 = \Psi(x_0)\Phi'(x_0) + \Phi(x_0)\Psi'(x_0). \tag{12.5}$$

If multiplication in Y is not commutative we have that

$$((\Phi\Psi)'(x_0))(h) = (\Phi'(x_0)(h))\Psi(x_0) + \Phi(x_0)(\Psi'(x_0)(h)) \tag{12.6}$$

[4]You really don't need the general chain rule in Theorem 11.4 to do so.

defines $(\Phi\Psi)'(x_0)$. It is Lipschitz continuous because, using $|yz|_Y \le |y|_Y\,|z|_Y$ and recalling (12.3), we have

$$\begin{aligned}|(\Phi\Psi)'(x_0)(h)|_Y &\le |(\Phi'(x_0)(h))\Psi(x_0)|_Y + |\Phi(x_0)(\Psi'(x_0)(h))|_Y\\ &\le |\Phi'(x_0)(h)|_Y\,|\Psi(x_0)|_Y + |\Phi(x_0)|_Y\,|(\Psi'(x_0)(h))|_Y\\ &\le |\Phi'(x_0)|_{L(X,Y)}\,|h|_X\,|\Psi(x_0)|_Y + |\Phi(x_0)|_Y\,|\Psi'(x_0)|_{L(X,Y)}\,|h|_X,\end{aligned}$$

whence

$$|(\Phi\Psi)'(x_0)|_{L(X,Y)} \le |\Phi'(x_0)|_{L(X,Y)}\,|\Psi(x_0)|_Y + |\Phi(x_0)|_Y\,|\Psi'(x_0)|_{L(X,Y)}.$$

Next we look at the remainder term $T_0(x)$, which is the sum of

$$\Phi(x_0)S_0(x) + R_0(x)\Psi(x_0),$$

$$A_0(x-x_0)B_0(x-x_0),$$

$$A_0(x-x_0)S_0(x) + R_0(x)B_0(x-x_0),$$

and

$$R_0(x)S_0(x).$$

Exercise 12.6. Prove in the general setting of normed spaces X and Y that (11.6) holds. That is, use

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\,\forall_{x\in X}\quad 0<|x-x_0|_X<\delta \implies |R_0(x)|_Y<\varepsilon|x-x_0|_X \tag{12.7}$$

and the same statement for $S_0(x)$ to prove the same statement for each of the above 6 terms in $T_0(x)$.

Exercise 12.7. The functions defined by

$$(x,y)\to x+y \quad\text{and}\quad (x,y)\to xy$$

are differentiable from $\mathbb{R}^2$ to $\mathbb{R}$. Why?

Remark 12.8. *Exercise 12.7 should lead you to reflect on the observation that the (general) chain rule below does in fact imply the sum and product rules in Section 11.1.*

We conclude this section with the observation that there is no difference between the arguments in the proof of Theorem 11.4 above for

$$D_f \subset \mathbb{R}, \quad f : D_f \to \mathbb{R}, \quad D_g \subset \mathbb{R}, \quad g : D_g \to \mathbb{R},$$

and the arguments for

$$D_\Phi \subset X, \quad \Phi : D_\Phi \to Y, \quad D_\Psi \subset Y, \quad \Psi : D_g \to Z,$$

$$x \xrightarrow{\Psi \circ \Phi} \Psi(\Phi(x))$$

in Theorem 12.9 below. To linearise this map around x_0 we combine

$$\Psi(y) = \Psi(y_0) + B_0(y - y_0) + S_0(y), \quad B_0 = \Psi'(y_0)$$

with

$$\Phi(x) = \Phi(x_0) + A_0(x - x_0) + R_0(x), \quad A_0 = \Phi'(y_0).$$

We assume both remainder terms $R_0(x)$ and $S_0(y)$ have the property needed for differentiability of Φ in x_0 and Ψ in y_0, namely (12.7) for Φ,

$$\forall_{\varepsilon>0}\, \exists_{\delta>0} \quad 0 < |x - x_0|_X < \delta \implies |R_0(x)|_Y < \varepsilon |x - x_0|_X,$$

and

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, 0 < |y - y_0|_Y < \delta \implies |S_0(y)|_Z < \varepsilon |y - y_0|_Y \tag{12.8}$$

for Ψ. Again these two statements provide us with $\delta > 0$ for which

$$B_\delta(x_0) = \{x \in X : |x - x_0|_X < \delta\} \subset D_\Phi$$

and

$$B_\delta(y_0) = \{y \in Y : |y - y_0|_Y < \delta\} \subset D_\Psi$$

hold. Writing

$$\Psi(\Phi(x)) = \Psi(y) = \Psi(y_0) + B_0(\underbrace{\Phi(x) - \Phi(x_0)}_{y - y_0}) + S_0(y) =$$

$$\Psi(y_0) + B_0 A_0(x - x_0) + \underbrace{B_0 R_0(x) + S_0(y)}_{T_0(x)},$$

in which the second term features the derivative of the composition. We note that the first term in $T_0(x)$ is now estimated via *an inequality*

$$|B_0 R_0(x)|_Z \le |B_0|_{L(Y,Z)}\, |R_0(x)|_Y.$$

The rest of the proof is copy-paste from the proof for $X = Y = Z = \mathbb{R}$, with f and g replaced by Φ and Ψ, and the appropiate subscripts on the norms. We paste only the inequalities. They read

$$|S_0(y)|_Z < \varepsilon |y - y_0|_Y \quad \text{if} \quad 0 < |y - y_0|_Y < \delta,$$

$$\begin{aligned} |y - y_0|_Y &= |\Phi(x) - \Phi(x_0)|_Y = |A_0(x - x_0) + R_0(x)|_Y \\ &\leq |A_0|_{L(X,Y)} |x - x_0|_X + |R_0(x)|_Y \\ &< (|A_0|_{L(X,Y)} + 1) |x - x_0|_X \quad \text{if} \quad 0 < |x - x_0|_X < \delta_R, \end{aligned}$$

$$0 < |x - x_0|_X < \frac{\delta}{|A_0|_{L(X,Y)} + 1} = \tilde{\delta}$$

$$|S_0(y)|_Z < \varepsilon |y - y_0|_Y < \underbrace{\varepsilon(|A_0|_{L(X,Y)} + 1)}_{\tilde{\varepsilon}} |x - x_0|_X.$$

The general chain rule is now given by the following theorem.

Theorem 12.9. *Let x_0 be an interior point of the domain of Φ, assume Φ differentiable in x_0, let $y_0 = \Phi(x_0)$ be an interior point of the domain of Ψ, and assume that Ψ differentiable in y_0. Then x_0 is in the interior of the domain of $\Psi \circ \Phi$ and $\Psi \circ \Phi$ is differentiable in x_0 with*

$$(\Psi \circ \Phi)'(x_0) = \Psi'(y_0)\Phi'(x_0). \tag{12.9}$$

Exercise 12.10. Derive and prove the differentiation rules for fg and $\frac{g}{f}$ if f and g are real valued functions from Exercise 12.7 and Theorem 11.4. Hint: use also $y \to \frac{1}{y}$.

13 Extra: Newton's method revisited

For the analysis of Newton's method we need the mean value theorem in integral form.

Exercise 13.1. Theorem 10.12 can be formulated for $F : [a, b] \to \mathbb{R}$ continuously differentiable, i.e. $F : [a, b] \to \mathbb{R}$ is differentiable and $x \to F'(x)$ defines a continuous function on $[a, b]$. Rewrite

$$F(b) - F(a) = \int_a^b F'(x)\,dx$$

via the substitution

$$x = (1-t)a + tb = a + t(b-a)$$

as

$$F(b)-F(a) = \int_0^1 F'((1-t)a+tb)(b-a)\,dt = \int_0^1 F'((1-t)a+tb)\,dt\,(b-a), \quad (13.1)$$

and prove the result directly from the definitions, without using the rule $dx = (b-a)dt$.

We note that if $x \to F'(x)$ is Lipschitz continuous is on $[a, b]$, the first integral in (13.1) with $b = x$ rewrites as

$$\int_0^1 F'(a)(x-a)\,dt + \int_0^1 (F'((1-t)a+tx) - F'(a))(x-a)\,dt,$$

so

$$F(x) = F(a) + F'(a)(x-a) + R(x;a) \quad (13.2)$$

with[1]

$$R(x;a) = R_a(x) = \int_0^1 (F'((1-t)a+tx) - F'(a))(x-a)\,dt.$$

If the Lipschitz constant of $x \to F'(x)$ is L then

$$|R(x;a)| \le \int_0^1 Lt|x-a|^2\,dt = \frac{L}{2}|x-a|^2. \quad (13.3)$$

In (13.2) we have a linear approximation with a remainder term estimated in (13.3) by a constant times $|x-a|^2$. We say that

$$R(x;a) = O(|x-a|^2)$$

is big O of $|x-a|$ squared as $x \to a$. This is just like what we had for power series with (9.9). Note that $O(|x-a|^2)$ implies $o(|x-a|)$ but in general it is not true that $o(|x-a|)$ implies $O(|x-a|^2)$.

[1]From here on we change form subscript a on $R(x)$ to $R(x;a)$.

13.1 The generalised mean value formula

Theorem 13.2. *Let X be complete metric vector space. For $f : [a, b] \to X$ continuous let the function $F : [a, b] \to X$ be defined*[2] *by*

$$F(x) = \int_a^x f(s)\, ds.$$

Then F is differentiable in every $x_0 \in [a, b]$ with $F'(x_0) = f(x_0)$.

As before Theorem 13.2 says that F is a primitive of f, and that for this primitive

$$\int_a^b f(s)\, ds = F(b) - F(a), \tag{13.4}$$

because $F(a) = 0$. If $\tilde{F}$ is another primitive of f then

$$G = F - \tilde{F} : [a, b] \to X$$

is differentiable with $G'(x) = 0$ for all $x \in [a, b]$.

Exercise 13.3. Show that for every linear Lipschitz continuous functions $\psi : X \to \mathbb{R}$ the real valued function

$$x \xrightarrow{g} \psi(G(x))$$

is differentiable on $[a, b]$ with $g'(x)$ for every $x \in [a, b]$ defined by

$$h \xrightarrow{g'(x)} \psi(G(x))G'(x)h = 0$$

for $h \in \mathbb{R}$. So $g(b) = g(a)$ by Theorem 10.7.

We conclude that $\psi(G(b)) - \psi(G(a)) = 0$ for every Lipschitz continuous linear function $\psi : X \to \mathbb{R}$. For $y = G(b) - G(a)$ it thus holds that $\psi(y) = 0$ for every linear Lipschitz continuous functions $\psi : X \to \mathbb{R}$. If this implies that $y = 0$ it follows that $F(b) - F(a) = \tilde{F}(b) - \tilde{F}(a)$. This completes the proof of the following theorem, in which $\tilde{F}$ is called F.

Theorem 13.4. *Let X be a complete metric vector space with the property*[3] *that $\psi(y) = 0$ for every Lipschitz continuous linear function $\psi : X \to \mathbb{R}$*

[2] See Theorem 8.15.

[3] Zorn's Lemma implies that this property holds.

implies that $y = 0$. *If* $f : [a, b] \to X$ *is continuous and* $F : [a, b] \to X$ *is a primitive*[4] *of* f, *then*

$$\int_a^b f(s)\,ds = F(b) - F(a) = \int_0^1 F'((1-t)a + tb)\,dt\,(b-a).$$

Such a primitive exists in view of Theorem 13.2.

Summing up, the mean value integral formula (13.1) also holds for X-valued functions and integrals. Only for $\mathbb{R}$-valued functions the integral can be seen as lying between the minimum and the maximum of the integrand, and is therefore equal to some value $F'(\xi)$ with $\xi \in [a, b]$, a slightly weaker statement than in Theorem 10.7, under a much stronger assumption than Theorem 10.7, exclusively for $\mathbb{R}$-valued functions.

For continuously differentiable $F : \mathcal{O} \to Y$, Y a complete metric vector space, $\mathcal{O}, x, y$ as in Theorem 12.4, we apply Theorem 13.4 with $a = 0$ and $b = 1$ to the function defined by (12.4), and conclude that

$$F(y) - F(x) = \int_0^1 F'((1-t)x + ty)(y-x)\,dt, \tag{13.5}$$

as a Y-valued integral, which we can write as

$$F(y) - F(x) = \int_0^1 F'((1-t)x + ty)\,dt(y-x), \tag{13.6}$$

an operator-valued integral acting on $y - x \in X$. This version of the mean value theorem will be used in the proof of Theorem 15.4.

13.2 Convergence of Newton's method

For $r > 0$ let $f : \mathbb{R} \to \mathbb{R}$ be differentiable on the open ball[5]

$$B_r = \{x \in \mathbb{R} : |x| < r\}.$$

If $x \to f'(x)$ is Lipschitz continuous on B_r, and x_n is a sequence in B_r, (13.2) rewrites as

$$f(x_n) = \underbrace{f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1})}_{\text{linear approximation}} + R(x_n; x_{n-1}), \tag{13.7}$$

[4] $F'(x) = f(x)$ for all $x \in [a, b]$.

[5] Generalises to $f : X \to X$, X a complete metric vector space (Theorem 8.15!).

in which

$$|R(x_n; x_{n-1})| \leq \frac{L}{2}|x_n - x_{n-1}|^2,$$

with L the Lipschitz constant of f' on B_r. Assume for all $x \in B_r$ that

$$|(f'(x))^{-1}| \leq C,$$

form some positive constant $C > 0$.

Let

$$p_n = |x_n - x_{n-1}| \quad \text{and} \quad q_n = |f(x_n)|, \tag{13.8}$$

and assume that x_n is defined by

$$x_n = x_{n-1} - (f'(x_{n-1}))^{-1} f(x_{n-1}) \quad (n \in \mathbb{N}), \tag{13.9}$$

with $x_0 = 0$. Then $x_n \in B_r$ as long as

$$p_1 + p_2 + \cdots + p_n < r, \tag{13.10}$$

in which case it follows that

$$p_n \leq C q_{n-1} \quad \text{and} \quad q_n \leq \frac{1}{2} L p_n^2, \tag{13.11}$$

because (13.9) puts the linear approximation in (13.7) equal to zero.

The inequalities in (13.11) can now be used beginning with

$$q_0 = |f(0)| \quad \text{and} \quad p_1 \leq C q_0 = C|f(0)|. \tag{13.12}$$

Combining (13.11) and (13.12) it follows that

$$p_n \leq \mu p_n^2 \quad \text{with} \quad \mu = \frac{1}{2} LC \quad \text{and} \quad p_1 \leq C|f(0)|. \tag{13.13}$$

The question then is for which P we can conclude that the implication

$$p_1 \leq C|f(0)| < P \implies \sum_{n=1}^{\infty} p_n < r \tag{13.14}$$

holds. If so then $x_n \in B_r$ for all $n \in \mathbb{N}$, x_n converges to a limit $\bar{x}$ which is also in B_r, and $f(x_n) \to 0$.

The larger P, the stronger the statement in the sense that larger values of $|f(0)|$ are allowed if we try to find a solution $x \in B_r$ of $f(x) = 0$ by means (13.9) starting from $x_0 = 0$. If we take equalities in (13.13) and (13.14) then

$$p_n = \mu p_{n-1}^2 \quad \text{for} \quad n \in \mathbb{N}; \quad p_1 = P\,; \quad \sum_{n=1}^{\infty} p_n = r. \tag{13.15}$$

Putting $\xi_n = \mu p_n$ so that $\xi_n = \xi_{n-1}^2$, this is equivalent to

$$G(\mu P) = \mu r \quad \text{with} \quad G(\xi) = \xi + \xi^2 + \xi^4 + \xi^8 + \xi^{16} + \cdots . \tag{13.16}$$

This defines P as a function of μ and r.

Exercise 13.5. Use

$$G(\xi) < \frac{\xi}{1-\xi}$$

to show that

$$|f(0)| \le \frac{2r}{(2+rLC)C}$$

guarantees $x_n \to \bar{x} \in B_r$ with $f(\bar{x}) = 0$.

Back to Heron's method. We we can scale the whole Heron procedure and put $x = y\sqrt{2}$, and likewise for $\tilde{x}, x_n, x_{n-1}$, to obtain

$$y_n = \frac{1}{2}(y_{n-1} + \frac{1}{y_{n-1}}),$$

which has $y_n \to 1$ as $n \to \infty$ if we start from $y_0 > 0$ with $y_0 \neq 1$.

Exercise 13.6. Put $y = 1 + z$ and see what you get for the sequence z_n to understand why the convergence is so fast.

Exercise 13.7. Put $e = x^2 - 2$, rewrite (2.2) in terms of e and $\tilde{e}$, examine the sequence e_n, and compare to Exercise 13.6.

14 Back to calculus

Most of this chapter should be part of any calculus course.

14.1 More on exp and ln

Exercise 14.1. Let $I \subset \mathbb{R}$ be an open interval, $F : I \to \mathbb{R}$ differentiable, $F'(x) = F(x)$ for all $x \in I$ and $(a, b) \subset I$ a maximal open interval on which $F(x) > 0$. Then $(a, b) = I$. Prove this via

$$F'(x) = F(x) \iff \frac{F'(x)}{F(x)} = 1 \iff \ln(F(x)) = x + C \iff F(x) = e^{x+c}.$$

Exercise 14.2. Same question as in Exercise 14.1 for $F : I \to \mathbb{R}$ satisfying $F'(x) = F(x)g(x)$ with $g : I \to \mathbb{R}$ continuous. Also solve the differential equation. Hint: use a primitive G of g.

Exercise 14.3. For $\alpha \in \mathbb{R}$ the function $F_\alpha : (-1, \infty) \to \mathbb{R}^+$ defined by $F_\alpha(x) = (1 + x)^\alpha$ solves $(1 + x)F'(x) = \alpha F(x)$, a differential equation like in Exercise 14.2. Determine a power series solution of the form

$$1 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots .$$

Write (the coefficients in) the solution in a form which for $\alpha = n \in \mathbb{N}$ reduces to Newton's binomium.The radius of convergence (for $\alpha \notin \mathbb{N}_0$) is $R = 1$. Why? How does it follow that for $|x| < 1$ the power series[1] just computed is equal to $F_\alpha(x)$?

14.2 Integrals with parameters

The results in this section will be needed later. They may be postponed, but at the risk of never being done at all. Let's consider

$$j(t) = \int_0^1 f(x, t)\, dx$$

[1] NB Take note of $\alpha = -1$, but also of $\alpha = \pm\frac{1}{n}$.

in which, for each t in a t-interval $[0,1]$, the function $x \to f(t,x)$ is continuous on the x-interval $[0,1]$. Then $j(t)$ is well-defined. What do we need to have j differentiable? Let's examine a follow your nose argument for what (the one-sided) derivative $j'(0)$ should be and see what we need to prove it.

If we use the mean value theorem in the form of Theorem 10.7 itself[2], for every fixed $x \in [0,1]$ applied to $t \to f(x,t)$, it follows that

$$f(t,x) = f(0,x) + f_t(\tau,x)\,t,$$

with $\tau = \tau(x) \in (0,t)$. This requires, for every x, differentiability of $f(x,t)$ on $[0,1]$ with respect to t, or on a smaller interval that contains $t = 0$ but does not depend on x. We can then write

$$f(t,x) = f(0,x) + f_t(0,x)t + \underbrace{f_t(\tau(x),x) - f_t(0,x)}_{R(t,x)}. \tag{14.1}$$

This defines $R(t,x)$. If in (14.1), with t fixed, everything is continuous in x, it follows that

$$\begin{aligned} j(t) = \int_0^1 f(t,x)\,dx &= \int_0^1 (f(0,x) + f_t(0,x)t + R(t,x))\,dx \\ &= j(0) + t\int_0^1 f_t(x,0)\,dx + \int_0^t R(t,x)\,dx \qquad (14.2) \\ &= j(0) + t\int_0^1 f_t(x,0)\,dx + r(t). \end{aligned}$$

Here

$$r(t) = \int_0^t R(t,x)\,dx \quad \text{with} \quad R(t,x) = (\underbrace{f_t(\tau(x),x) - f_t(0,x)}_{<\varepsilon\,?})t.$$

If we assume that

$$x \to f(t,x)$$

and

$$x \to f_t(0,x) = g(t,x)$$

are continuous on $[0,1]$ we don't have to worry about existence of the integrals. The integral $r(t)$ of $R(t,x)$ in (14.2) is then also continuous. The second expression with $\tau(x) \in (0,t)$ above (14.1) can now be used to establish $r(t) = o(t)$ as $t \to 0$.

[2]The integral form would require the use of not yet discussed double integrals.

Indeed, for the remainder term $r(t)$ we need $|r(t)| < \varepsilon t$ for t sufficiently small. Thus, if for $f_t(t,x) = g(t,x)$ it holds that

$$|g(t,x) - g(0,x)| < \varepsilon \tag{14.3}$$

if $t \in (0,\delta)$ for all $x \in [0,1]$ simultaneously for some $\delta > 0$, we will be happily done.

How can this uniform ε-statement fail to be true? Only if for some $\varepsilon > 0$ there exists a sequence of points (t_n, x_n) with $0 < t_n \to 0$ for which

$$|g(t_n,x_n) - g(0,x_n)| \geq \varepsilon.$$

But then the sequence x_n has a convergent subsequence x_{n_k} with limit $\bar{x} \in [0,1]$ and both sequences of points (t_n, x_n) and of points $(0, x_n)$ converge to $(0,\bar{x})$ preventing $(t,x) \to g(t,x)$ from being continuous in every point $(0,x)$ with $x \in [0,1]$. We have proved the following theorem.

Theorem 14.4. *Not so easy to memorise, let $(t,x) \to f(t,x)$ be defined for all $x \in [a,b] \subset \mathbb{R}$, with $a < b$, and all $t \in (t_0 - \delta, t_0 + \delta)$, with $t_0 \in \mathbb{R}$ and $\delta > 0$. Assume that for fixed $t \in (t_0 - \delta, t_0 + \delta)$ the function $x \to f(t,x)$ is continuous on $[a,b]$ and thus that*

$$j(t) = \int_a^b f(t,x)\,dx$$

exists. If for every fixed $x \in [a,b]$ the function $t \to f(t,x)$ is differentiable on $(t_0 - \delta, t_0 + \delta)$ and $(t,x) \to f_t(t,x)$ is continuous in every (t_0,x) with $x \in [a,b]$, then $t \to j(t)$ is differentiable in t_0 with derivative

$$j'(t_0) = \int_a^b f_t(t_0,x)\,dx.$$

Theorem 14.5. *A weaker statement easier to memorize: if f and f_t exist as continuous functions on $I \times [a,b]$, with I some t-interval, then $j : I \to \mathbb{R}$ is continuously differentiable with derivative*

$$j'(t) = \int_a^b f_t(t,x)\,dx.$$

Exercise 14.6. To prove the continuity of the derivative you need to prove: $t \to \int_a^b g(t,x)\,dx$ is continuous on I if $(t,x) \to g(t,x)$ is continuous on $I \times [a,b]$. Hint: use a uniform ε-argument.

14.3 Partial integration and Taylor polynomials

Theorem 14.7. *Let a real valued function f be twice continuously differentiable in a neighbourhood of $x = 0$, and $f(0) = 0$ and $f'(0) = 0$. Then*

$$f(x) = \int_0^x (x - s) f''(s)\, ds$$

for x in that neighbourhood.

This theorem follows from what we discuss below and is a special case of Exercise 14.10 below. You may consider to go for a direct proof instead, so that you can skip the rest of this section, which should be part of any calculus course. Theorem 14.7 is not really essential for the analysis of Newton's method in Chapter 13.2, but it is for the proof of Morse' Lemma in Chapter 16.

No new analysis is required for what follows. Via Theorem 10.12 the Leibniz rule in Theorem 11.1 has an immediate and important counter part which we state for continuously differentiable functions

$$x : [\alpha, \beta] \to \mathbb{R} \quad \text{and} \quad y : [\alpha, \beta] \to \mathbb{R}$$

as

$$\int_\alpha^\beta x(t)y'(t)\, dt = [x(t)y(t)]_\alpha^\beta - \int_\alpha^\beta x'(t)y(t)\, dt. \tag{14.4}$$

This *integration by parts formula* can and should never be forgotten. If you tend to forget important formulas do remember that it follows from Theorem 10.12 applied to a product of two continuously differentiable functions[3].

Here's a nice application. For given $f \in C([0, 1])$ we ask for a function u such that

$$-u''(x) = f(x) \quad \text{for all} \quad 0 \le x \le 1, \quad \text{and} \quad u(0) = u(1) = 0. \tag{14.5}$$

Taking the primitive on both sides we

$$u'(x) = u'(0) - \underbrace{\int_0^x f(s)\, ds}_{F(x)},$$

in which $u'(0)$ is unknown, and F a primitive of f with $F(0) = 0$. Taking primitives once more we have

$$u(x) = u'(0)x - \int_0^x F(s)\, ds,$$

[3]And in a much more general setting in fact.

with $u'(0)$ still unknown, $x \to \int_0^x F(s)\,ds$ the primitive of F which is 0 in $x = 0$, and $u(1) = 0$ not used yet.

Leibniz' product rule turns $F(s)$ into

$$\underbrace{1\,F(s)}_{G'(s)F(s)} = \underbrace{(s-a)'}_{G'(s)}F(s) = (\underbrace{(s-a)}_{G(s)}F(s))' - \underbrace{(s-a)}_{G(s)}F'(s)$$

$$= \underbrace{((s-a)F(s))'}_{(G(s)F(s))'} - \underbrace{(s-a)f(s)}_{G(s)F'(s)},$$

in which $1 = G'(s)$ with $G(s) = s - a$ and a free to choose.

The primitive of $F(x)$ then rewrites as

$$\int_0^x F(s)\,ds = [(s-a)F(s)]_0^x - \int_0^x (s-a)f(s)\,ds = \int_0^x (x-s)f(s)\,ds. \quad (14.6)$$

With $a = x$ it follows that

$$u(x) = u'(0)x - \int_0^x (x-s)f(s)\,ds$$

and $x = 1$ gives

$$u'(0) = \int_0^1 (1-s)f(s)\,ds.$$

Therefore

$$u(x) = \int_0^1 (1-s)f(s)\,ds\,x - \int_0^x (x-s)f(s)\,ds$$

$$= x\int_x^1 (1-s)f(s)\,ds + (1-x)\int_0^x sf(s)\,ds = \int_0^1 A(x,s)f(s)\,ds.$$

The expression

$$A(x,s) = \begin{cases} (1-x)s & \text{for } 0 \le s \le x \\ (1-s)x & \text{for } x \le s \le 1 \end{cases} \quad (14.7)$$

is called the kernel for the solution operator, which gives u in terms of f as

$$u(x) = \int_0^1 A(x,s)f(s)\,ds. \quad (14.8)$$

You may prefer to memorize the integration by parts formula as

$$\int_a^b F(x)G'(x)\,dx = [F(x)G(x)]_a^b - \int_a^b F'(x)G(x)\,dx. \quad (14.9)$$

It's handy for computing integrals, but also for taking primitives of primitives, as we just saw and see again below.

Exercise 14.8. For $f \in C([a,b])$ define

$$F_1(x) = F(x) = \int_a^x f(s)\,ds \quad \text{and} \quad F_2(x) = \int_a^x F_1(s)\,ds.$$

Use (14.9) to show that

$$F_2(x) = \int_a^x (x-s)f(s)\,ds.$$

Hint: the integration variabele is s and 1 is the derivative with respect to s of $s-x$.

Exercise 14.9. In the context of Exercise 14.8 let

$$F_{n+1}(x) = \int_a^x F_n(s)\,ds \quad (n = 1, 2, 3\dots).$$

Show that

$$F_n(x) = \frac{1}{(n-1)!}\int_a^x (x-s)^{n-1} f(s)\,ds.$$

Hint: for F_3 you need two integrations by parts, for F_4 three, et cetera.

Exercise 14.10. Modify the scheme in Exercise 14.9 as

$$F_0(x) = f(x), \quad F_n(x) = b_n + \int_a^x F_{n-1}(s)\,ds \quad (n = 1, 2, 3\dots), \tag{14.10}$$

and give a similar formula for $F_n(x)$ with more terms. By construction $F_n(a) = b_n$, $F_n'(a) = b_{n-1}$, $F_n''(a) = b_{n-2}$, $\dots$, and what you see is the Taylor approximation of order $n-1$ for a function whose first $n-1$ deravitives in a are given by the b's. Verify for every n times continuously differentiable function defined on an interval I which contains 0 that for all $x \in I$ it holds that

$$f(x) = f(a) + f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2!} + \cdots + f^{(n-1)}(a)\frac{(x-a)^{n-1}}{(n-1)!}$$

$$+\frac{1}{(n-1)!}\int_a^x (x-s)^{n-1} f^{(n)}(s)\,ds.$$

The last term is the remainder term. Let $M = M_n(x,a)$ and $m = m_n(x,a)$ be the maximum and minimum of $f^{(n)}(s)$ as s varies from $s = a$ to $s = x$. Then this term is between

$$\frac{M}{n!}(x-a)^n \quad \text{en} \quad \frac{m}{n!}(x-a)^n.$$

It follows that for some $s = \sigma$ between $s = a$ and $s = x$ the remainder terms is equal to

$$\frac{f^{(n)}(\sigma)}{n!}(x-a)^n.$$

So

$$f(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}(0)}{k!}(x-a)^k + \underbrace{\frac{f^{(n)}(\sigma)}{n!}(x-a)^n}_{\frac{1}{(n-1)!}\int_a^x (x-s)^{n-1} f^{(n)}(s)\,ds.} \tag{14.11}$$

for some σ between a and x.

The result in (14.11) holds in fact without the assumption that $f^{(n)}$ is continuous, with σ strictly between a and x, as a clever application of Theorem 10.7 shows. The case $n = 1$ reduces to Theorem 10.7.

14.4 Asymptotic formulas

This is not part of every standard calculus course. The notation

$$f(x) \sim g(x) \quad for \quad x \to a \tag{14.12}$$

means that

$$\frac{f(x)}{g(x)} \to 1 \quad \text{if} \quad x \to a,$$

in which often a is 0 or ∞. Similarly the statement

$$n! \sim (\frac{n}{e})^n \sqrt{2\pi n} \quad \text{as} \quad n \to \infty \tag{14.13}$$

means that the limit of the quotient of the terms on both sides of the *twiddle* is 1.

Exercise 14.11. Investigate $f : x \to x^x$ with $x \in \mathbb{R}^+$ using (11.19). Determine $g : \mathbb{R}^+ \to \mathbb{R}$ as simple as possible such that

$$f(x) - 1 \sim xg(x)$$

as $x \to 0$, i.e.

$$\frac{f(x)-1}{xg(x)} \to 1.$$

Put $f(0) = 1$. Is f differentiable from the right in $x = 0$?

Exercise 14.12. Since x^x is strictly increasing in x for x sufficiently large, $x \to x^x$ has an inverse function $y \to f(y)$ definied for y sufficiently large. Show that f is defined by $x \ln x = \ln y$, take $\ln x$ to the other side and use the resulting formula in the right hand side to get a simple $g(y)$ for which

$$f(y) \sim \frac{\ln y}{g(y)}$$

as $y \to \infty$.

14.5 Exercises

Exercise 14.13. Discuss the following formulas.

$$\int_\alpha^\beta x(t) \underbrace{y'(t)\,dt}_{dy} = [x(t)y(t)]_\alpha^\beta - \int_\alpha^\beta y(t) \underbrace{x'(t)\,dt}_{dx},$$

$$\int_\alpha^\beta \underbrace{F'(x(t))}_{f(x(t))} x'(t)\,dt = F(x(\beta)) - F(x(\alpha)) = \int_{x(\alpha)}^{x(\beta)} F'(x)\,dx = \int_a^b \underbrace{F'(x)}_{f(x)}\,dx,$$

$$\int_a^b f(x)\,dx = \int_\alpha^\beta f(x(t))x'(t)\,dt. \tag{14.14}$$

Exercise 14.14. Compute

$$\int_0^\infty \exp(-x)\,dx, \quad \int_0^\infty x\exp(-x)\,dx, \quad \int_0^\infty x^2\exp(-x)\,dx, \quad \int_0^\infty x^3\exp(-x)\,dx,$$

and derive an integral formula for $n!$ These are improper integrals, defined via

$$\int_0^\infty = \lim_{R\to\infty} \int_0^R.$$

Exercise 14.15. Sketch the graph $y = x^n e^{-x}$ (for n not too large) in the x, y-plane. Where's the top of the moumtain?

Exercise 14.16. Scale and shift the integral for $n!$ to conclude that

$$n! = (\frac{n}{e})^n \int_{-n}^{\infty} g_n(x)\, dx$$

with

$$g_n(x) = (1 + \frac{x}{n})^n e^{-x}$$

Sketch the graph defined by $y = g_n(x)$.

Exercise 14.17. Write

$$g_n(x) = e^{-\psi_n(x)} \quad \text{met} \quad \psi_n(x) = -\ln(g_n(x)),$$

and verify that

$$\psi_n(x) = x - n\ln(1 + \frac{x}{n}) = n(\frac{x}{n} - \ln(1 + \frac{x}{n})) = n\psi_1(\frac{x}{n}).$$

Put $x = s\sqrt{n}$ to conclude that

$$n! = (\frac{n}{e})^n \sqrt{n} \int_{-\sqrt{n}}^{\infty} e^{-n\Psi(\frac{s}{\sqrt{n}})} ds \tag{14.15}$$

and show that

$$\int_{-\sqrt{n}}^{\infty} e^{-n\Psi(\frac{s}{\sqrt{n}})} ds \to \int_{-\infty}^{\infty} e^{-\frac{1}{2}s^2} ds \tag{14.16}$$

as $n \to \infty$.

15 Implicit functions

If a function of two real variables, say

$$\mathbb{R}^2 = \{(x, y) : x, y \in \mathbb{R}\} \xrightarrow{F} \mathbb{R},$$

satisfies $F(0, 0) = 0$, then the equation

$$F(x, y) = 0 \tag{15.1}$$

usually has more solutions near $(x, y) = (0, 0)$. How do we find these other solutions? This chapter formulates an approach which generalises to the more general setting of $F : X \times Y \to Z$ for complete metric vector spaces X, Y and Z.

A special case is

$$F(x, y) = g(y) - x, \tag{15.2}$$

when the question concerns a possible inverse function f of a given function g, see Section 8.5. Note that for notational convenience we have then interchanged the roles of f and g and ask about the solution y of $g(y) = x$ rather than the solution x of $f(x) = y$. More important: we now choose for a local perspective and want to make assumptions that concern values of x and y close to 0 only. In Section 11.3, where we already had a global inverse, we also asked about behaviour in a single point.

In this chapter we ask both about the existence of an implicit function f, as well as its properties, but only near a given point. Thus we want to solve $F(x, y) = 0$ for given x close to $x = 0$, hoping that near $y = 0$ precisely one solution $y = f(x)$ can be shown to exist.

Before we formulate a local implicit function theorem we discuss Newton's method for solving equations[1]. We assume that for fixed x near $x = 0$ the function

$$y \to F(x, y)$$

is differentiable near $y = 0$. The derivative is denoted by $F_y(x, y)$. The special case $F(x, y) = g(y) - x$ with partial derivative $F_y(x, y) = g'(y)$ is not really different, and will lead to a local *inverse function* theorem.

For fixed x we take $y_0 = 0$ as starting value for Newton's method. Thus we put the linear expansion of $F(x, y)$ around $y = 0$ equal to 0, solve for $y = y_1$, and use the linear the linear expansion of $F(x, y)$ around $y = y_1$ to find y_2, and so on. In every step we need $F_y(x, y_{n-1})$ to be invertible[2]. The

[1] Fast convergence of this method will be shown in Section 13.2.

[2] Think of $F_y(x, y_{n-1})$ as the map $h \to F_y(x, y_{n-1})h$.

next y_n is uniquely defined by

$$F(x, y_{n-1}) + F_y(x, y_{n-1})(y_n - y_{n-1}) = 0.$$

For $n = 1, 2, \ldots$ we have

$$y_n = y_{n-1} - F_y(x, y_{n-1})^{-1} F(x, y_{n-1}), \quad \text{starting from} \quad y_0 = 0. \tag{15.3}$$

If this process, which is called Newton's method, defines a convergent sequence y_n, the x-dependent limit y defines a so-called implicit function

$$x \to y = f(x). \tag{15.4}$$

We then expect/hope that

$$F(x, f(x)) = 0, \tag{15.5}$$

and that $y = f(x)$ is the only solution of (15.1) near $y = 0$. If so we also ask which conditions will make f continuous and differentiable in $x = 0$.

15.1 A simpler version of Newton's method

A direct proof of (fast) convergence of the sequence y_n defined by (15.3) was given in Chapter 13.2 via an estimate of the form

$$|y_{n+1} - y_n| \leq C|y_n - y_{n-1}|^2 \tag{15.6}$$

and required a condition on the second derivative[3] of $y \to F(x, y)$. Here we avoid second derivatives of $y \to F(x, y)$ by simplifying the scheme: the derivative $F_y(x, y_{n-1})$ that has to be inverted in every step of Newton's scheme is replaced by $F_y(0, 0)$. The modified scheme reads

$$y_n = y_{n-1} - F_y(0, 0)^{-1} F(x, y_{n-1}), \tag{15.7}$$

and we look for an estimate which is very much like the estimate (3.6) for Heron's sequence: we lose the square in (15.6) but have to make sure that $C < 1$. To this end

a sufficienctly small bound on $|F(x, 0|$,
the invertibility of $F_y(0, 0)$,
and the continuity of $(x, y) \to F_y(x, y)$

will suffice.

[3] In fact Lipschitz continuity of $y \to F_y(x, y)$ will suffice, see Section 13.2.

Theorem 15.1. *Let $\bar{\delta} > 0$, $\bar{\varepsilon} > 0$,*

$$B = \{x \in \mathbb{R} : |x| < \bar{\delta}\}, \quad C = \{y \in \mathbb{R} : |y| < \bar{\varepsilon}\},$$

and suppose that $F : B \times C \to \mathbb{R}$ has the properties that

$F(0,0) = 0$;
$x \to F(x,0)$ is continuous in $x = 0$;
$(x,y) \to F_y(x,y)$ is continuous in $(0,0)$;
$F_y(0,0)$ is invertible;
$y \to F_y(x,y)$ is continuous on C for every $x \in B$.

Then there exists $\delta_0 > 0$ and $\varepsilon_0 > 0$ for which the statement

$$\forall (x,y) \in \bar{B}_{\delta_0} \times \bar{B}_{\varepsilon_0} : \quad F(x,y) = 0 \iff y = f(x)$$

holds, in which

$$B_{\delta_0} = \{x \in X : |x| \le \delta_0\}, \quad B_{\varepsilon_0} = \{y \in Y : |y| < \varepsilon_0\},$$

and $f : \bar{B}_{\delta_0} \to B_{\varepsilon_0}$ is constructed via (15.7) starting from $y_0 = 0$. In particular $f(0) = 0$ and f is continuous in 0.

In the proof we avoid a direct application of Theorem 3.16, which requires a map from a suitable closed and bounded set containing $y = 0$ to itself. Instead we focus on the single x-dependent sequence defined by (15.7) starting from $y_0 = 0$ only. Note that the unlikely event that $y_1 = y_0 = 0$ occurs only when $y = y_0 = 0$ and then automatically $y_0 = y_1 = y_2 = \cdots = 0$ solves $F(x,y) = 0$.

15.2 Estimating the steps: convergence

How large can y_1 be if $F(x,y_0) = F(x,0) \neq 0$? If we set

$$M_0 = |F_y(0,0)^{-1}| > 0. \tag{15.8}$$

then[4]

$$|y_1| = |F_y(0,0)^{-1} F(x,0)| \le M_0 \, |F(x,0)|. \tag{15.9}$$

If $F(x,y_1)$ is defined we can estimate the next step by

$$|y_2 - y_1| = |F_y(0,0)^{-1} F(x,y_1)| \le M_0 \, |F(x,y_1)|$$

[4]For future purposes we only use $|F_y(0,0)^{-1} k| \le M_0 \, |k|$.

using (15.7) with $n = 2$. The trick however is to use (15.7) with both $n = 1$ *and* $n = 2$ via

$$y_2 - y_1 = y_1 - F_y(0,0)^{-1}F(x,y_1) - y_0 + F_y(0,0)^{-1}F(x,y_0)$$

$$= F_y(0,0)^{-1}\left(F(x,y_0) - F(x,y_1) + F_y(0,0)y_1 - F_y(0,0)y_0\right),$$

in which we "factored" out $F_y(0,0)^{-1}$.

The first two terms in the remaining large factor are

$$F(x,y_0) - F(x,y_1) = \int_0^1 F_y(x, ty_0 + (1-t)y_1)\, dt\, (y_0 - y_1),$$

an integral we get by applying (13.1), the mean value theorem in integral form[5], to $y \to F(x,y)$ with $a = y_1$ and $b = y_0$, x fixed. Combined with the third and fourth term the whole large factor equals[6]

$$\int_0^1 \left(F_y(x, ty_0 + (1-t)y_1) - F_y(0,0)\right)\, dt\, (y_0 - y_1),$$

in which we brought the other two terms inside the integral. We conclude that

$$y_2 - y_1 = F_y(0,0)^{-1} \int_0^1 \left(F_y(x, ty_0 + (1-t)y_1) - F_y(0,0)\right)\, dt\, (y_0 - y_1)$$

if $y \to F_y(x,y)$ is continuous on[7]

$$[y_0, y_1] = \{ty_0 + (1-t)y_1 :\ 0 \le t \le 1\} \tag{15.10}$$

for fixed x. Therefore

$$|y_2 - y_1| \le M_0 \int_0^1 |(F_y(x, ty_0 + (1-t)y_1) - F_y(0,0)|\, dt\, |y_0 - y_1|. \tag{15.11}$$

We now ask that $(x,y) \to F_y(x,y)$ is continuous[8] in $(0,0)$. In particular this continuity requires the existence of $F_y(x,y)$ for (x,y) close to $(0,0)$. To be precise we assume that for every $\eta > 0$ an $\varepsilon > 0$ can be found such that for all x and y the implication

$$|x| \le \varepsilon \text{ en } |y| \le \varepsilon \implies |F_y(x,y) - F_y(0,0)| < \eta \tag{15.12}$$

[5]Which will also do for $F : X \times Y \to Y$.

[6]Look at (13.6), this argument is not restricted to $F : \mathbb{R}^2 \to \mathbb{R}$!

[7]This notation for $[y_0, y_1]$ does not require $y_0 < y_1$.

[8]For $F(x,y) = g(y) - x$ this means g' continuous in 0.

holds. Note that instead of an ε-δ-statement we used an η, ε-statement of continuity, with nonstrict inequalities on the left hand side of the implication arrow. In the end we want to have that $y = f(x)$, the limit of the x-dependent sequence y_n, satisfies $|y| < \varepsilon$ for all x with $|x| \leq \delta$, for some $\delta > 0$ depending on $\varepsilon > 0$ via the continuity of $x \to f(x, 0)$, and $\varepsilon > 0$ in turn depending on some $\eta > 0$ to be chosen to make what follows work

From (15.11) and (15.12) we have that $|x|, |y_0|, |y_1| \leq \varepsilon$ implies

$$|y_2 - y_1| < M_0\eta\,|y_1 - y_0| \quad \text{in which} \quad M_0 = |F_y(0,0)^{-1}| > 0.$$

The inequality is strict unless $y_0 = y_1$, which is why we assumed $y_0 \neq y_1$. Thus the second step has

$$|y_2 - y_1| < \theta\,|y_1 - y_0| = \theta\,|y_1| \quad \text{with} \quad \theta = M_0\eta.$$

By the same reasoning we have

$$|y_3 - y_2| \leq \theta\,|y_2 - y_1|,$$

provided $|y_2| < \varepsilon$, and so on.

Any $\theta < 1$ is now fine for our purposes[9]: as long as $|y_n| < \varepsilon$ it holds that[10]

$$|y_{n+1}| = |y_{n+1} - y_0| \leq \underbrace{|y_{n+1} - y_n|}_{\leq \theta|y_n - y_{n-1}|} + \cdots + \underbrace{|y_2 - y_1|}_{< \theta|y_1|} + |y_1| <$$

$$(\theta^n + \cdots + 1)\,|y_1| < \frac{|y_1|}{1-\theta} \leq \frac{M_0|F(x,0)|}{1-\theta},$$

so

$$|y_{n+1}| < \frac{M_0|F(x,0)|}{1-\theta} < \frac{M_0\tilde{\varepsilon}}{1-\theta} = \frac{M_0\tilde{\varepsilon}}{1 - M_0\eta} \tag{15.13}$$

if $|x| \leq \tilde{\delta}$. Here $\tilde{\varepsilon} > 0$ is still to be chosen and $\tilde{\delta} > 0$ corresponds to $\tilde{\varepsilon}$ via the definition[11] of continuity of $x \to F(x, 0)$ in $x = 0$.

Now choose

$$\eta_0 < \frac{1}{M_0}, \tag{15.14}$$

and then, given the corresponding ε_0 as in (15.12), a positive $\tilde{\varepsilon}_0$ such that

$$\frac{M_0\tilde{\varepsilon}_0}{1 - M_0\eta_0} < \varepsilon_0, \quad \text{i.e.} \quad \tilde{\varepsilon}_0 < (\frac{1}{M_0} - \eta_0)\varepsilon_0.$$

[9] In (3.6) we chose $\theta = \frac{1}{2}$ for the sake of simplicity only.

[10] In view of $1 + \theta + \theta^2 + \cdots = \frac{1}{1-\theta}$, see Section 1.4.

[11] With $\leq \tilde{\delta}$ instead of $< \tilde{\delta}$.

Then let $\tilde{\delta}_0 > 0$ correspond to $\tilde{\varepsilon}_0 > 0$ via the definition[12] of continuity of $x \to F(x,0)$ in $x = 0$.

Thus the chain of alternating choices and continuity arguments is

$$M_0 = |F_y(0,0)^{-1}| \quad \xrightarrow{choose} \quad \eta_0 < \frac{1}{M_0} \quad \xrightarrow[continuous\ in\ (0,0)]{(x,y)\to F_y(x,y)} \quad \varepsilon_0$$

$$\xrightarrow{choose} \quad \tilde{\varepsilon}_0 < (\frac{1}{M_0} - \eta_0)\varepsilon_0 \quad \xrightarrow[continuous\ in\ 0]{x\to F(x,0)} \quad \tilde{\delta}_0$$

and we finally let

$$\delta_0 = \min(\delta_0, \varepsilon_0).$$

Then the x-dependent sequence y_n converges to a limit for every x with $|x| \le \delta_0$, and the x-dependent limit $y = f(x)$ satisfies $|f(x)| < \varepsilon_0$.

Note that we used the map

$$y \xrightarrow{\Phi} y - F_y(0,0)^{-1}F(x,y), \tag{15.15}$$

and the estimate

$$|\Phi(x,y) - \Phi(x,\tilde{y})| \le \theta\,|y - \tilde{y}| \tag{15.16}$$

with $\theta < 1$ and strict inequality if $y \neq \tilde{y}$. Equation $F(x,y) = 0$ is via (15.15) equivalent to $y = \Phi(x,y)$ because $F_y(0,0)^{-1}$, being the inverse of $F_y(0,0)$, is invertible. For the limit $y = f(x)$ the continuity[13] of $y \to \Phi(x,y)$ implies

$$y = \lim_{n\to\infty} y_{n+1} = \lim_{n\to\infty} \Phi(x, y_n) = \Phi(x,y).$$

Thus

$$\forall (x,y) \in \bar{B}_{\delta_0} \times \bar{B}_{\varepsilon_0} : \quad F(x,y) = 0 \iff y = f(x), \tag{15.17}$$

and Theorem 15.1 is proved.

15.3 Differentiable implicit functions

The implicit function in Theorem 15.1 satisfies

$$|f(x)| \le \frac{M_0|F(x,0)|}{1 - M_0\eta_0}, \tag{15.18}$$

[12] With $\le \tilde{\delta}_0$ instead of $< \tilde{\delta}_0$.

[13] Continuity follows from differentiability.

in which η_0 was chosen at the beginning of Section 15.2, see (15.14). Estimate (15.18) immediately implies the continuity of f in 0 in view of the assumptions on $x \to F(x,0)$. What do we need to conclude that f is differentiable in 0?

Use (13.1) to write

$$0 = F(x, f(x)) = F(x,0) + F(x, f(x)) - F(x,0)$$

$$= F(x,0) + \int_0^1 F_y(x, tf(x)) f(x)\, dt = F(x,0) + F_y(0,0) f(x) + R(x),$$

$$\text{with} \quad R(x) = \int_0^1 (F_y(x, tf(x)) - F_y(0,0)) f(x)\, dt. \tag{15.19}$$

Clearly $x \to F(x,0)$ differentiable in $x = 0$ is the natural additional assumption, because then

$$0 = F(x, f(x)) = F_x(0,0)x + r(x) + F_y(0,0) f(x) + R(x), \tag{15.20}$$

with $r(x) = o(|x|)$ as $x \to 0$.

Theorem 15.2. *Let f be as in Theorem 15.1. If $x \to F(x,0)$ is differentiable in $x = 0$ then also f is differentiable in $x = 0$ and*

$$f'(0) = -F_y(0,0)^{-1} F_x(0,0).$$

The proof now follows the nose. Isolating $f(x)$ in (15.20) we have

$$f(x) = \underbrace{-F_y(0,0)^{-1} F_x(0,0)}_{f'(0)?}\, x \underbrace{-F_y(0,0)^{-1} r(x) - F_y(0,0)^{-1} R(x)}_{remainder}. \tag{15.21}$$

Since

$$|F_y(0,0)^{-1} r(x)| \le M_0 |r(x)| \quad \text{and} \quad |F_y(0,0)^{-1} R(x)| \le M_0 |R(x)|$$

it remains to be proved that $R(x) = o(|x|)$ as $x \to 0$. Given an arbitrary[14] $\varepsilon > 0$ we need to conclude that

$$|R(x)| < \varepsilon |x| \quad \text{if} \quad 0 < |x| < \delta$$

for some $\delta > 0$. Since $R(x)$ is given by (15.19) we use (15.12) again to conclude that

$$|R(x)| < \tilde{\eta}\, |f(x)| \quad \text{if} \quad |x| < \tilde{\varepsilon} \quad \text{and} \quad |f(x)| < \tilde{\varepsilon}. \tag{15.22}$$

[14]Earlier we only took one fixed ε_0 corresponding to one fixed η_0 as in (15.14).

The latter inequality will hold if $|x| < \tilde{\delta}$, $\tilde{\delta}$ corresponding to $\tilde{\varepsilon}$ in the established statement, via the construction and (15.18), that f is continuous in 0.

Restricting also to $|x| \le \delta_0$ we have

$$|R(x)| < \tilde{\eta}\,|f(x)| \le \frac{M_0\tilde{\eta}}{1 - M_0\eta_0}\,|F(x,0)|,$$

while

$$|F(x,0)| < (|F_x(0,0)| + \varepsilon_r)\,|x|$$

if $|x| < \delta_r$, where δ_r corresponds to some arbitrarily chosen but then fixed $\varepsilon_r > 0$ in the definition of $r(x) = o(|x|)$.

For given $\varepsilon > 0$ we then choose $\tilde{\eta} > 0$ such

$$\frac{M_0\tilde{\eta}}{1 - M_0\eta_0}(|F_x(0,0)| + \varepsilon_r) = \varepsilon,$$

take the corresponding $\tilde{\varepsilon}$ and $\tilde{\delta}$ as in and below (15.22). With $\delta = \min(\delta_0, \delta_r, \tilde{\delta})$ the implication

$$0 < |x| < \delta \implies |R(x)| < \varepsilon\,|x|$$

then holds. Since $\varepsilon > 0$ was arbitrary, this completes the proof that $R(x)$ and therebye the whole remainder term in (15.21) is $o(|x|)$ as $x \to 0$. This then completes the proof of Theorem 15.2.

Exercise 15.3. Actually the continuity of f in $x = 0$ follows directly from (15.21) and (15.19) if we assume that $|y| = |f(x)| \le \varepsilon_0$ with ε_0 chosen via (15.12) for (15.14). Use (15.20) in the form

$$0 = F(x,y) = F(x,0) + F_y(0,0)y + \underbrace{\int_0^1 (F_y(x,y) - F_y(0,0))y\,dt}_{\text{in norm less than } \eta_0|y| \quad \text{if } |x|,|y| \le \varepsilon_0}, \tag{15.23}$$

and derive that for solutions (x,y) of $F(x,y) = 0$ it holds that

$$|y| \le \frac{M_0|F(x,0)|}{1 - M_0\eta_0} \quad \text{if} \quad |x| \le \varepsilon_0 \text{ and } |y| \le \varepsilon_0. \tag{15.24}$$

Thus the existence of a solution of $F(x,y) = 0$ with $|y| \le \varepsilon_0$ for every x with $|x| < \delta_0 \le \varepsilon$ implies that $y \to 0$ if $F(x,0) \to 0$. Except for the choice of ε_0

this statement is independent of the construction of f and the uniqueness of the solution.

What about the other x-values in the domain $\bar{B}_{\delta_0}$ of f? We should have that f is differentiable in every x with $|x| \leq \tilde{\delta}_0$ for some $0 < \tilde{\delta}_0 < \delta_0$, and

$$f'(x) = -F_y(x, f(x))^{-1} F_x(x, f(x)). \tag{15.25}$$

For every $x \in \bar{B}_{\delta_0}$ the validity of (15.25) relies solely on the invertibility of $F_y(x, f(x))$. Note that $F_y(x, f(x))$ is continuous in $x = 0$ because F_y is continuous in $(0, 0)$ and f is continuous in 0. Since $F_y(0, f(0)) = F_y(0, 0)$ is invertible it follows that $F_y(x, f(x))$ is invertible for all x with $|x| \leq \tilde{\delta}_0 \leq \delta_0$ for some $\tilde{\delta}_0$.

The continuity of

$$x \to f'(x) = -(F_y(x, f(x)))^{-1} F_x(x, f(x))$$

in $x = x_0$ with $|x_0| \leq \tilde{\delta}_0$ requires the continuity of both $(x, y) \to F_x(x, y)$ and $(x, y) \to F_y(x, y)$ in (x_0, y_0), and the continuity of $A \to A^{-1}$ in every invertible $A_0 = F_y(x_0, y_0)$.

Theorem 15.4. *The Implicit Function Theorem. Let X, Y and Z be complete metric vector spaces, $\bar{\delta} > 0$, $\bar{\varepsilon} > 0$,*

$$B = \{x \in X : |x| < \bar{\delta}\}, \quad C = \{y \in Y : |y| < \bar{\varepsilon}\}.$$

Suppose that $F : B \times C \to Z$ is continuously differentiable, and that

$F(0, 0) = 0$*; $F_y(0, 0)$ is invertible.*

Then there exists $\tilde{\delta}_0 > 0$ and $\varepsilon_0 > 0$ for which

$$\forall (x, y) \in \bar{B}_{\tilde{\delta}_0} \times B_{\varepsilon_0} : \quad F(x, y) = 0 \iff y = f(x)$$

holds, in which

$$B_{\tilde{\delta}_0} = \{x \in X : |x| < \tilde{\delta}_0\}, \quad B_{\varepsilon_0} = \{y \in Y : |y| < \varepsilon_0\},$$

and $f : \bar{B}_{\tilde{\delta}_0} \to B_{\varepsilon_0}$ is differentiable on $\bar{B}_{\delta_0}$ with

$$x \to f'(x) = -(F_y(x, f(x)))^{-1} F_x(x, f(x))$$

continuous on $\bar{B}_{\tilde{\delta}_0}$.

This theorem builds on Theorems 15.1 and 15.2, which also hold in the general context of complete metric vector spaces. The proofs can be copy-pasted replacing absolute values by norms in X, Y, Z and provide us with δ_0 and ε_0. The existence and continuity of $f'(x)$ requires restriction to a possibly smaller $\bar{B}_{\tilde{\delta}_0}$, as explained above and formulated in the final theorem.

15.4 Application to integral equations

This concerns smooth dependence of the solution of (7.15) on ξ, and

$$x(t) = \xi + \int_0^t f(x(s))\,ds$$

as the integral equation corresponding to the differential equation $x' = f(x)$ with initial condition $x(0) = \xi$ for X-valued functions $t \to x(t)$. Assume the existence and uniform continuity of f'. Let $x = x(\xi)$ be the solution of (15.26). Then

$$\xi \to x(\xi)$$

is continuously differentiable, and x_ξ is the solution of the integral equation corresponding to

$$y'(t) = f'(x(t))y(t) \quad \text{with} \quad y(0) = 1.$$

This is a bit of a project[15]. The first steps are sketched below.

For $a, b \in \mathbb{R}$ met $0 \in [a, b]$ and $\xi \in \mathbb{R}$ introduce

$$x = \xi + \Phi(x) \quad \text{with} \quad (\Phi(x))(t) = \int_0^t f(x(s))\,ds, \tag{15.26}$$

defining a new $\Phi(x) \in C([a, b])$ given and ("old") function $x \in C([a, b])$. Theorem 15.4 is applicable if

$$\Phi : C([a, b]) \to C([a, b])$$

is continuously differentientable.

To see why and how, take $h \in C([a, b])$ and write

$$(\Phi(x + h))(t) = \int_0^t f(x(s) + h(s))\,ds = \int_0^t [f(x(s) + \tau h(s))]_0^1\,ds$$

$$= \int_0^t \int_0^1 f'(x(s) + \tau h(s))h(s)\,d\tau\,ds$$

$$= \int_0^t \int_0^1 f'(x(s))h(s)\,d\tau\,ds + \underbrace{\int_0^t \int_0^1 (f'(x(s) + \tau h(s)) - f(x(s))h(s)\,d\tau\,ds}_{R(h;x)(t)}$$

$$= (\Phi'(x)h)(t) + R(h; x)(t),$$

[15] We shall also deal with pararameters in f, e.g. $f(x, \mu, \varepsilon)$ or so, see Section 18.5.

in which

$$h \xrightarrow{\Phi'(x)} \Phi'(x)h \quad \text{with} \quad (\Phi'(x)h)(t) = \int_0^t f'(x(s))h(s)\,ds, \tag{15.27}$$

and

$$\begin{aligned}
|R(h;x)(t)| &= |\int_0^t \int_0^1 (f'(x(s)+\tau h(s)) - f(x(s))h(s)\,d\tau\,ds| \\
&\leq |\int_0^t |\int_0^1 (f'(x(s)+\tau h(s)) - f(x(s))h(s)\,d\tau|\,ds| \\
&\leq |\int_0^t |\int_0^1 |(f'(x(s)+\tau h(s)) - f(x(s))h(s)|\,d\tau|\,ds| \\
&\leq |\int_0^t |\int_0^1 \underbrace{|f'(x(s)+\tau h(s)) - f(x(s)|}_{\leq \varepsilon} \underbrace{|h(s)|}_{|h|_\infty}\,d\tau|\,ds| \leq (b-a)\varepsilon|h|_\infty
\end{aligned}$$

if $|h|_\infty \leq \delta$, with $\delta > 0$ corresponding to $\varepsilon > 0$ in the definition of uniform continuity of f'.

15.5 For later: partial differentiability $\Longrightarrow$?

Exercise 12.7 contained an example of a differentiable function $F : \mathbb{R}^2 \to \mathbb{R}$. Differentiability of F in (x_0, y_0) via linear expansion rewrites as

$$F(x,y) = F(x_0,y_0) + a(x-x_0) + b(y-y_0) + R_0(x,y),$$

with

$$|R_0(x,y)| < \varepsilon \max(|x-x_0|, |y-y_0|) \quad \text{if} \quad \max(|x-x_0|, |y-y_0|) < \delta,$$

$\delta > 0$ depending on ε.

Exercise 15.5. Put $x = x_0 + h$ and $y = y_0 + k$. Prove that

$$a = F_x(x_0,y_0) = \lim_{h\to 0} \frac{F(x_0+h,y_0) - F(x_0,y_0)}{h} = \lim_{x\to x_0} \frac{F(x,y_0) - F(x_0,y_0)}{x-x_0};$$

$$b = F_y(x_0,y_0) = \lim_{k\to 0} \frac{F(x_0,y_0+k) - F(x_0,y_0)}{k} = \lim_{y\to y_0} \frac{F(x_0,y) - F(x_0,y_0)}{y-y_0}.$$

These are called the partial derivatives of F in (x_0, y_0). It is possible for these derivatives to exist if the function is not differentiable. For instance, if $F : \mathbb{R}^2 \to \mathbb{R}$ is defined by $F(x, y) = 0$ if $xy = 0$ and $F(x, y) = 1$ if $xy \neq 0$ then $F_x(0, 0) = F_y(0, 0) = 0$, but F is not differentiable in $(0, 0)$, why?

What do we need of $x \to F(x, y)$ and $y \to F(x, y)$ to conclude that

$$F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$

is differentiable in (x_0, y_0)? We answer this question for

$$F : X \times Y \to \mathbb{R},$$

$x_0 \in X, y_0 \in Y$, and assume that $x \to F(x, y)$ and $y \to F(x, y)$ are differentiable, respectivily for fixed $y \in B_\delta(y_0)$ and fixed $x \in B_\delta(x_0)$ on $B_\delta(x_0)$ and $B_\delta(y_0)$, for some $\delta_0 > 0$.

Using Theorem 12.4 we have

$$F(x, y) = F(x_0, y_0) + F(x, y) - F(x_0, y_0) =$$

$$F(x_0, y_0) + \underbrace{F(x, y) - F(x_0, y)}_{\text{vary } x} + \underbrace{F(x_0, y) - F(x_0, y_0)}_{\text{vary } x} =$$

$$F(x_0, y_0) + F_x(\xi(y), y)(x - x_0) + F_y(x_0, \eta)(y - y_0),$$

for $x \in B_\delta(x_0)$ and $y \in B_\delta(y_0)$ with $\xi(y) \in (x_0, x)$ and $\eta \in (y_0, y)$. Therefore

$$F(x, y) = F(x_0, y_0) + F_x(x_0, y_0)(x - x_0) + F_y(x_0, y_0)(y - y_0) + R_0 \quad (15.28)$$

with remainder term

$$R_0 = (F_x(\xi(y), y) - F_x(x_0, y_0))(x - x_0) + (F_y(x_0, \eta) - F_y(x_0, y_0))(y - y_0).$$

If

$$(x, y) \to F_x(x, y) \quad \text{and} \quad y \to F_y(x_0, y)$$

are continuous in respectively (x_0, y_0) and y_0 then

$$|R_0| \leq |(F_x(\xi(y), y) - F_x(x_0, y_0))(x - x_0)| + |(F_y(x_0, \eta) - F_y(x_0, y_0))(y - y_0)| \leq$$

$$\underbrace{|F_x(\xi(y), y) - F_x(x_0, y_0)|}_{\leq \varepsilon} |x - x_0| + \underbrace{|F_y(x_0, \eta) - F_y(x_0, y_0)|}_{\leq \varepsilon} |y - y_0)|$$

$$\leq \varepsilon \max(|x - x_0|, |y - y_0|) = \varepsilon |(x, y) - (x_0, y_0)|$$

if $\delta > 0$ is sufficiently small. Thus F is differentiable in (x_0, y_0). A slightly stronger condition easier to remember is given in the following theorem.

Theorem 15.6. *Let X and Y be normed spaces. If $F : X \times Y \to \mathbb{R}$ has "partial" functions*

$$x \to F(x, y) \quad \text{en} \quad y \to F(x, y)$$

defined and differentiable for $x \in B_\delta(x_0)$ and $y \in B_\delta(y_0)$ with $x_0 \in X, y_0 \in Y, \delta > 0$, then continuity of

$$(x, y) \to F_x(x, y) \in X^* \quad \text{and} \quad (x, y) \to F_y(x, y) \in Y^*$$

in (x_0, y_0) implies that F is differentiable in (x_0, y_0), with $F'(x_0, y_0)$ defined by

$$(h, k) \xrightarrow{F'(x_0,y_0)} F_x(x_0, y_0)h + F_y(x_0, y_0)k.$$

Exercise 15.7. For X, Y, Z normed spaces and $\Phi : X \times Y \to Z$ the method via the mean value theorem fails. Write

$$\Phi(x, y) = \Phi(x_0, y_0) + \underbrace{\Phi(x, y) - \Phi(x_0, y)}_{\text{vary } x} + \underbrace{\Phi(x_0, y) - \Phi(x_0, y_0)}_{\text{vary } x}.$$

Assume Z is complete, $x \to \Phi(x, y_0)$ is continuously differentiable for $x \in X$ with $|x - x_0| < \delta_x$. If for each of these x the partial function $y \to \Phi(x, y)$ is continuously differentiable in $y \in Y$ with $|y - y_0| < \delta_y$, $\delta_x, \delta_y > 0$, and if $(x, y) \to \Phi_y(x, y)$ is continuous in (x_0, y_0), then Φ is differentiable in (x_0, y_0). Use (13.5) to prove this statement.

Exercise 15.8. If X, Y, Z are normed spaces, Z complete, and $\Phi : X \times Y \to Z$ has partial functions with partial derivatives Φ_x and Φ_y continuous on an open set O in $X \times Y$, then Φ is differentiable in every point of O and $\Phi' : O \to L(X \times Y, Z)$ is continuous and defined in every $(x_0, y_0) \in O$.

15.6 Stationary under a constraint

Suppose Φ and F are functions of x and y differentiable in $(x, y) = (0, 0)$, and f is a function of x differentiable in $x = 0$, for which it holds that

$$F_x(0, 0) + F_y(0, 0)f'(0) = 0. \tag{15.29}$$

In practice, f is the implicit function in Theorems 15.1 and 15.2. Then $y = f(x)$ describes the solution set of $F(x, y) = 0$ near $(0, 0)$, and we are interested in the restriction of Φ to the zero set of F. Clearly

$$x \xrightarrow{\phi} \phi(x) = \Phi(x, f(x))$$

is differentiable in $x = 0$, with

$$\phi'(0) = \Phi_x(0,0) + \Phi_y(0,0)f'(0). \tag{15.30}$$

If $F_y(0,0)$ is invertible it follows from (15.29) and (15.30) that

$$\phi'(0) = 0 \iff \Phi_x(0,0) = \Phi_y(0,0)F_y(0,0)^{-1}F_x(0,0). \tag{15.31}$$

Invertibility of $F_y(0,0) \in \mathbb{R}$ means that $F_y(0,0) \neq 0$, whence

$$\phi'(0) = 0 \iff \Phi_x(0,0)F_y(0,0) = \Phi_y(0,0)F_x(0,0),$$

equivalent to the existence of $\lambda \in \mathbb{R}$ for which it holds that

$$\begin{pmatrix} \Phi_x(0,0) \\ \Phi_y(0,0) \end{pmatrix} = \lambda \begin{pmatrix} F_x(0,0) \\ F_y(0,0) \end{pmatrix}.$$

This is a special case of the statement in Lagrange multiplier theorem which will be discussed elsewhere, starting from (15.31).

16 Quadratic functions and Morse' Lemma

This chapter is about a theorem which is not very special in the case of $X = \mathbb{R}$, when it says that a C^2-function $f : \mathbb{R} \to \mathbb{R}$ with $f(0) = f'(0) = 0$ is near $x = 0$ is just the function

$$x \to \frac{f''(0)}{2}x^2$$

in disguise[1], provided $f''(0) \neq 0$. But such a statement also holds for a C^3-function $f : \mathbb{R} \to \mathbb{R}$ with $f(0) = f'(0) = f''(0) = 0$ and

$$x \to \frac{f'''(0)}{6}x^6,$$

provided $f'''(0) \neq 0$, and so on.

Theorem 16.10 below does not generalise to any such other case. It can be formulated and proved exclusively for functions $F : X \to \mathbb{R}$ with $F(0) = 0$ in $\mathbb{R}$, $F'(0) = 0$ in $X^* = L(X, \mathbb{R})$, and $F''(0)$ invertible in a space to be introduced below[2]. So let X be a complete metric vector space. It's dual space X^* is the space of all Lipschitz continuous linear functions from X to $\mathbb{R}$. This space is itself a complete metric vector space, if we define the norm of $\phi \in X^*$ to be the smallest Lipschitz constant of ϕ. It is customary[3] to write

$$\langle \phi, x \rangle = \phi(x) \quad \text{for} \quad \phi \in X^* \quad \text{and} \quad x \in X.$$

For a function $F : X \to \mathbb{R}$ differentiable in $x = \xi \in X$ we thus write

$$F(x) = F(\xi) + \langle F'(\xi), x - \xi \rangle + R_\xi(x), \quad R_\xi(x) = o(|x - \xi|) \quad \text{as} \quad x \to \xi,$$

and we are interested in a local description of F near points where this holds with $F'(\xi) = 0$. For simplicity we assume that $\xi = 0$ and $F(0) = 0$.

The simplest nontrivial examples of such functions are then (purely) quadratic functions, i.e. functions $Q : X \to \mathbb{R}$ of the form

$$X \ni x \xrightarrow{Q} (Sx)(x) = \langle Sx, x \rangle \in \mathbb{R} \tag{16.1}$$

in which S is a Lipschitz continuous linear map [4]

$$X \ni x \xrightarrow{S} S(x) = Sx \in X^*$$

[1] Yes, we will make this statement explicit.

[2] We use the notation (12.3) introduced in Chapter 12.

[3] Though annoying at first.

[4] $L(X, X^*)$ is the complete metric vector space of all Lipschitz continuous linear maps

$$X \xrightarrow{S} X^*.$$

from X to X^*.

Exercise 16.1. Show that it is no restriction to assume that $\langle Sx, y\rangle = \langle Sy, x\rangle$ for all $x, y \in X$. Hint: assume that $Q(x,x) = \langle Ax, x\rangle$ with $A \in L(X, X^*)$ and write $B(x,y) = \langle Ax, x\rangle$ as in Section 17.3. Use $B(x,y)$ and $B(y,x)$ to construct such an $S \in L(X, X^*)$ with $\langle Ax, x\rangle = \langle Sx, x\rangle$.

Exercise 16.2. Show that Q is differentiable in 0 and that $Q'(0) = 0$ in X^*.

Now let $\mathcal{O} \subset X$ open, $0 \in \mathcal{O}$ and $F : \mathcal{O} \to \mathbb{R}$ differentiable, and assume $F(0) = 0$ in $\mathbb{R}$ and $F'(0) = 0$ in X^*. Under which conditions is it true that a coordinate transformation in X turns F into a quadratic function Q as in (16.1)? If so we say that F and Q are conjugate functions.

16.1 Intermezzo: second order partial derivatives

Theorem 16.3. *Let $g : \mathbb{R}^2 \to \mathbb{R}$ have partial derivatives*

$$(x,y) \to \frac{\partial g}{\partial x} = g_x(x,y) \quad \text{and} \quad (x,y) \to \frac{\partial g}{\partial y} = g_y(x,y)$$

differentiable in (x_0, y_0). Then the second order partial derivatives exist in (x_0, y_0) and

$$g_{yx}(x_0,y_0) = \frac{\partial}{\partial x}\frac{\partial g}{\partial y} = \frac{\partial}{\partial y}\frac{\partial g}{\partial x} = g_{xy}(x_0,y_0).$$

For the proof assume that $(x_0, y_0) = (0,0)$. The assumptions imply the existence of the first order partial derivatives near $(0,0)$. The differentiability of g_y in $(0,0)$ and Theorem 10.7 applied to

$$y \to g(x,y) - g(0,y)$$

for $x \neq 0$ and $y \neq 0$ small imply that for some x-dependent η between 0 and y we have

$$g(x,y) - g(0,y) - g(x,0) + g(0,0) = (g_y(x,\eta) - g_y(0,\eta))y$$

$$= (g_y(0,0) + g_{yx}(0,0)x + g_{yy}(0,0)\eta + R(x,\eta) - g_y(0,0) - g_{yy}(0,0)\eta - R(0,\eta))y$$

$$= (g_{yx}(0,0)x + R(x,\eta) - R(0,\eta))y,$$

in which

$$R(x,\eta) = o(\sqrt{x^2+\eta^2}) \quad \text{and so also} \quad R(0,\eta) = o(\eta) \quad \text{as} \quad \sqrt{x^2+\eta^2} \to 0.$$

The differentiability of

$$(x,y) \to g_y(x,y)$$

in $(0,0)$ has been used twice, with the same "remainder function" R. Since $|\eta| \le |y|$ it follows that

$$g(x,y) - g(0,y) - g(x,0) + g(0,0) = g_{yx}(0,0)xy + y\,o(r)$$

$$= g_{yx}(0,0)xy + o(r^2) = g_{xy}(0,0)xy + o(r^2) \tag{16.2}$$

for $r = \sqrt{x^2+y^2} \to 0$. The second version under (16.2) follows by interchanging the roles of x and y and implies $g_{yx}(0,0) = g_{xy}(0,0)$.

16.2 Second derivatives of functions on normed spaces

If we introduce $f(t) = F(tx)$ as a function of $t \in [0,1]$ for given small $x \in X$, then f is differentiable for t,

$$f'(t) = F'(tx)(x) = \langle F'(tx), x\rangle, \tag{16.3}$$

and $f(0) = 0 = f'(0)$ in $\mathbb{R}$. Now assume that also f' is differentiable with $f'' \in C([0,1])$. Then two integrations by parts show that

$$F(x) = f(1) = \int_0^1 (1-t)f''(t)\,dt, \tag{16.4}$$

see also Theorem 14.7.

Exercise 16.4. Give a direct proof of (16.3).

The differentiability of $t \to F'(tx)x = f'(t)$ will follow from differentiability of

$$x \to F'(x) \in X^*$$

in points ξ near 0, which means that

$$F'(x) = F'(\xi) + F''(\xi)(x-\xi) + R(x;\xi), \tag{16.5}$$

with $F''(\xi) : X \to X^*$ in $L(X, X^*)$ and

$$|R(x;\xi)|_{X^*} = o(|x-\xi|_X)$$

as $|x-\xi|_X \to 0$.

With $\xi = t_0 x$ and x replaced by tx in (16.5) this becomes

$$F'(tx) = F'(t_0x) + F''(t_0x)(tx - t_0x) + R(tx; t_0x)$$

$$= F'(t_0x) + (t-t_0)F''(t_0x)x + R(tx; t_0x)$$

in X^*, and (16.3) then gives

$$f'(t) = \langle F'(tx), x\rangle = \underbrace{\langle F'(t_0x, x\rangle}_{f'(t_0)} + (t-t_0)\langle F''(t_0x)x), x\rangle + \langle R(tx; t_0x), x\rangle.$$

We conclude that f' is differentiable in every $t \in [0,1])$ for which F' is differentiable in tx, with

$$f''(t) = \langle F''(tx)x, x\rangle. \tag{16.6}$$

Continuity of $F''(x)$ then implies the continuity of f''. So we assume that $x \to F'(x) \in L(X, X^*)$ is continuous in $\mathcal{O}$.

16.3 The second derivative as symmetric bilinear form

Theorem 16.5. *Let $x \to F'(x) \in X^*$ be differentiable in $x = \xi$. With $F''(\xi)h \in X^*$ for all $h \in X^*$ and then $(F''(\xi)h)k \in \mathbb{R}$ for all $k \in X^*$, we have that*

$$(h,k) \xrightarrow{F''(\xi)} (F''(\xi)h)k = \langle F''(\xi)h, k\rangle \in \mathbb{R} \tag{16.7}$$

is a bilinear form. This form is symmetric:

$$\langle F''(\xi)h, k\rangle = \langle F''(\xi)k, h\rangle \quad \text{for all} \quad h, k \in X^*.$$

Theorem 16.5 is proved by Exercise 16.6 and Theorem 16.3.

Exercise 16.6. For h and k in X and $x \to F'(x)$ differentiable in $x = 0$, the function

$$(s,t) \xrightarrow{g} F(sh + tk)$$

has mixed partial derivatives in $(0,0)$ given by $g_{st}(0,0) = F''(0)k\,h$ and $g_{ts}(0,0) = F''(0)h\,k$. Prove this directly from the definitions.

For $S = F''(\xi) \in L(X, X^*)$ it follows that

$$\langle Sh, k\rangle = \langle Sk, h\rangle,$$

which we see as the defining property of

$$S \in S(X, X^*) \subset L(X, X^*). \tag{16.8}$$

With also $F''(tx) \in S(X, X^*)$ we have from (16.4) that

$$F(x) = \int_0^1 (1-t)\langle F''(tx)x, x\rangle\, dt = \langle \int_0^1 (1-t)F''(tx)x\, dt, x\rangle,$$

whence

$$F(x) = \langle \int_0^1 (1-t)F''(tx)\, dt\, x, x\rangle = \langle \Phi_x x, x\rangle, \tag{16.9}$$

in which

$$\Phi_x = \int_0^1 (1-t)F''(tx)\, dt \in S(X, X^*). \tag{16.10}$$

Here we use a subscript to denote the x-dependence of the operator Φ_x which acts on X.

It follows that

$$F(x) = \frac{1}{2}\langle F''(0)x, x\rangle + \langle \int_0^1 (1-t)(F''(tx) - F''(0))\, dt\, x, x\rangle$$

$$= \langle \Phi_0 x, x\rangle + o(|x|_X^2), \tag{16.11}$$

as $|x|_X \to 0$ if F'' is continuous in $x = 0$. The quadratic function defined by

$$Q_0(x, x) = \langle \Phi_0 x, x\rangle = \frac{1}{2}\langle F''(0)x, x\rangle \tag{16.12}$$

=the obvious candidate for a conjugate to

$$F(x) = \langle \Phi_x x, x\rangle = \int_0^1 (1-t)\langle F''(tx)x, x\rangle\, dt.$$

Exercise 16.7. Check that ontinuity of F'' in 0 means that for every $\varepsilon > 0$ a $\delta > 0$ exists such that

$$0 < |x|_X < \delta \implies |(F''(x) - F''(0))y|_{X^*} < \varepsilon |y|_X$$

for all $0 \neq y \in X$.

Exercise 16.8. Show that

$$Q_0(x,x) = F_{x_1x_1}(0,0)x_1^2 + 2F_{x_1x_2}(0,0)x_1x_2 + F_{x_2x_2}(0,0)x_2^2$$

if $X = \mathbb{R}^2$ and $x = (x_1, x_2) \in \mathbb{R}^2$.

Exercise 16.9. Show there exists $r > 0$ such that

$$F(x) = \frac{1}{2}\langle F''(\theta(x))x, x\rangle$$

for some $\theta = \theta(x) \in [0,1]$ whenever $x \in X$ and $|x| < r$.

16.4 An equation for a change of coordinates

We ask if

$$x \to \langle \Phi_x x, x\rangle \quad \text{en} \quad y \to \langle \Phi_0 y, y\rangle$$

are the same functions, up to a change of coordinates, which we shall take of the special form

$$y = T_x x$$

with $T_x \in L(X,X)$. Again we use a subscript to denote the x-dependence, this time of T_x which acts act on X. Thus, given $x \to \Phi_x \in L(X, X^*)$, we look for $x \to T_x \in L(X,X)$ such that

$$\langle \Phi_x x, x\rangle = (\Phi_x x)\, x = (\Phi_0 y)\, y = \langle \Phi_0 y, y\rangle \tag{16.13}$$

for x close to $x = 0$.

Dropping the x-subscripts we need

$$\langle \Phi x, x\rangle = \langle \Phi_0 Tx, Tx\rangle = (\Phi_0 Tx)(Tx) = ((\Phi_0 Tx)\circ T)(x) = \langle (\Phi_0 Tx)\circ T, x\rangle,$$

which will certainly hold if

$$\Phi x = (\Phi_0 T\, x)\circ T$$

in X^* for all $x \in X$, or

$$\Phi h = (\Phi_0 T\, h)\circ T$$

for all $h \in X$ for that matter. Thus (16.13) holds if the map

$$h \to \Phi h \quad \text{is equal to the map} \quad h \to \Phi_0 Th \circ T = \kappa_0(T,T)\, h. \tag{16.14}$$

This is an $L(X, X^*)$-valued "quadratic" equation for $T \in L(X, X)$.

Abstractly we may write (16.14) as

$$\kappa_0(T, T) = \Phi, \tag{16.15}$$

in which

$$X \times X \xrightarrow{\kappa_0} L(X, X^*)$$

is the bilinear form defined by

$$h \to \kappa_0(T, U)\, h = \Phi_0 T h \circ U.$$

Clearly $T = I$ is a solution of (16.14) when $\Phi = \Phi_0$. We want a solution $T = T_x$ for $\Phi = \Phi_x$ given by (16.10) close to Φ_0. If you like you can skip Section 16.5 and jump to (16.26), or even Exercise 16.13. Just put $T = I + H$ in (16.14) and see what you can get[5].

16.5 A solution via the implicit function theorem?

The implicit function theorem is applicable if the derivative of

$$T \to \kappa_0(T, T)$$

is invertible in $T = I$. The continuity of $x \to \Phi_x$ in $x = 0$ is then the minimal assumption to obtain a solution T_x close to I for small x. Thus F'' continuous in 0 is a necessary condition to get started.

For the derivative with respect to T in I we write $T = I + H$, H small. Then (16.15) rewrites as

$$\underbrace{\Phi_0 H h + \Phi_0 h \circ H + \Phi_0 H h \circ H}_{\chi_0(H)h} = (\Phi_x - \Phi_0)h \tag{16.16}$$

for all $h \in X$. The left hand side defines an X^*-valued function

$$H \xrightarrow{\chi_0} \chi_0(H)$$

quadratic in H, with Φ_0 in the "coefficients" of the two linear terms and one quadratic term. Writing (16.16) as

$$\chi_0(H) = \Phi_x - \Phi_0, \tag{16.17}$$

the right hand side is in $S(X, X^*)$.

[5]But that's not how I found equation (16.27).

Look at (16.16). Clearly the derivative of χ_0 in $H = 0$ is given by

$$h \xrightarrow{\chi_0'(I)H} \Phi_0 Hh + \Phi_0 h \circ H.$$

Since $\chi_0'(0)H \in L(X, X^*)$ is characterised by

$$\langle \chi_0'(0)Hh, k\rangle = \langle \Phi_0 Hh, k\rangle + \langle \Phi_0 Hk, h\rangle, \tag{16.18}$$

we have that $\chi_0'(0)H \in S(X, X^*)$. Thus the invertibility condition cannot be that

$$\forall_{h\in X} : \chi_0'(I)H = \Phi_0 Hh + \Phi_0 h \circ H = Ch \tag{16.19}$$

is solvable for every $C \in L(X, X^*)$, while (16.19) is underdetermined for $C \in S(X, X^*)$.

A handy[6] extra condition on H is that $\Phi_0 H \in S(X, X^*)$. Then (16.18) reduces to

$$\langle \chi_0'(0)Hh, k\rangle = 2\langle \Phi_0 Hh, k\rangle, \tag{16.20}$$

and the invertibility condition (16.19) becomes

$$2\Phi_0 H = C, \tag{16.21}$$

which is solvable for H as

$$H = \frac{1}{2}\Phi_0^{-1} C \tag{16.22}$$

for every $C \in L(X, X^*)$.

Only $C \in S(X, X^*)$ can be relevant as we continue: we apply the implicit function theorem to

$$\{H \in L(X, X^*) : \Phi_0 H \in S(X, X^*)\} \xrightarrow{\chi_0} S(X, X^*)$$

around $H = 0$ and $x = 0$. With $K = \Phi_0 H$ as new independent variable this becomes[7]

$$2Kh + Kh \circ \Phi_0^{-1} K) = (\Phi_x - \Phi_0)h \tag{16.23}$$

for all $h \in X$, which amounts to the equation

$$2K + T_0(K) = C_x = \Phi_x - \Phi_0 \tag{16.24}$$

for $K \in S(X, X^*)$, in which the quadratic term is given by

$$T_0 : S(X, X^*) \to S(X, X^*), \quad T_0(K)h = Kh \circ (\Phi_0^{-1} K) \tag{16.25}$$

for all $h \in X$, and

$$X \ni x \to C_x \in S(X, X^*)$$

is continuous in $x = 0$ with $C_0 = 0$.

[6] As it turns out is how Duistermaat and Kolk put it.

[7] Equation (16.23) follows directly from (16.16).

16.6 Yes, but main result via power series instead

Theorem 16.10. *Let X be a complete metric vector space, $F : X \to \mathbb{R}$ twice continuously differentiable near $x = 0$. If $F'(0) = 0$ and $F''(0) \in L(X, X^*)$ is invertible with inverse in $L(X, X^*)$, then there is a transformation of the form*

$$y = T_x x = (I + \Phi_0^{-1} K_x)x,$$

in which

$$\Phi_0 = \frac{1}{2} F''(0)$$

and

$$x \to K_x \in S(X, X^*)$$

is continuous with $K_0 = 0$, such that

$$F(x) = \langle \Phi_0 T_x x, T_x x \rangle,$$

near $x = 0$.

Exercise 16.11. Prove Theorem 16.10 by applying the implicit function theorem to (16.23).

Remark 16.12. *If $F''(0)$ is positive definite in the sense that for some $\beta > 0$ it holds that*

$$\langle F''(0)(x), x \rangle \geq \beta |x|_X^2$$

for all $x \in X$, then X is really a Hilbert[8] space in disguise because

$$x \to \sqrt{\langle F''(0)(x), x \rangle}$$

then defines an equivalent[9] norm which comes from the symmetric bounded coercive bilinear form $(x, y) \to \langle F''(0)(x), y \rangle$. More on such forms in Section 17.3.

[8] See Chapter 17.

[9] Two norms are equivalent if there exists constants $M_1 > 0$ and $M_2 > 0$ such that

$$\frac{1}{M_1} |x|_{_2} \leq |x|_{_1} \leq M_2 |x|_{_2} \quad \text{for all } x.$$

In fact there's a direct way to solve (16.15) in the space

$$\{T \in L(X, X^*) : \Phi_0 T \in S(X, X^*)\}. \tag{16.26}$$

Via $T = I + H$ and (16.16) equation (16.15) was equivalent to (16.23) for

$$K = \Phi_0 H \in S(X, X^*).$$

We now return to an equation for H. Write (16.23) as

$$2Kh + Kh \circ (H) = (\Phi_x - \Phi_0)h$$

and apply it to $k \in X$. Then

$$\langle 2Kh, k\rangle + \underbrace{\langle Kh \circ (H), k\rangle}_{\langle Kh, Hk\rangle = \langle KHk, h\rangle} = \langle (\Phi_x - \Phi_0)h, k\rangle$$

for all $h, k \in X$. The first and the third term are symmetric in h and k. It follows that

$$2K + KH = \Phi_x - \Phi_0,$$

and applying Φ_0^{-1}, the equation to solve for H, still under the assumption that $\Phi_0 H \in S(X, X^*)$, is

$$2H + HH = \Phi_0^{-1}\Phi - I = P, \tag{16.27}$$

in which $P \in L(X, X^*)$ also has $\Phi_0 P \in S(X, X^*)$.

Exercise 16.13. Derive (16.27) directly from (16.15), the substitution $T = I + H$, and the assumption that $\Phi_0 H \in S(X, X^*)$.

In fact

$$P = \Phi_0^{-1}\Phi - I = \Phi_0^{-1}(\Phi - \Phi_0) = \Phi_0^{-1} \int_0^1 (1-t)(F''(tx) - F''(0))\, dt$$

$$= 2F''(0)^{-1} \int_0^1 (1-t)(F''(tx) - F''(0))\, dt = 2 \int_0^1 (1-t)(F''(0)^{-1}F''(tx) - I)\, dt,$$

and the equation for H to solve is

$$I + 2H + H^2 = I + P \quad \text{in} \quad L(X) = L(X, X). \tag{16.28}$$

It follows that $T = I + H$ is the square root of $I + P$, and we have some experience on solving that equation if P is not too large, see Exercise 11.12. The same power series tricks[10] give

$$T = I + H = I + \frac{1}{2}\,P - \frac{1}{2!}\,\frac{1}{2}\,\frac{1}{2}\,P^2 + \frac{1}{3!}\,\frac{1}{2}\,\frac{1}{2}\,\frac{3}{2}\,P^3 - \frac{1}{4!}\,\frac{1}{2}\,\frac{1}{2}\,\frac{3}{2}\,\frac{5}{2}\,P^4 + \cdots \tag{16.29}$$

if $|P| < 1$, and so $y = T_x x$ with

$$T_x = I + E_x - \frac{1}{2!}\,E_x^2 + \frac{1 \cdot 3}{3!}\,E_x^3 - \frac{1 \cdot 3 \cdot 5}{4!}\,E_x^4 + \cdots \tag{16.30}$$

and

$$E_x = \int_0^1 (1-t)(F''(0)^{-1}F''(tx) - I)\,dt, \tag{16.31}$$

which allows a more general setting[11]. In particular the assumption that $F''(0)$ is invertible may be relaxed. The basic assumption needed is that $|E_x| < \frac{1}{2}$, the norm being the norm in $L(X)$, i.e. the best Lipschitz constant.

Exercise 16.14. See if you can give a direct derivation of (16.30) and (16.31) as giving the transformation $y = T_x x$ that conjugates a real valued function $F(x)$ of $x \in \mathbb{R}$ having $F(0) = F'(0) = 0$ and $F''(0) \neq 0$ with the function $g(y) = \frac{1}{2}F''(0)y^2$. What do you need to assume on F?

[10] Copy/paste what you know by now for the case that $P, H \in \mathbb{R}$.

[11] Think of examples in which $F''(0)$ is not invertible in $L(X)$.

17 A short introduction to real Hilbert spaces

From another set of *do it yourself* notes in Dutch, translated only. A real Hilbert space H is a real vector space with an inner product, denoted

$$(x, y) \in H \times H \to (x, y)_H = x \cdot y,$$

in which Cauchy sequences are convergent. That is, if a sequence x_n in H has

$$(x_n - x_m) \cdot (x_n - x_m) \to 0$$

as $m, n \to \infty$, then there exists $\overline{x} \in H$ such that

$$(x_n - \overline{x}) \cdot (x_n - \overline{x}) \to 0$$

as $n \to \infty$.

Recall that the norm is given by

$$|x|_H^2 = (x, x)_H = x \cdot x,$$

and that the distance between x_n and x_m is

$$d_H(x_n, x_m) = |x_n - x_m|_H = \sqrt{(x_n - x_m) \cdot (x_n - x_m)}.$$

The map $d_H : H \times H \to \mathbb{R}^+ = [0, \infty)$ is the *metric* on H. Subscripts H will be dropped, unless they are needed to avoid confusion.

Exercise 17.1. Derive and prove the Cauchy-Schwarz[1] inequality

$$|x \cdot y| \leq |x|\,|y|,$$

and use it to prove the triangle inequality

$$|x + y| \leq |x| + |y|.$$

Formulate and prove the Pythagoras Theorem and the parallellogram law, i.e.

$$|x + y|^2 + |x - y|^2 = 2\,|x|^2 + 2\,|y|^2.$$

[1] See also Exercise 19.4.

17.1 Projections on closed convex sets

Exercise 17.2. Let H be a Hilbert space, $K \subset H$ a non-empty closed convex[2] subset, and $a \in H$. Show there exists a unique $p \in K$ that minimizes

$$|p - a| = \inf_{x \in K} |x - a| = d(a, K),$$

the distance from a to K, and show that $(p-a) \cdot (x-p) \geq 0$ for all $x \in K$. Hint: use the parallellogram law to show that a minimizing sequence is Cauchy. Also show that $P_K : H \to K$ defined by $P_K(a) = p$ has the property that $|P_K(a) - P_K(b)| \leq |a - b|$ for all $a, b \in H$.

Exercise 17.3. Let H be a Hilbert space, $L \subset H$ a closed linear subspace. Prove that $P_L : H \to L$ linear, and that

$$M = N(P_L) = \{x \in H : P_L(x) = 0\} = L^{\perp} = \{x \in H : x \cdot y = 0 \ \forall \, y \in L\},$$

the *null space* of P_L, is a closed linear subspace with $M \cap L = \{0\}$. Show that $M + L = H$ and conclude that $L \oplus M = H$: every $x \in H$ is uniquely written as $x = p + q$ with $p \in L$ and $q \in M$.

Exercise 17.4. Let H be Hilbert space, $K \subset H$ a non-empty closed convex subset. For all $b \in H$ the quadratic expression

$$|x|^2 + b \cdot x$$

has a unique minimizer on K.

We recall that a real valued function f defined on a normed vector space space X is called Lipschitz continuous if there exists a constant $L \geq 0$ such that

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2|$$

for all x_1 and x_2 in X. We shall call such functions Lipschitz functions.

Exercise 17.5. If such an L exists then there exists a smallest such L.

[2] If $a, b \in K$ then $[a, b] = \{ta + (1-t)b : t \in [0, 1]\} \subset K$.

17.2 Riesz representation of linear Lipschitz functions

Exercise 17.6. Let X be a normed vector space. The space of all Lipschitz (continuous) functions $f : X \to \mathbb{R}$ is denoted by $Lip(X)$. With

$$(f+g)(x) = f(x) + g(x) \quad \text{and} \quad (tf)(x) = tf(x)$$

it becomes a vector space. For every $f \in Lip(X)$ let $L = [f]_{Lip}$ be the smallest Lipschitz constant of f. Why is

$$f \to [f]_{Lip}$$

not a norm on $Lip(X)$? And why is it a norm on

$$Lip_0(X) = \{f \in Lip(X) : \ f(0) = 0\}?$$

Show that with this norm every Cauchy sequence $f_n \in Lip_0(X)$ is convergent. Hint: first for $X = \mathbb{R}$, then copy/paste for $X = X$.

The result in Exercise 46.8 is only of interest if there are such Lipschitz Lipschitz continuous functions on X. In case of $X = H$ a Hilbert space every $y \in H$ defines a linear ϕ_y in $Lip_0(H)$ by

$$\phi_y(x) = x \cdot y,$$

with smallest Lipschitz constant $|y|$. Thus $y \to \phi_y$ defines map

$$\Phi := H \to Lip_0(H).$$

and the range of Φ is contained in H^*, the (normed) space of all Lipschitz continuous *linear* functions $f : H \to \mathbb{R}$.

Exercise 17.7. Verify that $\Phi : H \to H^*$ satisfies

$$\Phi(x_1 + x_2) = \Phi(x_1) + \Phi(x_2) \quad \text{and} \quad \Phi(tx) = t\Phi(x)$$

for all $t \in \mathbb{R}$ and $x, x_1, x_2 \in H$, and that $[\Phi(x)]_{Lip} = |x|$. Thus Φ is linear.

Is Φ surjective, i.e. is every $f \in H^*$ of the form ϕ_y? Consider[3] its null space

$$N_f = \{x \in H : f(x) = 0\}.$$

[3]We write N_f instead of $N(f)$, to distinguish between f and P_L.

Exercise 17.8. Show that $N_f \subset H$ is a closed linear subspace.

Exercise 17.9. By 17.3 the projection

$$P_{N_f} : H \to N_f$$

is linear. Show that $M = N(P_{N_f}) = \{te : t \in \mathbb{R}\}$, in which $e \in N_f^{\perp}$ with $|e| = 1$. Then show that $f(x) = f(e)e \cdot x$.

Exercise 17.10. Explain why Exercise 17.8 says that $\Phi : H \to H^*$ a linear isometry.

The inverse of Φ is called the Riesz representation of H^*. We denote the inverse of Φ by R_H, and its domain is $H^* \subsetneq Lip_0(H)$.

Exercise 17.11. Use 17.2 to show that there are many nonlinear functions in $Lip_0(H)$.

Exercise 17.12. Show that

$$l^{(2)} = \{x = (x_1, x_2, x_3, \dots) : x_n \text{ a sequence in } \mathbb{R},\ \sum_{n=1}^{\infty} x_n^2 < \infty\} \quad \text{with} \quad x \cdot y = \sum_{n=1}^{\infty} x_n y_n$$

is a Hilbert space. We shall here write

$$x = \sum_{n=1}^{\infty} x_n e_n, \quad e_1 = (1, 0, 0, \dots), \quad e_2 = (0, 1, 0, \dots), \dots,$$

but we often prefer a notation with column vectors instead.

Every infinite dimensional separable[4] Hilbert space H can be identified with $l^{(2)}$. To see why take a sequence $a_1, a_2, a_2, \dots$ in H such that every element in H is a limit point of this sequence. Let

$$e_1 = \frac{1}{|a_1|} a_1$$

[4]See Section 5.5, this means that H contains a sequence a_n as in what follows.

of $a_1 \neq 0$, otherwise throw a_1 away and renumber the sequence until you have $a_1 \neq 0$. Then let

$$y_2 = a_2 - (a_2, e_1)e_1 \quad \text{and} \quad e_2 = \frac{1}{|y_2|} y_2$$

if $y_2 \neq 0$, but throw a_2 away if $y_2 = 0$ and renumber until you get $y_2 \neq 0$ and thereby e_2. Then put

$$y_3 = a_3 - (a_3, e_2)e_2 - (a_3, e_1)e_1 \quad \text{and} \quad e_3 = \frac{1}{|y_3|} y_3,$$

if $y_3 \neq 0$, but $\ldots$, and so on. This produces $e_1, e_2, e_3, \ldots$ with

$$(e_i, e_j) = \delta_{ij},$$

and

$$H = \{x = \sum_{n=1}^{\infty} x_n e_n : , x_n \text{ a sequence in } \mathbb{R}, \ \sum_{n=1}^{\infty} x_n^2 < \infty\}.$$

17.3 Bilinear forms and the Lax-Milgram theorem

This section is also from another set of notes. In Section 16.6 we mentioned that Theorem 16.10, the Morse lemma, is not restricted to the case that X is a Hilbert space in disguise[5]. In particular (16.31) does not require a Hilbert spaces setting. In this section we do require a Hilbert space setting, for a generalisation of the Riesz Representation Theorem[6].

Theorem 17.13. *Let H be a Hilbert space and $B : H \times H \to \mathbb{R}$ be a bounded coercive bilinear form, meaning that*

(a) for every $u \in H$ fixed $v \to B(u, v)$ is linear;
(b) for every $v \in H$ fixed $u \to B(u, v)$ is linear;
(c) $\exists_{\alpha \geq 0} \forall_{u,v \in H} : |B(u, v)| \leq \alpha |u| |v|$.
(d) $\exists_{\beta > 0} \forall_{u \in H} : B(u, u) \geq \beta |u|^2$.

Then every linear continuous $\phi : H \to \mathbb{R}$ is represented by a unique $u \in H$ via

$$\phi(v) = \langle \phi, v \rangle = B(u, v)$$

for all $v \in H$. This defines a continuous linear map

$$H^* \ni \phi \xrightarrow{S} u \in H$$

[5] A complete metric vector space which allows an equivalent inner product norm.
[6] This theorem is still somewhat hidden in Exercise 17.10.

with $|S| \le \frac{1}{\beta}$*, which is the inverse of the continuous linear map*

$$H \ni u \xrightarrow{A} \phi \in H^*$$

defined by

$$\langle \phi, v \rangle = \langle Au, v \rangle = B(u,v) \quad \text{for all} \quad v \in H, \tag{17.32}$$

which has $|A| \le \alpha$*.*

For the proof we observe that (17.32) and assumption (c) imply that

$$|\langle Au, v \rangle| = |B(u,v)| \le \alpha \, |u| \, |v|$$

for all u and v in H, and that for u fixed assumption (a) says that

$$Au : H \to \mathbb{R}$$

is linear. It follows that $Au \in H^*$ and

$$|Au| \le \alpha \, |u|.$$

Assumption (b) implies that the map

$$A : H \to H^*$$

is linear, and assumption (d) gives

$$\beta |u|^2 \le B(u,u) = \langle Au, u \rangle \le |Au| \, |u|$$

for all $u \in H$, whence

$$|Au| \ge \beta \, |u|.$$

We conclude that

$$H \xrightarrow{A} R(A) = \{Au : \, u \in H\}$$

is a linear bijection, continuous in both directions, because

$$\beta |u| \le |Au| \le \alpha |u| \tag{17.33}$$

for all $u \in H$. Thus $R(A)$ is complete because H is. In particular $R(A)$ is closed in H^*. It remains to show that $R(A) = H^*$.

Now let Φ be as in Riesz Representation Theorem and $L = \Phi^{-1}(R(A) \subset H$. If $L \ne H$ then

$$M = \{v \in H : \, v \cdot w = 0 \text{ for all } w \in L\} \ne \{0\}.$$

Choose $v \in M$ with $v \neq 0$. Then

$$\langle \Phi(w), v \rangle = w \cdot v = 0$$

for all $w \in R(A) = \{Au : u \in H\}$, whence $\langle Av, v \rangle = 0$, a contradiction with assumption (d). Thus $L = H$, whence $R(A) = H^*$. This completes the proof of Theorem 17.13.

If we start from the complete metric vector space perspective we find ourselves forced into the Hilbert space setting. Let's see why, while we formulate a result which is of independent interest.

Definition 17.14. *Let X be a normed space. A map $(u, v) \to B(u, v)$ from $X \times X$ to $\mathbb{R}$ is called a bounded bilinear form if*

(a) for every $u \in X$ fixed $v \xrightarrow{\phi} B(u, v)$ is linear;
(b) for every $v \in X$ fixed $u \xrightarrow{\psi} B(u, v)$ is linear;
(c) $\exists_{\alpha \geq 0} \forall_{u,v \in X} : |B(u, v)| \leq \alpha |u| |v|$.

If in addition

$$\exists_{\beta > 0} \forall_{u \in X} : B(u, u) \geq \beta |u|^2,$$

then B is called coercive.

Remark 17.15. *A bounded coercive bilinear form on a normed space X makes that X is an inner product space, with inner product defined by*

$$u \cdot v = \frac{1}{2}(B(u, v) + B(v, u)).$$

The corresponding inner product norm, defined by

$$|u|_B = \sqrt{B(u, u)},$$

is equivalent to the norm on X via

$$\beta |u|^2 \leq B(u, u) \leq \alpha |u|^2.$$

This makes any attempts to take the Lax-Milgram theorem out of the Hilbert space context futile. But it's good to know the statement of Theorem 17.16 below.

Theorem 17.16. *Every bounded bilinear form on a normed space X is of the form*

$$(u, v) \to B(u, v) = \langle Au, v \rangle \in \mathbb{R} \tag{17.34}$$

with $A \in L(X, X^*)$, and[7]

$$\sup_{u,v \in X \setminus \{0\}} \frac{|B(u,v)|}{|u|\,|v|} = |A|. \tag{17.35}$$

If X is complete and B is coercive then X is a Hilbert space in disguise, and A is a bijection[8] *between X and X^* with*

$$\beta\,|u| \le |Au| \le \alpha |u|$$

for all $u \in X$, $0 < \beta \le \alpha$, as in Definition 17.14.

For the proof we use (a) again to define A by $Au = \phi$, so (17.34) holds by definition. In particular Au is a linear functional on X for every $u \in X$. By (c) we have

$$|\langle Au, v\rangle| = |B(u,v)| \le \alpha\,|u|\,|v|$$

for all $v \in X$ whence $Au \in X^*$ with

$$|Au| \le \alpha\,|u|, \tag{17.36}$$

and (b) implies that $A : X \to X^*$ is linear. Thus $A \in L(X, X^*)$ with $|A| \le \alpha$.

Exercise 17.17. Prove (17.35) by showing that

$$\sup_{u,v \in X \setminus \{0\}} \frac{|\langle Au, v\rangle|}{|u|\,|v|} = |A|.$$

Hint: choose u with $|u| = 1$ and $|Au|$ close to $|A|$, and then v with $|v| = 1$ and $|\langle Au, v\rangle|$ close to $|Au|$.

Finally assume that X is complete and B is coercive. Then

$$\beta\,|u|^2 \le B(u,u) = \langle Au, u\rangle \le |\langle Au, u\rangle| \le |Au|,$$

whence (17.33) holds and

$$X \xrightarrow{A} R(A) = \{Au : u \in X\}$$

is a linear bijection, continuous in both directions. Thus $R(A)$ is a complete metric vector space because X is. In particular $R(A)$ is closed in X^*. Now write

$${}^0R(A) = \{v \in X : \forall_{\phi \in R(A)}\ \phi(v) = 0\} = \{v \in X : \forall_{u \in X}\ B(u,v) = 0\}.$$

[7] The norms have subscripts that we omit in this section.

[8] Lax-Milgram: $\forall_{\phi \in X^*}\, \exists_{u \in X}\, \forall_{v \in X} :\ B(u,v) = \phi(v) = \langle \phi, v\rangle$, u is unique for ϕ.

If we know that ${}^0R(A) \neq \{0\}$ then some $0 \neq v \in X$ has the property that

$$\langle Au, v\rangle = 0 \quad \text{for all} \quad u \in X,$$

impossible in view of $\langle Av, v\rangle \geq \beta|v|^2$. It follows that A is a linear bijection between X and X^* if X has the property[9] that closed subspaces $M \subset X^*$ with $M \neq X^*$ have ${}^0M \neq \{0\}$. Hilbert spaces (complete inner product spaces) have this property, and thus so does X. This completes the proof of Theorem 17.16.

[9]This property holds for reflexive spaces.

18 Analysis unpacked: more variables

In this chapter we are concerned with differential and integral calculus for functions from X to Y in which X and Y are Euclidean spaces. We begin with $X = Y = \mathbb{R}^2$, with (rectangular) coordinates $x, y \in \mathbb{R}$ for $X = \mathbb{R}^2$ and coordinates $u, v \in \mathbb{R}$ for $Y = \mathbb{R}^2$. Later we shall perhaps prefer $x_1, x_2 \in \mathbb{R}$ for $x = (x_1, x_2) \in X = \mathbb{R}^2$ and $y_1, y_2 \in \mathbb{R}$ for $y = (y_1, y_2) \in Y = \mathbb{R}^2$.

We frequently use polar coordinates r, θ and the transformation

$$x = r\cos\theta;$$

$$y = r\sin\theta,$$

to describe points $(x, y) \neq (0, 0)$ in the plane via their distance $r = \sqrt{x^2 + y^2}$ to the origin $(0, 0)$ and the angle θ between the halfline

$$\{(tx, ty) : t \geq 0\}$$

and the positive x-axis. Whenever convenient we identify $\mathbb{R}^2$ with the set $\mathbb{C}$ of *complex numbers*

$$z = x + iy,$$

and call $|z| = r$ the *absolute value* of z, the distance from z to the origin $z = 0$. The angle $\theta = \arg z$ is called the *argument* of z, uniquely determined modulo 2π for every $z \neq 0$.

Next to complex addition

$$w + z = (u + iv) + (x + iy) = u + x + i(v + y) = (u + x, v + y) = (u, v) + (x, y)$$

we also have complex mulitplication

$$wz = (u+iv)(x+iy) = ux-vy+i(uy+vx) = (ux-vy, uy+vx) = (u, v)(x, y),$$

based on the rule $i^2 = -1$, for $w = u + iv = (u, v)$ and $z = x + iy = (x, y) \in \mathbb{R}^2 = \mathbb{C}$. The rules for addition and multiplication in $\mathbb{C}$ are the same as the rules for addition and multiplication in $\mathbb{R}$. We also have

$$|w + z| \leq |w| + |z| \quad \text{and} \quad |wz| = |w|\,|z|.$$

Very important is the rule formulated in this exercise.

Exercise 18.1. The summation rules for cos and sin imply that

$$z_1 z_2 = r_1 r_2(\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)) \quad \text{for} \quad z_j = r_j(\cos\theta_j + i\sin\theta_j),\ j = 1, 2.$$

This rule is one many reasons to write

$$\cos\theta + i\sin\theta = \exp(i\theta) \quad \text{and} \quad \exp(z) = \exp(x)\exp(iy)$$

We note that polar coordinates are not needed to prove that for every nonzero γ the map

$$z \to \gamma z \tag{18.1}$$

is a rotation[1] around 0 followed by a point multiplication with 0 as fixed point, see (18.8) and Exercise 18.5.

18.1 Intermezzo: algebra's main theorem

The set $\mathbb{C}$ is algebraically closed: every polynomial

$$P(z) = \sum_{k=0}^{n-1} \alpha_k z^k + z^n \tag{18.2}$$

with $\alpha_0, \ldots, \alpha_{n-1} \in \mathbb{C}$ and $n \geq 2$ has a zero $z_1 \in \mathbb{C}$. Long division then gives that

$$P(z) = \sum_{k=0}^{n-1} \alpha_k z^k + z^n = (z - z_1)Q(z),$$

in which

$$Q(z) = \sum_{k=0}^{n-2} \beta_k z^k + z^{n-1},$$

with $\beta_0, \ldots, \beta_{n-2} \in \mathbb{C}$. In n steps it follows that

$$P(z) = (z - z_1)\cdots(z - z_n) \quad \text{with} \quad z_1, \ldots, z_n \in \mathbb{C}. \tag{18.3}$$

www-groups.dcs.st-and.ac.uk/history/HistTopics/Fund_theorem_of_algebra.html

Here's in modern language how Argand saw this. Consider the real valued function

$$(x, y) = x + iy = z \to |P(z)| = f(x, y).$$

If $P(z)$ does not have any zero's in $\mathbb{C}$, then f must have a global positive minimum and that's not possible.

[1] Unless $\gamma \in \mathbb{R}_+$.

Let's first show the latter statement. In terms of $P(z)$ this would mean that for some z_0 it holds that $|P(z)| \geq |P(z_0)| > 0$ for all $z \in \mathbb{C}$. Now use the algebra in $\mathbb{C}$ to write

$$w = z - z_0 \quad \text{and} \quad Q(w) = \frac{P(z)}{P(z_0)}.$$

Then

$$Q(w) = 1 + \sum_{k=1}^{n} \gamma_k w^k \tag{18.4}$$

and

$$w \to |Q(w)| = \frac{|P(z|)}{|P(z_0)|}$$

has a globale minimum $Q(0) = 1$. Thus $Q(w)$ cannot have values inside the unit disk. Now write $w = r\exp(i\theta)$ and $\gamma_k = c_k \exp(i\phi_k)$. Via Exercise 18.1 we have

$$Q(w) = 1 + \sum_{k=1}^{n} c_k r^k \exp(i(\phi_k + k\theta)), \tag{18.5}$$

an expression[2] in which the ϕ_k are parameters and $r > 0$ can be taken as small as we want. Exercise 18.2 below shows that all c_k are zero, meaning that $Q(w) = 1$ for all $w \in \mathbb{C}$ and hence $|P(z)| = |P(z_0)|$ for all $z \in \mathbb{C}$, contradicting (18.2).

Exercise 18.2. Assume some first c_k is nonzero. Show that $|Q(w)|$ has values smaller than 1. Hint: you may draw inspiration from the estimate in (18.6) below.

So why would f have a global minimum? Observe that f is continuous, so it has a minimum m_r and a maximum M_r on the closed disk

$$D_r = \{(x, y) : \, x^2 + y^2 \leq r^2\}.$$

Clearly m_r is nonincreasing in r. We wish to show that for r larger than some r_1 this minimum m_r does niet increase anymore, whence we can conclude that f has a global positive minimum on $\mathbb{R}^2$. This conclusion will follow from an easy large lower estimate for f on large circles.

Indeed, with $z = x + iy$ and $x^2 + y^2 = r^2$ we have for $|P(z)| = f(x, y)$ that

$$|P(z)| = |\sum_{k=0}^{n-1} \alpha_k z^k \, + z^n| \geq |z^n| - |\sum_{k=0}^{n-1} \alpha_k z^k| \geq r^n - \sum_{k=0}^{n-1} |\alpha_k| r^k. \tag{18.6}$$

[2]Ptolemaeus would have liked this.

On the circle defined by $x^2 + y^2 = r^2$ it then follows that

$$f(x,y) \geq r^{n-1}(r - \underbrace{\sum_{k=0}^{n-1} |\alpha_k|}_{r_0}) = \underline{M}_r,$$

a lower bound which is positive for r larger than

$$r_0 = \sum_{k=0}^{n-1} |\alpha_k|.$$

For $r = r_0$ we have $\underline{M}_{r_0} = 0 < m_{r_0}$. Clearly $\underline{M}_r$ increases to ∞ as r increases from r_0 to ∞. Thus for some $r_1 > r_0$ we have

$$\underline{M}_{r_1} > m_{r_0} \geq m_{r_1},$$

and then also

$$f(x,y) > m_{r_1} \quad \text{for all} \quad (x,y) \notin D_{r_1}.$$

It follows that m_{r_1} is the global minimum of f on the whole of $\mathbb{R}^2$ and the contradiction arises as explained above. This completes this truly remarkable proof in which elegant algebra, basically algebraic estimates, and rock solid analysis combine.

18.2 Complex and multivariate differential calculus

In Section 9.2 we saw, for every choice of coefficients $\alpha_n \in \mathbb{R}$ indexed by $n \in \mathbb{N}_0$, that

$$x \to \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots = \sum_{n=1}^{\infty} \alpha_n x^n$$

defines a function on

$$B_R = \{x \in \mathbb{R} : \, |x| < R\}$$

for some maximal $R \in [0, \infty]$, and that differential calculus for this function is just as differential calculus for polynomials.

The point to make now is that Theorem 9.3 and its proof carry over by copy-paste to complex valued power series with complex coefficients and variables. Also, differentiability via (10.1) becomes complex differentiability for functions[3]

$$H : \mathbb{C} \to \mathbb{C},$$

[3]For convenience we assume H is globally defined.

but now via

$$w = H(z) = H(z_0) + \gamma(z - z_0) + T(z; z_0) \tag{18.7}$$
$$= H(z_0) + H'(z_0)(z - z_0) + o(|z - z_0|)$$

as $z \to z_0$.

If we unpack (18.7), writing

$$z = x + iy,\ w = u + iv,\ H(z) = F(x, y) + iG(x, y),\ u = F(x, y),\ v = G(x, y),$$

we can view H, via the identification $\mathbb{C} = \mathbb{R}^2$, as a function

$$H : \mathbb{R}^2 \to \mathbb{R}^2$$

with components $H_1 = F$ and $H_2 = G$. With $h = x - x_0$ en $k = y - y_0$ the linear term (18.7) unpacks as

$$\gamma(z - z_0) = (\alpha + i\beta)(h + ik) = \alpha h - \beta k + i(\beta h + \alpha k).$$

This corresponds to

$$\begin{pmatrix} \alpha h - \beta k \\ \beta h + \alpha k \end{pmatrix} = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} h \\ k \end{pmatrix}, \tag{18.8}$$

in which the matrix describes the map (18.1).

The complex expansion (18.7) rewrites as

$$u = F(x, y) = F(x_0, y_0) + a(x - x_0) + b(y - y_0) + R(x, y; x_0, y_0);$$

$$v = G(x, y) = G(x_0, y_0) + c(x - x_0) + d(y - y_0) + S(x, y; x_0, y_0),$$

with remainder terms R and S defined via $T = R + iS$, and a special form of the 2×2 matrix A in the linear expansion around (x_0, y_0), namely

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = A = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}.$$

Changing to notation with indices,

$$\underbrace{\begin{pmatrix} H_1(x_1, x_2) \\ H_2(x_1, x_2) \end{pmatrix}}_{H(x)} = \underbrace{\begin{pmatrix} H_1(a_1, a_2) \\ H_2(a_1, a_2) \end{pmatrix}}_{H(a)} + \underbrace{\begin{pmatrix} A_{11}(x_1 - a_1) + A_{12}(x_2 - a_2) \\ A_{21}(x_1 - a_1) + A_{22}(x_2 - a_2) \end{pmatrix}}_{H'(a)(x-a)=A(x-a)} + R,$$

$$R = \begin{pmatrix} R_1(x_1, x_2; a_1, a_2) \\ R_2(x_1, x_2; a_1, a_2) \end{pmatrix},$$

we thus have the following theorem.

Theorem 18.3. *Let $H : \mathbb{R}^2 \to \mathbb{R}^2$ be differentiable in $a = (a_1, a_2)$ with $H'(a)$ given by the matrix A. Then $H_1 + iH_2 : \mathbb{C} \to \mathbb{C}$ is complex differentiable in $a_1 + ia_2$ if and only if*

$$A_{11} = A_{22} \quad \text{and} \quad A_{12} = -A_{21}.$$

Exercise 18.4. Prove Theorem 18.3.

Exercise 18.5. Examine (18.1) using (18.8).

So far for H. Returning to $F : \mathbb{R}^2 \to \mathbb{R}^2$ (possibly complex) differentiable in $a = (a_1, a_2)$, $F'(a)$ given by the matrix A, we write $h_1 = x_1 - a_1$, $h_2 = x_2 - a_2$ and

$$Ah = \begin{pmatrix} (Ah)_1 \\ (Ah)_2 \end{pmatrix} = \begin{pmatrix} A_{11}h_1 + A_{12}h_2 \\ A_{21}h_1 + A_{22}h_2 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \tag{18.9}$$

which we think of as $F'(a)$ acting on h.

A more algebraic point of view is to be fine with Ah as a product of A and h. Compare the notation[4] to on the one hand the notation with A_0 acting on h and the norm of A_0 in $L(X, Y)$, and on the other hand with A_0 algebraically multiplying h. In the latter context we can estimate

$$|(Ah)_1| = |A_{11}h_1 + A_{12}h_2| \leq \sqrt{A_{11}^2 + A_{12}^2}\,\sqrt{h_1^2 + h_2^2};$$

$$|(Ah)_2| = |A_{21}h_1 + A_{22}h_2| \leq \sqrt{A_{21}^2 + A_{22}^2}\,\sqrt{h_1^2 + h_2^2},$$

to conclude that

$$((Ah)_1)^2 + ((Ah)_2)^2 \leq (A_{11}^2 + A_{12}^2 + A_{21}^2 + A_{22}^2)(h_1^2 + h_2^2),$$

meaning for the product of A and h that[5]

$$|Ah|_2 \leq |A|_2 \, |h|_2. \tag{18.10}$$

In (18.10) the "Euclidean" lengths of $h = x - a$, Ah and A appear[6], in each case the square root of the sum of the squared entries. You may well prefer here to forget[7] all about the norm of

$$h \xrightarrow{A} Ah$$

[4] We dropped the zero-subcripts.

[5] This generalises, see (19.6).

[6] Actually this 2-norm of A is called the Frobenius norm of A.

[7] If not note that (18.10) says that this operator norm of A is at most equal to $|A|_2$.

in $L(\mathbb{R}^2, \mathbb{R}^2)$: going back to

$$F(x) = F(a) + A(x-a) + R(x;a) \quad \text{and} \quad |R(x;a)|_2 = o(|x-a|_2) \quad (18.11)$$

as $|x-a|_2 \to 0$, except for the subscript 2, the condition for differentiability is undistinguishable from differentiability of $F : \mathbb{R} \to \mathbb{R}$ and generalises to $F : \mathbb{R}^m \to \mathbb{R}^n$.

Looking at the "partial" functions

$$x_1 \to F_1(x_1, x_2),\ x_2 \to F_1(x_1, x_2),\ x_1 \to F_2(x_1, x_2),\ x_2 \to F_2(x_1, x_2)$$

we find

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(a_1, a_2) & \frac{\partial F_1}{\partial x_2}(a_1, a_2) \\ \frac{\partial F_2}{\partial x_1}(a_1, a_2) & \frac{\partial F_2}{\partial x_2}(a_1, a_2) \end{pmatrix} = F'(a) = DF(a) \quad (18.12)$$

in every point $x = (x_1, x_2) = (a_1, a_2) = a$ where F is differentiable.

We often identify the linear map[8] $F'(a) = DF(a)$ with its Jacobi matrix

$$\frac{\partial F}{\partial x} = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} \end{pmatrix}$$

evaluated in $x = a$, but the existence of this matrix is not sufficient for differentiability. We examined this issue in Section 15.5 for $F : \mathbb{R}^2 \to \mathbb{R}$ and $F : X \times Y \to \mathbb{R}$.

Exercise 18.6. State and prove a theorem for $F : \mathbb{R}^2 \to \mathbb{R}$ by specializing Theorem 15.6 to $X = Y = \mathbb{R}$ and generalise to $F : \mathbb{R}^m \to \mathbb{R}$ and $F : \mathbb{R}^m \to \mathbb{R}^n$.

18.3 Cauchy-Riemann equations, harmonic functions

Have another look at Theorem 18.3 and let H be complex differentiable[9] in $z_0 = x_0 + iy_0$. We use the correspondence

$$z = x + iy \in \mathbb{C} \leftrightarrow (x, y) \in \mathbb{R}^2 \quad \text{and} \quad w = u + iv \in \mathbb{C} \leftrightarrow (u, v) \in \mathbb{R}^2$$

and write

$$H'(z_0) = \alpha + i\beta.$$

[8] Both notations are widely used.

[9] We now prefer a notation with (x, y) and (x_0, y_0).

Exercise 18.7. Show that α and β are then given by

$$\alpha = \frac{\partial u}{\partial x} \quad \text{and} \quad \beta = -\frac{\partial u}{\partial y}, \tag{18.13}$$

evaluated in $(x, y) = (x_0, y_0)$.

Thus Theorem 18.3 says that u and v, as functions of x and y, must satisfy the so-called Cauchy-Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y} \tag{18.14}$$

in $(x, y) = (x_0, y_0)$.

If these partial derivatives exist and are by themselves differentiable, say for all $(x, y) \in \mathbb{R}^2$ in an open ball containing (x_0, y_0), then we would have

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial v}{\partial y} = \frac{\partial}{\partial y}\frac{\partial v}{\partial x} = -\frac{\partial}{\partial y}\frac{\partial u}{\partial y} = -\frac{\partial^2 u}{\partial y^2},$$

but only if the order of differentiation does not matter, and likewise for $v(x, y)$. If so, we conclude that in (x_0, y_0) it holds that

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = \Delta v, \tag{18.15}$$

in which the differential operator Δ, the *Laplacian*, occurs. This Δ is a feast to study, but not now. Here we want to be sure under what conditions (18.15) makes sense. We copy Theorem 16.3 from Section 16.1.

Theorem 18.8. *Let $v : \mathbb{R}^2 \to \mathbb{R}$ have the property that*

$$(x, y) \to \frac{\partial v}{\partial x} = v_x(x, y) \quad \text{and} \quad (x, y) \to \frac{\partial v}{\partial y} = v_y(x, y)$$

are differentiable in (x_0, y_0). Then the second order partial derivatives in (x_0, y_0) exist, and

$$v_{yx}(x_0, y_0) = v_{xy}(x_0, y_0).$$

Twice differentiable functions $u(x, y)$ and $v(x, y)$ that satisfy (18.15) on an open set $\mathcal{O} \subset \mathbb{R}^2$ are called harmonic. As an example, the functions

$$(x, y) \to \mathrm{Re}(x + iy)^n \quad \text{en} \quad (x, y) \to \mathrm{Im}(x + iy)^n$$

are harmonic on the whole of $\mathbb{R}^2$. These are the so-called homogeneous harmonic polynomials of degree $n \in \mathbb{N}$.

Referring to Section 16.3, twice differentiable means that the map

$$(x, y) \to (\frac{\partial u}{\partial x} \ \frac{\partial u}{\partial y})$$

is itself differentiable. With the chain rule it follows that

$$(x, y) \to (\frac{\partial u}{\partial x} \ \frac{\partial u}{\partial y}) \to \frac{\partial u}{\partial x} \quad \text{and} \quad (x, y) \to (\frac{\partial u}{\partial x} \ \frac{\partial u}{\partial y}) \to \frac{\partial u}{\partial y}$$

are differentiable. Thus $\Delta u = 0$ has a meaning as

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \tag{18.16}$$

without any v interfering[10].

There are many non-constant solutions of (18.16). Indeed, you should have noticed

$$x, y, x^2 - y^2, 2xy, x^3 - 3xy^2, 3x^2y - y^3, x^4 - 6x^2y^2 + y^4, 4x^3y - 4xy^3, \ldots \tag{18.17}$$

above.

Exercise 18.9. Unpack $w = \exp(z) = \exp(x + iy)$ starting from the power series for $\exp(z)$ and verify that $\exp(z) = \exp(x)\exp(iy)$ with $\exp(iy) = \cos x + i \sin x$. Explain why this leads to the concept of multivalued[11] functions

$$w \to \log w = \ln |w| + i \arg w.$$

.

18.4 Monomials and power series again

This should speak for itself. With

$$H = \frac{|x - a|}{r}$$

we have that

$$x^m = a^m + ma^{m-1}(x - a) + R_{a,m}(x), \quad |R_{a,m}(x)| \leq \frac{m(m-1)r^m}{2} H^2.$$

[10]Not a priori.

[11]Which are thereby not functions.

Likewise for $|y|, |b| \le s$, we have

$$y^n = b^n + nb^{n-1}(y-b) + R_{b,n}(y), \quad |R_{b,n}(y)| \le \frac{n(n-1)s^m}{2}K^2, \quad K = \frac{|y-b|}{s}.$$

Multiplication then gives[12]

$$x^m y^n = a^m b^n + \underbrace{ma^{m-1}b^n(x-a) + na^m b^{n-1}(y-b)}_{linear\ part} + \underbrace{R_2 + R_{21} + R_{12} + R_{2,2}}_{R_{a,b,m,n}(x,y)},$$

in which we identify

$$R_2 = b^n R_{a,m}(x) + mna^{m-1}b^{n-1}(x-a)(y-b) + a^m R_{b,n}(y),$$

$$R_{21} = ma^{m-1}(x-a)R_{b,n}(y),$$

$$R_{12} = nb^{n-1}R_{a,m}(x)(y-b),$$

$$R_{2,2} = R_{a,m}(x)R_{b,n}(y).$$

With rough but obvious estimates

$$|R_{2,2}| \le \frac{1}{4}m^2n^2r^m s^n H^2 K^2,$$

$$|R_{21}| \le \frac{1}{2}m^2 n r^m s^n H^2 K \le \frac{1}{2}m^2n^2r^m s^n H^2 K,$$

$$|R_{12}| \le \frac{1}{2}mn^2 r^m s^n H K^2 \le \frac{1}{2}m^2n^2r^m s^n H K^2,$$

and also, a little less obvious maybe,

$$|R_2| \le \frac{1}{4}(m^2+n^2)r^m s^n (H^2+K^2),$$

we conclude that

$$x^m y^n = a^m b^n + ma^{m-1}b^n(x-a) + na^m b^{n-1}(y-b) + \underbrace{R_{a,b,m,n}(x,y)}_{R}, \quad (18.18)$$

in which

$$|R| \le \frac{r^m s^n}{4}\left(m^2n^2HK(HK+2H+2K) + (m^2+n^2)(H^2+K^2)\right). \quad (18.19)$$

The perhaps less obvious estimate for R_2 follows via

$$|R_2 \le |s^n R_{a,m}(x)| + |mnr^{m-1}s^{n-1}(x-a)(y-b)| + |r^m R_{b,n}(y)| \le$$

[12]This is a bit like (11.5).

$$= \frac{m(m-1)r^m s^n}{2} H^2 + mnr^m s^n HK + \frac{n(n-1)r^m s^n}{2} K^2 =$$
$$\frac{r^m s^n}{4} \begin{pmatrix} m(m-1) & mn \\ mn & n(n-1) \end{pmatrix} \begin{pmatrix} H \\ K \end{pmatrix} \cdot \begin{pmatrix} H \\ K \end{pmatrix},$$

and the 2-norm of the matrix in this expression being less than $m^2 + n^2$.

We now multiply (18.18) by coefficients α_{mn} and the estimates for $R_{2,21,12,22}$ in

$$R = R_{a,b,m,n}(x,y) = R_2 + R_{21} + R_{12} + R_{2,2}$$

by coefficients $|\alpha_{mn}|$, and take the sum over $m, n \in \mathbb{N}_0$. Clearly a sufficient condition to conclude that on the rectangle

$$R_{rs} = \{(x,y) \in \mathbb{R}^2 : \, |x| < r, |y| < s\}$$

the power series

$$P(x,y) = \sum_{m,n \in \mathbb{N}_0} \alpha_{mn} x^m y^n$$

exists as a differentiable function, with

$$P_x(x,y) = \sum_{m,n \in \mathbb{N}_0} m\alpha_{mn} x^{m-1} y^n \quad \text{and} \quad P_y(x,y) = \sum_{m,n \in \mathbb{N}_0} n\alpha_{mn} x^m y^{n-1},$$

is that the series

$$\sum_{m,n \in \mathbb{N}_0} (m^2 + n^2) \, |\alpha_{mn}| r^m s^n \quad \text{and} \quad \sum_{m,n \in \mathbb{N}_0} m^2 n^2 \, |\alpha_{mn}| r^m s^n \tag{18.20}$$

converge. We then have

$$P(x,y) = P_x(a,b)(x-a) + P_y(a,b)(y-b) + R(x,y;a,b),$$

with $R(x,y;a,b)$ the sum of four remainder terms, each of which having the HK part factoring out, and the resulting coefficient bounded by (18.20).

Exercise 18.10. Fill in the details of the above proof. Show in addition that the convergence of

$$\sum_{m,n \in \mathbb{N}_0} (m^2 + n^2) \, |\alpha_{mn}| R^{m+n} \quad \text{and} \quad \sum_{m,n \in \mathbb{N}_0} m^2 n^2 \, |\alpha_{mn}| R^{m+n} \tag{18.21}$$

suffices to have $P(x,y)$ exist as a differentiable function on the disk

$$\{(x,y) \in \mathbb{R}^2 : \, x^2 + y^2 < R\}.$$

18.5 Application: the Hopf bifurcation

We examine the system of differential equations

$$\frac{dx}{dt} = \mu x - y + p_2(x,y) + p_3(x,y) + \cdots = P(x,y);$$

$$\frac{dy}{dt} = x + \mu y + q_2(x,y) + q_3(x,y) + \cdots = Q(x,y),$$

for real valued function $x(t)$ and $y(t)$, in which the functions

$$p_n(x,y) = a_{n0}x^n + a_{n1}x^{n-1}y + \cdots + a_{0n}y^n$$

and

$$q_n(x,y) = b_{n0}x^n + b_{n1}x^{n-1}y + \cdots + b_{0n}y^n$$

have real coefficients for every

$$n \in \mathbb{N}_2 = \{n \in \mathbb{N} : \, n \geq 2\},$$

and $\mu \in \mathbb{R}$ is a parameter. We shall call this family of systems the μ-systems.

In the special case that all the coefficients are zero the μ-systems reduce to

$$\frac{dx}{dt} = \mu x - y;$$

$$\frac{dy}{dt} = x + \mu y.$$

The reduced μ-system has nontrivial periodic solutions[13] if and only if $\mu = 0$. The plane defined by $\mu = 0$ and the line defined by $x = y = 0$ in μxy-space together form the set of all bounded solution orbits of the reduced μ-systems. We wish show that near $x = y = 0$ this family of periodic orbits persists as we add the nonlinear terms. Under the basic assumption that the coefficients are bounded we will show that there exists a locally defined smooth function $f(x,y)$ with $f_x(0,0) = f_y(0,0) = 0$ such that the graph $\mu = f(x,y)$ describes all the periodic solutions of the full system. In particular every level set

$$\Gamma_\mu = \{(x,y) \in \mathbb{R}^2, \, f(x,y) = \mu\}$$

is a periodic orbit of the full μ-system.

[13]Namely $x = \varepsilon \cos t, y = \varepsilon \sin t$, in which $\varepsilon > 0$ is not necessarily small.

Exercise 18.11. Assume that the coefficients a_{mn} and b_{mn} are bounded. Use Section 18.4 to conclude that

$$P(x,y) = \sum_{m,n\in\mathbb{N}_0} a_{mn}x^m y^n \quad \text{and} \quad Q(x,y) = \sum_{m,n\in\mathbb{N}_0} b_{mn}x^m y^n$$

are well-defined and smooth for x and y with $|x| < 1$ and $|y| < 1$.

Without loss of generality we now assume that

$$|a_{mn}| \leq 1 \quad \text{and} \quad |b_{mn}| \leq 1 \quad \text{for all} \quad m,n \in \mathbb{N} \quad \text{with} \quad m+n \geq 2, \tag{18.22}$$

and introduce polar coordinates $x = r\cos\theta, y = r\sin\theta$ to transform solutions of the μ-systems to solutions of

$$\frac{dr}{dt} = \mu r + \alpha_2(\theta)r^2 + \alpha_3(\theta)r^3 + \cdots ;$$

$$\frac{d\theta}{dt} = 1 + \beta_2(\theta)r + \beta_3(\theta)r^2 + \cdots .$$

Exercise 18.12. Use the chain rule[14] and Section 11.3 to determine the expressions for α_n and β_n expressed in terms of $c = \cos\theta, s = \sin\theta, p_n(c,s), q_n(c,s)$. Show that

$$|\alpha_n| \leq n \quad \text{and} \quad |\beta_n| \leq n \quad \text{for all} \quad n \in \mathbb{N}_2,$$

and denoting the r-dependent part of the right hand side of the θ-equation by

$$-\rho = \beta_2(\theta)r + \beta_3(\theta)r^2 + \cdots$$

that

$$|\rho| \leq 2r + 3r^2 + 4r^3 + \cdots = \frac{r(2-r)}{(1-r)^2} < 1,$$

if $0 < r < 2 - \sqrt{2}$.

Exercise 18.13. Use the chain rule and Section 11.3 again to show that, for

$$0 < r < 2 - \sqrt{2},$$

solutions can be seen as functions $r = r(\theta)$ of θ, and that

$$\frac{dr}{d\theta} = r_\theta = \mu r + A_3(\theta,\mu)r^2 + A_4(\theta,\mu)r^3 + A_5(\theta,\mu)r^4 + \cdots , \tag{18.23}$$

[14]Figure out how to use only the version with $X = Y = Z = \mathbb{R}$ from Section 11.2.

with $A_3, A_4, \ldots$ polynomials in $\cos\theta$ en $\sin\theta$ in which also μ appears. Hint:

$$\frac{1}{1-\rho} = 1 + \rho + \rho^2 + \rho^3 + \cdots = \sum_{n=0}^{\infty} \rho^n.$$

Exercise 18.14. Show directly from the differential equations for $r(t)$ and $\theta(t)$ that

$$|\frac{dr}{d\theta}| = |\frac{\frac{dr}{d\theta}}{\frac{d\theta}{d\theta}}| \leq \frac{r}{1-2r}(|\mu|(1-r^2) + r(2-r))$$

for $0 < r < \frac{1}{2}$.

Exercise 18.15. Show that

$$\int_0^{2\pi} A_3(\theta, \mu) d\theta = 0.$$

Exercise 18.16. Consider the truncated differential equation

$$r_\theta = \mu r + A_3(\theta, \mu) r^2$$

and do the Kepler trick: introduce $w = \frac{1}{r} > 0$ as a function of θ. Why can this equation have no 2π-periodic solutions? Hint: you should get an equation in which only $\frac{dw}{d\theta}$, w and A_3 appear. Integrate from 0 to 2π to derive a contradiction if $w(\theta)$ is a (positive) 2π-periodic solution.

Consider (18.23) with $r(0) = \varepsilon > 0$ as initial value. For the original μ-system this corresponds to the solution with $x(0) = \varepsilon, y(0) = 0$. Now scale r by setting $r = \varepsilon R$. Then (18.23) becomes

$$\frac{dR}{d\theta} = R_\theta = \mu R + \varepsilon A_3(\theta, \mu) R^2 + \varepsilon^2 A_4(\theta, \mu) R^3 + \varepsilon^3 A_5(\theta, \mu) R^4 + \cdots, \quad (18.24)$$

and we look for solutions with $R(0) = 1$. Note that the explicit estimate in Exercise 18.14 carries over. We have

$$|\frac{dR}{d\theta}| \leq \frac{R}{1-2\varepsilon R}(|\mu|(1-\varepsilon^2 R^2) + \varepsilon R(2-\varepsilon R))$$

for $0 < \varepsilon R < \frac{1}{2}$.

If this initial value problem has a solution $R(\theta; \mu, \varepsilon)$ for small μ and small ε, then we set

$$F(\mu, \varepsilon) = R(2\pi; \mu, \varepsilon) - 1$$

and examine the equation

$$F(\mu, \varepsilon) = 0.$$

Clearly we have $F(0, 0) = 0$. Can we apply Theorems 15.1 and 15.2? The answer is yes, via what we already started in Section 15.4.

19 Measures of parallelotopes

In this chapter we prove the spectral theorem[1] for compact linear symmetric operators. In fact this theorem is just a minor varation of a theorem for symmetric matrices S that we need for what follows next, starting from two 2-vectors[2]

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

spanning a parallelogram in the plane.

If you draw such a parallelogram you can easily deform it into a rectangle, while keeping its area fixed, and then it's clear what its area is. Have a look at

`https://en.wikipedia.org/wiki/Parallelogram`

to see how, and read to see how this can be turned into algebra.

We observe that there are two ways to put the two 2-vectors **a** and **b** into what we call a matrix. We choose for

$$A = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix},$$

with *transpose*

$$A^T = \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix}.$$

Likewise two 3-vectors **a** and **b** fit in A^T as

$$A^T = \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix}.$$

How does such a matrix provide us with the area spanned by **a** and **b**? The answer involves the matrix product

$$S = A^T A, \tag{19.1}$$

a symmetric matrix to which Section 19.4 applies.

[1] Essentially Theorem 19.6.

[2] We momentarily surrender to the boldface vector notation in physics.....

19.1 Matrix products

In general an $m \times n$ real matrix A is a block[3] with real entries a_{ij}. The vertical index i runs from 1 to m, the horizontal index j from 1 to n. Considered as a map[4] A sends an n-vector $x \in \mathbb{R}^n$ with coordinates $x_1, \ldots, x_n$ to an m-vector $y \in \mathbb{R}^m$ with coordinates

$$y_i = \sum_{j=1}^{n} a_{ij} x_j.$$

We say that

$$A \in L(\mathbb{R}^n, \mathbb{R}^m),$$

the space of linear maps from $\mathbb{R}^n$ to $\mathbb{R}^m$, and we write $y = Ax$.

If B is a real $n \times p$ matrix with entries b_{jk}, the vertical index j running from 1 up n, the horizontal index k from 1 up p, then AB is by definition the $m \times p$ matrix with entries

$$\sum_{j=1}^{n} a_{ij} b_{jk}, \tag{19.2}$$

with the corresponding linear map[5]

$$A \circ B : \mathbb{R}^p \xrightarrow{B} \mathbb{R}^n \xrightarrow{A} \mathbb{R}^m.$$

If we transpose both blocks A and B by numbering the first index horizontally, and the second index vertically, then we get *transposed* matrices A^T and B^T with entries $a^t_{ji} = a_{ij}$ and $b^t_{kj} = b_{jk}$, and (19.2) reads as

$$\sum_{j=1}^{n} b^t_{kj} a^t_{ji},$$

the entries of $B^T A^T$ in $(AB)^T = B^T A^T$.

In the special case that $m = n = p$ it can happen that $AB = I_n$, the $n \times n$ matrix with all diagonal entries equal to 1, and all off-diagonal entries equal to 0. This matrix corresponds to the linear map $I = I_n$ that sends every $x \in \mathbb{R}^n$ to itself . *What you really need to know from linear algebra*[6] is that the map $A \circ B$ being the same map as the map I_n is equivalent to

[3] With m and n in $\mathbb{N}$.

[4] A linear map in fact.

[5] So A is preceded by B.

[6] A proof should be given in one of the first hours of any course in Linear Algebra.

$AB = I$ for the corresponding matrices. We say that A and B are each others inverses, as linear maps because $A \circ B = B \circ A = I_n$, with $AB = I = BA$ for the matrices. And likewise for the transposes. We emphasise that these are statements about *square matrices*, and solutions of $Ax = y$ with A a square matrix.

If a third $p \times r$ matrix C has entries c_{kl} then $(AB)C$ is the matrix with entries

$$\sum_{k=1}^{p}(\sum_{j=1}^{n} a_{ij}b_{jk})c_{kl} = \sum_{k=1}^{p}\sum_{j=1}^{n} a_{ij}b_{jk}c_{kl}, \tag{19.3}$$

and these are also the entries of $A(BC)$: just change the order of the summations. Thus $(AB)C = A(BC)$ and we write ABC for the product of A, B and C. The corresponding linear map is $A \circ B \circ C$. Transposing we have $(ABC)^T = C^T B^T A^T$, which is what we will use in Section 20.2 for (20.16).

19.2 Matrix norms

The series

$$I + A + A^2 + A^3 + \cdots, \tag{19.4}$$

with A a square matrix[7], is important for the implicit function theorem with $F : \mathbb{R}^{n+m} \to \mathbb{R}^m$ in Section 27.1. You should also compare (19.4) to (16.29), and ask the question as to what is required to justify the manipulations that led to it. Estimates that do so can be best understood starting from a 2×2 matrix as in (18.9) and estimates for $A : \mathbb{R}^2 \to \mathbb{R}^2$ of the form (18.10).

Indeed you easily check[8] that for every $n \times n$ matrix and every real n-vector h it is true that

$$|Ah|_2 \leq M|h|_2, \tag{19.5}$$

if $M \geq 0$ is defined by

$$M^2 = \sum_{i,j=1}^{n} a_{ij}^2.$$

If you like this defines a kind of Pythagoras length of A, notation

$$M = |A|_2.$$

This norm has the property that

$$|A + B|_2 \leq |A|_2 + |B|_2 \quad \text{and} \quad |AB|_2 \leq |A|_2 \, |B|_2. \tag{19.6}$$

[7] A 2×2 matrix as in (18.9) for instance.

[8] Using proof by induction if you like.

holds[9].

If you like all of the above is just algebra with matrices. Recall though that the smallest M for which (19.5) holds is called the operator norm of A, notation $|A|_{op}$. It is for this latter definition that we we want see A as a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n$. Then the norm of A is the largest possible ratio between the norm of Ah and the norm of h.

We note that $L(\mathbb{R}^n) = L(\mathbb{R}^n, \mathbb{R}^n)$ is not only a vector space over $\mathbb{R}$, but also a normed algebra, because also the product operation

$$(A, B) \to AB$$

behaves as it should with respect to the norm

$$A \to |A|_{op},$$

namely, it holds that

$$|AB|_{op} \leq |A|_{op} \, |B|_{op}.$$

This is in addition to

$$|A|_{op} = 0 \iff A = 0, \quad |\lambda A|_{op} = |\lambda| \, |A|_{op} \geq 0, \quad |A + B|_{op} \leq |A|_{op} + |B|_{op}$$

for all $A, B \in L(\mathbb{R}^n)$ and $\lambda \in \mathbb{R}$.

As a vector space $L(\mathbb{R}^n)$ is just[10] $\mathbb{R}^{n^2}$, with the standard Pythagoraen norm[11] defined by

$$|A|_2^2 = \sum_{i,j=1}^{n} a_{ij}^2,$$

the (Frobenius) norm for which we have both inequalities in (19.6). Since $|A|_{op} \leq |A|_2$ for all $A \in L(\mathbb{R}^n)$ we prefer to use the smaller of the two norms[12].

Exercise 19.1. Prove there exists $\mu_n \in (0, 1]$ such that

$$\mu_n |A|_2 \leq |A|_{op} \leq |A|_2$$

for all $A \in L(\mathbb{R}^n)$. Hint[13]: if not then on

$$\{A \in L(\mathbb{R}^n) : |A|_{op} = 1\}$$

[9] Verify this. How does this generalise to non-square matrices?

[10] Entries in a block or in a column, what's the difference really?

[11] In the literature it is called the Frobenius norm.

[12] Which makes for a sharper statement than in Exercise 1.14.

[13] Hardy would dislike this proof, as you can do this with an explicit construction of μ_n.

the Pythagoras norm $|A|_2$ can be arbitrarily large, and therefore also the length of at least one of the column vectors. This is at odds with $|A|_{op} = 1$.

Exercise 19.2. If $A \in L(\mathbb{R}^n)$ has $|A|_{op} < 1$ then it holds for the series in (19.4) that

$$(I - A)(I + A + A^2 + A^3 + \cdots) = I.$$

Explain why and prove that

$$(I + A)^{-1} = I - A + A^2 - A^3 + \cdots = \sum_{j=0}^{\infty} (-A)^j.$$

Remark 19.3. *It should by now be clear that the whole machinery of power series carries over to Banach algebra's.*

19.3 Quadratic forms and operator norms

In (19.2) we can put $B = A^T$, the transpose of the matrix A with entries a_{ij} used in

$$y_i = \sum_{j=1}^{n} a_{ij} x_j,$$

which defined $A \in L(\mathbb{R}^n, \mathbb{R}^m)$. This gives[14]

$$S = AA^T \in L(\mathbb{R}^m, \mathbb{R}^m) \quad \text{with entries} \quad s_{ik} = \sum_{j=1}^{n} a_{ij} a_{kj} = s_{ki}. \tag{19.7}$$

Since

$$|A|_{op} = \max_{0 \neq x \in \mathbb{R}^n} \frac{|Ax|_2}{|x|_2} = \max_{|x|_2 = 1} |Ax|_2,$$

and likewise for $|A^T|_{op}$, we have

$$|A^T|_{op}^2 = \max_{|z|_2 = 1} \underbrace{|A^T z|_2^2}_{A^T z \cdot A^T z} = \max_{|z|_2 = 1} AA^T z \cdot z = \max_{|z|_2 = 1} Sz \cdot z = \max_{0 \neq z \in \mathbb{R}^m} \frac{Sz \cdot z}{z \cdot z}, \tag{19.8}$$

and we note that the bilinear mapping

$$(z, w) \to Sz \cdot w$$

[14] Don't let (19.1) confuse you.

from $\mathbb{R}^m \times \mathbb{R}^m$ to $\mathbb{R}$ then satisfies all the axioms of an inner product, except that $Sz \cdot z = 0$ does not imply that $z = 0$.

Exercise 19.4. Rederive the Cauchy-Schwarz inequality for $z, w \in \mathbb{R}^m$ by inspection of the minimum of the nonnegative function

$$\lambda \to |\lambda w - z|_2^2 = (\lambda w - z) \cdot (\lambda w - z),$$

and show that the same reasoning leads to

$$|Sz \cdot w| \leq \sqrt{Sz \cdot z}\,\sqrt{Sw \cdot w}.$$

Note the special case $m = n$ and $S = A = I$ and don't forget to discuss the possibility that the function you use is not a quadratic but a linear function.

For $S = AA^T$ as above we set

$$M = \max_{|z|_2 = 1} Sz \cdot z,$$

whereby we note that S is a symmetric matrix for which $Sz \cdot z \geq 0$ holds for all $z \in \mathbb{R}^m$. Just like it is easy to prove from the definition of the 2-norm via

$$|w|_2 = \sqrt{w \cdot w}$$

that

$$|z + w|_2^2 + |z - w|_2^2 = 2\,|z|_2^2 + 2\,|w|_2^2,$$

you easily verify that

$$S(z + w) \cdot (z + w) + S(z - w) \cdot (z - w) = 2\,Sz \cdot z + 2\,Sw \cdot w, \qquad (19.9)$$

an identity to play with, with $S = AA^T$as above, but also with $S = I$ the identity:

Exercise 19.5. The Cauchy-Schwarz inequality and the definition of the operator norm immediately imply that $M \leq |S|_{op}$. Write

$$4\,Sz \cdot w = S(z + w) \cdot (z + w) - S(z - w) \cdot (z - w)$$

and estimate the right hand side in terms of M to obtain that in particular for all $z, w \in \mathbb{R}^m$ with $|z|_2 = |w|_2 = 1$ it holds that $|Sz \cdot w| \leq M$. Conclude that $|S|_{op} = M$.

The map

$$z \to Q(z) = Sz \cdot z$$

defined by the symmetric matrix S is called a quadratic form. Observe that in Exercise 19.5 the assumption that $Sz \cdot z \geq 0$ can be dropped if M is defined by

$$M = \sup_{|z|_2 = 1} |Sz \cdot z|.$$

You should never forget the remarkable fact that the maxima of $z \to |Q(z)|$ and $z \to |Sz|$ on the unit ball coincide.

19.4 Eigenvalues of compact symmetric operators

The above carries over to $S : H \to H$ when H is any inner product space and $S : H \to H$ is linear and symmetric with respect to that inner product, and has the property that $Sz \cdot z \geq 0$ for all $z \in H$, except that we no longer know that the maxima exist. Introducing

$$|S|_{op} = \sup_{0 \neq z \in H} \frac{|Sz|}{|z|} = \sup_{0 \neq z \in H} \sqrt{\frac{Sz \cdot Sz}{z \cdot z}} = \sup_{z \cdot z = 1} \sqrt{Sz \cdot Sz}, \tag{19.10}$$

and

$$M = \sup_{z \cdot z = 1} Sz \cdot z, \tag{19.11}$$

it suffices to have that S is bounded on the unit ball in H to have

$$M = |S|_{op} < \infty. \tag{19.12}$$

Ignoring the trivial case that $M = 0$ we now observe that the Cauchy-Schwarz inequality in Exercise 19.4 also holds with S replaced by $M - S = MI - S$, I being the identity map, and it thus holds that

$$|(M - S)z \cdot w| \leq \sqrt{(M - S)z \cdot z}\, \sqrt{(M - S)w \cdot w}, \tag{19.13}$$

whence (varying w over the unit ball)

$$|(M - S)z| \leq \sqrt{(M - S)z \cdot z}\, \sqrt{|M - S|_{op}} \leq \sqrt{(M - S)z \cdot z}\, \sqrt{M + |S|_{op}}$$

Taking a sequence $z_n \in H$ with $|z_n| = 1$ and $Sz_n \cdot z_n \to M$, it then follows that the right hand side goes to zero, and thus

$$Mz_n - Sz_n \to 0.$$

If the sequence z_n can be chosen to have Sz_n converging to a limit $y \in H$, it follows that also $Mz_n \to y$ and that $M = |y| > 0$. But then $w = \frac{y}{M}$ is a unit eigenvector of S with eigenvalue M. We have therefore proved the following Theorem.

Theorem 19.6. *Let H be an inner product space and $S : H \to H$ linear, symmetric with $Sz \cdot z \geq 0$ for all $z \in H$, $Sz \neq 0$ for at least one $z \in H$. If for every bounded sequence z_n in H it holds that Sz_n has a convergent subsequence, then*

$$\lambda_1 = \max_{0 \neq z \in H} \frac{Sz \cdot z}{z \cdot z} > 0$$

exists, and λ_1 is an eigenvalue of S whose eigenvectors are the maximizers[15] *of the quotient under consideration.*

Remark 19.7. *In fact we only need one single sequence z_n with $z_n \cdot z_n = 1$ such that Sz_n converges and*

$$Sz_n \cdot z_n \to \sup_{0 \neq z \in H} \frac{Sz \cdot z}{z \cdot z}$$

to conclude that λ_1 exists, and is an eigenvalue of S whose eigenvectors are the maximizers. In particular this is the case when the supremum is a maximum.

Given an eigenvector w_1 with $|w_1| = 1$ it easily follows that S maps

$$H_1 = \{z \in H : \, z \cdot w_1 = 0\}$$

to itself. Unless H_1 is[16] the null space of S it then follows that

$$\lambda_2 = \max_{z \cdot w_1 = 0 \neq z \in H} \frac{Sz \cdot z}{z \cdot z} > 0$$

is also an eigenvalue of S with eigenvector w_2 with $|w_2| = 1$.

Repeating the argument with

$$H_2 = \{z \in H : \, z \cdot w_1 = z \cdot w_2 = 0\}$$

we obtain a sequence of eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots > 0,$$

which either terminates[17], or has the property that $\lambda_n \to 0$ as $n \to \infty$. The latter statement is a consequence of the convergent subsequences assumption: the corresponding mutually perpendicular unit eigenvectors

$$w_1, w_2, \ldots,$$

[15] Typically only multiples of one eigenvector.

[16] This includes the possibility that $H_1 = \{0\}$.

[17] If the range of H is spanned by $v_1, \ldots, v_N$ for some $N \in \mathbb{N}$.

terminating or not, have

$$|Sv_n - Sv_m|_2^2 = \lambda_n^2 + \lambda_m^2,$$

which prohibits Cauchy subsequences of Sv_n if the sequence $\lambda_n > 0$ does not terminate and decreases to a positive limit.

If we do *not* assume that $Sz \cdot z \geq 0$ for all $z \in H$ then the absolute value of the first eigenvalue is still obtained as

$$|\lambda_1| = \max_{0 \neq z \in H} \frac{|Sz \cdot z|}{z \cdot z} > 0,$$

because, changing from S to $-S$ if necessary, it is no restriction to assume that

$$M = \sup_{0 \neq z \in H} \frac{|Sz \cdot z|}{z \cdot z} = \sup_{0 \neq z \in H} \frac{Sz \cdot z}{z \cdot z},$$

and reason as above. With the Cauchy-Schwarz inequality in (19.13) still holding[18] while the version in Exercise 19.4 fails, the upshot is that we still obtain eigenvalues with

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq 0,$$

with eigenvectors as before. This is essentially the spectral theorem for compact symmetric linear operators S from an inner product space H to itself. It does not require any knowledge of the determinants which will become important next in the finite-dimensional case.

19.5 Singular values and measures of parallelotopes

In the case that $H = \mathbb{R}^m$ the subsequence argument is not needed as the maximizer w for the maximum in Theorem 19.6 exists in view of the compactness of the unit ball in $\mathbb{R}^m$. Now consider the matrix A defined by

$$A^T = \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} \tag{19.14}$$

and[19]

$$S = A^T A = \begin{pmatrix} a_1^2 + a_2^2 + a_3^2 & a_1 b_1 + a_2 b_2 + a_3 b_3 \\ b_1 a_1 + b_2 a_2 + b_3 a_3 & b_1^2 + b_2^2 + b_3^2 \end{pmatrix} = \begin{pmatrix} a \cdot a & a \cdot b \\ b \cdot a & b \cdot b \end{pmatrix} \tag{19.15}$$

[18] I first saw this Cauchy-Schwarz trick in the appendix of the PDE book of Craig Evans.

[19] Compared to (19.7) we switch from A to A^T, back to (19.1) for what comes next.

The outer product $a \times b$ of these two 3-column vectors a and b with, respectively, entries a_1, a_2, a_3 and entries b_1, b_2, b_3, is defined as the 3-vector with entries

$$a_2b_3 - a_3b_2, \quad a_3b_1 - a_1b_3, \quad a_1b_2 - a_2b_1,$$

and has squared length

$$|a \times b|^2 = (a_2b_3 - a_3b_2)^2 + (a_3b_1 - a_1b_3)^2 + (a_1b_2 - a_2b_1)^2 = \det(\mathrm{A^T A}),$$

as you should verify. That is to say, $\det(\mathrm{A^T A})$ is the sum of all the squares of all the 2×2-determinants of 2×2 submatrices of A. Here we count these 2×2 submatrices modulo the column permutations in (19.14).

As you may know, the length of the outer product $a \times b$ of a and b equals the area of the parallellogram spanned by a and b. Thus this area is the square root of the sum of the squares of the three 2×2-determinants in (19.14). It is precisely this statement that generalises to the n-dimensional measure of a parallelotope spanned by n vectors $x_1, \ldots, x_n$ in $\mathbb{R}^{\mathrm{N}}$.

Theorem 19.8. *Let $1 \leq n \leq N$. Consider the parallelotope P spanned by the vectors $x_1, \ldots, x_n$ in $\mathbb{R}^{\mathrm{N}}$. After putting these vectors in the columns of a matrix A, the n-dimensional measure $\mathcal{M}_n(x_1, \ldots, x_n)$ of P is the square root of the determinant of A^TA, and this determinant in turn is the sum of all the squares of the determinants of all $n \times n$ submatrices, and also equals the product $\sigma_1 \cdots \sigma_n$ of the singular values of A.*

Let us sketch a proof of this statement, first for (19.14), without using the outer product, using the invariance of the area under shear transformations. That is to say, the area of the parallelogram spanned by the vectors a and b is the same as that of the parallelogram spanned by the vectors $a + tb$ and b with $t \in \mathbb{R}$ arbitrary. The same statement holds for the determinant of $S = A^TA$ and the determinant of $S_t = A_t^TA_t$ where A_t is the matrix with column vectors $a + tb$ and b. Indeed, writing $A_t = A + tB$ we have

$$A_t^TA_t = (A + tB)^T(A + tB) = A^TA + tA^TB + tB^TA + t^2B^TB$$

$$= \underbrace{A^TA + tA^TB}_{C_t} + t(\underbrace{B^TA + tB^TB}_{D_t}) = S_t$$

The matrix C_t is the matrix obtained from $S = A^TA$ by adding t times the second (last) row of S to its first row. Therefore C_t and S have the same determinant. In turn, the matrix S_t is obtained from C_t by adding t times the second (last) column of C_t to its first column. Therefore S_t and C_t have the same determinant. It follows that S_t and S have the same

determinant. So both the area and the determinant are invariant under this shear transformation, which allows us to restrict our proof to the case in which $a \cdot b = 0$. Then the square of the area is equal to the product of the squares of the lengths of a and b, which is also the determinant of the diagonal matrix with entries $a \cdot a$ and $b \cdot b$. To prove the general statement in the theorem we use repeated shear transformations which leave both the determinant and the measure invariant and reduce the statement to be proved to the case that $x_i \cdot x_j = 0$ if $i \neq j$ and a corresponding diagonal matrix S with entries $x_1 \cdot x_1, \ldots, x_n \cdot x_n$. But this should be obvious from any formal definition of the n-dimensional measure of parallelotopes spanned by n vectors, a definition we happily leave here to be for what it is.

It remains to show that the determinant of the matrix S defined in (19.7) is also equal to the sum of the squares of the determinants of all the maximal square submatrices of A. These are also invariant under the shear transformations used above. Rather than using these transformations to reduce the statement to be proved to the case that the column vectors satisfy $x_i \cdot x_j = 0$ for $i \neq j$ we now use them diagonalise a maximal square part of the matrix A. Note that if the matrix A has no $n \times n$ submatrix with nonzero determinant, then the sum of the squared $n \times n$ determinants is zero, while also it cannot be the case that the column vectors are independent. Then our reduction to the case that the column vectors satisfy $x_i \cdot x_j = 0$ leads to one of these vectors being zero making the n-dimensional measure of P, and thereby the determinant of $A^T A$ zero as well.

Thus we may as well assume that the upper $n \times n$ part of A has nonzero determinant. It is a straightforward linear algebra exercise to show that, most likely after relabeling the first n coordinates, shear transformations bring A in the form

$$A = \begin{pmatrix} \Lambda \\ B \end{pmatrix}$$

where Λ is an $n \times n$ diagonal matrix with nonzero entries $\lambda_1, \ldots, \lambda_n$. Here we already assumed that $n < N$ because otherwise there was nothing to prove[20] in the first place. It now follows that

$$A^T A = \Lambda^2 + B^T B = \Lambda^2 + S,$$

where B is an $m \times n$ matrix with entries b_{ik} and S has entries

$$s_{ij} = \sum_{k=1}^{m} b_{ik} b_{jk}.$$

[20] If you know your determinants.

We therefore have, writing $B = [B_1, \ldots, B_n]$ with $B_1, \ldots, B_n$ the column vectors of B and using product notation, that

$$\det(\mathrm{A^T A}) = \underbrace{\Pi_j \lambda_j^2}_{\lambda_\mathrm{j}^1 \cdots \lambda_\mathrm{j}^\mathrm{n}} + \mathrm{s}_{11} \underbrace{\Pi_{j \neq 1} \lambda_j^2}_{\lambda_\mathrm{j}^2 \cdots \lambda_\mathrm{j}^\mathrm{n}} + \cdots + \mathrm{s}_{\mathrm{nn}} \Pi_{\mathrm{j} \neq \mathrm{n}} \lambda_\mathrm{j}^2$$

$$+ \det \begin{pmatrix} \mathrm{s}_{11} & \mathrm{s}_{12} \\ \mathrm{s}_{12} & \mathrm{s}_{22} \end{pmatrix} \Pi_{\mathrm{j} \neq 1,2} \lambda_\mathrm{j}^2 + \cdots + \det \mathrm{S} =$$

$$\Pi_j \lambda_j^2 + (B_1 \cdot B_1) \Pi_{j \neq 1} \lambda_j^2 + \cdots + \det \begin{pmatrix} \mathrm{B}_1 \cdot \mathrm{B}_1 & \mathrm{B}_1 \cdot \mathrm{B}_2 \\ \mathrm{B}_1 \cdot \mathrm{B}_2 & \mathrm{B}_2 \cdot \mathrm{B2} \end{pmatrix} \Pi_{\mathrm{j} \neq 1,2} \lambda_\mathrm{j}^2 + \cdots ,$$

in which we wrote the term of degree n and only the first terms of degree $2n-2$ and degree $2n-4$ in $\lambda_1, \ldots, \lambda_n$. It should be obvious what the remaining terms are.

On the other hand, the sum of the squared determinants of the $n \times n$ submatrices of A is

$$\Pi_j \lambda_j^2 + (b_{11}^2 + b_{21}^2 + \cdots + b_{m1}^2) \Pi_{j \neq 1} \lambda_j^2 + \cdots + \left(\det \begin{pmatrix} \mathrm{b}_{11} & \mathrm{b}_{12} \\ \mathrm{b}_{12} & \mathrm{b}_{22} \end{pmatrix}^2 + \cdots \right) \Pi_{j \neq 1,2} \lambda_j^2 + \cdots$$

It remains to show that

$$B_1 \cdot B_1 = b_{11}^2 + b_{21}^2 + \cdots + b_{m1}^2,$$

which is clearly the case, and then that

$$\det \begin{pmatrix} \mathrm{B}_1 \cdot \mathrm{B}_1 & \mathrm{B}_1 \cdot \mathrm{B}_2 \\ \mathrm{B}_1 \cdot \mathrm{B}_2 & \mathrm{B}_2 \cdot \mathrm{B2} \end{pmatrix} = \det \begin{pmatrix} \mathrm{b}_{11} & \mathrm{b}_{12} \\ \mathrm{b}_{12} & \mathrm{b}_{22} \end{pmatrix}^2 + \cdots + \det \begin{pmatrix} \mathrm{b}_{(\mathrm{m}-1)1} & \mathrm{b}_{(\mathrm{m}-1)2} \\ \mathrm{b}_{\mathrm{m}2} & \mathrm{b}_{\mathrm{m}2} \end{pmatrix}^2,$$

etcetera. These are the statements we set out to prove for A, before applying shear transformations, but with shorter columnvectors, namely of length $N - n$, respectively for two such vectors, up m such vectors. We can thus systematically reduce the statement we want to prove to lower dimensions of the matrix under consideration, until we reach the easy case that $m = 1$.

20 Stationary under constraints

This topic was started in Section 15.6 with the remarkable formula

$$\Phi_x = \Phi_y (F_y)^{-1} F_x \tag{20.16}$$

in $(x, y) = (0, 0)$ as the condition for

$$x \xrightarrow{\phi} \phi(x) = \Phi(x, f(x))$$

being stationary in $x = 0$, using the implicit function

$$y = f(x)$$

obtained in Section 15.2 to describe the solution set of $F(x, y) = 0$ near $(x, y) = (0, 0)$.

Continuity of the partials

$$(x, y) \to F_x(x, y) \quad \text{and} \quad (x, y) \to F_y(x, y)$$

in a neighbourhood of $(0, 0)$, and the invertibility of F_y in $(0, 0)$ sufficed for a proof that near $(x, y) = (0, 0)$ the level set

$$S = \{(x, y) : \, F(x, y) = F(0, 0)\} \tag{20.17}$$

is described as the graph of an implicitly definined continuously differentiable function f.

With this f the level set S is locally parameterised by

$$x \to X(x) = (x, f(x)),$$

which has a 2×1 Jacobi matrix $\frac{\partial X}{\partial x}$. The parameterisation is locally a bijection bewteen S and a neighbourhood of $x = 0$, which is due to the invertability of the 1×1 matrix

$$A = F_y \tag{20.18}$$

in $(0, 0)$. Differentiability of

$$(x, y) \to \Phi(x, y)$$

sufficed to have (20.16) as both necessary and sufficient for $\phi'(x) = 0$, not only in $x = 0$ but as long as $F_y(x, f(x))$ is invertertible on a whole neighbourhood of $x = 0$ in which $f(x)$ was constructed.

20.1 The method of Lagrange

This abstract section is part of a story line that started for the simplest concrete case in Section 15.6 and continues in Section 20.2 with the multivariate version of the Lagrange Multiplier Theorem. In the abstract setting with $x \in X$, $y \in Y$, $F : X \times Y \to Y$ and $\Phi : X \times Y \to \mathbb{R}$ consider

$$(x, y) \xrightarrow{F_x} F_x(x, y) \quad \text{and} \quad (x, y) \xrightarrow{F_y} F_y(x, y)$$

continuous near $(x, y) = (0, 0)$ with F_y invertible, and the continuously differentiable implicit function $y = f(x)$ as a local description of the set S defined by $F(x, y) = 0$. Now copy/paste (15.31) and read

$$\phi'(0) = 0 \iff \Phi_x(0, 0) = \Phi_y(0, 0)F_y(0, 0)^{-1}F_x(0, 0)$$

in the abstract setting. This formula will be unpacked in Section 20.1, for now we write it as (20.16), i.e.

$$\Phi_x = \Phi_y(F_y)^{-1}F_x.$$

If we can write $\Phi_y \in Y^*$ as

$$\Phi_y = \Lambda \circ F_y,$$

then

$$\Phi_x = \Phi_y(F_y)^{-1}F_x = \Lambda \circ F_y(F_y)^{-1}F_x = \Lambda \circ F_x,$$

and the criterion for stationarity becomes

$$\Phi' = \Lambda \circ F'. \tag{20.19}$$

What we need here is that every $A : Y \to Y$ and $\psi \in Y^*$ define a (unique) $\Lambda \in Y^*$ with $\psi = \Lambda \circ A$. This relates to what we discussed in Section 17.3. More details to follow perhaps.

20.2 The Lagrange multiplier method

With for instance

$$x \in \mathbb{R}^2,\ y \in \mathbb{R}^3,$$
$$F : \mathbb{R}^5 \to \mathbb{R}^3,\ \Phi : \mathbb{R}^5 \to \mathbb{R},$$
$$f : \mathbb{R}^2 \to \mathbb{R}^3,\ \phi : \mathbb{R}^2 \to \mathbb{R},$$

the theorems and proofs in Chapter 15 are essentially unchanged, beginning with (20.16) as the characterisation for

$$x \xrightarrow{\phi} \Phi(x, f(x))$$

being stationary, see (15.29) in Section 15.6.

Let's see how all this unpacks to give the method of Lagrange mulitpliers when we read (20.16) as a statement for Jacobi matrices and the corresponding linear maps. We write (20.16) in transposed form as

$$\nabla_x F \left(\nabla_y F\right)^{-1} \nabla \Phi_y = \nabla_x \Phi, \tag{20.20}$$

in which

$$\nabla_x F, \nabla_y F, \nabla_x \Phi, \nabla_y \Phi$$

are the transposes of the "partial" Jacobi matrices

$$\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}, \frac{\partial \Phi}{\partial x}, \frac{\partial \Phi}{\partial y}$$

corresponding to F_x, F_y, Φ_x, Φ_y.

Unpacking[1] the notation we have

$$\nabla_x F = \left(\nabla_x F_1 \ \nabla_x F_2 \ \nabla_x F_3\right) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_2}{\partial x_1} & \frac{\partial F_3}{\partial x_1} \\ \frac{\partial F_1}{\partial x_2} & \frac{\partial F_2}{\partial x_2} & \frac{\partial F_3}{\partial x_2} \end{pmatrix}$$

and likewise for $\nabla_y F$, which is a square 3×3 matrix, by assumption invertible in $(0, 0, 0, 0, 0)$. Its inverse sends the gradient vectors

$$\nabla_y F_1, \nabla_y F_2, \nabla_y F_3$$

back[2] to the column[3] base vectors e_1, e_2, e_3 in $\mathbb{R}^3$.

Now write $\nabla_y \Phi \in \mathbb{R}^3$ as linear combination[4]

$$\nabla_y \Phi = \lambda_1 \nabla_y F_1 + \lambda_2 \nabla_y F_2 + \lambda_3 \nabla_y F_3 \tag{20.21}$$

with $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$. It follows that $\nabla_x F \left(\nabla_y F\right)^{-1}$ in the left hand side of (20.20) acts on (20.21) as

$$\nabla_y \Phi \xrightarrow{(\nabla_y F)^{-1}} \lambda_1 e_1 + \lambda_2 e_2 + \lambda_3 e_3 \xrightarrow{\nabla_x F} \lambda_1 \nabla_x F_1 + \lambda_2 \nabla_x F_2 + \lambda_3 \nabla_x F_3 = \nabla_x \Phi$$

[1] It is really no more than that, check it!

[2] Since the column vectors of a matrix A are the images under A of the e's.

[3] As opposed to the convention in Exercise 17.12.

[4] This is possible in view of the invertibility condition imposed on F_y in $(0, 0, 0, 0, 0)$.

by (20.20) again. With (20.21) this combines as

$$\nabla\Phi = \lambda_1\nabla F_1 + \lambda_2\nabla F_2 + \lambda_3\nabla F_3, \tag{20.22}$$

simply[5] because it holds for ∇_x and ∇_y separately! The stationarity of

$$\Phi : S \to \mathbb{R}$$

in $(0,0)$ is thus equivalent with the existence of multiplicators $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$ for which (20.22) holds in $(0,0,0,0,0)$.

20.3 Application: Hölder's inequality

In (18.10) we had

$$|Ah|_2 \leq |A|_2 \, |h|_2$$

as a special case of

$$|AB|_2 \leq |A|_2 \, |B|_2 .$$

With $A = a$ a row matrix with entries a_i and $B = b$ a column matrix with entries b_i, this is the Cauchy-Schwarz inequality

$$|\sum_{i=1}^{n} a_i b_i| \leq \left(\sum_{i=1}^{n} |a_i|^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} |b_i|^2\right)^{\frac{1}{2}} .$$

This inequality is proved in every linear algebra course and then used to prove the triangle inequality for the Euclidean norm.

We now ask for which values of $p > 1$ and $q > 1$ we can also have that

$$|\sum_{i=1}^{n} a_i b_i| \leq |a|_p \, |b|_q , \tag{20.23}$$

if $|a|_p$ and $|b|_q$ are defined by

$$|a|_p^p = \sum_{i=1}^{n} |a_i|^p \quad \text{and} \quad |b|_q^q = \sum_{i=1}^{n} |b_i|^q . \tag{20.24}$$

Note that (20.23) is the Cauchy-Schwarz inequality of $p = q = 2$.

Exercise 20.1. Since (20.23) scales with a and b we can restrict the attention to vectors a and b for which $|a|_p = |b|_q = 1$. Explain!

[5] No 3×3 matrix inverted here.

Thus we introduce two boundary conditions

$$\phi(a_1,\ldots,a_n) = |a_1|^p + \cdots + |a_n|^p = 1;$$

$$\psi(b_1,\ldots,b_n) = |b_1|^q + \cdots + |b_n|^q = 1,$$

and max- and minimise

$$(a_1,\ldots,a_n,b_1,\ldots,b_n) \xrightarrow{F} a_1b_1 + \cdots + a_nb_n.$$

Exercise 20.2. Explain why the maximum and the minimum of F under the restriction $|a|_p = |b|_q = 1$ exist.

Exercise 20.3. Show that the functions ϕ and ψ are continuously differentiable if $p > 1$ and $q > 1$. Hint: if we redefine $x \to x^r$ to be odd for every $r > 0$ then the derivative of $x \to |x|^p$ is $x \to px^{p-1}$.

With two Lagrange multipliers λ en μ we arrive at $2n$ equations

$$b_i = \lambda p a_i^{p-1}; \quad a_i = \mu q b_i^{q-1} \quad (i = 1,\ldots,n)$$

to solve, together with

$$\sum_{i=1}^{n} |a_i|^p = \sum_{i=1}^{n} |b_i|^q = 1.$$

Exercise 20.4. Assume that $(p-1)(q-1) \neq 1$. Show that solutions have all $|a_i|$ equal and all $|b_i|$ equal, and therefore

$$\sum_{i=1}^{n} |a_ib_i| = n(\frac{1}{n})^{\frac{1}{p}+\frac{1}{q}} = n^{1-\frac{1}{p}-\frac{1}{q}}. \tag{20.25}$$

Deduce that (20.23) holds for $p > 1$ and $q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$.

21 Green's Theorem

We want to integrate continuous functions and partial derivatives of continuously differentiable functions over bounded sufficiently nice domains Ω. The goal is an early version of Green's Theorem (and thereby the Gauss Divergence Theorem), Theorem 21.6 in Section 21.3. To this end we need the integral calculus for continuous functions of two or more variables, beginning with integrals of $u = u(x, y)$. Integrating partial derivatives we discover the appropriate notion of boundary integrals.

We first integrate over closed rectangles $[a, b] \times [c, d]$, and over sets such as

$$\{(x, y) \in [a, b] \times [c, d] : y \leq f(x)\}, \quad \text{in which} \quad f \in C^1([a, b]) \tag{21.1}$$

and $f([a, b]) \subset (c, d)$. The integral calculus over closed blocks

$$[a, b] = [a_1, b_1] \times \cdots \times [a_N, b_N]$$

in $\mathbb{R}^N = \mathbb{R}^{n+1}$ is then completely similar, as well as integral calculus over sets described by

$$a_N \leq x_N \leq f(x_1, \ldots, x_{N-1}) < b_N \tag{21.2}$$

or

$$a_N \leq f(x_1, \ldots, x_{N-1}) \leq x_N < b_N, \tag{21.3}$$

and similar sets obtained by permutation of the variables.

To also integrate over closures of bounded open sets Ω in $\mathbb{R}^N = \mathbb{R}^{n+1}$ with $\partial\Omega \in C^1$, we understand $\partial\Omega \in C^1$ to mean that $M = \partial\Omega$ is the union of *patches*

$$P = M \cap W = M \cap (a, b),$$

each of which, after renumbering the variables, comes with a description of $[a, b] \cap \bar{\Omega}$ as given by (21.2) or (21.3). If so we say that Ω is a *bounded C^1-smooth domain*. We speak of

$$W = (a, b) = (a_1, b_1) \times \cdots \times (a_N, b_N)$$

as a *window*[1]. The boundary $\partial\Omega$ is denoted by M because it is a first example of a *manifold*, see Chapter 28.

Our characterisation of $\partial\Omega \in C^1$ implies in fact that[2] there exist finitely many such patches $P_i = M \cap W_i$ that cover $M = \partial\Omega$ completely,

$$M \subset P_1 \cup \cdots \cup P_k, \tag{21.4}$$

but in general $\bar{\Omega}$ is not a subset of $W_1 \cup \cdots \cup W_k$. However, there are then[3]

[1] Thus windows are open.

[2] This follows from the compactness of M.

[3] This follows from the compactness of $\bar{\Omega}$.

finitely many *other windows*

$$W_{k+1}, \ldots, W_m, \quad \overline{W}_i \subset \Omega \quad (i = k+1, \ldots, m),$$

that cover the part of Ω not yet covered by $W_1, \ldots, W_k$. Thus

$$\bar{\Omega} \subset W_1 \cup \cdots \cup W_m. \tag{21.5}$$

This covering of Ω will allow us to integrate continuous functions $u : \bar{\Omega} \to \mathbb{R}$ over $\bar{\Omega}$, using what we will call *fading*[4] *functions*.

21.1 Integrals over blocks

To integrate continuous functions

$$u : [a, b] \times [c, d] \to \mathbb{R}$$

we use partitions P as in (6.8) for $[a, b]$, and partitions

$$c = y_0 \leq y_1 \leq \cdots \leq y_M = b \tag{21.6}$$

for $[c, d]$. Lower- and undersums, or better, sums of the form[5]

$$S = \sum_{k=1}^{N} \sum_{l=1}^{M} u(\xi_k, \eta_l)(x_k - x_{k-1})(y_l - y_{l-1}), \tag{21.7}$$

with $\xi_k \in [x_{k-1}, x_k]$ and $\eta_l \in [y_{l-1}, y_l]$, then do the job. We skip the details and formulate the obvious theorem.

Theorem 21.1. *Let $u : [a, b] \times [c, d] \to \mathbb{R}$ be continuous. Then there exists a unique real number J such that for all $\varepsilon > 0$ there exists $\delta > 0$ such that for all sums S as in (21.7) it holds that*

$$|S - J| < \varepsilon,$$

provided

$$x_k - x_{k-1} < \delta \quad \textit{and} \quad y_l - y_{l-1} < \delta \quad \textit{for all} \quad k = 1, \ldots, N, \quad l = 1, \ldots M.$$

We define the intergal of f over $[a, b] \times [c, d]$ by

$$\int_{[a,b]\times[c,d]} u = J,$$

[4] The less friendly terms is cut-off functions.

[5] See Theorem 8.15.

and we have

$$J = \int_a^b \underbrace{\int_c^d u(x,y)\,dy}_{\text{continous function of } x}\, dx = \int_c^d \underbrace{\int_a^b u(x,y)\,dx}_{\text{continous function of } y}\, dy,$$

with

$$|J| = \left|\int_{[a,b]\times[c,d]} u\right| \le \int_{[a,b]\times[c,d]} |u|.$$

The repeated integrals are handled by the integration techniques for continuous functions on closed bounded intervals, see Theorems 8.6 and 8.15. Theorem 21.1 generalises to $u : [a,b] \to X$ with

$$[a,b] = [a_1,b_1] \times \cdots \times [a_N,b_N],$$

a bounded closed block in $\mathbb{R}^N$, and X a complete metric vector space.

Exercise 21.2. This is more or less Exercises 8.16 continued. Prove Theorem 21.1 without using lower- and upper sums for $X = \mathbb{R}$. Then explain why it is also a proof for X.

21.2 Integrals over bounded smooth domains

Next we consider windows such as used in (21.4), for example (21.1). The following theorem ties the obvious outcome to the perhaps slightly technical but not less obvious definition. Here we only need continuity of the function f that describes the upper boundary.

Theorem 21.3. *Let $f : [a,b] \to (c,d)$ be continuous and let*

$$u : A = \{(x,y) :\ a \le x \le b,\ c \le y \le f(x)\} \to \mathbb{R}$$

be continuous. Then

$$\int_a^b \underbrace{\int_c^{f(x)} u(x,y)\,dy}_{\text{continuous function of } x}\, dx$$

is equal to

$$J = \int_A u.$$

This integral J is uniquely defined as in Theorem 21.1, with approximating sums (21.7) in which we put $u(\xi_k,\eta_l) = 0$ whenever $\eta_l > f(\xi_k)$.

Again we leave the proof to the reader, and we note that the obvious generalisations with

$$f : [a_1, b_1] \times \cdots \times [a_{N-1}, b_{N-1}] \to (a_N, b_N),$$

and, if you like, u taking values in a complete metric vector space X hold true.

Next we consider $u : \bar{\Omega} \to \mathbb{R}$ with $\partial\Omega \in C^1$, and windows as in (21.5). It is then possible to choose[6] functions $\zeta_1 \in C_c^1(W_1), \ldots, \zeta_m \in C_c^1(W_m)$ with $0 \le \zeta_i \le 1$ for $i = 1, \ldots, m$, such that

$$\zeta_1 + \cdots + \zeta_m \equiv 1 \quad \text{on a neighbourhood of } \bar{\Omega}. \tag{21.8}$$

We use each ζ_i to *fade out* u towards the boundary of the corresponding window: each function $u_i = \zeta_i u$ has its support strictly within W_i, and as the natural definition of the integral of u_i over $\bar{\Omega}$ we take[7]

$$\int_\Omega u_i = \int_{\bar{\Omega}\cap\overline{W}_i} u_i.$$

Since

$$u = u_1 + \cdots + u_m,$$

and integrals are bound to be linear functionals on $C(\bar{\Omega})$, the obvious definition of the integral of u over Ω is

$$\int_\Omega u = \int_\Omega u_1 + \cdots + \int_\Omega u_m = \sum_{i=1}^m \int_{\bar{\Omega}\cap\overline{W}_i} \zeta_i u. \tag{21.9}$$

Exercise 21.4. A bit tedious perhaps: show the outcome in (21.9) does not depend on the choice of patches and windows. Hint: given also patches $M \cap V_1, \ldots, M \cap V_l$ in windows $V_1, \ldots, V_l$ and additional windows $V_{l+1}, \ldots, V_r$, with fading functions χ_j, $j = 1, \ldots, r$, write

$$u = \sum_{i=1}^m \zeta_i \sum_{j=1}^r \chi_j u = \sum_{i=1}^m \sum_{j=1}^r \zeta_i \chi_j u = \sum_{j=1}^r \sum_{i=1}^m \chi_j \zeta_i u = \sum_{j=1}^r \chi_j \sum_{i=1}^m \zeta_i u,$$

and evaluate the individual integrals

$$\int_\Omega \zeta_i \chi_j u$$

in two ways.

---

[6] See Chapter 29, we can make sure that $\zeta_i \in C_c^\infty(W_i)$ in fact.

[7] In accordance with $\int_a^b = \int_{[a,b]} = \int_{(a,b)}$ we just put Ω as a subscript on $\int$.

Remark 21.5. *It is also possible to give such a definition if we only assume $\partial\Omega \in C$, meaning that, possibly*[8] *after a rotation, every point of the boundary is contained in a patch described by the graph of a continuous fuction. The windows we started with are an example.*

21.3 Green's Theorem

We now *integrate partial derivatives to discover a theorem*, and in particular the right hand side in (21.11) below. It involves the outwards pointing unit normal vector ν on $\partial\Omega$, as we will see from the local calculations we do in the separate boundary windows. In particular we discover[9]

$$dS_n = \sqrt{1 + \left(\frac{\partial f}{\partial x_1}\right)^2 + \cdots + \left(\frac{\partial f}{\partial x_n}\right)^2}\, dx_1 \cdots dx_n \tag{21.10}$$

as the natural generalisation of

$$ds = \sqrt{1 + f'(x)^2}\, dx,$$

which you should recognise from the high school formula

$$\int_a^b \sqrt{1 + f'(x)^2}\, dx$$

for the length of the graph of a function $f \in C^1([a,b])$.

Theorem 21.6. *Let Ω be a bounded open set in $\mathbb{R}^N = \mathbb{R}^{n+1}$ with $\partial\Omega \in C^1$, let $v : \bar{\Omega} \to \mathbb{R}$ be continuously differentiable. Then the integral of every partial derivative v_{x_j} of v evaluates as*

$$\int_\Omega v_{x_j} = \int_{\partial\Omega} \nu_j v\, dS_n. \tag{21.11}$$

The integral on the right hand side will be defined in the proof, as well as ν_j, the j^{th} component of the outwards pointing unit normal vector ν on $\partial\Omega$.

We start the proof in the case that $N = 2$ and $n = 1$. Consider a piece of the boundary described by $y = f(x)$, with $f \in C^1([a,b])$ and $c < f(x) < d$ for all $x \in [a,b]$, such that

$$\tilde{\Omega} = \Omega \cap ([a,b] \times [c,d]) = \{(x,y) : a \le x \le b,\ f(x) < y \le d\}, \tag{21.12}$$

[8] This makes a bit more technical.

[9] More on this later: Chapter 26.

and multiply v by a function $\zeta \in C^1(\mathbb{R}^2)$ which is zero outside a subset $[\tilde{a}, \tilde{b}] \times [\tilde{c}, \tilde{d}]$ of $(a, b) \times (c, d)$. Denoting the resulting product by $\tilde{v} = \zeta v$ we have from a minor variant of Theorem 21.3 that

$$\int_{\tilde{\Omega}} \tilde{v}_y = \int_a^b \left(\int_{f(x)}^d \tilde{v}_y(x, y) dy \right) dx =$$

(by Theorem 10.12)

$$-\int_a^b \tilde{v}(x, f(x))\, dx$$

$$= \int_a^b \underbrace{\frac{-1}{\sqrt{1 + f'(x)^2}}}_{\nu_y} \tilde{v}(x, f(x)) \underbrace{\sqrt{1 + f'(x)^2} dx}_{ds = dS_1} = \int_\Phi \nu_y \tilde{v}\, ds,$$

in which the subscript Φ indicates that we use the *parameterisation*

$$\Phi(x) = (x, f(x))$$

for the boundary integral. There are of course many other[10] parameterisations that can be used to compute integrals over (this part of) the boundary.

In the above calculations we recognised the y-component of the (let's call it) *normal* unit vector

$$\nu = \frac{1}{\sqrt{1 + f'(x)^2}} \begin{pmatrix} -f'(x) \\ 1 \end{pmatrix}$$

and

$$ds = |\Phi'(x)|^2\, dx = \sqrt{1 + f'(x)^2}\, dx$$

evaluated via the parameterisation $\Phi(x) = (x, f(x))$.

For the integral of $\tilde{v}_x$ we use new coordinates ξ, η defined by

$$\xi = x,\ \eta = y - f(x), \quad \text{whence} \quad x = \xi,\ y = \eta - f(\xi) \quad \text{and} \quad dx\, dy = d\xi\, d\eta$$

when transforming an integral over $(x, y) \in \tilde{\Omega}$ to an integral over

$$(\xi, \eta) \in D = \{(x, y - f(x)) :\ a \le x \le b,\ f(x) \le y \le d\}.$$

Indeed, defining $\phi(\xi, \eta)$ by

$$\phi(\xi, \eta) = \tilde{v}(x, y) \quad \text{we have} \quad \tilde{v}_x(x, y) = \phi_\xi(\xi, \eta) - f'(\xi)\phi_\eta(\xi, \eta)$$

[10] With issues for later worries.

via the chain rule, whence[11]

$$\int_{\tilde\Omega} \tilde v_x = \int_D (\phi_\xi - f'\phi_\eta) = \int_D \phi_\xi - \int_D f'\phi_\eta$$

$$= \int_0 \left(\int_a^b \phi_\xi(\xi,\eta)\,d\xi \right) d\eta - \int_a^b \left(\int_0 f'(\xi)\phi_\eta(\xi,\eta)\,d\eta \right) d\xi$$

$$= \int_0 (\phi(b,\eta) - \phi(a,\eta))\,d\eta - \int_a^b f'(\xi) \int_0 \phi_\eta(\xi,\eta)\,d\eta\,d\xi$$

$$= \int_a^b f'(\xi)\phi(\xi,0)\,d\xi = \int_a^b v(x,f(x))f'(x)\,dx$$

$$= \int_a^b \underbrace{\frac{f'(x)}{\sqrt{1+f'(x)^2}}}_{\nu_x} \tilde v(x,f(x)) \underbrace{\sqrt{1+f'(x)^2}dx}_{ds=dS_1} = \int_\Phi \nu_x \tilde v\,ds,$$

after inserting $\sqrt{1+f'(x)^2}$ to get $ds = dS_1$ and recognising the x-component of the normal vector ν. The subscript Φ indicates again that we use the parameterisation $\Phi(x) = (x, f(x))$ for the boundary integral.

In conclusion we have

$$\int_\Omega \tilde v_x = \int_\Phi \nu_x \tilde v\,ds = \int_{\partial\Omega} \nu_x \tilde v\,ds \quad \text{and} \quad \int_\Omega \tilde v_y = \int_\Phi \nu_y \tilde v\,ds = \int_{\partial\Omega} \nu_y \tilde v\,ds,$$

in which we have taken the integrals with subscript Φ as definition of the boundary integrals over $\partial\Omega$.

Likewise we have, for the general case with $n \geq 1$, that

$$\int_\Omega (\zeta_i v)_{x_j} = \int_{\partial\Omega} \nu_j \zeta_i v\,dS_n, \tag{21.13}$$

for all $j = 1,\dots,m$ and all $i = 1,\dots,N = n+1$, with expressions like (21.10) and

$$\nu_1 = \frac{1}{\sqrt{1+|\nabla f|^2}} \frac{\partial f}{\partial x_1}, \dots, \nu_N = \frac{1}{\sqrt{1+|\nabla f|^2}} \frac{\partial f}{\partial x_n},$$

$$\nu_N = \nu_{n+1} = \frac{-1}{\sqrt{1+|\nabla f|^2}}$$

for the normal unit vector ν. Note that in (21.13) the integrals with $i = k+1,\dots,m$ all vanish.

[11] We drop a conveniently chosen fixed upper bound in the η-integrals from the notation.

We now use the fading functions to conclude that

$$\int_\Omega v_{x_j} = \int_\Omega \sum_{i=1}^m (\zeta_i v)_{x_j} = \sum_{i=1}^m \int_\Omega (\zeta_i v)_{x_j} = \sum_{i=1}^m \int_{\partial\Omega} \nu_j \zeta_i v \, dS_n.$$

This latter expression is what we take as the definition of the boundary integral

$$\int_{\partial\Omega} \nu_j v \, dS_n,$$

in much the same way as in (21.9). We can then conclude that

$$\int_\Omega v_{x_j} = \int_{\partial\Omega} \nu_j v \, dS_n,$$

which is (21.11) in Theorem 21.6.

Remark 21.7. *If we put a subscript j on v, and view v_j as the coordinates of a vector field V, we obtain*

$$\int_\Omega \nabla \cdot V = \int_{\partial\Omega} \nu \cdot V \, dS_n, \tag{21.14}$$

the statement of the Gauss Divergence Theorem.

Remark 21.8. *Applying (21.11) to the product of v and some other function $\zeta \in C^1(\Omega)$ we obtain the integration by parts formula*

$$\int_\Omega v_{x_i} \zeta = \int_{\partial\Omega} \zeta v \, \nu_i dS_{N-1} - \int_\Omega \zeta_{x_i} v. \tag{21.15}$$

For ζ we may take a function such as one of the ζ_i in (21.8) to have integrals of functions supported in one single block $[a, b]$.

Remark 21.9. *The above approach avoids reparameterisations and the use of other parameterisations to define and compute integrals over manifolds such as $M = \partial\Omega$ and other manifolds, see Chapter 28. Of course we need these later too, which requires Chapters 23 and 27.*

Exercise 21.10. In physics results like (21.6) are usually taken for granted in view of the trivial case that

$$\Omega = (a, b) = (a_1, b_1) \times (a_2, b_2)$$

is a rectangle parallel to the axes[12]. Verify directly that (21.14) holds for $v : [a, b] \to \mathbb{R}^2$ continuously differentiable.

[12] https://www.quora.com/What-is-the-plural-of-axis

Exercise 21.11. Suppose that the boundary of a bounded open set $\Omega \subset \mathbb{R}^2$ is given by a periodic solution of a system of differential equations $\dot{x} = P(x, y)$ and $\dot{y} = Q(x, y)$, with $P, Q : \mathbb{R}^2 \to \mathbb{R}$ continuously differentiable on Ω. Show that

$$\iint_{\Omega} \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y}\right) dxdy = 0.$$

22 Fourier theory

In another set of notes this began with the odd function defined by

$$f_7(x) = \sin x - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \frac{\sin 4x}{4} + \frac{\sin 5x}{5} - \frac{\sin 6x}{6} + \frac{\sin 7x}{7},$$

which is periodic with period 2π. On the interval $(-\pi, \pi)$ the graph of f_7 is close to the graph of $f(x) = \frac{1}{2}x$. Replace 7 by N, take larger and larger N, and conclude that apparently

$$\frac{x}{2} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\sin kx}{k} \tag{22.1}$$

for these values of x. For lots of other examples see Section 22.9 below, but in Section 22.2 we first cut a long story short. Another story not told here could start from these two exercises.

Exercise 22.1. Connection with power series: The right hand side of (22.1) is the imaginairy part of

$$\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \zeta^k, \quad \zeta = e^{ix}.$$

Determine the sum of this power series for $|\zeta| < 1$. Hint: differentiate with respect to ζ, take the sum and then the primitive.

Exercise 22.2. The complex version of the Leibniz criterion says that the series in Exercise 22.1 converges for all ζ with $|\zeta| = 1$ except $\zeta = -1$. Assume that the sum you found in Exercise 22.1 is valid for all such ζ. Verify (22.1).

22.1 The sawtooth function

In the spirit of Exercise 22.2, and perhaps Section 9.6, consider

$$1 + \zeta + \zeta^2 + \cdots + \zeta^{N-1} \quad \text{and its primitive} \quad \zeta + \frac{1}{2}\zeta^2 + \cdots + \frac{1}{N}\zeta^N,$$

put

$$\zeta = e^{ix} = \cos x + i \sin x,$$

and take the imaginairy part multiplied by a cosmetic 2. What you get is what we will call

$$Z_N(x) = \sum_{n=1}^{N} \frac{2}{n} \sin nx. \tag{22.2}$$

Exercise 22.3. Show that

$$Z_N(x) = x - \int_0^x D_N(s)ds = \pi - x + \int_x^\pi D_N(s)ds,$$

in which[1]

$$D_N(s) = \frac{\sin(N + \frac{1}{2})s}{\sin \frac{s}{2}} = \sum_{n=-N}^{N} e^{inx}.$$

Exercise 22.4. (continued) Prove that as

$$Z_N(x) \to \pi - x \quad \text{as} \quad N \to \infty$$

uniformly on every interval $[\delta, \pi]$ with $\delta > 0$. Then determine

$$Z(x) = \lim_{N\to\infty} Z_N(x)$$

for every $x \in \mathbb{R}$.

Exercise 22.5. (continued) The integral

$$\int_0^x D_N(s)ds$$

has extrema in the zero's of D_N. The first maximum M_N to the right of $x = 0$ is in

$$x = \frac{\pi}{N + \frac{1}{2}}.$$

Show that

$$M_N = \int_0^{\frac{\pi}{N+\frac{1}{2}}} \frac{\sin(N + \frac{1}{2})s}{\sin \frac{s}{2}} ds = 2 \int_0^\pi \frac{\sin t}{t} \frac{\frac{t}{2N+1}}{\sin \frac{t}{2N+1}} dt.$$

[1] See (22.9).

Exercise 22.6. (continued) Show that

$$\frac{\frac{t}{2N+1}}{\sin\frac{t}{2N+1}} \to 1,$$

uniformly on $t \in [0, \pi]$.

Exercise 22.7. (continued) Show that

$$M_N \to 2\int_0^\pi \frac{\sin t}{t} dt$$

as $N \to \infty$.

Exercise 22.8. (continued) You must have seen[2] the thoroughly improper integral

$$\int_0^\infty \frac{\sin t}{t}\, dt = \frac{\pi}{2}.$$

So now explain why the first maximum of $Z_N(x)$ to the right of $x = 0$ converges to

$$2\int_0^\pi \frac{\sin t}{t} dt > \pi = \lim_{x\downarrow 0} Z(x)$$

as $N \to \infty$.

Remark 22.9. *Conclusion: the function sequence Z_N converges pointwise to the sawtooth function Z which is defined by being 2π-periodic, odd[3], and $Z(x) = \pi - x$ for x between 0 and[4] π, but its maxima and minima near 0 and multiples of 2π over- and undershoot the values $Z(0^\pm) = \pm\pi$ by a factor of about 1.178979744.*

22.2 Fourier series

We first consider complex valued 2π-periodic continuous functions. Piecewise continuous functions as usually considered in this context, are de facto functions of the form

$$g(x) = f(x) + \sum_{k=1}^{m} A_k Z(x - \xi_k) \quad \text{with} \quad f \in C_{2\pi} \quad \text{and} \quad A_k \in \mathbb{C},\ \xi_k \in (0, \pi),$$

[2] Computed using the complex function $\frac{e^{iz}}{z}$.
[3] So odd, draw a sawtooth picture of its graph.
[4] And thus also for $0 < x < 2\pi$.

and since we already understand Z and its Fourier series we may just as well restrict the attention to $f \in C_{2\pi}$.

Definition 22.10. *The space of 2π-periodic continuous functions $f : \mathbb{R} \to \mathbb{C}$ is denoted by $C_{2\pi}$, a complete metric (complex vector) space with respect to the metric*[5]

$$d(f,g) = \max_{x\in\mathbb{R}} |f(x) - g(x)|.$$

For $f \in C_{2\pi}$ we consider the *Fourier series* of f, namely the right hand side of the $\sim$ symbol in

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{inx} = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx), \tag{22.3}$$

in which

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx, \quad b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx, \tag{22.4}$$

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}\,dx \tag{22.5}$$

are the *Fourier coefficients* of f. In (22.3) we use the symbol $\sim$ because it's hard to say in which sense the left and the right hand side are equal to one another. We sometimes write

$$f(x) \sim \sum_{n=-\infty}^{\infty} \hat{f}(n)e^{inx}, \quad \hat{f}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-inx}\,dx, \tag{22.6}$$

and we choose *not to modify* this notation. We're fine with 2π appearing only in the integral formula, it's the length of the interval of integration.

The formulas for the coefficients may be derived from considerations involving the L^2-inner product in (22.48) below, but in what follows we take them for granted and see what we can say about the N-th partial sum

$$S_N f(x) = \sum_{n=-N}^{N} c_n e^{inx} = \frac{a_0}{2} + \sum_{n=1}^{N} (a_n \cos nx + b_n \sin nx) \tag{22.7}$$

of the Fourier series of f in (22.3). A miraculous calculation[6] with complex exponential geometric series then first tells us that

$$S_N f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} D_N(y) f(x-y)\,dy = \frac{1}{2\pi}\int_{-\pi}^{\pi} D_N(y) f(x+y)\,dy, \tag{22.8}$$

[5] Just like in Section 4.2.

[6] You did it in Exercise 22.3, we'll do it below for Fourier series of functions $f(x,y)$.

in which

$$D_{_N}(x) = \frac{\sin(N+\frac{1}{2})x}{\sin\frac{1}{2}x} = \sum_{k=-N}^{N} e^{ikx} \tag{22.9}$$

is called the Dirichlet kernel. We say that $2\pi\, S_{_N} f$ is the convolution of $D_{_N}$ and f. The 2π-periodic function $D_{_N}$ is called the Dirichlet kernel. For larger and larger N it concentrates near 0, with a narrower and narrower peak, while its integral remains constant and equal to 2π. That's good. What's bad is that away from 0 it oscillates between maxima and minima which in absolute value remain larger than 1 as N gets large. These properties will only make $S_{_N} f(x)$ converge to $f(x)$ if f is nicer than just being in the good space $C_{2\pi}$, while the proofs of such statements are a bit cumbersome.

The average of $D_0, \ldots, D_{_N}$ however, which via another miraculous calculation is equal to

$$F_{_N}(x) = \frac{1}{N+1}\left(D_0(x) + \cdots + D_{_N}(x)\right) = \frac{1}{N+1}\frac{\sin^2\frac{(N+1)x}{2}}{\sin^2\frac{x}{2}}, \tag{22.10}$$

the 2π-periodic Féjer kernel, is much nicer. It is nonnegative, has integral 2π, and concentrates in 0 as N gets large, thereby forcing it to be small away from mulitples of 2π. Tanja[7] called such functions *good kernels*. This not so very hard theorem explains why.

Theorem 22.11. *Define*

$$\sigma_{_N} f = \frac{1}{N+1}(S_0 f + S_1 f + \cdots + S_{_N} f),$$

the Cesàro sums of f. Then

$$\sigma_{_N} f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_{_N}(\xi) f(x-\xi) d\xi,$$

and $\sigma_{_N} f \to f$ in $C_{2\pi}$ if $f \in C_{2\pi}$. That is, the convergence is uniform.

Exercise 22.12. Let $f \in C_{2\pi}$, let $M = |f|_{\max}$ be the maximum of $|f(x)|$ on $\mathbb{R}$, and let $\varepsilon > 0$. Explain why there exists $\delta > 0$ such that

$$|\sigma_{_N} f(x) - f(x)| = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_{_N}(\xi)|f(x+\xi) - f(x)| d\xi \le \varepsilon + \frac{2M}{N+1}\frac{1}{\sin^2\frac{\delta}{2}}$$

if $|\xi| < \delta$. Hint: split the integral in 3 parts. Then prove Theorem 22.11.

[7] She's no longer at the UvA unfortunately.

Exercise 22.13. Let $\xi \in (0, \pi)$. Let $Z(x)$ as in Exercise 22.4 and further be the sawtooth function. Determine the Fourier coefficients of $Z(x - \xi)$ and show that the partial sums are equal to $Z_N(x - \xi)$. Describe their behaviour as $N \to \infty$.

22.3 Fourier series with multiple variables

Thanks to the multiplicative property of

$$\exp(z) = e^z$$

these results generalise to functions of more variables, with remarkably nice multiplicative properties of the two convolution kernels (22.9) and (22.10) in (22.11) and (22.13) below. To see how let f be in $C_{2\pi}(\mathbb{R}^2)$, i.e. $f(x, y)$ is continuous in (x, y), and 2π-periodic in both x and y seperately . As before we write

$$f(x, y) \sim \sum_{m,n=-\infty}^{\infty} c_{mn} e^{i(mx+ny)},$$

but now with

$$c_{mn} = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f(x, y) e^{-i(m\xi+n\eta)} \, d\xi \, d\eta.$$

We prepared for the arguments below by using dummy variables ξ and η instead of x and y.

It follows that

$$S_{MN} f(x, y) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} c_{mn} e^{i(mx+ny)} =$$

$$\sum_{m=-M}^{M} \sum_{n=-N}^{N} \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f(\xi, \eta) e^{-i(m\xi+n\eta)} \, d\xi \, d\eta \, e^{i(mx+ny)} =$$

$$\frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f(\xi, \eta) \sum_{m=-M}^{M} \sum_{n=-N}^{N} e^{im(x-\xi)} \, e^{im(y-\eta)} \, d\xi \, d\eta =$$

$$\frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f(\xi, \eta) \underbrace{\sum_{m=-M}^{M} e^{im(x-\xi)}}_{D_M(x-\xi)} \underbrace{\sum_{n=-N}^{N} e^{in(y-\eta)}}_{D_N(y-\eta)} \, d\xi \, d\eta,$$

so[8]

$$S_{MN}f(x,y) = \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x+\xi, y+\eta)D_M(\xi)D_N(\eta)\,d\xi\,d\eta, \qquad (22.11)$$

in which

$$D_M(\xi) = \sum_{m=-M}^{M} e^{im\xi} = \frac{e^{-iM\xi} - e^{i(M+1)\xi}}{1-e^{i\xi}} = \frac{\sin(M+\frac{1}{2})\xi}{\sin\frac{1}{2}\xi}, \qquad (22.12)$$

and likewise

$$D_N(y) = \frac{\sin(N+\frac{1}{2})\eta}{\sin\frac{1}{2}\eta}.$$

The averages

$$\sigma_{MN}f(x,y) = \frac{1}{(M+1)(N+1)}\sum_{m=0}^{M}\sum_{n=0}^{N} S_{mn}f(x,y) =$$

$$\frac{1}{(2\pi)^2}\frac{1}{(M+1)(N+1)}\sum_{m=0}^{M}\sum_{n=0}^{N}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x-\xi, y-\eta)D_m(\xi)D_n(\eta)\,d\xi\,d\eta$$

$$\frac{1}{(2\pi)^2}\frac{1}{(M+1)(N+1)}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x-\xi, y-\eta)\sum_{m=0}^{M}D_m(\xi)\sum_{n=0}^{N}D_n(\eta)\,d\xi\,d\eta$$

rewrite as

$$\sigma_{MN}f(x,y) = \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} F_M(\xi)F_N(\eta)f(x+\xi, y+\eta)\,d\xi\,d\eta, \qquad (22.13)$$

in which

$$F_M(\xi) = \frac{1}{M+1}\sum_{m=0}^{M} D_m(\xi) = \frac{1}{M+1}\frac{\sin^2\frac{(M+1)\xi}{2}}{\sin^2\frac{\xi}{2}}, \qquad (22.14)$$

as you should verify, and likewise for $F_N(\eta)$. Again it follows that

$$\sigma_{MN}f \to f \quad \text{in} \quad C_{2\pi}(\mathbb{R}^2) \quad \text{as} \quad M, N \to \infty,$$

and for sufficiently smooth f in $C_{2\pi}(\mathbb{R}^2)$ that $S_{MN}f \to f$ in $C_{2\pi}(\mathbb{R}^2)$ because once both limits exist[9] they have to be the same. Clearly all this generalises to $f(x_1, \ldots, x_n)$, $f \in C_{2\pi}(\mathbb{R}^n)$.

[8] See Remark **??**.

[9] An easy variant of Exercise 3.63 is needed here.

22.4 Derivation of the integral Fourier transform

I first discuss Fourier transform for functions of one variable and start from the intuitive presentation in §7.1 of Olver's very nice PDE book, which I slightly modify and then merge with the rigorous approach in Folland's wonderful Real Analysis book. It should be clear from the last part of the previous section that for more variables the story is much the same. What I like about the arguments above and below is that the theory can be built on Riemann integrals[10].

For a start let $f = f(x)$ be defined and continuous on the real line, and $f(x) = 0$ for $|x| \geq l$. If the function F is defined by

$$F(y) = f(x), \quad \frac{x}{l} = \frac{y}{\pi},$$

we can write

$$F(y) \sim \sum_{n=-\infty}^{\infty} C_n e^{iny} \tag{22.15}$$

just as in (22.3) for f. Now assume that f and thus F is smooth. We write the Fourier coefficients $C_n = \hat{F}(n)$ of $F(y)$ as η-integrals. It follows for $x \in [-l, l]$ that

$$f(x) = F(y) = \sum_{n=-\infty}^{\infty} \underbrace{\frac{1}{2\pi} \int_{-\pi}^{\pi} F(\eta) e^{-in\eta} d\eta}_{\hat{F}(n)} \, e^{iny},$$

in which the series is uniformly convergent, uniformly in $y \in \mathbb{R}$ that is.

The Fourier series of f on the interval $[-l, l]$ is obtained via scaling from the uniformly convergent Fourier series of the 2π-periodic smooth extension of the smooth function F. This gives

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \frac{\pi}{l} \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-l}^{l} f(\xi) e^{-i\frac{n\pi}{l}\xi} \, d\xi}_{\text{this is } \widehat{f}(\frac{n\pi}{l}) \text{ if supp} f \subset [-l,l].} \, e^{i\frac{n\pi}{l}x} \tag{22.16}$$

for $x \in [-l, l]$, in which as indicated the underbraced term is equal to

$$\begin{array}{c} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\xi) e^{-i\frac{n\pi}{l}\xi} \, d\xi \\ \uparrow \\ n\Delta k \end{array}$$

[10] Not that I dislike measure theory and Lebesgue integrals.

for l sufficiently large. Note that ξ is just a convenient dummy variable, and that we recognise

$$\frac{n\pi}{l} = n\Delta k \quad \text{as an integer multiple of} \quad \frac{\pi}{l} = \Delta k.$$

Introducing the *Fourier integral transform*[11] $\widehat{f}$ of f by

$$\widehat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\xi) e^{-ikx}\, d\xi, \tag{22.17}$$

the terms in the sum on the right hand side of (22.16) are

$$\Delta k\, \widehat{f}(n\Delta k)\, e^{ixn\Delta k}.$$

We then *see* that

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \widehat{f}(n\Delta k) e^{ixn\Delta k}\, \Delta k, \tag{22.18}$$

in which the sum looks like a Riemann sum for

$$\int_{-\infty}^{\infty} \widehat{f}(k) e^{ikx}\, dk.$$

Note that we have changed the prefactor in order to have

$$\widehat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx}\, dx \quad \text{and} \quad f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \widehat{f}(k) e^{ikx}\, dk$$

as the outcome of a limit argument for $\Delta k \to 0$.

Likewise it follows for l sufficiently large that

$$\int_{-\infty}^{\infty} |f(x)|^2\, dx = \sum_{n=-\infty}^{\infty} |\widehat{f}(n\Delta k)|^2\, \Delta k, \tag{22.19}$$

which looks like a Riemann sum for

$$\int_{-\infty}^{\infty} |\widehat{f}(k)|^2\, dk,$$

and the identities (22.18) and (22.19) remain valid if we increase l and thereby decrease the step size Δk. Both Riemann sums are independent of Δk in the limit $\Delta k \to 0$. Can we conclude that then both

$$f(x) \quad \text{and} \quad \int_{-\infty}^{\infty} |f(x)|^2\, dx$$

[11] Note the notational difference between $\widehat{f}(k)$ here and $\hat{f}(n)$ for Fourier coeffients.

are also equal to the corresponding k-integrals? The answer is yes if $\widehat{f}(k)$ is continuous and decays sufficiently fast as $|k| \to \infty$, so as to make the tails of both the k-integrals and the Riemann sums small. Then we can restrict the convergence argument to integrals and Riemann sums on bounded k-intervals.

For smooth compactly supported functions $f : \mathbb{R} \to \mathbb{C}$ such decay rates are obtained using integration by parts. Since

$$\int_{-\infty}^{\infty} f(x)e^{-ikx}\,dx = \frac{1}{ik}\int_{-\infty}^{\infty} f'(x)e^{-ikx}\,dx = \frac{1}{(ik)^2}\int_{-\infty}^{\infty} f''(x)e^{-ikx}\,dx,$$

we have[12]

$$\widehat{f}(k) = \frac{\widehat{f'}(k)}{ik} = \frac{\widehat{f''}(k)}{(ik)^2}, \tag{22.20}$$

and so on for the Fourier transforms of the derivatives of f, which remain dry[13] under the wide hat in the notation. Therefore

$$|\widehat{f}(k)| \le \frac{1}{\sqrt{2\pi}k^2}\int_{-\infty}^{\infty} |f''(x)|\,dx.$$

For the limit $\Delta k \to 0$ in both (22.16) and (22.19) this suffices. The continuity of f'' and the compact support of f thus imply all but the smoothness statements in the following theorem.

Theorem 22.14. *Let $f : \mathbb{R} \to \mathbb{C}$ be in C_c^2, i.e. f and f' are differentiable on $\mathbb{R}$, f'' is continuous, and f has compact support[14]. Then (22.17) defines a smooth function $\widehat{f} : \mathbb{R} \to \mathbb{C}$ satisfying the estimate*

$$|\widehat{f}(k)| \le \frac{1}{\sqrt{2\pi}\,k^2}\int_{-\infty}^{\infty} |f''(x)|\,dx = \frac{|f''|_1}{\sqrt{2\pi}\,k^2}$$

for every real $k \neq 0$, so

$$|\widehat{f}\,|_1 = \int_{-\infty}^{\infty} |\widehat{f}(k)|\,dk < \infty.$$

Moreover,

$$f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \widehat{f}(k)e^{ikx}\,dk \quad \text{and} \quad \int_{-\infty}^{\infty} |f(x)|^2\,dx = \int_{-\infty}^{\infty} |\widehat{f}(k)|^2\,dk.$$

[12] Because $|e^{i\xi x}| = 1$.

[13] Only barely in case of f'...

[14] So $f, f, f'' \in C_c$.

Proof. It remains to show that $\hat{f}$ is smooth. In fact $f \in C_c$ suffices to have $\widehat{f}$ *smooth*, thanks to Theorem 14.5. By differentiation under the integral we have

$$(\widehat{f})'(k) = -ik\,\widehat{f}(k), \quad (\widehat{f})''(k) = (-ik)^2\,\widehat{f}(k), \tag{22.21}$$

and so on[15] for the derivatives of $\widehat{f}$. □

Remark 22.15. *We say that $\widehat{f}$ is in C^∞. If f is also in C_c^∞, the class of smooth compactly supported $\mathbb{C}$-valued functions*[16] *on $\mathbb{R}$, then all derivatives of $\widehat{f}$ go to zero with a uniform decay rate faster than every p-th power of $|k|$, because (22.20) and (22.21) imply*

$$|(\widehat{f})^{(n)}(k)| \leq \frac{|k|^n\,|f^{(m)}|_1}{\sqrt{2\pi}\,|k|^m}$$

for $k \neq 0$. For $f \in C_c^\infty$ we thus have

$$\forall_{n\in\mathbb{N}_0}\,\forall_{p\in\mathbb{N}_0}\,\exists_{C>0}\,\forall_{k\in\mathbb{R}}\,:\,|k|^p\,|(\widehat{f})^{(n)}(k)| \leq C. \tag{22.22}$$

The class of functions in C^∞ that satisfy (22.22) is called the Schwarz class. If f is in C_c^∞ then $\hat{f}$ is in $\mathcal{S}$, but in general not in C_c^∞.

Theorem 22.16. *The implication*

$$f \in \mathcal{S} \implies \widehat{f} \in \mathcal{S}$$

holds for all $f : \mathbb{R} \to \mathbb{C}$.

Proof. For $f \in C_c^\infty$ this follows from Theorem 14.5 in Section 14.2. Now watch

`https://canvas.vu.nl/courses/38607/pages/differentiation-under-the-integral`

to see that the proof for $f \in \mathcal{S}$ follows from a variant of Theorem 14.5 still to be included in Section 14.2. The upshot is that the chains (22.20) and (22.21) are also valid for $f \in \mathcal{S}$. □

22.5 The Fourier transform as a bijection

The pairing[17]

$$\widehat{f}(\xi) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)e^{-i\xi x}\,dx \quad \text{and} \quad f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \widehat{f}(\xi)e^{i\xi x}\,d\xi \tag{22.23}$$

[15] Not the same formulas as in (22.20) to begin with, but in the end they really are....

[16] The test functions used in the theory of *distributions*.

[17] I now prefer ξ as name for the Fourier variable, so much for wave numbers.

discovered with (22.20) should of course define bijections between suitable pairs of function spaces. Which function spaces X allow $f \leftrightarrow \widehat{f}$ as a bijection between X and X itself?

To answer this question we first re-examine the definition

$$\widehat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-i\xi x}\, dx \tag{22.24}$$

of $\hat{f}$. For $\widehat{f}$ to be well defined it certainly suffices that f is in $C \cap L^1$, i.e.

$$f : \mathbb{R} \to \mathbb{C} \text{ is continuous and } |f|_1 = \int_{\mathbb{R}} |f| < \infty. \tag{22.25}$$

This is because

$$|\widehat{f}(\xi)| = \left| \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-i\xi x}\, dx \right| \leq \frac{|f|_1}{\sqrt{2\pi}}.$$

Moreover, this estimate implies that if a sequence f_n in $C \cap L^1$ is a Cauchy sequence with respect to the 1-norm, then $\widehat{f_n}$ is a Cauchy sequence with respect to the ∞-norm.

Theorem 22.17. *The space C_0 of continuous functions $f : \mathbb{R} \to \mathbb{C}$ with $f(x) \to 0$ as $|x| \to \infty$ is a Banach space with respect to the norm defined by*

$$|f|_{max} = \max_{x \in \mathbb{R}} |f(x)|.$$

The space C_0 is a closed subspace of the space C_b of bounded continuous functions $f : \mathbb{R} \to \mathbb{C}$, on which

$$|f|_\infty = \sup_{x \in \mathbb{R}} |f(x)|$$

defines the norm. This space is also a Banach space, its norm reduces to the maximum norm for $f \in C_0$. The space C_b is contained in the vector space C of all continuous functions $f : \mathbb{R} \to \mathbb{C}$, which is not a Banach space for any reasonable choice of a norm. We write $C_0 \cap L^1$ for $C_0 \cap C \cap L^1$, the class of functions f that satisfy (22.25).

Proposition 22.18. *If $f \in C \cap L^1$ then $\widehat{f} \in C_0$. If we write*

$$f(x) \quad \widehat{\to} \quad \widehat{f}(\xi)$$

then for every $y, \eta \in \mathbb{R}$ and $a > 0$ it holds that

$$e^{i\eta x} f(a(x-y)) \quad \widehat{\to} \quad \frac{1}{a} e^{-iy\xi} \widehat{f}\left(\frac{\xi - \eta}{a}\right), \tag{22.26}$$

and

$$e^{-\frac{1}{2}x^2} \quad \widehat{\to} \quad e^{-\frac{1}{2}\xi^2}.$$

Proof. To prove that $\widehat{f} \in C_0$ we take a sequence of compactly supported continuously differentiable functions f_n with $|f_n - f|_1 \to 0$. Then

$$\sqrt{2\pi}\,|\widehat{f}_n(\xi) - \widehat{f}(\xi)| = \left|\int_{-\infty}^{\infty} f_n(x)e^{-i\xi x}\,dx - \int_{-\infty}^{\infty} f(x)e^{-i\xi x}\,dx\right| =$$

$$\left|\int_{-\infty}^{\infty} (f_n(x) - f(x))e^{-i\xi x}\,dx\right| \le \int_{-\infty}^{\infty} |f_n(x) - f(x))| = |f_n - f|_1 \to 0,$$

so $\widehat{f}_n \to \widehat{f}$ uniformly on $\mathbb{R}$. Since for each n the integral

$$\int_{-\infty}^{\infty} f_n(x)e^{-i\xi x}\,dx$$

reduces to an integral over a bounded closed interval $[-R_n, R_n]$, the functions $\widehat{f}_n$ are certainly continuous and thus $\widehat{f}_n \to \widehat{f}$ in C_b. Integration by parts shows that

$$\widehat{f}_n(\xi) = \frac{\widehat{f}_n'(\xi)}{i\xi}, \quad \text{whence} \quad |\widehat{f}_n(\xi)| \le \frac{|f_n'|_1}{|\xi|} \quad \text{and} \quad \widehat{f}_n \in C_0.$$

Since C_0 is closed in C_b it follows that $\widehat{f} \in C_0$. The statement in (22.26) is easily checked. □

Proposition 22.19. *Let $f, g \in C \cap L^1$. Then $\widehat{f}, \widehat{g} \in C_0$ and*

$$\int_{-\infty}^{\infty} \widehat{f}(\xi)g(\xi)\,d\xi = \int_{-\infty}^{\infty} f(x)\widehat{g}(x)\,dx,$$

in which

$$\widehat{g}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(\xi)e^{-i\xi x}\,d\xi.$$

Proof. We have that

$$\sqrt{2\pi}\int_{-\infty}^{\infty} \widehat{f}(\xi)g(\xi)\,d\xi = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)e^{-i\xi x}\,dx\,g(\xi)\,d\xi$$

exists because $\widehat{f} \in C_0$ and $g \in C \cap L^1$. But

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)e^{-i\xi x}\,dx\,g(\xi)\,d\xi = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)g(\xi)e^{-i\xi x}\,dx\,d\xi.$$

because $f, g \in C \cap L^1$. Changing the orde of integration and using that $\widehat{g} \in C_0$ and $f \in C \cap L^1$ the statement in the proposition follows by reversing the roles of x and ξ, and of f and g. □

Theorem 22.20. *Let*

$$X = \{f \in C_0 \cap L^1 : \widehat{f} \in C_0 \cap L^1\}. \tag{22.27}$$

Then $C_c^2 \subset X$, so X is nonempty. The Fourier transform is a bijection between X and itself in the sense that

$$g(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-i\xi x}\, dx \iff f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(\xi) e^{ix\xi}\, d\xi \tag{22.28}$$

for all $f, g \in X$. In particular the equalities in (22.28) and (22.23) hold pointwise for every x and every ξ when f is in X. Moreover,

$$\int_{-\infty}^{\infty} |\widehat{f}(\xi)|^2\, d\xi = \int_{-\infty}^{\infty} |f(x)|^2\, dx \leq |f|_{max} \, |f|_1 .$$

Proof. We observe that X contains C_c^2 because of Theorem 22.14 in Section 22.4, a statement which is independent of the rest of the theorem, which we prove next. So let g be defined by the left hand side of (22.28). Note that $\hat{f} \in C_0$ because of Proposition 22.18. The extra assumption $\widehat{f} \in L^1$ is needed when we apply Proposition 22.19 with the function $g(\xi)$ and $\hat{g}(x)$ that appear there[18] replaced by

$$e^{i\eta\xi} e^{-\frac{1}{2}a^2\xi^2} \quad \widehat{\to} \quad \frac{1}{a}\, e^{-\frac{1}{2a^2}(x-\eta)^2}.$$

The left hand side of the equality Proposition 22.19 then evaluates as

$$\int_{-\infty}^{\infty} \widehat{f}(\xi) e^{i\eta\xi} e^{-\frac{1}{2}a^2\xi^2}\, d\xi \to \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{i\eta\xi}\, d\xi$$

for $a \to 0$, the limit statement holding thanks to $\widehat{f} \in L^1$.

The right hand side of the equality Proposition 22.19 becomes is

$$\int_{-\infty}^{\infty} f(x) \frac{1}{a}\, e^{-\frac{1}{2a^2}(x-\eta)^2}\, dx.$$

Put $a^2 = 2t$ to recognise, up to the usual factor, the solution formula for the heat equation $u_t = u_{xx}$ with initial data $u(0, x) = f(x)$, by convolution with the heat kernel

$$E_t(x) = \frac{1}{2\sqrt{\pi t}} e^{-\frac{x^2}{4t}}.$$

For this good kernel we have that $E_t * f \to f$ uniformly[19] as $t \downarrow 0$ for every $f \in C_0$. Get the prefactor right to conclude that the the right hand side of (22.28) holds. This finishes the proof of $\implies$ in (22.28). The proof of $\impliedby$ in (22.28) is of course similar. □

[18] Not to be confused with the g in (22.28)!

[19] For the pointwise convergence $f \in C_b \cap L^1$ suffices!

Remark 22.21. *If we denote the space of measurable complex valued Lebesgue measurable integrable functions by L^1, then*

$$f \in L^1 \implies \widehat{f} \in C_0.$$

Thus there is no point in considering possible versions of Theorem 22.20 with C_0 replaced by C_b or even C in (22.27). The assumption $f \in C \cap L^1$ is sufficient to conclude

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{ix\xi}\, d\xi \quad \textit{and} \quad \int_{-\infty}^{\infty} |\widehat{f}(\xi)|^2\, d\xi = \int_{-\infty}^{\infty} |f(x)|^2\, dx. \tag{22.29}$$

If also the (weak) derivative f' exists in L^1 then

$$\widehat{f}(\xi) = \frac{1}{i\xi} \widehat{f'}(\xi) \quad \textit{for} \quad \xi \neq 0,$$

with $\widehat{f'} \in C_0$ again. Denoting the class of measurable complex valued functions f with bounded 1-norm

$$\int_{-\infty}^{\infty} |f(x)|\, dx = |f|_1$$

by L^1 and the class with bounded 2-norm

$$\int_{-\infty}^{\infty} |f(x)|^2\, dx = |f|_2^2$$

by L^2 we have that C_c^∞ is dense in L^2.

Remark 22.22. *In fact the bijection*

$$\mathcal{F} : X \to X, \quad \mathcal{F}(f) = \widehat{f}$$

extends to an isometry

$$\mathcal{F} : L^2 \to L^2$$

because X is dense in the Hilbert space L^2 of complex Lebesgue measurable functions with finite 2-norm. Upto a reflection in x, this map is its own inverse. For all $f, g \in L^2$ we have

$$\int_{-\infty}^{\infty} f(x)\, \overline{g(x)}\, dx = \int_{-\infty}^{\infty} \widehat{f}(\xi)\, \overline{\widehat{g}(\xi)}\, d\xi, \tag{22.30}$$

which you should compare to Theorem 22.19. If $f \in L^1 \cap L^2$ we can copy (22.17), replacing the integral by a Lebesgue integral.

Theorem 22.23. *The space X contains the Schwarz class $\mathcal{S}$. Thus (22.23) defines a bijection on $\mathcal{S}$ and we have*

$$g(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-i\xi x}\, dx \iff f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(\xi) e^{ix\xi}\, d\xi$$

for all f, g in $\mathcal{S}$, just as in (22.28) for f, g in X.

Proof. Theorem 22.16 implies that $\mathcal{S}$ itself maps to $\mathcal{S}$. Therefore both f and $\widehat{f}$ are in $C_0 \cap L^1$ if $f \in S$, and (22.28) holds for $f, g \in S$. □

22.6 Connection with probability theory

Considering the Fourier transform

$$f \to \phi = \widehat{f}$$

on L^2 we have that

$$|f|_2 = 1 \iff |\phi|_2 = 1,$$

in which case both $x \to |f(x)|^2$ and $\xi \to |\phi(\xi|^2$ are probability distributions, say of the stochastic variables X and Ξ, with possibly great expectations

$$EX = \int_{-\infty}^{\infty} x\, |f(x)|^2\, dx \quad \text{and} \quad E\,\Xi = \int_{-\infty}^{\infty} \xi\, |\phi(\xi)|^2\, d\xi,$$

if these integrals exist. If so, then the exponential factors in

$$e^{i\eta x} f(x - y) \quad \widehat{\to} \quad e^{-iy\xi} \phi(\xi - \eta)$$

don't change X and Ξ but the shifts do. They change X and Ξ in $y + X$ and $\eta + \Xi$ and can therefore be chosen to put the expectations equal to zero.

In questions about variances we can thus restrict our attention to stochastic variables X and Ξ with zero expectation. Integrating the integral for the squared 2-norm of f by parts with the 1-trick[20] to get

$$1 = \int_{-\infty}^{\infty} |f(x)|^2\, dx = \Big[x\, f(x)\overline{f(x)}\Big]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} x \left(f'(x)\overline{f(x)} + f(x)\overline{f'(x)} \right) dx$$

$$= -(f', f_1) - \overline{(f', f_1)} \le 2|f'|_2\, |f_1|_2 = 2\, |\phi_1|_2\, |f_1|_2,$$

in which for the moment f_1, ϕ_1 are defined by

$$f_1(x) = xf(x) \quad \text{and} \quad \phi_1(x) = x\phi(x).$$

[20]Recall $\int_1^x \ln s\, ds = \int_1^x 1\, \ln s\, ds = [s \ln s]_1^x - \int_1^x s\, \frac{1}{s}\, ds = x \ln x + x - 1$.

Note that with this notation the rules for the derivatives of $\phi \in \mathcal{S}$ are

$$\widehat{\phi}' = -i\,\widehat{\phi_1} \quad \text{and} \quad \widehat{\phi'} = i\,\widehat{\phi}_1.$$

Since

$$\int_{-\infty}^{\infty} |f_1(x)|^2\,dx = \int_{-\infty}^{\infty} |x|^2\,|f(x)|^2\,dx,$$
$$\int_{-\infty}^{\infty} |\phi_1(x)|^2\,dx = \int_{-\infty}^{\infty} |x|^2\,|\phi(x)|^2\,dx,$$

this establishes the estimate

$$4\,EX^2\,E\,\Xi^2 \geq 1$$

for the product of the variations of X and Ξ. For the standard deviations the conclusion is that

$$2\sigma(X)\,\sigma(\Xi) \geq 1, \tag{22.31}$$

which corresponds to the Heisenberg Uncertainty Principle.

22.7 Convolutions and Fourier solution methods

Both Fourier series and Fourier integrals are called Fourier transforms. In both cases we can ask about the Fourier transform of a convolution $f*g$ and of a product fg. Statements about products can of course be obtained using the inverse transform but below we discuss a more direct approach. Statements about convolutions are somewhat easier, and in the context of solving linear differential equations with constant coefficients they are extremely useful.

For example, the ordinary differential quation

$$-u''(x) + u(x) = f(x)$$

transforms to the algebraic equation

$$(\xi^2 + 1)\,\widehat{u}(\xi) = \widehat{f}(\xi), \quad \text{so} \quad \widehat{u}(\xi) = \frac{1}{1+\xi^2}\,\widehat{f}(\xi).$$

As we explain below, the solution of the ODE is found by taking the convolution of the inverse of

$$\frac{1}{1+\xi^2}$$

with f itself, with some π-dependent prefactor. And likewise for problems in which x is taken modulo 2π, which we discuss first. Indeed, if we solve the same equation for $f \in C_{2\pi}$, then the solution will have to be in $C^2_{2\pi}$,

because differentiability of u' implies that u' is continuous and thereby that u is continuous. But then[21] $u'' = f - u$ is also continuous, so we know *a priori* that the Fourier coefficients $\hat{u}(n)$ are going to have a quadratic decay. We have that

$$\hat{u''}(n) = n^2 \hat{u}(n)$$

and therefore the differential equation becomes the algebraic equation

$$(1+n^2)\,\hat{u}(n) = \hat{f}(n) \quad \text{whence} \quad \hat{u}(n) = \underbrace{\frac{1}{1+n^2}}_{=\hat{g}(n)?}\,\hat{f}(n).$$

Now recall that the convolution of two 2π-periodic integrable functions is defined by

$$(f*g)(x) = \int_{-\pi}^{\pi} f(x-y)g(y)\,dy = \int_{-\pi}^{\pi} f(y)g(x-y)\,dy \tag{22.32}$$

whenever one of these integrals has a meaning for (almost) all x, which is certainly the case if $f, g \in C_{2\pi}$. *Alternatively, read the next calculation backwards to discover why we introduce $f*g$.* Either way, we have

$$\begin{aligned}
\int_{-\pi}^{\pi} (f*g)(x)\,e^{-inx}\,dx &= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x-y)g(y)\,dy\,e^{-inx}\,dx \\
&= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x-y)g(y)\,e^{-inx}\,dy\,dx \\
&= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x-y)g(y)\,e^{-in(x-y)}\,e^{-iny}\,dx\,dy \\
= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x)g(y)\,e^{-inx}\,e^{-iny}\,dx\,dy &= \underbrace{\int_{-\pi}^{\pi} f(x)\,e^{-inx}\,dx}_{2\pi\hat{f}(n)}\,\underbrace{\int_{-\pi}^{\pi} g(y)\,e^{-iny}\,dy}_{2\pi\hat{g}(n)}.
\end{aligned}$$

Upto an annoying factor 2π the Fourier coefficients of $f*g$ are the products of the Fourier coefficients of f and of g. Moreover, Corollary **??** allows to conclude that the statement in this theorem holds.

Theorem 22.24. *Let $f, g \in C_{2\pi}$. Then*

$$(f*g)(x) = 2\pi \sum_{n=-\infty}^{\infty} \hat{f}(n)\hat{g}(n)e^{inx} \quad \textit{for all} \quad f, g \in C_{2\pi}, \tag{22.33}$$

[21] This line of reasoning only works for ordinary differential equations unfortunately.

in which $2\pi\hat{f}(n)\hat{g}(n)$ are the complex Fourier coefficients of $f * g$. The right hand side is uniformly convergent because

$$|\hat{f}\hat{g}|_1 = \sum_{n=-\infty}^{\infty} |\hat{f}(n)\hat{g}(n)| \leq \sqrt{\sum_{n=-\infty}^{\infty} |\hat{f}(n)|^2} \sqrt{\sum_{n=-\infty}^{\infty} |\hat{g}(n)|^2} = |\hat{f}|_2 \, |\hat{g}|_2 .$$

Summing up, the Fourier coefficients of $f * g$ followed by direct calculation and the Fourier series converges uniformly because

$$|\hat{f}\hat{g}|_1 \leq |\hat{f}|_2 \, |\hat{g}|_2 .$$

For our above solution u all this leads to

$$u = G * f, \quad G(x) = \frac{1}{2\pi} \sum_{n \in \mathbb{Z}} \frac{1}{1+n^2} \, e^{inx} = \frac{1}{\pi} \left(\frac{1}{2} + \sum_{n=1}^{\infty} \frac{\cos nx}{1+n^2} \right).$$

Note how we thus avoid the use of solutions with f replaced by the Dirac δ-function. In principal we can then avoid the distributions altogether when using Fourier transformations to solve linear differential equations with constant coefficients. But if we don't we come to realise that the solutions of equations such as $-u''(x)+u(x) = \delta(x)$ fit nicely in the mathematical theory that combines Fourier transformations and distributions.

We very briefly touch upon this theory in Section 22.8 and illustrate the advantage of the use of the δ-function with an example. Usually way before the theory is understood, solutions of equations with δ as the inhomogeneous term on the right hand side are computed using smooth solutions of the homogeneous equation with a singularity in $x = 0$. In the 2π-periodic case for $-u''(x) + u(x) = \delta(x)$ this means a negative jump in the first derivative at $x = 0$ (and in the other integer multiples of 2π), because $u'' = u - \delta$, and the "integral" of $\delta(x)$ over any interval $(-\varepsilon, \varepsilon)$ is equal to 1. Combined with symmetry and 2π-periodicity this implies that[22]

$$G(x) = \frac{\cosh(x-\pi)}{2\sinh \pi} \quad \text{for} \quad 0 \leq x \leq 2\pi,$$

and you easily check that indeed[23]

$$G(x) = \sum_{n \in \mathbb{Z}} \frac{e^{inx}}{n^2+1}.$$

[22] You may like to compare $-G'(x+\pi)$ to the saw tooth in $Z(x)$ in Section 22.1.

[23] Draw the 2π-periodic graph and compute $2\pi\hat{G}(n)$ as $\int_0^{2\pi} \cosh(x-\pi)\, e^{-inx}\, dx$.

Next we consider the Fourier coefficients of fg. This is more difficult. Again Corollary **??** tells us that

$$f(x) \sim \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{inx} \quad \text{and} \quad g(x) \sim \sum_{n=-\infty}^{\infty} \hat{g}(n) e^{inx}$$

have a clear meaning if

$$|\hat{f}|_1 = \sum_{n=-\infty}^{\infty} \hat{|}f(n)| < \infty \quad \text{and} \quad |\hat{g}|_1 = \sum_{n=-\infty}^{\infty} |\hat{g}(n)| < \infty, \tag{22.34}$$

because then the right hand sides in

$$f(x) = \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{inx} \quad \text{and} \quad g(x) = \sum_{n=-\infty}^{\infty} \hat{g}(n) e^{inx},$$

are both uniformly absolutely convergent Fourier series. This justifies the calculation $f(x)g(x) =$

$$\sum_{k=-\infty}^{\infty} \hat{f}(k) e^{ikx} \sum_{m=-\infty}^{\infty} \hat{g}(m) e^{imx} = \sum_{n=-\infty}^{\infty} \underbrace{\sum_{k+m=n} \hat{f}(k)\hat{g}(m)}\, e^{inx}, \tag{22.35}$$

so if (22.34) holds it must be that the underbraced factor is the n-th Fourier coefficient of fg. We rewrite this factor as

$$\sum_{k+m=n} \hat{f}(k)\hat{g}(m) = \sum_{k=-\infty}^{\infty} \hat{f}(k)\hat{g}(n-k), \tag{22.36}$$

a *discrete convolution*. For its partial sums we have that

$$\begin{aligned}
(2\pi)^2 \sum_{k=-N}^{N} \hat{f}(k)\hat{g}(n-k) &= \sum_{k=-N}^{N} \int_{-\pi}^{\pi} f(x) e^{-ikx}\, dx \int_{-\pi}^{\pi} g(y) e^{-i(n-k)y}\, dy \\
&= \sum_{k=-N}^{N} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f(x) e^{-ik(x-y)}\, dx\, g(y)\, e^{-iny}\, dy \\
&= \int_{-\pi}^{\pi} \underbrace{\int_{-\pi}^{\pi} f(x+y) \sum_{k=-N}^{N} e^{-ikx}\, dx}_{\text{in view of (22.8) this is } 2\pi\, S_N f(y)} g(y)\, e^{-iny}\, dy,
\end{aligned} \tag{22.37}$$

so the conclusion should be that

$$\sum_{k=-N}^{N} \hat{f}(k)\hat{g}(n-k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_N f(y) g(y)\, e^{-iny}\, dy \tag{22.38}$$

$$\to \frac{1}{2\pi} \int_{-\pi}^{\pi} f(y) g(y)\, e^{-iny}\, dy$$

as $N \to \infty$. This only requires

$$\int_{-\pi}^{\pi} (S_N f(y) - f(y)) g(y)\, e^{-iny}\, dy \to 0 \quad \text{as} \quad N \to \infty,$$

which is the case if $|S_N f - f|_1 \to 0$, which in turn is a consequence of $|S_N f - f|_2 \leq |\sigma_N f - f|_2 \to 0$. We have proved the following theorem.

Theorem 22.25. *Let $f, g \in C_{2\pi}$. Then the complex Fourier coefficients of fg are given by*

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} f(y) g(y)\, e^{-iny}\, dy = \sum_{k=-\infty}^{\infty} \hat{f}(k)\hat{g}(n-k). \tag{22.39}$$

Now let $f, g \in C_b \cap L^1$. Then for the Fourier transform of the convolution

$$f * g(x) = \int_{-\infty}^{\infty} f(x-y) g(y)\, dy = \int_{-\infty}^{\infty} f(y) g(x-y)\, dy \tag{22.40}$$

we need to examine

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x-y) g(y)\, dy\, e^{-i\xi x}\, dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x-y) g(y)\, e^{-i\xi x}\, dy\, dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x-y) g(y)\, e^{-i\xi(x-y)}\, e^{-i\xi y}\, dx\, dy =$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x) g(y)\, e^{-i\xi x}\, e^{-i\xi y}\, dx\, dy =$$

$$\int_{-\infty}^{\infty} f(x)\, e^{-i\xi x}\, dx \int_{-\infty}^{\infty} g(y)\, e^{-i\xi y}\, dy = 2\pi \widehat{f}(\xi)\widehat{g}(\xi).$$

We will need some estimates for convolutions that follow from estimates for

$$F(x) = Kf(x) = \int_{-\infty}^{\infty} K(x,y) f(y)\, dy, \tag{22.41}$$

in which $K(x,y)$ is continuous with

$$\int_{-\infty}^{\infty} |K(x,y)|\,dx \le C \quad \text{and} \quad \int_{-\infty}^{\infty} |K(x,y)|\,dy \le C$$

for all $x, y \in \mathbb{R}$ and some fixed $C > 0$. For $f \in C_b \cap L^1$ we have

$$|F(x)| \le \int_{-\infty}^{\infty} |K(x,y)|\,|f(y)|\,dy \le \int_{-\infty}^{\infty} |K(x,y)|\,dy\,|f|_\infty \le C\,|f|_\infty,$$

and also

$$|F|_1 = \int_{-\infty}^{\infty} |\int_{-\infty}^{\infty} K(x,y)f(y)\,dy|\,dx \le \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |K(x,y)f(y)|\,dy\,dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |K(x,y)f(y)|\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |K(x,y)|\,dx\,|f(y)|\,dy$$

$$\le C\int_{-\infty}^{\infty} |f(y)|\,dy = C|f|_1.$$

You should check that in fact $F \in C_b \cap L^1$. These estimates can be applied to (22.40) with $K(x,y) = f(x-y)$ or $K(x,y) = g(x-y)$, all this proves the following theorem about $f * g$.

Theorem 22.26. *Let $f, g \in C \cap L^1$. Then the convolution $f * g$, defined by*

$$(f*g)(x) = \int_{-\infty}^{\infty} f(x-y)g(y)\,dy = \int_{-\infty}^{\infty} f(y)g(x-y)\,dy, \tag{22.42}$$

is in $C \cap L^1$ as well, and its Fourier transform is given by

$$\widehat{f*g}(\xi) = \sqrt{2\pi}\,\widehat{f}(\xi)\widehat{g}(\xi). \tag{22.43}$$

Since both $\widehat{f}$ and $\widehat{g}$ are in C_0 also their product is. If g is bounded then $f * g$ is bounded and in L^1, $\widehat{f*g}$ is integrable and

$$|\widehat{f*g}|_1 \le \sqrt{2\pi}|\widehat{f}\widehat{g}|_2 = \sqrt{2\pi}|\widehat{f}g|_2 \le \sqrt{2\pi}|\widehat{f}|_1\,|g|_\infty,$$

and finally Remark 22.21 applies to give

$$f*g(x) = \int_{-\infty}^{\infty} \widehat{f}(\xi)\widehat{g}(\xi)\,e^{ix\xi}\,d\xi.$$

Next we consider the Fourier transform of the product fg. For $f, g \in X$ as in (22.27) we have

$$\begin{aligned} 2\pi f(x)g(x) &= \int_{-\infty}^{\infty} \widehat{f}(k)e^{ikx}\,dk \int_{-\infty}^{\infty} \widehat{g}(m)e^{imx}\,dm \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \widehat{f}(k)\widehat{g}(m)\,e^{i(k+m)x}\,dm\,dk \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \widehat{f}(k)\widehat{g}(m-k)\,e^{i(k+m-k)x}\,dm\,dk \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \widehat{f}(k)\widehat{g}(m-k)\,dk\,e^{imx}\,dm \\ &= \int_{-\infty}^{\infty} (\widehat{f} * \widehat{g})(m)\,e^{imx}\,dm, \end{aligned}$$

so

$$f(x)g(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}}(\widehat{f} * \widehat{g})(m)}\ dm,$$

Let us check when the underbraced term is indeed

$$\widehat{fg}(m) = \frac{1}{\sqrt{2\pi}}(\widehat{f} * \widehat{g})(m). \tag{22.44}$$

So let $f, g \in C \cap L^1$.

A direct calculation in the spirit of what followed after (22.36) gives that

$$\begin{aligned} 2\pi \int_{-R}^{R} \widehat{f}(k)\widehat{g}(m-k)\,dk &= \int_{-R}^{R}\int_{-\infty}^{\infty} f(x)e^{-ikx}\,dx \int_{-\infty}^{\infty} g(y)e^{-i(m-k)y}\,dy\,dk \\ &= \int_{-R}^{R}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)e^{-ikx}g(y)e^{-i(m-k)y}\,dx\,dy\,dk \\ &= \int_{-R}^{R}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x+y)e^{-ikx}g(y)e^{-imy}\,dx\,dy\,dk \\ &= \int_{-\infty}^{\infty}\int_{-R}^{R}\int_{-\infty}^{\infty} f(x+y)e^{-ikx}\,dx\,dk\,g(y)e^{-imy}\,dy, \\ &= \int_{-\infty}^{\infty} \underbrace{\int_{-\infty}^{\infty} f(x+y)\int_{-R}^{R} e^{-ikx}\,dk\,dx}\ g(y)e^{-imy}\,dy, \end{aligned}$$

which you should compare to (22.37), in which the underbraced factor equals

$$\int_{-\pi}^{\pi} f(x+y) \sum_{k=-N}^{N} e^{ikx}\, dx = \int_{-\pi}^{\pi} f(x+y) \frac{\sin(N+\frac{1}{2})x}{\sin\frac{x}{2}}\, dx \quad \text{with} \quad f \in C_{2\pi}.$$

Here the underbraced factor equals

$$\int_{-\infty}^{\infty} \int_{-R}^{R} f(x+y) e^{ikx}\, dk\, dx = \int_{-\infty}^{\infty} f(x+y) \frac{\sin Rx}{\frac{x}{2}}\, dx \quad \text{with} \quad f \in C \cap L^1.$$

It's a nice exercise to generalise the analysis of $S_N f$, which used

$$\frac{1}{N+1} \sum_{n=0}^{N} D_n(x)$$

to the analysis of

$$\int_{-\infty}^{\infty} f(x+y) \frac{\sin Rx}{\frac{x}{2}}\, dx,$$

using also

$$\frac{1}{R} \int_{0}^{R} \frac{\sin rx}{\frac{x}{2}}\, dr,$$

and arrive at the conclusion that (22.44) holds for $f, g \in C \cap L^1$.

22.8 Remark on Fourier transforms of distributions

This section could build on an earlier section not written yet about the distributional definition of generalised functions such as the δ-function. With Proposition 22.19 we showed that $f, \phi \in C \cap L^1$ implies $\widehat{f}, \widehat{\phi} \in C_0$ and

$$\langle \widehat{f}, \phi \rangle = \int_{-\infty}^{\infty} \widehat{f}(\xi)\phi(\xi)\, d\xi = \int_{-\infty}^{\infty} f(x)\widehat{\phi}(x)\, dx = \langle f, \widehat{\phi} \rangle, \tag{22.45}$$

in which we use the notation

$$\langle f, g \rangle = \int_{-\infty}^{\infty} fg. \tag{22.46}$$

So certainly (22.45) holds for all $\phi \in \mathcal{S}$ if $f \in C \cap L^1$. We now take (22.45) as the defining property of $\widehat{f} : \mathcal{S} \to \mathbb{C}$ if $f : \mathcal{S} \to \mathbb{C}$ is a linear functional

$$\phi \to \langle f, \phi \rangle$$

defined for $\phi \in \mathcal{S}$. This is very similar to the definition of f' which copies

$$\langle f', \phi \rangle = \int_{-\infty}^{\infty} f'(x)\phi(x)\,dx = -\int_{-\infty}^{\infty} f(x)\phi'(x)\,dx = -\langle f, \phi' \rangle$$

for e.g. $f, \phi \in C^1 \cap C_0$ to define the linear functional $f' : C_c^\infty \to \mathbb{C}$ by[24]

$$\langle f', \phi \rangle = -\langle f, \phi' \rangle, \tag{22.47}$$

if f is a linear functional on C_c^∞.

An example of such a linear functional is δ_s defined by

$$\langle \delta_s, \phi \rangle = \phi(s).$$

We note that δ_s is often written as (not the) function

$$\delta_s(x) = \delta(x - s) = \delta(s - x),$$

with the convolution rule that

$$\int f(s)\delta(x - s)\,ds = f(x),$$

a rule we may like to make precise as the outcome of

$$\int f(s)\delta_s\,ds \quad \text{being equal to} \quad f \quad \text{when acting on} \quad \phi.$$

This requires the integral to be defined in some dual space of the space for ϕ, via its action on the space for ϕ as

$$\langle \int f(s)\delta_s\,ds, \phi \rangle = \int \langle f(s)\delta_s, \phi \rangle\,ds = \int f(s)\phi(s)\,ds = \langle f, \phi \rangle.$$

If so then we have

$$\int f(s)\delta_s\,ds = f, \quad \text{informally written in turn as} \quad \int f(s)\delta(x-s)\,ds = f(x).$$

[24]For (22.45) the assumption on f is stronger, it needs to be a linear functional on $\mathcal{S}$.

22.9 Examples, details and inner product approach

Form old notes again, to be adapted. The series in (22.1) is called a Fourier sine series. If we change the minus signs into plus signs in the definition of f_7 we get the function defined by

$$h_7(x) = \sin x + \frac{\sin 2x}{2} + \frac{\sin 3x}{3} + \frac{\sin 4x}{4} + \frac{\sin 5x}{5} + \frac{\sin 6x}{6} + \frac{\sin 7x}{7},$$

which is close to $h(x) = \frac{\pi - x}{2}$ for $(0, 2\pi)$.

The function

$$g_7(x) = \cos x - \frac{\cos 2x}{4} + \frac{\cos 3x}{9} - \frac{\cos 4x}{16} + \frac{\cos 5x}{25} - \frac{\cos 6x}{36} + \frac{\cos 7x}{49}$$

is close to

$$g(x) = \frac{\pi^2}{12} - \frac{x^2}{4}$$

on the interval $(-\pi, \pi)$. Apparently

$$\frac{x^2}{4} = \frac{\pi^2}{12} + \sum_{k=1}^{\infty} (-1)^k \frac{\cos kx}{k^2}.$$

The right hand side is called a Fourier cosine series. Substituting $x = 0$ we find

$$\frac{\pi^2}{12} = 1 - \frac{1}{4} + \frac{1}{9} - \frac{1}{16} + \frac{1}{25} - \cdots.$$

Exercise 22.27. Let f be an integrable[25] 2π-periodic function. Show that

$$\sigma_N f(x) - f(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F_N(y) \left(f(x - y) - f(x) \right) \, dy$$

Show that

$$\sigma_N f(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F_N(y) f(x - y) \, dy.$$

Exercise 22.28. Derive the equality in (22.10) by writing

$$\sin \frac{x}{2} + \cdots + \sin \frac{(N+1)x}{2}$$

[25] Riemann or Lebesgue integral.

as imaginairy part of a finite geometric sum. Verify that

$$\int_{-\pi}^{\pi} F_{N}(x)\,dx = 2\pi,$$

and that $F_{N}(x) \to 0$ als $N \to \infty$, except in integer multiples of 2π. To be precise

$$0 < \delta \leq x \leq \pi \implies 0 \leq F_{N}(x) \leq \frac{1}{N+1}\frac{1}{\sin^2 \frac{\delta}{2}}.$$

For fixed δ this upper bound is small when N is large. Note that $F_{N}(x)$ is even and 2π-periodic. Make plots of F_{N} for some values of N.

Exercise 22.29. Let f be 2π-periodic and continuous. Then is f uniformly continuous and bounded. Why? Prove that $\sigma_{N} f$ converges uniformly to f as $N \to \infty$.

Exercise 22.30. Let f be 2π-periodic, bounded and piecewise continuous, with the property that in every point the limits from the left and from the right exist. Show that for every x the sequence $\sigma_{N} f(x)$ converges as $N \to \infty$. What's the limit? Hint: split de integral in 4 parts.

Exercise 22.31. Let $f : [-\pi, \pi] \to \mathbb{R}$ be twice continuously differentiable with $f(\pm\pi) = f'(\pm\pi) = f''(\pm\pi) = 0$. Show that f is the sum of its (uniformly convergent) Fourier series in every $x \in [-\pi, \pi]$. Hint: use partial integration to show the Fourier coefficients a_n en b_n make for summable series.

Exercise 22.29 shows that in the space of continuous functions 2π-periodic functions equipped with the maximum norm

$$|f|_{\max} = \max_{x \in \mathbb{R}} |f(x)|$$

the Cesàro sums of f converge to f: $|\sigma_{N} f - f|_{\max} \to 0$ als $N \to \infty$.

Fourier series can be traced back to Daniel Bernouilli, who used them to solve the wave equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}.$$

Fourier was perhaps the first to give integral expressions for the coefficients, when he tried to solve the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}.$$

Nowadays we see the functions

$$\frac{1}{2}, \cos x, \sin x, \cos 2x, \sin 2x, \ldots,$$

and

$$\ldots, e^{-3ix}, e^{-2ix}, e^{-ix}, e^{0ix} = 1, e^{ix}, e^{i2x}, e^{3ix}, \ldots$$

as orthonormal bases in a (Hilbert) space of functions, and the Fourier coefficients as coordinates with respect to these bases. For a large class of functions $f : (-\pi, \pi) \to \mathbb{R}$ the Fourier coefficienten a_n, b_n and c_n as coordinates of f are thus well-defined.

Exercise 22.32. Compute

$$\int_{-\pi}^{\pi} \cos nx \cos mx \, dx \quad \text{and} \quad \int_{-\pi}^{\pi} \cos nx \sin mx \, dx$$

for integer m and n. Hint: if $f''(x) + \lambda f(x) = 0$ and $g''(x) + \mu g(x) = 0$ then

$$\int (gf'' - fg'')$$

evaluates as ..... using integration by parts.

Exercise 22.33. Use Exercise 22.32 to show that for f defined by

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{N} (a_n \cos nx + b_n \sin nx),$$

it holds that a_n and b_n are given by (22.4) for $n \leq N$.

The following programme is meant to get you aquainted with Fourier series. Use Maple/Mathematica for the plots. The integrals you should do by hand.

Exercise 22.34. Let $f : (0, \pi) \to \mathbb{R}$ be given by $f(x) = 1$ and choose a 2π-periodic even extension $f : \mathbb{R} \to \mathbb{R}$. Determine all Fourier coefficients a_n and b_n.

Exercise 22.35. Let $f : (0, \pi) \to \mathbb{R}$ be given by $f(x) = 1$ and choose a 2π-periodic even extension $f : \mathbb{R} \to \mathbb{R}$.

1. Determine all Fourier coefficients a_n en b_n.
2. Plot f and $S_N f$ (for some values of N) in one graph.
3. Investigate numerically what happens to location and value of the maximum of $S_N f$ as $N \to \infty$.
4. Simplify $S_N f$ in $x = \frac{\pi}{2}$ and compare to $f(\frac{\pi}{2})$. Which sum of which series, assuming $S_N f(\frac{\pi}{2}) \to f(\frac{\pi}{2})$, do you obtain?
5. Same question for $x = \frac{\pi}{4}$.

Exercise 22.36. Let $f : (0, \pi) \to \mathbb{R}$ be given by $f(x) = \sin x$. Choose an even 2π-periodic extension $f : \mathbb{R} \to \mathbb{R}$.

1. Determine all Fourier coefficients a_n and b_n.
2. Plot f and $S_N f$ (voor een aantal waarden van N) in een grafiek.
3. Simplify $S_N f$ in $x = 0$. Compare with $f(0)$. Which sum of which series, assuming $S_N f(0) \to f(0)$, do you obtain?
4. Idem for $x = \frac{\pi}{2}$.
5. Idem for $x = \frac{\pi}{4}$.

Exercise 22.37. Let $f : (0, \pi) \to \mathbb{R}$ be given by $f(x) = \cos x$ and choose an odd 2π-periodic extension $f : \mathbb{R} \to \mathbb{R}$.

1. Determine all Fourier coefficients a_n and b_n, and plot f and $S_N f$ (for some values of N) in one graph.
2. Compare the behaviour near $x = 0$ for N large with that in Exercise 22.35.
3. Now take the odd 2π-periodic extension of $f(x) = 1 - \cos x$ (the difference of the function in Exercise 22.35 and the function here). Investigate numerically the behaviour of $S_N f$ near $x = 0$ for large N.

Exercise 22.38. Let $f : (0, \pi) \to \mathbb{R}$ be given by $f(x) = \pi - x$ and choose an odd 2π-periodic extension $f : \mathbb{R} \to \mathbb{R}$.

1. Determine the Fourier coefficients a_n and b_n, and plot f and $S_N f$ (for some values of N) in one graph.
2. Differentiate $S_N f(x)$ with respect to x and call the derivative $d_N(x)$. Are there values of x for which $d_N(x)$ converges as $N \to \infty$?

Exercise 22.39. Let $f : (0, \pi) \to \mathbb{R}$ be given by $f(x) = x(\pi - x)$ and choose an odd 2π-periodic extension $f : \mathbb{R} \to \mathbb{R}$.

1. Determine the Fourier coefficients a_n and b_n, and plot f and $S_N f$ (for some values of N) in one graph.
2. Simplify $S_N f$ in $x = \frac{\pi}{2}$. Compare with $f(\frac{\pi}{2})$. Which sum of which series do you get if $S_N f(x) \to f(x)$?
3. Differentiate $S_N f(x)$ with respect to x and call the derivative $g_N(x)$. Show that $g_N(x)$ op $\mathbb{R}$ converges uniformly on $\mathbb{R}$ to a limit function.
4. Determine the limit function numerically.
5. Compare $g_N(0)$ with its limit. Which sum of which series do you obtain?

Convergence of Fourier series in the mean was not really discussed so far. Let a_n and b_n be the Fourier coefficients of a 2π-periodic integrable real valued function, that is

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)\,dx \quad \text{and} \quad b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)\,dx.$$

To answer questions about convergence, i.e. about whether or not

$$\sum_{n=-N}^{N} c_n e^{inx} = \underbrace{\frac{a_0}{2} + \sum_{n=1}^{N}(a_n \cos nx + b_n \sin nx)}_{S_N(f(x)} \to f(x)$$

as $N \to \infty$, convolutions are of great importance.

But what about $S_N f$ if we don't have decay rates for the coefficients? We consider convergence in the 2-norm. The integral

$$f \cdot g = \int_{-\pi}^{\pi} f(x)g(x)\,dx \tag{22.48}$$

is called the inner product of the functions f and g. If $f \cdot g = 0$ we say that f and g are *perpendicular*. The 2-norm of f is defined by

$$|f|_2 = \sqrt{f \cdot f}, \tag{22.49}$$

the length of f considered as a vector. Pythagoras could have told us told us that

$$f \cdot g = 0 \quad \Rightarrow \quad |f+g|_2^2 = |f|_2^2 + |g|_2^2. \tag{22.50}$$

Below we write

$$S_N g(x) = \frac{c_0}{2} + \sum_{k=1}^{N} (c_k \cos kx + d_k \sin kx). \tag{22.51}$$

Now let f have real Fourier coefficients a_k and b_k, and let c_k and d_k be the real Fourier coefficients of g. Do the following exercises.

Exercise 22.40. The Cauchy-Schwartz inequality says that $|f \cdot g| \leq |f|_2 |g|_2$.

1. Prove this inequality for functions f and g with $|f|_2 = |g|_2 = 1$ by evaluating $0 \leq \int_{-\pi}^{\pi} (f(x) - g(x))^2 \, dx = \ldots$.
2. Prove the Cauchy-Schwartz inequality. Hint: apply 1 to $f(x)/|f|_2$ and $g(x)/|g|_2$.
3. Prove that
$$|f+g|_2 \leq |f|_2 + |g|_2. \tag{22.52}$$

Exercise 22.41. Show that

$$|S_N f|_2^2 = \pi \left(\frac{1}{2} a_0^2 + \sum_{k=1}^{N} (a_k^2 + b_k^2) \right)$$

Exercise 22.42. Show that

$$S_N f \cdot S_N g = \pi \left(\frac{1}{2} a_0 c_0 + \sum_{k=1}^{N} (a_k c_k + b_k d_k) \right)$$

Exercise 22.43. Define $R_N f = f - S_N f$ and, with

$$\sigma_N f = \frac{1}{N+1}(S_0 f + S_1 f + \cdots S_N f),$$

let $\rho_N f = f - \sigma_N f$.

1. Show that $R_N f \cdot S_N f = 0$.
2. Show that $R_N f \cdot \sigma_N f = 0$.
3. Show that
$$|S_N f|_2^2 + |R_N f|_2^2 = |f|_2^2,$$
whence $|S_N f|_2 \leq |f|_2$ and (Bessel's inequality)
$$\frac{1}{2}a_0^2 + \sum_{k=1}^{N}(a_k^2 + b_k^2) \leq \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)^2\,dx. \tag{22.53}$$
4. Show that
$$|R_N f|_2^2 + |\sigma_N f - S_N f|_2^2 = |\rho_N f|_2^2.$$
5. In Exercise 22.29 we showed for f continuous and 2π-periodic that $\sigma_N f \to f$ uniformly on $\mathbb{R}$ as $N \to \infty$. Prove that then also $|R_N f|_2 \to 0$, so that (Parceval equality)
$$\frac{1}{2}a_0^2 + \sum_{k=1}^{\infty}(a_k^2 + b_k^2) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)^2\,dx. \tag{22.54}$$
Hint: use part 4.
6. Show that
$$\begin{aligned} f \cdot g &= (S_N f + R_N f) \cdot (S_N g + R_N g) \\ &= S_N f \cdot S_N g + R_N f \cdot R_N g. \end{aligned}$$
7. For f and g continuous and 2π-periodic show that
$$\frac{1}{2}a_0 c_0 + \sum_{k=1}^{\infty}(a_k c_k + b_k d_k) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)g(x)\,dx = \frac{1}{\pi}\, f \cdot g. \tag{22.55}$$
Hint: use part 6, Exercise 22.42 and apply the Cauchy-Schwartz inequality to $R_N f \cdot R_N g$.

Exercise 22.44. We want to show that Parceval's equality (22.54) holds for 2π-periodic peiceswise continuous functions. Let $f : \mathbb{R} \to \mathbb{R}$ be such a function. Show that there exists a sequence of 2π-periodic continuous functions $f_k : \mathbb{R} \to \mathbb{R}$ with

$$|f_k - f|_2^2 = \int_{-\pi}^{\pi} (f_k(x) - f(x))^2 \, dx \to 0$$

as $k \to \infty$. Hint: if f is discontinuous in x_0, replace $f(x)$ on the interval $(x_0 - \frac{1}{k}, x_0 + \frac{1}{k})$ by a linear function, such that the new function f_k is continuous and linear on $(x_0 - \frac{1}{k}, x_0 + \frac{1}{k})$.

Exercise 22.45. Prove (22.54) for f. Hint: the desired equality is equivalent with $|R_N f|_2 \to 0$. Write

$$R_N f = f - f_k + f_k - S_N f_k + S_N f_k - S_N f = (f - f_k) + R_N f_k + S_N(f_k - f)$$

and use (22.52) and Exercise 3 for $S_N(f_k - f)$ to make $|R_N f|_2$ small. Let $\varepsilon > 0$, choose k large as needed, etc.

TO DO. The abstract construction[26] of a Hilbert space H from $C(\mathbb{R}_{2\pi})$ is via Cauchy sequences $f_1, f_2, \ldots$, using the 2-norm, i.e. sequences with

$$|f_n - f_m|_2 \to 0$$

as $m, n \to \infty$. We think of such sequences as approximating some f in the space H under construction. This is just like decimal or binary expansions approximating real numbers, by which different expansions can define the same real number, which we can picture on a number line if we like. Of course the abstract construction by itself is completely independent of the pictures.

The standard way to visualise a function is as the graph of that function, in case of $f : \mathbb{R} \to \mathbb{R}$ a subset G of

$$\mathbb{R}^2 = \{(x, y) : \, x, y \in \mathbb{R}\}$$

with the property that

$$\forall_{x \in \mathbb{R}} \, \exists!_{y \in \mathbb{R}} : \, (x, y) \in G.$$

Here $\exists!$ means *there exists precisely one* with (in this case) the property that $y \in \mathbb{R}$ and $(x, y) \in G$. This unique y may then be denoted by $f(x)$. The formal definition of a graph in $\mathbb{R}^2$ is de facto equivalent with the definition of a function from $\mathbb{R}$ to $\mathbb{R}$.

[26] Choices to be made in relation to Section 33.4.

Fourier series in Olver's book

More old notes to be adapted. Olver's Section 3.1 derives

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx) e^{-k^2 t} \tag{22.56}$$

as the general form for 2π-periodic (in x) solutions $u(t, x)$ of the one-dimensional heat equation with unit diffusion coeffient:

$$u_t = u_{xx}$$

Important is the eigenvalue perspective, culminating in the table on page 68 and, in the 2π-periodic context, (3.27). Note that (3.18) is a solution of *separated variables*, the topic of Chapter 4.

The initial data of the solution in (3.27) is $f(x)$ given by (3.28). Section 3.2 explains how the left hand side $f(x)$ and the Fourier series on the right hand side correspond to one another. The starting point is (3.34) and (3.35) with

$$f(x) \sim \frac{a_0}{2} + \sum \dots$$

This notation is introduced because a priori we do not know that or in what sense

$$f(x) = \frac{a_0}{2} + \sum \dots,$$

except when f is a trigonometric polynomial, see Exercise 3.2.4.

Note that we interested in both directions. If we compute solutions $u(t, x)$ we find t-dependent coefficients of the solution and ask how well the solution is defined and how smooth it is. If we fit the general solution $u(t, x)$ to initial data f we do this by fitting the coefficients to the coefficients computed from f.

There are two equivalent forms of the Fourier series, real and complex. The real cos/sin form of the Fourier series allows a nice and useful distinction between even and odd functions $f(x)$: Prop. 3.14. The complex form (3.64) allows for smoother formula's and proofs and a much smoother transition to the Fourier integral transform in Chapter 7. Both forms are intrinsically related to a suitable inner product for functions: (3.30)/(3.61).

One can prove that C^1 2π-periodic functions f are sums of *uniformly convergent* Fourier series computed through (3.35). A stronger localised statement is given in Thm 3.30. The uniform convergence result remains valid if $f'(x)$ has finitely many jump discontinuities (corners in the graph of f) on its interval of periodicity. However, if f itself has jump discontinuities the convergence *cannot be uniform*. It still holds pointwise, but only under the

strong assumption that f' is piecewise continuous: Theorem 3.8, proved at the very end of Section 3.5. A slight variant of that proof, using the mean value theorem, shows the convergence is indeed uniform if f is continuous and f' is (piecewise) continuous.

Remark: jump discontinuities may be removed by subtracting suitably scaled and shifted sawtooth functions. The resulting continuous piecewise C^1-function has a uniformly convergent Fourier series. For sawtooths we can examine the behaviour of their Fourier series by direct calculations which are somewhat reminiscent of the calculations in the convergence proof, and which also clarify the Gibbs overshoot phenomenon illustrated in Figure 3.7. The nicest sawtooth to illustrate what's going on is the odd 2π-periodic extension of

$$\pi - x = \frac{2\sin x}{1} + \frac{2\sin 2x}{2} + \frac{2\sin 3x}{3} + \frac{2\sin 4x}{4} + \frac{2\sin 5x}{5} + \dots \quad (0 < x < \pi).$$

If you differentiate the right hand side you get an expression which diverges! Its truncation after n terms is easily related to the outcome of (3.128) in the convergence proof.

Exercise. Integrate the resulting expression to conclude that

$$\frac{2\sin x}{1} + \dots + \frac{2\sin nx}{n} = \int_0^x D_n(s)ds - x, \quad D_n(s) = \frac{\sin(n+\frac{1}{2})s}{\sin\frac{s}{2}}$$

This D_n is called the Dirichlet kernel. Examine the integral using the n-dependent scaling $y = (n+\frac{1}{2})x$ (and likewise for s). Identify π as a limit of the integral for $n \to \infty$ when $0 < x \leq \pi$ and

$$2\int_0^\pi \frac{\sin t}{t}dt > \pi = 2\int_0^\infty \frac{\sin t}{t}dt$$

as the limit of the largest maximum of $\int_0^x D_n(s)ds$, which occurs in $\frac{\pi}{n+\frac{1}{2}}$.

Basis general facts about uniform convergence are Thm 3.26 and 3.27 (which implies Thm 3.29), Prop. 3.28, and the discussion about interchanging sums and integrals just above Prop. 3.28. I will assume these facts are known. Not mentioned here in the book is that uniform convergence, illustrated in Fig. 3.11, also comes with a norm, called the maximum or supremum norm,

$$||f||_m = \sup_x |f(x)|,$$

which is certainly defined if f is 2π-periodic and piecewise continuous in the sense of Def. 3.6. Note that the inner product norm is controlled by the

maximum norm: if the truncation error goes to zero in the maximum norm, it certainly goes to zero in the inner product norm (but not the other way around!). It is in general not true that the partial sums s_n of the Fourier series of a 2π-periodic continuous function f satisfy

$$||s_n - f||_m \to 0,$$

as it is not even clear that $s_n(x) \to f(x)$ pointwise.

Section 3.5 should be studied in mathematical detail. It discusses not only uniform convergence, and pointwise convergence, but also convergence in the mean. The latter is by definition equivalent to convergence in the inner product norm (3.102) and allows one to think of and work with 2π-periodic functions (for which the inner product norm is defined) as column vectors, and the Fourier modes as unit base vectors. Think of straight angles and the Pythagorean theorem here, we will see expressions like

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots = \frac{\pi^2}{6}.$$

The relevant formula's are given in the complex notation (3.124), the real version is discussed in 3.5.38. The proof relies on the essential observation made in Thm 3.36 that in terms of the inner product norm the n-th partial sum s_n of the Fourier series is the best trigonometric approximation of degree n to f, and that the error goes to zero as $n \to \infty$, which amounts to convergence in the mean.

Convergence in the mean is in fact equivalent to the first Plancherel (Pythagoras) formula in (3.124). On page 115 the author avoids the consideration of different norms with a direct proof of the Plancherel formula for the sum function of a uniformly convergent Fourier series. For general square integrable functions the Plancherel formula then follows by approximation arguments (the details are not given in the book). Thus, although for continuous functions even pointwise convergence may fail, as complicated examples show, one can be content that convergence in the mean always holds.

Section 3.2 emphasises that Fourier series are *not* like power series. Therefore Section 3.3 on what is allowed in differentiation and integration of Fourier series should be read with care. The statements are only about the coefficients, not about convergence of the Fourier series, and they follow from integration by parts. Indeed, the formula's for the Fourier coefficients can be integrated by parts whenever $f'(x)$ exists, giving similar formula's with prefactors $\frac{1}{k}$ and thus better (and faster) convergence of the Fourier series than expected from the defining formula's. The smoother f, the more times

we can integrate by parts and gain a prefactor $\frac{1}{k}$, and thus the better and faster the convergence of the Fourier series. However, for uniform convergence we need (3.99) with $\alpha > 1$, which requires more than one derivative for f, so this trick cannot compete with the direct convergence proof which only needs the first order derivative.

Nevertheless, the larger we can choose n in (3.100), the smoother the sum function of the Fourier series: Thm 3.31. Solving PDE's, we compute Fourier series solutions

$$u(t,x) \sim \sum \dots$$

Solutions of the heat equation $u_t = u_{xx}$ are easily seen to be very well-behaved in this respect as soon as $t > 0$, even with ugly initial data, because of the exponentials in (3.27). However, interesting (fractal) issues with respect to lack of smoothness will appear in Chapter 7 when we solve $u_t = u_{xxx}$ with e.g. piecewise constant initial data.

Read also Section 3.4 yourself. It discusses simple changes of scale needed to deal with l-periodic functions $f(x)$, $l > 0$. The limit $l \to \infty$ leads to the Fourier integral transform in Chapter 7.

Exercise. Assume that f and f' are piecewise continuous, so that f' is bounded by a fixed constant M. Show that the function $g(y)$ defined immediately after (3.131) is bounded in terms of M. Then split the integral of $g(y)\sin(n+\frac{1}{2})y$ in $\int_0^\delta$ and $\int_\delta^\pi$. The first is bounded in terms of δ and M. Integrate the second one by parts and estimate in terms of δ, M and n which appears in the denominator. Conclude the pointwise convergence proof without using the Riemann-Lebesgue lemma and show that the convergence is uniform if f is continuous and f' is (piecewise) continuous.

23 Transformation theorem

This chapter is only a sketch of what we want. Let's see what that is from an example. For $R \subset \mathbb{R}^2$ and continuous $f : R \to \mathbb{R}$ we have that

$$\iint_R f(x,y)\,dxdy = \iint_Q f(x(r,\theta),y(r,\theta))\,(\frac{\partial x}{\partial r}\frac{\partial y}{\partial \theta} - \frac{\partial y}{\partial r}\frac{\partial x}{\partial \theta})drd\theta,$$

if

$$Q \xrightarrow{(r,\theta)\to(x,y)} R$$

is reasonably nice. We explore how we can prove such statements.

If

$$(x,y) \xrightarrow{\Phi} (u,v)$$

is a bijection between $R \subset \mathbb{R}^2_{x,y}$ and $A \subset \mathbb{R}^2_{u,v}$ we would like to have that the integral

$$\iint_A g(u,v)\,dudv$$

relates to an integral with $g(u(x,y),v(x,y))$ and $dxdy$ over R, perhaps with the convention that $dudv = -dvdu$ en $dxdy = -dydx$. Let's assume that R is a rectangle, e.g. $R = [0,1] \times [0,1]$.

Have a look at (15.2) and read

$$F(x,y,u,v) = \begin{pmatrix} \Phi_1(x,y) - u \\ \Phi_2(x,y) - v \end{pmatrix} \quad \text{instead of} \quad F(x,y) = g(y) - x.$$

Unpacking[1] the theorem we obtain an inverse function theorem which says that if the Jacobi matrix in in (x_0,y_0), i.e.

$$J(x,y) = \begin{pmatrix} \frac{\partial \Phi_1}{\partial x} & \frac{\partial \Phi_1}{\partial y} \\ \frac{\partial \Phi_2}{\partial x} & \frac{\partial \Phi_2}{\partial y} \end{pmatrix}$$

is invertible, in some neightbourhood of $(u_0,v_0) = (\Phi_1(x_0,y_0),\Phi_2(x_0,y_0))$ the inverse function

$$(u,v) \xrightarrow{\Phi^{-1}} (x,y)$$

exists and continuously differentiable. The Jacobi matrix of the inverse map is the inverse of the Jacobi matrix of Φ.

For a transformation theorem we therefore assume that the Jacobi matrix $J(x,y)$ is invertible in every point of R. This makes A a region in $\mathbb{R}^2_{u,v}$ with four boundary parts parameterised by

$$x \to \Phi(x,0), \quad y \to \Phi(1,y), \quad x \to \Phi(x,1), \quad y \to \Phi(0,y).$$

[1] Chapter 18 explained how to unpack.

Partitions

$$(P) \quad 0 = x_0 \leq x_1 \leq \cdots \leq x_N = 1 \quad \text{met} \quad N \in \mathbb{N},$$

$$(Q) \quad 0 = y_0 \leq y_1 \leq \cdots \leq y_M = 1 \quad \text{met} \quad M \in \mathbb{N},$$

then give $(M+1)(N+1)$ parameterisations

$$x \to \Phi(x, y_j) \quad \text{en} \quad y \to \Phi(x_i, y) \quad (i = 0, \ldots, \mathrm{M},\, j = 0 \ldots, N),$$

which form a grid of deformed rectangles S_{ij} in A.

A proper definition of Riemann integrability of $g : A \to \mathbb{R}$ should[2] give that with

$$M_{ij} = \sup_{S_{ij}} g \quad \text{and} \quad m_{ij} = \inf_{S_{ij}} g$$

it follows that

$$\sum_{ij} m_{ij}|S_{ij}| \leq \iint_A g \leq \sum_{ij} M_{ij}|S_{ij}|,$$

in which S_{ij} is the area of S_{ij}. We then rewrite this as

$$\sum_{ij} m_{ij} \frac{|S_{ij}|}{|R_{ij}|}|R_{ij}| \leq \iint_A g \leq \sum_{ij} M_{ij} \frac{|S_{ij}|}{|R_{ij}|}|R_{ij}|,$$

and note that

$$M_{ij} = \sup_{R_{ij}} f \quad \text{and} \quad m_{ij} = \inf_{R_{ij}} f$$

with $f = g \circ \Phi$.

It remains to make precise[3] that

$$\frac{|S_{ij}|}{|R_{ij}|} \sim |\mathrm{det} J(x_i, y_i)| \tag{23.1}$$

as $M, N \to \infty$ to obtain the Riemann integrability of

$$(x, y) \to f(x, y)|J(x, y)|$$

over R and conclude that

$$\iint_R f\, |\mathrm{det} J| = \iint_A g. \tag{23.2}$$

[2] To do, note that J is constant if Φ is linear.

[3] See e.g. Section 5 of Chapter III in the Advanced Calculus book of Edwards.

24 Differential forms

Have a look at Section 10.3 and then look at the first part of the proof of Theorem 21.6. Dropping the tildes we found that

$$\int_\Omega v_x = \int_a^b v(x, f(x)) f'(x)\, dx \quad \text{and} \quad \int_\Omega v_y = -\int_a^b v(x, f(x))\, dx$$

for a function $v \in C^1(\bar{\Omega})$ vanishing outside a window in which we (locally) describe the boundary as a graph $y = f(x)$. It is tempting to write

$$\int_\Omega v_x = -\int_{\partial\Omega} v dy \quad \text{and} \quad \int_\Omega v_y = \int_{\partial\Omega} v dx, \tag{24.1}$$

in which the right hand sides are evaluated using the parameterisation[1]

$$x = x(t) = t \quad \text{and} \quad y = y(t) = f(t). \tag{24.2}$$

$$\underbrace{x(t)}_{x}, \quad \underbrace{y = f(t)}_{y}, \quad \underbrace{dx = x'(t)\, dt = dt}_{dx} \quad \text{and} \quad \underbrace{dy = y'(t)\, dt = f'(t)\, dt}_{dy}.$$

We have skipped the spaces in front of dx and dy to allow $v = v(x, y) = v(t, f(t))$ to cozy up with dx and dy. This reminds us of notation in and below (10.6). Can we see the right hand sides of (24.1) as

$$\int_{\partial\Omega} \quad \text{acting on the 1-forms} \quad v dy = v(x, y) dy \quad \text{and} \quad v dx = v(x, y) dx?$$

If so, how should we see the (double) integrals on the left hand sides then? Recall that in Theorem 21.1 we read the repeated integral

$$\int_c^d \underbrace{\int_a^b u(x, y)\, dx}_{\text{function of } y}\, dy$$

as

$$\int_c^d \left\{ \int_a^b u(x, y)\, dx \right\} dy$$

and wrote

$$\int_{[a,b]\times[c,d]} u = \int_c^d \underbrace{\int_a^b u(x, y)\, dx}_{\text{function of } y}\, dy = \int_a^b \underbrace{\int_c^d u(x, y)\, dy}_{\text{function of } x}\, dx,$$

[1] We only need a local parameterisation because v was localised by a fading function.

with a little space in front of dx and dy. This is *not yet* a notation with 2-forms $udxdy$ as hinted at under Theorem 10.12.

We shall now agree[2] that

$$\int_c^d \underbrace{\int_a^b u(x,y)\,dx}_{\text{function of } y}\,dy = \underbrace{\int_{[a,b]\times[c,d]} u}_{J \text{ as in Theorem 21.1}} = \underbrace{\int_{[a,b]\times[c,d]}}_{\text{integral acting on}} \underbrace{udxdy}_{\text{2-form}},$$

in which we view

$$\int_{[a,b]\times[c,d]} \quad \text{as acting on the 2-form } udxdy, \quad u = u(x,y).$$

The result of this action is equal to

$$-\int_{[a,b]\times[c,d]} u(x,y)dydx$$

if we adopt[3] the rule $dxdy = -dydx$. Likewise, we can then see

$$\int_\Omega \quad \text{as acting on both} \quad v_x dydx \quad \text{and} \quad v_y dxdy,$$

so that (24.1) can now be read with the forms

$$v_x dydx, \quad vdy, \quad v_y dxdy, \quad vdx$$

having (two different) integrals acting on them. Let's look at the formal algebra first to see which rules will make the d-algebra work for the expressions with x, y, d, dx, dy that we encounter.

24.1 Formal d-algebra

The algebra for such "differential" *forms* developes itself. After $du = u'(x)dx$ for $u = u(x)$ what else could we have but

$$du = u_x dx + u_y dy = \frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy \tag{24.3}$$

for the d of the 0-vorm $u = u(x,y)$? This expression is of the form[4]

$$fdx + gdy = f(x,y)dx + g(x,y)dy,$$

[2]Here we avoid the notation $dx \wedge dy$ used when defining an action of forms on vectors.
[3]Definition 7.8 already led us to consider the sign of dx and also dy in relation to $\int$.
[4]As it happens, a differential form.

a 1-form that in turn must be ready and willing to have d acting upon it. Here's the obvious action:

$$d(fdx+gdy) = d(fdx)+d(gdy) \quad \textit{(a sum rule for d)}$$

$$= \underbrace{dfdx+fddx}_{d(fdx)} + \underbrace{dgdy+gddy}_{d(gdy)} \quad \textit{(twice a Leibniz rule for d)}$$

$$= \underbrace{(f_xdx+f_ydy)}_{\textit{definition of df}}dx + fddx + \underbrace{(g_xdx+g_ydy)}_{\textit{definition of dg}}dy + gddy.$$

$$= f_xdxdx + f_ydydx + g_xdxdy + g_ydydy + fddx + gddy \quad \textit{(bye bye brackets)}$$

$$= f_xdxdx - f_ydxdy + g_xdxdy + g_ydydy + fddx + gddy \quad \textit{(if we use } dydx = -dxdy\textit{)}$$

$$= (g_x - f_y)dxdy + fddx + gddy \quad \textit{(if we use } dxdx = 0 = dydy\textit{)}.$$

The Leibniz rules we used were

$$d(fdx) = (df)dx + f(ddx), \quad \text{which mimics} \quad d(fg) = (df)g + f(dg),$$

and likewise

$$d(gdy) = (dg)dy + g(ddy).$$

Both rules can then be evaluated using the earlier definition of df and dg, and a convenient rule for ddx and ddy. Let's take the simplest choice, we just *introduce*[5] the rule that

$$ddx = ddy = 0.$$

Following old and new rules we then obtain

$$f(x,y)dx + g(x,y)dy \xrightarrow{d} (g_x(x,y) - f_y(x,y))dxdy,$$

as the action of d on a 1-form. If we're fine with this action it follows that

$$u(x,y) \xrightarrow{d} u_xdx + u_ydy \xrightarrow{d} (u_{yx} - u_{xy})dxdy = 0$$

if $u_{xy} = u_{yx}$. We're fine with that. Apparently the rules imply that $d^2 = 0$. Using a notation with differential quotients the rules for d-algebra with two variables are

$$\begin{gathered} f \xrightarrow{d} \tfrac{\partial f}{\partial x}dx + \tfrac{\partial f}{\partial y}dy, \quad fdx + gdy \xrightarrow{d} (\tfrac{\partial g}{\partial x} - \tfrac{\partial f}{\partial y})dxdy, \\ fdxdy \xrightarrow{d} 0, \quad gdxdy \xrightarrow{d} 0, \end{gathered} \tag{24.4}$$

in which $f = f(x,y)$, $g = g(x,y)$. The (zero) action of d on 2-forms is a consequence of the rules if we have only two variables.

[5]Recall we *decided* that $dxdx = 0$ because $dxdy = -dydx$.

Exercise 24.1. Look at the forms in (24.1) and see how they are related by the d-algebra just developed.

We will be looking for a formulation in which the result is

$$\int_\Omega d\omega = \int_{\delta\Omega} d\omega \tag{24.5}$$

for a 1-form ω and a bounded domain Ω with sufficiently nice boundary $\partial\Omega$. This result will generalise to $(n-1)$-forms and $\Omega \subset \mathbb{R}^n$, and is in fact equivalent to the Gauss Divergence Theorem, Remark 21.7 in Section 21.11.

Exercise 24.2. Do the algebra for

$$f \xrightarrow{d} \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz,$$

$$fdx + gdy + hdz \xrightarrow{d} (\frac{\partial h}{\partial y} - \frac{\partial g}{\partial z})dydz + (\frac{\partial f}{\partial z} - \frac{\partial h}{\partial x})dzdx + (\frac{\partial g}{\partial x} - \frac{\partial f}{\partial y})dxdy,$$

$$fdydz + gdzdx + hdxdy \xrightarrow{d} (\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} + \frac{\partial h}{\partial z})dxdydz,$$

$$hdxdydz \xrightarrow{d} 0,$$

with $f = f(x,y,z), g = g(x,y,z), h = h(x,y,z)$. Use the sum and Leibniz rule for d, the anti-symmetry rules $dxdy = -dydx, dxdz = -dzdx, dydz = -dzdy$, then also $dxdx = dydy = dzdz = 0$, and $ddx = ddy = ddz = 0$. Verify $ddf = 0$ and also $dd(fdx + gdy + hdz) = 0$. If dd kills x,y and z, then dd kills all forms. We like d.

Exercise 24.3. Do it again for $F \xrightarrow{d} F'(x)dx$ and $f(x)dx \xrightarrow{d} 0$ with $f(x)$ and $F(x)$.

Remark 24.4. *The notation is consistent with*

$$dx \wedge dy = -dy \wedge dx$$

in Adams' calculus book and his treatment of such objects as acting on (pairs of) vectors[6]*. For now we find it easier not to write wedges between the dx, dy, etc.*

[6]Tangent vectors really, written as xy-dependent linear combinations of $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$.

24.2 Pull backs

If $x \to f(x) = F'(x)$ and $t \to x'(t)$ are continuous, say with $x(0) = a$ and $x(1) = b$, then

$$\int_a^b f(x)\,dx = \int_a^b F'(x)\,dx = \int_a^b dF = F(b) - F(a) = F(x(1)) - F(x(0)) =$$

$$[F(x(t))]_0^1 = \int_0^1 F'(x(t))x'(t)\,dt = \int_0^1 f(x(t))\,x'(t)dt.$$

In

$$dx = x'(t)dt \tag{24.6}$$

we recognise the d-algebra from Section 24.1, and we see that a 1-form $f(x)dx$ in x is *pulled back* by $t \to x(t)$ to a 1-form

$$f(x)dx = f(x(t))x'(t)dt \tag{24.7}$$

in t. Likewise the 1-form

$$f(x,y)dx + g(x,y)dy$$

is pulled back by $t \to x(t)$ and $t \to y(t)$ to a 1-form

$$f(x,y)dx + g(x,y)dy = (f(x(t),y(t))x'(t) + g(x(t),y(t))y'(t))dt \tag{24.8}$$

in t. So $t \to (x(t), y(t))$ leads to

$$f(x,y)dx + g(x,y)dy \xrightarrow{pull\ back} (f(x(t),y(t))x'(t) + g(x(t),y(t))y'(t))dt$$

for a general 1-form, while for the 2-form $dxdy$ we find

$$dxdy \to x'(t)dt\,y'(t)dt = x'(t)y'(t)dtdt = 0,$$

not of much use, but

$$(r,\theta) \to (x(r,\theta), y(r,\theta))$$

gives

$$dxdy \to (\frac{\partial x}{\partial r}dr + \frac{\partial x}{\partial \theta}d\theta)(\frac{\partial y}{\partial r}dr + \frac{\partial y}{\partial \theta}d\theta) = (\frac{\partial x}{\partial r}\frac{\partial y}{\partial \theta} - \frac{\partial y}{\partial r}\frac{\partial x}{\partial \theta})drd\theta, \tag{24.9}$$

with the determinant[7] of the Jacobi matrix.

[7]Plus or minus the area spanned by the two vectors, compare to (23.2) in Chapter 21.9.

In case of $x = r\cos\theta, y = r\sin\theta$ this reads

$$dxdy \to rdrd\theta.$$

Note that (24.8) does not correspond to a coordinate transformation but (24.9) does. Have another look at the derivation of (24.8) and replace t by ϕ in $[0, 2\pi]$. With $x(0) = x(2\pi)$ and $y(0) = y(2\pi)$ this compares to (24.9) with r fixed, and you discover how the pull back algebra works for

$$(\theta, \phi) \to (x(\theta, \phi), y(\theta, \phi), z(\theta, \phi) \tag{24.10}$$

and 2-forms in x, y, z.

Exercise 24.5. Pull $f(x, y, z)dx + g(x, y, z)dy + h(x, y, z)dz$ back to a 2-form in θ and ϕ.

Remark 24.6. *Note the notational*[8] *space that separates dx and dy in the common notation with $dx\,dy = dy\,dx$. With x_1 and x_2 replacing x and y the notation*

$$\int_\Omega v_{x_1} = \int_\Omega v_{x_1}\,dx = \iint_\Omega v_{x_1}(x_1, x_2)\,d(x_1, x_2),$$

and likewise for the other integral, would be more to my liking but everybody writes $dx_1\,dx_2$ and $dx\,dy$, rather than $dx = d(x_1, x_2)$ and $d(x, y)$. Without the notational space we have forms $dxdy = -dydx$ that are usually written with wedges, namely $dx \wedge dy = -dy \wedge dx$.

[8]Section 24 introduced notation with dx and dy not separated and $dxdy = -dydx$.

25 Some integral equations in two variables

Have a look at Section 7.6 before you read on. The integral equations in this section relate to *partial differential equations* (PDE's).

Exercise 25.1. Let $F : \mathbb{R} \to \mathbb{R}$ be continuous and suppose that $u \in C^2(\mathbb{R}^2)$ is a solution of

$$u_{xy} = F(u) \tag{25.1}$$

with $u = 0$ of the both the axes. Show that

$$u(x,y) = \underbrace{\int_0^x \int_0^y F(u(\xi,\eta)\, d\eta\, d\xi}_{\Phi(u)(x,y)} \quad (x,y \in \mathbb{R}). \tag{25.2}$$

This is like a two variable version of (7.18) with $u_0 = 0$ which we solved in Exercise 7.41 using weighted norms.

Exercise 25.2. For $F : \mathbb{R} \to \mathbb{R}$ Lipschitz continuous we solve (25.2) first in $C(\bar{B}_R)$ using the norm

$$|u|_{\mu,R} = \max_{x^2+y^2 \le R^2} \frac{|u(x,y)|}{\exp(\mu(x^2+y^2))}.$$

Show that for every $R > 0$ there exists $\mu > 0$ such that Φ is a contraction. Use this to show that (25.2) has a unique solution in $C(\mathbb{R}^2)$.

Remark 25.3. *Did we solve (25.1)? The solution of (25.2) does have some differentiability properties, but it is not so clear whether it is in $C^2(\mathbb{R}^2)$. Note that (25.1) is the nonlinear one-dimensional wave equation*

$$v_{tt} = v_{xx} + G(v) \tag{25.3}$$

in disguise. For the linear inhomogeneous wave equation

$$v_{tt} = v_{xx} + F(t,x) \tag{25.4}$$

there exists the d'Alembert solution formula

$$v(t,x) = \frac{1}{2}(f(x-t) + f(x+t)) + \frac{1}{2}\int_{x-t}^{x+t} g + \frac{1}{2}\iint_{C(t,x)} F \tag{25.5}$$

for the solution with initial data

$$v(0,x) = f(x), \quad v_t(0,x) = g(x) \quad (x \in \mathbb{R}). \tag{25.6}$$

In (25.5)

$$C(t,x) = \{(\tau,\xi) : 0 \le \tau \le t, x + \tau - t \le \xi \le x + t - \tau\}$$

is (part of) the backwards light cone starting from (t,x), namely the triangle with vertices $(0, x \pm t)$ and (t,x). Its measure (area) is t^2. Here we restrict the attention to $t \ge 0$. The smoothness of v defined by (25.5) depends on the smoothness of f, g, F.

Exercise 25.4. Consider the integral equation

$$v(t,x) = \frac{1}{2}(f(x-t) + f(x+t)) + \frac{1}{2}\int_{x-t}^{x+t} g + \frac{1}{2}\iint_{C(t,x)} G(v), \tag{25.7}$$

which would correspond to the solution of (25.3) with initial data given by (25.6). Assume that $G : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant $L > 0$, and $f, g :\to \mathbb{R}$ are continuous and bounded. Let C_T be the space of all continuous bounded functions $v : \mathbb{R} \times [0,T] \to \mathbb{R}$ equipped with the supremum norm, i.e.

$$|v|_T = \sup_{\substack{x\in\mathbb{R} \\ 0\le t\le T}} |v(t,x)|.$$

Prove that (25.7) has a unique solution in C_T for every T with $LT^2 < 1$.

Exercise 25.5. (continued) Modify the argument in the spirit of Exercise 25.1 using weighted norms

$$|v|_{\mu,T} = \sup_{\substack{x\in\mathbb{R} \\ 0\le t\le T}} \frac{|v(t,x)|}{\exp(\mu t)}$$

to establish that (25.7) has a unique continuous solution $v : \mathbb{R} \times [0,\infty) \to \mathbb{R}$ which is in every C_T.

Exercise 25.6. Let $F : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous. Rewrite

$$u_{xyz} = F(u) \tag{25.8}$$

with $u = 0$ if $xyz = 0$ as an integral equation and show that the integral equation has a unique continuous solution $u : \mathbb{R}^3 \to \mathbb{R}$. Generalise to $\mathbb{R}^n$.

26 Parameterisations and integrals

Part of this chapter was already done in Chapter 21. Let us recall the main result, which concerns a bounded open set $\Omega \subset \mathbb{R}^N$ with $\partial\Omega \in C^1$ and a function $v \in C^1(\Omega) \cap C(\overline{\Omega})$. In the proof of Theorem 21.11 we explained how

$$\int_\Omega v_{x_i} = \int_{\partial\Omega} \nu_i \, v \tag{26.1}$$

follows from local calculations, in which the boundary integrals are actually defined using parameterisations of a very special form, in the notation of most of this chapter, $u \to \Phi(u) = (u, f(u))$, with $f : [a,b] \to \mathbb{R}$. The local statements led to the global statement via arguments which involved cut-off functions[1] and partitions of unity, which will be discussed as an independent topic in Section 29.

26.1 The length of a curve

In the 1-dimensional case I now follow Edwards[2] and write $x = \gamma(t)$ with $t \in [a,b]$ and $\gamma : [a,b] \to \mathbb{R}^N$. For any such γ the natural definition of the length would be the smallest upper bound on the set of numbers obtained via

$$\sum_{j=1}^{m} |\gamma(t_j) - \gamma(t_{j-1})|_2$$

with

$$a = t_0 < t_1 < \cdots < t_m = b.$$

Clearly this definition of length is invariant under reparameterisation of γ via strictly monotone bijections $\phi : [a,b] \to [c,d]$ as in Section 8.5. It's not a very hard exercise to show that for continuously differentiable $\gamma : [a,b] \to \mathbb{R}^N$ the length is given by

$$s(\gamma) = \int_a^b |\gamma'(t)|_2 \, dt,$$

and the change of variables formula applied to $u = \phi(t)$ with $\phi \in C^1([a,b])$ with $\phi'(t) \neq 0$ confirms that

$$\Phi : \, u \to \gamma(\phi^{-1}(u)) \tag{26.2}$$

[1] I called them fading functions.

[2] This was written while teaching from his book Advanced Calculus of Several Variables.

is in $C^1([c,d])$ with $[c,d] = \phi([a,b])$, and has the same length[3]. Also, if $f = f(x)$ is continuous on $\gamma([a,b]) = \Phi([c,d])$, it follows that

$$\int_\gamma f = \int_\gamma f ds = \int_a^b f(\gamma(t))|\gamma'(t)|_2 \, dt = \int_c^d f(\Phi(u))|\Phi'(u)|_2 \, du = \int_\Phi f ds. \tag{26.3}$$

As a special case we have that

$$s = \phi(t) = \int_a^t |\gamma'(\tau)|_2 \, d\tau$$

defines a reparameterisation for which $\hat{\gamma} = \Phi$ defined by $\gamma(t) = \hat{\gamma}(s) = \Phi(s)$ has

$$|\hat{\gamma}'(s)|_2 = |\Phi'(s)|_2 = 1.$$

Such a reparametrised $\tilde{\gamma}$ is called a unit speed path.

26.2 Line integrals of vector fields along curves

Besides (26.3) as a 1-dimensional example of what is to come in (26.13) we can also define an integral for $F = F(x) \in \mathbb{R}^n$ continuous on $\gamma([a,b])$, namely

$$\int_\gamma F \cdot ds = \int_a^b F(\gamma(t)) \cdot \gamma'(t) \, dt = \int_a^b F(\gamma(t)) \cdot \underbrace{\frac{\gamma'(t)}{|\gamma'(t)|_2}}_{T(t)} |\gamma'(t)|_2 \, dt, \tag{26.4}$$

but Edwards avoids the commonly used notation in the left hand side of (26.4). Instead he writes

$$\int_\gamma F \cdot T \, ds,$$

with T the unit tangent vector[4] defined by

$$T(t) = \frac{\gamma'(t)}{|\gamma'(t)|_2}.$$

For reparametrisations $u = \phi(t)$ with $\phi \in C^1([a,b])$ and $\phi'(t) > 0$ and Φ defined as in (26.2) above you easily verify that the *work*

$$W = \int_\gamma F \cdot ds = \int_\gamma F \cdot T \, ds = \int_\Phi F \cdot T \, ds = \int_\Phi F \cdot ds.$$

[3]The condition that $\gamma'(t) \neq 0$ also carries over to $\Phi'(u) \neq 0$.

[4]I will use $\tau = T$.

done by the *force field* F does not change under reparametrisations $u = \phi(t)$ with $\phi'(t) > 0$. Of course

$$W = \int_\gamma F \cdot ds = \int_\gamma F \cdot T\, ds = \int_a^b (F_1(\gamma(t))\gamma_1'(t) + \cdots + F_N(\gamma(t))\gamma_N'(t))\, dt$$

$$= \int_a^b F_1(\gamma(t))\, \underbrace{\gamma_1'(t)\, dt}_{dx_1} + \cdots + \int_a^b F_N(\gamma(t))\, \underbrace{\gamma_N'(t)\, dt}_{dx_N}$$

leads to the notational convention

$$\int_\gamma F \cdot ds = \int_\gamma F_1 dx_1 + \cdots + \int_\gamma F_N dx_N = \int_\gamma F_1 dx_1 + \cdots + F_N dx_N. \quad (26.5)$$

If $F = \nabla f$ it is common to write

$$\int_\gamma df = \int_\gamma \underbrace{\frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_N} dx_N}_{df} = \int_\gamma \nabla f \cdot ds =$$

$$\int_a^b \nabla f(\gamma(t)) \cdot \gamma'(t)\, dt = f(\gamma(t))\big|_a^b = f(\gamma(b)) - f(\gamma(a)),$$

a notation which generalises (10.6), after which d was seen[5] as acting on f to produce $df = f'(x)dx$. Here we have d acting on f as[6]

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_N} dx_N. \quad (26.6)$$

These 1-forms act on vectors. Whereas the x-dependent vector[7]

$$F(x) = F_1(x)e_1 + \cdots + F_N(x)e_N \quad (26.7)$$

The vector

$$v = v_1 e_1 + \cdots + v_N e_N$$

have and x-dependent inner product

$$F(x) \cdot v = F_1(x)v_1 + \cdots + F_N(x)v_N.$$

the 1-form

$$\omega = F_1(x)dx_1 + \cdots + F_N(x)dx_N \quad (26.8)$$

[5]Writing f instead of F again.

[6]Compare to (24.3) in Section 24.1.

[7]As in Section 20.2 we consider the e_i as column vectors.

assigns to the same vector v the same x-dependent scalar

$$F_1(x)v_1 + \cdots + F_N(x)v_N,$$

in which we can insert $x = \gamma(t)$ and $v_i = \gamma_i'(t)$ to get a t-dependent quantity that we can integrate from $t = a$ to $t = b$ to define

$$\int_a^b (F_1(\gamma(t))\gamma_1'(t) + \cdots + F_N(\gamma(t))\gamma_N'(t))\, dt = \int_\gamma \omega.$$

Thus, ω evaluated in $x = \gamma(t)$ acts on $\gamma'(t)$ and is integrated from $t = a$ to $t = b$ to define $\int_\gamma \omega$. Note that a reparameterisation of γ with $u = \phi(t)$ and $\phi'(t) < 0$ changes the sign of the integral.

The notation for ω hides the x-dependence, just like the abuse of notation in $f = f(x)$. In conclusion we have $\int_\gamma f = \int_\gamma f\, ds$ defined for continuous scalar functions $f = f(x)$ and $\int_\gamma \omega$ for 1-forms $\omega = F_1(x)dx_1 + \cdots + F(x)dx_N$.

26.3 Surface area

We need some linear algebra[8] for integrals over more general surface patches then the ones encountered in Section 21.3. *We now understand a surface patch* to be a set in $\mathbb{R}^3$ parameterised by a continuously differentiable injective map

$$\Phi : [0,1] \times [0,1] \to \mathbb{R}^3, \tag{26.9}$$

with

$$\nabla\Phi = (\nabla\Phi_1\ \nabla\Phi_2\ \nabla\Phi_3) = \begin{pmatrix} \frac{\partial\Phi_1}{\partial u_1} & \frac{\partial\Phi_2}{\partial u_1} & \frac{\partial\Phi_3}{\partial u_1} \\ \frac{\partial\Phi_1}{\partial u_2} & \frac{\partial\Phi_2}{\partial u_2} & \frac{\partial\Phi_3}{\partial u_2} \end{pmatrix}$$

denoting the matrix of which the columns are the gradients of the $N = 3$ components Φ_1, Φ_2, Φ_3 of Φ with respect to the $n = 2$ variables[9] u_1, u_2 in $\Phi = \Phi(u) = \Phi(u_1, u_2)$, consistent with the notation in Section (20).

Momentarily switching to a notation with Φ_1, Φ_2, Φ_3 as functions of u, v, $\nabla\Phi$ is the transpose of the Jacobian matrix

$$(\ \frac{\partial\Phi}{\partial u}\ \frac{\partial\Phi}{\partial v}\ ),$$

which has column vectors $\frac{\partial\Phi}{\partial u}$, $\frac{\partial\Phi}{\partial v}$. In the special linear case with

$$\Phi_i(u, v) = a_i u + b_i v \tag{26.10}$$

[8] Theorem 19.8.

[9] Everything that follows should generalise or trivialise to $1 \le n \le N$.

the Jacobian matrix is the transpose of

$$\nabla\Phi = A^T = \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix},$$

the matrix example in (19.14) starting the discussion in Section 19.5 below on

the area $\mathcal{M}_2(a,b)$ of a parallelogram spanned by two vectors a and b

with entries a_1, a_2, a_3 and b_1, b_2, b_3 respectively. This parallelogram is then the image of $[0,1] \times [0,1]$ under Φ defined by (26.10), and its area is then equal to

$$\int_0^1 \int_0^1 \mathcal{M}_2(\frac{\partial\Phi}{\partial u}, \frac{\partial\Phi}{\partial v})\, du\, dv, \tag{26.11}$$

the integrand being independent of u, v, as $a = \frac{\partial\Phi}{\partial u}$ and $b = \frac{\partial\Phi}{\partial v}$ are constant vectors is the linear case (26.10).

It will be no surprise that (26.11) will also be used to define the area of the surface patch defined by Φ if Φ is not a linear map from $[0,1]^2$ to $\mathbb{R}^3$, and that everything generalises to $\Phi : [0,1]^n \to \mathbb{R}^N$ with $1 \leq n < N$. We expand on the linear case of this generalisation next.

26.4 Surface integrals

I now return to (26.11). Generalising to $1 \leq n \leq N$ we consider

$$\int_{[0,1]^n} \mathcal{M}_n(\frac{\partial\Phi}{\partial u})\, du = \int_0^1 \cdots \int_0^1 \mathcal{M}_n(\frac{\partial\Phi}{\partial u_1}, \ldots, \frac{\partial\Phi}{\partial u_n})\, du_1 \cdots du_n \tag{26.12}$$

in which

$$\mathcal{M}_n(\frac{\partial\Phi}{\partial u_1}, \ldots, \frac{\partial\Phi}{\partial u_n}) = \mathcal{M}_n(\Phi_{u_1}, \ldots, \Phi_{u_n})$$

is given by Theorem 19.8. Here $du = du_1 \cdots du_n$ and $\int_{[0,1]^n} = \int_0^1 \cdots \int_0^1$ are just notational conventions.

In the special case that $n = 1$ we have

$$\mathcal{M}_1(\Phi_u) = \sqrt{\Phi'_1(u)^2 + \cdots + \Phi'_n(u)^2},$$

and

$$ds = \mathcal{M}_1(\Phi_u)\, du = \sqrt{\Phi'_1(u)^2 + \cdots + \Phi'_n(u)^2}\, du$$

is a common notation, introduced[10] after a change of coordinates defined by

$$\frac{ds}{du} = \sqrt{\Phi'_1(u)^2 + \cdots + \Phi'_n(u)^2}.$$

[10] In Edwards Section V.1, his $\gamma(t)$ would correspond $\Phi(u)$.

While not corresponding to a change of coordinates the notation

$$dS = \mathcal{M}_2(\Phi_u, \Phi_v)\, du\, dv,$$

with the S of surface, is also common. Here I will use dS_n for

$$dS_n = \mathcal{M}_n(\frac{\partial \Phi}{\partial u})\, du = \mathcal{M}_n(\Phi_{u_1}, \dots, \Phi_{u_n})\, du_1 \cdots du_n$$

in (26.12), i.e.

$$\int_\Phi dS_n = \int_0^1 \cdots \int_0^1 \mathcal{M}_n(\frac{\partial \Phi}{\partial u_1}, \dots, \frac{\partial \Phi}{\partial u_n})\, du_1 \cdots du_n.$$

For a function $f = f(x) = f(x_1, \cdots, x_n)$ which is continuous on

$$\{x = \Phi(u) : \, u \in [0,1]^n\},$$

we write

$$\int_\Phi f dS_n = \int_{[0,1]^n} f(\Phi(u)) \mathcal{M}_n(\frac{\partial \Phi}{\partial u})\, du = \tag{26.13}$$

$$\int_0^1 \cdots \int_0^1 f(\Phi(u_1, \dots, u_n)) \mathcal{M}_n(\frac{\partial \Phi}{\partial u_1}, \dots, \frac{\partial \Phi}{\partial u_n})\, du_1 \cdots du_n.$$

The subscript Φ on the integral is consistent with the case $n = 1$ and $ds = dS_1$, and coincides with the notation in the second part of (26.3). Personally I often drop the dS_n from the notation and just write $\int_\Phi f$ instead of $\int_\Phi f dS_n$, and $\int_\gamma f$ if $n = 1$ and $\gamma = \Phi$ is a path in $\mathbb{R}^N$. Af course we can also allow general closed blocks

$$[a, b] = [a_1, b_1] \times \cdots \times [a_n, b_n]$$

in stead of $[0, 1]^n$.

27 Varieties in Euclidean space

In this chapter we think of manifolds as solution sets of systems of equations in $\mathbb{R}^N$. In Chapter 28 this will bother us a bit when we get the topology on M only from the topology on $\mathbb{R}^N$. Think of lines and planes as nontrivial examples in $\mathbb{R}^3$ of linear varieties $\mathcal{M}$. Along $\mathcal{M}$ something varies, and the variations are linear: by definition linear varieties in $\mathbb{R}^N$ are solution sets of systems[1] of linear equations, which upon solving these systems are described as graphs of linear functions[2]. The typical example[3] of $\mathcal{M}$ is the graph defined by[4]

$$y = Ax + b, \tag{27.1}$$

in which $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, $A : \mathbb{R}^n \to \mathbb{R}^m$ linear, $b \in \mathbb{R}^m$, and $N = n + m$ with $n, m \in \mathbb{N}$.

Exercise 27.1. Use your knowledge of linear algebra to show that a linear variety $\mathcal{M}$ is always the graph of a linear function, unless $\mathcal{M}$ is a singleton, and then there is no reason to call it a variety. After relabelling the variables $\mathcal{M}$ is given by (27.1).

If we see x and y as column vectors then (27.1) reads as

$$(A \ -I)\begin{pmatrix} x \\ y \end{pmatrix} = b \in \mathbb{R}^m,$$

with $C = (A \ -I)$ a somewhat special matrix with m rows and N columns. The first n columns form the matrix A, the last m columns the diagonal matrix with entries -1. The matrix C acts on column vectors

$$z = \begin{pmatrix} x \\ y \end{pmatrix}$$

in $\mathbb{R}^N$. Thus (27.1) is a system of m linear equation for Nunknowns $z_1, z_2, \ldots, z_N$:

$$C_{11}z_1 + C_{12}z_2 + \cdots + C_{1N}z_N = b_1;$$

$$C_{21}z_1 + C_{22}z_2 + \cdots + C_{2N}z_N = b_2;$$

$$\vdots$$

[1] That is, $Ax = b$ with A a given matrix, b a given vector, and x the unknown vector.
[2] You may prefer to call them maps.
[3] Unless they are empty, a singleton or the whole space, you must have seen this.
[4] For some other matrix A and some other vector b of course.

$$C_{m1}z_1 + C_{m2}z_2 + \cdots + C_{mN}z_N = b_m.$$

In the example the coefficient matrix C has maximale rank, which means that you can choose m column of C which together form an invertible square matrix, in this example the last m columns. More generally, if $C = (A\ B)$ with B invertible, then the system is solved for y via $y = B^{-1}(b - Ax)$, which defines a graph, just like (27.1). We have

$$Cz = b \iff y = Ax + b \tag{27.2}$$

as equivalent descriptions of non-trivial linear varieties in $\mathbb{R}^{\mathrm{N}}$, under the assumption that C hax maximal rank.

27.1 Implicit function theorem in Euclidean spaces

Referring to Theorem 15.4 we use the notation

$$x \in X = \mathbb{R}^n,\ y \in Y = \mathbb{R}^m, \quad (x, y) \in Z = X \times Y = \mathbb{R}^{n+m}$$

to formulate the implicit function theorem in the neighbourhood of a point $(x, y) = (a, b)$. Aiming for a vector version of (15.25) we assume that $(x, y) \to F_x(x, y)$ and $(x, y) \to F_y(x, y)$ are continuous near $(x, y) = (a, b)$. Equivalently: F is continuously differentiable in a neighbourhood of $(x, y) = (a, b)$.

Theorem 27.2. *(Implicit function theorem) For $r > 0$ let the $\mathbb{R}^m$-valued function F be continuously differentiable on $B_r(a) \times B_r(b)$. If $F_y(a, b)$ is invertible then there exist $\delta_0 > 0$ and $\varepsilon_0 > 0$, and a continuously differentiable function*

$$f : \bar{B}_{\delta_0}(a) \to B_{\varepsilon_0}(b),$$

such that

$$\{(x, y) \in \bar{B}_{\delta_0}(a) \times \bar{B}_{\varepsilon_0}(b) :\ F(x, y) = F(a, b)\} = \{(x, f(x)) :\ x \in \bar{B}_{\delta_0}(a)\}.$$

It holds that

$$f'(x) = -(F_y(x, f(x)))^{-1}F_x(x, f(x)) \quad \textit{for all} \quad x \in \bar{B}_{\delta_0}(a).$$

The proof can be copied from the proofs of Theorems 15.1 and 15.2. Recall that the function $x \to F(x, f(x))$ is never differentiated tot derive the expression for $f'(x)$ but differentiation of this function does help to remember the result. The construction of $y = f(x)$ requires first a choice of $0 < \varepsilon_0 \leq r$ and then a choice of $\delta_0 > 0$ sufficiently small, which in the end has to be chosen even smaller to also have $f'(x) = -(F_y(x, f(x)))^{-1}F_x(x, f(x))$ for

$|x| \leq \delta_0$. In general it will not be the case that $\delta_0 > \varepsilon$. Thus Theorem 27.2 can be read as stating the existence of $0 < \delta_0 \leq \varepsilon_0 \leq r$ for which the assertions hold.

Applying Theorem 27.2 to

$$F(x, y) = x - g(y)$$

we obtain the inverse function theorem via the statement

$$\{(x, y) \in \bar{B}_{\delta_0}(a) \times \bar{B}_{\varepsilon_0}(b) : g(y) = x\} = \{(x, f(x)) : x \in \bar{B}_{\delta_0}(a)\},$$

with $f'(x) = (g'(f(x))^{-1}$ for all $x \in \bar{B}_{\delta_0}(a)$. The solution $y = f(x)$ of $x = g(y)$ is constructed with the scheme

$$y_{n+1} = y_n - g'(0)^{-1}(g(y_n) - x),$$

starting from $y_0 = 0$. We formulate the result for $X = Y = \mathbb{R}^n$ en $g : Y \to Y$.

Theorem 27.3. *(Inverse function theorem) For $r > 0$ let $g : Y \to Y$ be continuously differentiable on $\bar{B}_r(b)$ and let $a = g(b)$. If $g'(b)$ is invertible there exist $0 < \delta_0 \leq \varepsilon_0 \leq r$ and a continuously differentiable injective function $f : \bar{B}_{\delta_0}(a) \to B_{\varepsilon_0}(b)$, such that for all $(x, y) \in \bar{B}_{\delta_0}(a) \times \bar{B}_{\varepsilon_0}(b)$ it holds that $x = g(y) \iff y = f(x)$, and $f'(x) = (g'(f(x))^{-1}$ for all $x \in \bar{B}_{\delta_0}(a)$.*

N.B. Theorem 27.2 gives $f : \bar{B}_{\delta_0}(a) \to B_{\varepsilon_0}(b)$ in Theorem 27.3 only as continuously differentiable function. Because $y = f(x)$ for $x \in \bar{B}_{\delta_0}(a)$ it follows that $x = g(y) = g(f(x))$, so f is injective on $\bar{B}_{\delta_0}(a)$, and in view of $f'(x) = (g'(f(x))^{-1}$ it must be that $f'(x)$ is invertible in every $x \in \bar{B}_{\delta_0}(a)$.

This argument does not immediately apply to g: to insert $x = g(y)$ in $y = f(x)$ we must have $g(y)$ in the domain of f. But Theorem 27.3 can be applied once more (interchange the roles of x and y) to obtain $0 < \varepsilon_1 \leq \delta_1 \leq \delta_0$ and a continuously differentiable $g_1 : \bar{B}_{\varepsilon_1}(b) \to \bar{B}_{\delta_1}(a)$ such that for $(x, y) \in \bar{B}_{\delta_1}(a) \times \bar{B}_{\varepsilon_1}(b)$ it holds again that $x = g_1(y) \iff y = f(x)$. From the earlier equivalence $x = g(y) \iff y = f(x)$ for all $(x, y) \in \bar{B}_{\delta_0}(a) \times \bar{B}_{\varepsilon_0}(b)$ we have that $g_1 = g$ on $\bar{B}_{\varepsilon_1}(b)$. Just as earlier for $f : \bar{B}_{\delta_0}(a) \to \bar{B}_{\varepsilon_0}(b)$ it follows that g_1 and therefore g is injective on $\bar{B}_{\varepsilon_1}(b)$.

Summarizing we conclude that in the chain

$$\bar{B}_{\varepsilon_1}(b) \xrightarrow{g} \bar{B}_{\delta_1}(a) \to \bar{B}_{\delta_0}(a) \xrightarrow{f} \bar{B}_{\varepsilon_0}(b) \xrightarrow{g} X = Y = \mathbb{R}^n$$

not only f but also the g in te first link is injective. The second link is the inclusion map. The keten can be extended to the left. Starting from a met continuously differentiable

$$\mathbb{R}^n \supset \bar{B}_{\delta_0}(a) \xrightarrow{f} \mathbb{R}^n \tag{27.3}$$

with $f'(a)$ invertible, we have with $b = f(a)$ a diagram that goes on forever:

$$\begin{array}{ccc}
\bar{B}_{\delta_0}(a) & \xrightarrow{f} & \mathbb{R}^n \\
\uparrow & & \uparrow \\
\bar{B}_{\delta_1}(a) & \xleftarrow{g} & \bar{B}_{\varepsilon_1}(b) \\
\uparrow & & \uparrow \\
\bar{B}_{\delta_2}(a) & \xrightarrow{f} & \bar{B}_{\varepsilon_2}(b) \\
\uparrow & & \uparrow \\
\bar{B}_{\delta_3}(a) & \xleftarrow{g} & \bar{B}_{\varepsilon_3}(b) \\
\uparrow & & \uparrow
\end{array}$$

Every image is contained in the open ball. Except for the first top link, every link is injective but in general not surjective, with invertible $f'(x)$ and $g'(y)$ (because of $f'(x) = (g'(f(x))^{-1}$ and $g'(y) = (f'(g(y))^{-1})$. Going down the epsilons and deltas get smaller.

Exercise 27.4. Derive 27.2 from Theorem 27.3. Hint: use F to construct a function $\tilde{F} : \mathbb{R}^{\mathrm{N}} = \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}^m$ which has its last m components given by $F(x, y)$ and its first n components by x itself.

27.2 General subvarieties

For in general nonlinear subvarieties[5] we ask about an equivalence simliar to (27.2), starting from the nonlinear version $Cz = b$, written in Theorem 27.2 as[6]

$$F(z) = F(x, y) = 0,$$

with $F : \mathbb{R}^{\mathrm{N}} \to \mathbb{R}^m$ continuously differentiable. We use the nonlinear version of (27.1) to agree what we mean by a subvariety $\mathcal{M} \subset \mathbb{R}^{\mathrm{N}}$:

Definition 27.5. *Let $n \in \{1, \ldots, N-1\}$. An n-dimensional C^1-subvariety $\mathcal{M} \subset \mathbb{R}^{\mathrm{N}}$ is a set that in a neigbourhood of any of its points can be written like the level set $F(x, y) = F(a, b)$ in Theorem 27.2: possibly after renumbering the coordinates it must be that every point $p \in \mathcal{M}$ has*

$$p = (a, b) \in \mathcal{M} \cap \bar{B}_{\delta_0}(a) \times \bar{B}_{\varepsilon_0}(b) = \{(x, f(x)) : x \in \bar{B}_{\delta_0}(a)\}.$$

[5] Not defined yet!

[6] We prefer to have y to the right of x in the notation.

for some $\delta_0 > 0$ and $\varepsilon_0 > 0$, and some f continuously differentiable from $\bar{B}_{\delta_0}(a)$ to $B_{\varepsilon_0}(b)$. If $n = N - 1$ then $\mathcal{M}$ is called a hypersurface.

Exercise 27.6. Let $F : \mathbb{R}^N \to \mathbb{R}^m$ be continuously differentiable. Assume that for all $z \in \mathbb{R}^N$ with $F(z) = 0$ the derivative $F'(z)$, seen as matrix, has maximal rank. Prove that $\{z \in \mathbb{R}^N : F(z) = 0\}$ is an n-dimensional subvariety of $\mathbb{R}^N$, with $n + m = N$.

Exercise 27.7. Give an example of an n-dimensional subvariety $\mathcal{M} \subset \mathbb{R}^N$ which is not given by a function F as in Exercise 27.6.

The standard example for Exercise 27.6 is the boundary of a ball in $\mathbb{R}^n$ with center $(a_1, a_2, \ldots, a_n)$ and radius $\delta > 0$:

$$(x_1 - a_1)^2 + \cdots + (x_n - a_n)^2 - \delta^2 = 0. \tag{27.4}$$

There are three equivalent ways to say that $\mathcal{M} \subset \mathbb{R}^N$ is an n-dimensional subvariety:

(A) $\mathcal{M}$ is locally the graph of a continuously differentiable function

$$f : \mathbb{R}^n \to \mathbb{R}^m \quad (n + m = N),$$

given by $y = f(x)$ after renumbering $z = (x, y)$.

(B) $\mathcal{M}$ is locally the zero level set of

$$F : \mathbb{R}^N \to \mathbb{R}^m \quad (n + m = N),$$

a continuously differentiable function with, after renumbering, F_y invertible in the points $z = (x, y) \in M$ under consideration.

(C) M is locally the image[7] of a continuously differentiable function

$$\Phi : \mathbb{R}^n \to \mathbb{R}^N,$$

which is injective and has Φ' of maximal rank.

Theorem 27.2 showed that (B) $\implies$ (A), and (A) $\implies$ (B) because (A) is a special case of B with $F(x, y) = g(y) - x$. Likewise (A) is a special case of

[7]The inverse map of Φ is called a chart on M.

C with $\Phi(x) = (x, f(x))$. To complete the circle with a proof that (C) $\Longrightarrow$ (A) we use Theorem 27.3 and the chain rule.

To wit, consider Φ as

$$\Phi : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^m,$$

with, after renumbering, $\Phi(x) = (\Psi(x), \chi(x))$, $\Psi : \bar{B}_r(a) \to \mathbb{R}^n$ and $\chi : \bar{B}_r(a) \to \mathbb{R}^m$ continuously differentiable, and $\Psi'(a)$ invertible in a. This is possible because we assumed that $\Phi'(x)$ is of maximal rank in $x = a$. Theorem 27.3, applied to $g = \Psi$ with $y = x$, provided us with a continuously differentiable injective function f renamed here as ϕ, $\phi : \bar{B}_{\delta_0}(\Psi(a)) \to \mathbb{R}^n$, with $\phi'(\xi)$ invertible[8] for all $\xi \in \bar{B}_{\delta_0}(\Psi(a))$, and $\Psi(\phi(\xi)) = \xi$ for all $\xi \in \bar{B}_{\delta_0}(\Psi(a))$. Thus

$$\xi \to \Phi(\phi(\xi)) = (\Psi(\phi(\xi)), \chi(\phi(\xi))) = (\xi, f(\xi)),$$

with $f(\xi) = \chi(f(\xi))$, parameterises $\mathcal{M}$ in a neigbourhood of $b = \Psi(a)$ and hence $\mathcal{M}$ is locally given as the graph of $f : \bar{B}_{\delta_0}(b) \to \mathbb{R}^{\mathrm{N}}$. The continuous differentiability of f follows from the chain rule, the first time we use it actually. The proof of

$$(\mathrm{A}) \Longleftrightarrow (\mathrm{B}) \Longleftrightarrow (\mathrm{C})$$

is now complete.

Exercise 27.8. Let $\mathcal{M} \subset \mathbb{R}^n$ be a subvariety and $f : \mathbb{R}^n \to \mathbb{R}^n$ continuously differentiable in a neighbourhood of each and very point of $\mathcal{M}$. If $f'(x)$ is invertible for every $x \in \mathcal{M}$ and f is injective on M, then the image of $\mathcal{M}$ under f is again a subvariety. Why?

Exercise 27.9. As Exercise 27.8, but with $f : \mathbb{R}^n \to \mathbb{R}^m$ and $f'(x)$ of maximal rank in every $x \in \mathcal{M}$.

[8] Not used here.

27.3 Images of ball boundaries

With $0 < \varepsilon_1 \leq \delta_1 \leq \delta_0 \leq \varepsilon_0$ Theorem 27.3 provided us with a chain

$$\bar{B}_{\varepsilon_1}(b) \xrightarrow{g} \bar{B}_{\delta_1}(a) \xrightarrow{f} \bar{B}_{\varepsilon_0}(b)$$

in which both links are injective but not surjective as every mage is contained in the open ball. The smaller δ_1 and ε_1 were needed for the injectivity of g on the smaller closed ball $\bar{B}_{\varepsilon_1}(b)$.

The images of the boundaries $\partial B_{\varepsilon_1}(b)$ and $\partial B_{\varepsilon_0}(a)$ are the subvarieties $g(\partial B_{\varepsilon_1}(b))$ and $f(\partial B_{\varepsilon_0}(a))$. In case g and f are linear maps and $a = b = 0$, it is easy to see that these images are graphs over the unit sphere

$$S^{n-1} = \{x \in \mathbb{R}^n : |x| = 1\}.$$

If $A : \mathbb{R}^n \to \mathbb{R}^n$ is such an invertible linear map, then a height function $h : S^{n-1} \to \mathbb{R}^+$ can be constructed to make that the image of $\partial B_1(0)$ under A is of the form

$$S_h = \{h(x)x : x \in S^{n-1}\}. \tag{27.5}$$

The function h is contructed by intersecting the half lines

$$\{\lambda x : \lambda > 0\}$$

through $x \in S^{n-1}$ with $A(\partial B_1(0))$. You may prefer to use another name for x here if you think in terms of $y = Ax$.

Exercise 27.10. Let $A : \mathbb{R}^n \to \mathbb{R}^n$ be an invertible linear map Prove that every $\xi \in S^{n-1}$ has a unique $\lambda > 0$ such that $\lambda\xi \in A(\partial B_1(0))$. Setting $\lambda = h_A(\xi)$ defines $h_A : S^{n-1} \to \mathbb{R}^+$. Prove that the image of $\partial B_\delta(0)$ under A has heigth function $\xi \to \delta h_A(\xi)$.

These questions and answers about $g(\partial B_{\varepsilon_1}(b))$ and $f(\partial B_{\varepsilon_0}(a))$ lead to the question if the statements in Exercise 27.10 also hold for the image of a small ball boundery $\bar{B}_{\delta_1}(0)$ under a continously differentiable map $F : \bar{B}_\delta(0) :\to \mathbb{R}^n$ of the form

$$F(x) = Ax + R(x) \quad \text{with} \quad R(x) = o(|x|) \quad \text{for} \quad |x| \to 0.$$

Theorem 27.3 tells us that F os injective is on a smaller ball $\bar{B}_{\delta_0}(0)$ with $F'(x)$ invertible (not only for $x = 0$ but also) for all $x \in \bar{B}_{\delta_0}(0)$. The next exercise is a small project that also requires Theorem 27.2, to be expanded on.

Exercise 27.11. Prove the statement in Exercise 27.10 for the image $F(\partial B_{\delta_1}(0))$ of a small ball boundary $\bar{B}_{\delta_1}(0)$. Establish the continous differentiability of the height function h you construct in a neighbourhood of every point of S^{n-1}, as function of suitable chosen local coordinates.

27.4 Coordinate transformations

If a point P on an n-dimensional subvariety $\mathcal{M}$ of $\mathbb{R}^N$ lies in the image of a Φ and a Ψ as in (C) in Section 27.2, say with $\Phi(\xi)$ and $\Psi(\eta)$, and $P = \Phi(0) = \Psi(0)$, with 0 an interior point of the domains of Φ and Ψ, then ξ are η are related by statements as in Theorem 27.3 in a neighbourhood of 0.

27.5 Higher order derivatives of the implicit function

Apply the implicit function theorem to

$$\tilde{F} : (x, h) \to (F(x), F'(x)h)$$

and obtain statements about the second derivatives of the implicit function f constructed before or simultaneously to describe the level set of F as a graph.

28 Integration over manifolds

Section 27.2 and Section 29.3 below will concern 3 descriptions of what it means for $M \subset \mathbb{R}^N$ to be an n-dimensional manifold in $\mathbb{R}^N$. We now use characterisation (C), and assume in addition that there exist finitely many injective continuously differentiable

$$\Phi_i : [a_i, b_i] \to \mathbb{R}^N$$

defined on blocks $[a_i, b_i]$ as in the elaboration on (C) in Section 29.3 above[1], such that

$$M = \Phi_1((a_1, b_1)) \cup \cdots \cup \Phi_m((a_n, b_n)) = \Phi_1([a_1, b_1]) \cup \cdots \cup \Phi_m([a_n, b_n]), \quad (28.1)$$

and moreover that there exist corresponding smooth functions

$$\zeta_i : \mathbb{R}^N \to [0, 1]$$

with

$$\zeta_1 + \cdots + \zeta_n \equiv 1 \text{ on } M \text{ and } \operatorname{supp} \zeta_i \circ \Phi_i \subset (a_i, b_i)$$

for every $i = 1, \ldots, I$. Here $\operatorname{supp} \zeta_i \circ \Phi_i$ is the support of the function $u \to \zeta_i(\Phi_i(u))$, defined as the closure of the set

$$\{u \in (a_i, b_i) : \zeta_i(\Phi_i(u)) \neq 0\}.$$

We say that $u \to \zeta_i(\Phi_i(u))$ belongs to $C_c^1((a_i, b_i))$, the class of C^1-functions with support contained in the open set (a_i, b_i).

You can think of each function ζ_i as fading the patch $\Phi_i((a_i, b_i))$, making it fade away completely near its boundary where $\zeta_i \equiv 0$, while together the ζ_i leave the whole of M as *bright* as it was before. Such *fading* functions ζ_i can be chosen to vanish outside a neighbourhood in $\mathbb{R}^N$ of the image $\Phi_i(K_i)$, and the collection $\zeta_1, \ldots, \zeta_m$ is called a finite partition of unity on M, which is then (turning[2] a theorem around which says that such partitions exist if M is compact) a closed and bounded subset of $\mathbb{R}^N$.

If $f : M \to \mathbb{R}$ is continuous we now wish to define

$$\int_M f dS_n = \int_{\Phi_1} f\zeta_1 dS_n + \cdots + \int_{\Phi_m} f\zeta_m dS_n, \quad (28.2)$$

which requires a theorem that says this is independent of the choice of patches and fading functions. We leave this issue[3] for now.

[1] The index i numbering the blocks now.

[2] Following Steenbrink in his exposition of the Poincaré conjecture in Noordwijkerhout.

[3] But see later sections.

Of course the exposition above involves the change of variables theorem and Section 27.4. At the end of the day every theorem that we may wish to prove involving integrals of functions over M may be proved by restating and proving a local form only.

Finally we note that if the blocks $[a_i, b_i]$ and the injective continuously differentiable functions $\Phi_i : [a_i, b_i] \to \mathbb{R}^N$ with $\Phi'(u)$ of maximal rank can be chosen such that[4]

$$M = \Phi_1([a_1, b_1]) \cup \cdots \cup \Phi_m([a_n, b_n]) \quad \text{with} \quad \Phi_i((a_i, b_i)) \cap \Phi_j((a_j, b_j)) = \emptyset \tag{28.3}$$

for $i \neq j$, then

$$\int_M f dS_n = \int_{\Phi_1} f \, dS_n + \cdots + \int_{\Phi_m} f dS_n \tag{28.4}$$

is the obvious definition which Edwards uses, and which is what you do in examples. Usually there are many ways to choose the patches.

28.1 More integration of differential forms

We look again at the right hand side of (21.11) with $N = n+1$, evaluated for $\tilde{v}_i = \zeta v_i$ with ζ a cut-off function vanishing outside and near the boundary of some window

$$[a, b] = [a_1, b_1] \times \cdots \times [a_N, b_N],$$

in which we now assume a local representation of $\Omega \cap [a, b]$ given by[5]

$$(x_1, \ldots, x_n) \in [a_1, b_1] \times \cdots \times [a_n, b_n] \quad \text{and} \quad a_N \leq x_N < f(x_1, \ldots, x_n),$$

with $f \in C^1([a_1, b_1] \times \cdots \times [a_n, b_n])$ taking values in (a_N, b_N), and

$$\Phi(u_1, \ldots, u_n) = (u_1, \ldots, u_n, f(u_1, \ldots, u_n)) \tag{28.5}$$

parameterising $M \cap [a, b] = \partial\Omega \cap [a, b]$. We denote the unit basis vectors by $e_1, \ldots, e_N$.

For $n = 2$ the vector obtained by the formal determinant manipulation

$$\begin{vmatrix} \frac{\partial \Phi_1}{\partial u_1} & \frac{\partial \Phi_2}{\partial u_1} & \frac{\partial \Phi_3}{\partial u_1} \\ \frac{\partial \Phi_1}{\partial u_2} & \frac{\partial \Phi_2}{\partial u_2} & \frac{\partial \Phi_3}{\partial u_2} \\ e_1 & e_2 & e_3 \end{vmatrix} = \begin{vmatrix} \frac{\partial \Phi_1}{\partial u_1} & \frac{\partial \Phi_2}{\partial u_1} \\ \frac{\partial \Phi_1}{\partial u_2} & \frac{\partial \Phi_2}{\partial u_2} \end{vmatrix} e_3 + \begin{vmatrix} \frac{\partial \Phi_2}{\partial u_1} & \frac{\partial \Phi_3}{\partial u_1} \\ \frac{\partial \Phi_2}{\partial u_2} & \frac{\partial \Phi_3}{\partial u_2} \end{vmatrix} e_1 + \begin{vmatrix} \frac{\partial \Phi_3}{\partial u_1} & \frac{\partial \Phi_1}{\partial u_1} \\ \frac{\partial \Phi_3}{\partial u_2} & \frac{\partial \Phi_1}{\partial u_2} \end{vmatrix} e_2 \tag{28.6}$$

[4]Edwards: a hard theorem says this can be done.

[5]Like in Section 21.3.

is commonly called the cross product of the vectors Φ_{u_1} and Φ_{u_2}, and for $\Phi(u_1, u_2) = (u_1, u_2, f(u_1, u_2))$ it evaluates as[6]

$$e_3 - \frac{\partial f}{\partial u_1} e_1 - \frac{\partial f}{\partial u_2} e_2 = -\frac{\partial f}{\partial u_1} e_1 - \frac{\partial f}{\partial u_2} e_2 + e_3, \tag{28.7}$$

which is a positive multiple of the unit vector ν characterised by having its last component positive and being perpendicular to the graph defined by $u_3 = f(u_1, u_2)$. For any continuously differentiable

$$\Phi : [a_1, b_1] \times [a_2, b_2] \to \mathbb{R}^3$$

with Φ_{u_1} and Φ_{u_2} linearly independent, the vector defined by (28.6) is perpendicular to the plane spanned by Φ_{u_1} and Φ_{u_2}, and can be normalised by dividing it by its length, which we recognise as

$$\mathcal{M}_2(\Phi_{u_1}, \Phi_{u_2})$$

in view of Theorem 19.8. If we call this normalised vector ν, which in case of (28.7) is simply[7]

$$\nu = \frac{1}{\sqrt{1 + f_{u_1}^2 + f_{u_2}^2}} \left(-\frac{\partial f}{\partial u_1} e_1 - \frac{\partial f}{\partial u_2} e_2 + e_3 \right), \tag{28.8}$$

and consider $\tilde{v}_i$ as the i^{th} component of a vector field $\tilde{v} = \zeta v$ defined on $M \cap [a, b]$, with v a vector field on M, then

$$\int_M \nu \cdot \tilde{v} \; dS_2 =$$

$$\iint_{[a_1,b_1]\times[a_2,b_2]} \left(\tilde{v}_1 \begin{vmatrix} \frac{\partial \Phi_2}{\partial u_1} & \frac{\partial \Phi_3}{\partial u_1} \\ \frac{\partial \Phi_2}{\partial u_2} & \frac{\partial \Phi_3}{\partial u_2} \end{vmatrix} + \tilde{v}_2 \begin{vmatrix} \frac{\partial \Phi_3}{\partial u_1} & \frac{\partial \Phi_1}{\partial u_1} \\ \frac{\partial \Phi_3}{\partial u_2} & \frac{\partial \Phi_1}{\partial u_2} \end{vmatrix} + \tilde{v}_3 \begin{vmatrix} \frac{\partial \Phi_1}{\partial u_1} & \frac{\partial \Phi_2}{\partial u_1} \\ \frac{\partial \Phi_1}{\partial u_2} & \frac{\partial \Phi_2}{\partial u_2} \end{vmatrix} \right) \underbrace{du_1 \, du_2}_{du},$$

in which we use the short hand notation $du = du_1 \, du_2 = du_2 \, du_1$. We may be inclined to write this as

$$\int_\Phi \tilde{v}_1 \, dx_2 dx_3 + \tilde{v}_2 \, dx_3 dx_1 + \tilde{v}_3 \, dx_1 dx_2 = \int_\Phi \omega, \tag{28.9}$$

with

$$\omega = \tilde{v}_1 \, dx_2 dx_3 + \tilde{v}_2 \, dx_3 dx_1 + \tilde{v}_3 \, dx_1 dx_2,$$

[6]Denoting the partials with subscripts u_1 and u_2.

[7]Please allow the simultaneous use of both expressions in $f_{u_i} = \frac{\partial f}{\partial u_i}$.

using formal rules such as[8]

$$dx_2dx_3 = \begin{vmatrix} \frac{\partial x_2}{\partial u_1} & \frac{\partial x_3}{\partial u_1} \\ \frac{\partial x_2}{\partial u_2} & \frac{\partial x_3}{\partial u_2} \end{vmatrix} \underbrace{du_1du_2}_{\neq du} = \begin{vmatrix} \frac{\partial \Phi_2}{\partial u_1} & \frac{\partial \Phi_3}{\partial u_1} \\ \frac{\partial \Phi_2}{\partial u_2} & \frac{\partial \Phi_3}{\partial u_2} \end{vmatrix} \underbrace{du_1du_2}_{\neq du}.$$

We then have that (28.9) is equal to

$$\int_\Omega \nabla \cdot \tilde{v} = \int_\Omega \nabla \cdot \tilde{v}(x)\, dx = \iiint_\Omega \left(\frac{\partial \tilde{v}_1}{\partial x_1} + \frac{\partial \tilde{v}_2}{\partial x_2} + \frac{\partial \tilde{v}_3}{\partial x_3} \right) \underbrace{dx_1\, dx_2\, dx_3}_{dx},$$

which we will wish to write as an integral of the differential form

$$d\omega = \left(\frac{\partial \tilde{v}_1}{\partial x_1} + \frac{\partial \tilde{v}_2}{\partial x_2} + \frac{\partial \tilde{v}_3}{\partial x_3} \right) \underbrace{dx_1dx_2dx_3}_{\neq dx},$$

in which $dx_1dx_2dx_3$ is part of a 3-form and not to be read as $dx = dx_1\, dx_2\, dx_3$.

All of the above generalises[9] to arbitrary $N = n + 1$, e.g. we also have

$$\int_\Omega \left(\frac{\partial \tilde{v}_1}{\partial x_1} + \frac{\partial \tilde{v}_2}{\partial x_2} + \frac{\partial \tilde{v}_3}{\partial x_3} + \frac{\partial \tilde{v}_4}{\partial x_4} \right) dx \tag{28.10}$$

$$= \int_\Phi \tilde{v}_1\, dx_2dx_3dx_4 + \cdots \text{ (cyclicly permutated terms) } \cdots = \int_\Phi \omega,$$

using rules like

$$dx_2dx_3dx_4 = \begin{vmatrix} \frac{\partial x_2}{\partial u_1} & \frac{\partial x_3}{\partial u_1} & \frac{\partial x_4}{\partial u_1} \\ \frac{\partial x_2}{\partial u_2} & \frac{\partial x_3}{\partial u_2} & \frac{\partial x_4}{\partial u_2} \\ \frac{\partial x_2}{\partial u_3} & \frac{\partial x_3}{\partial u_3} & \frac{\partial x_4}{\partial u_3} \end{vmatrix} \underbrace{du_1du_2du_3}_{\neq du},$$

and (28.10) should be the integral of the 4-form

$$d\omega = \left(\frac{\partial \tilde{v}_1}{\partial x_1} + \frac{\partial \tilde{v}_2}{\partial x_2} + \frac{\partial \tilde{v}_3}{\partial x_3} + \frac{\partial \tilde{v}_4}{\partial x_4} \right) dx_1dx_2dx_3dx_4.$$

Clearly such a d-calculus requires rules such as $dx_idx_j = -dx_jdx_i$. I played with the formal rules that one might like to have in Chapter 24, see also the discussion after Theorem 10.12. The notation, used in Edwards, is cumbersome as the difference between spaces or no spaces between dx_i and dx_j is hardly visible, which is a reason to write $dx_i \wedge dx_j$ instead of dx_idx_j.

[8] Compare this to (24.9) in Section 24.2.

[9] This is why we put the unit vectors in the last row of the determimant in (28.6).

We conclude with the simplest but slightly confusing case, $n = 1$ and $N = 2$, when (28.6) should be replaced by

$$\begin{vmatrix} \frac{\partial \Phi_1}{\partial u} & \frac{\partial \Phi_2}{\partial u} \\ e_1 & e_2 \end{vmatrix} = \frac{\partial \Phi_2}{\partial u}\, e_1 - \frac{\partial \Phi_1}{\partial u}\, e_2, \tag{28.11}$$

which for

$$\Phi(u) = (u, f(u))$$
$$e_2 - f'(u)e_1,$$

and leads to

$$\int_M \nu \cdot \tilde{v}\; dS_1 = \int_{[a,b]} \left(-\tilde{v}_1 \frac{\partial \Phi_2}{\partial u} + \tilde{v}_2 \frac{\partial \Phi_1}{\partial u} \right) du = \iint_\Omega \left(\frac{\partial \tilde{v}_1}{\partial x_1} + \frac{\partial \tilde{v}_2}{\partial x_2} \right) dx_1\, dx_2,$$

in which we dropped the subscripts in a_1, b_1, u_1. Here we have

$$\omega = -\tilde{v}_1\, dx_2 + \tilde{v}_2\, dx_1 \quad \text{with} \quad d\omega = \left(\frac{\partial \tilde{v}_1}{\partial x_1} + \frac{\partial \tilde{v}_2}{\partial x_2} \right) dx_1 dx_2,$$

and

$$\int_{\partial\Omega} \omega = \int_\Omega d\omega.$$

In x, y notation for $\omega = p(x,y)dx + q(x,y)dy$ we have $d\omega = (q_x - p_y)dxdy$ and

$$\int_{\partial\Omega} p(x,y)dx + q(x,y)dy = \int_\Omega (q_x - p_y)dxdy, \tag{28.12}$$

which should make you wonder about

$$\int_\gamma p(x,y,z)dx + q(x,y,z)dy + r(x,y,z)dz,$$

for $\gamma : [a,b] \to \mathbb{R}^3$ as in Section 26.2. Section 28.2 below explores what's going on here.

Note that in all these examples the N-form $\omega = f(x)dx_1 \cdots dx_N$ integrated over the domain Ω should sensibly be agreed to give[10]

$$\int_\Omega \omega = \int_\Omega f(x)dx_1 \cdots dx_N = \int_\Omega f.$$

[10]Don't confuse this f with f in the local description of the boundary of a domain.

28.2 From Green's to Stokes' curl theorem

Now consider (28.5) as a local description of a manifold M and forget about Ω as being a domain with $M = \partial\Omega$. Instead let Ω be as in (26.1) with $N = 2$ and let M be the graph of $f : \Omega \to \mathbb{R}$. Assume for simplicity that $\partial\Omega$ is parameterised by a 1-periodic continuously differentiable function $t \to u(t) = (u_1(t), u_2(t))$. Then

$$t \xrightarrow{\gamma} (u_1(t), u_2(t), f(u_1(t), u_2(t))) \tag{28.1}$$

parameterises the "boundary"

$$\partial M = \{\underbrace{(u, f(u))}_{\Phi(u)} : u \in \partial\Omega\},$$

and

$$u \xrightarrow{\Phi} (u, f(u)) \tag{28.2}$$

parameterises M, with $u = (u_1, u_2) \in \Omega$.

For

$$F(x) = F_1(x)e_1 + F_2(x)e_2 + F_3(x)e_3$$

we introduce

$$\omega = F_1(x)dx_1 + F_2(x)dx_2 + F_3(x)dx_3$$

as in (26.7) and (26.8) and consider the integral

$$\int_{\partial M} \omega$$

as in (26.5). It evaluates as

$$\int_{\partial M} \omega = \int_0^1 (F_1(\gamma(t))\gamma_1'(t) + F_2(\gamma(t))\gamma_2'(t)) + F_3(\gamma(t))\gamma_3'(t))) \, dt$$

$$= \int_0^1 (F_1(u(t), f(u(t))u_1'(t) + F_3(u(t), f(u(t)))f_{u_1}(u(t)u_1'(t)) \, dt$$

$$+ \int_0^1 (F_2(u(t), f(u(t))u_2'(t) + F_3(u(t), f(u(t)))f_{u_2}(u(t)u_2'(t)) \, dt =$$

$$\int_{\partial\Omega} \zeta = \int_\Omega d\zeta, \tag{28.3}$$

in which

$$\zeta = \left(F_1 + F_3\frac{\partial f}{\partial u_1}\right) du_1 + \left(F_2 + F_3\frac{\partial f}{\partial u_2}\right) du_2$$

Next we compute

$$d\zeta = \left(\frac{\partial F_1}{\partial x_2} + \frac{\partial F_1}{\partial x_3}\frac{\partial f}{\partial u_2} + \frac{\partial F_3}{\partial x_2}\frac{\partial f}{\partial u_1} + \frac{\partial F_3}{\partial x_3}\frac{\partial f}{\partial u_2}\frac{\partial f}{\partial u_1} + F_3\frac{\partial^2 f}{\partial u_2 \partial u_1}\right) du_2 du_1$$

$$+ \left(\frac{\partial F_2}{\partial x_1} + \frac{\partial F_2}{\partial x_3}\frac{\partial f}{\partial u_1} + \frac{\partial F_3}{\partial x_1}\frac{\partial f}{\partial u_2} + \frac{\partial F_3}{\partial x_3}\frac{\partial f}{\partial u_1}\frac{\partial f}{\partial u_2} + F_3\frac{\partial^2 f}{\partial u_1 \partial u_2}\right) du_1 du_2,$$

which in view of $du_2 du_1 = -du_1 du_2$ reduces to

$$d\zeta = \phi(u_1, u_2) du_1 du_2 \tag{28.4}$$

with $\phi(u_1, u_2)$ given by

$$\phi = -\underbrace{\left(\frac{\partial F_3}{\partial x_2} - \frac{\partial F_2}{\partial x_3}\right)}_{G_1}\frac{\partial f}{\partial u_1} - \underbrace{\left(\frac{\partial F_1}{\partial x_3} - \frac{\partial F_3}{\partial x_1}\right)}_{G_2}\frac{\partial f}{\partial u_2} + \underbrace{\left(\frac{\partial F_2}{\partial x_1} - \frac{\partial F_1}{\partial x_2}\right)}_{G_3} \tag{28.5}$$

$$= -G_1\frac{\partial f}{\partial u_1} - G_2\frac{\partial f}{\partial u_2} + G_3.$$

You should note that the *second order derivatives* of (28.2) are dropouts in the calculations that lead to (28.5).

Now compare (28.5) to ν in (28.8) and recall that for Φ given by (28.2) we know that

$$\mathcal{M}_2(\Phi_{u_1}, \Phi_{u_2}) = \sqrt{1 + f_{u_1}^2 + f_{u_2}^2}.$$

Summing up we thus have

$$\int_{\partial M} (F \cdot \tau)\, dS_1 =$$

(hello forms)

$$\int_{\partial M} \omega = \int_{\partial \Omega} \zeta = \int_{\Omega} d\zeta = \int_{\Omega} \underbrace{\phi\, du_1 du_2}_{d\zeta} =$$

(goodbye forms)

$$\int_{\Omega} \phi = \int_{\Omega} (G \cdot \nu)\, \mathcal{M}_2(\Phi_{u_1}, \Phi_{u_2}) = \int_{M} (G \cdot \nu)\, dS_2,$$

with G derived from F as indicated in (28.5), and commonly denoted as $G = \nabla \times F$, i.e.

$$\int_{\partial M} (F \cdot \tau)\, dS_1 = \int_{M} (G \cdot \nu)\, dS_2 \quad \text{with} \quad G = \nabla \times F, \tag{28.6}$$

using the parameterisations as indicated[11]. But don't say goodbye:

28.3 Pullbacks and the action of d

We already saw in the reasoning from (26.5) to (26.6) that d acting on a C^1-function $f = f(x_1, \ldots, x_N)$ produces a 1-form

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_N} dx_N = \frac{\partial f}{\partial x_i} dx_i, \tag{28.7}$$

using the convention that we sum over repeated indices. With $f(x_1, \ldots, x_N)$ replaced by $u(x, y)$ this is (24.3) in Section 24.1. There I played with the d-algebra that emerges whenever you do integration using formal notations such as (10.6), which is just (28.7) with $n = 1$ and $f(x_1, \cdots, x_N)$ replaced by $F(x)$.

Now consider a parameterisation $x = \Phi(u)$ as in (C) in Section 27.2. We use Φ to pull back expressions with x and $dx_1, \ldots, dx_N$ back to expressions with u and $du_1, \ldots, du_n$, in a way that is consistent with the discussion leading to (28.9) and the formal rules that emerge in the calculations to do so. Thus we certainly want to deal with

$$f(x) = \phi(u) \quad \text{via} \quad x = \Phi(u). \tag{28.8}$$

A mathematician's way to do so is to introduce

$$\phi = \Phi^*(f) = f \circ \Phi, \tag{28.9}$$

the pullback of f via Φ, which then also provides us with

$$d\phi = \frac{\partial \phi}{\partial u_1} du_1 + \cdots + \frac{\partial \phi}{\partial u_n} du_n. \tag{28.10}$$

If g is another function of x then clearly

$$\Phi^*(f + g) = \Phi^*(f) + \Phi^*(g), \quad \Phi^*(fg) = \Phi^*(f)\Phi^*(g),$$

which suggests as a definition of the pullback of a 1-form $\omega = f_i dx_i$ that

$$\Phi^*(f_i dx_i) = \underbrace{\Phi^*(f_i)}_{\phi_i} \Phi^*(dx_i), \tag{28.11}$$

[11] Figure out that annoying $\pm$ afterwards? We have, depending on the parameterisation:

$$\int_{\partial M} (F \cdot \tau)\, dS_1 = \pm \int_M (G \cdot \nu)\, dS_2.$$

in which $\phi_i(u) = f_i(\Phi(u))$ as before. This definition would imply that

$$\Phi^*(df) = \underbrace{\frac{\partial f}{\partial x_i}((\Phi(u))}_{\Phi^*(D_i f)(u)}\ \Phi^*(dx_i). \tag{28.12}$$

Note that $D_i f$ as notation for the i^{th} first order partial derivative of f has the advantage of not using the variable x in the notation.

On the other hand (28.10) implies via the chain rule that

$$d(\Phi^*(f)) = \frac{\partial}{\partial u_j}(f(\Phi(u))\, du_j = \frac{\partial f}{\partial x_i}(\Phi(u))\frac{\partial \Phi_i}{\partial u_j}\, du_j, \tag{28.13}$$

and comparing to (28.12) we see that, if we define the pullback of dx_i under Φ to be

$$\Phi^*(dx_i) = \frac{\partial \Phi_i}{\partial u_j}\, du_j, \tag{28.14}$$

it follows that

$$\Phi^*(df) = \Phi^*(df). \tag{28.15}$$

The definition of $\Phi^*(dx_i)$ by (28.14) is just a formalisation of the familiar "rule"

$$dx_i = \frac{\partial x_i}{\partial u_j}\, du_j$$

for expressing dx_i in u, $du_1, \ldots, du_n$, just like expressing $f(x)$ in u via (28.8) is formalised by (28.9). It implies that the pullback of the 1-form in (28.11) evaluates as

$$\underbrace{\Phi^*(f_i dx_i) = \phi_i\, \frac{\partial \Phi_i}{\partial u_j}\, du_j}_{with\ \ \phi_i(u)=f_i(\Phi(u))} = f_i(\Phi(u))\, \frac{\partial \Phi_i}{\partial u_j}\, du_j = f_i(\Phi(u))\, D_j\Phi_i(u)\, du_j. \tag{28.16}$$

Next we observe that d acting on the resulting 1-form in (28.16) may be evaluated, using the chain rule and $du_k du_j = -du_j du_k$, as

$$d(\Phi^*(f_i dx_i)) = d(f_i(\Phi(u))\, \frac{\partial \Phi_i}{\partial u_j}\, du_j) = \frac{\partial}{\partial u_k}(f_i(\Phi(u))\, \frac{\partial \Phi_i}{\partial u_j})\, du_k du_j$$

$$= (\frac{\partial}{\partial u_k}(f_i(\Phi(u)))\, \frac{\partial \Phi_i}{\partial u_j}\, du_k du_j + f_i(\Phi(u))\ \underbrace{\frac{\partial^2 \Phi_i}{\partial u_k \partial u_j}\, du_k du_j}_{zero\ the\ hero!} =$$

$$\frac{\partial f_i}{\partial x_k}(\Phi(u))\, \frac{\partial \Phi_k}{\partial u_k}\, \frac{\partial \Phi_i}{\partial u_j}\, du_k\, du_j = \Phi^*(D_k f_i)\, \frac{\partial \Phi_k}{\partial u_k}\, \frac{\partial \Phi_i}{\partial u_j}\, du_k\, du_j, \tag{28.17}$$

in which we used

$$d(f_i dx_i) = \frac{\partial f_i}{\partial x_k}\, dx_k\, dx_i \tag{28.18}$$

in the u-variables. Recall that this was the definition[12] in Section 24.1 of the action of d on 1-forms. With

$$\Phi^*(f_{ij} dx_i dx_j) = \underbrace{\Phi^*(f_{ij})}_{\phi_{ij}}\, \Phi^*(dx_i dx_j) \tag{28.19}$$

as the obvious defining analog of (28.11), we have that

$$\Phi^*(d(f_i dx_i)) = \Phi^*(\frac{\partial f_i}{\partial x_k}\, dx_k dx_i) = \Phi^*(D_k f_i)\, \Phi^*(dx_k dx_i). \tag{28.20}$$

Comparing to (28.20) to (28.17) it follows that

$$\Phi^*(d(f_i dx_i)) = d(\Phi^*(f_i dx_i)), \tag{28.21}$$

provided we define

$$\Phi^*(dx_k dx_i) = \underbrace{\frac{\partial \Phi_k}{\partial u_k}\, \frac{\partial \Phi_i}{\partial u_j}\, du_k\, du_j}_{sum\ over\ 1\le k,j\le n} = \underbrace{(\frac{\partial \Phi_k}{\partial u_k}\, \frac{\partial \Phi_i}{\partial u_j} - \frac{\partial \Phi_k}{\partial u_k}\, \frac{\partial \Phi_i}{\partial u_j})\, \underline{du_k\, du_j}}_{sum\ over\ 1\le k<j\le n}$$

$$= \frac{\partial(\Phi_k, \Phi_i)}{\partial u_k \partial u_j}\, \underline{du_k\, du_j}, \tag{28.22}$$

in which the underline indicates that we sum over all k, j with $1 \le k < j \le n$. Just as in (28.15) we see that the actions of d and Φ^* commute.

Note that the second order derivatives have disapperared in (28.17). The derivation is typically done under the assumption that $\Phi \in C^2$, also in Edwards, and an additional analysis argument is needed[13] to give meaning to the results if Φ is only in C^1, because the determinants in (28.22) are exactly the determinants that showed up in (28.6) and the subsequent derivation of (28.9), where effectively $dx = dx_1\, dx_2\, dx_3$ is first replaced by a 3-form $dx_1 dx_2 dx_3$ pulled back to a 2-form $du_1 du_2$, which in turn is replaced by $du = du_1\, du_2$ again.

The step by step generalisation to the action of d and Φ^* on k-forms of any order k is easily made once the reasoning above is understood. For any k-form

$$\omega = f_{i_1,\dots,i_k} dx_{i_1} \cdots dx_{i_k}$$

[12] Recall the choice to set $ddx_i = 0$, leading to $dd\omega = 0$ for any form ω.

[13] Using approximation arguments.

we have

$$\Phi^*(d\omega) = d(\Phi^*(\omega)) \tag{28.23}$$

Every such form may be written as

$$\omega = f_{i_1,\ldots,i_k} dx_{i_1} \cdots dx_{i_k} = \tilde{f}_{i_1,\ldots,i_k} \underline{dx_{i_1} \cdots dx_{i_k}}, \tag{28.24}$$

where in the second expression we sum only over those $i_1, \ldots, i_k$ for which $1 \leq i_1 < \cdots < i_k \leq N$. For instance

$$\omega = f_{ij}\, dx_i dx_j = \underbrace{(f_{ij} - f_{ji})}_{\tilde{f}_{ij}}\, \underline{dx_i dx_j},$$

but this is not compulsory, as the examples

$$\omega = f_1\, dx_1 + f_2\, dx_2 + f_3\, dx_3$$

with cyclic notation for

$$d\omega = \underbrace{(\frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3})}_{g_1}\, dx_2 dx_3 + \underbrace{(\frac{\partial f_1}{\partial x_3} - \frac{\partial f_3}{\partial x_1})}_{g_2}\, dx_3 dx_1 + \underbrace{(\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2})}_{g_3}\, dx_1 dx_2$$

and

$$\zeta = g_1\, dx_2 dx_3 + g_2\, dx_3 dx_1 + g_3\, dx_1 dx_2$$

with

$$d\zeta = (\frac{\partial g_1}{\partial x_1} + \frac{\partial g_2}{\partial x_2} + \frac{\partial g_3}{\partial x_3})\, dxdydz$$

in Section 28.4 show.

Finally we observe that if we put the coefficients f_1, f_2, f_3 of this ω in a vector $F = f_1 e_1 + f_2 e_2 + f_3 e_3$ and the coefficients g_1, g_2, g_3 in this cyclic representation of $d\omega$ in a vector $G = g_1 e_1 + g_2 e_2 + g_3 e_3$, we obtain that

$$G = \nabla \times F,$$

the curl of F, whereas with the coefficients of η we obtain the coefficient of $d\zeta$ as

$$\frac{\partial g_1}{\partial x_1} + \frac{\partial g_2}{\partial x_2} + \frac{\partial g_3}{\partial x_3} = \nabla \cdot G,$$

the divergence of G. These appear in the Gauss divergence and the Stokes curl theorems for vectorfields in $\mathbb{R}^3$ in Section 28.4 below[14]. The general statement is also called Stokes Theorem. It has both theorems in $\mathbb{R}^3$ and Green's Theorem in $\mathbb{R}^3$ as special cases.

[14]The statement that $dd\omega = 0$ corresponds to the div of a curl being always zero:

$$\nabla \cdot \nabla \times F = 0.$$

28.4 From Gauss' to general Stokes' Theorem

From Section 28.1 and partitions of unity arguments we have that for $\Omega \subset \mathbb{R}^N = \mathbb{R}^{n+1}$ open and bounded, with $\partial\Omega$ a compact $(N-1)$-dimensional C^1-manifold, and in every $p \in M$, after renumbering, a local description of $\Omega \cap [a,b]$ given by

$$a_N \leq x_N < f(x_1, \ldots, x_n) < b_N$$

or

$$a_N < f(x_1, \ldots, x_n) \leq x_N < b_N,$$

with $f \in C^1$ and $p \in (a,b)$, that there exists a globally defined normal vectorfield $\nu : \partial\Omega \to \mathbb{R}^N$ with $\nu(p)$ pointing out of Ω in every patch as above. For every continuously differentiable $V : \Omega \to \mathbb{R}^N$ it now holds that

$$\int_\Omega \nabla \cdot V = \int_{\partial\Omega} \nu \cdot V \, dS_{N-1}, \tag{28.25}$$

and this statement is called the Gauss Divergence Theorem.

We now use the reformulation with differential forms and pullbacks of forms with $\Phi : \mathbb{R}^{n+1} \to \mathbb{R}^N$ with $N > n+1$ to formulate Stokes' Theorem for integral n-forms over $\Phi(M)$ considered as the boundary of $\Phi(\Omega)$, first for $n+1 = 2$ and $N = 3$. So let

$$\omega = f_1(x)dx_1 + f_2(x)dx_2 + f_3(x)dx_3 \tag{28.26}$$

and $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$. Then

$$\Phi^*(dx_1) = \frac{\partial\Phi_1}{\partial u_1}du_1 + \frac{\partial\Phi_1}{\partial u_2}du_2; \quad \Phi^*(dx_2) = \frac{\partial\Phi_2}{\partial u_1}du_1 + \frac{\partial\Phi_2}{\partial u_2}du_2;$$

$$\Phi^*(dx_3) = \frac{\partial\Phi_3}{\partial u_1}du_1 + \frac{\partial\Phi_3}{\partial u_2}du_2,$$

and with $\phi_1 = \Phi^* f_1$, $\phi_2 = \Phi^* f_2$, $\phi_3 = \Phi^* f_3$ we have

$$\Phi^*(F) = (\phi_1\frac{\partial\Phi_1}{\partial u_1} + \phi_2\frac{\partial\Phi_2}{\partial u_1} + \phi_3\frac{\partial\Phi_3}{\partial u_1})\,du_1 + (\phi_1\frac{\partial\Phi_1}{\partial u_2} + \phi_2\frac{\partial\Phi_2}{\partial u_2} + \phi_3\frac{\partial\Phi_3}{\partial u_2})\,du_2$$

$$= p_1(u_1,u_2)du_1 + p_2(u_1,u_2)du_2 = \zeta,$$

a 1-form that can be integrated over $M = \partial\Omega$, and to which (28.12) applies, whence

$$\int_{\partial\Omega} \zeta = \int_{\partial\Omega} p_1(u_1,u_2)du_1 + p_2(u_1,u_2)du_2 = \int_\Omega (\frac{\partial p_2}{\partial u_1} - \frac{\partial p_1}{\partial u_2})du_1du_2 = \int_\Omega d\zeta. \tag{28.27}$$

Observe that the second equality in (28.27) holds in view of (28.12), which is a rewritten version of (28.25) with $N = 2$, while the first and the third merely substitute $\omega = p_1 du_1 + p_2 du_2$ and evaluate $d\omega$ according to (28.18).

We need

$$\int_{\partial\Omega} \zeta = \int_{\partial\Omega} \Phi^*\omega = \int_{\Phi(\partial\Omega)} \omega, \tag{28.28}$$

and

$$\int_{\Omega} d\zeta = \int_{\Omega} d\,\Phi^*\omega = \int_{\Omega} \Phi^*(d\omega) = \int_{\phi(\Omega)} d\omega \tag{28.29}$$

to conclude for ω given by (28.26) that

$$\int_{dS} \omega = \int_{dS} f_1 dx_1 + f_2 dx_2 + f_3 dx_3 = \int_S d\omega, \tag{28.30}$$

in which $S = \Phi(\Omega)$. It is the last equality in each of (28.28) and (28.29) that has to be checked, the other equalities follow from our d-algebra and the commutation of d and Φ^*.

Let us once more spell out the d-algebra by which (28.18) evaluates as

$$d\omega = (\frac{\partial f_1}{\partial x_1}dx_1 + \frac{\partial f_1}{\partial x_2}dx_2 + \frac{\partial f_1}{\partial x_3}dx_3)dx_1$$

$$+(\frac{\partial f_2}{\partial x_1}dx_1 + \frac{\partial f_2}{\partial x_2}dx_2 + \frac{\partial f_2}{\partial x_3}dx_3)dx_2$$

$$+(\frac{\partial f_3}{\partial x_1}dx_1 + \frac{\partial f_3}{\partial x_2}dx_2 + \frac{\partial f_3}{\partial x_3}dx_3)dx_3 =$$

$$\frac{\partial f_1}{\partial x_2}dx_2dx_1 + \frac{\partial f_2}{\partial x_1}dx_1dx_2 + \frac{\partial f_1}{\partial x_3}dx_3dx_1 + \frac{\partial f_3}{\partial x_1}dx_1dx_3 + \frac{\partial f_2}{\partial x_3}dx_2 + \frac{\partial f_3}{\partial x_2}dx_2dx_3$$

$$= (\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2})dx_1dx_2 + (\frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3})dx_2dx_3 + (\frac{\partial f_1}{\partial x_3} - \frac{\partial f_3}{\partial x_1})dx_3dx_1$$

$$(\frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3})dx_2dx_3 + (\frac{\partial f_1}{\partial x_3} - \frac{\partial f_3}{\partial x_1})dx_3dx_1 + (\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2})dx_1dx_2$$

$$= g_1 dx_2dx_3 + g_2 dx_3dx_1 + g_3 dx_1dx_2.$$

Comparing to (28.9) we recognise for $F(x) = f_1(x)e_1 + f_2(x)e_2 + f_3(x)e_3$ that

$$\int_{dS} f_1 dx_1 + f_2 dx_2 + f_3 dx_3 =$$

$$\int_S (\frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3})dx_2dx_3 + (\frac{\partial f_1}{\partial x_3} - \frac{\partial f_3}{\partial x_1})dx_3dx_1 + (\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2})dx_1dx_2$$

$$= \int_S G \cdot \nu = \int_S (\nabla \times F) \cdot \nu, \tag{28.31}$$

in which $g_1(x)e_1 + g_2(x)e_2 + g_3(x)e_3 = G(x) = \nabla \times F$ and ν is the normal vector on $S = \Phi(\Omega)$ defined by (28.6).

It thus remains to check the two analytical statements

$$\int_{\partial\Omega} \Phi^*\omega = \int_{\Phi(\partial\Omega)} \omega \quad \text{and} \quad \int_\Omega \Phi^*(d\omega) = \int_{\Phi(\Omega)} d\omega, \tag{28.32}$$

which complement the d-algebra presented above, and which are both of the form

$$\int_M \Phi^*\omega = \int_{\Phi(M)} \omega, \tag{28.33}$$

with respectively $M = \partial\Omega$ and $M = \Omega$. For this we need again Section 23 combined with the usual localisations via partitions of unity. Not very hard but still to be done.

It will be convenient here to have $\Phi(M)$ described by compositions of Φ and patches of M, see the remark at the end of Section 29.4. Also, we still have to deal with integrals over manifolds with boundaries, to obtain

$$\int_{\Phi(\partial\Omega)} \omega = \int_{\Phi(\Omega)} d\omega, \tag{28.34}$$

as the final result in which $M = \Phi(\Omega)$ is a manifold with boundary $\partial M = \Phi(\partial\Omega)$, with $\Omega \in \mathbb{R}^n$ as described at the beginning of this section, Φ a continuously differentiable injective map from Ω to $\mathbb{R}^N$ with Jacobian matrix of rank n throughout Ω, and ω an n-form with continuously differentiable coefficients. Generalisations to piecewise C^1-boundaries then still have to be discussed.

28.5 More exercises

Let Ω be the open unit disk. Then its boundary $\partial\Omega$ is the circle defined by

$$x^2 + y^2 = 1.$$

Graph parameterisations such as

$$x \to (x, \sqrt{1-x^2}), \quad x \to (x, -\sqrt{1-x^2}), \tag{28.35}$$

$$y \to (\sqrt{1-y^2}, y), \quad y \to (-\sqrt{1-y^2}, y)$$

are ugly for calculations. Much nicer and more in common is of course

$$\phi \to (\cos\phi, \sin\phi), \tag{28.36}$$

but parameterisations obtained from substitutions like $y = tx$ in the defining equation $x^2 + y^2 = 1$ for $\partial\Omega$ are also handy: from $x^2 + t^2x^2 = 1$ we have

$$x = \frac{1}{\sqrt{1+t^2}},\ y = \frac{t}{\sqrt{1+t^2}} \quad \text{and} \quad x = -\frac{1}{\sqrt{1+t^2}},\ y = -\frac{t}{\sqrt{1+t^2}}$$

parameterising two semicircles if we let t run from $-\infty$ to $+\infty$. With

$$t = \frac{s}{1-s} \tag{28.37}$$

this gives

$$x = \frac{1-s}{\sqrt{1-2s+2s^2}},\ y = \frac{s}{\sqrt{1-2s+2s^2}}$$

parameterising $\{(x,y) \in \mathbb{R}^2 : x \geq 0,\ y \geq 0,\ x^2 + y^2 = 1\}$ with $s \in [0,1]$.

Exercise 28.1. Use the t-parameterisations above to calculate the area of the unit disk via integrals such as $\int x dy$ of $\int y dx$ over $\partial\Omega$. You should get and evaluate integrands[15] like

$$\frac{t^2}{(1+t^2)^2} = \frac{1}{1+t^2} - \frac{1}{(1+t^2)^2}.$$

Exercise 28.2. Referring to line integral notation with 1-forms, consider the form

$$\omega = (a_{20}x^2 + a_{11}xy + a_{02}y^2)dx + (b_{20}x^2 + b_{11}xy + b_{02}y^2)dy$$

and evaluate $\int_{\partial\Omega} \omega$ for $\Omega = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$ with $\partial\Omega$ parameterised such that (28.11) defines a vector pointing out of Ω.

Exercise 28.3. Same as Exercise 28.2 but with

$$\omega = (a_{30}x^3 + a_{21}x^2y + a_{12}xy^2 + a_{03}y^3)dx + (b_{30}x^3 + b_{21}x^2y + b_{12}xy^2 + b_{03}y^3)dy$$

Which coefficients disappear in the calculations? Generalise to the obvious n^{th} order case.

[15] Recall $\int_{-\infty}^{\infty} \frac{1}{1+t^2}\,dt = \pi$, $\int_{-\infty}^{\infty} \frac{1}{(1+t^2)^2}\,dt = \frac{\pi}{2}$, $\int_{-\infty}^{\infty} \frac{1}{(1+t^2)^3}\,dt = \frac{3\pi}{8}$, ...

Edwards has a nice exercise about Descartes' Folium from which I lifted the $y = tx$-trick above. It allows to find the solutions of

$$F(x, y) = x^3 + y^3 - 3xy = 0, \tag{28.38}$$

in the form

$$x = x(t) = \frac{3t}{1+t^3};\ y = y(t) = \frac{3t^2}{1+t^3}, \tag{28.39}$$

with $t \in (0, \infty)$, $t \in (-1, 0)$ and $t \in (-\infty, -1)$ giving the smooth parts of the curve. The origin $(0, 0)$ is the intersection of two solution curves, one given by (28.39) with $t \in (-1, 1)$, the other by (28.39) with x and y interchanged. Exercise 2.3 in Chapter V of Edwards is about

$$\Omega = \{(x, y) \in \mathbb{R}^2 :\ x > 0,\ y > 0,\ F(x, y) = x^3 + y^3 - 3xy < 0\}. \tag{28.40}$$

with $\partial\Omega$ given by (28.39) and $t \in [0, \infty)$. You should examine the graphs of x and y as functions of t in (28.39). You can get the area of Ω as

$$-\int_0^\infty y(t)x'(t)\,dt = \int_0^\infty x(t)y'(t)\,dt, \tag{28.41}$$

or the average of the two integrals, which may turn out to be easier, using Green's Theorem the way we derived it. Edwards tells you to cut the folium along the diagonal $y = x$, in which case you have the boundary consisting of two curves, the part described by (28.39) with $0 \le t \le 1$, and the diagonal part given by $y = x = t$ with $0 \le t \le \frac{3}{2}$, which you should parameterise as $y = x = \frac{3}{2} - t$ if you think about it. Still, I wonder whether Edwards actually did the exercise:

Exercise 28.4. Substitute $y = t^{\frac{1}{3}}x$ in the equation for the folium to get x and y in terms of t and evaluate (28.41) above to obtain the value $\frac{3}{2}$ for the area of $\{(x, y) \in \mathbb{R}^2 :\ x > 0,\ y > 0,\ x^3 + y^3 - 3xy < 0\}$.

Exercise 28.5. As Exercise 28.4 but use (28.37) to get the boundary parameterised with $0 \le s \le 1$.

In the last exercise you see that the boundary of (28.40) is actually given by one single parameterisation with the parameter s in the unit interval $[0, 1]$, with $s = 0$ and $s = 1$ both mapped to the origin where the condition for the

local description as used in Section 27.2 fails. The same issue occurs in the trivial case of Exercise 21.10.

Note that (28.40) is a special case of an obvious general question with two parameters, these being $p = 3$ and $n = 2$ here[16]. Dropping the coefficient of xy we have for general $p > 2$ that

$$x = s^{\frac{1}{p(p-2)}}\,(1-s)^{\frac{p-1}{p(p-2)}}\,;\; y = s^{\frac{p-1}{p(p-2)}}\,(1-s)^{\frac{1}{p(p-2)}}\,, \tag{28.42}$$

parameterises the loop in the solution set of $x^p + y^p = xy$, with

$$x\frac{dy}{ds} - y\frac{dx}{ds} = \frac{1}{p}\,s^{\frac{1}{p-2}-1}(1-s)^{\frac{1}{p-2}-1}, \tag{28.43}$$

which looks much better than the individual terms $x\frac{dy}{ds}$ and $y\frac{dx}{ds}$. With the β-function[17] defined by

$$B(x,y) = \int_0^1 s^{x-1}(1-s)^{y-1}\,ds,$$

the area surrounded by $[0,1] \ni s \to (x(s), y(s))$, the loop in

$$x^p + y^p = xy \tag{28.44}$$

is thus equal to

$$A_p = \frac{1}{2p}\,B(\frac{1}{p-2}, \frac{1}{p-2}), \tag{28.45}$$

which gives $\frac{1}{6}$ for $p = 3$ and differs from Exercise 28.4 by a factor 3^2, consistent with (28.39).

Note that in deriving (28.43) from (28.42) you may get lost if you don't introduce

$$\alpha = \frac{1}{p(p-2)} \quad \text{and} \quad \beta = \frac{p-1}{p(p-2)} = (p-1)\alpha$$

and continue your calculations with α and β. I also suggest to write derivatives such as

$$\frac{d}{ds}\,s^\alpha(1-s)^\beta = (\frac{\alpha}{s} - \frac{\beta}{1-s})\,s^\alpha(1-s)^\beta = (\alpha - (\alpha+\beta)s)\,s^{\alpha-1}(1-s)^{\beta-1},$$

which will help you to factor out common factors when such expressions have to be combined later on, as you will notice if you tackle this question: how about the volume V_p in $\{(x,y,z) \in \mathbb{R}^3 : x \geq 0,\, y \geq 0,\, z \geq 0\}$ surrounded by $x^p + y^p + z^p = xyz$ when $p > 3$?

[16] See Exercise 28.12.

[17] More on the β-function in [HM].

Exercise 28.6. Substitute $y = s^{\frac{1}{p}}x$ and $z = t^{\frac{1}{p}}x$ in $x^p + y^p + z^p = xyz$ to obtain a parameterisation of the solutions with $x, y, z > 0$ in the form

$$x = s^\alpha t^\alpha (1+s+t)^{-p\alpha}, \quad y = s^{(p-2)\alpha} t^\alpha (1+s+t)^{-p\alpha}, \quad z = s^\alpha t^{(p-2)\alpha} (1+s+t)^{-p\alpha},$$

and evaluate

$$xdydz = x(\frac{\partial y}{\partial s}\frac{\partial z}{\partial t} - \frac{\partial y}{\partial t}\frac{\partial z}{\partial s})$$

as xyz times a factor that you have to computer carefully, to find the correct double integral in s and t that gives the desired volume. The integral is the difference of two similar terms each of which is st to some power times $(1 + s + t)$ to some power. Substituting $t = (1 + s)x$ both integrals reduce to products of single integrals that reduce to β-functions again.

Just in case, I arrived via

$$xyz = \frac{(st)^{\frac{1}{p-3}}}{(1+s+t)^{\frac{3}{p-3}}}$$

and

$$\frac{1}{yz}(\frac{\partial y}{\partial s}\frac{\partial z}{\partial t} - \frac{\partial y}{\partial t}\frac{\partial z}{\partial s}) = \frac{1}{p^2(p-3)st}\left(\frac{p}{1+s+t} - 1\right)$$

at

$$\frac{1}{p(p-3)} \underbrace{\int_0^\infty \int_0^\infty \frac{(st)^{\frac{1}{p-3}-1}\, dsdt}{(1+s+t)^{\frac{p}{p-3}}}}_{S(\frac{1}{p-3}, \frac{p}{p-3})} - \frac{1}{p^2(p-3)} \underbrace{\int_0^\infty \int_0^\infty \frac{(st)^{\frac{1}{p-3}-1}\, dsdt}{(1+s+t)^{\frac{3}{p-3}}}}_{S(\frac{1}{p-3}, \frac{3}{p-3})}.$$

These integrals are known. With

$$B(a,b) = \int_0^1 s^{a-1}(1-s)^{b-1}\, ds$$

we have[18]

$$T(a,b) = \int_0^\infty \frac{s^{a-1}\, ds}{(1+s)^b} = B(a, b-a)$$

and[19]

$$S(a,b) = \int_0^\infty \int_0^\infty \frac{(st)^{a-1}\, dsdt}{(1+s+t)^b} = T(a,b)T(a,b-a),$$

so V_p can be expressed in p via β-functions. It should lead to what we get in Exercise 28.11, which is really nice[20].

[18] Via $s = \frac{t}{1-t}$, a substitution I avoided for (28.6).

[19] Via $t = (1+s)\tau$.

[20] There were mistakes in an earlier version and then it did not, but now it does.

Exercise 28.7. How general is the $y = tx$-trick in $\mathbb{R}^2$? Let $F : \mathbb{R}^2 \to \mathbb{R}$ be continuously differentiable, and suppose that $F(x_0, y_0)$ for some $(x_0, y_0) \in \mathbb{R}^2$ with $x_0 \neq 0$. Define t_0 by $y_0 = t_0 x_0$ and apply the implicit function theorem to derive a condition that guarantees the existence of a C^1-solution curve of the form $t \to (x(t), y(t) = (x(t), tx(t))$ defined on an t-interval which has t_0 as an interior point.

Don't forget you want to have nonzero speed, which is a second condition on top of the usual condition from the the implicit function theorem. The latter condition will involve a simple combination of x, y, F_x, F_y in (x_0, y_0) with a clear (but local) geometric interpretation.

Verify that in the end the nonzero speed condition follows from $x \neq 0$ and the condition from the implicit function theorem. Note that if $(x_0, y_0) \neq (0, 0)$ you always realise at least one of $t \to (x(t), y(t) = (x(t), tx(t))$ and $t \to (x(t), y(t) = (ty(t), y(t))$ if this condition is satisfied. Relate your results to polar coordinates.

Exercise 28.8. Verify that computing the area of (28.40) using polar coordinates is even a bigger pain than using the $y = tx$-trick.

Exercise 28.9. In Exercise 28.7 you must have computed the time derivatives of $x(t)$ and $y(t) = tx(t)$. Verify[21] that the derivative of

$$\frac{y(t)}{x(t)}$$

is what it should be, and that the area of such a curve parameterised by $t \in \mathbb{R}$ with $(x(t), y(t) \to (0, 0)$ as $t \to 0$ and $t \to \infty$ is given by[22]

$$\frac{1}{2} \int_0^\infty x(t)^2 \, dt,$$

and compute again the area in Exercise 28.4 from the formula for $x(t)$ in (28.39).

Exercise 28.10. Verify (28.45) by putting $y = tx$ in (28.44), solve for x, and set $t^p = s$ in the integral you get from Exercise 28.9 and convert to β-functions.

[21] You should have got $\dot{x} = -\frac{x^2 F_y}{xF_x + yF_y}$, $\dot{y} = \frac{x^2 F_x}{xF_x + yF_y}$

[22] Compare this to a similar formula with polar coordinates.

Exercise 28.11. See Exercise 28.9. How would $F(x,y,z)=0$ lead to

$$\frac{1}{3}\int_0^\infty \int_0^\infty x(s,t)^3\,dsdt?$$

Hint: in relation to

$$x^p+y^p+z^p=xyz$$

and for

$$x=x(s,t)=\left(\frac{st}{1+s^p+t^p}\right)^{\frac{1}{p-3}}$$

this integral is equal to[23]

$$V_p=\frac{1}{3p^2}\,B(\frac{1}{p-3},\frac{1}{p-3})\,B(\frac{1}{p-3},\frac{2}{p-3}),$$

and you might see a pattern emerge.

Exercise 28.12. Let $p>4$. The 4-dimensional measure of the bounded open set in $\mathbb{R}^4$ with all coordinates positive and bounded by

$$x_1^p+x_2^p+x_3^p+x_4^p=x_1x_2x_3x_4$$

is

$$\frac{1}{4p^3}\,B(\frac{1}{p-4},\frac{1}{p-4})\,B(\frac{1}{p-4},\frac{2}{p-4})\,B(\frac{1}{p-4},\frac{3}{p-4}),$$

and likewise for

$$\sum_{j=1}^n x_j^p=\Pi_{j=1}^n x_j$$

in $\mathbb{R}^n$ for $p>n$.

[23]Earlier mistakes have have been corrected....

29 Cut-off functions and partitions of unity

This chapter explains the basic tools for cutting up functions in smaller parts which are localised. This involves two tricks, each of which you can play with.

The first trick concerns an open set $O \subset \mathbb{R}^N$ and a compact subset $K \subset O$ which should be non-empty[1]. Then every $a \in K$ is contained in an open ball B centered at a such that the closed ball with the same center but twice the radius is contained in O. We denote this larger ball by $2B$. Thus we have

$$K \ni a \in B \subset 2B \subset O.$$

These balls cover K and the (sequential) compactness[2] of K implies that K is covered bij finitely many of such balls, i.e.

$$K \subset B_1 \cup \cdots \cup B_k \subset 2B_1 \cup \cdots \cup 2B_k \subset O.$$

On each such ball $2B_i$ we choose a smooth function $\eta_i \in C_c^\infty(2B)$ with $0 \leq \eta_i \leq 1$ and $\eta_i \equiv 1$ on B_i, and we extend these functions[3] to the whole of $\mathbb{R}^N$ by setting $\eta_i \equiv 0$ outside $2B_i$. Then $\eta_i \in C_c^\infty(\mathbb{R}^N)$ for $i = 1, \ldots, k$ and a new function $\chi \in C_c^\infty(\mathbb{R}^N)$ may be defined by[4]

$$1 - \chi(x) = (1 - \eta_1(x)) \cdots (1 - \eta_k(x)). \tag{29.1}$$

Indeed, if x is not contained in the union of the supports of $\eta_1, \ldots, \eta_k$ then all factors in the right hand side of (29.1) are equal to 1 and hence $\chi(x) = 0$. On the other hand, if x is contained in one of the balls B_i then the corresponding factor in the right hand side of (29.1) is equal to zero making the right hand side vanish whence $\chi(x) = 1$. In particular $\chi(x) \equiv 1$ on K. Moreover, since all factors take values in $[0, 1]$ the same holds for $\chi(x)$, for any $x \in \mathbb{R}^N$. We conclude that

$$\chi \in C_c^\infty(O), \quad \forall x \in \Omega \quad \chi(x) \in [0, 1], \quad \forall x \in K \quad \chi(x) = 1, \tag{29.2}$$

and this is why χ is called a cut-off function for K in O.

The second trick applies the first trick to a finite collection of such sets

$$\emptyset \neq K_1 \subset O_1, \ldots, \emptyset \neq K_m \subset O_m \quad \text{with} \quad \eta_j \in C_c^\infty(O_j) \tag{29.3}$$

cut-off functions as in (29.2). We define $\zeta_j \in C_c^\infty(O_j)$ by

$$\zeta_j(x) = \frac{\chi_j(x)}{\chi_1(x) + \cdots + \chi_m(x)} \tag{29.4}$$

[1] The set K could be the closure of a bounded domain Ω, or its boundary.

[2] This characterisation of compactness was not discussed yet in these notes.

[3] We can use the p-norm to our liking, the choice $p = 2$ allows radially symmetric η_i.

[4] I first saw this elegant trick in Folland's Real Analysis book.

and extend ζ_j to $\mathbb{R}^N$ via $\zeta_j(x) \equiv 0$ outside O_i. Note that below we don't really use the last part of (29.2) as $\chi_j(x) > 0$ for all $x \in K_i$ suffices to obtain the essential properties of the collection $\zeta_1, \ldots, \zeta_m$, which is called a partition of unity. For every x for which one of the $\chi_j(x) > 0$ it follows that

$$\zeta_1(x) + \cdots + \zeta_m(x) = 1. \tag{29.5}$$

Certainly this holds for x in $K_1 \cup \cdots \cup K_m$. On the other hand, outside the union of $O_1, \ldots, O_m$ this sum is by definition equal to zero.

Any function

$$f : K_1 \cup \cdots \cup K_m \to \mathbb{R}$$

splits up via

$$f(x) = f_1 + \cdots + f_m = \zeta_1(x)f(x) + \cdots + \zeta_m(x)f(x),$$

with the smaller parts $f_j = \zeta_j f$ compactly supported in O_j, and ζ_j not harming any smoothness the original function f may enjoy. Adding more K_j to the collection changes the functions ζ_j only via (29.4), with (29.5) remaining valid.

29.1 Partitions of compact manifolds

This section was written before Section 29.3. For Ω and $M = \partial\Omega$ you can jump to the end of this section. We now apply the techniques in Section 29 to a non-empty compact set $M \subset \mathbb{R}^N$ for which (C) in Section 27.2 applies in a sense we made more precise in Section 29.3 specifying blocks $[\tilde{a}, \tilde{b}] \subset \mathbb{R}^N$ in which the description (A) of Section 27.2 can be given, see (29.13). Below we rather choose blocks $[\tilde{a}_i, \tilde{b}_i] \subset \mathbb{R}^n$, given Φ_i as in (28.1). Thus for each $p \in M$ there exists a continuously differentiable injective

$$\Phi_i : [a, b] = [a_1, b_1] \times \cdots \times [a_n, b_n] \to \mathbb{R}^N$$

with $\mathcal{M}(\frac{\partial\Phi}{\partial u}) > 0$ such that $p \in \Phi((\tilde{a}, \tilde{b}))$ for some $[\tilde{a}, \tilde{b}] \subset (a, b)$, and in some open neighbourhood O of the compact set $K = \Phi([\tilde{a}, \tilde{b}])$ it holds that

$$x \in M \iff x \in \Phi((a, b)) \tag{29.6}$$

We would now like to consider the sets $\Phi((\tilde{a}, \tilde{b}))$ as open sets covering M, so that by compactness

$$M \subset \Phi_1((\tilde{a}_1, \tilde{b}_1)) \cup \cdots \cup \Phi_m((\tilde{a}_m, \tilde{b}_m)), \tag{29.7}$$

for some finite collection Φ_j, but clearly the sets $\Phi((\tilde{a},\tilde{b}))$ are not open[5] in $\mathbb{R}^N$, unless $n = N$. Nevertheless such a finite subcover exists.

To see this first choose $[\underline{a},\underline{b}] \subset (\tilde{a},\tilde{b})$ with $p \in \Phi((\underline{a},\underline{b}))$ and a suitable open neighbourhood $\underline{O}$ of $\underline{K} = \Phi([\underline{a},\underline{b}])$ with $\underline{O} \subset O$ to have the characterisation in (29.6) hold for all $x \in \underline{O}$ as well, and such that $\underline{O}$ does not intersect the (compact) image under Φ of the compact set $[a,b]\backslash(\tilde{a},\tilde{b})$. It then follows that $M \cap \underline{O} \subset \Phi((\tilde{a},\tilde{b}))$ because Φ is injective.

Varying $p \in M$ the open sets $\underline{O}$ cover M and by compactness there exists a finite collection $O_1,\dots,O_m$ such that

$$M \subset \underline{O}_1 \cup \cdots \cup \underline{O}_m \subset \Phi_1((\tilde{a}_1,\tilde{b}_1)) \cup \cdots \cup \Phi_m((\tilde{a}_m,\tilde{b}_m)),$$

which is the desired finite covering (29.7) consisting of patches.

We can now put $K_j = \Phi_j([\tilde{a}_j,\tilde{b}_j])$ and the corresponding open neighbourhoods O_j of K_j in which (29.6) characterises the elements of M. The description following (28.1) in Section 28 with unit blocks then results from Section 29.

We note that we can also have our partition of unity defined using cut-off functions $\chi = \chi(u)$ for $[\tilde{a},\tilde{b}] \subset (a,b)$, such as the blocks appearing in (29.7), but it is then slightly more complicated to formulate (29.4), because each χ_j is then a function of u. This allows us to deal with manifolds which are not necessarily embedded in $\mathbb{R}^N$.

Finally we observe that any Ω and $M = \partial\Omega$ as in Section **??** allow a choice of functions $\zeta_1,\dots,\zeta_n \in C_c^\infty((a_i,b_i))$ with $0 \le \zeta_i \le 1$ and $\zeta_1 + \cdots + \zeta_n \equiv 1$ on a neighbourhood of Ω such that every for every i either $[a_i,b_i] \subset \Omega$ holds, or $P_i = M \cap (a_i,b_i)$ is a patch such as in Section **??**.

29.2 Changing partitions

We still have to check that the integrals do not depend on the choice of the partitioning functions $\zeta_1,\dots,\zeta_n$. We observe that (28.2) defines a linear map

$$f \xrightarrow{L} \int_M f\,dS_n \tag{29.8}$$

from $X = C(M)$, the space of continuous real valued functions on M, to $\mathbb{R}$. Note that L is bounded in the sense that $|Lf| \le C|f|_\infty$, just as in Section 7.5, but we will not be using this below[6].

The partition naturally defines linear subspaces

$$X_i = \{\zeta_i f :\ f \in C(M)\},$$

[5] Of course they should be open in M.

[6] But we will need it to get rid of the annoying assumption $\Phi \in C^2$ in Section 28.3.

and the same holds for any other partition of M, given by say $\eta_1, \ldots, \eta_J$, which also defines a linear map

$$f \xrightarrow{K} \int_M f \, dS_n \tag{29.9}$$

via (28.2), and corresponding linear subspaces Y_j. Now let $\underline{\zeta}_{ij} = \zeta_i \eta_j$, with $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Then

$$f = f \sum_{i=1}^{I} \zeta_i = \sum_{i=1}^{I} \zeta_i f = \sum_{i=1}^{I} \zeta_i f \sum_{j=1}^{J} \eta_j = \sum_{i=1}^{I} \sum_{j=1}^{J} \zeta_i \eta_j f, \tag{29.10}$$

whence

$$Lf = L(\sum_{i=1}^{I} \zeta_i f) = \sum_{i=1}^{I} \int_{\Phi_i} \zeta_i f \, dS_n = \sum_{i=1}^{I} \sum_{j=1}^{J} \int_{\Phi_i} \zeta_i \eta_j f \, dS_n,$$

and likewise

$$Kf = \sum_{j=1}^{J} \sum_{i=1}^{I} \int_{\Psi_j} \eta_j \zeta_i f \, dS_n,$$

and thus it remains to show that

$$\int_{\Phi_i} \zeta_i \eta_j f \, dS_n = \int_{\Psi_j} \eta_j \zeta_i f \, dS_n \tag{29.11}$$

The integral on the left is defined via (26.13) as

$$\int_{\Phi_i} \zeta_i \eta_j f dS_n = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} \zeta_i(\Phi_i(u)) \eta_j(\Psi_j(v)) f(\Phi_i(u)) \mathcal{M}_n(\frac{\partial \Phi_i}{\partial u}) \, du_1 \cdots du_n.$$

It should be equal to the integral on the right which is defined via (26.13) as

$$\int_{\Psi_j} \eta_j \zeta_i f dS_n = \int_{c_1}^{d_1} \cdots \int_{c_n}^{d_n} \eta_j(\Psi_j(v)) \zeta_i(\Phi_i(u)) f(\Psi_j(v)) \mathcal{M}_n(\frac{\partial \Psi_j}{\partial v}) \, dv_1 \cdots dv_n.$$

The coordinates v have to be expressed in u and vice versa via coordinate transformations such as the ones in Section 29.4. These were defined in neighbourhoods of a given points $p \in \Phi_i((a,b)) \cap \Psi_j((c,d))$ only. Therefore we need another localisation argument[7] before we can apply Section 23 to conclude that the two integrals are the same.

[7] Try this one by yourself.

29.3 Again: local descriptions of a manifold

Let us be very precise in what we established for the local descriptions as in (A), (B) and (C) of Section 27.2, which correspond to (a,b,c) in III.4 of Edwards. Writing $z = (x, y)$ we take as a starting point that $F = F(z)$ is continuously differentiable on a block

$$[a, b] = [a_1, b_1] \times \cdots \times [a_n, b_n] \times [a_N, b_N] \times \cdots \times [a_N, b_N] \subset \mathbb{R}^{\mathrm{N}}$$

and that for some $p \in (a, b)$ the derivative $F'(p)$ is of maximal rank. Renaming and relabeling the variables in $z = (x, y)$ we can then arrange for the "partial" derivative $F_y(p)$ to be invertible. Theorem 27.2 then implies that there exists $(\tilde{a}, \tilde{b}) \subset (a, b)$ with $p \in (\tilde{a}, \tilde{b})$ and a continuously differentiable function

$$f : [\tilde{a}_x, \tilde{b}_x] \to (\tilde{a}_y, \tilde{b}_y)$$

such that $p = (p_x, p_y) \in (\tilde{a}, \tilde{b})$ and

$$F^{-1}(p) \cap [\tilde{a}, \tilde{b}] = \{(x, f(x)) : x \in [\tilde{a}_x, \tilde{b}_x]\} \subset [\tilde{a}_x, \tilde{b}_x] \times (\tilde{a}_y, \tilde{b}_y), \qquad (29.12)$$

with subscripts indicating the x and the y-parts of p, $\tilde{a}$ and $\tilde{b}$. Thus in the smaller block $[\tilde{a}, \tilde{b}]$ the level set of $F(p)$ coincides with the graph of f, and in the same block $[\tilde{a}, \tilde{b}]$ this graph then coincides with the zero-level set of $\tilde{F}(z) = \tilde{F}(x, y) = y - f(x)$.

As for (C), if we have, with subscripts denoting the x and the y-parts of Φ, that $\Phi(u) = (\Phi_x(u), \Phi_y(u))$ is continuously differentiable on $[a, b]$ with $0 \in (a, b)$ and $p = \Phi(0)$, then via Theorem 27.3 the invertibility of $\Phi'_x(0)$ is sufficient for the existence of $[\underline{a}_x, \underline{b}_x]$ with $p_x \in (\underline{a}_x, \underline{b}_x)$ and a continuously differentiable function $\phi : [\underline{a}_x, \underline{b}_x] \to (a, b)$ such that $\Phi_x(\phi(x)) = x$ for all $x \in [\underline{a}_x, \underline{b}_x]$. Moreover[8], we can choose $[\underline{a}_x, \underline{b}_x]$ such that $\phi((\underline{a}_x, \underline{b}_x))$ is an open set as the inverse image of $(\underline{a}_x, \underline{b}_x)$ under Φ_x.

The function f defined by $f(x) = \Phi_y(\phi(x))$ now defines a graph

$$\{(x, f(x)) : x \in [\underline{a}_x, \underline{b}_x]\}$$

which is a subset of $\Phi([a, b])$. If in addition Φ is injective on $[a, b]$ then the image under Φ of the closed bounded set $[a, b] \backslash \phi((\underline{a}_x, \underline{b}_x))$ is bounded and closed, and does not contain p. Thus there exists a block $[\tilde{a}, \tilde{b}]$ with $p \in (\tilde{a}, \tilde{b})$ such that $[\tilde{a}_x, \tilde{b}_x] \subset (\underline{a}_x, \underline{b}_x)$ with

$$\Phi([a, b] \backslash \phi((\underline{a}_x, \underline{b}_x))) \cap [\tilde{a}, \tilde{b}] = \emptyset.$$

[8] See the discussion after Theorem 27.3.

The continuity of f implies that we can restrict $\tilde{a}_x$ and $\tilde{b}_x$ a bit further to ensure that $f([\tilde{a}_x, \tilde{b}_x]) \subset (\tilde{a}_y, \tilde{b}_y)$. We note we also have that

$$\Phi([a,b]\backslash\phi((\tilde{a}_x, \tilde{b}_x))) \cap [\tilde{a}, \tilde{b}] = \emptyset,$$

since the additional points in the larger image $\Phi([a,b]\backslash\phi((\tilde{a}_x, \tilde{b}_x)))$ are on the graph of f outside $[\tilde{a}_x, \tilde{b}_x]$. Thus we have arrived from (C) to exactly the same formulation of (A) as above starting from (B): $p \in (\tilde{a}, \tilde{b})$ and

$$\Phi([a,b] \cap [\tilde{a}, \tilde{b}] = \{(x, f(x)) : x \in [\tilde{a}_x, \tilde{b}_x]\} \subset [\tilde{a}_x, \tilde{b}_x] \times (\tilde{a}_y, \tilde{b}_y). \qquad (29.13)$$

The two statements (29.12) and (29.13) should be compared to the definition Edwards gives in Section 4 of his Chapter III for $M \subset \mathbb{R}^N$ to be an n-dimensional manifold. Every $p \in M$ should, after relabeling and renaming in $z = (x, y)$, be contained in an open set O in which

$$P = O \cap M = \{(x, f(x)) : x \in U\},$$

with $U \subset \mathbb{R}^n$ open and $f : U \to \mathbb{R}^m$ continuously differentiable, is called a C^1-patch of M. Of course it is then clear that $U \supset [\tilde{a}_x, \tilde{b}_x] \supset (\tilde{a}_x, \tilde{b}_x)$ and $O \supset [\tilde{a}, \tilde{b}] \supset (\tilde{a}, \tilde{b})$ for some $(\tilde{a}, \tilde{b}) \ni p$, and thus it is completely equivalent to ask that $p \in (\tilde{a}, \tilde{b})$ and

$$p \in M \cap [\tilde{a}, \tilde{b}] = \{(x, f(x)) : x \in [\tilde{a}_x, \tilde{b}_x]\} \subset [\tilde{a}_x, \tilde{b}_x] \times (\tilde{a}_y, \tilde{b}_y) \qquad (29.14)$$

for some closed block $[\tilde{a}, \tilde{b}]$ with $(\tilde{a}_x, \tilde{b}_x) \ni p_x$, and some continuously differentiable $f : [\tilde{a}_x, \tilde{b}_x] \to (\tilde{a}_y, \tilde{b}_y)$, exactly as in (29.12,29.13), the patch being

$$M \cap (\tilde{a}, \tilde{b}) = \{(x, f(x)) : x \in (\tilde{a}_x, \tilde{b}_x)\} \ni p = (p_x, f(p_x)). \qquad (29.15)$$

In the closed N-block $[\tilde{a}, \tilde{b}]$ there are no other points of M than the points on the graph of $f : [\tilde{a}_x, \tilde{b}_x] \to (\tilde{a}_y, \tilde{b}_y)$.

29.4 Coordinate transformations

By definition every $p \in M$ is in such a patch as above and typically patches overlap. If p is in two such patches, say with functions f and g, it may happen that f and g are functions of the x-part of z. In that case the patches are parameterised by

$$u \to \Phi(u) = (u, f(u)) \quad \text{and} \quad v \to \Psi(v) = (v, g(v)) \qquad (29.16)$$

defined on overlapping blocks with p_x in the interior of the intersection of the blocks, which is an open block itself. The common part of M is then contained in the intersection of the two N-blocks.

Viewing the n-tuples u and v as local coordinates on M near p, a transformation of these coordinates is simply given by $v = u$. In all other cases, we may renumber the variables of $\mathbb{R}^N$ to have the patches parameterised as

$$u \to \Phi(u) = (u_1, u_2, f_3(u_1, u_2), f_4(u_1, u_2));$$

$$v \to \Psi(v) = (v_1, g_2(v_1, v_3), v_3, g_4(v_1, v_3)),$$

with (u_1, u_2) and (v_1, v_3) in some open block in $\mathbb{R}^n$, or as

$$u \to \Phi(u) = (u_1, f_2(u_1), f_3(u_1));$$

$$v \to \Psi(v) = (g_1(v_2), v_2, g_3(v_2)),$$

with u_1 and v_3 in some open block in $\mathbb{R}^n$. Note that the first case above cannot occur if $N = n + 1$.

To rewrite Ψ in the form Φ we need the invertibility of respectively

$$\frac{\partial g_2}{\partial v_3} \quad \text{and} \quad \frac{\partial g_1}{\partial v_2}, \tag{29.17}$$

in which case we we obtain respectively

$$w \to \tilde{\Phi}(w) = (w_1, w_2, h_3(w_1, w_2), h_4(w_1, w_2))$$

and

$$w \to \tilde{\Phi}(w) = (w_1, h_2(w_1), h_3(w_1))$$

as local descriptions of the Ψ-patches near p. The definition of what a manifold is then implies that

$$\tilde{\Phi} \equiv \Phi$$

on an open block containing (p_1, p_2) in the first case and p_1 in the second case. It then follows as above that $u = w$ is a coordinate transformation just as $u = v$ for (29.16) while w is obtained from v via a coordinate transformation just as x from u in the proof of (A) from (C) above.

It thus remains to establish the invertibility of the partial Jacobian matrices in (29.17) in p to conclude there exists a local C^1-transformation from u to v near p. Note that these are also the conditions for solving part[9] of $\Phi(u) = \Psi(v)$ via

$$v_1 = u_1,\ v_3 = f_3(u_1, u_2) \quad \text{and} \quad u_1 = v_1,\ u_2 = g_2(v_1, v_3) \tag{29.18}$$

in the first case, and

$$u_1 = g_1(v_2) \quad \text{and} \quad v_2 = f_2(u_1) \tag{29.19}$$

[9]All equations but the last one, which then requires some argument to hold as well.

in the second case. The invertiblity of the partial Jacobian matrices in (29.17) in p follows because otherwise the Ψ-patch cannot achieve all respectively (u_1, u_2)-directions and u_1-directions that occur in the Φ-patch, contradicting the assumption that the Ψ-patch covers all of M in its defining neighbourhood.

The restriction to patches of the form (29.15) looks like an obvious choice for simplicity, but may bother us later when dealing with (28.33), we'll see.

30 Applications

Still in Dutch. Part of the this was initially written for students in chemistry. Wat bruggetjes naar hoe de natuurkundigen en scheikundigen het doen, en aan het eind wat complexe functietheorie, met de lijnintegralen alleen maar over rechte lijnstukjes. Voldoende voor *an early introduction of the functional calculus* waarmee voor z in $f(z)$ ook iets heel anders mag worden ingevuld, bijvoorbeeld een vierkante matrix.

30.1 Integraalrekening in poolcoördinaten

Merk op dat we *in het echte leven* over meer verzamelingen zullen willen integreren dan over rechthoeken. Bijvoorbeeld over heel $\mathbb{R}^2$. Voor niet-negatieve functies $u : \mathbb{R}^2 \to \mathbb{R}$ is

$$\iint_{\mathbb{R}^2} u = \lim_{R\to\infty} \underbrace{\iint_{[-R,R]\times[-R,R]} u(x,y)d(x,y)}_{J(R)} = \lim_{R\to\infty} J(R) \tag{30.1}$$

op natuurlijke manier gedefinieerd in $[0,\infty]$ als limiet van een niet-dalende functie $R \to J(R) \geq 0$.

Er is natuurlijk geen enkele reden om een integraal over heel $\mathbb{R}^2$ per se als een limiet van integralen in rechthoekige coördinaten over in dit geval vierkanten te introduceren. Poolcoördinaten zijn vaak veel handiger. Voor Riemannsommen in poolcoördinaten ten behoeve van de rechtstreekse definitie en uitwerking van

$$\begin{aligned}\iint_{x^2+y^2\leq R^2} u(x,y)d(x,y) &= \int_0^{2\pi}\int_0^R u(r\cos\theta, r\sin\theta)\, r dr\, d\theta \\ &= \int_0^R \int_0^{2\pi} u(r\cos\theta, r\sin\theta)\, d\theta\, r dr \end{aligned} \tag{30.2}$$

gebruiken we

$$0 = r_0 \leq r_1 \leq \cdots \leq r_M = R \quad \text{met} \quad M \in \mathbb{N} \tag{30.3}$$

en

$$0 = \theta_0 \leq \theta_1 \leq \cdots \leq \theta_N = 2\pi \quad \text{met} \quad N \in \mathbb{N}, \tag{30.4}$$

en tussensommen van de vorm

$$\sum_{k=1}^{M}\sum_{l=1}^{N} u(\rho_k\cos\phi_l, \rho_k\sin\phi_l)\, \underbrace{\frac{1}{2}(r_k^2 - r_{k-1}^2)(\theta_l - \theta_{l-1})}_{\text{waarom dit dan?}} =$$

$$\sum_{k=1}^{M}\sum_{l=1}^{N} u(\rho_k\cos\phi_l,\rho_k\sin\phi_l)\underbrace{\frac{r_k+r_{k-1}}{2}}_{\tilde\rho_k}(r_k-r_{k-1})(\theta_l-\theta_{l-1}),$$

met tussenwaarden $\rho_k,\tilde\rho_k\in[r_{k-1},r_k]$ en $\phi_l\in[\theta_{l-1},\theta_l]$. De details zijn zelf in te vullen. Leuker is deze mooie toepassing van (30.2) in de volgende stelling over harmonische functies.

Exercise 30.1. Een twee keer continu differentieerbare functie $(x,y)\to u(x,y)=u(r\cos\theta,r\sin\theta)$ heet harmonisch als $\Delta u=0$. Laat zien dat

$$u(0,0)=\frac{1}{2\pi}\int_0^{2\pi}u(r\cos\theta,r\sin\theta)\,d\theta,$$

en dat harmonische functies dus in elk punt het gemiddelde van hun waarden op een diskvormige omgeving zijn. Hint: gebruik Stelling 14.5 als je de integraal van Δu over $\bar B_R$ hebt vertaald naar een integraal met alleen maar $d\theta$.

Ook leuk is dat voor niet-negatieve continue functies $u:\mathbb{R}^2\to\mathbb{R}$ de integraal

$$\iint_{\mathbb{R}^2}u=\lim_{R\to\infty}\iint_{x^2+y^2\le R^2}u(x,y)d(x,y) \tag{30.5}$$

nu net zo natuurlijk gedefinieerd is in $[0,\infty]$ als door (30.1). Alleen een wiskundige vraagt zich dan af dit consistent is. Dat moet en dat mag hoor:

Exercise 30.2. Voor niet-negatieve continue $u:\mathbb{R}^2\to\mathbb{R}$ geldt

$$\lim_{R\to\infty}\iint_{x^2+y^2\le R^2}u(x,y)d(x,y)=\lim_{R\to\infty}\iint_{[-R,R]\times[-R,R]}u(x,y)d(x,y).$$

In de formule van Stirling stond nog een integraal die we nu netjes kunnen uitrekenen met behulp van Opgave 30.2 en de functie

$$(x,y)\xrightarrow{u}e^{-\frac12(x^2+y^2)}$$

Kort door de bocht opgeschreven concluderen we dat

$$\left(\int_{-\infty}^{\infty}e^{-\frac12x^2}\,dx\right)^2=\int_{-\infty}^{\infty}e^{-\frac12x^2}\,dx\int_{-\infty}^{\infty}e^{-\frac12y^2}\,dy=\iint_{\mathbb{R}^2}e^{-\frac12(x^2+y^2)}\,d(x,y)$$
$$=\int_0^{\infty}\int_0^{2\pi}e^{-\frac12r^2}\,rd\theta dr=\int_0^{\infty}2\pi e^{-\frac12r^2}\,rdr=2\pi[-e^{-\frac12r^2}]_0^{\infty}=2\pi.$$

Exercise 30.3. Laat met Opgave 30.2 zien dat

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx = \sqrt{2\pi}.$$

Bijgevolg hebben

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \text{en} \quad u(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} \tag{30.6}$$

dus de eigenschap dat ze (positief zijn en) en totale integraal gelijk aan 1 hebben. We noemen zulke functies *kansdichtheden*. De dichtheid $f(x)$ hoort bij een stochastische grootheid X waarvoor geldt dat de kans op uitkomst $X \in [a,b]$ gelijk is aan

$$P(X \in [a,b]) = \int_a^b f(x)\,dx,$$

en

$$P(X \leq x) = \int_{-\infty}^{x} f(s)\,ds$$

wordt de cumulatieve verdelingsfunctie van X genoemd.

Een van X onafhankelijke stochastische grootheid Y kan een kansdichtheid $g(y)$ hebben die beschrijft dat de kans op $Y \in [c,d]$ gelijk is aan

$$P(Y \in [c,d]) = \int_c^d g(y)\,dy.$$

De simultane kansdichtheid $u(x,y) = f(x)g(y)$ geeft dan de kans op $X \in [a,b]$ en $Y \in [c,d]$ als

$$\iint_{\mathbb{R}^2} u = \int_a^b f(x)\,dx \int_c^d g(y)\,dy.$$

De kansdichtheden in (30.6) worden de 1-en 2-dimensionale standaard *normale verdeling* genoemd. Is de functie g hetzelfde als de functie f in (30.6), dan zijn X en Y allebei standaard normaal verdeeld. De twee stochastische grootheden X en Y kunnen op elkaar gedeeld worden. De kans op

$$Q = \frac{Y}{X} \in [a,b]$$

is dan gelijk aan de integraal van $u(x,y)$ over het gebied ingesloten door de lijnen $y = ax$ en $y = bx$.

In het geval dat X en Y standaard normaal verdeeld en onderling onafhankelijk zijn, bestaat die integraal uit twee identieke stukken waarvan er één gegeven wordt door

$$\{(x,y) : x \geq 0,\, ax \leq y \leq bx\},$$

een gebied dat in poolcoördinaten beschreven wordt door θ in een deelinterval van $(-\frac{\pi}{2}, \frac{\pi}{2})$.

We willen concluderen dat

$$P(Q \in [a,b]) = 2\int_0^\infty \int_{ax}^{bx} \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}\, dy\, dx$$

$$= \frac{1}{\pi}\int_0^\infty \int_{\arctan a}^{\arctan b} e^{-\frac{1}{2}r^2}\, r d\theta dr = \frac{1}{\pi}(\arctan b - \arctan a) = \int_a^b \frac{1}{\pi}\frac{1}{1+q^2}\, dq.$$

De stochastische grootheid Q heeft dan een kansdichtheid gegeven door de functie

$$q \to \frac{1}{\pi}\frac{1}{1+q^2}.$$

Exercise 30.4. Hierboven manipuleerden we met meervoudige oneigenlijke integralen over "taartpunten" in $\mathbb{R}^2$. De daarvoor benodigde theorie vraagt om een uitbreiding van de theorie van integralen over het hele vlak in poolcoördinaten. Dat kun je ook zelf proberen precies te maken nu.

30.2 Gradient, kettingregel, coördinatentransformaties

De *kettingregel* generaliseert de regel in Opgave 11.2. Met de opmerking dat de formules gelezen moet worden met matrices[1] is de regel met bewijs en al over te schrijven en nu meteen toepasbaar.

We spellen een en ander nu uit in het geval van coördinatentransformaties, met als belangrijk voorbeeld de overgang op poolcoördinaten die we al gebruikten om $\mathbb{C}$ te beschrijven en in $\mathbb{C}$ te rekenen: ieder punt $(x,y) \in \mathbb{R}^2$ kunnen we via

$$x = r\cos\theta \quad \text{en} \quad y = r\sin\theta \tag{30.7}$$

zien als gegeven door poolcoördinaten $r, \theta \in \mathbb{R}$ voor $(x,y) \neq (0,0)$.

[1]Beter: lineaire afbeeldingen, in dit hele hoofdstuk de facto matrices.

Een differentieerbare scalaire functie $F(x, y)$ van x en y is zo automatisch ook een differentieerbare functie van r en θ. In wat volgt zien we (30.7) als transformatie

$$Z : \mathbb{R}^2 \to \mathbb{R}^2$$

van de onafhankelijke plaatsvariabelen, en $F(x, y) = F(Z(r, \theta))$ als de *afhankelijke* variabele. Buiten de wiskunde, met name in de natuurkunde, is het gebruikelijk om de afhankelijke variabele met hetzelfde symbool te noteren als alleen de onafhankelijke variabelen worden getransformeerd.

30.2.1 Gradient, divergentie en Laplaciaan

Voor $F : \mathbb{R}^2 \to \mathbb{R}$ is de definitie van differentieerbaarheid in de gewone rechthoekige coördinaten x en y en $h = x - x_0$, $k = y - y_0$ te lezen als

$$F(x_0 + h, y_0 + k) = F(x_0, y_0) + ah + bk + R(h, k; x_0, y_0), \tag{30.8}$$

met $a, b \in \mathbb{R}$ en

$$\frac{R(h, k; x_0, y_0)}{\sqrt{h^2 + k^2}} \to 0 \quad \text{als} \quad \sqrt{h^2 + k^2} \to 0, \tag{30.9}$$

vergelijk met het eerdere uitpakken. De volgende opgave is misschien nu wat dubbelop, maar dat kan geen enkel kwaad.

Exercise 30.5. Neem voor $F : \mathbb{R}^2 \to \mathbb{R}$, $(x_0, y_0) \in \mathbb{R}^2$ en $a, b \in \mathbb{R}$ aan dat (30.8) geldt met (30.9). Dan volgt dat

$$\frac{F(x_0 + h, y_0) - F(x_0, y_0)}{h} \to a \quad \text{en} \quad \frac{F(x_0, y_0 + k) - F(x_0, y_0)}{k} \to b$$

als $h, k \to 0$. Laat dit zien.

Meerdere notaties worden gebruikt, zoals

$$a = F_x(x_0, y_0) = \frac{\partial F}{\partial x}(x_0, y_0) = (\delta_x F)(x_0, y_0) = (D_1 F)(x_0, y_0); \tag{30.10}$$

$$b = F_y(x_0, y_0) = \frac{\partial F}{\partial y}(x_0, y_0) = (\delta_y F)(x_0, y_0) = (D_2 F)(x_0, y_0), \tag{30.11}$$

waarbij (x_0, y_0) en haakjes vaak worden weggelaten want

$$a = F_x = \frac{\partial F}{\partial x} = \delta_x F = D_1 F \quad \text{en} \quad b = F_y = \frac{\partial F}{\partial y} = \delta_y F = D_2 F$$

ziet er gewoon fijner uit.

Als kolomvector schrijven we ook, met

$$e_x = \begin{pmatrix}1\\0\end{pmatrix} \quad \text{en} \quad e_y = \begin{pmatrix}0\\1\end{pmatrix}, \tag{30.12}$$

$$\begin{pmatrix}a\\b\end{pmatrix} = \nabla F = \frac{\partial F}{\partial x}e_x + \frac{\partial F}{\partial y}e_y = e_x\frac{\partial F}{\partial x} + e_y\frac{\partial F}{\partial y}, \tag{30.13}$$

de *gradient* van F in (x_0, y_0), geschreven zonder (x_0, y_0). Merk op dat het lineaire gedeelte in (30.8) te schrijven is als

$$ah + bk = \begin{pmatrix}a\\b\end{pmatrix} \cdot \begin{pmatrix}h\\k\end{pmatrix} = \begin{pmatrix}h\\k\end{pmatrix} \cdot \begin{pmatrix}a\\b\end{pmatrix} = h\frac{\partial F}{\partial x} + k\frac{\partial F}{\partial y}, \tag{30.14}$$

het inprodukt[2] van ∇F en de verschilvector $\binom{h}{k}$.

We zien dus hoe de gradiënt de vector is die de lineaire afbeelding $DF : \mathbb{R}^2 \to \mathbb{R}$ via het inprodukt representeert als

$$\begin{pmatrix}h\\k\end{pmatrix} \xrightarrow{DF} \nabla F \cdot \begin{pmatrix}h\\k\end{pmatrix},$$

maar ook dat (30.14) te lezen is als de *differentiaaloperator*

$$h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y} \quad \text{werkend op} \quad F.$$

Evenzo zien we ∇ als

$$\nabla = e_x\frac{\partial}{\partial x} + e_y\frac{\partial}{\partial y} \quad \text{werkend op} \quad F \quad \text{geeft} \quad \nabla F, \tag{30.15}$$

een vectorwaardige differentiaaloperator.

Middels het inprodukt kan ∇ ook werken op een vectorwaardige differentieerbare functie

$$(x, y) \to \begin{pmatrix}V_x(x,y)\\V_y(x,y)\end{pmatrix} = \begin{pmatrix}V_x\\V_y\end{pmatrix} = V_x e_x + V_y e_y,$$

en wel als

$$\nabla \cdot V = (e_x\frac{\partial}{\partial x} + e_y\frac{\partial}{\partial y}) \cdot (V_x e_x + V_y e_y) = \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y}, \tag{30.16}$$

[2]Het inprodukt van twee vectoren in $\mathbb{R}^2$ wordt gegeven door $\binom{a}{b} \cdot \binom{h}{k} = ah + bk$.

de *divergentie* van V.

We schrijven hier nu V met subscripten[3] x, y voor de x- en y-coördinaten V_x en V_y van V t.o.v. de orthonormale vectoren (30.12) die samen de standaardbasis van $\mathbb{R}^2$ vormen. Merk wel op dat V_x en V_y van x en y afhangen maar e_x en e_y niet. De indices x en y staan voor de x-richting en de y-richting, en die richtingen zijn overal in het x, y-vlak hetzelfde.

Elk van de twee termen in ∇ werkt nu alleen op V_x en V_y, en omdat

$$e_x \cdot e_x = e_y \cdot e_y = 1 \quad \text{en} \quad e_x \cdot e_y = e_y \cdot e_x = 0, \tag{30.17}$$

blijven er maar twee termen over in (30.16). Omdat e_x en e_y niet van x en y afhangen geeft elk van de vier termen

$$e_x \frac{\partial}{\partial x} \cdot V_x e_x, \quad e_x \frac{\partial}{\partial x} \cdot V_y e_y, \quad e_y \frac{\partial}{\partial y} \cdot V_x e_x, \quad e_y \frac{\partial}{\partial x} \cdot V_y e_y$$

die we krijgen bij het uitwerken van (30.16) maar één term, te weten

$$e_x \frac{\partial}{\partial x} \cdot V_x e_x = e_x \frac{\partial}{\partial x} \cdot V_x e_x = e_x \cdot \frac{\partial}{\partial x} V_x e_x = e_x \cdot \frac{\partial V_x}{\partial x} e_x = \frac{\partial V_x}{\partial x} e_x \cdot e_x = \frac{\partial V_x}{\partial x}$$

voor de eerste,

$$e_x \frac{\partial}{\partial x} \cdot V_y e_y = e_x \frac{\partial}{\partial x} \cdot V_y e_y = e_x \cdot \frac{\partial}{\partial x} V_y e_y = e_x \cdot \frac{\partial V_x}{\partial x} e_y = \frac{\partial V_x}{\partial y} e_x \cdot e_y = 0$$

voor de tweede, en

$$e_y \frac{\partial}{\partial y} \cdot V_x e_x = 0, \quad e_y \frac{\partial}{\partial x} \cdot V_y e_y = \frac{\partial V_y}{\partial y}$$

voor de derde en vierde. Van de vier termen worden er dus nog twee nul vanwege $e_x \cdot e_y = 0$ in (30.17) en de andere twee vereenvoudigen en blijven in die vorm over in (30.16).

Als $V = \nabla F$ differentieerbaar is dan volgt zo dat

$$\nabla \cdot \nabla F = (e_x \frac{\partial}{\partial x} + e_y \frac{\partial}{\partial y}) \cdot (\frac{\partial F}{\partial x} e_x + \frac{\partial F}{\partial y} e_y) = \frac{\partial}{\partial x} \frac{\partial F}{\partial x} + \frac{\partial}{\partial y} \frac{\partial F}{\partial y} = \Delta F, \tag{30.18}$$

de Laplaciaan van F, die weer gezien kan worden als

$$\Delta F \quad \text{is} \quad \Delta = \frac{\partial}{\partial x} \frac{\partial}{\partial x} + \frac{\partial}{\partial y} \frac{\partial}{\partial y} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \quad \text{werkend op} \quad F. \tag{30.19}$$

Omschrijven van gradiënt, divergentie en Laplaciaan naar poolcoördinaten is nu een nuttige oefening waarvoor de volgende subsecties van belang zijn. Het is handig om daarbij naar twee net iets anders uitgewerkte notaties voor de kettingregel te kijken.

[3] Niet te verwarren met het gebruik van subscripten voor partiële afgeleiden!

30.2.2 Kettingregel uitgeschreven voor transformaties

We weten dat we de kettingregel toe mogen passen op

$$(r,\theta) \to (r\cos\theta, r\sin\theta) = (X(r,\theta), Y(r,\theta)) = (x,y) \to F(x,y) = G(r,\theta),$$

door de lineaire benadering van

$$(r,\theta) \to (r\cos\theta, r\sin\theta)$$

rond (r_0,θ_0) in te vullen in de lineaire benadering van

$$(x,y) \to F(x,y)$$

rond (x_0,y_0). We doen dit nu met $\tilde{h} = r - r_0$ en $\tilde{k} = \theta - \theta_0$, met weglating van (r_0,θ_0) in de partiële afgeleiden.

Omdat we in deze sectie $F(x,y) = G(r,\theta)$ als onbekende afhankelijke grootheid willen zien, bijvoorbeeld de oplossing van een partiële differentiaalvergelijking, kiezen we nu eerst voor de schrijfwijze zoals rechts in (30.14). De lineaire termen in de expansies

$$X(r_0+\tilde{h},\theta_0+\tilde{k}) = X(r_0,\theta_0) + \tilde{h}\frac{\partial X}{\partial r} + \tilde{k}\frac{\partial X}{\partial \theta} + \cdots,$$

$$Y(r_0+\tilde{h},\theta_0+\tilde{k}) = Y(r_0,\theta_0) + \tilde{h}\frac{\partial Y}{\partial r}\tilde{h} + \tilde{k}\frac{\partial Y}{\partial \theta} + \cdots$$

moeten dan als

$$h = \tilde{h}\frac{\partial X}{\partial r} + \tilde{k}\frac{\partial X}{\partial \theta} \quad \text{en} \quad k = \tilde{h}\frac{\partial Y}{\partial r} + \tilde{k}\frac{\partial Y}{\partial \theta}$$

in (30.14) worden ingevuld[4], en het resultaat

$$(\tilde{h}\frac{\partial X}{\partial r} + \tilde{k}\frac{\partial X}{\partial \theta})\frac{\partial F}{\partial x} + (\tilde{h}\frac{\partial Y}{\partial r} + \tilde{k}\frac{\partial Y}{\partial \theta})\frac{\partial F}{\partial y} =$$

$$\tilde{h}(\frac{\partial X}{\partial r}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial r}\frac{\partial F}{\partial y}) + \tilde{k}(\frac{\partial X}{\partial \theta}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial \theta}\frac{\partial F}{\partial y})$$

is dan volgens de kettingregel gelijk aan

$$\tilde{h}\frac{\partial G}{\partial r} + \tilde{k}\frac{\partial G}{\partial \theta}.$$

[4]We gaan er nu niet echt vanuit dat de lezer al met matrices heeft leren rekenen.

Er volgt dus dat

$$\frac{\partial G}{\partial r} = \frac{\partial X}{\partial r}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial r}\frac{\partial F}{\partial y} \tag{30.20}$$

$$\frac{\partial G}{\partial \theta} = \frac{\partial X}{\partial \theta}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial \theta}\frac{\partial F}{\partial y}, \tag{30.21}$$

in vector-matrixnotatie te schrijven als

$$\begin{pmatrix} \frac{\partial G}{\partial r} \\ \frac{\partial G}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \frac{\partial X}{\partial r}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial r}\frac{\partial F}{\partial y} \\ \frac{\partial X}{\partial \theta}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial \theta}\frac{\partial F}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{\partial X}{\partial r} & \frac{\partial Y}{\partial r} \\ \frac{\partial X}{\partial \theta} & \frac{\partial Y}{\partial \theta} \end{pmatrix} \begin{pmatrix} \frac{\partial F}{\partial x} \\ \frac{\partial F}{\partial y} \end{pmatrix}, \tag{30.22}$$

waarin we links een 2 bij 1 matrix zien met de partiële afgeleiden van G, en rechts net zo'n matrix voor F, en een 2 bij 2 matrix voor

$$(r,\theta) \xrightarrow{Z} (X(r,\theta), Y(r,\theta)),$$

met $Z : \mathbb{R}^2 \to \mathbb{R}^2$ via (30.7) gedefinieerd door

$$Z(r,\theta) = (X(r,\theta), Y(r,\theta)) = (r\cos\theta, r\sin\theta).$$

Horizontaal worden deze matrices genummerd met de variabele grootheden in het beeld, verticaal met die in het domein van de betreffende afbeelding. Andersom als voorheen, omdat we de schrijfwijze rechts in (30.14) hebben gebruikt.

De kolomvectoren in (30.22) zien er uit als gradiënten, maar dat is slechts misleidende schijn, zoals we in Sectie 30.2.4 zullen zien.

30.2.3 Kettingregel met Jacobimatrices

Mooie voorbeelden van matrixprodukten als in (19.3) zien we als we in (30.22) aan beide kanten links $(\tilde{h}\ \tilde{k})$ erbij zetten. Dan is

$$(\tilde{h}\ \tilde{k}) \begin{pmatrix} \frac{\partial G}{\partial r} \\ \frac{\partial G}{\partial \theta} \end{pmatrix} = (\tilde{h}\ \tilde{k}) \begin{pmatrix} \frac{\partial X}{\partial r} & \frac{\partial Y}{\partial r} \\ \frac{\partial X}{\partial \theta} & \frac{\partial Y}{\partial \theta} \end{pmatrix} \begin{pmatrix} \frac{\partial F}{\partial x} \\ \frac{\partial F}{\partial y} \end{pmatrix}, \tag{30.23}$$

nu links en rechts uit te werken tot een 1 bij 1 matrix, met daarin precies de twee lineaire stukken die we hierboven aan elkaar gelijkstelden bij het uitwerken van de kettingregel, om tot (30.20) en (30.21) te komen.

Via links en rechts transponeren is (30.23) equivalent met

$$\left(\frac{\partial G}{\partial r}\ \ \frac{\partial G}{\partial \theta}\right) \begin{pmatrix} \tilde{h} \\ \tilde{k} \end{pmatrix} = \left(\frac{\partial F}{\partial x}\ \ \frac{\partial F}{\partial y}\right) \begin{pmatrix} \frac{\partial X}{\partial r} & \frac{\partial X}{\partial \theta} \\ \frac{\partial Y}{\partial r} & \frac{\partial Y}{\partial \theta} \end{pmatrix} \begin{pmatrix} \tilde{h} \\ \tilde{k} \end{pmatrix}, \tag{30.24}$$

waarin we de *Jacobimatrices* van G, F en Z herkennen, waarin de beeldvariabelen niet horizontaal maar verticaal genummerd worden. Ook zien we dat de volgorde in (30.24) nu prettig is als we $\tilde{h}$ en $\tilde{k}$ zien als variabel.

Als $F(x,y) = G(r,\theta)$ een grootheid is met twee componenten

$$F_1(x,y) = G_1(r,\theta) \quad \text{en} \quad F_2(x,y) = G_2(r,\theta),$$

dan kan een en ander voor beide componenten in één keer opgeschreven worden als

$$\begin{pmatrix} \frac{\partial G_1}{\partial r} & \frac{\partial G_1}{\partial \theta} \\ \frac{\partial G_2}{\partial r} & \frac{\partial G_2}{\partial \theta} \end{pmatrix} \begin{pmatrix} \tilde{h} \\ \tilde{k} \end{pmatrix} = \begin{pmatrix} \frac{\partial F_1}{\partial x} & \frac{\partial F_1}{\partial y} \\ \frac{\partial F_2}{\partial x} & \frac{\partial F_2}{\partial y} \end{pmatrix} \begin{pmatrix} \frac{\partial X}{\partial r} & \frac{\partial X}{\partial \theta} \\ \frac{\partial Y}{\partial r} & \frac{\partial Y}{\partial \theta} \end{pmatrix} \begin{pmatrix} \tilde{h} \\ \tilde{k} \end{pmatrix}, \tag{30.25}$$

en zien we hoe de kettingregel toegepast op

$$\mathbb{R}^2 \xrightarrow{Z} \mathbb{R}^2 \xrightarrow{F} \mathbb{R}^2$$

de Jacobimatrix van G produceert via het matrixprodukt van de Jacobimatrices van F en Z.

Deze notatie suggereert om de afhankelijke grootheid $F(x,y) = G(r,\theta)$ als 2-vector te zien, dus

$$F(x,y) = \begin{pmatrix} F_1(x,y) \\ F_2(x,y) \end{pmatrix} \quad \text{en} \quad G(r,\theta) = \begin{pmatrix} G_1(r,\theta) \\ G_2(r,\theta) \end{pmatrix},$$

en dus ook x, y en r, θ als componenten van de 2-vectoren

$$\begin{pmatrix} x \\ y \end{pmatrix} \quad \text{en} \quad \begin{pmatrix} r \\ \theta \end{pmatrix}.$$

We blijven echter $F = F(x,y)$ en $G = G(r,\theta)$ schrijven.

30.2.4 Omschrijven van differentiaaloperatoren

De notatie (30.23) is handiger als we zoals gebruikelijk in de natuurkunde aan $F(x,y) = G(r,\theta)$ denken als één en dezelfde afhankelijke grootheid, en niet als een functie zoals gebruikelijk in de wiskunde.

In dat geval ligt het voor de hand om die grootheid af te splitsen uit de notatie in (30.22) en de kettingregel voor coördinatentransformaties te schrijven als

$$\begin{pmatrix} \frac{\partial}{\partial r} \\ \frac{\partial}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \frac{\partial X}{\partial r} & \frac{\partial Y}{\partial r} \\ \frac{\partial X}{\partial \theta} & \frac{\partial Y}{\partial \theta} \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix}, \tag{30.26}$$

hetgeen de matrixnotatie is voor

$$\frac{\partial}{\partial r} = \frac{\partial X}{\partial r}\frac{\partial}{\partial x} + \frac{\partial Y}{\partial r}\frac{\partial}{\partial y};$$

$$\frac{\partial}{\partial \theta} = \frac{\partial X}{\partial \theta}\frac{\partial}{\partial x} + \frac{\partial Y}{\partial \theta}\frac{\partial}{\partial y},$$

waaruit de differentiaaloperatoren

$$\frac{\partial}{\partial x} \quad \text{en} \quad \frac{\partial}{\partial y}$$

kunnen worden opgelost in termen van de coëfficiënten

$$\frac{\partial X}{\partial r}, \frac{\partial Y}{\partial r}, \frac{\partial X}{\partial \theta}, \frac{\partial Y}{\partial \theta} \quad \text{en de differentiaaloperatoren} \quad \frac{\partial}{\partial r}, \frac{\partial}{\partial \theta}.$$

Exercise 30.6. In het concrete geval van poolcoördinaten geeft dit

$$\frac{\partial}{\partial x} = \cos\theta\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial \theta};$$

$$\frac{\partial}{\partial y} = \sin\theta\frac{\partial}{\partial r} + \frac{\cos\theta}{r}\frac{\partial}{\partial \theta}.$$

Laat dit zien.

Met Opgave 30.6 zijn we nog niet klaar als we in (30.15)

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix} = e_x\frac{\partial}{\partial x} + e_y\frac{\partial}{\partial y}$$

willen omschrijven naar r en θ. De vraag is ook hoe we e_x en e_y omschrijven naar e_r en e_θ, en daarvoor komt de vraag wat e_r en e_θ eigenlijk zijn.

Een natuurkundige zal hier niet lang over nadenken Teken maar een plaatje en het is evident dat *Teken plaatje!*

$$e_r = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \quad \text{en} \quad e_\theta = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix},$$

en

$$\nabla = e_x\frac{\partial}{\partial x} + e_y\frac{\partial}{\partial y} = e_r\frac{\partial}{\partial r} + \frac{1}{r}e_\theta\frac{\partial}{\partial \theta} \tag{30.27}$$

de gradiënt in poolcoördinaten geeft. Daar had'ie de hele kettingregel überhaupt niet voor nodig. Omdat

$$e_r \cdot e_r = e_\theta \cdot e_\theta = 1 \quad \text{en} \quad e_r \cdot e_\theta = e_\theta \cdot e_r = 0,$$

staan de vectoren e_r en e_θ in ieder punt onderling loodrecht[5], met elk lengte 1, en wijzen in de richtingen waarin het punt $(x, y) = (r\cos\theta, r\sin\theta)$ loopt als je r respectievelijk θ varieert. De voorfactor $\frac{1}{r}$ compenseert de met r evenredige snelheid bij gelijkmatige toename van θ.

Exercise 30.7. In (30.27) staan twee representaties van dezelfde operator. Door e_x en e_y in e_r en e_θ uit te drukken en Opgave 30.6 te gebruiken kun je zien dat ze inderdaad hetzelfde zijn. Doe dat. Schrijf ook $V = V_x e_x + V_y e_y$ om als $V = V_r e_r + V_\theta e_\theta$.

Exercise 30.8. Laat zien dat de divergentie in poolcoördinaten wordt gegeven door

$$\nabla \cdot V = (e_r \frac{\partial}{\partial r} + \frac{1}{r} e_\theta \frac{\partial}{\partial \theta}) \cdot (V_r e_r + V_\theta e_\theta) = \frac{\partial V_r}{\partial r} + \frac{V_r}{r} + \frac{1}{r}\frac{\partial V_\theta}{\partial \theta}.$$

Hint: Omdat e_r en e_θ van θ afhangen werkt met de produktregel van Leibniz de $e_\theta \frac{\partial}{\partial \theta}$ in de factor links nu ook op e_r en e_θ in de factor rechts, en één van die twee geeft na inprodukt met de voorfactor e_θ een bijdrage.

Exercise 30.9. Pas de regel in Opgave 30.8 nu toe op ∇ zelf en laat zien dat

$$\Delta = \nabla \cdot \nabla = \underbrace{\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}}_{\Delta_r} + \frac{1}{r^2}\underbrace{\frac{\partial^2}{\partial \theta^2}}_{\Delta_S}$$

Hint: wellicht eerst Opgave 30.8 toepassen op ∇ als werkend op de afhankelijke grootheid $G = F$, waarvoor de natuurkundige dezelfde letter gebruikt en de wiskundige dan met $G(r, \theta) = F(r\cos\theta, r\cos\theta)$ in de war raakt, omdat G en F niet dezelfde functies zijn.

In Opgave 30.9 zien we

$$\Delta = \Delta_r + \frac{1}{r^2}\Delta_S, \tag{30.28}$$

[5]Wiskundig is dit per definitie en consistent met wat je ziet als je pijltjes tekent.

waarin Δ_r de radiële Laplaciaan is, die ook werkt op functies $R = R(r)$, en Δ_S de Laplace-Beltrami operator op

$$S = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\},$$

uitgedrukt in de hoekvariabele θ als

$$\Delta_S = \frac{\partial^2}{\partial\theta^2}.$$

Het aardige nu is dat de integraal van de Laplaciaan van een nette functie

$$u(x, y) = u(r\cos\theta, r\sin\theta)$$

over een disk B_R met straal $R > 0$ in poolcoördinaten meteen tot een belangrijke conclusie leidt, maar daarvoor moeten we eerst weten wat meervoudige integralen zijn.

30.3 Harmonische polynomen

We vinden deze polynomen ook als we de Laplace vergelijking

$$u_{xx} + u_{yy} = 0$$

voor $u = u(x, y)$ met *scheiding van variabelen* in poolcoördinaten oplossen door de operator in Opgave 30.9 los te laten op

$$u(x, y) = R(r)\Theta(\theta), \tag{30.29}$$

en het resultaat gelijk aan nul te stellen. Dit geeft

$$\begin{aligned} 0 &= (\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2})R(r)\Theta(\theta) \\ &= \Theta(\theta)(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r})R(r) + \frac{R(r)}{r^2}\frac{\partial^2}{\partial\theta^2}\Theta(\theta) \\ &= \Theta(\theta)(R''(r) + \frac{1}{r}R'(r)) + \frac{R(r)}{r^2}\Theta''(\theta). \end{aligned}$$

Als Θ'' een veelvoud is van Θ, zeg

$$-\Theta'' = \mu\Theta \tag{30.30}$$

dan volgt Euler's vergelijking

$$R''(r) + \frac{1}{r}R'(r) = \mu\frac{R(r)}{r^2} \tag{30.31}$$

voor $R(r)$.

Merk op dat (30.30) gezien kan worden als een (eigenwaarde)probleem voor

$$-\Delta_S = -\frac{d^2}{d\theta}$$

op de eenheidscirkel waar bij Θ een 2π-periodieke functie moet zijn om een functie op de cirkel

$$S = \{(x, y) : x^2 + y^2 = 1\}$$

te definiëren.

Exercise 30.10. Welke μ zijn toegestaan in (30.30) voor oplossingen (30.29) die op heel $\mathbb{R}^2$ zijn gedefinieerd? Leg uit dat je die waarden ook meteen[6] aan de harmonische polynomen kunt zien zonder de precieze vorm van (30.30) te kennen. Schrijf die harmonische polynomen in gescheiden variabelen r en θ als $R(r)\Theta(\theta)$ en verifieer dat $R(r)$ een oplossing is van (30.31) met de bijbehorende μ.

Exercise 30.11. Voor elke $N \in \mathbb{N}$ en $a_0, \ldots, a_N$, $b_1, \ldots, b_N$ in $\mathbb{R}$ is

$$\frac{a_0}{2} + \sum_{k=1}^{N} (a_k \cos k\theta + b_k \sin k\theta) r^n$$

via $x = r\cos\theta, y = r\sin\theta$ een harmonische functie. Overtuig jezelf van de juistheid van de informele uitspraak dat deze oplossing in $(0, 0)$ gelijk is aan zijn gemiddelde op elke disk met middelpunt $(0, 0)$.

Opgave 30.11 suggereert

$$u(x, y) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos k\theta + b_k \sin k\theta) r^n$$

als een algemene oplossing voor de Laplacevergelijking op de eenheidsdisk met randvoorwaarde

$$u(\cos\theta, \sin\theta) = f(\theta) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos k\theta + b_k \sin k\theta), \tag{30.32}$$

een zogenaamde *Fourierreeks*[7] voor een 2π-periodieke functie $\theta \to f(\theta)$. Ook deze $u(x, y)$ is dan in $(x, y) = (0, 0)$ het gemiddelde van $u(x, y)$ op elke disk met middelpunt $(0, 0)$ en straal voldoende klein, kleiner dan 1 in dit geval.

[6] In $\mathbb{R}^3$ eigenwaarden en -functies van Laplace-Beltrami operator ook via polynomen.

[7] Uitgebreid behandeld in de mamanotes van vorig jaar.

Exercise 30.12. In $\mathbb{R}^3$ gebruiken we *bolcoördinaten*

$$x = r\sin\theta\cos\phi;$$
$$y = r\sin\theta\sin\phi;$$
$$z = r\cos\theta,$$

en

$$e_r = \sin\theta\cos\phi\, e_x + \sin\theta\sin\phi\, e_y + \cos\theta\, e_z$$
$$e_\theta = \cos\theta\cos\phi\, e_x + \cos\theta\sin\phi\, e_y - \sin\theta\, e_z$$
$$e_\phi = -\sin\phi\, e_x + \cos\phi\, e_y.$$

Schrijf e_r, e_θ, e_ϕ al of niet als kolomvectoren, en verifieer dat

$$e_r \cdot e_r = e_\theta \cdot e_\theta = e_\phi \cdot e_\phi = 1; \quad e_r \cdot e_\theta = e_r \cdot e_\phi = e_\theta \cdot e_\phi = 0.$$

Overtuig jezelf van

$$\nabla = e_r\frac{\partial}{\partial r} + \frac{1}{r}e_\theta\frac{\partial}{\partial\theta} + \frac{1}{r\sin\theta}e_\phi\frac{\partial}{\partial\phi}, \tag{30.33}$$

en gebruik (30.33) om voor

$$V = V_r e_r + V_\theta e_\theta + V_\phi e_\phi$$

eerst af te leiden dat

$$\nabla\cdot V = \frac{\partial V_r}{\partial r} + \frac{2}{r}V_r + \frac{1}{r}(\frac{\partial V_\theta}{\partial\theta} + \frac{\cos\theta}{\sin\theta}V_\theta + \frac{1}{\sin\theta}\frac{\partial V_\phi}{\partial\phi}),$$

en vervolgens via $V = \nabla F$ dat

$$\Delta = \underbrace{\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}}_{\Delta_r} + \frac{1}{r^2}(\underbrace{\frac{\partial^2}{\partial\theta^2} + \frac{\cos\theta}{\sin\theta}\frac{\partial}{\partial\theta} + \frac{1}{\sin\theta}\frac{\partial^2}{\partial\phi^2}}_{\Delta_S}).$$

Wederom zien we hier (30.9), maar nu met Δ_S gedefinieerd op

$$S = \{(x,y) \in \mathbb{R}^2 :\ x^2 + y^2 + z^2 = 1\},$$

De formules in $\mathbb{R}^n$ laten zich nu raden, afgezien wellicht van de exacte vorm van Δ_S in de hoekvariabelen $\theta_1, \ldots, \theta_{n-1}$, maar met

$$\Delta_r = \frac{\partial^2}{\partial r^2} + \frac{n-1}{r}\frac{\partial}{\partial r}$$

voor het radiële gedeelte.

Exercise 30.13. In $\mathbb{R}^3$ gebruiken we bolcoördinaten

$$x = r\sin\theta\cos\phi;$$

$$y = r\sin\theta\sin\phi;$$

$$z = r\cos\theta,$$

en

$$e_r = \sin\theta\cos\phi\, e_x + \sin\theta\sin\phi\, e_y + \cos\theta\, e_z$$

$$e_\theta = \cos\theta\cos\phi\, e_x + \cos\theta\sin\phi\, e_y - \sin\theta\, e_z$$

$$e_\phi = -\sin\phi\, e_x + \cos\phi\, e_y.$$

Schrijf e_r, e_θ, e_ϕ al of niet als kolomvectoren en verifieer dat

$$e_r \cdot e_r = e_\theta \cdot e_\theta = e_\phi \cdot e_\phi = 1; \quad e_r \cdot e_\theta = e_r \cdot e_\phi = e_\theta \cdot e_\phi = 0.$$

Overtuig jezelf van

$$\nabla = e_r\frac{\partial}{\partial r} + \frac{1}{r}e_\theta\frac{\partial}{\partial\theta} + \frac{1}{r\sin\theta}e_\theta\frac{\partial}{\partial\phi} \tag{30.34}$$

en gebruik (30.34) om voor

$$V = V_r e_r + V_\theta e_\theta + V_\phi$$

eerst af te leiden dat

$$\nabla V = \frac{\partial V_r}{\partial r} + \frac{2}{r}V_r + \frac{1}{r}(\frac{\partial V_\theta}{\partial\theta} + \frac{\cos\theta}{\sin\theta}V_\theta + \frac{1}{\sin\theta}\frac{\partial V_\phi}{\partial\phi}),$$

en vervolgens via $V = \nabla F$ dat

$$\Delta = \underbrace{\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}}_{\Delta_r} + \frac{1}{r^2}(\underbrace{\frac{\partial^2}{\partial\theta^2} + \frac{\cos\theta}{\sin\theta}\frac{\partial}{\partial\theta} + \frac{1}{\sin\theta}\frac{\partial^2}{\partial\phi^2}}_{\Delta_S}).$$

Wederom zien we hier (30.9), maar nu met Δ_S gedefinieerd op

$$S = \{(x,y) \in \mathbb{R}^2 :\ x^2 + y^2 + z^2 = 1\},$$

De formules in $\mathbb{R}^n$ laten zich nu raden, afgezien wellicht van de exacte vorm van Δ_S in de hoekvariabelen $\theta_1, \ldots, \theta_{n-1}$, maar met

$$\Delta_r = \frac{\partial^2}{\partial r^2} + \frac{n-1}{r}\frac{\partial}{\partial r}$$

voor het radiële gedeelte.

30.4 Derivation of the heat equation

For[8] some bounded domain $\Omega \subset \mathbb{R}^n$ we denote by $\varepsilon(t,x)$ the thermal energy density and by w the heat flux. Given any ball or box $B \subset \Omega$ with outside normal ν on ∂B this means that

$$\frac{d}{dt}\int_B \varepsilon = -\int_{\partial B} \nu \cdot w.$$

The Gauss Divergence Theorem[9] turns the integral on the right into an integral over B. The term on the left becomes the integral of the partial time derivative of ε, basically as in Section 14.2. This leads to

$$\int_B (\varepsilon_t + \nabla \cdot w) = 0$$

for every such B and therefore[10] to

$$\varepsilon_t + \nabla \cdot w = 0 \tag{30.35}$$

in Ω.

Physics also tells us that the energy density is given by

$$\varepsilon(t,x) = \sigma(x)u(t,x), \quad \text{in which } u \text{ is temperature and}$$

$$\sigma(x) = \rho(x)\chi(x), \quad \text{with } \chi \text{ the specific heat capacity and } \rho \text{ the density.}$$

Fourier's cooling law says that

$$w = -\kappa \nabla u, \quad \kappa = \kappa(x) > 0 \text{ thermal conductivity.} \tag{30.36}$$

If ν is an outward pointing normal vector on the (smooth) boundary $\partial\Omega$, then

$$\nu \cdot w = -\nu \cdot \kappa \nabla u$$

is the outward heat flux at the boundary.

Equations (30.35) and (30.36), and a possible heat source $h = h(t,x)$, lead to the linear partial differential equation

$$(\chi\rho u)_t = \nabla \cdot \kappa \nabla u + h \tag{30.37}$$

[8] This section relates to Section 4.1 in Olver's PDE book.

[9] See Chapter 21.3.

[10] Certainly if the integrand is continuous.

for the temperature. *Only* if χ, ρ, κ are independent of x this reduces, in the absence of heat sources, to

$$u_t = \gamma \Delta u, \quad \gamma = \frac{\kappa}{\chi \rho}. \tag{30.38}$$

Before we scale the variables to have $\gamma = 1$ we discuss the boundary conditions.

Either u or the outward heat flux $\nu \cdot \kappa \nabla u$ prescribed as function of $x \in \partial\Omega$ and $t > 0$ lead to the Dirichlet and Neumann initial boundary value problems for (30.37). The standard homogeneous boundary conditions are therefore

$$u = 0 \quad \text{on} \quad \partial\Omega \qquad \text{(Dirichlet)} \tag{30.39}$$

and

$$\nu \cdot \nabla u = 0 \quad \text{on} \quad \partial\Omega \qquad \text{(Neumann)} \tag{30.40}$$

for $t > 0$.

The Robin boundary condition prescribes the flux in terms of also the temperature, e.g. the homogenous boundary condition

$$\nu \cdot \kappa \nabla u + \beta u = 0 \tag{30.41}$$

is Newton's cooling law. If the outside temperature is equal to zero, it relates the outward flux to the temperature inside via some heat exchange constant $\beta > 0$. With $\beta = 0$ it reduces to (30.40). Note that Olver writes this condition with $\kappa = 1$. Initial data for $u(0, x)$, $x \in \Omega$, complete the *initial boundary value problem* formulation.

Each of the natural homogenous boundary conditions above allows for separation of variables to solve (30.38). Without loss of generality we now assume that $\gamma = 1$ and $\kappa = 1$. We then have that

$$u(t, x) = e^{-\lambda t} v(x)$$

is a solution of

$$u_t = \Delta u$$

if

$$-\Delta v = \lambda v. \tag{30.42}$$

There are now two natural boundary conditions to choose from,

$$u = 0 \qquad \text{(Dirichlet) and} \qquad \nu \cdot \nabla u + \beta u = 0 \qquad \text{(Robin)}. \tag{30.43}$$

The Neumann boundary condition corresponds to $\beta = 0$.

30.5 Intermezzo: het waterstofatoom

Met

$$V(r) = -\frac{e^2}{r}$$

is de stationaire Schrödinger vergelijking voor het waterstofatoom

$$\frac{\hbar^2}{2m}\Delta\psi - \frac{e^2}{r}\psi = E\psi, \tag{30.44}$$

waarin m de massa van het electron is, e de lading van het electron, $\hbar$ de constante van Planck. De negatieve waarden van E waarvoor (30.44) een oplossing met

$$\iiint_{\mathbb{R}^3} |\psi(x,y,z)|^2 \, d(x,y,z) = 1$$

heeft zijn de energieniveaus die het electron in gebonden toestand kan aannemen.

We hebben gezien dat

$$\Delta = \underbrace{\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}}_{\Delta_r} + \frac{1}{r^2}\Big(\underbrace{\frac{\partial^2}{\partial\theta^2} + \frac{\cos\theta}{\sin\theta}\frac{\partial}{\partial\theta} + \frac{1}{\sin\theta}\frac{\partial^2}{\partial\phi^2}}_{\Delta_S}\Big).$$

Via

$$\psi(x,y,z) = R(r)P_l\Big(\frac{x}{r}, \frac{y}{r}, \frac{z}{r}\Big),$$

waarin $P_l(x,y,z) = Y(\theta,\phi)$ een harmonisch homogeen polynoom van graad l in x,y,z is, en een nieuwe x en n gedefinieerd door

$$x = \frac{1}{\hbar}\sqrt{-2mE}\, r \quad \text{en} \quad -E = \frac{me^4}{2\hbar^2 n^2},$$

leidt dit tot

$$\frac{d^2R}{dx^2} + \frac{2}{x}\frac{dR}{dx} - \frac{l(l+1)}{x^2}R + \frac{2n}{x}R = R$$

met $R(x) \sim x^l$ voor $x \to 0$ en $R(x) \sim e^{-x}$ voor $x \to \infty$.

Substitueer daarom $R(x) = x^l e^{-x} u(x)$ en leidt voor $u(x)$ af dat

$$\frac{d^2u}{dx^2} + \Big(\frac{4l}{x} - 2\Big)\frac{du}{dx} = 2\frac{n-l-1}{x}u.$$

Exercise 30.14. Corrigeer eventuele typo's hierboven. De machtreeksoplossing[11]

$$u(x) = 1 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots$$

breekt af voor een n die van l afhangt. Welke n is dat?

[11] Instructief om eerst $\frac{d^2R}{dx^2} + \frac{2}{x}\frac{dR}{dx} = R$ op te lossen.

31 Functional calculus

31.1 Lijnintegralen over polygonen en Goursat

We beginnen met formule (13.1) uit Sectie 10.1, al of niet via de "verboden" operaties daarboven, geschreven als

$$F(x_1) - F(x_0) = \int_0^1 \underbrace{F'((1-t)x_0 + tx_1)}_{f(x(t))} \underbrace{(x_1 - x_0)dt}_{dx} = \int_{x_0}^{x_1} f(x)\,dx,$$

waarin, met $x, x(t), x_0, x_1, f(x)$ vervangen door $z, z(t), z_0, z_1, f(z)$,

$$t \to z(t) = (1-t)z_0 + tz_1 \tag{31.1}$$

het interval $[z_0, z_1]$ parametriseert, en

$$\int_0^1 f(z(t))z'(t)dt =$$

$$\int_0^1 f((1-t)z_0 + tz_1)dt\,(z_1 - z_0) = \int_{z_0}^{z_1} f(z)\,dz. \tag{31.2}$$

Dit is een formule die we, zonder dat (13.1) daarvoor nog nodig is, nu kunnen lezen met $z_0, z_1 \in \mathbb{C}$ en $f : \mathbb{C} \to \mathbb{C}$, met het linkerlid als ondubbelzinnige definitie van het rechterlid: de lijnintegraal

$$\int_{z_0}^{z_1} f(z)\,dz$$

over het rechte lijnstuk van z_0 naar z_1, van de functie $z \to f(z)$. Niet meer praten over andere parametervoorstellingen van $[z_0, z_1]$ dan (31.1), tenzij het nodig[1] is zou ik zeggen. Merk op dat $[z_0, z_1]$ voor alle $z_0, z_1 \in \mathbb{C}$ is gedefinieerd, dus $[1, 0]$ heeft nu ook betekenis. Je moet er even aan wennen maar het spreekt vanzelf. Het ligt voor de hand om aan z_0 als het begin- en z_1 als het eindpunt van $[z_0, z_1]$ te denken. Daarmee wordt $[z_0, z_1]$ meer dan alleen een verzameling: $[z_0, z_1]$ is zo een georienteerd lijnstuk.

Exercise 31.1. Laat zien dat

$$\int_{z_0}^{z_1} z\,dz = \frac{1}{2}(z_1^2 - z_0^2).$$

Evalueer vervolgens $\int_{z_0}^{z_1} z^n\,dz$ voor alle $n \in \mathbb{N}$.

[1] Quod non.

Exercise 31.2. Evalueer $\int_{z_0}^{z_1} z^n\,dz$ voor alle $n \in -\mathbb{N} = \{-1, -2, -3, \cdots\}$. Doe $n = -1$ als laatste[2]. Welke voorwaarde moet je leggen op z_0 en z_1?

Exercise 31.3. Laat zien dat

$$|\int_{z_0}^{z_1} f(z)\,dz| \leq |z_0 - z_1| \max_{z\in[z_0,z_1]} |f(z)|.$$

Integralen gedefinieerd als hierboven door (31.2) kunnen we rijgen tot een integraal over een *polygonaal pad*

$$z_0 \to z_1 \to \cdots \to z_n$$

middels

$$\int_{z_0}^{z_1} f(z)\,dz + \int_{z_1}^{z_2} f(z)\,dz + \cdots + \int_{z_{n-1}}^{z_n} f(z)\,dz = \int_{z_0,\ldots,z_n} f(z)\,dz, \quad (31.3)$$

als $z \to f(z)$ een continue functie

$$[z_0, z_1] \cup \cdots \cup [z_{n-1}, z_n] \xrightarrow{f} \mathbb{C}$$

definieert. In het bijzondere geval dat $z_0 = z_n$ is eenvoudig na te gaan dat hier NUL uitkomt als $n = 2$.

Exercise 31.4. Ga dit na. Hint: neem eerst $z_0 = z_2, z_1 \in \mathbb{R}$ om te zien hoe het moet werken.

Wat deze opgave zegt is dat heen en weer lopen geen bijdrage aan een keten als in (31.3) geeft omdat vrijwel per definitie

$$\int_{z_0,z_1,z_0} f(z)\,dz = \int_{z_0}^{z_1} f(z)\,dz + \int_{z_1}^{z_0} f(z)\,dz = 0,$$

voor *iedere* $[z_0, z_1] \xrightarrow{f} \mathbb{C}$ continu. We kunnen integralen dus vereenvoudigen door heen en weer stukjes weg te laten, ook als die niet achter elkaar zitten in het polygonale pad.

[2]De eersten zullen de laatsten zijn.

Neem nu in je complexe vlak z_0, z_1, z_2, not all on a line[3], en neem aan dat

$$\Delta = \Delta_{z_0,z_1,z_2} = \{t_0 z_0 + t_1 z_1 + t_2 z_2 : t_0, t_1, t_2 \geq 0, t_0 + t_1 + t_2 = 1\} \xrightarrow{f} \mathbb{C}$$

continu is. De verzameling Δ bestaat dus uit alle convexe combinaties van z_0, z_1, z_2 gezien als punten in het complexe vlak. Dat is een driehoekige tegel waarvan de rand een driehoek[4] is.

Teken plaatje!

Laat z_3, z_4, z_5 de middens zijn van $[z_0, z_1]$, $[z_1, z_2]$, $[z_2, z_0]$, die je krijgt door achtereenvolgens t_2, t_0, t_1 nul, en steeds de andere twee t-tjes $\frac{1}{2}$ te kiezen. Dan is

$$\int_{z_0,z_1,z_2,z_0} f(z)\,dz = \int_{z_0,z_3,z_5,z_4,z_3,z_1,z_4,z_2,z_5,z_0} f(z)\,dz =$$

$$\int_{z_0,z_3,z_5,z_0} f(z)\,dz + \int_{z_3,z_4,z_5,z_3} f(z)\,dz + \int_{z_3,z_1,z_4,z_3} f(z)\,dz + \int_{z_5,z_4,z_2,z_5} f(z)\,dz.$$

De integraal over het gesloten[5] pad

$$z_0 \to z_3 \to z_5 \to z_4 \to z_3 \to z_1 \to z_4 \to z_2 \to z_5 \to z_0$$

is zo enerzijds gelijk aan de integraal over het gesloten pad

$$z_0 \to z_1 \to z_2 \to z_0$$

rond de grote driehoek, en anderzijds de som van vier integralen over de vier gesloten paden

$$z_0 \to z_3 \to z_5 \to z_0, \quad z_3 \to z_4 \to z_5 \to z_3,$$

$$z_3 \to z_1 \to z_4 \to z_3, \quad z_5 \to z_4 \to z_2 \to z_5$$

rond de vier kleinere driehoeken.

Als de oorspronkelijke integraal niet nul was, zeg gelijk[6] aan 1, dan is tenminste één van de vier integralen in absolute waarde minstens gelijk aan $\frac{1}{4}$, en kan daarna met die integraal het argument herhaald worden om een rij geneste driehoeken

$$\Delta = \Delta_{z_0,z_1,z_2} \supset \Delta_{z_0^{(1)},z_1^{(1)},z_2^{(1)}} \supset \Delta_{z_0^{(2)},z_1^{(2)},z_2^{(2)}} \supset \Delta_{z_0^{(3)},z_1^{(3)},z_2^{(3)}} \supset \cdots$$

te maken met

$$\left|\int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} f(z)\,dz\right| \geq \frac{1}{4^k}$$

voor $k = 0, 1, 2, 3, \ldots$.

[3]Sommigen horen hierbij de stem van Erdös.

[4]Vereniging van 3 lijnstukjes: $[z_0, z_1] \cup [z_1, z_2] \cup [z_2, z_0]$, met zo gewenst een orientatie.

[5]Teken het in je plaatje.

[6]Door $f(z)$ te delen door zijn integraal over de rand van Δ_{z_0,z_1,z_2}.

Exercise 31.5. Bewijs dat de rijen $z_0^{(k)}, z_1^{(k)}, z_2^{(k)}$ convergeren naar een limiet in Δ_{z_0,z_1,z_2} als $k \to \infty$.

Zonder beperking der algemeenheid mogen we nu wel aannemen[7] dat deze limiet gelijk is aan 0, i.e.

$$z_0^{(k)}, z_1^{(k)}, z_2^{(k)} \to 0.$$

Kan het zo zijn dat f complex differentieerbaar is in 0? Zo ja, dan geldt voor $z \in \Delta$ dat

$$f(z) = f'(0)z + R(z)$$

met $R(z) = o(|z|)$ als $|z| \to 0$.

Maar dan is

$$\int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} f(z)\,dz = \int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} f'(0)z\,dz + \int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} R(z)\,dz$$

$$= \int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} R(z)\,dz,$$

omdat de eerste integraal nul is vanwege (31.1) toegepast op $\int_{z_0}^{z_1} z\,dz$, $\int_{z_1}^{z_2} z\,dz$, $\int_{z_2}^{z_0} z\,dz$. Dus volgt

$$\frac{1}{4^k} \leq |\int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} R(z)\,dz| \leq \frac{|z_0 - z_1| + |z_1 - z_2| + |z_2 - z_0|}{2^k} \max_{z \in \delta\Delta^{(k)}} |R(z)|,$$

waarin $\delta\Delta^{(k)}$ de rand is van

$$\Delta^{(k)} = \Delta_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} \ni 0.$$

Omdat $|z|$ op zijn hoogste gelijk is aan de grootste afstand tussen twee punten in $\Delta^{(k)}$ geldt

$$|z| \leq \frac{d}{2^k} \quad \text{met} \quad d = \max_{z,w \in \Delta} |z - w|,$$

en samen met de 2^k die we al hadden krijgen nu ook een bovengrens voor de integraal, met een 4^k in de noemer en een $\varepsilon > 0$ naar keuze in de teller:

Exercise 31.6. Gebruik (de definitie van) $|R(z)| = o(|z|)$ als $|z| \to 0$ om met de laatste twee ongelijkheden een tegenspraak te forceren.

[7] Schuif de boel anders op.

Theorem 31.7. *Voor iedere $f : \Delta_{z_0,z_1,z_2} \to \mathbb{C}$ die complex differentieerbaar is op de gesloten*[8] *driehoek Δ_{z_0,z_1,z_2} met hoekpunten $z_0, z_1, z_2 \in \mathbb{C}$ geldt dat*

$$\oint_{z_{012}} f(z)\,dz = \int_{z_0,z_1,z_2,z_0} f(z)\,dz = 0,$$

nu ook met een notatie[9] *die wellicht hierboven al voor de hand lag.*

Eenzelfde uitspraak geldt voor iedere n-hoek ($n > 3$) gemaakt uit n driehoeken

$$\Delta_{w_0,z_0,z_1}, \Delta_{w_0,z_1,z_2}, \ldots, \Delta_{w_0,z_{n-1},z_0},$$

met

$$\Delta_{w_0,z_0,z_1} \cap \Delta_{w_0,z_1,z_2} = [w_0, z_1], \quad \Delta_{w_0,z_1,z_2} \cap \Delta_{w_0,z_2,z_3} = [w_0, z_2],$$

$$\ldots$$

$$\Delta_{w_0,z_{n-2},z_{n-1}} \cap \Delta_{w_0,z_{n-1},z_n} = [w_0, z_{n-1}], \quad \Delta_{w_0,z_{n-1},z_0} \cap \Delta_{w_0,z_0,z_1} = [w_0, z_0].$$

Als de lijnstukjes waarmee (31.3) is gemaakt de zijden zijn van zo'n door

$$z_n = z_0, \ldots, z_{n-1} \tag{31.4}$$

gemaakte n-hoek met een punt w_0 in de n-hoek waarvoor alle $[w_0, z_k]$ in de veelhoek liggen[10], dan is

$$\oint_{z_0,\ldots,z_n} f(z)\,dz = \sum_{k=1}^{n} \oint_{w_0,z_{k-1},z_k,w_0} f(z)\,dz.$$

Als $z \to f(z)$ complex differentieerbaar is in elk punt van

$$P_{z_0,\ldots,z_{n-1}} = \cup_{k=1}^{n} \Delta_{w_0,z_{k-1},z_k},$$

dan volgt zo dat

$$\oint_{z_n=z_0,\ldots,z_n} f(z)\,dz = 0. \tag{31.5}$$

Bovendien kan ieder punt z_k dan met de andere punten vastgehouden naar binnen geschoven worden naar een $\tilde{z}_k$ in de door (31.4) gemaakte veelhoek, waarbij (31.5) niet verandert met eenzelfde argument waarin driehoeken met

[8] Voor w op de rand $f(z) = f(w) + f'(w)(z-w) + R(z;w)$ met $z \in \Delta_{z_0,z_1,z_2}$ lezen.

[9] Zie het rondje door de integraalslang heen maar als driehoekje.

[10] De veelhoek is dan convex.

hoekpunten $z_{k-1}, z_k, \tilde{z}_k, z_{k+1}$ voorkomen. Kortom, (31.5) geldt voor iedere complex differentieerbare

$$f : P_{z_0,\ldots,z_{n-1}} \to \mathbb{C}$$

op ieder domein $P_{z_0,\ldots,z_{n-1}}$ begrensd door lijnstukjes $[z_{k-1}, z_k] \subset P_{z_0,\ldots,z_{n-1}}$.

Merk op dat we op een vanzelfsprekende manier kunnen praten over het links- of rechtsom genummerd zijn van de hoekpunten (31.4), ook als de n-hoek niet gemaakt is zoals boven (31.4) beschreven is. Als er er (maar) eindig veel punten $\zeta_1, \ldots, \zeta_p$ zijn in $P_{z_0,\ldots,z_{n-1}}$ (maar niet op de rand van) waar f niet complex differentieerbaar is, dan het niet moeilijk om na te gaan dat (31.5) gelijk is aan de som van de integralen over de randen van kleine driehoekjes

$$\Delta^{(j)} = \Delta_{z_0^{(j)},z_1^{(j)},z_2^{(j)},z_0^{(j)}}$$

in $P_{z_0,\ldots,z_{n-1}}$ waar ζ_j echt in ligt, als die allemaal maar met dezelfde orientatie genomen worden. In dat geval is dus

$$\oint_{z_n=z_0,\ldots,z_n} f(z)\,dz = \sum_{j=1}^{p} \oint_{z_2^{(j)}=z_0^{(j)},z_1^{(j)},z_2^{(j)},} f(z)\,dz, \tag{31.6}$$

en we werken dat idee in het geval dat $p = 1$ met een hele bijzondere integrand nu uit in de volgende sectie, teneinde uiteindelijk ook te zien dat de termen in het rechterlid van (31.6) een verrassend simpele vorm krijgen.

31.2 Machtreeksen via een Cauchy integraalformule

Neem nu aan dat

$$D = \{z \in \mathbb{C} : |z| < 1\} \xrightarrow{f} \mathbb{C}$$

een complex differentieerbare functie is op de open eenheidsdisk. Neem een $\zeta \in \mathbb{C}$ met $|\zeta| = \rho < 1$. We laten nu rechtstreeks zien dat

$$f(\zeta) = \sum_{n=0}^{\infty} a_n \zeta^n,$$

met integraalformules voor de coëfficiënten $a_n \in \mathbb{C}$. De integralen zijn daarbij over regelmatige polygonen in D, met hoekpunten dicht bij de cirkelvormige[11] rand van D.

We leiden de formules of door Stelling 31.7 toe te passen op integralen van

$$z \to \frac{f(z) - f(\zeta)}{z - \zeta}$$

[11]Denk aan deze contouren: http://www.fi.uu.nl/publicaties/literatuur/7244.pdf

over geschikt gekozen driehoeken. Voor iedere $r \in (0,1)$ en $n \in \mathbb{N}$ met $n \geq 3$ definiëren de punten

$$z_k = r \exp(2\pi i \frac{k}{n})$$

de hoekpunten van een regelmatig n-hoek $C_{r,n}$ in D met middelpunt 0. Maak een plaatje met $0 < |\zeta| < r$ en $n = 8$ of zo. Laat k van 0 tot en met n lopen om het kringetje rond[12] te maken. *Teken plaatje!*

Op dezelfde manier maken

$$w_k = \zeta + \rho \exp(2\pi i \frac{k}{n})$$

de hoekpunten van een regelmatig n-hoek $\zeta + C_{\rho,n}$, met middelpunt ζ dat binnen $C_{r,n}$ ligt als $\rho < r - |\zeta|$. Teken beide polygonen in je plaatje en merk op dat voor iedere complex differentieerbare functie

$$\{z \in D : z \neq \zeta\} \xrightarrow{g} \mathbb{C}$$

nu geldt dat

$$\oint_{z_{0-n}} g(z)\,dz =$$

$$\int_{z_0,z_1,\ldots,z_n=z_0} g(z)\,dz = \int_{w_0,w_1,\ldots,w_n=w_0} g(z)\,dz \tag{31.7}$$

$$= \oint_{w_{0-n}} g(z)\,dz,$$

voor elke $n \geq 3$, waarbij we ook hier de voor de hand liggende notatie[13] met $\oint$ gebruiken voor de integralen van $g(z)$ over de linksom[14] doorlopen n-hoeken.

Exercise 31.8. Bewijs (31.7) door de zigzagintegraal

$$\int_{w_0,z_0,w_1,z_1,\ldots,w_n,z_n} g(z)\,dz$$

mee te nemen in de beschouwingen en de stelling van Goursat toe te passen op in totaal $2n$ driehoekjes.

[12] Nou ja, rond...

[13] Voor grote n is de veelhoek bijna een rondje.

[14] Dat is maar een woord hier, de punten bepalen de richting.

Nu nemen we voor $g(z)$ het differentiaalquotiënt

$$\frac{f(z)-f(\zeta)}{z-\zeta}$$

en concluderen dat

$$\oint_{w_{0-n}} \frac{f(z)-f(\zeta)}{z-\zeta}\,dz = \oint_{z_{0-n}} \frac{f(z)-f(\zeta)}{z-\zeta}\,dz. \tag{31.8}$$

Exercise 31.9. De integraal in het linkerlid van (31.8) hangt van ρ af. Gebruik de differentieerbaarheid van f in ζ om te laten zien dat

$$\oint_{w_{0-n}} \frac{f(z)-f(\zeta)}{z-\zeta}\,dz \to 0$$

als $\rho \to 0$.

Maar de integraal in het linkerlid van (31.8) is gelijk aan de integraal in het rechterlid en hing niet van ρ af als $\rho < r - |\zeta|$. Hij ging[15] dus niet naar 0 want hij was al 0. Zo reduceert (31.8) tot

$$0 = \oint_{z_{0-n}} \frac{f(z)}{z-\zeta}\,dz - \oint_{z_{0-n}} \frac{f(\zeta)}{z-\zeta}\,dz,$$

en volgt

$$f(\zeta)\oint_{z_{0-n}} \frac{1}{z-\zeta}\,dz = \oint_{z_{0-n}} \frac{f(z)}{z-\zeta}\,dz.$$

Exercise 31.10. Laat zien dat

$$\oint_{z_{0-n}} \frac{1}{z-\zeta}\,dz = \oint_{w_{0-n}} \frac{1}{z-\zeta}\,dz = 2\pi i.$$

Hint: de eerste gelijkheid volgt als in Opgave 31.8 en de tweede integraal hangt niet van ζ of ρ af. In het linkerlid kan dus $\zeta = 0$ genomen worden. Op elke $[z_{k-1}, z_k]$ heeft $\frac{1}{z}$ een primitieve: de meerwaardige functie gedefinieerd in Opgave 18.9 die je als het goed is in Opgave 31.2 als laatste hebt gebruikt.

[15] Letterlijk gesproken.

Theorem 31.11. *Als ζ ligt binnen een n-hoek zoals hierboven in de open eenheidsdisk D, en $f : D \to \mathbb{C}$ complex differentieerbaar is, dan is*

$$f(\zeta) = \frac{1}{2\pi i} \oint_{z_{0-n}} \frac{f(z)}{z - \zeta}\, dz.$$

This is the Cauchy Integral Formula, maar dan met n-hoeken in plaats van de gebruikelijke cirkels met middelpunt 0 en straal $r < 1$ groot genoeg.

Tenslotte volgt na het invullen van[16] de meetkundige reeksontwikkeling

$$\frac{1}{z - \zeta} = \frac{1}{z} \frac{1}{1 - \frac{\zeta}{z}} = \frac{1}{z} + \frac{\zeta}{z^2} + \frac{\zeta^2}{z^3} + \frac{\zeta^3}{z^4} + \cdots$$

de machtreeksontwikkeling in de vorm als aangekondigd, via

$$f(\zeta) = \frac{1}{2\pi i} \oint_{z_{0-n}} f(z)(\frac{1}{z} + \frac{\zeta}{z^2} + \frac{\zeta^2}{z^3} + \frac{\zeta^3}{z^4} + \cdots)\, dz =$$

$$\frac{1}{2\pi i} \oint_{z_{0-n}} \frac{f(z)}{z}\, dz + \frac{1}{2\pi i} \oint_{z_{0-n}} \frac{f(z)}{z^2}\, dz\, \zeta + \frac{1}{2\pi i} \oint_{z_{0-n}} \frac{f(z)}{z^3}\, dz\, \zeta^2 + \cdots,$$

met

$$a_j = \frac{1}{2\pi i} \oint_{z_{0-n}} \frac{f(z)}{z^{j+1}}\, dz \tag{31.9}$$

voor alle $j \in \mathbb{N}_0$.

Theorem 31.12. *Als $f : D \to \mathbb{C}$ complex differentieerbaar is, dan geldt*

$$f(z) = \sum_{j=0}^{\infty} a_n z^n$$

met a_j gegeven door (31.9), waarin de integraal nu door $z_k = r \exp(2\pi i \frac{k}{n})$ $(k = 0, 1, \ldots, n)$ wordt gedefinieerd. En achteraf zijn dan zowel $r \in (0, 1)$ als $n \geq 3$ arbitrair.

Het rechtvaardigen van het verwisselen van $\oint$ en $\sum$ is wezen niets anders dan opmerken dat voor elke α en β in $\mathbb{C}$ de verzameling

$$\{f : [\alpha, \beta] \to \mathbb{C} : \ f \text{ is continu}\}$$

een (complexe) Banachruimte is, net zoals $C([a, b])$ een reële Banachruimte is. Het wordt dus tijd voor het volgende hoofdstuk.

[16] We prefereren weer de notatie met puntjes natuurlijk.

Voor hier is het nog de vraag of we de limiet $n \to \infty$ willen nemen in Stelling 31.11 en (31.9) teneinde de $\oint$ te nemen over de cirkel geparametriseerd door

$$z = r\exp(i\theta) \quad \text{met} \quad dz = ir\exp(i\theta)d\theta \tag{31.10}$$

und so weiter.

Dat laatste kan komen na de observatie dat voor $0 \leq \rho < R$ en complex differentieerbare

$$A_{\rho,R} = \{z \in \mathbb{C} : \rho < |z| < R\} \xrightarrow{f} \mathbb{C}$$

geldt dat $f(z)$ te schrijven is als een zogenaamde *Laurentreeks*, i.e.

$$f(z) = \sum_{j=-\infty}^{\infty} a_j z^j = \sum_{j=0}^{\infty} a_j z^j + \sum_{j=1}^{\infty} \frac{a_{-j}}{z^j}, \tag{31.11}$$

met a_j gegeven door (31.9) voor *alle* $j \in \mathbb{Z}$, maar $n \geq 3$ en $r \in (\rho, R)$ wel zo gekozen dat met de punten $z_k = r\exp(2\pi i\frac{k}{n})$ de n-hoek in de annulus $A_{\rho,R}$ ligt.

Je bewijst dit met drie veelhoeken in $A_{\rho,R}$, waar ζ dan tussen twee van de drie in moet liggen, en in de kleinste. Als je eenmaal op het idee[17] bent gekomen wijst het zich vanzelf. De integraal over de nieuwe veelhoek wordt ook weer via een net iets andere meetkundige reeks in een machtreeks vertaald, nu met $\frac{1}{\zeta}$ die naar buiten gehaald wordt uit $\frac{1}{z-\zeta}$.

Deze zo verkregen spectaculaire uitspraak wordt gewoonlijk bewezen *na* het invoeren van lijnintegralen over echte krommen[18] zoals gegeven door (31.10) en de hele machinerie die nodig is om netjes te beschrijven wat krommen[19] eigenlijk zijn, waarbij vaak ook de continuiteit van $z \to f'(z)$ wordt gebruikt om de nog te bespreken stellingen van Green te kunnen gebruiken.

Die laatste stellingen zijn dus hier niet nodig. En de veelhoeken bieden veel meer mogelijkheden. Bijvoorbeeld voor functies die gedefinieerd en complex differentieerbaar zijn op gebieden ingesloten door twee geneste veelhoeken. Dat is misschien nog leuk om uit te zoeken.

Exercise 31.13. Natuurlijk geldt Stelling 31.12 niet alleen voor de eenheidsdisk, en kan r zowel zo klein als zo groot mogelijk gekozen worden voor het polygon waarmee de coefficienten worden berekend. Gebruik dit om te bewijzen dat er geen niet-constante begrensde complex differentieerbare functies $f : \mathbb{C} \to \mathbb{C}$ zijn.

[17] Gauss en Cauchy gingen ons voor.....

[18] Denk aan die goal van nummer 14 in het Zuiderpark en de enige echte Kromme.

[19] Denk ook aan hoe Murre dit woord uitspreekt.

31.3 De Cauchy Integraal Transformatie

De formule in Stelling 31.11 schrijven we met $1 = I$ en $\zeta = A$ als

$$f(A) = \frac{1}{2\pi i} \oint_{z_{0-n}} f(z)(zI - A)^{-1}\, dz, \tag{31.12}$$

nu voor een willekeurig polygon waar A binnen ligt en waarop

$$z \to (zI - A)^{-1} \tag{31.13}$$

dus bestaat als zeker een continue functie. Het polygon hoeft ook niet per se in de eenheidsdisk te liggen. Van de functie $z \to f(z)$ hoeven we bij nadere beschouwing alleen maar aan te nemen dat f complex differentieerbaar is op het gebied begrensd door een polygon, inclusief het polygon[20] zelf.

Let op, de hoekpunten van het polygon moeten wel "linksom" genummerd worden, hetgeen ondubbelzinnig gedefinieerd kan worden aan de hand van de vergelijkingen voor de lijnen door de opeenvolgende hoekpunten, met iedere $z_k = x_k + iy_k$ opgevat als $(x_k, y_k) \in \mathbb{R}^2$, waarbij je wil formuleren dat het binnengebied van het polygon steeds links van ieder georienteerde interval $[z_{k-1}, z_k]$ ligt.

Met deze notatie kunnen we (31.12) nu ook lezen lezen met A een vierkante eerst nog reële matrix gezien als een continue lineaire afbeelding van $X = \mathbb{R}^n$ naar zichzelf, afbeeldingen die een algebra[21] vormen. Hier is nog het een en ander mee te doen, met behulp ook van

$$(zI - A)^{-1} = \frac{1}{z}(I + \frac{1}{z}A + \cdots)$$

als $|z|$ voldoende groot is, misschien beter meteen maar voor algemene X in Sectie 31.5.

De vraag is natuurlijk wel eerst wat we precies onder *A binnen het polygon gedefinieerd door*

$$z_0 \to z_1 \to \cdots \to z_n = z_0$$

moeten verstaan, als we (31.12) zomaar overschrijven met ζ vervangen door een lineaire operator $A : X \to X$. Voor de hand ligt dat A zo moet zijn dat met een groter polygon de zigzagtruc weer werkt, en de integrand als $L(X)$-waardige functie complex differentieerbaar is op het gebied tussen de twee polygonen, en ook op de twee polygonen zelf, en dat weer voor ieder groter polygon.

[20] Via een zigzagintegraal als in Opgave 31.8 volgt de geldigheid van (31.12).

[21] Een Banachalgebra zelfs, zie Charlie's teleurstelling in Flowers for Algernon.

Daartoe moeten X en ook $L(X)$ zelf eerst complex uitgebreid worden, hetgeen abstract een constructie vereist maar in voorbeelden automatisch[22] gaat. En daarna is dan de natuurlijke eis dat (31.13) op het polygon en zijn buitengebied moet bestaan in de complexe versie van $L(X)$. Lees wat dit betreft verder in Sectie 31.5.

31.4 Kromme lijnintegralen

Met behulp van (31.2) is in (31.3)

$$\int_{z_0,\ldots,z_n} f(z)\,dz = \sum_{k=1}^{n} \int_{z_{k-1}}^{z_k} f(z)\,dz \tag{31.14}$$

gedefinieerd voor een rij punten die we (nog) niet als partitie zien, waarvoor we ook (nog) niet Riemanntussensommen als

$$\sum_{k=1}^{n} f(\zeta_k)(z_k - z_{k-1}) \quad \text{met} \quad \zeta_k \in [z_{k-1}, z_k] \tag{31.15}$$

hebben ingevoerd. Maar als de "incrementen" $z_k - z_{k-1}$ klein zijn ligt gezien (31.2) iedere term in (31.15) voor de hand als benadering voor de overeenkomstige term in het rechterlid van (31.14) via

$$\int_{z_{k-1}}^{z_k} f(z)\,dz = \int_0^1 f((1-t)z_{k-1} + tz_k)dt\,(z_k - z_{k-1}) \approx f(\zeta_k)(z_k - z_{k-1}).$$

De vraag wat er gebeurt als $n \to \infty$ is echter nog niet goed gesteld, want de rij "partities" kan in principe willekeurig zijn.

In iedere schatting die het limietgedrag onder controle moet krijgen zal, behalve het klein worden van de incrementen, ook het gedrag van

$$\sum_{k=1}^{n} |z_k - z_{k-1}|$$

een rol spelen, met

$$z_k = z_k^{(n)}$$

zinvol afhankelijk van n gekozen, maar wat is zinvol? Hieronder wat overwegingen en een aanzet tot een uitgewerkt antwoord.

[22] Denk hier even over na.

Het stuksgewijs lineaire pad P_n van $z_0^{(n)}$ via $z_1^{(n)}, \ldots, z_{n-1}^{(n)}$, naar $z_n^{(n)}$ voor $n \to \infty$ moet een nog te formuleren limietgedrag hebben, waarmee in ieder geval voor continue $z \to f(z)$ volgt dat

$$\int_{P_n} f(z)\,dz = \sum_{k=1}^{n} \int_{z_{k-1}^{(n)}}^{z_k^{(n)}} f(z)\,dz \quad \text{en} \quad \sum_{k=1}^{n} f(\zeta_k^{(n)})(z_k^{(n)} - z_{k-1}^{(n)}) \tag{31.16}$$

convergeren naar een limiet die we $\int_P f(z)\,dz$ zouden willen noemen.

Voor het verschil van deze sommen geldt

$$|\sum_{k=1}^{n} \int_{z_{k-1}^{(n)}}^{z_k^{(n)}} f(z)\,dz - \sum_{k=1}^{n} f(\zeta_k^{(n)})(z_k^{(n)} - z_{k-1}^{(n)})| =$$

$$|\sum_{k=1}^{n} \int_0^1 f((1-t)z_{k-1}^{(n)} + tz_k^{(n)})dt\,(z_k^{(n)} - z_{k-1}^{(n)}) - \sum_{k=1}^{n} f(\zeta_k^{(n)})(z_k^{(n)} - z_{k-1}^{(n)})| ==$$

$$|\sum_{k=1}^{n} \int_0^1 (f((1-t)z_{k-1}^{(n)} + tz_k^{(n)}) - f(\zeta_k^{(n)}))dt\,(z_k^{(n)} - z_{k-1}^{(n)})| \leq$$

$$\max_{k=1,\ldots,n} |f((1-t)z_{k-1}^{(n)} + tz_k^{(n)}) - f(\zeta_k^{(n)})| \sum_{k=1}^{n} |z_k^{(n)} - z_{k-1}^{(n)}| \leq$$

$$\max_{k=1,\ldots,n} \sup_{z,w \in [z_{k-1}^{(n)}, z_k^{(n)}]} |f(z) - f(w)| \sum_{k=1}^{n} |z_k^{(n)} - z_{k-1}^{(n)}|,$$

en dat zou klein moeten zijn als f uniform continu is op een geschikt gekozen domein dat alle paden P_n bevat. In dat geval zijn de aannames dat

$$\mu_n = \max_{k=1,\ldots,n} |z_k^{(n)} - z_{k-1}^{(n)}| \to 0 \tag{31.17}$$

en

$$L_n = \sum_{k=1}^{n} |z_k^{(n)} - z_{k-1}^{(n)}| \quad \text{begrensd} \tag{31.18}$$

is als $n \to \infty$ voldoende om het verschil tussen de termen in (31.16) naar 0 te doen gaan als $n \to \infty$.

Voor we een definitie geven bekijken we wat we langs deelrijen sowieso kunnen bereiken kwa convergentie van P_n onder de aanname dat (31.17) en (31.18) gelden, en

$$z_0^{(n)} = a \quad \text{en} \quad z_n^{(n)} = b \tag{31.19}$$

vastgehouden worden in $\mathbb{C}$. We kijken dus naar mogelijke limieten van stuksgewijs lineaire paden van a naar b.

Het ligt voor de hand meteen een deelrij te nemen waarlangs L_n convergent is, zeg $L_{n_k} \to L \geq |b-a|$ met n_k een stijgende rij in $\mathbb{N}$. Vanaf zekere zulke n is er dan steeds een eerste $j = j_n$ waarvoor geldt dat de totale lengte langs P_n van a tot $z_{j_n}^{(n)}$ minstens $\frac{L}{2}$ is, en langs een verdere deelrij convergeren dan zowel $z_{j_n}^{(n)}$ als $z_{j_n-1}^{(n)}$ naar een limiet $z_{\frac{1}{2}}$.

Maar dit argument werkt niet alleen voor $\frac{1}{2}$. Voor elke $t \in (0,1)$ kunnen we vanaf zekere n een eerste $j = j_n^t$ vinden waarvoor de totale lengte langs P_n van a tot $z_{j_n^t}^{(n)}$ minstens tL is. Doen we dit voor

$$t = \frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8}, \ldots,$$

dan geeft een diagonaalrijargument dat, voor elke rationale $t \in (0,1)$ met een noemer die een pure macht van 2 is, dat langs de geconstrueerde deelrij geldt dat

$$z_{j_n^t}^{(n)}$$

convergeert naar een limiet z_t voor al zulke t. Dit definieert een afbeelding

$$t \to z(t) = z_t,$$

waarvoor per constructie geldt[23] dat

$$|z(t_1) - z(t_2)| \leq |t_1 - t_2|L, \tag{31.20}$$

en die uniek uitbreidt tot een afbeelding $z : [0,1] \to \mathbb{C}$ met dezelfde eigenschap.

Onze eerste geparametriseerde kromme die niet per se van de vorm (31.1) is. Een kromme waarvan de lengte nog niet gedefinieerd maar wel gelijk aan L is, als alles goed is[24], en waarlangs we kunnen integreren, middels benaderingen met Riemannsommen van de vorm

$$\sum_j f((z(\tau_j))(z(t_j) - z(t_{j-1}).$$

Wat we van dit alles hier willen uitwerken is nog de vraag, maar voor continu differentieerbare zulke $t \to z(t)$ is

$$\int_P f(z)\, dz = \int_0^1 f(z(t))z'(t)\, dt$$

[23] Wel even nagaan!

[24] En de lengte van het stuk tussen t_1 en t_2 gelijk aan $|t_1 - t_2|L$.

een uitspraak die we willen hebben, waarbij het linkerlid gedefinieerd is als

$$\lim_{n\to\infty} \int_{P_n} f(z)\, dz$$

en de limiet langs de deelrij wordt genomen en moet bestaan. Dat vergt nog een stelling voor bijvoorbeeld continue $z \to f(z)$.

31.5 Calculus in Banachalgebras van operatoren

Deze sectie is nog wat schetsmatig maar niettemin precies. We willen (31.12) uitwerken voor $A \in L(X)$ en schrijven met z vervangen door λ

$$f(A) = \frac{1}{2\pi i} \oint_P f(\lambda)(\lambda - A)^{-1}\, d\lambda, \tag{31.21}$$

nu voor een willekeurig polygon[25] met hoekpunten $\lambda_1, \ldots, \lambda_n = \lambda_0$, waarop en waarbuiten[26]

$$\lambda \to (\lambda - A)^{-1} = (\lambda I - A)^{-1} \tag{31.22}$$

gedefinieerd is. Het complement van het domein van (31.22) in $\mathbb{C}$ heet het spectrum van A, notatie $\sigma(A)$. Het domein zelf heet de resolvente verzameling, notatie $\rho(A)$, en de afbeelding in (31.22) heet de resolvente van A.

Exercise 31.14. Gebruik berekeningen met meetkundige reeksen om te laten zien dat iedere $\lambda \in \mathbb{C}$ met $\lambda > |A|$ in $\rho(A)$ ligt en dat $\rho(A)$ open is. Bewijs ook dat (31.22) complex differentieerbaar is op $\rho(A)$. Wat is de afgeleide?

Exercise 31.15. Kan het zijn dat $\rho(A) = \mathbb{C}$? Het antwoord is nee, maar dat vergt nog een argument dat we weer zo licht mogelijk willen houden. Uit het ongerijmde, we zouden dan hebben dat (31.22) een $L(X)$-waardige functie definieert die naar $0 \in L(X)$ gaat als $\lambda \to \infty$ en dat moet niet kunnen, met een argument dat over te schrijven zou moeten zijn van wat we voor gewone complexwaardige functies weten, zie Opgave 31.13.

In deze opgaven heb je niet gebruikt dat met $AB = BA = I$ en $A \in L(X)$ ook volgt dat $B \in L(X)$, een wat diepere stelling voor Banachruimten, die ook maar eens heel kort en clean moet worden uitgelegd. Dat komt nog wel een keer. Denk in het vervolg voorlopig bijvoorbeeld eerst aan $X = \mathbb{C}^2$ als complexe uitbreiding van $\mathbb{R}^2$ en A een lineaire afbeelding gegeven door een 2×2 matrix, met complexe of reële entries. In dat geval bestaat $\sigma(A)$ meestal uit 2 punten, en met twee disjuncte driehoekjes Δ_1 en Δ_2 om die punten heen kunnen we al aan de slag met ieder paar complex differentieerbare functies

$$f_1 : \Delta_1 \to \mathbb{C} \quad \text{en} \quad f_1 : \Delta_1 \to \mathbb{C}$$

[25] Of een vereniging daarvan.

[26] Wat bedoelen we daarmee?

die samen één functie

$$f : \Delta_1 \cup \Delta_2 \to \mathbb{C}$$

maken waarvan de twee stukken elkaar niet zien. Maar ook het rechterlid van (14.8) gezien als afbeelding van een gecomplexificeerde $X = C([0,1])$ naar zichzelf is een voorbeeld.

In het algemeen kan $\sigma(A)$ van alles zijn en daarom kijken we nu eerst wat voor gebieden we met eindig veel disjuncte polygonen kunnen maken. Elk polygon P heeft op natuurlijke manier een binnengebied C en een buitengebied U, waar we steeds de rand bijnemen, dus

$$P = U \cap C.$$

Als binnen een polygon P_0 een aantal kleinere polygonen $P_1, \ldots, P_n$ ligt, wier binnengebieden onderling disjunct zijn, dus

$$C_i \cap C_j = \emptyset \quad \text{als} \quad i \neq j \quad \text{voor} \quad i, j = 1, \ldots, n,$$

dan kan het zijn dat

$$\sigma(A) \subset K_{int} \subset K = C_0 \cap U_1 \cap \cdots \cap U_n, \tag{31.23}$$

waarbij we K zien als begrensd door de buitenkant P_0 naarbuiten en door binnenkanten $P_1, \ldots, P_n$ naar binnen, en

$$K_{int} = K \cap P_0^c \cap \cdots \cap P_n^c$$

de doorsnijding van K met de complementen van de polygonen $P_0, \ldots, P_n$ is, dus alles in K dat niet op de rand ligt. Als we polygonen *altijd* als linksom doorlopen zien dan schrijven we in dit geval

$$f(A) = \frac{1}{2\pi i} \oint_{\delta K} f(\lambda)(\lambda - A)^{-1}\, d\lambda =$$

$$\frac{1}{2\pi i} \left(\oint_{P_0} f(\lambda)(\lambda - A)^{-1}\, d\lambda - \sum_{j=1}^{n} \oint_{P_j} f(\lambda)(\lambda - A)^{-1}\, d\lambda \right) \tag{31.24}$$

voor $f : K \to \mathbb{C}$ complex differentieerbaar.

Ligt $\sigma(A)$ in een disjuncte eindige vereniging

$$K_1 \cup \cdots \cup K_m$$

van zulke K_j, en zijn

$$f_j : K_j \to \mathbb{C} \quad (j = 1, \ldots, m)$$

complex differentieerbaar, dan vormen die samen weer een complex differentieerbare functie

$$f : K = K_1 \cup \cdots \cup K_n \to \mathbb{C}$$

waarvoor we

$$f(A) = \frac{1}{2\pi i} \sum_{j=1}^{m} \oint_{\delta K_j} f(\lambda)(\lambda - A)^{-1}\, d\lambda \tag{31.25}$$

met iedere term in de som gedefinieerd als in (31.24) als definitie van $f(A)$ gebruiken.

Elk van de K_j kan van de vorm alleen maar $K_j = C_j$ zijn, en één $K = C$ is altijd mogelijk om dat $\sigma(A)$ begrensd is, maar hoe kleiner K gekozen wordt, hoe meer speelruimte er is. De mogelijk steeds grotere[27] uitdrukkingen voor K moet daarbij graag op de koop toe worden genomen, en als $K_j \neq C_j$ kunnen de bijbehorende buitenkanten ook genest liggen. In het simpele geval dat $\sigma(A)$ een eindige discrete puntverzameling is kunnen we natuurlijk toe met $K = C_j = \Delta_j$, met de driehoekjes Δ_j zo klein als we maar willen en

$$\sigma(A) \subset K^{int} \subset K = \Delta_1 \cup \cdots \cup \Delta_m.$$

Wat we nu sowieso in alle gevallen willen is dat, als we de hoekpunten van de polygonen een beetje naar binnen schuiven, K in dus, de integralen die in (31.24) en (31.25) de nieuwe lineaire afbeelding $f(A) : X \to X$ moeten maken, niet veranderen. En hetzelfde als we K groter maken door de punten naar buiten te schuiven, zolang we maar niet uit het definitiegebied van de continu differentieerbare complexwaardige f lopen. Bij het verder kleiner of groter maken kan de structuur van K versimpelen als twee polygonen elkaar ontmoeten en vervolgens samen één polygon vormen. Strict genomen hebben we niet nodig hoe dat precies kan gaan, maar het is toch aardig om daar even over na te denken.

Exercise 31.16. Het is een aardige project om dat versimpelen precies te maken. Bij het groter maken van K kunnen twee buitenkanten van twee K_j-tjes elkaar ontmoeten waarna verder groter maken tot één nieuwe buitenkant leidt waarmee de bijbehorende binnenkanten dan samen de nieuwe binnenkanten van een nieuwe K_j worden. Ook kan uit een groeiende buitenkant die binnen een krimpende binnenkant ligt meteen na het eerste contact één nieuwe binnenkant ontstaan. Bij kleiner maken kunnen een binnen- en een buitenkant van eenzelfde K_j-tje elkaar ontmoeten en daarna een nieuwe buitenkant vormen, en ook kunnen twee binnenkanten elkaar ontmoeten en een nieuwe binnenkant vormen. Ga in alle gevallen na wat de nieuwe structuur wordt en welke

[27] Als we alle zijden van alle poygonen dicht bij $\sigma(A)$ willen hebben.

andere scenarios er nog zijn, zoals ondermeer polygonen tot een punt laten krimpen en verdwijnen.

Via de inmiddels vertrouwde zigzagkrommen vernandert bij het geschuif met de hoekpunten (31.25) niet, mits de Stelling van Goursat geldt voor driehoekjes waarop en waarbinnen (31.22) complex differentieerbaar is. De betreffende integralen bestaan weer uit integralen over lijnstukjes. Continue $L(X)$-waardige functies van $t \in [0,1]$ zijn integreerbaar via de tussensommen van Riemann, en

$$t \to f((1-t)\lambda_{k-1} + t\lambda_k)((1-t)\lambda_{k-1} + t\lambda_k - A)^{-1}$$

is zo'n functie waarmee $L(X)$-waardige integralen als

$$\int_{\lambda_{k-1}}^{\lambda_k} f(\lambda)(\lambda - A)^{-1}\, d\lambda$$

nu gedefinieerd zijn.

Mooi, dan kan voor

$$\lambda \to f(\lambda)(\lambda - A)^{-1}$$

de Stelling van Goursat met bewijs en al worden overgeschreven[28] en is (31.24) een goede definitie van $f(A)$. Voorlopig houden we nu A vast en kijken naar nog zo'n f, een g dus, waarbij we eerst aannemen dat we het allersimpelste geval hebben, één polygon rond $\sigma(A)$ waarmee de berekeningen gedaan worden. In dat geval is de samenstelling van de afbeeldingen $f(A)$ en $g(A)$ te schrijven als

$$f(A)g(A) = \frac{1}{2\pi i} \oint_{\lambda_{0-n}} f(\lambda)(\lambda - A)^{-1}\, d\lambda\, \frac{1}{2\pi i} \oint_{\mu_{0-n}} g(\mu)(\mu - A)^{-1}\, d\mu,$$

met in de Cauchyintegraal voor $g(A)$ de hoekpunten μ_l een klein beetje naar binnen geschoven hebben, niet omdat het moet, maar omdat het kan, iets minder ver naar binnen dan de hoekpunten λ_k. Het μ-polygon komt zo binnen het λ-polygon te liggen.

Omdat

$$f(A) = \frac{1}{2\pi i} \sum_{k=1}^{n} \int_{\lambda_{k-1}}^{\lambda_k} f(\lambda)(\lambda - A)^{-1}\, d\lambda,$$

$$g(A) = \frac{1}{2\pi i} \sum_{l=1}^{n} \int_{\mu_{l-1}}^{\mu_l} g(\mu)(\mu - A)^{-1}\, d\mu,$$

[28] Op detail nog te bespreken.

wordt $f(A)g(A)$ afgezien van de voorfactoren dankzij overwegingen als bij (21.3) een som van produkten

$$\int_{\lambda_{k-1}}^{\lambda_k} f(\lambda)(\lambda - A)^{-1}\, d\lambda \int_{\mu_{l-1}}^{\mu_l} g(\mu)(\mu - A)^{-1}\, d\mu =$$

$$\int_{\mu_{l-1}}^{\mu_l} \int_{\lambda_{k-1}}^{\lambda_k} f(\lambda)g(\mu)(\lambda - A)^{-1}(\mu - A)^{-1}\, d\lambda\, d\mu =$$

$$\int_{\lambda_{k-1}}^{\lambda_k} \int_{\mu_{l-1}}^{\mu_l} f(\lambda)g(\mu)(\lambda - A)^{-1}(\mu - A)^{-1}\, d\mu\, d\lambda.$$

Dankzij wat fraaie algebra, te weten

$$(\lambda - A)^{-1}(\mu - A)^{-1} = \frac{1}{\mu - \lambda}(\lambda - A)^{-1} + \frac{1}{\lambda - \mu}(\mu - A)^{-1},$$

kunnen de integralen gesplitst worden in

$$\int_{\lambda_{k-1}}^{\lambda_k} f(\lambda)(\lambda - A)^{-1} \int_{\mu_{l-1}}^{\mu_l} \frac{g(\mu)}{\mu - \lambda}\, d\mu\, d\lambda$$

en

$$\int_{\mu_{l-1}}^{\mu_l} g(\mu)(\mu - A)^{-1} \int_{\lambda_{k-1}}^{\lambda_k} \frac{f(\lambda)}{\lambda - \mu}\, d\lambda\, d\mu,$$

en in beide herhaalde integralen zien we een bij sommeren over de index in de binnenste integraal een gewone complexwaardige lijnintegraal verschijnen waar nul uit komt als de noemer niet nul is in het binnengebied, en een functiewaarde anders, kijk maar naar de Cauchy integraalformule. Sommeren over l in de eerste geeft derhalve 0, en sommeren over k in de tweede

$$\int_{\mu_{l-1}}^{\mu_l} g(\mu)(\mu - A)^{-1}\, 2\pi i\, f(\mu)\, d\mu = 2\pi i \int_{\mu_{l-1}}^{\mu_l} f(\mu)\, g(\mu)(\mu - A)^{-1}\, d\mu,$$

en nog een keer sommeren vervolgens $(2\pi i)^2(fg)(A)$. We concluderen dat

$$(fg)(A) = \frac{1}{2\pi i} \oint_{\lambda_{0-n}} f(\lambda)g(\lambda)(\lambda - A)^{-1}\, d\lambda = f(A)g(A) = g(A)f(A), \tag{31.26}$$

en daar is nog veel mee te spelen.

Exercise 31.17. Ga na dat in het algemene geval (31.25), wanneer $f(A)$ en $g(A)$ de som zijn van een eindig aantal integralen over links- dan wel rechtsom[29] doorlopen polygonen P_j, er in de compositie alleen bijdragen zijn van de vorm zoals juist behandeld en dat ook in dat geval volgt dat $(fg)(A) = f(A)g(A)$.

De tweede gelijkheid in (31.26) is een gelijkheid in de niet-commutatieve Banachalgebra $L(X)$ van continue lineaire afbeeldingen van X naar zichzelf, en $f \to f(A)$ is een afbeelding die gedefinieerd is voor een klasse van functies gedefinieerd op een omgeving van het $\sigma(A)$. Die omgeving mag van f afhangen, dus met f en g moeten we ons beperken tot de doorsnede van de twee definitiegebieden. Wat we nog willen laten zien is dat een schrijfwijze met (31.24) en (31.25) altijd mogelijk is met alle polygonen zo dicht bij $\sigma(A)$ als we maar willen. Daarmee bewijzen we dan ook meteen de volgende stelling.

Theorem 31.18. *Laat voor $A \in L(X)$ en een complexwaardige f de operator $f(A)$ gedefinieerd zijn via (31.24) en (31.25). Dan geldt*

$$\sigma(f(A)) = f(\sigma(A)).$$

Om deze stelling te bewijzen maken we nu precies hoe we K kiezen. Kies daartoe een triangulatie van het complexe vlak opgespannen door $\rho > 0$ en $\rho \exp(\frac{\pi i}{6})$. De verzameling van al deze driehoekjes noemen we I. Voor elke $\Delta \in I$ maken we onderscheid tussen

$$\Delta \cap \sigma(A) = \emptyset, \quad \Delta \cap \sigma(A) \neq \emptyset = \delta\Delta \cap \sigma(A), \quad \delta\Delta \cap \sigma(A) \neq \emptyset,$$

waarmee $I = I_0 \cup I_1 \cup J$, met I_0, I_1, J de onderling disjuncte deelverzamelingen waarvoor respectievelijk de eerste, tweede dan wel derde karakterisatie geldt. Zowel I_1 als J hebben maar eindig veel elementen omdat $\sigma(A)$ begrensd is. Iedere $\Delta \in I_1$ kan als een K_j genomen worden in (31.25).

De driehoekjes in I_0 zijn niet relevant voor (31.25), maar iedere $\Delta \in J$ heeft 12 buren[30] waarvan er tenminste één ook in J ligt, zeg $\tilde{\Delta}$, gekarakteriseerd door

$$\delta\tilde{\Delta} \cap \delta\Delta \cap \sigma(A) \neq \emptyset,$$

en in dat geval noemen we Δ en $\tilde{\Delta}$ fijne buren in J. Twee zulke fijne buren die verder geen andere fijne buren hebben vormen samen een fijn duo verenigd in

$$\Delta \cup \tilde{\Delta},$$

[29] Lees: linksom, maar met een min voor het integraalteken.

[30] Waarvan er drie een zijde met Δ gemeen hebben en de rest alleen een hoekpunt.

en I_2 is per definitie de verzameling van zulke verder geisoleerde fijne buren, die verenigd steeds een ruit vormen, een ruit die als een K_j kan worden meegenomen in (31.25).

Een paar niet geisoleerde fijne buren kan nog 1 of meerdere fijne buren hebben, en als het maar 1 is, zeg $\hat{\Delta}$, dan kan het zijn dat die verder zelf geen fijne buren meer heeft. Dan vormen ze een fijn triootje waarbij verschillende standjes denkbaar zijn. Dit definieert de verzameling I_3, alle driehoeken Δ die onderdeel vormen van een fijn trio verenigd in

$$\Delta \cup \tilde{\Delta} \cup \hat{\Delta},$$

dat een parallelogram of een halve zeshoek is.

En zo gaat dat door met fijne quatrootjes, fijne quintootjes, etc totdat J op is, waarbij het aantal standjes flink maar niet oneindig toe kan nemen. Kortom, met I gepartioneerd als

$$I = I_0 \cup I_1 \cup I_2 \cup \cdots \cup I_p$$

is het nu nog de vraag wat de mogelijke onderlinge standjes zijn: als $\Delta_1 \in I_k$ met $k-1$ andere driehoeken in I_k een fijn k-stel vormt hoe kan de vereniging

$$\Delta_1 \cup \Delta_2 \cup \cdots \cup \Delta_k$$

er dan uitzien?

Antwoord: als een binnengebied van een polygon, of als het rechterlid van (31.23). Dat moet dus nog door iemand[31] bewezen worden, als dat niet al eens gebeurd is. Maar verder zijn we nu wel klaar met de beschrijving van $f(A)$. Dat kan altijd met eindig veel polygonen die willekeurig dicht bij $\sigma(A)$ liggen door ρ klein te kiezen. Hoe dichter bij $\sigma(A)$ hoe meer je er nodig hebt en hoe wilder de standjes kunnen worden.

We zijn nu klaar voor het bewijs van Stelling 31.18. Neem een $\mu \notin f(\sigma(A))$ en definieer g door

$$\lambda \xrightarrow{g} \frac{1}{\mu - f(\lambda)},$$

met f complex differentieerbaar op een omgeving van $\sigma(A)$. Kies een mogelijk kleinere omgeving waarop $f(\lambda) \neq \mu$. Uit de functional calculus volgt nu dat $g(A)$ gedefinieerd is en de algebra geeft

$$g(A)(\mu - f(A) = (\mu - f(A)g(A) = I,$$

[31] Ik pas, maar dat is voor even.

waarmee $\mu \in \rho(f(A))$. Dus $\sigma(f(A)) \subset f(\sigma(A))$.

Kan de inclusie strict zijn? In dat geval is er een $\mu_0 = f(\lambda_0) \in \sigma(f(A))$ waarvoor $\mu_0 - f(A)$ inverteerbaar is terwijl $\lambda_0 - A$ het niet is. Door schuiven en schalen van f, en schuiven van A en λ kunnen we zonder beperking der algemeenheid wel aannemen dat $\lambda_0 = 0 = \mu_0$ en dat de machtreeks van f begint met λ^n voor zekere $n \in \mathbb{N}$ omdat $f(0) = 0$. In dat geval is

$$f(\lambda) = \lambda^n g(\lambda) \quad \text{met} \quad g(\lambda) = 1 + b_1\lambda + b_2\lambda^2 + \cdots$$

en dus is $g(A)$ inverteerbaar, net als $f(A)$. Maar de algebra geeft

$$f(A) = A^n g(A).$$

Voor $n = 1$ is de tegenspraak onmiddellijk. Voor $n > 1$ niet helemaal. Pas daarom het argument hierboven aan en concludeer eerst dat $\mu_0 = f(\lambda_0)$ zo gekozen kan worden dat $f'(\lambda_0) \neq 0$. Hiermee is het bewijs van de stelling wel klaar. Als g een andere functie is die complex differentieerbaar is op een omgeving van $\sigma(f(A)) = f(\sigma(A))$ dan volgt ook vrij direct uit de definities dat

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

Exercise 31.19. Bewijs dit.

Nog een expliciet voorbeeld. Als

$$\lambda = \lambda \sum_{j=1}^{N} \chi_j(\lambda) = \sum_{j=1}^{N} \lambda\chi_j(\lambda),$$

met

$$\chi_j(\lambda) = \delta_{ij} \quad \text{voor} \quad \lambda \in K_i,$$

dan

$$I = \sum_{j=1}^{N} I\chi_j.$$

Definieer de "spectraalprojecties"

$$P_j = \chi_j(A).$$

Exercise 31.20. Laat zien dat $P_iP_j = \delta_{ij}P_j$, $AP_j = P_jA$, $\sigma(AP_j) = \sigma(A) \cap K_j$,

$$I = \sum_{j=1}^{N} P_j \quad \text{en} \quad A = \sum_{j=1}^{N} AP_j = \sum_{j=1}^{N} P_jA.$$

Zo wordt

$$X = R(P_1) \oplus \cdots \oplus R(P_n),$$

en beeld A iedere $X_i = R(P_i)$) op zich zelf af, en volgt voor

$$A_j : X_i \xrightarrow{AP_j} X_i$$

dat $\sigma(A_j) = \sigma(A) \cap K_j$.

Zo, en dat alles met een beetje lijnintegreren.

32 Standing at the crossroads of PDE and FA

This chapter relates to the courses in Functional Analysis and Partial Differential Eqations as given in the bachelor pogrammes in Amsterdam, as well as courses run under the same name in the national mastermath programme. Have a look at Section 14.3, in particular at the solution method for (14.5), which clearly does not generalise to the problem of solving

$$-\Delta u = f \quad \text{in} \quad \Omega \quad \text{with} \quad u = 0 \quad \text{on} \quad \partial\Omega \tag{32.1}$$

for given $f : \Omega \to \mathbb{R}$ and $\Omega \subset \mathbb{R}^{\mathrm{N}}$ on a bounded open set with boundary $\partial\Omega$.

Moreover, even is $\partial\Omega$ is smooth, it does not in general hold that (32.1) has a twice differentiable solution which solves the partial differential equation $-\Delta u = f$ for the given $f \in C(\bar{\Omega})$. Below we show another way of solving (14.5) which does generalise to a large class of problems including (32.1). This technique is based on integration by parts[1] and the theory of Hilbert spaces, mainly the unique and obvious generalisation[2] of the inner product space $\mathbb{R}^2$ with the dimension 2 replaced by the first infinite cardinal.

Als we de vergelijking in (14.5) vermenigvuldigen met een functie $v \in C^1[0,1]$ dan bestaan onder de aanname dat $u \in C^2[0,1]$ en $f \in C^0[0,1]$ beide integralen

$$-\int_0^1 u''(x)v(x)\,dx = \int_0^1 f(x)v(x)\,dx$$

en kan de linkerkant partieel geintegreerd worden. Het resultaat is

$$-[u'(x)v(x)]_0^1 + \underbrace{\int_0^1 u'(x)v'(x)\,dx}_{\text{symmetrisch in } u,v} = \underbrace{\int_0^1 f(x)v(x)\,dx}_{\text{symmetrisch in } f,v}, \tag{32.2}$$

[1] Now have a look at (21.15).

[2] In different guises.

waarbij de lelijke eerste term verdwijnt als we de extra aanname maken dat $v(0) = v(1) = 0$.

De oplossing $u \in C^2[0,1]$ van (14.5) heeft dus de eigenschap dat $u(0) = u(1) = 0$ en

$$\int_0^1 u'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx \quad \forall v \in C_0^1[0,1], \tag{32.3}$$

waarin

$$C_0^1[0,1] = \{v \in C^1[0,1] : v(0) = v(1) = 0\},$$

en de gelijkheid in (32.3) kan voor elke $u \in C_0^1[0,1]$ geverifieerd worden. Kortom, we zouden dus kunnen afspreken om $u \in C_0^1[0,1]$ een oplossing van (14.5) te noemen als aan (32.3) voldaan is.

Rechts in (32.3) zien we een integraal die te zien is als een inwendig product van f en v, dat we kunnen noteren als

$$f \cdot v = \int_0^1 f(x)v(x)\,dx, \tag{32.4}$$

waarmee (14.5) zich uiteindelijk herschrijft als

$$u \in C_0^1[0,1], \quad \underbrace{u' \cdot v'}_{((u,v))} = \underbrace{f \cdot v}_{(f,v)} \quad \forall v \in C_0^1[0,1], \tag{32.5}$$

een uitdrukking waarin twee inwendige producten voorkomen en $u \in C_0^1[0,1]$ en $f \in C^0[0,1]$ vast zijn, en $v \in C_0^1[0,1]$ willekeurig.

Exercise 32.1. Waarom is $(u,v) \to u' \cdot v'$ wel een inprodukt op $C_0^1[0,1]$ en niet op $C^1[0,1]$?

Exercise 32.2. Laat $L > 0$, bijvoorbeeld $L = 2\pi$ of $L = 1$. De vectorruimte van continue L-periodieke functies noemen we $C(\mathbb{R}_L) = C^0(\mathbb{R}_L)$, en de deelruimte van k keer ($k \in \mathbb{N}$) continu differentieerbare functies noemen we $C^k(\mathbb{R}_L)$. Voor welke $f \in C(\mathbb{R}_L)$ is de vergelijking $-u'' = f$ oplosbaar met u in $C^2(\mathbb{R}_L)$? Is de oplossing uniek? Hint: neem eerst $L = 1$ natuurlijk.

Exercise 32.3. Neem $L = 1$ en los de vergelijking $-u'' = f$ voor f continu, 1-periodiek met $\int_0^1 f(x)\,dx = 0$: geef een uitdrukking van de vorm (14.8) voor de

oplossing u die (ook) voldoet aan $\int_0^1 u(x)\,dx = 0$. Hint: zonder deze laatste conditie is de oplossing niet uniek bepaald en evenzo is de (wederom symmetrische!) kern $A(x,s)$ niet uniek bepaald. Maar wel onder de conditie dat $\int_0^1 A(x,s)\,dx = \int_0^1 A(x,s)\,ds = 0$.

Exercise 32.4. Laat

$$\bar{C}^k(\mathbb{R}_L) = \{u \in C^k(\mathbb{R}_L) : \int_0^L u(x)dx = 0\}$$

en geef een herformulering van $-u'' = f$ voor $f \in \bar{C}^0(\mathbb{R}_L)$ en $u \in \bar{C}^1(\mathbb{R}_L)$ zoals in (32.5).

Hopelijk is duidelijk dat de twee laatste opgaven kwa bewerkelijkheid nogal uiteenliepen. Formuleringen als in (32.5) gebaseerd op *integration by parts*, zonder dat het doel daarvan het uitrekenen van getallen is, bieden een ander en vaak algemener perspectief om eigenschappen van oplossingsoperatoren te begrijpen dan expliciete oplossingsmethoden gebaseerd op primitiveren. In wat volgt zullen we daartoe v als een variabele zien en

$$v \to ((u,v)) \quad \text{en} \quad f \to (f,v)$$

als dezelfde lineaire functie, maar anders gepresenteerd.

Lineaire functies en inprodukten, hoe zit dat? Hoe weet je dat er bij f via dezelfde lineaire functie van v een u hoort? En kan dat algemener? Voor later[3]. Dit hoofdstuk besluiten we met een paar opgaven die laten zien hoe intrinsiek de herformulering als (32.5) verbonden is met de eigenschappen van de oplossingsoperator

$$A : f \xrightarrow{\forall v((u,v))=(f,v)} u, \tag{32.6}$$

waarbij het van het specifieke probleem afhangt welke inproducten en welke functieruimten gedefinieerd moeten worden om een A te maken die in simpele gevallen samenvalt met expliciet uitgerekende integraaloperatoren als in (14.8).

Exercise 32.5. Laat zien dat een solver als A in (32.6) de eigenschap heeft dat $((Au,v)) = ((u,Av))$ en $(Af,g) = (f,Ag)$ voor alle u,v en f,g in de nog te kiezen ruimten $V \subset H$ en H waarop de inproducten zijn gedefinieerd met de eigenschappen die we nodig hebben om alles precies te maken.

[3]See Chapter 34.

De operator A is dus symmetrisch[4] met betrekking tot twee inproducten, waaronder het 'gewone' inproduct (32.4) dat in eerste instantie was opgeschreven onder verschillende aannames voor f en v.

Exercise 32.6. Gebruik de symmetrie van A om te laten zien dat eigenvectoren[5] van A bij verschillende eigenwaarden van A loodrecht op elkaar staan.

Exercise 32.7. Gebruik de vorige opgave en Opgave 32.4 om zonder rekenwerk te laten zien dat

$$\int_{-\pi}^{\pi} \sin nx \sin mx\, dx = 0 = \int_{-\pi}^{\pi} \cos nx \cos mx\, dx \quad (m, n \in \mathbb{N},\ m \neq 0)$$

$$\int_{-\pi}^{\pi} \sin nx \cos mx\, dx = 0 \quad (m, n \in \mathbb{N})$$

Natuurlijk wist je dit al, waarschijnlijk via $\exp(inx) = \cos nx + i \sin nx$ en de gebruikelijke rekenregeltjes gebaseerd op de somformules[6] voor $\cos(a + b)$ en $\sin(a + b)$ die niet meer tot de tegenwoordig zelden precies gerechtvaardigde basiskennis van de gemiddelde β-student horen. De functie sin kan gedefinieerd worden als de unieke oplossing van het beginwaardeprobleem

$$u'' + u = 0; \quad u(0) = 0; \quad u'(0) = 1, \tag{32.7}$$

en cos als de afgeleide van sin. Alle eigenschappen van cos en sin, i.h.b. de somformules volgen uit deze definities en kunnen gebruikt worden voor de volgende opgave.

Exercise 32.8. Bepaal alle $\lambda > 0$ waarvoor $u'' + \lambda u = 0$ oplossingen van periode 2π heeft en bepaal alle even en oneven oplossingen voor die waarden λ.

De even oplossingen die je zo vindt zijn veelvouden van $c_1 : x \to \cos x$, $c_2 : x \to \cos 2x$, $c_3 : x \to \cos 3x$, ..., en de oneven oplossingen zijn veelvouden van $s_1 : x \to \sin x$, $s_2 : x \to \sin 2x$, $s_3 : x \to \sin 3x$, ..., en ieder tweetal van deze functies staat loodrecht op elkaar, zoals we in Opgave 32.7 gezien

[4] We praten nog niet over complexwaardige functies hier.

[5] $A\phi = \lambda\phi$, $\lambda \in \mathbb{R}$, ϕ een (eigen)functie.

[6] Zie Wiskunde in je Vingers, sectie 10.4.

hebben. En ze zijn gemiddeld allemaal nul, hetgeen betekent dat ze loodrecht staan op de functie $\mathbf{1} : x \to 1$, bijvoorbeeld

$$(\mathbf{1}, s_1) = \mathbf{1} \cdot s_1 = \int_{-\pi}^{\pi} 1 \sin x \, dx = \int_{-\pi}^{\pi} \sin x \, dx = 0.$$

Wat we in vervolg gaan doen is $1, c_1, c_2, c_3, \ldots, s_1, s_2, s_3, \ldots$ zien als vectoren die lijnen door de oorsprong[7] definiëren. En die lijnen zien we als een assenkruis waarmee we een oneindig-dimensionale ruimte opspannen, een ruimte waarin we willen werken zoveel mogelijk als we dat in het platte vlak doen.

[7]Die oorsprong is de nulfunctie $\mathbf{0} : x \to 0$.

33 Lebesgue spaces

If you are already familiar with Lebesgue spaces you may like to jump to Section 33.3 and flip back when needed. The following definitions usually come at the end of a course on measure theory and Lebesgue integration.

Definition 33.1. *Let $U \subset \mathbb{R}^N$ be open and $p \geq 1$. A measurable function $u : U \to \mathbb{R}$ is said to be in $L^p_{loc}(U)$ if*

$$\int_B |f|^p < \infty$$

for every open ball $B \subset U$, and in $L^p(U)$ if the p-norm of u defined by

$$|f|_p^p = \int_U |f|^p < \infty \tag{33.1}$$

exists.

Modulo hassle needed to deal with $|f|_p = 0$ not implying that $f(x) = 0$ for all $x \in U$, but only that

$$\{x \in \mathbb{R}^N : \, f(x) \neq 0\}$$

is a set of zero measure[1], the normed space $L^p(U)$ is Banach space with its norm defined by (33.1).

Remark 33.2. *Every $f \in L^p(U)$ extends to $f \in L^p(\mathbb{R}^N)$ by setting $f(x) = 0$ for $x \notin U$. No such general[2] statement holds for $f \in L^p_{loc}(U)$ and $L^p_{loc}(\mathbb{R}^N)$.*

We recall from the discussion in Section 20.3 about (20.23,20.24) that

$$|\sum_{i=1}^n a_i b_i| \leq |a|_p \, |b|_q \quad \text{for} \quad p, q > 1 \quad \text{with} \quad \frac{1}{p} + \frac{1}{q} = 1, \tag{33.2}$$

Hölder's inequality for finite sums of real numbers. Memorise that

$$\frac{1}{p} + \frac{1}{q} = 1 \iff (p-1)(q-1) = 1 \iff q = \frac{p}{p-1} \iff p = \frac{q}{q-1},$$

and convince yourself that via any definition of the integral it also holds that

$$|\int_U fg| \leq |f|_p \, |g|_q \tag{33.3}$$

with the norms defined by (33.1).

[1] Here $|A|$ denotes the Lebesgue measure of a Lebesgue measurable subset $A \subset \mathbb{R}^N$.

[2] Example: $p = N = 1$, $f(x) = \frac{1}{x}$, $U = \mathbb{R}_+$.

Exercise 33.3. Explain why the spaces $L^p_{loc}(U)$ are nested: $L^p_{loc}(U) \subset L^q_{loc}(U) \subset L^1_{loc}(U)$ if $p \geq q \geq 1$. Hint: use (33.3) with $g \equiv 1$ to show that the spaces $L^p(U)$ are nested if U is bounded.

Exercise 33.4. No estimate of the type

$$|u|_p \leq C_{pqN}|u|_q$$

with $p > q \geq 1$ can hold for all $u \in C_c(\mathbb{R}^N)$. Why? Show that the spaces $L^p(\mathbb{R}^N)$ are not nested.

Exercise 33.5. Apply (33.3) to f^a and f^b to show that

$$|f|^{a+b}_{a+b} \leq |f|^a_{ap}|f|^b_{bq}$$

and solve the equations $1 \leq ap = r < a + b = s < bq = t$ and $(p-1)(q-1) = 1$ to obtain an (interpolation) inequality for $|f|_s$ in terms of $|f|_r$ and $|f|_t$. This shows that

$$L^r(U) \cap L^t(U) \subset L^s(U) \quad \text{for} \quad r < s < t.$$

Discuss the limit case $t = \infty$.

33.1 The Lebesgue's Differentiation Theorem

Since Lebesgue we see every $f \in L^1(\mathbb{R}^N)$ as an equivalence class[3] F of integrable measurable functions $f : \mathbb{R}^N \to \mathbb{R}$ for the equivalence relation

$$f \sim g \iff |\{x \in \mathbb{R}^N : f(x) \neq g(x)\}| = 0 \iff \int_{\mathbb{R}^N} |f - g| = 0.$$

In this chapter we shall also explore and perhaps prefer the alternative approach, similar tot the construction[4] of $\mathbb{R}$ out of $\mathbb{Q}$, namely as equivalence classes of Cauchy sequences f_n in $C_c(\mathbb{R}^N)$ with respect to the p-norm. This will completely avoid the notion of Lebesgue measure and all integrals will be limits of integrals of continuous functions.

[3] Often denoted as $[f]$, so $f \in [f] = F$.

[4] Which for $\mathbb{R}$ may obscure what was already obvious.

Remark 33.6. *We have that $|A| = 0$ if and only if for every $\varepsilon > 0$ there exist a sequence of open balls $B_n = B(x_n, r_n)$ indexed by $n \in \mathbb{N}$ such that*

$$A \subset \cup_{n\in\mathbb{N}} B_n \quad \textit{and} \quad \sum_{n\in\mathbb{N}} |B_n| < \varepsilon,$$

in which

$$|B_n| = \omega_N r_n^N.$$

Note that this zero measure concept does not even involve the Lebesgue measure of the covering countable union of (open) balls, but it does contain the fundamental idea that measure theory should deal with countable unions.

Exercise 33.7. Prove that a countable union of zero measure sets is again a zero measure set. Hint: every small $\varepsilon > 0$ is the sum of countably many smaller positive epsilons; this goes back to Zeno and Section 1.4.

Remark 33.8. *No matter how we define $L^1(\mathbb{R}^N)$, a fundamental truth is that for every $f \in L^1(\mathbb{R}^N)$ and every open ball $B(x, r)$ the integral*

$$\int_{B(x,r)} f$$

is independent of the choice of $f \in F = [f]$ for whatever concept of equivalence and equivalence classes used to define $L^1(\mathbb{R}^N)$, and varies[5] continuously with $x \in \mathbb{R}^N$ and $r \geq 0$. Moreover

$$|\int_{B(x,r)} f| \leq \int_{B(x,r)} |f| \to |f|_1 \quad as \quad r \to \infty,$$

$$B(x, r) \subset B(y, s) \implies \int_{B(x,r)} |f| \leq \int_{B(y,s)} |f|,$$

and[6]

$$\int_{B_1\cup\cdots\cup B_n} |f| = \int_{B_1} |f| + \cdots + \int_{B_n} |f|$$

for balls $B_1, \ldots, B_n$ with $B_i \cap B_j = \emptyset$ if $i \neq j$.

[5]This follows from the dominated convergence theorem for Lebesgue integrals in fact.

[6]The finite additivity of the integral over disjoint unions of open balls.

As a consequence also the average

$$A_f(x,r) = \fint_{B(x,r)} f = \frac{1}{|B(x,r)|}\int_{B(x,r)} f = A_{x,r}f \tag{33.4}$$

of f over $B(x,r)$ varies continuously with $x \in \mathbb{R}^N$ and $r > 0$. The function

$$(x,r) \xrightarrow{A_f} \fint_{B(x,r)} f$$

is continuous from $\mathbb{R}_+ \times \mathbb{R}^N$ to $\mathbb{R}$, and independent of the choice of $f \in F$. For fixed $x \in \mathbb{R}^N$ and $r > 0$ the map

$$f \xrightarrow{A_{x,r}} \fint_{B(x,r)} f$$

is linear and continuous from $L^1(\mathbb{R}^N)$ to $\mathbb{R}$, and the estimate

$$|A_{x,r}f| \le A_{x,r}|f|$$

holds for every $f \in L^1(\mathbb{R}^N)$.

Definition 33.9. *The good set of a function $f \in L^1_{loc}(\mathbb{R}^N)$ is defined by*

$$G_f = \{x \in \mathbb{R}^N : \lim_{r\downarrow 0} A_f(x,r) = f(x)\}. \tag{33.5}$$

Clearly the existence and value of the limit does not rely on the choice of $f \in F$. If we set

$$\mathcal{N}_f = \{x \in \mathbb{R}^N : \lim_{r\downarrow 0} A_f(x,r) \text{ does not exist}\},$$

then the complement of the set $\mathcal{N}_f$ contains the good set G_f of every $f \in F$.

Theorem 33.10. *For every $f \in L^1_{loc}(\mathbb{R}^N)$ the good set G_f has a complement with zero measure*[7]*. This complement contains the set $\mathcal{N}_f$, which is therefore not that bad: it also has zero measure, and it is natural to choose the unique $f \in F$ for which*

$$f(x) = \lim_{r\downarrow 0} A_f(x,r)$$

for all $x \notin \mathcal{N}_f$, and $f(x) = 0$ for all $x \in \mathcal{N}_f$. In that case $\mathbb{R}^N$ is the disjoint union of G_f and $\mathcal{N}_f$. For every $x \in G_f$ the value of $f(x)$ is what it should be, and for every $x \in \mathcal{N}_f$ the value of $f(x)$ is irrelevant as far as integrals are concerned, and chosen to be 0.

[7]This is the Lebesgue Differentiation Theorem, a name we do not explain yet.

Exercise 33.11. Explain why it suffices to prove Theorem 33.10 for $f \in L^1(\mathbb{R}^N)$.

Exercise 33.12. Prove the statements in Theorem 33.10 for $f \in C_c(\mathbb{R}^N)$ by showing that $G_f = \mathbb{R}^N$: the limit exists for every $x \in \mathbb{R}^N$ and is what it should be.

Remark 33.13. *Check out the literature to understand why this theorem is called the Lebesgue Differentation Theorem. We prove the theorem in Section 33.2 the way Folland does it, and provide an alternative proof starting from Section 33.3, avoiding measure theory and topology.*

33.2 The proof of the good set theorem

The proof of Theorem 33.10 invokes the Hardy-Littlewood function, defined for $f \in L^1(\mathbb{R}^N)$ by

$$H_f(x) = \sup_{r>0} A_{|f|}(x,r) = \sup_{r>0} \frac{1}{|B(x,r)|} \int_{B(x,r)} |f| \in [0,\infty],$$

the largest possible average of $|f|$ on balls centered in x. Since for every fixed $x \in \mathbb{R}^N$ we have

$$0 \le A_{|f|}(x,r) \le \frac{|f|_1}{\omega_N r^N},$$

the supremum $H_f(x)$ is finite unless $A_{|f|}(x,r) \to \infty$ as $r \to 0$. Note that

$$|A_f(x,r)| \le A_{|f|}(x,r) \le H_f(x) \le \infty. \tag{33.6}$$

We examine A_f via H_{f-g} with $g \in C_c(\mathbb{R}^N)$ and small $|f-g|_1 > 0$. Writing

$$A_f(x,r) - f(x) = \underbrace{A_f(x,r) - A_g(x,r)}_{A_{f-g}(x,r)} + \underbrace{A_g(x,r) - g(x)}_{\to 0 \text{ as } r\to 0} + g(x) - f(x),$$

we use (33.6) with $f-g$ and observe that in the resulting inequality

$$|A_f(x,r) - f(x)| \le H_{f-g}(x) + \underbrace{|A_g(x,r) - g(x)|}_{\to 0 \text{ as } r\to 0} + |g(x) - f(x)| \tag{33.7}$$

the role of r disappears as $r \to 0$. So if the left hand side is not small then the first or the third term is not small. Or both.

We therefore consider the sets[8]

$$O^\varepsilon = \{x \in \mathbb{R}^N : H_{f-g}(x) > \varepsilon\} \quad \text{and} \quad S^\varepsilon = \{x \in \mathbb{R}^N : |f(x) - g(x)| > \varepsilon\},$$

[8] O is for open, as we shall see.

and let W_ε be the set of all points $x \in \mathbb{R}^N$ for which the statement

$$\exists_{\delta>0}\, \forall_{r\in(0,\delta)} : \, |A_f(x,r) - f(x)| \le 2\varepsilon$$

fails. Then it must be that

$$W^\varepsilon \subset O^\varepsilon \cup S^\varepsilon. \tag{33.8}$$

These sets are nested,

$$0 < \delta < \varepsilon \implies O^\varepsilon \subset O^\delta, \quad S^\varepsilon \subset S^\delta \quad \text{and} \quad W^\varepsilon \subset W^\delta,$$

and via

$$\int_{\mathbb{R}^N} |f-g| \ge \int_{S^\varepsilon} |f-g| \ge \varepsilon |S^\varepsilon|$$

we have

$$\int_{\mathbb{R}^N} |f-g| > \varepsilon |S^\varepsilon|$$

unless both sides are zero. Note that this statement requires to have the Lebesgue measure of S^ε be well-defined. Referring to Exercise 33.12 we may as well assume that $|f-g|_1 > 0$ and conclude from (33.8) that

$$|W^\varepsilon| < |O^\varepsilon| + \frac{1}{\varepsilon}|f-g|_1. \tag{33.9}$$

Now suppose that

$$|O^\varepsilon| \le \frac{C_N}{\varepsilon}|f-g|_1 \tag{33.10}$$

for some universal N-dependent constant C_N. We can then choose $|f-g|_1$ as small we like and thereby establish that

$$|W^\varepsilon| = 0$$

for every $\varepsilon > 0$. This will complete the proof because G_f is the complement of the union of the sets

$$W_1, W_{\frac{1}{2}}, W_{\frac{1}{3}}, W_{\frac{1}{4}}, W_{\frac{1}{5}}, W_{\frac{1}{6}}, \ldots,$$

and thereby, see Exercise 33.7, the complement of a set of measure zero.

It remains to estimate $H_{f-g}(x)$ and establish (33.10), but this argument will not depend on the choice of g. So we take $g \equiv 0$ and note that the set

$$O^\varepsilon = \{x \in \mathbb{R}^N : \, H_f(x) > \varepsilon\}$$

is open because

$$x \in O^\varepsilon \iff \exists r > 0 : \underbrace{\int_{B(x,r)} |f|}_{\substack{continuous \\ in\ r\ en\ x}} > \varepsilon |B(x,r)|. \tag{33.11}$$

How big can O^ε be? Every compact $K \subset O^\varepsilon$ is covered[9] by only finite many balls as in (33.11), say

$$K \subset B_1 \cup \cdots \cup B_m.$$

If, for every such $K \subset O^\varepsilon$, these balls were disjoint, then

$$\varepsilon |K| \leq \varepsilon(|B_1 + \cdots + |B_m|) < \int_{B_1} |f| + \cdots + \int_{B_m} |f| = \int_{B_1 \cup \cdots \cup B_m} |f| \leq \int_{\mathbb{R}^N} |f|.$$

and (33.10) would follow with $C_N = 1$.

Remark 33.14. *So in addition to (33.10) we need the statement that the measure of an open set O is the supremum of all the measures of compact subsets K of O. I will come back to get this rid of this issue in Section 33.3.*

It is of course highly unlikely that such disjoint coverings are possible, but with an extra N-dependent factor the estimate does indeed hold. We show below that

$$\varepsilon |O^\varepsilon| \leq 3^N |f|_1, \tag{33.12}$$

as a consequence of what is known as Vitali's covering lemma. This gives (33.10) with $C_N = 3^N$ and will complete the proof.

To wit, choose the[10] largest ball, say B_{j_1}, take it out of the collection and make it the first ball in a new collection. The balls B_i in the old collection for which $B_i \cap B_{j_1} \neq \emptyset$ are all contained in $3B_{j_1}$, the ball with the same center as B_{j_1} but 3 times it radius. Take these B_i out of the old collection and throw them away. If there are any balls left, let B_{j_2} be the largest of these remaining balls in the collection and take it as second ball in the new collection. Repeat the procedure until, say after choosing B_{j_k} and having thrown away all the remaining balls intersecting it, there are no more balls left in the old collection. Then[11]

$$B_{j_1}, \ldots, B_{j_k}$$

[9] Both compactness and measurability rely on definitions with coverings.

[10] Better: a ball which maximizes the radius in the collection.

[11] This is Vitali's covering lemma.

are disjoint, most likely don't cover K of course, but we do have

$$B_1 \cup \cdots \cup B_m \subset 3B_{j_1} \cup \cdots \cup 3B_{j_k},$$

whence

$$|f|_1 \geq \int_{B_1 \cup \cdots \cup B_m} |f| \geq \int_{B_{j_1} \cup \cdots \cup B_{j_k}} |f| = \int_{B_{j_1}} |f| + \cdots + \int_{B_{j_k}} |f|$$

$$> \varepsilon(|B_{j_1}| + \cdots + |B_{j_k}|) = 3^{-N}\varepsilon(|3B_{j_1}| + \cdots + |3B_{j_k}|)$$

$$\geq 3^{-N}\varepsilon|3B_{j_1} \cup \cdots \cup 3B_{j_k}| \geq 3^{-N}\varepsilon|B_1 \cup \cdots \cup B_m| \geq 3^{-N}\varepsilon|K|$$

for all compact $K \in O^\varepsilon$. It follows that (33.12) holds and this then completes the proof that the complement of the good set (33.5) has zero measure.

Theorem 33.15. *Let $f \in L^1(\mathbb{R}^N)$. Then for almost all x it holds that*

$$⨍_{B(x,r)} |f - f(x)| \to 0 \quad \text{as} \quad r \to 0.$$

Exercise 33.16. Apply Theorem 33.10 to the function $s \to |f(s) - q|$ for every $q \in \mathbb{Q}$ to prove Theorem 33.15. Hint: with the integration variable in

$$|f(s) - f(x)| \leq |f(s) - q| + |q - f(x)|$$

being s, it follows that

$$⨍_{B(x,r)} |f - f(x)| \leq ⨍_{B(x,r)} |f - q| + |q - f(x)|.$$

Given x you can take $|q - f(x)|$ as small as you like. Show that the complement of the intersection of all the good sets $G_{|f-q|}$ is a set of measure zero and conclude.

33.3 Vitali coverings and Hardy-Littlewood's again

Both compactness and measurability were defined in terms of properties of coverings of the sets under consideration. In this and later sections we avoid these notions but do use countable coverings with open balls. For convenience we restrict the attention to $f \in L^1(\mathbb{R}^N)$.

Theorem 33.17. *For $f \in L^1(\mathbb{R}^N)$ and $\varepsilon > 0$ let*

$$H_f(x) = \sup_{r>0} \fint_{B(x,r)} |f| \quad \textit{and} \quad O^\varepsilon = \{x \in \mathbb{R}^N : H_f(x) > \varepsilon\}.$$

Then there exists an at most countable family of balls B_i indexed by a subset I of $\mathbb{N}$ such that

$$O^\varepsilon \subset \cup_{i\in I} B_i \quad \textit{with} \quad \sum_{i\in I} |B_i| \leq \frac{6^N}{\varepsilon} |f|_1 .$$

We use Remark 33.8 to prove Theorem 33.17. The set O^ε is open because

$$x \in O^\varepsilon \iff \exists r > 0 : \underbrace{\int_{B(x,r)} |f|}_{\textit{continuous in } r \textit{ en } x} > \varepsilon |B(x,r)|.$$

Thus O^ε is contained in the union of all such balls $B(x,r)$ and close to every $B(x,r)$ there is a ball B with rational center and rational radius such that

$$\int_B |f| > \varepsilon |B| \quad \text{and} \quad x \in B.$$

We conclude that there is a countable family of open balls B_n such that

$$O^\varepsilon \subset \cup_{n\in\mathbb{N}} B_n \quad \text{with} \quad \int_{B_n} |f| > \varepsilon |B_n|,$$

and we may of course assume that non of these balls are concentric. We will show that a subcollection of enlarged balls will do the job.

To see how let r_n be the corresponding sequence of radii and denote the distances between the centers of the balls B_m and B_n by $d_{mn} > 0$. Since

$$\varepsilon |B_n| < |f|_1 ,$$

the sequence r_n is bounded. Let R_1 be its supremum, choose $n_1 \in \mathbb{N}$ with

$$r_{n_1} > \frac{R_1}{2},$$

and let $\tilde{B}_1 = B_{n_1}$. Every ball B_n with $d_{nn_1} \leq 2R_1$ is contained in the ball concentric with $\tilde{B}_1$ with six times its radius, because the radius of this ball $6\tilde{B}_1$ is larger than $3R_1$, and the distance from any point in B_n to the center of $\tilde{B}_1$ is at most $R_1 + 2R_1 = 3R_1$. Throw all these balls away. If there are

any balls left consider the supremum R_2 of the remaining radii and choose n_2 with[12]

$$r_{n_2} > \frac{R_2}{2},$$

and let $\tilde{B}_2 = B_{n_2}$, and throw away all B_n with $d_{nn_2} \leq 2R_2$. And so on. This gives a possibly infinite sequence of disjoint[13] open balls $\tilde{B}_k$ indexed by k, and for every finite sum indexed by a finite subset K of $\mathbb{N}$ we have

$$\varepsilon \sum_{k\in K} |B_k| < \sum_{k\in K} \int_{B_k} |f| = \int_{\cup_{k\in K} B_k} |f| \leq |f|_1 .$$

If the process to choose the balls $\tilde{B}_k$ did not stop at some $k = n \in \mathbb{N}$ it follows that $R_k \to 0$, and thus every ball not chosen as a $\tilde{B}_k$ is eventually thrown away, whence

$$O^\varepsilon \subset \cup_{k\in\mathbb{N}} 6\tilde{B}_k,$$

which we view as $n = \infty$ in

$$O^\varepsilon \subset \cup_{k=1}^{n} 6\tilde{B}_k \tag{33.13}$$

for the case that the process does stop, at some $k = n \in \mathbb{N}$.

For every $m \in \mathbb{N}$ with $m \leq n$ we now have that

$$\varepsilon \sum_{k=1}^{m} |6\tilde{B}_k| = 6^N \varepsilon \sum_{k=1}^{m} |\tilde{B}_k| < 6^N \sum_{k=1}^{m} \int_{\tilde{B}_k} |f| = 6^N \int_{\cup_{k=1}^{m} \tilde{B}_k} |f| \leq 6^N |f|_1 ,$$

so we conclude that

$$O^\varepsilon \subset \cup_{k=1}^{n} 6\tilde{B}_k \quad \text{with} \quad \sum_{k=1}^{n} |6\tilde{B}_k| \leq \frac{6^N}{\varepsilon} |f|_1 \quad \text{and} \quad n \in \mathbb{N} \cup \{\infty\}. \tag{33.14}$$

This completes the proof of Theorem 33.17.

Remark 33.18. *Note that the number* 6 *appears as* $2 \cdot 3$. *Choosing* $p > 1$ *instead of* 2 *it may be replaced by any number*[14] *larger than* 3. *Thus we have shown that the Lebesgue outer measure of* O^ε *is at most*

$$\frac{3^N}{\varepsilon} |f|_1 .$$

[12]The $2 > 1$ in the denominator leads to $3 \cdot 2 = 6 > 3$, any other 2 will also do.

[13]Nontouching because $d_{kl} > 2R_k \geq R_k + R_l$ for $l > k \geq 1$.

[14]If 6 was π...

33.4 Via Cauchy sequences instead?

Observe that Theorem 33.10 identified the in some sense unique best choice $f \in F$ when F is an equivalence class of functions. In hindsight this would justify the sloppy notation

$$f \underbrace{\in [f] = F}_{skipped} \in L^1(\mathbb{R}^N),$$

properly taking into account that $f \in L^1(\mathbb{R}^N)$ is not a space of functions but a space of equivalence classes of functions. From here on we skip F as the notation for the equivalence class $[f]$, as we will be needing a symbol in an alternative approach for introducing the space $L^1(\mathbb{R}^N)$, as consisting of equivalence classes of Cauchy sequences of compactly supported continuous functions f_n.

Definition 33.19. *Two sequences f_n and g_n in $C_c(\mathbb{R}^N)$ are called equivalent if $|f_n - g_n|_1 \to 0$ as $n \to \infty$.*

Cauchy sequences are characterised by the property that

$$|f_n - f_m|_1 = \int_{\mathbb{R}^N} |f_n - f_m| \to 0$$

as $m, n \to \infty$. If $f_n \in C_c(\mathbb{R}^N)$ a Cauchy sequence with respect to the 1-norm and $f_n \sim g_n$ then also g_n is a Cauchy sequence with respect to the 1-norm. Moreover, for every $x \in \mathbb{R}^N$ and every $r > 0$ the sequences

$$\int_{B(x,r)} f_n \quad \text{and} \quad \int_{B(x,r)} g_n$$

are (equivalent) Cauchy sequences in $\mathbb{R}$ with the same limit. Writing

$$F = [f_n]$$

for such am equivalence class we see that

$$\int_{B(x,r)} F = \lim_{n\to\infty} \int_{B(x,r)} f_n, \tag{33.15}$$

is the natural definition of the integral of F over the ball $B(x, r)$, and thus

$$A_F(x, r) = \fint_{B(x,r)} F = \frac{1}{|B(x, r)|} \int_{B(x,r)} F \tag{33.16}$$

is well-defined for every equivalence class.

Theorem 33.20. *Let F be an equivalence class of Cauchy sequences in $C_c(\mathbb{R}^N)$ with respect to the 1-norm, and let $\mathcal{N}_F$ be the set of points x for which*

$$\lim_{r\to 0} A_F(x,r) \tag{33.17}$$

does not exist. Then $\mathcal{N}_F$ is a zero measure set.

For the proof we examine $\mathcal{N}_F$ again using

$$H_F(x) = \sup_{r>0} A_{|F|}(x,r) = \sup_{r>0} \frac{1}{|B(x,r)|}\int_{B(x,r)} |F| \in [0,\infty],$$

in which

$$\int_{B(x,r)} |F| = \lim_{n\to\infty}\int_{B(x,r)} |f_n| \geq |\lim_{n\to\infty}\int_{B(x,r)} f_n| = |\int_{B(x,r)} F|.$$

Note that (33.16) defines a quantity which is continuous as a function of $r>0$ and $x\in\mathbb{R}^N$. This is because

$$|\int_{B(x,r)} F - \int_{B(y,s)} F| \leq$$

$$\underbrace{|\int_{B(x,r)} F - \int_{B(x,r)} f_n|}_{\leq\varepsilon} + |\int_{B(x,r)} f_n - \int_{B(y,s)} f_n| + \underbrace{|\int_{B(y,s)} f_n - \int_{B(y,s)} F|}_{\leq\varepsilon}$$

$$\leq \underbrace{\int_{B(x,r)} |F-f_n|}_{\leq\varepsilon} + |\int_{B(x,r)} f_n - \int_{B(y,s)} f_n| + \underbrace{|\int_{B(y,s)} |f_n - F|}_{\leq\varepsilon}$$

for $n\geq N$, if N corresponds to $\varepsilon>0$ via the definition of f_n being a Cauchy sequence with respect to the 1-norm.

Clearly the definition of the integral of the class F in (33.15) as the limit of the Cauchy sequence

$$\int_{B(x,r)} f_n,$$

has that same N doing the job for $\varepsilon>0$ for all ball $B(x,r)$ simultaneously. In the second middle term we then fix $n=N$ and ask for that difference to be at most $\varepsilon>0$. Since $f_N\in C_c(\mathbb{R}^N)$, this can be done uniformly in terms of the smallness of $|r-s|$ and $|x-y|$. As a result the function

$$(x,r)\to\int_{B(x,r)} F$$

is (uniformly) continuous, just as in (33.11).

We can now consider the existence issue for the limit in (33.17), before we have even identified what its limit value should be, for x in the good set of F, the set for which the limit exists. Most of these limit values will come from Theorem 33.23 in Section 33.5 below, but first we reason as in Theorem 33.10, replacing the basic estimate (33.7) via

$$|A_F(x,r) - A_F(x,s)| \le$$

$$|A_F(x,r) - A_{f_m}(x,r)| + |A_{f_m}(x,r) - A_{f_m}(x,s)| + |A_{f_m}(x,s) - A_F(x,s)|$$
$$\le A_{|F-f_m|}(x,r) + |A_{f_m}(x,r) - A_{f_m}(x,s)| + A_{|f_m-F|}(x,s)$$

for $0 < s < r$ by

$$|A_F(x,r) - A_F(x,s)| \le 2H_{F-f_m}(x) + \underbrace{|A_{f_m}(x,r) - A_{f_m}(x,s)|}_{\to 0 \ as \ r \to 0}. \qquad (33.18)$$

The first term on the right hand side of (33.18) is twice the upper bound $H_{F-f_m}(x)$ for

$$A_{|F-f_m|}(x,r) = A_{|f_m-F|}(x,s),$$

in which $|F - f_m|$ with m fixed denotes the equivalence class of the Cauchy sequence[15] $|f_n - f_m|$.

Let W_ε be the set of all points $x \in \mathbb{R}^N$ for which the statement

$$\exists_{\delta>0} \forall_{r,s\in(0,\delta)} : |A_F(x,r) - A_F(x,s)| \le 2\varepsilon$$

fails. Then (33.18) gives

$$W_\varepsilon \subset O_m^\varepsilon = \{x \in \mathbb{R}^N : H_{F-f_m}(x) > \varepsilon\},$$

and similar to (33.11) we have

$$x \in O_m^\varepsilon \iff \exists r > 0 : \underbrace{\int_{B(x,r)} |F - f_m|}_{\substack{continuous \\ in \ r \ en \ x}} > \varepsilon |B(x,r)|. \qquad (33.19)$$

Exercise 33.21. Modify the proof of Theorem 33.17 to show that W_ε is set of zero measure for every $\varepsilon > 0$. Thus the limit in (33.17) exists outside a set of measure zero. Hint: use (33.14).

[15] Indexed by n.

Remark 33.22. *In view of Theorem 33.20 the function f defined by*

$$f(x) = \lim_{r\to 0} A_F(x,r) \quad \textit{for} \quad x \notin \mathcal{N}_F \quad \textit{and} \quad f(x) = 0 \quad \textit{for} \quad x \in \mathcal{N}_F \tag{33.20}$$

has to be examined next. Is it in $L^1(\mathbb{R}^N)$ and does it coincide with the f chosen in Theorem 33.10?

33.5 Pointwise limits of the Cauchy sequence?

In relation to Remark 33.20 we first try to extract a function from the Cauchy sequence f_n. It remains to be seen if this can be avoided, and give a direct formulation and proof of the desired interpretation of the function f defined in Remark 33.20.

Theorem 33.23. *Given a sequence $f_n \in C_c(\mathbb{R}^N)$ with $|f_n - f_m|_1 \to 0$ and a number $\eta > 0$, we can extract a subsequence along which the sequence converges uniformly on the complement of a union of open balls*

$$U = \cup_{k\in\mathbb{N}} B_k \quad \textit{with} \quad \sum_{k\in\mathbb{N}} |B_k| < \eta.$$

For the proof we observe that

$$|f_n(x) - f_m(x)| \leq$$

$$\underbrace{|f_n(x) - A_{f_n}(x,r)|}_{\to 0 \ \text{ as } \ r\to 0} + \underbrace{|A_{f_n}(x,r) - A_{f_m}(x,r)|}_{\leq H_{f_n - f_m}(x)} + \underbrace{|A_{f_m}(x,r) - f_m(x)|}_{\to 0 \ \text{ as } \ r\to 0},$$

so

$$|f_n(x) - f_m(x)| \leq H_{f_n - f_m}(x), \tag{33.21}$$

and with

$$O^\varepsilon_{mn} = \{x \in \mathbb{R}^N : \, H_{f_n - f_m}(x) > \varepsilon\} \tag{33.22}$$

we have

$$|f_n(x) - f_m(x)| \leq \varepsilon \quad \text{for} \quad x \notin O^\varepsilon_{mn}.$$

We now cover O^ε_{mn} with finitely many open balls such that sum of the measures of these balls is bounded as in (33.10), with $|f-g|_1$ replaced by $|f_n - f_m|_1$, and this norm we can make as small as we like by taking m, n larger then some $N \in \mathbb{N}$. This is an argument that only uses the function $g = f_n - f_m \in C_c(\mathbb{R}^N)$, and is independent of how we arrived at the particular choice of g.

So consider $g \in C^c(\mathbb{R}^N)$ and let

$$x \in O^\varepsilon = \{x \in \mathbb{R}^N : \, H_g(x) > \varepsilon\},$$

in which for every x the supremum

$$H_g(x) = \sup_{r>0} A_{|g|}(x,r) \in [0,\infty]$$

is possibly realised as a maximum by some $r > 0$. If not then the continuity of g in x implies that $H_g(x) = |g(x)|$ because

$$A_{|g|}(x,r) \to |g(x)| \quad \text{as} \quad r \to 0 \quad \text{and} \quad A_{|g|}(x,r) \to 0 \quad \text{as} \quad r \to \infty,$$

the latter being a consequence of $|g|_1 = \int |g|$ existing as the Riemann integral of the continuous function $x \to |g(x)|$ over some large ball.

That being said we only use that $x \in O^\varepsilon$ means that for some radius $r = r_x > 0$ it must be that

$$\int_{B(x,r)} |g| > \varepsilon |B(x,r_x)|.$$

This property is classifying for $x \in O^\varepsilon$. It follows that

$$O^\varepsilon = \{x \in \mathbb{R}^{\mathrm{N}} : \int_{B(x,r)} |g| > \varepsilon |B(x,r)| \quad \text{for some } r > 0\}$$

is open and bounded, and its closure is compact. For every boundary point $\bar{x}$ of O^ε there is a sequence of points $x_n \in O^\varepsilon$ with $x_n \to \bar{x}$ and the sequence[16] r_n of corresponding radii converging to either 0 or to a positive limit $\bar{r} > 0$. In both cases it follows that $H_g(\bar{x}) \geq \varepsilon$, possibly as $H_g(\bar{x}) = |g(\bar{x})|$, but it cannot be that $H_g(\bar{x}) > \varepsilon$, so $H_g(\bar{x}) = \varepsilon$. We can thus cover the closure of O^ε with finitely many balls open balls $\tilde{B}_1, \ldots, \tilde{B}_m$ such that

$$\int_{\tilde{B}_i} |g| > \frac{\varepsilon}{2} |\tilde{B}_i|, \tag{33.23}$$

and using Vitali's covering lemma again we choose $j_1, \ldots, j_k$ such that

$$O^\varepsilon \subset \tilde{B}_1 \cup \cdots \cup \tilde{B}_m \subset 3\tilde{B}_{j_1} \cup \cdots \cup 3\tilde{B}_{j_k} = B_1 \cup \cdots \cup B_k,$$

with $\tilde{B}_{j_1}, \ldots, \tilde{B}_{j_k}$ disjoint, implying that

$$O^\varepsilon \subset B_1 \cup \cdots \cup B_k \quad \text{with} \quad |B_1| + \cdots + |B_k| < \frac{2 \cdot 3^N}{\varepsilon} |g|_1,$$

and likewise for O^ε_{mn}.

[16] Taking a subsequence.

Given $\varepsilon > 0$ and $m, n \in \mathbb{N}$ we have a finite collection of open balls $B_1, \ldots, B_k$ such that

$$\{x \in \mathbb{R}^N : |f_n(x) - f_m(x)| > \varepsilon\} \subset B_1 \cup \cdots \cup B_k \tag{33.24}$$

with

$$|B_1| + \cdots + |B_k| < \frac{2 \cdot 3^N}{\varepsilon} |f_n - f_m|_1 < \eta \tag{33.25}$$

if we choose $m, n \geq N$, N depending on $\frac{1}{2}\eta\varepsilon 3^{-N}$ via the Cauchy-property of the sequence f_n.

Now choose $\varepsilon_k \downarrow 0$ and $\eta_k \downarrow 0$ with

$$\sum_{k=1}^{\infty} \varepsilon_k = \varepsilon, \quad \sum_{k=1}^{\infty} \eta_k = \eta,$$

and corresponding N_k as above. Then f_{N_k} converges uniformly on the complement of the countable union U of all the balls used in (33.24) with $m = N_{k+1}$ and $n = N_k$ for $k = 1, 2, 3, \ldots$.

This completes the proof, but we can do slightly better. After applying the theorem with $\eta > 0$ we obtain a countable union U of open balls. Outside U the constructed subsequence converges uniformly. On the open set U we repeat the proof, which still concerns functions defined on the whole of $\mathbb{R}$ with compact supports. These supports most likely are not subsets of U, but this is of no importance. The required modifications are only minor. We look for coverings

$$B_1 \cup \cdots \cup B_k \supset \{x \in U : |f_n(x) - f_m(x)| > \varepsilon\} \tag{33.26}$$

via coverings of

$$\{x \in U : H_g(x) > \varepsilon\} = O^\varepsilon \cap U.$$

For boundary points of $O^\varepsilon \cap U$ in O^ε the adjustment in (33.23) with the factor $\frac{1}{2}$ is not necesssary, and for the other boundary points we reason as before to conclude that U contains an open union $\tilde{U}$ of balls $\tilde{B}_1, \tilde{B}_2, \ldots$, for which $|\tilde{B}_1| + |\tilde{B}_2| + \cdots < \tilde{\eta}$, outside of which there is uniform convergence. With $\delta_1 = \eta$, $U_1 = U$, $\delta_2 = \tilde{\eta} < \delta_1$, $U_2 = \tilde{U} \subset U = U_1$, we have the first numbers and unions in the sequences in the following theorem.

Theorem 33.24. *Given a sequence $f_n \in C_c(\mathbb{R}^N)$ with $|f_n - f_m|_1 \to 0$ as $m, n \to \infty$, and a sequence of positive real numbers*

$$\delta_1 > \delta_2 > \delta_3 \cdots \to 0,$$

there exists a sequence

$$U_1 \supset U_2 \supset U_3 \supset \cdots$$

of countable unions

$$U_j = \cup_{k\in\mathbb{N}} B_{jk} \quad \textit{with} \quad \sum_{k\in\mathbb{N}} |B_{jk}| < \delta_j,$$

such that for some sequence

$$n_1 < n_2 < n_3 < \cdots$$

of natural numbers it holds that $f_{n_l}(x)$ converges to a limit for every x in the complement of the intersection

$$\mathcal{N} = \cap_{j\in\mathbb{N}} U_j,$$

uniformly on the complement of every U_j. The limit is denoted by

$$f(x) = \lim_{l\to\infty} f_{n_l}(x).$$

The set $\mathcal{N}$ has measure zero and empty interior.

We still have to relate this limit function f to the equivalence class f we started with.

34 Riesz or no Riesz?

Recall that in every Hilbert space the Riesz representation Theorem is applicable, so also in $l^{(2)}$, the standard Hilbert space $H = l^{(2)} = L^2(\mathbb{N})$ with the counting measure on $\mathbb{N}$. Elements u in this H are functions

$$u : \mathbb{N} \to \mathbb{R}.$$

If we denote the values of u in $n \in \mathbb{N}$ by u_n then we can also think of $u \in H = l^{(2)}$ as a sequence $u_1, u_2, \ldots$. We can put these in a column vector

$$u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{pmatrix}$$

and then every such vector has length given by

$$|u| = \sqrt{u \cdot u} = \sqrt{u_1^2 + u_2^2 + \cdots},$$

defined via the inner product

$$u \cdot v = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \end{pmatrix} = u_1 v_1 + u_2 v_2 + u_3 v_3 + \cdots = \sum_{k=1}^{\infty} u_k v_k = (u, v)_H.$$

This inner product is the integral of the product function uv with respect to the counting measure on $\mathbb{N}$. Note that in general uv is not[1] in $l^{(2)} = L^2(\mathbb{N})$.

Exercise 34.1. Give a direct proof of Riesz Representation Theorem for $H = l^{(2)}$. Hint: take a fixed $\phi \in H^*$ and determine what the representing u should be.

Remark 34.2. *Every separable Hilbert space has an orthonormal basis via the Gramm-Schmidt procedure, and is therefore isometrically linearly isomorphic with $H = l^{(2)}$. Thus for separable Hilbert spaces the Riesz Representation Theorem is immediate from Exericise 34.1.*

[1] Many function spaces are not algebra's and this is one of them.

34.1 Other standard Hilbert spaces

The example starts from the observation that there are other measures on $\mathbb{N}$: every sequence of positive numbers

$$0 < \lambda_1 \leq \lambda_2 \leq \lambda_2 \leq \cdots \tag{34.1}$$

definines a measure on $\mathbb{N}$ by assigning measure λ_n to the singleton $\{n\}$. The corresponding integral of the product of two functions $u, v : \mathbb{N} \to \mathbb{R}$ is

$$((u, v)) = (u, v)_V = \sum_{n=1}^{\infty} \lambda_n u_n v_n,$$

defined on a subspace V of our standard space $H = l^{(2)}$. This subspace is not closed in H if $\lambda_n \to \infty$.

Exercise 34.3. Why not? Assume that (34.1) holds. Show that V with $((\cdot, \cdot))$ is a Hilbert space, and that $V = H$ if and only if λ_n is a bounded sequence.

So $V \subset H$, and the norm on V is given by

$$|u|_V = ||u|| = \sqrt{\sum_{n=1}^{\infty} \lambda_n u_n^2},$$

and

$$||u||^2 \geq \lambda_1 |u|^2$$

for all $u \in V$. Here $|u|$ is the standard norm of u. The map

$$i : u \in V \to u \in H$$

is linear and continuous. For all $u \in V \subset H$ we have

$$\underbrace{|i(u)|}_{u \in V} = \underbrace{|u|}_{u \in H} \leq \frac{1}{\sqrt{\lambda_1}} \underbrace{||u||}_{u \in V},$$

in which we think of u in V as lying in both V and H. It follows that

$$|i| = \frac{1}{\sqrt{\lambda_1}}$$

is the norm of i in $L(V, H)$. There is no smaller constant L for which the bound $|u| \leq L||u||$ holds.

Exercise 34.4. Check that $\overline{i(V)} = \overline{V} = H$.

34.2 Double dealing with Riesz

We say that V is dense in H because $\overline{V} = H$. By the Riesz representation Theorem every continuous linear function $\phi : H \to \mathbb{R}$ is of the form

$$\phi(v) = (f, v)$$

with $f = R_H(\phi)$, and of course $\phi(v) = (f, v)$ is also defined for $v \in V$. The map

$$\phi \circ i : v \in V \xrightarrow{i} v \in H \xrightarrow{\phi} (f, v) \in \mathbb{R}$$

is thus continuous and linear, and represented by $u = R_V(\phi \circ i) \in V$. It follows that

$$\phi(v) = (f, v) = ((u, v)) \quad \forall v \in V.$$

The linear continuous functions

$$V \ni u \xrightarrow{f \in H} (f, u)_H \in \mathbb{R}$$

and

$$V \ni u \xrightarrow{u \in V} (u, u)_V \in \mathbb{R}$$

are exactly the same, but given by different (Riesz) representations: we have two different vectors u and f representing the same map via two different inner products.

Exercise 34.5. Assume $0 < \lambda_1 \leq \lambda_2, \ldots$ is unbounded. Why is not every continuous linear $\psi : V \to \mathbb{R}$ of the form $\phi(u) = (f, u)$ with $f \in H$?

It is easier[2] for a linear function on V to be continuous with respect to the norm on V than with respect to the norm on H: there are more continuous lineair functions on V then just the functions

$$v \in V \to (f, v)_H \in \mathbb{R}.$$

If we choose to identify H^* with H via R_H, then

$$V \subsetneq H = H^* \subsetneq V^*,$$

which conflicts with an identification of V^* and V via R_V.

[2] If $\lambda_n \to \infty$.

Nevertheless

$$H \ni f \xrightarrow{R_H^{-1}} \underbrace{\phi \in H^* \xrightarrow{i^*} \phi \circ i \in V^*}_{i^*(\phi)=\phi\circ i} \xrightarrow{R_V} u \in V,$$

is linear and continuous, because the first and third link in this chain are both isometries, and the second link, which is called the adjoint i^* of i, is continuous.

Exercise 34.6. Prove that $i^* : H^* \to V^*$ is linear and continuous. Hint: consider the norm of $i^*(\phi) = \phi \circ i$.

For V and H Hilbert spaces with $i : V \to H$ an injective, continuous linear map with $i(V) \subsetneq \overline{i(V)} = H$ the story is the same.

34.3 A more general abstract perspective

We do not need[3] to assume that $V \subset H$. It's instructive to see how the injectivity of $i : V \to H$ and the density of its range being dense come into play.

Exercise 34.7. Assume H and V are Hilbert spaces and that $i : V \to H$ is linear and continuous. Prove that $S : H \to V$ defined by $f \in H \to u = Sf \in V$ and

$$(u, v)_V = (f, i(v))_H$$

for all $v \in H$ is given by

$$S = R_V \circ i^* \circ (R_H)^{-1},$$

and has norm $|S| = |i^*|$.

Remark 34.8. *We think of S as a solution operator. In the more general case in which the inner product on V is replaced by a (in general) nonsymmetric coercive bilinear form. The Lax-Milgram theorem then replaces the Riesz representation theorem. In Section 17.3 we discuss why this approach still requires a Hilbert space setting. In Section 34.4 we consider the operator S in Exercise 34.7 as a solution operator as map from H to H and as a map from V to V. Look very carefully at the four quotients in Exercise 34.14 and how they are used in Exercise 34.16. They all relate to the solution operator, but one of them does not need the solution operator.*

[3]Note that in our applications to elliptic boundary values problems we do have $V \subset H$.

Exercise 34.9. (continued) Show that

$$|i^*|_{L(H^*,V^*)} = |i|_{L(V,H)}.$$

Hint: we have that i^* is defined by $i^*(\phi) = \phi \circ i$ for every $\phi \in H^*$. This means that

$$\langle i^*(\phi), v\rangle = \langle \phi, i(v)\rangle \tag{34.2}$$

for every $v \in V$ and every $\phi \in H^*$. In case we identify H and H^* this reads

$$\langle i^*(\phi), v\rangle = (\phi, i(v))_H. \tag{34.3}$$

Now

$$|\langle i^*(\phi), v\rangle| = |\langle \phi, i(v)\rangle| \leq |\phi|_{H^*} |i(v)|_H \leq |\phi|_{H^*} |i|_{L(V,H)} |v|_V$$

means that

$$|\langle i^*(\phi)|_{V^*} \leq |\phi|_{H^*} |i|_{L(V,H)},$$

which in turn means that

$$|i^*|_{L(H^*,V^*)} \leq |i|_{L(V,H)}.$$

To bound $|i^*|$ from below take suitable choices of $\phi \in H^*$ and $v \in V$ with $|\phi|_{H^*} = 1$ and $|v|_V = 1$ in the chain

$$|i|_{L(V,H)} \geq |\langle i^*(\phi)|_{V^*} \geq |\langle i^*(\phi), v\rangle| = |\langle \phi, i(v)\rangle|.$$

To wit, take a sequence $v_n \in V$ with $|v_n|_V = 1$ and $|i(v_n)|_H \to |i|$, and then[4] $\phi_n \in H^*$ with $|\phi_n|_{H^*} = 1$ and $\phi_n(i(v_n)) = |i(v_n)|_H$. Conclude that also

$$|i^*|_{L(H^*,V^*)} \geq |i|_{L(V,H)}.$$

Exercise 34.10. Prove that S is injective if $\overline{i(V)} = H$. Hint: this concerns the second equivalence in S injective $\iff$ i^* injective $\iff$ $\overline{i(V)} = H$. Hint: use (34.2) to characterise the null space of i^* in H^*. We have $i^*(\phi) = 0$ if and only if

$$\langle i^*(\phi), v\rangle = \langle \phi, i(v)\rangle$$

for all $v \in V$.

[4]This is really the Hahn-Banach property.

Exercise 34.11. Assume H and V Hilbert spaces, $i : V \to H$ linear and continuous. Let $S : H \to V$ be given via Exercise 34.7 and $f \in H \to u = Sf \in V$ with

$$(u, v)_V = (f, i(v))_H$$

for all $v \in V$. Show that

$$N(i) = \{v \in V : i(v) = 0\} = S(H)^{\perp} = \{v \in V : (u, v)_V = 0 \text{ for all } u \in S(H)\}.$$

Thus the range of S is dense in V if and only if i is injective. Hint: use that $i(v) = 0$ in H if and only if $(f, i(v))_H = 0$ for all $f \in H$.

34.4 The operator remains the same?

Exercise 34.12. Assume $i : V \to H$ linear and continuous. Prove that

$$S_0 = i \circ S : H \to H$$

is symmetric, i.e.

$$(S_0 f_1, f_2)_H = (f_1, S_0 f_2)_H$$

for all $f_1, f_2 \in H$, and

$$(S_0 f, f)_H = |Sf|_V^2 .$$

Exercise 34.13. Assume $i : V \to H$ linear and continuous. Prove that

$$S_1 = S \circ i : V \to V$$

is symmetric, i.e.

$$(S_1 u_1, u_2)_V = (u_1, S_1 u_2)_V$$

for all $u_1, u_2 \in V$, and

$$(S_1 u, u)_V = |i(u)|_H^2 .$$

Exercise 34.14. Show that

$$\frac{(S_0 f, f)_H}{(f,f)_H} = \frac{(Sf, Sf)_V}{(f,f)_H} \quad \text{and} \quad \frac{(i(u), i(u))_H}{(u,u)_V} = \frac{(S_1 u, u)_V}{(u,u)_V}.$$

At the end of Section 19.3, see also (19.12), we showed that taking suprema we obtain the norms of S_0 and S_1 for the left hand sides, and the right hand sides give the squares of norms of i and S. Thus

$$|S_0|_{L(H,H)} = |S|^2_{L(H,V)} = |i^*|^2_{L(H^*,V^*)} = |i|^2_{L(V,H)} = |S_1|_{L(V,V)}$$

via Exercises 34.7 and 34.9.

Exercise 34.15. (continued) If the first supremum is a maximum then its maximizer ϕ is an eigenvector with eigenvalue $\lambda = |S_0|_{L(H,H)}$. You should give a direct proof of this, but see Remark 19.7. Same statement for S_1 and the second supremum of course.

Exercise 34.16. Any eigenvector ϕ of S_0 makes for an eigenvector $\psi = S\phi$ of S_1 with the same eigenvalue, unless $S\phi = 0$. Likewise, any eigenvector ψ of S_1 makes for an eigenvector $\phi = i(\psi)$ of S_0 with the same eigenvalue, unless $i(\psi) = 0$. Show that if one of the suprema in Exercise 34.14 for the norm of S_0 is a maximum, then so is the supremum for the norm of S_1 and vice versa.

Remark 34.17. *Each linear, injective, continuous*[5] *compact*

$$i : V \to H \quad \textit{with} \quad \overline{i(V)} = H$$

defines via Exercises 34.7, 34.12 and 34.13 two strictly positive definite symmetric compact linear mappings $S_0 : H \to H$ and $S_1 : V \to V$ with the same eigenvalues, by dropping either the first or the last link in

$$V \xrightarrow{i} H \xrightarrow{(R_H)^{-1}} H^* \xrightarrow{i^*} V^* \xrightarrow{R_V} V \xrightarrow{i} H.$$

The triple

$$V \subset H = H^* \subset V^*$$

with V and H Hilbert spaces, $i : V \to H$ injective and $V = i(V)$ dense in H is the standard framework in the French PDE school, see the Brézis book on functional analysis.

[5]Follows from compactness of i.

35 Sobolev spaces

This chapter is roughly based on the section about mollifiers in the appendix of Evans' PDE book and the chapter about Sobolev spaces. We let U be an open set in $\mathbb{R}^{\mathrm{N}}$ and consider functions u and v rather than f and g. For $1 \leq p < \infty$ the Lebesgue p-norm of a function $v \in C_c(U)$ is defined by

$$|v|_p^p = \int_{\mathbb{R}^{\mathrm{N}}} |v|^p, \tag{35.1}$$

and the Sobolev $W^{1,p}$-norm of a function $v \in C_c^1(U)$ by

$$|v|_{1,p}^p = |v|_p^p + |v_{x_1}|_p^p + \cdots + |v_{x_{\mathrm{N}}}|_p^p, \tag{35.2}$$

The spaces $L^p(\mathbb{R}^{\mathrm{N}})$ and $W_0^{1,p}(U)$ are the closures of $C_c^1(U)$ with respect to the p-norm and the $W^{1,p}$-norm. In case of the $W^{1,p}$-norm the closure is either taken in a not yet defined larger space, or in the abstract sense with equivalence classes of Cauchy sequences. In what follows partial derivatives will be taken in the distributional sense[1], with the functions in $C_c^1(U)$, the space of continuously differentiable functions with compact support in U, acting as test functions.

35.1 Mollifiers and density tricks

We first restrict the attention to $L^p(\mathbb{R}^{\mathrm{N}})$. We note that in case of the p-norm the closure of $C_c^1(U)$ is the same as the closure of $C_c(U)$, and it may also be taken in the abstract sense with equivalence classes of Cauchy sequences. One way or another, Theorem **??** allows for every $u \in L^1_{loc}(\mathbb{R}^{\mathrm{N}})$ the introduction of its ε-mollified version

$$u^\varepsilon(x) = (\eta_\varepsilon * u)(x) = \int_{\mathbb{R}^{\mathrm{N}}} \eta_\varepsilon(x-y)u(y)\,dy = \int_{\mathbb{R}^{\mathrm{N}}} \eta_\varepsilon(y)u(x-y)\,dy$$

$$= \int_{|y|\leq\varepsilon} \eta_\varepsilon(y)u(x+y)\,dy,$$

in which

$$\eta_\varepsilon(x) = \frac{1}{\varepsilon^{\mathrm{N}}}\eta(\frac{x}{\varepsilon}), \quad \eta(x) = \eta(|x|) \geq 0, \quad \eta \in C_c^\infty(B), \quad \int_{\mathbb{R}^{\mathrm{N}}} \eta = 1,$$

B denoting the open unit ball. In practice we prove statements about u^ε and u via calculations with v^ε, $v \in C_c^1(U)$, but we do note that u^ε is smooth.

[1]The definition of (unique) weak derivatives relies on Theorem 33.15.

Exercise 35.1. Use the techniques presented in Section 14.2 to show that u^ε is smooth.

I try to restrict the use of mollifiers to globally defined locally integrable functions u for which

$$u^\varepsilon(x) = \int_{\mathbb{R}^N} \eta_\varepsilon(x-y)u(y)\,dy = \int_{\mathbb{R}^N} \eta_\varepsilon(y)u(x-y)\,dy = \int_{|y|\le\varepsilon} \eta_\varepsilon(y)u(x+y)\,dy.$$

For $u \in L^p(\mathbb{R}^N)$ Hölder's inequality with

$$\frac{1}{p} + \frac{1}{q} = 1$$

gives

$$|u^\varepsilon(x)| \le \int_{|y|\le\varepsilon} \eta_\varepsilon(y)^{\frac{1}{q}+\frac{1}{p}} |u(x+y)|\,dy \le \left(\int_{|y|\le\varepsilon} \eta_\varepsilon(y)|u(x+y)|^p\,dy\right)^{\frac{1}{p}},$$

whence, taking the p-th power, integration gives[2] the following result for $u \in L^p(\mathbb{R}^N)$, which is complemented by a convergence result in case $v \in C_c^1(U)$.

Theorem 35.2. *Let $u \in L^p(\mathbb{R}^N)$. Then*

$$\int_{\mathbb{R}^N} |u^\varepsilon(x)|^p\,dx \le \int_{\mathbb{R}^N} |u(x)|^p\,dx, \text{ i.e. } \quad |u^\varepsilon|_p \le |u|_p. \tag{35.3}$$

Theorem 35.3. *Let $v \in C_c^1(U)$ and $1 \le p < \infty$. Then*

$$|v^\varepsilon(x) - v(x)| \le \left(\int_{\mathbb{R}^N} |v(x+\varepsilon y) - v(x)|^p\,\eta(y)dy\right)^{\frac{1}{p}}. \tag{35.4}$$

for $v \in C_c^1(U)$, and

$$|v^\varepsilon - v|_p \le \varepsilon\,|\nabla v|_p, \tag{35.5}$$

in which the right hand side is the p-norm of the Euclidean length of ∇v.

Before we prove Theorem 35.3 we note that it is via these two theorems, namely (35.3) with $u - v$, the splitting

$$|u - u^\varepsilon|_p \le |u - v|_p + |v - v^\varepsilon|_p + |v^\varepsilon - u^\varepsilon|_p,$$

the density of $C_c^1(U)$ in $L^p(\mathbb{R}^N)$ and (35.5), that the following theorem follows.

[2]Changing the order of integration, same trick as in Exercise 35.22.

Theorem 35.4. *Let $u \in L^p(\mathbb{R}^N)$, then $|u^\varepsilon|_p \le |u|_p$ and $|u^\varepsilon - u|_p \to 0$.*

Exercise 35.5. Prove Theorem 35.4 using Theorems 35.2 and 35.3.

Remark 35.6. *We shall also use (35.5) in Section 35.3 for a direct proof that every bounded sequence u_n in $W_0^{1,p}(U)$ has a subsequence which, considered as a sequence in $L^p(U)$, is convergent. In fact this could already be an exercise here.*

Exercise 35.7. Let u_n be a bounded sequence in $W_0^{1,p}(U)$. Prove that there is a subsequence of u_n which is Cauchy with respect to the p-norm. Hint: use the splitting

$$|u_n - u_m|_p \le |u_n - v_n|_p + |v_n - v_n^\varepsilon|_p + |v_n^\varepsilon - v_m^\varepsilon|_p + |v_m^\varepsilon - v_m|_p + |v_m - u_m|_p,$$

deal with the second and fourth term by (35.5), with the first and fifth term by density of $C_c^1(U)$ in $W_0^{1,p}(U)$, and finally with the third term by Theorem 8.13 (with $[0,1]$ replaced by a large closed box) and ε-dependent bounds for ε fixed. A diagonal argument completes the proof.

For the proof of Theorem 35.3 write

$$v^\varepsilon(x) - v(x) = \int_{\mathbb{R}^N} \eta_\varepsilon(y) v(x+y)\, dy - v(x) = \int_{\mathbb{R}^N} \eta_\varepsilon(y)\, (v(x+y) - v(x))\, dy$$

$$= \int_{\mathbb{R}^N} \eta(y)\, (v(x+\varepsilon y) - v(x))\, dy = \int_{\mathbb{R}^N} \eta(y)^{\frac{1}{q}} \eta(y)^{\frac{1}{p}}\, (v(x+\varepsilon y) - v(x))\, dy.$$

Hölder's inequality gives

$$|v^\varepsilon(x) - v(x)| \le \underbrace{\left(\int_{\mathbb{R}^N} \eta(y)\, dy \right)^{\frac{1}{q}}}_{=1} \left(\int_{\mathbb{R}^N} \eta(y) |v(x+\varepsilon y) - v(x)|^p\, dy \right)^{\frac{1}{p}}.$$

whence (35.4) follows.

We next use the mean value theorem in integral form, see (13.1), and the Hölder estimate

$$|\int_0^1 f|^p \le |\int_0^1 |f|^p \qquad (1 \le p < \infty).$$

The x-integral in the right hand side of (35.4) is estimated for $|y| \le \varepsilon$ as

$$\int_{\mathbb{R}^N} |v(x+\varepsilon y) - v(x)|^p\, dx = \int_{\mathbb{R}^N} |[v(x+t\varepsilon y)]_{t=0}^{t=1}|^p\, dx$$

$$= \int_{\mathbb{R}^N} |\int_0^1 \nabla v(x+\varepsilon t y)\cdot \varepsilon y\, dt|^p\, dx$$

$$\le \varepsilon^p \int_{\mathbb{R}^N} \left(\int_0^1 |\nabla v(x+\varepsilon t y)|\, dt\right)^p dx \le \varepsilon^p \int_{\mathbb{R}^N}\int_0^1 |\nabla v(x+\varepsilon t y)|^p\, dt\, dx$$

whence (35.4) gives

$$\int_{\mathbb{R}^N} |v^\varepsilon(x)-v(x)|^p\, dx \le \int_{\mathbb{R}^N} \eta(y) \int_{\mathbb{R}^N} |v(x+\varepsilon t y)-v(x)|^p\, dx\, dy \le$$

$$\le \int_{\mathbb{R}^N} \eta(y)\, \varepsilon^p \int_{\mathbb{R}^N}\int_0^1 |\nabla v(x+\varepsilon t y)|^p\, dt\, dx\, dy,$$

and (35.5) follows by changing the order of integration from $dt\, dx\, dy$ to $dx\, dt\, dy$. This completes the proof.

Theorem 35.8. *Let $u \in L^p(U)$, $U \subset \mathbb{R}^N$ open, and $V \subset\subset U$. Then $u^\varepsilon \to u$ in $L^p(V)$.*

Theorem 35.8 follows by extending u to $\tilde{u}$ defined on the whole of $\mathbb{R}^N$ via $\tilde{u}(x) = u(x)$ for $x \in U$ and $\tilde{u}(x) = 0$ for $x \notin U$. Then Theorem 35.4 implies $\tilde{u}^\varepsilon \to \tilde{u}$ in $L^p(\mathbb{R}^N)$ and thus also in $L^p(V)$, but on V we have $\tilde{u} = u$ and $\tilde{u}^\varepsilon = u^\varepsilon$ if ε is small.

In Appendix C.5 Theorem 7 Evans proves a similar result for functions $u \in L^p_{loc}(U)$ via basically[3] $v \in C_c(U)$. Such functions are uniformly continuous, i.e

$$\forall_{\delta>0}\, \exists_{\eta>0}\, \forall_{x,y\in U} : |x-y| < \eta \implies |v(x)-v(y)| < \delta.$$

Then

$$|v^\varepsilon(x)-v(x)| = |\int_{|y|\le\varepsilon} \eta_\varepsilon(y)(v(x+y)-v(x))\, dy| < \delta,$$

provided $\varepsilon \le \eta$. This proves that $v^\varepsilon \to v$ uniformly, i.e $|v^\varepsilon - v|_\infty \to 0$ as $\varepsilon \to 0$.

Theorem 35.9. *Let $v \in C_c(\mathbb{R}^N)$, then $v^\varepsilon \in C_c(\mathbb{R}^N)$, $|v^\varepsilon|_\infty \le |v|_\infty$ and $|v^\varepsilon - v|_\infty \to 0$.*

Remark 35.10. *Evans' proof that $u \in L^1_{loc}(\mathbb{R}^N)$ implies $u^\varepsilon(x) \to u(x)$ for almost every x relies on Theorem 33.15.*

[3]He has u defined on a subset only.

35.2 Sobolev spaces of functions with weak derivatives

The concept of weak derivative comes as a theorem.

Theorem 35.11. *Suppose that*

$$\int_U v\phi = -\int_U u\phi_{x_i} \tag{35.6}$$

for every $\phi \in C_c^1(U)$, for some given u and v in $L^1_{loc}(U)$, $U \subset \mathbb{R}^N$. Then v is unique in $L^1_{loc}(U)$ for $u \in L^1_{loc}(U)$. We say that v is the weak derivative of u with respect to its i^{th} variable, notation $v = D_i u = u_{x_i}$.

For the proof suppose that some other v, say $\tilde{v} \in L^1_{loc}(U)$ also satisfies this condition then the difference $w = v - \tilde{v}$ is in $L^1_{loc}(U)$ and satisfies

$$\int_U w\phi = 0$$

for every $\phi \in C_c^1(U)$. Take an open ball $B \subset U$ and redefine $w(x) = 0$ for $x \notin B$. The mollifier w^ε is then identically zero on $\mathbb{R}^N$ for all $\varepsilon > 0$. By Theorem 35.8 we have

$$|w|_1 = |w^\varepsilon - w|_1 \to 0$$

as $\varepsilon \to 0$. But $|w|_1$ doesn't go anywhere. So $|w|_1 = 0$, and Theorem 33.10 tells us what w is, as a function: zero! It follows that $v = \tilde{v}$ in B outside a set of measure zero, for every $B \subset U$ open. Thus $v = \tilde{v}$ in U outside a set of measure zero.

Definition 35.12. *For $1 \leq p < \infty$ and $U \subset \mathbb{R}^N$ the space $W^{1,p}(U)$ is defined as the space of functions $u \in L^p(U)$ for which the weak derivatives $u_{x_1}, \ldots, u_{x_N}$ exist and are in $L^p(U)$.*

Exercise 35.13. Prove that $W^{1,p}(U)$ is a Banach space with the norm defined by

$$|u|^p_{1,p} = |u|^p_p + |u_{x_1}|^p_p + \cdots + |u_{x_N}|^p_p. \tag{35.7}$$

Hint: use that $L^p(U)$ is a Banach space.

Remark 35.14. *You may now prefer to define $W^{1,p}_0(U)$ as the closure of $C^1_c(U)$ in $W^{1,p}(U)$. As a closure in a Banach space the space $W^{1,p}_0(U)$ is itself then also complete.*

Exercise 35.15. Every $u \in W_0^{1,p}(U)$ extends to a $u \in W_0^{1,p}(\mathbb{R}^N)$ by setting u equal to zero outside U. Hint: a similar statement holds for $u \in C_c^1(U)$ and $C_c^1(\mathbb{R}^N)$.

Remark 35.16. *Every $C_c^k(U)$ with $k \in \mathbb{N}$ is dense in $W_0^{1,p}(U)$ and so is $C_c^\infty(U)$, the space of test functions used throughout in the literature.*

Exercise 35.17. A bit harder and to do somewhere along the road in this chapter:

$$W_0^{1,p}(\mathbb{R}^N) = W^{1,p}(\mathbb{R}^N).$$

Remark 35.18. *The definitions of $W^{2,p}(U)$ and of $W_0^{2,p}(U)$ (the closure of $C_c^k(U)$ with $k \geq 2$) should be obvious, starting from the definition of weak second order derivatives $u_{x_ix_j}$ and $u_{x_jx_i}$, the inevatible observation that $u_{x_ix_j} = u_{x_jx_i}$ under assumptions you should figure out, and the norm defined by*

$$|u|_{2,p}^p = |u|_p^p + \sum_{1\leq i\leq N} |u_{x_i}|_p^p + \sum_{1\leq i\leq j\leq N} |u_{x_ix_j}|_p^p.$$

Exercise 35.19. Fill in the details of Remark 35.18 and generalise to $W^{k,p}(U)$ and $W_0^{k,p}(U)$ with $k \geq 2$.

35.3 Compactness for $W_0^{1,p}(U)$

In Section 35.8 we use calculus to derive the Gagliardo-Nirenberg-Sobolev[4] and Morrey estimates and identify $p = N$ as a critical exponent. We will have

Theorem 35.20. *Let $p > N$ and $u \in W_0^{1,p}(\mathbb{R}^N)$. Then, after redefining u on a zero measure set,*

$$u : \mathbb{R}^N \to \mathbb{R}$$

is contininous, and, as a consequence of the uniform continuity,

$$\max_{\substack{x\in\mathbb{R}^N \\ |x|=R}} |u(x)| \to 0 \quad as \quad R \to \infty.$$

[4]GNS-estimates for short.

The Morrey estimates come with a uniform modulus of continuity which via the Ascoli-Arzelà theorem[5] *implies that the inclusion map* $W_0^{1,p}(U) \to C(\bar{U})$ *is compact if* U *is bounded*[6].

What follows is not restricted to $p > N$, and was announced in Remark 35.6 of Section 35.1. The proof of Theorem 35.21 below is also based on the AA Theorem, which states for compact metric spaces X that a sequence u_n in $C(X)$, the space of $\mathbb{R}$-valued continuous functions on X with norm

$$|u|_\infty = \max_{x \in X} |u(x)|,$$

has a convergent[7] subsequence u_{n_k}, provided the sequence is bounded in $C(X)$ and has the property[8] that

$$\forall \varepsilon > 0 \, \exists \delta > 0 \, \forall n \in \mathbb{N} \, \forall x, y \in X : \, d(x,y) < \delta \implies |f_n(x) - f_n(y)| < \varepsilon.$$

The other ingredient in the proof of Theorem 35.21 is the use of mollifiers.

Theorem 35.21. *We have for all* $p \in [1, \infty)$ *that*

$$W_0^{1,p}(U) \to L^p(U) \quad \textit{is compact if} \quad U \quad \textit{is bounded.}$$

NB: the weaker statement that the embedding is bounded will be characterised by the Poincaré inequality: $|u|_p \leq C_{pU} |\nabla u|_p$ *for all* $u \in W_0^{1,p}(U)$, C_{pU} *a constant depending on* p *and* U *only*[9].

For a large part the proof has already been done in Section 35.1. Consider a bounded sequence $u_n \in W_0^{1,p}(U)$, choose $v_n \in C_c^1(U)$ and extend v_n to $v_n \in C_c^1(\mathbb{R}^N)$, such that $|u_n - v_n|_{1,p} \to 0$, and consider mollified versions of v_n defined by

$$v_n^\varepsilon(x) = (\eta_\varepsilon * v)(x) = \int_{\mathbb{R}^N} \eta_\varepsilon(x-y) v_n(y) \, dy = \int_{\mathbb{R}^N} \eta_\varepsilon(y) v_n(x-y) \, dy$$

$$= \int_{|y| \leq \varepsilon} \eta_\varepsilon(y) v_n(x+y) \, dy,$$

in which

$$\eta_\varepsilon(y) = \frac{1}{\varepsilon^N} \eta(\frac{y}{\varepsilon}) \quad \text{with} \quad 0 \leq \eta \in C_c^\infty(B) \quad \text{radial}, \quad \int_B \eta = 1,$$

[5] See Section 8.3, AA Theorem for short.

[6] For $U = \mathbb{R}^N$ compactness fails for the same reason as Theorem 8.13. Which reason?

[7] Here convergence means uniform convergence.

[8] This is called the (uniform) equicontinuity of the sequence u_n.

[9] And thereby also on the dimension N.

where B is the open unit ball. Then split $|u_n - u_m|_p$ as

$$|u_n - u_m|_p \le |u_n - v_n|_p + |v_n - v_n^\varepsilon|_p + |v_n^\varepsilon - v_m^\varepsilon|_p + |v_m^\varepsilon - v_m|_p + |v_m - u_m|_p.$$

The first and fifth term converge to zero in view of $|u_n - v_n|_{1,p} \to 0$, the second and fourth are controlled by, with $v = v_n$ and $v = v_m$, the estimate

$$\int_{\mathbb{R}^N} |v^\varepsilon(x) - v(x)|^p \, dx \le \varepsilon^p \int_{\mathbb{R}^N} |\nabla v(x)|^p \, dx, \tag{35.8}$$

and therefore bounded by $C\varepsilon$, with C depending only on the bound for the sequence $|v_n|_{1,p}$, and thus only on the original bound for the sequence u_n in $W_0^{1,p}(U)$,

For every $\varepsilon > 0$ fixed the third term converges to zero along a subsequence in view of the AA Theorem applied to the sequence v_n^ε in $C(\bar{V})$, with V bounded and slight larger than U so as to have all v_n^ε in $C_c^\infty(V)$. We thus have that $|u_n - u_m|_p$ is asymptotically controlled by $C\varepsilon$ along that subsequence, along which $|v_n^\varepsilon - v_m^\varepsilon|_\infty \to 0$. A standard diagonal argument now produces a subsequence which is a Cauchy sequence in $L^p(V)$. The following exercises[10] fill in the details and are all that's needed to conclude the proof of Theorem 35.21.

Exercise 35.22. Show that

$$\int_{\mathbb{R}^N} |v^\varepsilon(x) - v(x)|^p \, dx \le \int_{\mathbb{R}^N} \eta(y) \int_{\mathbb{R}^N} |v(x + \varepsilon y) - v(x)|^p \, dx \, dy. \tag{35.9}$$

Hint: write the difference $v^\varepsilon(x) - v(x)$ as a single y-integral over $|y| \le \varepsilon$, scale y to have the integral over the unit ball, use

$$\eta(y) = \eta(y)^{\frac{1}{p'}} \eta(y)^{\frac{1}{p}},$$

and apply Hölder's inequality. Then use this estimate for the integral of $|v^\varepsilon(x) - v(x)|^p$ and change the order of integration to conclude.

Exercise 35.23. Write the difference in the right hand side of (35.9) as an s-integral over the interval $[0, 1]$, use Hölder's inequality and $|y| \le 1$ to arrive at an integral over s and x only, and obtain (35.8).

[10] We already did most of them in Section 35.1.

Exercise 35.24. Show that for fixed $\varepsilon > 0$ the sequence v_n^ε has a convergent subsequence in $C(\bar{V})$.

We apply the AA Theorem. The bounds

$$|v_n^\varepsilon(x)| = |\int_{\mathbb{R}^N} \eta_\varepsilon(x-y)v_n(y)\,dy| \le |\eta_\varepsilon|_{p'}|v_n|_p$$

and

$$|\nabla v_n^\varepsilon(x)| = |\int_{\mathbb{R}^N} \nabla\eta_\varepsilon(x-y)v_n(y)\,dy| \le |\nabla\eta_\varepsilon|_{p'}|v_n|_p$$

imply that the conditions of the AA Theorem are satisfied if v_n is merely a bounded sequence in $L^p(U)$ extended by $v_n(x) = 0$ for $x \notin U$, whence v_n has a convergent subsequence on the compact closure of an ε-neighbourhood of U. Since v_n is in fact bounded in $W_0^{1,p}(U)$ we are done (also for $p = 1$).

35.4 The need for extension operators

To extend the results for $W_0^{1,p}(U)$ to $W^{1,p}(U)$ we need a well behaved extension operator that maps $W^{1,p}(U)$ into $W_0^{1,p}(\tilde{U})$ with $\tilde{U}$ slightly larger than U. Boundary straightenings and partitions of unity[11] will play a crucial role here, just like in the proof of the local version of the Gauss divergence or Green's theorem in (21.13), and the step to the global version in (14.8). The extension operator is first defined for $u \in C^1(\bar{U})$ and requires the boundary ∂U to be bounded and C^1 (locally the graph of a C^1-function).

Partitions of unity are first used to establish that $W^{1,p}(U)$ itself is the closure of $C^1(\bar{U})$ if ∂U is bounded and C^1. This is Theorem 35.33, which is pretty explicit in how the approximations are constructed. It shows that the assumptions on the boundary can be weakened. But we apply Theorem 35.33 to domains which are assumed to have ∂U bounded and C^1 for other reasons.

Remark 35.25. *The elements of the afore mentioned partitions are suitably chosen $\zeta \in C_c^\infty(\mathbb{R}^N)$ which, when multiplied by $u \in W^{1,p}(U)$, produce products $\zeta u \in W^{1,p}(U)$ to which the Leibniz rule applies*[12]*. Any statement that we want make about a function $u \in W^{1,p}(U)$ can be localised using partitions of unity, splitting u via $\zeta_0, \zeta_1, \ldots, \zeta_n \in C_c^1(\mathbb{R}^N) \subset C_c^\infty(\mathbb{R}^N)$ with $\zeta_0 + \zeta_1 + \cdots + \zeta_n \equiv 1$ on $\bar{U}$, writing*

$$u = u_0 + u_1 + \cdots + u_n = \zeta_0 u + \zeta_1 u + \cdots + \zeta_n u,$$

*in which we take $\zeta_0 \in C_c^1(U)$ and $\zeta_1, \ldots, \zeta_n \in C_c^1(\mathbb{R}^N)$, just like in (**??**).*

[11] Introduced in (**??**) and explained in Chapter 29.

[12] Evans' stronger assumption $\zeta \in C_c^\infty(U)$ for Leibniz' rule leads to cumbersome details.

35.5 Mollifiers and weak derivatives

The following remarks summarise what we have and what we don't have yet towards the density of $C^1(\bar{U})$ in $W^{1,p}(U)$.

Remark 35.26. *The basic estimates*[13]

$$|u^\varepsilon|_p \le |u|_p \quad \textit{and} \quad |v^\varepsilon - v|_p \le \varepsilon\, |\nabla v|_p$$

for $u \in L^p(\mathbb{R}^N)$ and $v \in C^1_c(\mathbb{R}^N)$ sufficed to show that

$$u^\varepsilon \to u \quad \text{in} \quad L^p(\mathbb{R}^N) \quad \textit{as} \quad \varepsilon \to 0 \tag{35.10}$$

for every $u \in L^p(\mathbb{R}^N)$ via

$$|u^\varepsilon - u|_p \le \underbrace{|u^\varepsilon - v^\varepsilon|_p}_{\le |u-v|_p} + |v^\varepsilon - v|_p + |v - u|_p, \tag{35.11}$$

in the proof of Theorem 35.4: first choose $v \in C^1_c(\mathbb{R}^N)$ such that $|u - v|_p$ is small, say less than δ, and then choose $\varepsilon > 0$ to make also $|v^\varepsilon - v|_p$ less than δ. The same statements hold with $\mathbb{R}^N$ replaced by an open set $U \subset \mathbb{R}^N$.

Remark 35.27. *If $u \in W^{1,p}(\mathbb{R}^N)$ and $v \in C^2_c(\mathbb{R}^N)$ then Remark 35.26 applies to*[14] *$w_i = u_{x_i} = D_i u \in L^p(\mathbb{R}^N)$ and $D_i v = v_{x_i} \in C^1_c(\mathbb{R}^N)$. Since*

$$D_i(u^\varepsilon) = (u^\varepsilon)_{x_i} = (u_{x_i})^\varepsilon = (D_i u)^\varepsilon \tag{35.12}$$

it follows that

$$u^\varepsilon \to u \quad \text{in} \quad W^{1,p}(\mathbb{R}^N) \quad \textit{as} \quad \varepsilon \to 0.$$

The same statements do not hold with $\mathbb{R}^N$ replaced by an open set $U \subset \mathbb{R}^N$. We prove (35.12) below and then worry about what to do for U.

We have[15]

$$w_i^\varepsilon(x) = (u_{x_i})^\varepsilon(x) = (D_i u)^\varepsilon(x) = \int_{\mathbb{R}^N} \eta_\varepsilon(x-y)(D_i u)(y)\, dy \tag{35.13}$$

$$= \int_{\mathbb{R}^N} (D_i \eta_\varepsilon)(x-y) u(y)\, dy = \Big(\int_{\mathbb{R}^N} \eta_\varepsilon(x-y) u(y)\, dy\Big)_{x_i} = (D_i u^\varepsilon)(x),$$

which is and proves (35.12).

[13] Applied with u replaced by $u - v$.

[14] In practice: $u_{x_i} - v_{x_i}$.

[15] Note that D_i acts on u to give $D_i u$ which we can evaluate in x, $x - y$, y and so on.

Exercise 35.28. Explain why the inequalities in the above chain hold. Hint: you first need the techniques from Section 14.2, then the definition of weak derivatives in Theorem 35.11, and then again the techniques from Section 14.2.

We record the positive result in Remark 35.27 as

Theorem 35.29. *Let $u \in W^{1,p}(\mathbb{R}^N)$, $1 \le p < \infty$. Then*

$$u^\varepsilon \to u \quad \text{in} \quad W^{1,p}(\mathbb{R}^N) \quad \textit{as} \quad \varepsilon \to 0.$$

Now what about $u \in W^{1,p}(U)$ if U is an open subset of $\mathbb{R}^N$? Note that we can extend not only u but also $w_1 = D_1 u, \dots, w_N = D_N u$ to $\mathbb{R}^N$ by setting $u(x) = w_1(x) = \cdots = w_N(x) = 0$ for $x \notin U$, but then we don't have that $w_i = D_i u$ in $L^p(\mathbb{R}^N)$ for $i = 1, \dots, N$. We can only conclude that

$$w_i^\varepsilon = D_i u^\varepsilon \quad \text{in} \quad L^p(U_\varepsilon), \quad U_\varepsilon = \{x \in U : B(x, \varepsilon) \subset U\}. \tag{35.14}$$

It is only on this issue that the reasoning in Remark 35.27 towards

$$u^\varepsilon \to u \quad \text{in} \quad W^{1,p}(U) \quad \text{as} \quad \varepsilon \to 0 \tag{35.15}$$

fails. We still have that

$$u^\varepsilon \to u \quad \text{and} \quad w_i^\varepsilon \to w_i \quad \text{for} \quad i = 1, \dots, N \quad \text{in} \quad L^p(\mathbb{R}^N) \quad \text{as} \quad \varepsilon \to 0,$$

and then also in $L^p(U)$. We next use localisations and translates to prove that, provided the boundary ∂U is sufficiently nice, there is a family[16] $u_\varepsilon \in C^1(\bar{U})$ with $u_\varepsilon \to u$ in $W^{1,p}(U)$. To do so we first establish equality in (35.14) for mollified translates of localised functions $\tilde{w}$ and $\tilde{u}$ obtained from $u \in W^{1,p}(U)$.

35.6 Shifts and localisation

The translation trick goes as follows. Given $h > 0$, a unit vector e and a function $u \in L^p(\mathbb{R}^N)$ we define $u_{he} \in L^p(\mathbb{R}^N)$ by[17]

$$u_h(x) = u(x + he)$$

and consider u_h^ε. Since clearly

$$(u_h)^\varepsilon = (u^\varepsilon)_h$$

[16] Do note ε is a subscript now, so $u_\varepsilon \neq u^\varepsilon$.

[17] Dropping e from the subscript notation.

it follows from (35.10) that

$$|u_h^\varepsilon - u_h|_p = |u^\varepsilon - u|_p \to 0 \quad \text{in} \quad L^p(\mathbb{R}^N) \quad \text{as} \quad \varepsilon \to 0.$$

But then

$$|u_h^\varepsilon - u|_p \le \underbrace{|u_h^\varepsilon - u_h|_p}_{=|u^\varepsilon - u|_p} + |u_h - u|_p = \underbrace{|u^\varepsilon - u|_p}_{\to 0} + \underbrace{|u_h - u|_p}_{\to 0}, \tag{35.16}$$

the latter because

$$|u_h - u|_p \le \underbrace{|u_h - v_h|_p}_{=|v-u|_p} + |v_h - v|_p + |v - u|_p \tag{35.17}$$

for every $v \in C_c(\mathbb{R}^N)$, similar to (35.11).

Exercise 35.30. Use (35.17) with $v \in C_c(\mathbb{R}^N)$ and $|v-u|_p$ as small as desired to prove that $u_h \to u$ in $L^p(\mathbb{R}^N)$. Hint: use the uniform continuity of each such v.

We conclude from (35.16) that

$$u_h^\varepsilon \to u \quad \text{in} \quad L^p(\mathbb{R}^N) \quad \text{as} \quad h \to 0 \quad \text{and} \quad \varepsilon \to 0, \tag{35.18}$$

which will be used in next both for $\tilde{u}$ and $\tilde{w}_i$ as functions extended by zero outside their original domain.

Recall that the limitation to U_ε in (35.14) kept us from concluding that $u^\varepsilon \to u$ in $W^{1,p}(U)$. We now localise u by multiplying it by a function $\zeta \in C_c^1(\mathbb{R}^N)$ and consider shifts $\tilde{u}_h$ of $\tilde{u} = \zeta u$ defined by the choice of a fixed unit vector e to be chosen in relation to ζ, and the local description of U and its boundary in and near the support of ζ. Note that

$$\tilde{u} \in W^{1,p}(\tilde{U}), \quad \tilde{U} = U \cup (\operatorname{supp}\zeta)^c,$$

and that

$$\tilde{u}_h \in W^{1,p}(\tilde{U}_h)$$

is defined by

$$\tilde{u}_h(x) = \tilde{u}(x + eh) \quad \text{for} \quad x \in \tilde{U}_h = \{x \in \mathbb{R}^N : x + eh \in \tilde{U}\}.$$

We extend $\tilde{u}$ and $\tilde{w}_i = D_i\tilde{u}$, defined in $L^p(\tilde{U})$, to $L^p(\mathbb{R}^N)$ by setting

$$\tilde{u}(x) = \tilde{w}_i(x) = 0 \quad \text{for} \quad x \notin \tilde{U} \quad \text{and} \quad i = 1, \dots, N$$

as before, and know from (35.18) that $u^\varepsilon, w_i^\varepsilon \in C_c^\infty(\mathbb{R}^N)$ have the property that

$$u_h^\varepsilon \to u \quad \text{and} \quad w_{ih}^\varepsilon \to w_i \quad \text{in} \quad L^p(\mathbb{R}^N) \quad \text{as} \quad h \to 0 \quad \text{and} \quad \varepsilon \to 0.$$

To conclude that

$$\tilde{u}_h^\varepsilon \to \tilde{u} \quad \text{in} \quad W^{1,p}(U)$$

we need

$$\tilde{w}_{ih}^\varepsilon = D_i \tilde{u}_h^\varepsilon \quad \text{in} \quad L^p(U), \tag{35.19}$$

and in view of (35.14) this will follow if

$$U \subset \tilde{U}_{h\varepsilon} = \{x \in \tilde{U}_h : B(x, \varepsilon) \subset \tilde{U}_h\}. \tag{35.20}$$

35.7 Global density of smooth functions

We use a partition of unity as in (**??**) with each $\zeta_1, \ldots, \zeta_n$ taking care of some part of the boundary of U, and $\zeta_0 \in C_c^\infty(U)$.

Exercise 35.31. Use Theorem 35.29 to show that $(\zeta_0 u)^\varepsilon \to \zeta_0 u$ in $W^{1,p}(U)$ as $\varepsilon \to 0$.

Without loss of generality we continue the reasoning in the 2-dimensional setting, with

$$\text{supp } \zeta = [\tilde{a}, \tilde{b}] = [\tilde{a}_1, \tilde{b}_1] \times [\tilde{a}_2, \tilde{b}_2],$$
$$a_1 < \tilde{a}_1 < \tilde{b}_1 < b_1, \quad a_2 < \tilde{a}_2 < \tilde{b}_2 < b_2, \quad f : [a_1, b_1] \to (\tilde{a}_2, \tilde{b}_2)$$

Lipschitz continuous with Lipschitz constant L,

$$U \cap (a, b) = \{(x, y) : a_1 < x < b_1,\ f(x) < y < b_2\},$$

e the second unit vector, whence

$$\tilde{u}_h(x, y) = \tilde{u}(x, y + h)$$

and

$$\tilde{U}_h \supset \{(x, y) :\ a_1 < x < b_1,\ y > f(x) - h\}.$$

Now let $\lambda = \sqrt{1 + L^2}$ and $h = \lambda\varepsilon$. Then every point in $[\tilde{a}, \tilde{b}] \cap U$ with

$$x_N \geq f(x_1, \ldots, x_{N-1}) + \lambda\varepsilon$$

is the center of an open ball with radius $\varepsilon > 0$ that is contained in $(a, b) \cap U$, provided ε is smaller than the distance from $[\tilde{a}, \tilde{b}]$ to the boundary of (a, b). This implies (35.20) holds with $h = \lambda\varepsilon$, whence

$$\tilde{u}_{\lambda\varepsilon}^\varepsilon \to \tilde{u} \quad \text{in} \quad W^{1,p}(U) \tag{35.21}$$

Exercise 35.32. To convince yourself of the statement preceding (35.21) draw a picture in the xy-plane with the line $y = Lx$ and find the point $P_\varepsilon = (0, \lambda\varepsilon)$ on the positive y-axis with distance ε to that line, and the point Q_ε on that line which realises this distance. Shift the origin $O = (0,0)$ to a point on the graph $y = f(x)$ contained in $[\tilde{a}, \tilde{b}]$, and pull the triangle $OP_\varepsilon Q_\varepsilon$ along. Specify the smallness condition on ε.

The above argument applies to every $\zeta_1, \ldots, \zeta_n$. Combined with Exercise 35.31 this allows to conclude that the following theorem has been proved.

Theorem 35.33. *Assume*[18] *that U allows a partition of unity $\zeta_0, \zeta_1, \ldots, \zeta_n$ such as used above. For every $u \in W^{1,p}(U)$ there exists a family $u_\varepsilon \in C_c^\infty(\mathbb{R}^N)$ with $u_\varepsilon \to u$ in $W^{1,p}(U)$. We can take*[19]

$$u_\varepsilon = (\zeta_0 u)^\varepsilon + \sum_{i=1}^{n} (\zeta_i u)^\varepsilon_{\lambda\varepsilon e_i},$$

in which λ is the largest $\sqrt{1+L^2}$ that occurs in the construction. Thus the result is valid under the assumption that ∂U is compact and uniformly Lipschitz continuous[20]*. If $u \in W^{k,p}(U)$ with $k \in \mathbb{N}$ and $1 \le p < \infty$ then $u_\varepsilon \to u$ in $W^{k,p}(U)$.*

35.8 Estimates and embeddings for $W_0^{1,p}(U)$

Estimates derived for functions in $C_c^1(U)$ carry over to functions in $W_0^{1,p}(U)$, and there are two basic estimates to which this principle is applied. The first Gagliardo-Nirenberg-Sobolev estimate is

$$|u|_q \le C_{p,N} |\nabla u|_p \quad \text{for} \quad \frac{1}{q} = \frac{1}{p} - \frac{1}{N} \quad \text{if} \quad 1 \le p < N, \tag{35.22}$$

in which the norm[21] of ∇u is evaluated via an integral over the whole of U, and over the whole of $\mathbb{R}^N$ in the derivation, via repeated application of the one-dimensional estimate[22]

$$|u|_\infty \le \frac{1}{2} |u|_1 \quad \text{for} \quad u \in C_c^1(\mathbb{R}), \tag{35.23}$$

[18] See Chapter 29 for C^1-boundaries. TO DO: non smooth boundaries!

[19] Bringing unit vectors e_i back into the notation, these do not have to the basis vectors.

[20] Give a definition of what this should mean.

[21] For the p-norm of ∇u the p-norm of any vector norm of $\nabla u(x)$ can be used.

[22] The case $p = N = 1, q = \infty$ in (35.22), does not generalise to $p = N > 1, q = \infty$.

first[23] in the special case that $p = 1$, with a clever use of the Hölder's inequality with exponents satisfying

$$\frac{1}{p_1} + \cdots + \frac{1}{p_{n-1}} = 1.$$

The general case in (35.22) follows from putting u^γ for u and a follow your nose estimate invoking Hölder's inequality for the integral of $\gamma|u|^{\gamma-1}u_{x_i}$, which involves a particular choice of γ to get the exponents right. The constant $C_{p,\mathrm{N}}$ blows up as $p \to \mathrm{N}$ (from below).

The second (Morrey) estimate[24] is usally stated as

$$|u(x_1) - u(x_2)| \le C_{p,\mathrm{N}} |\nabla u|_p |x_1 - x_2|^\alpha \quad \text{for} \quad \alpha = 1 - \frac{\mathrm{N}}{p} \quad \text{if} \quad p > \mathrm{N},$$

but the p-norm of ∇u may be restricted to the intersection of the two balls centered in x_1 and x_2 with radius $|x_1 - x_2|$. That is

$$|u(x_1) - u(x_2)| \le C_{p,N} |\nabla u|_{L^p(W_{x_1x_2})} |x_1 - x_2|^{1-\frac{N}{p}}, \tag{35.24}$$

in which

$$W_{x_1x_2} = B(x_1, |x_1 - x_2|) \cap B(x_2, |x_1 - x_2|).$$

This estimate is derived from the inequality

$$\int_{C_R} |u - u(0)| \le \frac{R^\mathrm{N}}{\mathrm{N}} \int_{C_R} \frac{|u_r|}{r^{\mathrm{N}-1}} \tag{35.25}$$

for cones described in polar coordinates as

$$C_R = \{x = r\omega : 0 \le r \le R,\, \omega \in A\},$$

with A a nice subset of the unit sphere, and u_r denoting the radial derivative. The r-part of the integral in (35.25) is in some sense the counter part of (35.23), and integral on the right hand side is estimated using a follow your nose estimate invoking Hölder's inequality.

The Morrey estimate (35.24) is then proved estimating

$$|u(x_1) - u(x_2)| \le |u(x_1) - u(x)| + |u(x) - u(x_2)|,$$

and integrating over the intersection of the two cones C_1 and C_2 centered in x_1 and x_2, chosen to have $C_1 \cup C_2$ equal to the union of the two balls mentioned earlier. Again the constant $C_{p,\mathrm{N}}$ blows up as $p \to \mathrm{N}$ (from above).

[23] See Section **??**.

[24] See Section **??**.

Exercise 35.34. Let $1 \le p < \mathrm{N}$. Prove that $W_0^{1,p}(U) \subset L^q(U)$ if $\frac{1}{q} \ge \frac{1}{p} - \frac{1}{\mathrm{N}}$ if U is bounded, and that in that case $|\nabla u|_p$ defines an equivalent norm on $W_0^{1,p}(U)$. Theorem 35.21 already stated the desired estimate. Make C_{pU} as explicit as possible in terms of p, N and the measure of U.

Exercise 35.35. Let $p > \mathrm{N}$. Prove that $W_0^{1,p}(U) \subset C^\alpha(U)$ for $\alpha = 1 - \frac{\mathrm{N}}{p}$, in which

$$C^\alpha(U) = \{u \in C(U) : [u]_\alpha < \infty\} \quad \text{where} \quad [u]_\alpha = \sup_{x_1 \neq x_2} \frac{|u(x_1) - u(x_2)|}{|x_1 - x_2|^\alpha},$$

the supremum taken over the whole of U.

Exercise 35.36. This is to convince you that it is better to rename $C^\alpha(U)$ and write $C^\alpha(\bar{U})$: show that for U bounded and $\alpha \in (0, 1]$, every $u \in C^\alpha(U)$ extends to a continuous function on $\bar{U}$, and that the space $C^\alpha(U)$ is a Banach space with norm defined by[25] $|u|_\alpha = |u|_\infty + [u]_\alpha$.

Exercise 35.37. Let $p > \mathrm{N}$, U bounded, $\alpha = 1 - \frac{\mathrm{N}}{p}$. Use the AA Theorem to prove that every bounded sequence u_n in $W_0^{1,p}(U)$ has a subsequence that, considered as a sequence in $C(\bar{U})$, converges uniformly to a limit u, which is also in $C_0^\alpha(U)$, the subspace consisting of functions $u \in C^\alpha(U)$ which vanish on ∂U. Verify that this subspace has the seminorm $[\cdot]_\alpha$ as an equivalent norm.

In view of Exercise 35.37 the embedding

$$W_0^{1,p}(U) \to C(\bar{U})$$

is compact for $p > \mathrm{N}$ if U is bounded. This is Theroem 35.20. Since $C(\bar{U}) \subset L^p(U)$, with the obvious bound on the norms, it then also holds that

$$W_0^{1,p}(U) \to L^p(U) \quad \text{is compact if} \quad U \quad \text{is bounded,} \tag{35.26}$$

but this holds for all $p \ge 1$ because of Theorem 35.21 via a different argument[26].

[25] If you don't use Greek letters for Lebesgue norms this will not confuse.

[26] But it is still the AA Theorem after all.

Exercise 35.38. Let $1 \leq p < \mathrm{N}$ and $U \subset \mathbb{R}^{\mathrm{N}}$ open and bounded. Prove that the embedding

$$W_0^{1,p}(U) \to L^q(U)$$

is compact if

$$\frac{1}{q} > \frac{1}{p} - \frac{1}{\mathrm{N}}.$$

Hint: use Theorem 35.21 and interpolation inequalities with the p-norms.

35.9 Statements for $W^{1,p}(U)$ via extension; traces

Given a bounded domain U we look for a slightly larger domain $\tilde{U}$ such that every $u \in W^{1,p}(U)$ extends to a $\tilde{u} \in W_0^{1,p}(\tilde{U})$ in the sense that $\tilde{u}(x) = u(x)$ for (almost) all $x \in U$. The extension map

$$u \in W^{1,p}(U) \xrightarrow{E} \tilde{u} \in W_0^{1,p}(\tilde{U})$$

should be linear and bounded. The extensions are first defined for $u \in C^1(\bar{U})$ and require U bounded and $\partial U \in C^1$.

I usually followed Evans' approach in which the extensions are first defined locally for u and then glued together using a partition of unity, but it came to me that here too it is in fact simpler to first cut up u in smaller pieces $\zeta_i u$ and choose globally defined extensions of $\zeta_i u$ rather than locally defined extensions of u. This requires a suitable set of functions

$$\zeta_0 \in C_c^\infty(O_0), \zeta_1 \in C_c^\infty(O_1), \ldots, \zeta_n \in C_c^\infty(O_n),$$

with $0 \leq \zeta_i \leq 1$ and

$$\zeta_0(x) + \zeta_1(x) + \cdots + \zeta_n(x) = 1 \quad \text{for all} \quad x \in \bar{U}$$

with $O_0 \subset \bar{O}_0 \subset U$ and $O_1, \ldots, O_n$ chosen so as to allow globally defined extensions $\tilde{u}_1, \ldots, \tilde{u}_n$ of $u_1 = \zeta_1 u, \ldots, u_n = \zeta_n u$ which define

$$u \xrightarrow{E_i} \tilde{u}_i$$

as a linear map with

$$|\tilde{u}_i|_{1,p} \leq C_i \, |u|_{1,p},$$

allowing to define

$$u \in C_c^1(\bar{U}) \xrightarrow{E} C_c^1(\tilde{U}) \quad \text{by} \quad u \to \zeta_0 u + \tilde{u}_1 + \cdots + \tilde{u}_n.$$

In view of the Leibniz rule

$$(\zeta u)_{x_j} = \zeta u_{x_j} + \zeta_{x_j} u$$

it follows that

$$|Eu|_{1,p} \leq C\,|u|_{1,p},$$

with C some horrible constant depending on $\tilde{U}$ and U via the norms of ζ_i in C^1. The functions $\zeta_1, \ldots, \zeta_n$ are chosen to allow a C^1-coordinate transformation similar to the ones used by Evans. It all looks a bit cleaner if the O_i are taken, after a permutation of coordinates, in cylindrical form as $C_i = B_i \times I_i$, with B_i an open ball, I_i a bounded interval, and

$$U \cap C_i = \{x = (x_1, \ldots, x_{N-1}, x_N) \in C : \, x_N > \gamma_i(x_1, \ldots, x_{N-1})\},$$

in which $\gamma_i : \bar{B}_i \to I_i$ is C^1, and the supports of the ζ_i are contained in smaller cylinders $\tilde{C}_i = \tilde{B}_i \times \tilde{I}_i \subset\subset C_i$. The extensions of $\zeta_i u$ are then defined via the transformations in Appendix C.1 and the higher order reflection in Section 5.4.

Finally

$$u \in C^1(\bar{U}) \xrightarrow{E} \tilde{u} \in C_c^1(\tilde{U}),$$

as then defined with the desired $W^{1,p}$-estimates, has to be extended as a map from $W^{1,p}(U)$ to $W_0^{1,p}(\tilde{U})$ using the density of $C^1(\bar{U})$ in $W^{1,p}(U)$.

We needed the boundedness of U and the C^1-regularity of ∂U to define an extension operator

$$C^1(\bar{U}) \xrightarrow{E} C_c^1(\tilde{U})$$

with the $W^{1,p}(\tilde{U})$-norm bound of Eu controlled by the $W^{1,p}(U)$-norm of u. The domain $\tilde{U}$ can be taken as close to U as desired by taking the cylinders C_i small, and E extends to

$$W^{1,p}(U) \xrightarrow{E} W_0^{1,p}(\tilde{U})$$

via the density result in Theorem 35.33. The other important operator is the bounded linear trace operator

$$W^{1,p}(U) \xrightarrow{T} L^p(\partial U)$$

in Section 5.4 of Evans, which extends

$$u \in C^1(\bar{U}) \to u|_{\partial U} \in C^1(\partial U).$$

Evans defines it locally, first under the assumption that ∂U is flat and $u \in C^1(\bar{U})$. The same splitting as in Theorem 35.33 can be used to first define $T(\zeta_i u)$ instead, for $u \in C^1(\bar{U})$, so

$$u \in C^1(\bar{U}) \to \zeta_i u = u_i \in C^1(\bar{U} \cap \bar{C}_i) \to u_i|_{\partial U} \in C^1(\partial U).$$

The local coordinate transformation flattening $\partial U \cap \bar{C}_i$ is not even needed, as u_i is defined for all $x_N \geq \gamma(x_1, \ldots, x_{\mathrm{N}-1})$ with $(x_1, \ldots, x_{\mathrm{N}-1}) \in B_i$ and vanishes for x_{N} large. Thus

$$Tu_i(x_1, \ldots, x_{\mathrm{N}-1}) = u_i(x_1, \ldots, x_{\mathrm{N}-1}, \gamma(x_1, \ldots, x_{\mathrm{N}-1})) = -\int_\gamma^\infty (u_i)_{x_{\mathrm{N}}},$$

and the p-norm on B_i is estimated by the p-norm of ∇u_i, the factor

$$\left(1 + \gamma_{x_1}^2 + \cdots + \gamma_{x_{\mathrm{N}-1}}^2\right)^{\frac{1}{2}}$$

being irrelevant for the estimate. The characterisation of the kernel of T as in Theorem 2 of Section 5.4 is also done locally then, as Evans observes in (6), in which the flattening avoids cumbersome notation in the already technical proof that follows. Actually the proof is not so hard. It relies on this estimate, formulated in $\mathbb{R}^2$ without loss of generality for $u \in C_c^2(\mathbb{R}^2)$:

$$\int_{-\infty}^\infty |u(x,y)|^p \, dx \leq 2^p \left(\int_{-\infty}^\infty |u(x,0)|^p \, dx + y^{p-1} \int_0^y \int_{-\infty}^\infty |u_y|^p \right). \quad (35.27)$$

Exercise 35.39. Prove (35.27) and explain why it holds for $u \in W^{1,p}(\mathbb{R}^2_+)$ with compact support in $\mathbb{R} \times [0, \infty)$.

Exercise 35.40. If such a u has $Tu = 0$, then the functions u_m defined by $u_m(x,y) = (1 - \zeta(my))u(x,y)$ with $\zeta \in C_c^\infty([0,2))$ and $\zeta \equiv 1$ on $[0,1]$, $\zeta' \leq 0$ on $[0,2)$ are in $W_0^{1,p}(\mathbb{R}^2_+)$ and converge to u in $W^{1,p}(\mathbb{R}^2_+)$. Prove this and conclude that $u \in W_0^{1,p}(\mathbb{R}^2_+)$. Hint: you have to use Exercise 35.27.

Exercise 35.41. Let U be a bounded domain in $\mathbb{R}^2 = \{(x,y) : x, y \in \mathbb{R}\}$ and $\zeta \in C^1(\mathbb{R}^2)$. Prove that $\zeta u \in W^{1,p}(U)$ if $u \in W^{1,p}(U)$.

Exercise 35.42. Introduce new coordinates ξ, η by

$$x = x_0 + a\xi + b\eta, \quad y = y_0 + c\xi + d\eta,$$

with $ad \neq bc$, define V by $(\xi, \eta) \in V \iff (x, y) \in U$, and write $v(\xi, \eta) = u(x, y)$. Show that

$$u \in W^{1,p}(U) \iff v \in W^{1,p}(V)$$

and that this correspondence defines a linear homeomorphism between the two Sobolev spaces.

Exercise 35.43. Assume $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$ is C^1, injective on $\bar{U}$, with invertible Jacobian matrix in every $(x, y) \in \bar{U}$. Then $u \to u \circ \Phi^{-1} = v$ defines a bijective map from $C^1(\bar{U}) \to C^1(\bar{V})$ where $V = \Phi(U)$. Show that

$$\frac{1}{C}|v|_{W^{1,p(V)}} \leq |u|_{W^{1,p(U)}} \leq C|v|_{W^{1,p(V)}}$$

for some $C > 1$.

Exercise 35.44. Explain why this map uniquely extends to a bijection from $W^{1,p}(U)$ to $W^{1,p}(V)$ if $\partial U \in C^1$.

Exercise 35.45. The intersection of $W^{1,p}(U)$ and $C(\bar{U})$ is a Banach space with norm e.g. $|u|_\infty + |u_x|_p + |u_y|_p$. Explain why this Banach space is the closure of $C^1(\bar{U})$ with respect to this norm.

Exercise 35.46. Let $u \in C^1(\mathbb{R}^3)$. Write $u(x, y, z) = v(r, \theta, \phi)$, with

$$x = r\sin\theta\cos\phi,\ y = r\sin\theta\sin\phi,\ z = r\cos\theta,$$

and let $C_{R,\psi}$ be the set in $\mathbb{R}^3$ defined by $0 \leq r \leq R$, $0 \leq \theta \leq \psi$ and ϕ free. If $0 < \psi < \frac{\pi}{2}$ and $R > 0$, then ψ is called[27] the opening angle of the closed cone $C_{R,\psi}$, and R is called the radius of the cone. For $\psi > \frac{\pi}{2}$ we don't call $C_{R,\psi}$ a cone. It's a half ball with radius R if $\psi = \frac{\pi}{2}$ and a ball if $\psi = \pi$. Evans only integrates over balls. This is to clarify and improve the remark after the proof of Theorem 4 in his Section 5.6.2 on Morrey's inequality. Assuming that $u(0, 0, 0) = v(0, \theta, \phi) = 0$ we used

$$|v(r, \theta, \phi)| \leq \int_0^r |v_r(\rho, \theta, \phi)|\, d\rho$$

[27] And not 2ψ.

to estimate

$$\int_{C_{R,\psi}} |u| = \int_0^{\psi}\int_0^{2\pi}\int_0^{R} |v(r,\theta,\phi)|\, r^2 \sin\theta \, dr\, d\phi\, d\theta$$
$$\leq \int_0^{\psi}\int_0^{2\pi}\int_0^{R}\int_0^{r} |v_r(\rho,\theta,\phi)|\, d\rho\, r^2 \sin\theta \, dr\, d\phi\, d\theta$$

(interchanging the order of the integrations with respect to r and ρ, throwing a away one negative term and replacing ρ by r again)

$$\leq \frac{R^3}{3}\int_0^{\psi}\int_0^{2\pi}\int_0^{R} \frac{|v_r|}{r^2}\, r^2 \sin\theta \, dr\, d\phi\, d\theta = \frac{R^3}{3}\int_{C_{R,\psi}} \frac{|u_r|}{r^2},$$

in which $u_r = xu_x + yu_y + zu_z = v_r$. Use generalised polar coordinates

$$x_1 = r\omega_1 = r\cos\theta_1 = rc_1,\ x_2 = r\omega_2 = r\sin\theta_1\cos\theta_2 = rs_1c_2,\ x_3 = r\omega_3 = rs_1s_2c_3,$$
$$\ldots, x_{N-2} = r\omega_{N-1} = rs_1\cdots s_{N-2}c_{N-1},\ x_N = r\omega_N = rs_1\cdots s_{N-2}s_{N-1}$$

to generalise and improve this estimate as

$$\int_{C_{R,\psi}} |u| + \frac{1}{N}\int_{C_{R,\psi}} r|u_r| \leq \frac{R^N}{N}\int_{C_{R,\psi}} \frac{|u_r|}{r^{N-1}} \tag{35.28}$$

for $C_{R,\psi}$ in $\mathbb{R}^{\mathrm{N}}$ defined by $0 \leq r \leq R$, $0 \leq \theta_1 \leq \psi$ and $\theta_2, \ldots, \theta_{N-1}$ free.

Exercise 35.47. (continued) Let $\omega \in \mathbb{R}^{\mathrm{N}}$ with $|\omega| = 1$. Explain why estimate (35.28) holds with $C_{R,\psi}$ replaced by the closed cone

$$C_{R,\omega,\psi} = \{x \in \mathbb{R}^{\mathrm{N}} : |x| \leq R,\ x\cdot\omega \geq |x|\cos\psi\}, \tag{35.29}$$

a cone with direction[28] ω and opening angle $\psi \in (0, \frac{\pi}{2})$.

Exercise 35.48. Explain why the measure $|C_{1,\omega,\psi}|$ of the cone $C_{1,\omega,\psi}$ defined by (35.29) is given by

$$\int_{C_{1,\omega,\psi}} 1 = \frac{2\pi}{N}\int_0^{\psi} \sin^{N-2}\theta_1\, d\theta_1 \int_0^{\pi} \sin^{N-1}\theta_2\, d\theta_2 \cdots \int_0^{\pi} \sin\theta_{N-2}\, d\theta_{N-2}. \tag{35.30}$$

if $R = 1$. Call this number $C_{N\psi}$. Show that

$$C_{N\psi} = \frac{\omega_{N-1}}{N}\int_0^{\psi} \sin^{N-2}\theta\, d\theta,$$

in which ω_{N-1} is the measure of the unit ball in $\mathbb{R}^{N-1}$. Correct my mistakes. Is there a quicker way? Does the integral simplify if $\psi = \frac{\pi}{3}$?

[28] Note $\omega = e_3$ in the 3-dimensional example, $\omega = e_1$ in the N-dimensional example.

Exercise 35.49. We use Hölder's inequality[29] to estimate

$$\int_{C_{R,\omega,\psi}} \frac{|u_r|}{r^{N-1}} \le \left(\int_{C_{R,\omega,\psi}} \left(\frac{1}{r^{N-1}} \right)^{p'} \right)^{\frac{1}{p'}} \underbrace{\left(\int_{C_{R,\omega,\psi}} |u_r|^p \right)^{\frac{1}{p}}}_{|u_r|_{L^p(C_{R,\omega,\psi})}} \quad \text{with} \quad \frac{1}{p} + \frac{1}{p'} = 1.$$

Show that

$$\int_{C_{R,\omega,\psi}} \frac{|u_r|}{r^{N-1}} \le \left(C_{N\psi} \, \frac{p-1}{p-N} \right)^{1-\frac{1}{p}} |u_r|_{L^p(C_{R,\omega,\psi})} \, R^{1-\frac{N}{p}} \tag{35.31}$$

if $p > N$. Explain why the estimate holds for all $u \in C^1(C_{R,\omega,\psi})$. Why does the estimate fail for $p \le N$?

Combining (35.28) and (35.31) we have

$$\frac{N}{R^N} \int_{C_{R,\omega,\psi}} |u| \le \left(C_{N\psi} \, \frac{p-1}{p-N} \right)^{1-\frac{1}{p}} |u_r|_{L^p(C_{R,\omega,\psi})} \, R^{1-\frac{N}{p}}, \tag{35.32}$$

in which ψ can have any value in $[0, \pi]$. This estimate is hidden in Step 2 of the proof of Theorem 4 in Evans' Section 5.6.2, and only given there for $\psi = \pi$.

Exercise 35.50. For $R > 0$, $x_1, x_2, \omega_1, \omega_2 \in \mathbb{R}^N$ with $|\omega_1| = |\omega_2| = 1$ and angles ψ_1, ψ_2, consider $C_1 = x_1 + C_{R,\omega_1,\psi_1}$ and $C_2 = x_2 + C_{R,\omega_2,\psi_2}$. Use

$$|C_1 \cap C_2| \, |u(x_1) - u(x_2)| \le \int_{C_1} |u(x_1) - u(x)| dx + \int_{C_2} |u(x) - u(x_2)| dx$$

and (35.32) to show that

$$|C_1 \cap C_2| \, |u(x_1) - u(x_2)| \le \frac{R^N}{N} \left(C_{N\psi_1}^{1-\frac{1}{p}} + C_{N\psi_2}^{1-\frac{1}{p}} \right) \left(\frac{p-1}{p-N} \right)^{1-\frac{1}{p}} |\nabla u|_{L^p(C_1 \cup C_2)} \, R^{1-\frac{N}{p}}.$$

Exercise 35.51. In Exercise 35.50 take[30]

$$R = |x_1 - x_2|, \; \omega_1 = \frac{x_2 - x_1}{R} = -\omega_2, \; \psi = \frac{\pi}{3},$$

[29] Which for integrals follows from the inequality in Section 20.3.

[30] Sketch the balls with centers x_1 and x_2 and radius $|x_1 - x_2|$ to see what's going on.

and prove that

$$|u(x_1) - u(x_2)| \leq C(N,p)|\nabla u|_{L^p(B_{|x_1-x_2|}(x_1)\cap(B_{|x_1-x_2|}(x_2))}\,|x_1 - x_2|^{1-\frac{N}{p}}, \quad (35.33)$$

in which

$$C(N,p) = c_{_N}\frac{\left(\int_0^{\frac{\pi}{3}}\sin^{N-2}\theta\,d\theta\right)^{1-\frac{1}{p}}}{\omega_{N-1}{}^{\frac{1}{p}}} \quad (35.34)$$

with $c_{_N}$ to be specified by you. Hint: show first that the measure A_N of the set described by

$$x_1 \geq 0,\, x_2 = r\cos\theta_1,\, x_3 = r\sin\theta_1\cos\theta_2, \ldots,\, x_N = r\sin\theta_1\cdots\sin\theta_{N-2},\, r \geq 0,$$

and

$$x_1 + \frac{r}{\sqrt{3}} \leq \frac{1}{2}$$

is

$$A_N = \frac{\omega_{N-1}3^{\frac{N-1}{2}}}{N(N-1)2^N},$$

and explain why $2A_NR^N$ is the measure of the intersection of the two cones C_1 and C_2. Another hint in hindsight: for $N = 2$ the value of A_2 is immediate from a picture. Do A_3 first with high school calculus and guess the formula for $N > 3$.

Thus we have improved the estimate stated by Evans without proof in the remark following the proof of Theorem 4. Now recall the definitions of $W^{1,p}(U)$ and $W_0^{1,p}(U)$ for U in $\mathbb{R}^{\mathrm{N}}$ bounded, open and connected,

$$u \in W^{1,p}(U) \iff u, u_{x_1}, \ldots, u_{x_N} \in L^p(U),$$

and, for $1 \leq p < \infty$, the space $W_0^{1,p}(U)$ being the closure of $C_c^1(U)$ in the Banach space $W^{1,p}(U)$.

Remark 35.52. *Every statement we will ever be able to make about $W^{1,p}(U)$ will be based on a statement about $W_0^{1,p}(\tilde{U})$ for a slightly larger $\tilde{U}$ and some extension $\tilde{u}$ of u from U to $\tilde{U}$, which will heavily depend on the properties of the boundary of U.*

Exercise 35.53. Let $u \in W_0^{1,p}(U)$, U in $\mathbb{R}^{\mathrm{N}}$ bounded, open and connected, $N < p < \infty$ and let $\alpha = 1 - \frac{N}{p}$. Take a sequence $u_n \in C_c^\infty(U)$ with $u_n \to u$ in $W^{1,p}(U)$. Prove that u_n is a Cauchy sequence in $C^\alpha(\bar{U})$, and that its limit $\bar{u}$ in $C^\alpha(\bar{U})$ has the property that $|u - \bar{u}|_{_p} = 0$. Prove that the map $u \to \bar{u}$ is linear and continuous from $W_0^{1,p}(U)$ to $C^\alpha(\bar{U})$.

Exercise 35.54. (continued) A rough estimate for the seminorm

$$[u]_\alpha = \sup_{\substack{x,y\in U\\ x\neq y}} \frac{|u(x)-u(y)|}{|x-y|^\alpha}$$

with $\alpha = 1 - \frac{N}{p}$: show that

$$[u]_{1-\frac{N}{p}} \leq C(p,N)|\nabla u|_{L^p(U)},$$

and also show that

$$|u|_\infty \leq \tilde{C}(p,N,U)|\,|\nabla u|_{L^p(U)}$$

for some constant $\tilde{C}(p,N,U)$ you can make as precise as you want. Give just one. Hint: first for $u \in C_c^1(U)$, reason as in Exercise 35.53 to get the estimate for all $u \in W_0^{1,p}(U)$ if $p > N$.

Exercise 35.55. Show that $C^\alpha(\bar{U})$ is a Banach space.

Exercise 35.56. Show that $W_0^{1,p}(U)$ is compactly embedded in $C_0(U)$. Hint: use the Ascoli-Arzéla theorem via Exercise 35.54.

Exercise 35.57. Let $0 < \beta < \alpha < 1$. Show that

$$[u]_\beta \leq [u]_\alpha + C_{\alpha\beta}|u|_\infty,$$

in which $C_{\alpha\beta}$ is a constant depending on α and β only. Hint: it is easy to estimate $[u]_\alpha$ by a product of powers of $[u]_\beta$ and $|u|_\infty$. Use Young's inequality

$$ab \leq \frac{\varepsilon^p a^p}{p} + \frac{b^q}{q\varepsilon^q} \quad \text{for} \quad \varepsilon > 0,\, a,b \geq 0,\, p,q \geq 1 \quad \text{with} \quad \frac{1}{p}+\frac{1}{q} = 1$$

to conclude.

Exercise 35.58. Use Exercises 35.56 and 35.57 to conclude that $W_0^{1,p}(U)$ is compactly embedded in $C^\beta(\bar{U})$ is $0 < \beta < 1 - \frac{N}{p}$.

Exercise 35.59. Show that $W_0^{1,p}(U)$ is embedded in $h^\alpha(\bar{U})$, the closed subspace[31] of $C^\alpha(\bar{U})$ for which

$$\sup_{\substack{x,y\in U\\ 0<|x-y|\le\varepsilon}} \frac{|u(x)-u(y)|}{|x-y|^\alpha} \to 0 \quad \text{as} \quad \varepsilon \to 0,$$

if $\alpha = 1 - \frac{N}{p}$ and $p > N$.

It is easy to see that

$$|f(x)| \le \frac{1}{2}\int_{-\infty}^{\infty} |f'(x)|dx$$

for $f \in C_c^1(\mathbb{R})$ and apply it to $x \to u(x,y)$ and $y \to u(x,y)$ to derive an estimate for the 2-norm of $u \in C_c^1(\mathbb{R}^2)$ in terms of the 1-norms of u_x and u_y. The result is

$$\iint_{\mathbf{R}^2} |u|^2 \le \frac{1}{4}\iint_{\mathbf{R}^2} |u_x| \iint_{\mathbf{R}^2} |u_y| \quad \text{from which} \quad |u|_2 \le \frac{1}{2}\max_{i=1,2} |u_{x_i}|_1$$

follows (I'm writing single bars with subscript p for the p-norm in L^p).

The same trick with $x \to u(x,y,z)$, $y \to u(x,y,z)$ and $z \to u(x,y,z)$ and Hölder's inequality applied 3 times with exponents $p_1 = p_2 = \frac{1}{2}$ applied to the successive integrations with respect to x, y, z gives

$$\begin{aligned}
\iiint_{\mathbf{R}^3} |u|^{\frac{3}{2}} &\le \iiint_{\mathbf{R}^3} (\frac{1}{2}\int_x |u_x|)^{\frac{1}{2}} (\frac{1}{2}\int_y |u_y|)^{\frac{1}{2}} (\frac{1}{2}\int_z |u_z|)^{\frac{1}{2}} \\
&= (\frac{1}{2})^{\frac{3}{2}} \int_z\int_y\int_x (\int_x |u_x|)^{\frac{1}{2}} (\int_y |u_y|)^{\frac{1}{2}} (\int_z |u_z|)^{\frac{1}{2}} \\
&\le (\frac{1}{2})^{\frac{3}{2}} \int_z\int_y (\int_x |u_x|)^{\frac{1}{2}} (\int_x\int_y |u_y|)^{\frac{1}{2}} (\int_x\int_z |u_z|)^{\frac{1}{2}} \\
&\le (\frac{1}{2})^{\frac{3}{2}} \int_z (\int_y\int_x |u_x|)^{\frac{1}{2}} (\int_x\int_y |u_y|)^{\frac{1}{2}} (\int_y\int_x\int_z |u_z|)^{\frac{1}{2}} \\
&\le (\frac{1}{2})^{\frac{3}{2}} (\int_z\int_y\int_x |u_x|)^{\frac{1}{2}} (\int_z\int_x\int_y |u_y|)^{\frac{1}{2}} (\int_y\int_x\int_z |u_z|)^{\frac{1}{2}}
\end{aligned}$$

(in each integration one of the 3 factors does not depend on the integration variable).

[31] These are the so-called little Hölder spaces, unlike $C^\alpha(\bar{U})$ they are separable.

Exercise 35.60. Prove that

$$|u|_{\frac{3}{2}} \le \frac{1}{2}|u_x|_1^{\frac{1}{3}}|u_y|_1^{\frac{1}{3}}|u_z|_1^{\frac{1}{3}} \le \frac{1}{2}\max_{i=1,2,3}|u_{x_i}|_1$$

generalises to

$$|u|_{\frac{N}{N-1}} \le \frac{1}{2}\max_{i=1,\dots,N}|u_{x_i}|_1$$

for $u \in C_c^1(\mathbb{R}^N)$ via N integrations and Hölder's inequality with $p_1 = \dots = p_N = \frac{1}{N-1}$ applied in every step (in each integration one of the N factors does not depend on the integration variable).

Applied to $u^\gamma = |u|^{\gamma-1}u$ it follows via Hölder's inequality with

$$\frac{1}{p} + \frac{1}{p'} = 1$$

that

$$|u|^\gamma_{\frac{\gamma N}{N-1}} = |u^\gamma|_{\frac{N}{N-1}} \le \frac{1}{2}\max_{i=1,\dots,N}|\gamma u^{\gamma-1}u_{x_i}|_1 \le \frac{\gamma}{2}\max_{i=1,\dots,N}|u^{\gamma-1}|_{p'}|u_{x_i}|_p$$

$$= \frac{\gamma}{2}|u|^{\gamma-1}_{\frac{(\gamma-1)p}{p-1}}\max_{i=1,\dots,N}|u_{x_i}|_p$$

in which γ can be chosen to have equal subscripts of $|u|$ in the first and last expression in this chain.

Exercise 35.61. For $1 \le p < N$ you should check that this gives

$$q = \frac{\gamma N}{N-1} = \frac{(\gamma-1)p}{p-1} = \frac{pN}{N-p},$$

which you may prefer to memorise as

$$\frac{1}{q} = \frac{1}{p} - \frac{1}{N}.$$

What's the value of γ? Dividing by $|u|_q^{\gamma-1}$ on both sides you get

$$|u|_q \le C_{Np}\max_{i=1,\dots,N}|u_{x_i}|_p$$

with an explicit constant C_{Np}. Give this value. Check again that $1 \le p < N$ is the assumption to make here.

Exercise 35.62. In fact we have

$$|u|_q \le C_{Np}|u_{x_1}|_p^{\frac{1}{N}} \cdots |u_{x_N}|_p^{\frac{1}{N}}$$

Prove that this estimate holds for all $u \in W_0^{1,p}(U)$ if $1 \le p < N$ and

$$\frac{1}{q} = \frac{1}{p} - \frac{1}{N}.$$

Exercise 35.63. Show for $N > 2$ that[32]

$$|u|_{\frac{N}{N-2}} \le \frac{1}{4} \max_{i \ne j} |u_{x_i x_j}|_1$$

for $u \in C_c^2(\mathbb{R}^N)$. Only the mixed derivatives are needed[33].

Exercise 35.64. Let $u \in W_0^{1,p}(U)$, U bounded, $1 \le p < N$. Prove that

$$|u|_q \le C_{p,q,N,|U|} |\nabla u|_p$$

if

$$\frac{1}{q} > \frac{1}{p} - \frac{1}{N},$$

with a constant depending only on p, q, N and the measure $|U|$ of the domain. Hint: estimate the q-norm in terms of the 1-norm and the p-norm using Hölder's inequality[34].

Exercise 35.65. A special case in Exercise 35.64 is $q = p$, and the inequality for $p = q = 2$ is called Poincaré's inequality. For $1 \le p < N$ it makes that

$$u \to |\nabla u|_p$$

is an equivalent norm on $W_0^{1,p}(U)$, which was defined as the closure of $C_c^1(U)$ in $W^{1,p}(U)$ with respect to the norm defined by

$$|u|_{1,p}^p = |u|_p^p + |u_{x_1}|_p^p + \cdots + |u_{x_N}|_p^p$$

Show that these norms are also equivalent for $N \le p < \infty$.

[32] There are similar estimates for $u \in C_c^3(\mathbb{R}^N)$, $u \in C_c^4(\mathbb{R}^N)$,....

[33] Nice project: versions similar to Exercise 35.62 for $W_0^{2,p}$ with only mixed derivatives.

[34] Interpolation between 1 and p similar to the interpolation in Exercise 35.57.

36 Evans' Chapter 6 and Navier-Stokes

This chapter is still under (rearranging) construction. The two large exercises in Section 36.4 relate to the Navier-Stokes equations. Problem (**N**) is like the first exercise to Exercise 6.6.4 in Evans. Exercises 3,4,5,6 in his Section 6.6 are variants on the general theme in Section 6.2. Note that 4 requires Theorem 1 in Section 5.8.1 of Evans. Each of these 4 exercises has $B(u,v)$ symmetric and the solution operator S compact and symmetric with respect to both the L^2-inner product, and an inner product defined by B which replaces the inner product with double brackets on V in Section 34.2. The space V depends on the problem. It may be $H_0^1(U)$, $H^1(U)$, or some other Sobolev-Hilbert space. Evans uses H in his formulation of the Lax-Milgram theorem. The space with which it is applied corresponds to V in Remark 34.17.

If S is the inverse of L, the formula's for the eigenvalues of L follow as in Theorem 2 of Section 6.5.1, see Exercise 34.14 and thereafter. Evaluate these eigenvalue formula's for the problems in Exercises 3,4,5,6. NB if numbers are not clickable they refer to Evans.

36.1 Existence of weak solutions via Lax-Milgram

Consider first the equation

$$Lu = -(a_{ij}u_{x_i})_{x_j} + cu = f \quad \text{with boundary condition} \quad u = 0 \tag{36.1}$$

for $u = u(x)$, $x \in U$, U a bounded domain in $\mathbb{R}^N$, ∂U at least continuous, i.e. locally the graph of a continuous function.

Compared to (1) in Section 6.1, I drop the summation signs, use subscripts for the coefficients, and omit the first order terms. Existence of classical solutions, i.e. solutions u with $u \in C^2(U)$ to have equation (36.1) make sense in U, and $u \in C(\bar{U})$ to have the homogeneous (Dirichlet) boundary condition $u = 0$ have a meaning, requires conditions on the coefficients $a_{ij} = a_{ij}(x)$ and $c = c(x)$, and on the right hand side f.

36.1.1 Weak solutions

In the weak solution approach we multiply (36.1) by a $v \in C^1(\bar{U})$, integrate over U and use integration by parts to rewrite the terms with a_{ij} as

$$-\int_U \underbrace{(a_{ij}u_{x_i})_{x_j}}_{w_{x_j}} v = -\int_{\partial U} \nu_j \underbrace{a_{ij}u_{x_i}}_{w} v + \int_U \underbrace{a_{ij}u_{x_i}}_{w} v_{x_j}, \tag{36.2}$$

in which ν_j is the j^{th} component of the outward normal on ∂U. This requires ∂U to be piecewise C^1, the coefficients $a_{ij} \in C^2(\bar{U})$, $c \in C(\bar{U})$, the right hand side $f \in C(\bar{U})$, and $u \in C^2(\bar{U})$. The boundary integral disappears if $v = 0$ on ∂U, leading to the identity

$$\underbrace{\int_U a_{ij} u_{x_i} v_{x_i} + \int_U cuv}_{B[u,v]} = \underbrace{\int_U fv}_{(f,v)} \tag{36.3}$$

for all $v \in C^1(\bar{U})$ with $v = 0$ on ∂U.

This identity that make sense for $u \in C^1(\bar{U})$, with u still required do have $u = 0$ on ∂U. The assumptions of a_{ij}, c, f can of course be weakened now. The weak solution approach works with weak solutions which have their first order derivatives in $L^2(U)$, the natural (Hilbert) space for u, v to live is $H_0^1(U)$ and $u \in H_0^1(U)$ is called a weak solution if (36.3) holds for all $v \in H_0^1(U)$.

This formulation requires $a_{ij}, c \in L^\infty(U)$ only, and $f \in L^2(U)$ then suffices for the right hand side of (36.3) to makes sense because $H_0^1(U) \subset L^2(U)$. Recall that we defined $H_0^1(U)$ as the closure of $C_c^1(U)$ in $H^1(U) = W^{1,2}(U)$. The right hand side of (36.3) is equal to the inner product of f and v in $L^2(U)$ and it defines a linear functional

$$v \in H_0^1(U) \xrightarrow{F} \int_U fv = (f,v)_{L^2(U)} = \underbrace{F(v) = \langle F, v \rangle}_{\substack{\text{different notations} \\ \text{for same functional } F}}, \tag{36.4}$$

the latter being the notation used in the Lax-Milgram Theorem in Section 6.2.1.

In (36.4) it is tempting to write f for F in $\langle F, v \rangle$, as it is really f that acts on v, but *not* via the $H^1(U)$ inner product, as the $H^1(U)$ inner product is defined by the left hand side of (36.3) with $a_{ij} = \delta_{ij}$ and $c = 1$, i.e.

$$(u,v)_{H^1(U)} = \underbrace{\int_U u_{x_i} v_{x_i}}_{\substack{\text{highest order} \\ \text{terms}}} + \int_U uv = \int_U \nabla u \cdot \nabla v + \int_U uv, \tag{36.5}$$

which is the bilinear form corresponding to the partial differential equation $\Delta u + u = f$.

36.1.2 The Lax-Milgram Theorem

The Lax-Milgram Theorem has already been done in Section 17.3. The symmetric case is also discussed in Section 34.2. The f in Theorem 1 in

Section 6.2.1 is really the F in (36.4) if the theorem is applied to the bilinear form in (36.3) with $V = H_0^1(U)$. The H below corresponds to V in Section 34.4.

If the bilinear form is bounded on H, i.e.

$$\forall u, v \in H \quad |B[u,v]| \leq \alpha |u|\,|v|,$$

them for each $u \in H$ the map

$$v \in H \xrightarrow{Au} B[Au, v] = \underbrace{(Au)(v) = \langle Au, v\rangle}_{\substack{\text{different notations} \\ \text{for same functional } Au}} \tag{36.6}$$

is linear and bounded, since

$$|\langle Au, v\rangle| = |B[u,v]| \leq \alpha\,|u|\,|v|,$$

implying that

$$|Au| \leq \alpha\,|u|$$

for all $u \in H$. Thus $A : H \to H^*$, where

$$H^* = \{f : H \to \mathbb{R} :\ f \text{ is linear and bounded}\}$$

normed by

$$|f| = \sup_{0 \neq v \in H} \frac{|\langle f, v\rangle|}{|v|},$$

is linear and bounded. This dual space H^* of H can be identified with H via the Riesz Representation Theorem and $\langle f, v\rangle = f \cdot v = (f, v)_H$, considering $f \in H = H^*$, but in the application to $H = H_0^1(U)$ this is not the inner product in the right hand side to (36.3).

If the bilinear form is also coercive on H, i.e.

$$\forall u \in H \quad B[u,u] \geq \beta\,|u|^2,$$

then

$$\beta\,|u|^2 \leq B[u,u] = \langle Au, u\rangle \leq |Au|\,|u| \quad \text{whence} \quad |Au| \geq \beta\,|u|$$

for all $u \in H$ and it follows that A is a bijection between H and $A(H)$, a subspace of H^*, bounded in both directions. Thus $A(H)$ is complete, and thereby a closed subspace of H^* which[1] coincides with H^*. The (linear) solution operator[2] is then defined by

$$F \in H^* \xrightarrow{S} u \in H \quad \text{defined by} \quad B[u,v] = \langle F, v\rangle \quad \text{for all} \quad v \in H \tag{36.7}$$

[1] Via the Riesz Representation or the Hahn-Banach Theorem and the reflexivity of H.

[2] In Section 34.4 I distinguised between S, S_0, S_1. Check which S this is about!

and has the property that

$$|u|_H = |SF|_H \leq \frac{1}{\beta}\,|F|_{H^*}$$

In the application to boundary value problems any right hand side of (36.1) that defines an F in the dual of the Sobolev space used is allowed. In the case of $H_0^1(U)$ this dual space is denoted by $H^{-1}(U)$ and may be viewed as the space consisting of functions in L^2 as well as their first order distributional derivatives, see Section 5.9.1 in Evans.

36.1.3 Lax-Milgram; boundedness condition

It is usually easy to show that the bilinear form derived from the boundary value problem formulation via integration by parts is bounded, also for other boundary conditions, such as the Neumann condition

$$\nu_j a_{ij} u_{x_i} = 0 \quad \text{on} \quad \partial U, \tag{36.8}$$

which is a special case of the Robin boundary condition

$$\nu_j a_{ij} u_{x_i} + bu = 0 \quad \text{on} \quad \partial U, \tag{36.9}$$

in which $b = b(x)$ is assumed to bounded and integrable for instance. See Exercises 6.6.4 and 6.6.5. The latter condition is called Newton's cooling law in the case that a_{ij} is a positive multiple of the identity matrix and u is the temperature in a body U with heat exchange at the boundary[3]. In (36.2) this condition gives the additional term

$$\int_{\partial U} buv$$

which should now be included in the left hand side of (36.3), and the natural Sobolev space to pose

$$\underbrace{\int_U a_{ij} u_{x_i} v_{x_i}}_{\text{highest order terms}} + \int_U cuv + \int_{\partial U} buv = \underbrace{\int_U fv}_{(f,v)} \tag{36.10}$$

in is now $H^1(U)$.

In the case of the Neumann condition (36.8) this extra term is not there and the only difference with the Dirichlet problem is the choice of the Sobolev

[3]This physical context forces the exchange coefficient to be positive.

space. Boundary conditions which are used in the integration by parts derivation of the weak formulation are sometimes called natural boundary conditions. The Dirichlet boundary condition is not such a natural boundary condition, it has to be forced on the solution by the choice of the smaller Sobolev space $H_0^1(U)$.

36.1.4 Lax-Milgram; coercivity

It is usually more delicate to show the coercivity of the bilinear form. The basic (ellipticity) assumption on the coefficients a_{ij} is (4) in Section 6.1.1 of Evans. With $v = u$ it bounds the highest order terms from below by the highest order terms in (36.5). In the case of $H_0^1(U)$ the Poincaré inequality

$$\int_U u^2 \leq C_{_U} \int_U |\nabla u|^2 \tag{36.11}$$

helps. In particular the bilinear form

$$B[u, v] = \int_U \nabla u \cdot \nabla v$$

used for solving

$$-\Delta u = f \quad \text{with boundary condition} \quad u = 0$$

is coercive on $H_0^1(U)$ considered as a subspace of $H^1(U)$ with the norm derived from (36.5).

The Neumann problem for $-\Delta u = f$ is very instructive. It requires a condition on f for solvability, as well as the same condition on u to have a unique condition, choosing

$$\tilde{H}^1(U) = \{u \in H^1(U) : \int_U u = 0\}$$

as the Sobolev space to be used in the weak formulation.

You should compare the role of b in the Robin boundary condition to that of c in the partial differential equation, as should be clear from (36.10). Coercivity requires some positivity. **It's easy to cook up exam questions on this theme.**

Also, the higher order problem for the bi-Laplacian in Exercise 6.6.3 is only one of the problems of this type. It has two "unnatural" boundary conditions, which are forced upon the solution by the choice of $H_0^2(U)$. Can you think of natural boundary conditions that lead to a formulation in $H^2(U)$, or a mix of natural and unnatural boundary conditions that require $H^2(U) \cap H_0^1(U)$ as the space to be used? Note that for the coercivity of the bilinear form you need the regularity theory in Section 6.3.

36.1.5 The general case with first order terms

The treatment of the Dirichlet problem in Section 6.2.2 should be easy to follow after the discussion above. The main issue is how to deal with the terms in $B[u, v]$ that come from the first order derivatives in the Lu. I did not discuss Section 6.2.3 with the adjoint operator and the Fredholm alternative but read Theorem 4. It is proved via an application of the Fredholm alternative to the solution operator S_μ for the bilinear form

$$B_\mu[u, v] = B[u, v] + \gamma \int_U uv$$

with γ chosen to make B_m coercive.

36.2 The selfadjoint case

See again Chapter 34. The first order terms in L typically prevent the bilinear form from being symmetric. Without these first order terms the symmetry of a_{ij} makes the bilinear form symmetric. This symmetry is usually assumed, see the opening statements in Section 6.5.1. In the case that $B[u, v]$ is a symmetric bounded coercive bilinear form, it defines an equivalent norm on the Sobolev space (used in in the weak formulation) via

$$|u| = \sqrt{B[u, u]}.$$

The solution operator

$$f \xrightarrow{S} u$$

then satisfies both

$$(Sf, g)_{L^2(U)} = (f, Sg)_{L^2(U)} \quad \text{and} \quad B(Su, v) = B(u, Sv)$$

as you should satisfy, and it is compact form $L^2(U)$ to $L^2(U)$ as well as from the Sobolev space to itself. The eigenvalue formula's I discussed for the solution operator using $B[u, v]$ then lead to eigenvalue formula's of which the first is stated in the remark following Theorem 2 in Section 6.5.1.

36.2.1 Second hand in homework set

I would have restricted a second homework set to Exercises like 6.6.3, 6.6.4, 6.6.5, 6.6.6 in the second edition of Evans. In all exercises also write down the eigenvalue formula for the first eigenvalue when f is replaced by λu.

36.2.2 Maximum principles

Evans Section 6.4. More on those principles in Chapters 5 and 10 in

`http://www.few.vu.nl/~jhulshof/NOTES/ellpar.pdf`

36.3 The Navier-Stokes equations

You can read about these equations in Dutch on a very introductionary and informal level in

`http://www.math.vu.nl/~jhulshof/handoutNS.pdf`

Consider the Navier-Stokes equations on a bounded domain $\Omega \subset \mathbb{R}^2$ for $t \geq 0$ with smooth boundary $\partial\Omega$, given initial data for the velocity

$$u = \begin{pmatrix} u_1(t, x_1, x_2) \\ u_2(t, x_1, x_2) \end{pmatrix}$$

at $t = 0$ and no-slip boundary conditions $u = 0$ on $\partial\Omega$ for all $t \geq 0$. For the exercises below you may restrict your attention to the case that

$$\Omega = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 < 1\} \quad \text{with outer normal} \quad n = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{in} \quad x \in \partial\Omega.$$

The Navier-Stokes equations read (with kinematic viscosity equal to unity)

$$u_t + (u \cdot \nabla)u + \nabla p = \Delta u, \quad \nabla \cdot u = 0.$$

The second zero divergence equation has to be imposed on the initial data for u at $t = 0$ as well. In view of the Laplacian in the equation and the boundary condition $u = 0$ on $\partial\Omega$ the natural spaces for solutions to live in as functions of t are

$$H_0^1(\Omega)^2 = \{ \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} : \, u_1, u_2 \in H_0^1(\Omega) \} \subset (L^2(\Omega))^2 = \{ \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} : \, u_1, u_2 \in L^2(\Omega) \},$$

but the zero divergence equation imposes an a priori restriction as explained next.

If $u \in (L^2(\Omega))^2$ satisfies $\nabla \cdot u \in L^2(\Omega)$ then the normal component $n \cdot u$ of the velocity is well defined in $L^2(\partial\Omega)$ by a theorem similar to the trace theorems in Evans, and the Gauss divergence formula

$$\int_\Omega \nabla \cdot u = \int_{\partial\Omega} n \cdot u$$

holds true for such u. Solutions with finite kinetic energy

$$E(u) = \frac{1}{2}\int_\Omega (u_1^2 + u_2^2)$$

actually live in

$$H = \{u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} : \, u_1, u_2 \in L^2(\Omega), \, \nabla \cdot u = 0 \text{ on } \Omega, \, n \cdot u = 0 \text{ on } \partial\Omega\}.$$

If also the first order spatial weak derivatives exist with

$$\mathcal{E}(u) = \int_\Omega |Du|^2 = \int_\Omega \left((\frac{\partial u_1}{\partial x_1})^2 + (\frac{\partial u_1}{\partial x_2})^2 + (\frac{\partial u_2}{\partial x_1})^2 + (\frac{\partial u_2}{\partial x_2})^2 \right) < \infty,$$

then $u \in H^1(\Omega)^2$ and it is possible to speak of u on $\partial\Omega$ as the trace of u and in particular of its tangential component $n \times u = n_1 u_2 - n_2 u_1$ in the usual sense.

36.4 Navier-Stokes related exercises

1. This exercise concerns the projection of

$$L^2_{div}(\Omega) = \{w \in (L^2(\Omega))^2 : \nabla \cdot w \in L^2(\Omega)\}$$

on the space H above (the subscript *div* stands for divergence). For $w \in L^2_{div}(\Omega)$ let $f = -\nabla \cdot w \in L^2(\Omega)$ and $g = n \cdot w \in L^2(\partial\Omega)$, and consider the Neumann problem

$$(\mathbf{N}) \qquad -\Delta p = f \quad \text{in} \quad \Omega \quad \text{with} \quad \frac{\partial p}{\partial n} = g \quad \text{on} \quad \partial\Omega.$$

You may think of p in (**N**) as related to the pressure in the Navier-Stokes equations.

   (a) What is the natural condition on arbitrary $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ to have a solution of (**N**)? Hint: use the divergence theorem, you may argue as if f, g and p are smooth. Does your condition hold for the particular choice of f and g above? If so, why? Explain why then the solution p is never unique and can be chosen to have $\int_\Omega p = 0$.

   (b) Explain why for $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ we say that $p \in H^1(\Omega)$ is a weak solution of (**N**) if

$$(\mathbf{N_{weak}}) \qquad \int_\Omega \nabla p \cdot \nabla \phi = \int_\Omega f\phi + \int_{\partial\Omega} g\phi \quad \text{for all} \quad \phi \in H^1(\Omega).$$

Check that this can only hold if your condition in (a) is satisfied, in which case it suffices to show that that the identity in ($\mathbf{N_{weak}}$) holds for every $\phi \in \tilde{H}^1(\Omega) = \{p \in H^1(\Omega) : \int_\Omega p = 0\}$.

(c) Let $\tilde{H}^1(\Omega)$ be as in (b). Show that

$$((p, \phi)) = \int_\Omega \nabla p \cdot \nabla \phi$$

defines an inner product on $\tilde{H}^1(\Omega)$ with an inner product norm that is equivalent on $\tilde{H}^1(\Omega)$ to the full H^1-norm defined by

$$(p, \phi)_{H^1(\Omega)} = \int_\Omega p\phi + \int_\Omega \nabla p \cdot \nabla \phi, \quad |p|^2_{H^1(\Omega)} = (p, p)_{H^1(\Omega)},$$

(d) Explain why for every $f \in L^2(\Omega)$ and every $g \in L^2(\partial\Omega)$ satisfying your condition in (a) there is a unique $p \in \tilde{H}^1(\Omega)$ that satisfies ($\mathbf{N_{weak}}$).

(e) Recall that

$$H = \{u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} : u_1, u_2 \in L^2(\Omega),\ \nabla \cdot u = 0 \text{ on } \Omega,\ n \cdot u = 0 \text{ on } \partial\Omega\}.$$

Explain why every $w \in L^2_{div}(\Omega)$ can be written as $w = \nabla p + u$ with $u \in H$ and $p \in H^1(\Omega)$, and that u is uniquely determined by w. This u is called the Leray projection of w.

2. In this exercise we consider smooth solutions of the Navier-Stokes equations with zero slip boundary conditions as above (so you can forget about weak derivatives and all that now).

(a) Write u_0 for the initial velocity field of a smooth solution u with pressure p: then $u(x, 0) = u_0(x)$ and u_0 must satisfy $\nabla \cdot u_0 = 0$. We write $u(t)$ for the function $x \to u(x, t)$. Integrate the inner product of

$$u_t + (u \cdot \nabla)u + \nabla p - \Delta u$$

with u over Ω and derive that

$$\frac{d}{dt} E(u(t) + \mathcal{E}(u(t)) = 0,$$

where $E(u)$ and $\mathcal{E}(u)$ are as in the introduction above. In other words, show that

$$\frac{1}{2}\frac{d}{dt}\int_\Omega |u|^2 + \int_\Omega |Du|^2 = 0.$$

Why does it follow that

$$\int_0^T \int_\Omega |Du|^2 \leq \frac{1}{2}\int_\Omega |u_0|^2 \,?$$

Hint: write terms out in coordinates, e.g.

$$u \cdot (u \cdot \nabla)u = \sum_{j,k=1}^{2} u_k u_j \frac{\partial u_k}{\partial x_j},$$

and use integration by parts (the boundary terms disappear, as well as $\nabla \cdot u$).

(b) For smooth solutions u and v with pressures p and q respectively let $w = u - v$. Subtract the equations for u and v, take the inner product with w and integrate over Ω to derive that

$$\frac{1}{2}\frac{d}{dt}\int_\Omega |w|^2 + \int_\Omega |Dw|^2 = -\int_\Omega w \cdot (w \cdot \nabla)v \leq \int_\Omega |Dv|\,|w|^2.$$

Hint: (i) use integration by parts for the equality, the boundary terms disappear, as well as $\nabla \cdot u$, $\nabla \cdot v$, $\nabla \cdot w$, if they show up. Check that the terms coming from the nonlinear terms in the equations may be rewritten as a term giving the integral with w and v, and another integral with u and w which disappears; (ii) for the subsequent inequality use $Ax \cdot x \leq |A|\,|x|^2$ for 2×2 matrices A and 2-vectors x, with $|A|^2 = A_{11}^2 + A_{12}^2 + A_{21}^2 + A_{22}^2$.

(c) Derive from (b) that

$$\frac{d}{dt}\int_\Omega |w|^2 + 2\int_\Omega |Dw|^2 \leq 2\Big(\int_\Omega |Dv|^2\Big)^{\frac{1}{2}}\Big(\int_\Omega |w|^4\Big)^{\frac{1}{2}}.$$

(d) Insert the inequality

$$\int_\Omega |w|^4 \leq \int_\Omega |w|^2 \int_\Omega |Dw|^2$$

in (c) to derive that

$$\frac{d}{dt}\int_\Omega |w|^2 \leq \frac{1}{2}\int_\Omega |Dv|^2 \int_\Omega |w|^2.$$

Hint: in the right hand side you get a product which contains the factor $a = \int_\Omega |Dw|^2$. Use the inequality $2ab \leq a^2 + b^2$ and observe that a^2 als appears on the left hand side.

(e) Derive from (d) and (a) with u replaced by v that

$$\int_\Omega |w(t)|^2 \leq \int_\Omega |w_0|^2 e^{\frac{1}{4}\int_\Omega |v_0|^2}.$$

(f) Prove the inequality used in (d) for compactly supported smooth vectorfields on $\mathbb{R}^2$ by first showing that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x_1,x_2)^4\, dx_1 dx_2 \leq$$

$$(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u^2)\,(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u_{x_1}^2)^{\frac{1}{2}}\,(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u_{x_2}^2)^{\frac{1}{2}}$$

for compactly supported smooth functions u. In short

$$|u|_4^4 \leq |u|_2^2\, |u_{x_1}|_2\, |u_{x_2}|_2.$$

Hint: write $u(x_1,x_2)^4 = u(x_1,x_2)^2 u(x_1,x_2)^2$ and show first that

$$u(x_1,x_2)^2 \leq \int_{-\infty}^{\infty} u(\xi,x_2) u_{x_1}(\xi,x_2) d\xi$$

and likewise

$$u(x_1,x_2)^2 \leq \int_{-\infty}^{\infty} u(x_1,\eta) u_{x_2}(x_1,\eta) d\eta.$$

37 Geostuff

I will use L for the Lagrangian and not F. We assume that $L = L(t, u, p)$ is as smooth as we need. Chapter 1 of [J&J] concerned Euler-Lagrange equations for $u = u(t) \in \mathbb{R}^n$. We saw how minimizing

$$I(u) = \int_a^b L(t, u(t), \dot{u}(t))\, dt \tag{37.1}$$

for sufficiently smooth functions $u : [a, b] \to \mathbb{R}^n$ (with $u(a)$ and $u(b)$ prescribed) leads to the Euler-Lagrange system of differential equations:

$$\frac{d}{dt}\frac{\partial L}{\partial p^i} - \frac{\partial L}{\partial u^i} = 0 \quad (i = 1, \ldots, n) \tag{37.2}$$

We also saw the Jacobi equations, obtained from (1.3.6) and the linearised Lagrangian

$$\phi = \frac{\partial^2 L}{\partial p^i \partial p^j}\pi^i\pi^j + 2\frac{\partial^2 L}{\partial u^i \partial p^j}\pi^i\eta^j + \frac{\partial^2 L}{\partial u^i \partial u^j}\eta^i\eta^j \tag{37.3}$$

The Euler-Lagrange equations of (37.3) are the Jacobi equations

$$\frac{d}{dt}\frac{\partial \phi}{\partial \pi^i} - \frac{\partial \phi}{\partial \eta^i} = 0 \quad (i = 1, \ldots, n) \tag{37.4}$$

These Jacobi equations are the linearised Euler-Lagrange equations. Verify this!

For Lagrangians independent of t we noticed a conservation law. When you multiply (37.2) by $p^i(t) = \dot{u}^i(t)$ you get

$$0 = p^i(t)\frac{d}{dt}\frac{\partial L}{\partial p^i} - \dot{u}^i(t)\frac{\partial L}{\partial u^i} = \frac{d}{dt}\left(p^i\frac{\partial L}{\partial p^i}\right) \underbrace{-\dot{p}^i(t)\frac{\partial L}{\partial p^i} - \dot{u}^i(t)\frac{\partial L}{\partial u^i}}_{-\frac{dL}{dt}}$$

$$= \frac{d}{dt}\left(p^i\frac{\partial L}{\partial p^i} - L\right)$$

37.1 Submanifolds of $\mathbb{R}^d$ are Riemannian

Chapter 2 deals with the problem of finding the shortest connecting curve between two given points in an n-dimensional submanifold M of $\mathbb{R}^d$ with $d > n$. For this will need knowledge of the concept of covariant differentiation

on M. The nonabstract introduction with submanifolds below provides a machinery that also works in the abstract setting of general Riemannian manifolds.

Locally M is given by smooth parameterisations

$$x = f(u)$$

(coordinate charts) defined on open connected sets $U \subset \mathbb{R}^n$ with smooth transitions between u and $\tilde{u}$ on $U \cap \tilde{U}$ if $f : U \to M$ and $\tilde{f} : \tilde{U} \to M$ are two different coordinate patches. A (preferably finite[1]) collection with this property that describes the whole of M is called an atlas for M.

Every such parameterisation provides us with locally defined tangent vector fields

$$x_1 = \frac{\partial x}{\partial u^1}, \cdots, x_n = \frac{\partial x}{\partial u^n},$$

since for every $u \in U$ the vectors $x_i(u)$ are tangent to M in $x(u) \in M$. The inner products

$$g_{ij} = g_{ij}(u) = x_i \cdot x_j$$

are locally defined scalar fields, the coefficients of the Riemannian metric on M inherited from the inner product in the ambient space $\mathbb{R}^d$.

In terms of local coordinates $u^1, \ldots, u^n$ tangent vector fields V on M are described by

$$V = V^i x_i = V^i(u) x_i(u) = V^1(u) x_1(u) + \cdots + V^n(u) x_n(u), \tag{37.5}$$

in which we use a summation convention for repeated lower and upper indices. Two such vectors fields have inner product

$$V \cdot W = V^i x_i \cdot W^j x_j = V^i W^j x_i \cdot x_j = V^i W^j g_{ij}$$

Don't forget the u-dependence which is usually dropped from the notation and pay attention to the double use of subscripts: as indices in g_{ij} and as derivatives in x_i. The inner product of two tangent vector fields on M defines a scalar field[2] on M. The map

$$(V, W) \to V \cdot W$$

is well defined, independent of the choice of coordinates, and multilinear over the scalar fields[3]. In particular, if $\phi, \psi : M \to \mathbb{R}$ are (smooth) functions, then

$$(\phi V) \cdot (\psi W) = \phi\psi \, (V \cdot W)$$

[1] This is related to the concept of compactness

[2] A real valued function

[3] Tensor property

37.2 Covariant differentiation

If we differentiate a vector field V as given by (37.5) we get contributions from u-dependence in $V^i(u)$ and from u-dependence in $x_i(u)$. The tangential part of the resulting derivative is what is by definition the covariant derivative. The partial derivative of (37.5) with respect to u^j can be written as

$$\frac{\partial V}{\partial u^j} = \frac{\partial V^i}{\partial u^j} x_i + V^i x_{ij}, \quad x_{ij} = \frac{\partial x_i}{\partial u^j} = \frac{\partial^2 x}{\partial u^j \partial u^i} = \frac{\partial^2 x}{\partial u^i \partial u^j} = x_{ji} \tag{37.1}$$

In the case that $M = \mathbb{R}^n = \mathbb{R}^d$ with $x^i = u^i$, the tangent vectors x_i are the unit base vectors e_i so that $x_{ij} = 0$ and the covariant partial derivatives of V are just the partial derivatives V. The same holds if $x(u)$ in linear in u. In all other cases we decompose x_{ij} as

$$x_{ij} = \Gamma^l_{ij} x_l + \textit{ normal parts}$$

and take the inner product with x_k to get

$$\Gamma_{ijk} := x_{ij} \cdot x_k = \Gamma^l_{ij} x_l \cdot x_k = \Gamma^l_{ij} g_{lk}$$

Thus Γ_{ijk} is obtained from Γ^l_{ij} using g_{lk}. Introducing $g^{kl} = g^{lk}$ by

$$g_{lk} g^{km} = \delta^m_l,$$

we also obtain Γ^m_{ij} from Γ_{ijk}:

$$g^{mk} \Gamma_{ijk} = \Gamma^l_{ij} g_{lk} g^{km} = \Gamma^l_{ij} \delta^m_l = \Gamma^m_{ij}$$

The relation between both Γ-symbols is given by

$$\Gamma_{ijk} = \Gamma^l_{ij} g_{lk}, \quad \Gamma^m_{ij} = g^{mk} \Gamma_{ijk}$$

The metric coefficients are used to raise and lower the exponents[4].

Next we determine Γ_{ijk}. Differentiating g_{ij} with respect to u^k we get

$$g_{ij,k} = \frac{\partial g_{ij}}{\partial u^k} = \frac{\partial}{\partial u^k}(x_i \cdot x_j) = x_{ki} \cdot x_j + x_{jk} \cdot x_i = \Gamma_{kij} + \Gamma_{jki}$$

Note the two cyclic permutations kij and jki of ijk on the right. Using cyclic permutation, we have the following three equivalent forms of the resulting statement:

$$g_{ij,k} = \Gamma_{kij} + \Gamma_{jki}$$

[4]Just as with tensor coefficients, though the Γ's are not tensor coefficients

$$g_{jk,i} = \Gamma_{ijk} + \Gamma_{kij}$$
$$g_{ki,j} = \Gamma_{jki} + \Gamma_{ijk}$$

Multiplying by $-\frac{1}{2}$, $\frac{1}{2}$ and $\frac{1}{2}$ and adding up we get

$$\Gamma_{ijk} = \frac{1}{2}\left(g_{jk,i} + g_{ki,j} - g_{ij,k}\right)$$

Using the symmetry $g_{ij} = g_{ji}$ it follows that

$$\Gamma_{ijk} = \frac{1}{2}\left(g_{jk,i} + g_{ik,j} - g_{ij,k}\right), \quad \Gamma^m_{ij} = \frac{1}{2}g^{mk}\left(g_{jm,i} + g_{im,j} - g_{ij,m}\right) \tag{37.2}$$

These formula's define the *Christoffel symbols* $\Gamma^k_{ij} = \Gamma^k_{ji}$ in terms of the metric and its first order derivatives and can be used to write (37.1) as

$$\frac{\partial V}{\partial u^j} = \frac{\partial V^i}{\partial u^j}x_i + V^i\Gamma^l_{ij}x_l + \textit{normal parts}$$

The tangential part is thus

$$D_{u^j}V := (\frac{\partial V}{\partial u^j})_T = \left(\frac{\partial V^l}{\partial u^j} + V^i\Gamma^l_{ij}\right)x_l, \quad V = V^i x_i \tag{37.3}$$

This is called the covariant derivative of V with respect to u^j. Both V and $D_{u^j}V$ are tangent vector fields, with components

$$V^i \quad \text{and} \quad (D_{u^j}V)^l = \frac{\partial V^l}{\partial u^j} + V^i\Gamma^l_{ij}$$

37.3 Tangent vectors as derivatives

Next we introduce the modern view point on tangent vectors. Since every tangent vector defines a directional derivative, it has become customary to identify such first order differential operators with their direction vectors. In short, we think of

$$x_i = \frac{\partial x}{\partial u^i} \quad \text{and} \quad \frac{\partial}{\partial u^i}$$

as essentially the same objects. To see how this works in a point $x_0 \in M$ we use integral curves starting at x_0, that is, solutions of

$$\dot\gamma(t) = X(\gamma(t)), \quad \gamma(0) = x_0 \in M, \tag{37.1}$$

where X is a tangent vector field defined near x_0. The differential equation in (37.1) is called the *flow equation* for X. Using coordinates u, with $u = u_0$ corresponding to x_0, the expressions in (37.1) evaluate as

$$\gamma(t) = x(u(t)), \quad \dot\gamma(t) = \frac{\partial x}{\partial u^i}(u(t))\dot u^i(t) = \dot u^i(t)x_i, \quad X(\gamma(t)) = X^i(u(t))x_i,$$

so the system to be solved for $u = u(t)$ to obtain the integral curves is

$$\dot u^i = X^i(u), \quad u(0) = u_0. \tag{37.2}$$

The solution $u = u(t)$ exists locally and is unique. We have $\dot u^i(0) = X^i(u_0)$ and $X_0 := X(x_0) = \dot\gamma(0) = \dot u^i(0)x_i = X^i(u_0)x_i$. On scalar fields (functions) $\phi : M \to \mathbb{R}$, given in local coordinates as

$$\phi = \phi(u^1, \dots, u^n),$$

the vector field X now acts through

$$\frac{d}{dt}|_{t=0}\phi(u(t)) = \frac{\partial \phi}{\partial u^i}(u_0)\dot u^i(0) = X_0^i\frac{\partial \phi}{\partial u^i}(u_0)$$

at ϕ in $u = u_0$, i.e. as the directional derivative

$$X_0^i\frac{\partial}{\partial u^i} \quad \text{corresponding to the direction vector} \quad X_0^i x_i$$

in $u = u_0$. The derivative only depends on the value of the vector field in x_0. Since the point $x_0 = x(u_0)$ was arbitrary we have

$$X = X^i\frac{\partial}{\partial u^i} \quad \text{corresponding to the tangent field} \quad X = X^i x_i = X^i\frac{\partial x}{\partial u^i}.$$

The two expressions above are merely different representations of the tangent vector field X (both in local coordinates):

The components

$$X^i\frac{\partial x^k}{\partial u^i}$$

of the *tangent field* X multiply

$$\frac{\partial \phi}{\partial x^k}$$

in the chain rule formula if ϕ is extended to a neighbourhood of M in $\mathbb{R}^d$. As *differential operator*

$$X = X^i\frac{\partial}{\partial u^i}$$

X acts on scalar fields like $\phi = \phi(u)$ and produces a scalar field $X\phi$, the derivative of ϕ in the direction of X. This directional derivative is denoted by

$$\nabla_X \phi = X\phi, \quad \text{replacing the notation} \quad \frac{\partial \phi}{\partial X}$$

in calculus texts. We already use the notation ∇_X customary for covariant differentiation. For reasons that should be clear, covariant differentiation of scalar fields is by definition the same as differentiation of scalar fields.

37.4 Commutators of tangent vector fields

If X and Y are scalar fields on M then the commutator of X and Y is defined as

$$[X, Y] = XY - YX$$

meaning that

$$\nabla_{[X,Y]}\phi = [X, Y]\phi = X(Y\phi) - Y(X\phi) = \nabla_X(\nabla_Y \phi) - \nabla_Y(\nabla_X \phi).$$

This commutator has a meaning by itself. If $\gamma(t)$ is the solution of (37.1), then the linearised flow equation transports the vector $Y(x_0)$ along $\gamma(t)$. Denoting the transported vector as $\xi(t)$, we may differentiate the difference of $\xi(t)$ and $Y(\gamma(t))$ with respect to t and evaluate the derivative in $t = 0$. This defines

$$(\mathcal{L}_X Y)(x_0) = \lim_{t \to 0} \frac{\xi(t) - Y(\gamma(t))}{t},$$

the Lie derivative of Y with respect to X in x_0.

In coordinates $\xi(t) = \xi^i(t) x_i$ with $\xi^i(t)$ is a solution of the linearizsation of (37.2) around $u(t)$,

$$\dot{\xi}^i = \underbrace{(\frac{\partial X^i}{\partial u^j})}_{\text{in } (u(t)} \xi^j(t), \quad \xi^j(0) = Y^i(u_0) \tag{37.1}$$

Writing

$$\xi(t) - Y(\gamma(t)) = \xi(t) - Y(x_0) - (Y(\gamma(t)) - Y(x_0))$$

you should verify that

$$(\mathcal{L}_X Y)(x_0) = (XY)(x_0) - (YX)(x_0)$$

so that

$$[X, Y] = \mathcal{L}_X Y \tag{37.2}$$

Note that $[X, Y]$ is bilinear over de scalar fields. Verify that

$$[X, Y]^j = X^k Y^j_k - Y^k X^j_k$$

and that the Jacobi identity

$$[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0 \tag{37.3}$$

holds.

37.5 Covariant differentiation of tangent vectors

Next we observe that

$$X = X^i \frac{\partial}{\partial u^i}$$

naturally acts covariantly on tangent fields V, if we replace

$$\frac{\partial}{\partial u^i} \quad \text{by} \quad D_{u^j},$$

as defined in (37.3) through

$$D_{u^j} V := \left(\frac{\partial V^l}{\partial u^j} + V^i \Gamma^l_{ij} \right) x_l \quad \text{for} \quad V = V^i x_i.$$

The result of this action is

$$X^j \left(\frac{\partial V^l}{\partial u^j} + V^i \Gamma^l_{ij} \right) x_l$$

and is denoted as

$$\nabla_X V = X^j \left(\frac{\partial V^l}{\partial u^j} + V^i \Gamma^l_{ij} \right) \frac{\partial}{\partial u^i} \tag{37.1}$$

in the modern notation for tangent vectors as differential operators.

The map

$$V \to \nabla_X V$$

is *not* linear over the scalar fields because

$$\nabla_X \phi V = X^j \left(\frac{\partial \phi V^l}{\partial u^j} + \phi V^i \Gamma^l_{ij} \right) x_l$$

$$= \phi X^j \left(\frac{\partial V^l}{\partial u^j} + V^i \Gamma^l_{ij} \right) x_l + X^j \frac{\partial \phi}{\partial u^j} V^l = \phi \nabla_X V + (\nabla_X \phi) V.$$

The latter term in this Leibniz rule destroys the tensor property of linearity over the scalar fields.

Convince yourself that in the non-abstract approach

$$\nabla_X V = X^j \left(\frac{\partial V^l}{\partial u^j} + V^i \Gamma^l_{ij} \right) x_l$$

is the tangential[5] *component of the derivative of* V *in the direction of* X and verify that

$$\nabla_X (V \cdot W) = \nabla_X V \cdot W + V \cdot \nabla_X W$$

if W is another tangent vector field on M.

37.6 Second fundamental form

The normal part of the derivative of V in the direction of X is denoted by $\mathrm{I\!I}(X, V)$, in which $\mathrm{I\!I}$ is called the *second fundamental form* of M. Verify that it is bilinear over the smooth fields on M. Since the normal part essentially comes from the mixed derivatives x_{ij}, the *second fundamental form must be symmetric*. Moreover, if N is a normal vector field on M and N, X, V are extended smoothly[6] to the ambient space $\mathbb{R}^d$ then

$$\bar{\nabla}_X (N \cdot Y) = \bar{\nabla}_X N \cdot Y + N \cdot \bar{\nabla}_X Y, \tag{37.1}$$

in which $\bar{\nabla}$ is the (standard covariant) derivative in $\mathbb{R}^d$. On M the left hand side of (37.1) is zero, and the second term $N \cdot \bar{\nabla}_X Y$ on the right hand side only sees the normal part of $\bar{\nabla}_X Y$ which is $\mathrm{I\!I}(X, Y)$. It follows that

$$\bar{\nabla}_X N \cdot Y = -N \cdot \mathrm{I\!I}(X, Y) \quad \text{on } M. \tag{37.2}$$

This is called Weingarten's relation. Note that in the codimension 1 case $d = n + 1$ we can choose a unit normal field N and define

$$h(X, Y) = N \cdot \mathrm{I\!I}(X, Y) = -\bar{\nabla}_X N \cdot Y = h_{ij} X^i Y^j \tag{37.3}$$

37.7 Curvature

The equality

$$\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z = \nabla_{[X,Y]} Z + R(X, Y) Z \tag{37.1}$$

[5] to M

[6] This can be done, certainly locally, why?

defines $R(X,Y)Z$ for tangent vector fields X, Y, Z. You may verify that $R(X,Y)Z$ is multilinear in X, Y, Z over the scalar fields on M. In the case $M = \mathbb{R}^n = \mathbb{R}^d$ you will find that $R(X,Y)Z \equiv 0$. The standard way to write $R(X,Y)Z$ in local coordinates u is

$$(R(X,Y)Z)^\alpha = R^\alpha_{ijk} Z^i X^j Y^k. \tag{37.2}$$

So Z comes first[7] and then X and Y. Using (37.1) and writing

$$\Gamma^\alpha_{ij,k} = \frac{\partial \Gamma_{ij}}{\partial u^k}$$

you should verify that[8]

$$R^\alpha_{ijk} = \Gamma^\beta_{ik}\Gamma^\alpha_{\beta j} - \Gamma^\beta_{ij}\Gamma^\alpha_{\beta k} + \Gamma^\alpha_{ik,j} - \Gamma^\alpha_{ij,k} \tag{37.3}$$

and the zero ijk and jk cyclic sums

$$R^\alpha_{ijk} + R^\alpha_{kij} + R^\alpha_{jki} = 0 = R^\alpha_{ijk} + R^\alpha_{ikj} \tag{37.4}$$

If W is another tangent field then[9]

$$Rm(X,Y,Z,W) = R(X,Y)Z \cdot W = R^\alpha_{ijk} Z^i X^j Y^k g_{\alpha l} W^l = R_{lijk} W^l Z^i X^j Y^k, \tag{37.5}$$

which has the symmetries

$$Rm(X,Y,Z,W) + Rm(Y,Z,X,W) + Rm(Z,X,Y,W) = 0,$$

$$Rm(X,Y,Z,W) + Rm(Y,X,Z,W) = 0 = Rm(X,Y,Z,W) + Rm(X,Y,W,Z)$$

(the second one obtained from $Rm(X,Y,Z,Z) = 0$), implying

$$Rm(X,Y,Z,W) = Rm(Z,W,X,Z)$$

In the 2-dimensional case $n = 2$ the only possible nonzero entries of R_{ijk} are

$$R_{1212} = R_{2121} = -R_{1221} = -R_{2112}$$

In the codimension 1 case

$$R_{lijk} = h_{ik}h_{lj} - h_{ij}h_{lk}$$

[7] As if we would have prefered the notation $ZR(X,Y)$

[8] note the order ijk in the minus terms and the $j \leftrightarrow k$ relation with the plus terms

[9] $lijk$ = dead body, as if we would have prefered the notation $W \cdot ZR(X,Y)$

consists of all the 2×2 determinants you can get from the matrix h_{ij}. Note that similarly

$$(W \cdot X)(Y \cdot Z) - (W \cdot Y)(X \cdot Z) = \underbrace{(g_{ik}g_{lj} - g_{ij}g_{lk})}_{G_{lijk}} W^l Z^i X^j Y^k, \qquad (37.6)$$

in which G_{lijk} has the same symmetry properties as R_{lijk} (and depends only on G_{1212} if $n = 2$).

For submanifolds you can verify from the definitions that

$$Rm(X, Y, Z, W) = \mathrm{I\!I}(X, W)\mathrm{I\!I}(Y, Z) - \mathrm{I\!I}(X, Z)\mathrm{I\!I}(Y, W), \qquad (37.7)$$

which in the codimension 1 case (37.3) reduces to

$$Rm(X, Y, Z, W) = h(W, X)h(Y, Z) - h(W, Y)h(X, Z),$$

so

$$Rm(X, Y, Z, W) = \underbrace{(h_{ik}h_{lj} - h_{ij}h_{lk})}_{R_{lijk}} W^l Z^i X^j Y^k, \qquad (37.8)$$

Gauss computed this expression for R_{lijk} from $x_{ijk} = x_{ikj}$, see Chapter 10 in Schaum's Differential Geometry book by Martin Lipschutz. The Gauss curvature of a surface in $\mathbb{R}^d$ is the scalar ratio between (37.7) and (37.6). In $\mathbb{R}^3$ this is the scalar ratio between (37.8) and (37.6).

37.8 Geodesic curves

A smooth curve $\gamma(t) \in M$ may require several coordinate patches to describe it. For the moment we assume that it can be described by one coordinate patch. If

$$\gamma : [a, b] \ni t \to u(t) \to x(u(t)) \in M$$

is such a curve in M, then its velocity is given by

$$\dot{\gamma} = \frac{\partial x}{\partial u^1}\dot{u}^1 + \cdots + \frac{\partial x}{\partial u^n}\dot{u}^n = \sum_{i=1}^{n} \dot{u}^i \frac{\partial x}{\partial u^i} = \sum_{i=1}^{n} \dot{u}^i x_i.$$

Think of $\dot{\gamma}$ as a vector at the point $x = \gamma(t)$ in M. For every t this vector is tangent to M, and written as a linear combination of the tangent vectors obtained from the parameterisation:

$$x_1 = \frac{\partial x}{\partial u^1}, \ldots, x_n = \frac{\partial x}{\partial u^n}.$$

Its length l is given by

$$l = \int_a^b |\dot\gamma(t)|\,dt = \int_a^b \sqrt{\dot\gamma(t)\cdot\dot\gamma(t)}\,dt = \int_a^b \sqrt{x_i\dot u^i\cdot x_j\dot u^j}\,dt$$

$$= \int_a^b \sqrt{\dot u^i\dot u^j g_{ij}(u)}\,dt$$

We will work with another quantity, called the energy, which involves an L as in Chapter 1. Since I prefer to have u in L, my u's are the γ's in the book. My $\gamma(t)$ is what is $c(t)$ in the book. The energy is defined by

$$E = \frac12\int_a^b |\dot\gamma(t)|^2\,dt = \frac12\int_a^b \dot\gamma(t)\cdot\dot\gamma(t)\,dt = \frac12\int_a^b x_i\dot u^i\cdot x_j\dot u^j\,dt$$

$$= \frac12\int_a^b \dot u^i\dot u^j g_{ij}(u)\,dt = \int_a^b L(u(t),\dot u(t))\,dt,$$

in which

$$L = L(u,p) = \frac12 p^i p^j g_{ij}(u). \tag{37.9}$$

Playing with the estimate

$$\int_a^b |\dot\gamma(t)|\,dt = \int_a^b 1\,|\dot\gamma(t)|\,dt \le \sqrt{\int_a^b 1^2\,dt}\sqrt{\int_a^b |\dot\gamma(t)|^2\,dt}$$

and reparameterisation of γ to make $|\dot\gamma|$ constant you should easily conclude that minimizers of l are minimizers of E and vice versa if we keep $[a,b]$ fixed.

The Euler-Lagrange equations for E involve the derivatives of g_{ij} and come out as

$$\ddot u^i + \Gamma^i_{\alpha\beta}\dot u^\alpha\dot u^\beta = 0 \tag{37.10}$$

and are called the geodesic equations. Indeed,

$$\Gamma^i_{\alpha\beta} = \frac12 g^{ik}\left(g_{\alpha k,\beta} + g_{\beta k,\alpha} - g_{\alpha\beta,k}\right),$$

the symbols computed in (37.2). You should repeat this calculation without looking at the notes above. What is the conservation law for this system?

A nice example is a surface M which is described by a single set of coordinates $u\in\mathbb{R}^2$ with a metric

$$g_{ij}(u) = g(|u|)\delta_{ij} \tag{37.11}$$

in which $u \to g(|u|)$ is smooth and positive[10]. You can write the geodesic equations as in the book (2.1.27). In a special case the example is related to stereographic projection through

$$u^1 = \frac{x^1}{1-x^3}, \quad u^2 = \frac{x^2}{1-x^3},$$

which you may prefer as

$$u = \frac{x}{1-z}, \quad v = \frac{y}{1-z}$$

without indices.

- Verify that large circles on $x^2 + y^2 + z^2$ correspond to circles in the uv-plane. Hint: describe the large circles as $z = ax + by$ and avoid goniometric functions.
- The large circles not contained in this description are the vertical great circles which correspond to lines through the origin in the uv-plane. Assuming unit speed for both the vertical great circles and lines through the origin derive the formula for $g(|u|)$.

We return to (37.11) with general $g(|u|)$.

- Why are geodesics through the origin straight lines?
- Take a geodesic line parametrized by t such that $t = 0$ corresponds to $(0, 0)$ and that the speed in $(0, 0)$ is equal to 1. Use the conservation law to derive a first order equation for $R(t) = |u(t)|$ and solve it.
- Examine how long it takes for the geodesic curve to reach infinity. What is the condition on $g(|u|)$ to reach infinity in finite time? This should involve some integral with g. Do the same in dimension $n > 2$? Is there a difference?
- Can you cook up an example for which the geodesic cannot cross $|u| = 1$? Can you classify these examples?
- Incidentally, what is the Gauss curvature for metrics of the form (37.11) in $\mathbb{R}^2$?

[10] implying $0 = g'(0) = g'''(0) = g''''''(0) = \cdots$

37.9 The Jacobi equations

Consider the Lagrangian (37.9).

- Show that the Jacobi equations (37.4) for (37.9) are

$$\ddot{\eta}^i + 2\Gamma^i_{jk}\dot{u}^j\dot{\eta}^k + \Gamma^i_{jk,l}\dot{u}^j\dot{u}^k\eta^l = 0 \tag{37.12}$$

Both $\dot{u}^i(t)$ and $\eta^i(t)$ define vector fields along $\gamma(t) = x(u(t))$ in $M \in \mathbb{R}^d$ tangent to M through

$$\dot{\gamma}(t) = \dot{u}^i(t)x_i(u(t)), \quad \eta(t) = \eta^i(t)x_i(u(t))$$

The Jacobi equations are much more transparent if we work with the tangential parts D_tV of the time derivatives of such vector fields

$$V(\gamma(t)) = V^i(t)x_i(u(t))$$

- Derive that

$$D_tV = (D_tV)^j x_j \quad \text{with} \quad \dot{V}^j + V^\alpha\Gamma^j_{\alpha\beta}\dot{u}^\beta$$

- Derive that the geodesic equation (37.10) may be written as

$$D_t\dot{\gamma} = 0, \quad \dot{\gamma} = \dot{u}^i x_i$$

- Derive that (37.12) may be written as

$$(D_t^2\eta)^i + \dot{u}^\alpha R^i_{\alpha\beta k}\eta^\beta\dot{u}^k = 0, \quad \text{i.e.} \quad D_t^2\eta + R(\eta,\dot{\gamma})\dot{\gamma} = 0$$

38 Newton's method the hard way

Some time ago I was asked to give a talk on the work of Nash. I apologise for doing something else instead. On a family of theorems that bear his name and proofs Nash never wrote. In these notes I describe how Newton's method can be adapted in the case that the map

$$u \to u - f'(u)^{-1} f(u) \tag{38.1}$$

is not defined as a map from a Banach space X to itself. The resulting theorems are called HARD Implicit Function Theorems. My purpose here is to demystify the terminology and present a simple proof of convergence for a modification of Newton's method in such a case. Observe that a direct proof of the Inverse Function Theorem for a continuously differentiable function f amounts to solving the equation $f(u) = v$ for u given small v under the assumption that $f(0) = 0$, using the map

$$u \to u + f'(0)^{-1}(v - f(u)) \tag{38.2}$$

which is contractive if $f'(0)^{-1} : X \to X$ exists as a continuous linear map.

The proof of the Implicit Function Theorem for solving equations like $f(u, v) = 0$ in the form $u = u(v)$ if $f(0, 0) = 0$ and the partial derivative of f with respect to u is invertible in $(u, v) = (0, 0)$ is similar. To show that (38.2) produces a local solution $u = u(v)$ which is continuously differentiable the only regularity on f that has to be assumed is that $u \to f'(u)$ is continuous, as only $f'(u)$ is needed in the calcutions and estimates. Newton's method, which employs a suitable inverse of $f'(u)$ for all u in some (say the unit) ball B in X, relies an Taylor's theorem with a quadratic remainder and therefore the assumption that also $u \to f''(u)$ be continuous is required.

38.1 Newton's method: a convergence proof

I will modify the treatment in [KP][1] which begins with a somewhat alternative treatment of Newton's method in the standard case. So to warm up consider an equation of the form $f(u) = 0$ in which $f : B \to X$ is a twice continuously differentiable function defined on the open unit ball B in a Banach space X, with first and second order derivative satisfying bounds

$$|f'(u)| \leq M_1 \quad \text{and} \quad |f''(u)| \leq M_2 \quad \forall u \in B. \tag{38.3}$$

The general case of Banach spaces is really not that different from the case in which $X = \mathbb{R}$, which you may think of in what follows below. Simply take $B = (-1, 1)$ and replace all norms by absolute values.

[1] Krantz & Parks, The Implicit Function Theorem, Birkhäuser 2003.

What we need is that Taylor's theorem with a second order remainder,

$$f(u_n) = \underbrace{f(u_{n-1}) + f'(u_{n-1})(u_n - u_{n-1})}_{\text{linear aproximation}} + Q_f(u_{n-1}, u_n), \tag{38.4}$$

in which

$$|Q_f(u_{n-1}, u_n)| \leq \frac{M_2}{2}|u_n - u_{n-1}|^2, \tag{38.5}$$

applies to a sequence of iterates $u_n \in B$. For the standard Newton method one does not explicitly need the bound on $f'(u)$ in (38.3) which says that the linear map $f'(u) : X \to X$ satisfies

$$|f'(u)v| \leq M_1|v| \quad \forall u \in B \quad \forall u \in X, \tag{38.6}$$

but a similar bound

$$|L(u)| \leq C \tag{38.7}$$

for maps $L(u)$, that act as right inverses of $f'(u)$ in the sense that

$$f'(u_{n-1})L(u_{n-1})f(u_{n-1}) = f(u_{n-1}), \tag{38.8}$$

is essential. Writing

$$p_n = |u_n - u_{n-1}| \text{ and } q_n = |f(u_n)| \tag{38.9}$$

the Newton scheme

$$u_n = u_{n-1} - L(u_{n-1})f(u_{n-1}) \quad (n \in \mathbb{N}), \tag{38.10}$$

starting with $u_0 = 0$, then defines $u_n \in B$ as long as

$$p_1 + p_2 + \cdots + p_n < 1, \tag{38.11}$$

and the inequalities

$$p_n \leq Cq_{n-1} \text{ and } q_n \leq \frac{1}{2}M_2 p_n^2 \tag{38.12}$$

are immediate from (38.4,38.5,38.10). Note that (38.10) kills the linear approximation in (38.4). The inequalities in (38.12) are complemented by

$$q_0 = |f(0)| \quad \text{and} \quad p_1 \leq Cq_0 = C|f(0)|. \tag{38.13}$$

38.2 The optimal result

Clearly (38.12) and (38.13) combine as

$$p_n \le \mu p_n^2 \quad \text{with} \quad \mu = \frac{1}{2}MC \quad \text{and} \quad p_1 \le C|f(0)|, \tag{38.14}$$

and the condition to be stated is which $\bar{P} = \bar{P}(\mu)$ guarantees that the implication

$$C|f(0)| < \bar{P} \implies \sum_{n=1}^{\infty} p_n < 1 \tag{38.15}$$

holds. The larger $\bar{P}$ the stronger the statement in the sense that larger $|f(0)|$ are allowed to obtain a solution $u = \bar{u} \in B$ of $f(u) = 0$ via (38.10) with $u_0 = 0$. Note that with $C|f(0)| \le \bar{P}$ the same conclusion will hold if only one of all the inequalites in the estimates below is strict, which will inevatibly be the case of course.

Obviously the smallest $\bar{P}$ we can get follows from replacing the three inequalities in (38.14) and (38.15) by inequalities. This leads to

$$p_n = \mu p_{n-1}^2 \ \text{ for } \ n \in \mathbb{N}; \quad p_1 = \bar{P}\,; \quad \sum_{n=1}^{\infty} p_n = 1. \tag{38.16}$$

Via $\xi_n = \mu p_n$ and $\xi_n = \xi_{n-1}^2$ this is easily seen to be equivalent to

$$\mu = G(\mu\bar{P}) \quad \text{with} \quad G(\xi) = \xi + \xi^2 + \xi^4 + \xi^8 + \xi^{16} + \cdots \tag{38.17}$$

but this does not yield a simple formula for $\bar{P} = \bar{P}(\mu)$.

38.3 A suboptimal result

A rough estimate

$$G(\xi) < \xi + \xi^2 + \xi^3 + \xi^4 + \xi^5 + \cdots = \frac{\xi}{1-\xi} \tag{38.18}$$

leads to a simple but suboptimal formula:

$$\bar{P} = \frac{1}{1+\mu} \quad \text{or} \quad \mu = \frac{1}{\bar{P}} - 1. \tag{38.19}$$

38.4 Alternative proof of convergence

The alternative approach to (38.12) and (38.13) in [KP] is not to solve the corresponding system with equalities but to derive an estimate of the form

$$p_n \le e^{-\gamma\lambda^n} \tag{38.20}$$

via induction starting from

$$p_1 \le C|f(0)| < \bar{P} = e^{-\gamma\lambda}, \tag{38.21}$$

with choices of γ and λ that guarantee both

$$\sum_{n=1}^{\infty} e^{-\gamma\lambda^n} \le 1 \tag{38.22}$$

as well as that the induction step can be done via

$$p_{n-1} \le e^{-\gamma\lambda^{n-1}} \implies p_n \le \mu p_{n-1}^2 \le \underbrace{\mu e^{-2\gamma\lambda^{n-1}} \le e^{-\gamma\lambda^n}}_{\text{should hold for all } n\ge 1},$$

which is the case if

$$\ln\mu \le \gamma\lambda^{n-1}(2-\lambda) \quad \forall n \ge 1.$$

38.5 The optimal alternative result

For a given μ this is equivalent to

$$\ln\mu \le \gamma\lambda(2-\lambda) \quad \text{and} \quad \lambda \le 2 \tag{38.23}$$

if we make the obvious restriction that γ and λ be positive. Conditions (38.21) and (38.23) suggest $\alpha = \gamma\lambda$ and λ as the more relevant parameter so we have to pick $\alpha > 0$ and $1 < \lambda \le 2$ with

$$\ln\mu \le \alpha(2-\lambda), \quad \sum_{n=0}^{\infty} e^{-\alpha\lambda^n} \le 1 \quad \text{and} \quad \bar{P} = e^{-\alpha} \quad \text{maximal.} \tag{38.24}$$

For $\mu > 1$ the inequalities define a set in the first quadrant of the λ, α-plane bounded by the two curves given by

$$\ln\mu = \alpha(2-\lambda) \ \text{ and } \ \sum_{n=0}^{\infty} e^{-\alpha\lambda^n} = 1, \tag{38.25}$$

which intersect in one point.

This point defines the minimal value of $\alpha = -\ln \bar{P}$ via

$$1 = \sum_{n=0}^{\infty} e^{-\alpha\lambda^n} = \sum_{n=0}^{\infty} \bar{P}^{\lambda^n} = \sum_{n=0}^{\infty} \bar{P}^{(2+\frac{\ln\mu}{\ln\bar{P}})^n}$$

if $\mu > 1$. The curve defined by

$$1 = \sum_{n=0}^{\infty} \bar{P}^{(2+\frac{\ln\mu}{\ln\bar{P}})^n} \quad \text{and} \quad \mu \geq 1 \tag{38.26}$$

hits the curve defined by (38.17) in $\mu = 1$ and lies below (38.17) of course, but above (38.19) in view of

$$\mu = \frac{1}{\bar{P}} - 1 \implies \sum_{n=0}^{\infty} \bar{P}^{(2+\frac{\ln\mu}{\ln\bar{P}})^n} = \sum_{n=0}^{\infty} \bar{P}^{(1+\frac{\ln(1-\bar{P})}{\ln\bar{P}})^n} < \underbrace{\sum_{n=0}^{\infty} \bar{P}^{1+n\frac{\ln(1-\bar{P})}{\ln\bar{P}}}}_{\text{a geometric series}} = 1.$$

For $\mu \leq 0$ the optimal choice of $\bar{P}$ via (38.24) is given by

$$\sum_{n=0}^{\infty} \bar{P}^{2^n}.$$

38.6 A suboptimal alternative result

A more explicit formula is again obtained via a rough estimate

$$\sum_{n=1}^{\infty} e^{-\gamma\lambda^n} \leq \underbrace{\sum_{n=1}^{\infty} e^{-\gamma(1+n(\lambda-1))}}_{\text{a geometric series}} = \frac{e^{-\gamma\lambda}}{1-e^{-\gamma(\lambda-1)}} = \frac{e^{-\alpha}}{1-e^{\gamma}e^{-\alpha}} \tag{38.27}$$

and replacing (38.24) by

$$\ln\mu \leq \alpha(2-\lambda), \quad \lambda \geq \frac{\alpha}{\ln(e^{\alpha}-1)} \quad \text{and} \quad \bar{P} = e^{-\alpha} \quad \text{maximal.}$$

This leads to

$$\mu = e^{\alpha(2-\lambda)} = \frac{1}{\bar{P}^{2-\lambda}} = \frac{1}{\bar{P}^{2+\frac{\ln\bar{P}}{\ln(\frac{1}{\bar{P}}-1)}}} = \bar{P}^{\frac{\ln(\bar{P})-2\ln(1-\bar{P})}{\ln(1-\bar{P})-\ln(\bar{P})}}$$

so that

$$1 \leq \mu = \frac{1}{\bar{P}^{2+\frac{\ln\bar{P}}{\ln(\frac{1}{\bar{P}}-1)}}} < \frac{1}{\bar{P}} - 1 \tag{38.28}$$

defines another curve with

$$\bar{P} \leq \frac{3-\sqrt{5}}{2},$$

which is below the three curves above, but to leading coincides with them in the limit $\mu \to \infty$ and $\bar{P} \to 0$.

38.7 A lousy alternative result

The even rougher estimate used in [KP] via

$$\sum_{n=1}^{\infty} e^{-\gamma\lambda^n} \leq \sum_{n=1}^{\infty} e^{-n\gamma(\lambda-1)}$$

is to be avoided as at some point below the treatment of ill-behaved Newton's methods will show.

38.8 A much better suboptimal alternative result

Actually the first rough estimate above works better with α dan with γ, as I only noticed May 21. Directly in terms of γ and λ we have

$$\sum_{n=1}^{\infty} e^{-\gamma\lambda^n} = \sum_{n=1}^{\infty} e^{-\gamma\lambda\lambda^{n-1}} \leq \sum_{n=1}^{\infty} e^{-\gamma\lambda(1+(n-1)(\lambda-1))} = \frac{e^{-\gamma\lambda}}{1-e^{-\gamma\lambda(\lambda-1)}} \leq 1 \tag{38.29}$$

if

$$2-\lambda \leq \frac{\ln(e^{\gamma\lambda}-1)}{\gamma\lambda} = \frac{\ln(e^{\alpha}-1)}{\alpha},$$

so that we arrive at

$$\ln\mu \leq \alpha(2-\lambda), \quad \alpha(2-\lambda) \leq \ln(e^{\alpha}-1) \quad \text{and} \quad \bar{P} = e^{-\alpha} \quad \text{maximal.} \tag{38.30}$$

This is the optimal estimate using the Bernoulli type inequality

$$\lambda^n \geq 1 + n(\lambda-1). \tag{38.31}$$

With equality in the final inequality in (38.29) we arrive at

$$\ln\mu \leq \ln(e^{\alpha}-1) = \ln(\frac{1}{\bar{P}}-1),$$

which for $\mu > 1$ coincides with (38.19) and we can forget about the annoying (38.28) above. Note that factoring out another λ in the exponent in (38.29)

will and cannot help to improve this result, which says that if $\mu > 1$ the bound

$$|f(0)| \leq \frac{1}{C(\mu+1)}$$

suffices.

This bound may be compared with the bound in [KP], where all constants are named M, for unclear reasons $M > 2$ is assumed, and the $\frac{1}{2}$-coefficient in the Taylor-remainder term is omitted. Since $\mu = \frac{1}{2}CM$ our bounds looks similar to their bound $|f(0)| \leq M^{-5}$. In the next section the comparison will be a true pain, as [KP] have a formulation in which again all constants are called M with apparently $M > 1$, and the bound on some norm of $f(0)$ (the wrong norm actually) involving M^{-307}. Comparing to the lectures notes of Schwartz from 60 years ago this is hardly an improvement as Schwartz had M^{-202} (also for the wrong norm).

39 Nash' modification of Newton's method

Now that we have seen several small variants of the method to obtain convergence for Newton's method, we consider the problem of solving $f(u) = 0$ in $B \subset X$ in the case that $f : B \to Z$ and $L(u) : Z \to Y$ with X, Y and Z *different* Banach spaces that we assume to belong to a family of spaces denoted by C^k, which we think of as function spaces. Here k denotes the number of possibly fractional derivatives that elements $u \in C^k$ have. Think of k for X, l for Z and m for Y. The goal is to have conditions that guarantee the existence of a solution to $f(u) = 0$ with k-norm smaller than 1, provided $f(0)$ has a norm bounded by some power of M, where M is a universal bound for all constants related to the derivatives of f.

Both [KP] and Schwartz require a very strong norm of $f(0)$ to be bounded, but the treatment below will show that a bound on the l-norm suffices. It should be noted that [KP] more or less copied from Schwartz with some additional details explained. Both formulate a statement for the case that $k > l$, but give a not completely correct proof for the case that $k = l > m$ (without mentioning the difference). The main additional assumption is a natural affine bound for $|L(u)f(u)|_{\bar{m}}$ in terms of $|u|_{\bar{k}}$, for $\bar{m}$ and $\bar{k}$ sufficiently large and $\bar{k} - \bar{m} = k - m$. The ratio

$$N = \frac{\bar{k} - k}{k - m} \tag{39.1}$$

measures the required higher regularity of the Newton map for the modified scheme described below to still do the job.

Below the norms $u \to |u|_k$ on C^k are assumed to be monotone increasing in k and we assume that there are linear so-called smoothing operators $S(t)$ parametrized by $t \geq 1$ that satisfy

$$|S(t)u|_k \leq K_{kl} t^{k-l} |u|_l \quad \text{and} \quad |(I - S(t))u|_l \leq \frac{K^{kl}}{t^{k-l}} |u|_k \tag{39.2}$$

for all $k > l$ in a sufficiently large range as needed in the particular implementation of the modified Newton method presented next. Thus $S(t)$ maps C^l to C^k, with an estimate for the ratio between the norms that grows worse as $S(t)$ approaches the identity I for $t \to \infty$, when I is considered as the embedding $I : C^k \to C^l$. It is convenient to write the norms of $S(t)$ and $I - S(t)$ with subscripts indicating the norms used for u, $S(t)u$ and $(I - S(t))u$. Thus (39.2) says that

$$|S(t)|_{kl} \leq K_{kl} t^{k-l} \quad \text{and} \quad |(I - S(t))|_{lk} \leq \frac{K^{kl}}{t^{k-l}}. \tag{39.3}$$

Besides (39.3) we assume (now also) a bound M_1^{lk} on $|f'(u)|_{lk}$ and, as before, bounds M_2^{lk} on $|f''(u)|_{lk}$ and C_{ml} on $|L(u)|_{ml}$ for $|u|_k \leq 1$.

39.1 The modified scheme

The idea of Nash was to modify Newton's scheme into

$$u_n = u_{n-1} - S(t_{n-1})L(u_{n-1})f(u_{n-1}), \tag{39.4}$$

with a suitable choice of $t_n \to \infty$ as $n \to \infty$. In (39.4) the new factor $S_{n-1} = S(t_{n-1})$ maps $L(u_{n-1})f(u_{n-1})$ back to (the strict subset of smooth functions of) the original domain of f. This comes with a cost which is estimated using the norm of the smoothing operator S_{n-1} in the chain

$$u_{n-1} \in X = C^k \xrightarrow{f} Z = C^l \xrightarrow{L(u_{n-1})} Y = C^m \xrightarrow{S_{n-1}} u_n \in X = C^k.$$

Before we do so let's examine how (38.4) is modified when combined with (39.4). We have

$$f(u_n) = \underbrace{f(u_{n-1}) + f'(u_{n-1})(u_n - u_{n-1})}_{\text{vanishes with (38.10)}} + Q_f(u_{n-1}, u_n)$$

$$= \underbrace{f'(u_{n-1})(I - S_{n-1})L(u_{n-1})f(u_{n-1})}_{\text{because of (39.4)}} + Q_f(u_{n-1}, u_n),$$

so that, with

$$p_n = |u_n - u_{n-1}|_k \quad \text{and} \quad q_n = |f(u_n)|_l,$$

the estimate

$$q_n \leq \underbrace{M_1^{lk}|I - S_{n-1}|_{km}|L(u_{n-1})f(u_{n-1})|_m}_{\text{new error like term}} + \frac{1}{2}M_2^{lk}p_n^2 \tag{39.5}$$

holds.

39.2 The new error term

The third factor in the error like term in (39.5) will have to be controled using some assumption on the map

$$u \to L(u)f(u)$$

which was not needed in the case of (38.10) and that should guarantee that quadratic term in (39.5) will still allow us to establish a conclusion like

(38.15). Clearly this is impossible if $m \leq k$ because we can only make $|I - S_n|_{km}$ small if $k < m$. Nash' solution was to replace m by a (much) larger $\bar{m}$ and assume an otherwise natural affine estimate of the form

$$|L(u)f(u)|_{\bar{m}} \leq A_{\bar{m}\bar{k}}(1 + |u|_{\bar{k}}) \tag{39.6}$$

with

$$\bar{k} - \bar{m} = k - m,$$

which requires an additional estimate for

$$r_n = 1 + |u_n|_{\bar{k}} \tag{39.7}$$

to be used in combination with

$$q_n \leq M_1^{lk} \underbrace{|I - S_{n-1}|_{k\bar{m}}}_{\text{controled by (39.3)}} r_{n-1} + \frac{1}{2} M_2^{lk} p_n^2 \tag{39.8}$$

and the estimate for p_n. Via (39.4) the latter now reads

$$p_n \leq |S_{n-1}|_{km} C_{ml} q_{n-1} \tag{39.9}$$

because $|L(u_{n-1})f(u_{n-1})|_m \leq C_{ml} q_{n-1}$.

The additional estimate needed for r_n also follows from (39.4). In view of

$$|u_n - u_{n-1}|_{\bar{k}} \leq |S_{n-1}|_{\bar{k}\bar{m}} |L(u_{n-1}f(u_{n-1})|_{\bar{m}} \leq |S_{n-1}|_{\bar{k}\bar{m}} A_{\bar{m}\bar{k}}(1 + |u_{n-1}|_{\bar{k}})$$

we have

$$1 \leq r_n \leq 1 + A_{\bar{m}\bar{k}} \sum_{j=1}^{n} |S_{j-1}|_{\bar{k}\bar{m}} r_{j-1}. \tag{39.10}$$

The "error" terms accumulate but can be kept under control as we shall see below.

The system of inequalities (39.9,39.8,39.10) and initial inequalities for q_0, $r_0 = 1$ and r_1 allows again estimates of the form (38.20), provided $\bar{k} - k = \bar{m} - m$ is sufficiently large in terms of (39.1). The idea is to get the first term in (39.8) controled by the right hand side of

$$p_n^2 \leq e^{-2\gamma\lambda^n}$$

in the induction argument, so that the norm $|S_n|_{km}$ in (39.9) can be chosen not to large so as still to have (38.20) with n if it already holds with $n-1$. To

do so we need a control on $|S_{n-1}|_{km}$ of the same form and this is established by setting

$$t_{n-1} = e^{\beta\lambda^{n-1}} \tag{39.11}$$

with $\beta > 0$ to be chosen in terms of γ. Note that this gives λ^n in the exponents of the exponential bounds for S_n and $I - S_n$.

Here we choose to keep λ as a parameter in a range as large as possible, like we did in the analysis of (38.10). Clearly we can only complete the argument if we also specify a bound on r_n to be established in the course of the argument, and this bound has to be of the same form as the bound chosen for S_n. Thus we look for a proof that

$$p_n \le e^{-\gamma\lambda^n} \quad \text{and} \quad r_n \le e^{\delta\lambda^n} \tag{39.12}$$

with $\delta > 0$. We note that the proof presented in [KP] the choice $\delta = \gamma$ and $\lambda = \frac{3}{2}$ dates back to Schwartz's lecture notes. As we shall see below this is not quite the optimal choice.

39.3 The system of inequalities

With (39.11) we have the system of inequalities

$$p_n \le K_{km} e^{(k-m)\beta\lambda^{n-1}} C_{ml} q_{n-1}; \tag{39.13}$$

$$q_n \le M_1^{lk} K^{k\bar{m}} e^{(k-\bar{m})\beta\lambda^{n-1}} A_{\bar{m}\bar{k}} \underbrace{r_{n-1}}_{\le e^{\delta\lambda^{n-1}}} + \frac{1}{2} M_2^{lk} \underbrace{p_n^2}_{\le e^{-2\gamma\lambda^n}}; \tag{39.14}$$

$$1 \le r_n \le 1 + \underbrace{A_{\bar{m}\bar{k}} K_{\bar{k}\bar{m}}}_{\mu_3} \sum_{j=1}^{n} e^{(\bar{k}-\bar{m})\beta\lambda^{j-1}} \underbrace{r_{j-1}}_{\le e^{\delta\lambda^{j-1}}}, \tag{39.15}$$

and we aim for a proof of (39.12) via induction, using the underbraced estimates in the three inequalities above as induction hypothesis. In (39.14) the estimate of the first term is controled by the estimate of the second term if

$$e^{(k-\bar{m})\beta\lambda^{n-1}} e^{\delta\lambda^{n-1}} \le e^{-2\gamma\lambda^n},$$

requiring

$$(\bar{m} - k)\beta \ge \delta + 2\gamma\lambda, \tag{39.16}$$

which says that in the λ, β-plane we must be above a line that comes down as $\bar{m}$ is increased.

Combining the first two inequalities we arrive at

$$p_n \le e^{(k-m)\beta\lambda^{n-1}}(\mu_1 e^{(k-\bar{m})\beta\lambda^{n-2}} r_{n-2} + \mu_2 p_{n-1}^2) \quad r_n \le 1 + \mu_3 \sum_{j=0}^{n-1} e^{(\bar{k}-\bar{m})\beta\lambda^j} r_j, \tag{39.17}$$

the constants μ_{123} given by

$$\mu_1 = \underbrace{K_{km}C_{ml}}_{C} M_1^{lk} \underbrace{K^{k\bar{m}}A_{\bar{m}\bar{k}}}_{A}, \quad \mu_2 = \frac{1}{2}\underbrace{K_{km}C_{ml}}_{C} M_2^{lk}, \quad \mu_3 = \underbrace{K_{\bar{k}\bar{m}}A_{\bar{m}\bar{k}}}_{\bar{A}}. \tag{39.18}$$

39.4 Estimating the increments

Under the assumption that (39.16) holds, the induction hypotheses for p_{n-1} and r_{n-2} produce the desired inequality for p_n from (39.17) if

$$(\mu_1 + \mu_2) e^{(k-m)\beta\lambda^{n-1}} e^{-2\gamma\lambda^{n-1}} \le e^{-\gamma\lambda^n}.$$

Thus we must have

$$\ln(\mu_1 + \mu_2) \le -(k-m)\beta\lambda^{n-1} + 2\gamma\lambda^{n-1} - \gamma\lambda^n$$

for all $n \ge 2$. As in the case of the standard Newton scheme, this leads to

$$\ln(\mu_1 + \mu_2) \le \lambda(\gamma(2-\lambda) - (k-m)\beta) \quad \text{with} \quad (k-m)\beta \le \gamma(2-\lambda), \tag{39.19}$$

a sharp upper bound for β that we need to stay away from if we don't want to impose that $\mu_1 + \mu_2 \le 1$.

As sufficient condition for

$$\sum_{n=1}^{\infty} p_n < 1$$

we can use the optimal condition found using Bernoulli's inequality, namely

$$\lambda\gamma(2-\lambda) \le \ln(e^{\gamma\lambda} - 1). \tag{39.20}$$

39.5 Estimating the error terms

For the inductive construction of the upper bound for r_n we set

$$b = (\bar{k} - \bar{m})\beta = (k-m)\beta > 0 \tag{39.21}$$

and conclude from the inequality in (39.17) that (shifting the index)

$$r_n \le 1 + \mu_3 \sum_{j=0}^{n-1} e^{b\lambda^j} r_j \le 1 + \mu_3 \sum_{j=0}^{n-1} e^{b\lambda^j} e^{\delta\lambda^j}$$

in view of the induction assumption for (all) smaller n. Thus we need the inequality

$$1 + \mu_3 \sum_{j=0}^{n-1} e^{(b+\delta)\lambda^j} \le e^{\delta\lambda^n} \tag{39.22}$$

for all $n \ge 2$. Recall that we start with $r_0 = 1 \le e^\delta$ and

$$1 \le r_1 \le e^{\delta\lambda} \quad \text{(and also } p_1 \le e^{\gamma\lambda} \text{ of course)} \tag{39.23}$$

via a smallness assumptions on q_0 still to be discussed.

Dividing by the right hand side, (39.22) is equivalent to

$$e^{-\delta\lambda^n} + \mu_3 (e^{(b+\delta-\delta\lambda)\lambda^{n-1}} + e^{-\delta\lambda^n} \sum_{j=0}^{n-2} e^{(b+\delta)\lambda^j}) \le 1 \tag{39.24}$$

in which we have separated the probably dominant term with $j = n-1$ from the sum. Neglecting the sum in (39.24) a sufficient (and in any case necessary) condition for the induction step to work for all $n \ge 2$ would be that

$$\ln \mu_3 + (b + \delta - \delta\lambda)\lambda^{n-1} \le 0 \quad \text{with} \quad b \le \delta(\lambda - 1), \tag{39.25}$$

so that in particular we now need to impose two inequalities on b, namely

$$b < \delta(\lambda - 1) \quad \text{and} \quad b < \gamma(2 - \lambda), \tag{39.26}$$

the latter being the (strict) inequality from (39.19).

These two bounds severely restrict the bound in (39.16), which in terms of b becomes

$$\frac{\bar{m} - k}{k - m} b \ge \delta + 2\gamma\lambda, \tag{39.27}$$

and this does not really depend on how we turn the necessary condition (39.25) into a sufficient condition, which we do next, rewriting it as

$$e^{-\delta\lambda^n} + \mu_3 (e^{(b+\delta-\delta\lambda)\lambda^{n-1}} + \sum_{j=0}^{n-2} e^{(b+\delta-\delta\lambda^{n-j})\lambda^j}) \le 1.$$

In view of (39.26) and using Bernoulli's inequality (38.31) the left hand side is smaller than

$$e^{-\delta\lambda^2} + \mu_3(e^{(b+\delta-\delta\lambda)\lambda} + \sum_{j=0}^{n-2} e^{(b+\delta-\delta\lambda^2)\lambda^j}) <$$

$$e^{-\delta\lambda^2} + \mu_3(e^{(b+\delta-\delta\lambda)\lambda} + \sum_{j=0}^{\infty} e^{(b+\delta-\delta\lambda^2)(1+j(\lambda-1))}) <$$

$$e^{-\delta\lambda^2} + \mu_3(e^{(b+\delta-\delta\lambda)\lambda} + \frac{e^{b+\delta-\delta\lambda^2}}{1-e^{(\lambda-1)(b+\delta-\delta\lambda^2)}}),$$

in which we used that $b+\delta-\delta\lambda^2 < b+\delta-\delta\lambda < 0$. Thus we arrive at

$$e^{-\delta\lambda^2} + \mu_3 e^{(b+\delta-\delta\lambda)\lambda}(1 + \frac{e^{-(b+\delta)(\lambda-1)}}{1-e^{(\lambda-1)(b+\delta-\delta\lambda^2)}}) \leq 1 \tag{39.28}$$

Note that the first term on the right hand side of (38.31) is essential here. Without this first term the numerator, which is the first term ($j=0$) in the geometric series, would be 1 and we be stuck, as there would be no way to get a statement without an a priori bound on μ_3. We note that in [KP] the proof is without the 1 in (38.31) but an accidental mistake of computing the series with $j=1$ as the first term "allows" to conclude. Technically speaking that proof is incorrect[1].

The quickest way to finish is to estimate the sum of the geometric series by a fixed constant, rewriting it as

$$\frac{e^{-s}}{1-e^{s-S}} = \frac{e^S}{e^s(e^S-e^s)}$$

with

$$s = (b+\delta)(\lambda-1) \leq \delta\lambda(\lambda-1) = s_0 < S = \delta\lambda^2(\lambda-1).$$

Provided

$$2e^{s_0} \leq e^S \quad \text{or} \quad \ln 2 \leq \delta\lambda(\lambda-1)^2,$$

this expression is monotone decreasing in s on $[0, s_0]$ and thus

$$\frac{e^{-(b+\delta)(\lambda-1)}}{1-e^{(\lambda-1)(b+\delta-\delta\lambda^2)}} \leq \frac{1}{1-e^{-\delta(\lambda-1)\lambda^2}} \leq 2.$$

We conclude that

$$e^{-\delta\lambda^2} + 3\mu_3 e^{(b+\delta-\delta\lambda)\lambda} \leq 1 \quad \text{suffices if} \quad \ln 2 \leq \delta\lambda(\lambda-1)^2, \tag{39.29}$$

[1]And it is not a proof of the theorem actually stated.

and the first inequality in (39.29) certainly holds if it holds with the first exponential replaced by the larger second exponential. Thus we arrive at

$$\ln(1+3\mu_3) \le \lambda(\delta(\lambda-1)-b) \quad \text{and} \quad \ln 2 \le \delta\lambda(\lambda-1)^2 \tag{39.30}$$

as the final condition needed.

39.6 Sufficient conditions for a convergence result

Summing up, with the condition on q_0 still to be imposed we arrive at

$$\lambda\gamma(2-\lambda) \le \ln(e^{\gamma\lambda}-1), \tag{39.31}$$

$$\ln 2 \le \delta\lambda(\lambda-1)^2, \tag{39.32}$$

$$(\bar{m}-k)\beta \ge \delta + 2\gamma\lambda, \tag{39.33}$$

$$(k-m)\beta < \gamma(2-\lambda) \quad \text{and} \quad (\bar{k}-\bar{m})\beta < \delta(\lambda-1) \tag{39.34}$$

as conditions on the parameters that we still have to choose.

The first inequality, (39.31), is to have the sum of the increments, and thereby the solution, bounded by 1 in the l-norm. Of course it can be replaced by just asking that

$$\sum_{n=1}^{\infty} e^{-\gamma\lambda^n} \le 1.$$

The second, (39.32), was a technical condition to bound the sum of the geometric series in (39.28) by 2. The third, (39.33), allows to bound the error term in estimate (39.14) for q_n by the bound on p_n^2 that has to be established. It involves the choice of sufficiently large $\bar{m}$ and $\bar{k}$ with $\bar{k}-\bar{m} = k-m$.

The last two conditions are strict inequalities that have be chosen sufficiently strict depending on the constants related to f, to allow for an inductive proof of the desired estimates (39.12) for p_n and r_n. Thus, given μ_1, μ_2, μ_3, we need to choose $1 < \lambda < 2$ and $\gamma, \beta, \delta > 0$ such that

$$\lambda(\gamma(2-\lambda) - (m-k)\beta) \ge \ln(\mu_1+\mu_2); \tag{39.35}$$

$$\lambda(\delta(\lambda-1) - (\bar{m}-\bar{k})\beta) \ge \ln(1+3\mu_3). \tag{39.36}$$

After a simultaneous rescaling of γ, β, δ, this is always possible once the first 5 conditions are satisfied. The inequalities in (39.34) being strict is essential for convergence of Nash' modified Newton scheme.

Of course we still have to formulate the necessary sufficient bound on $q_0 = |f(0)|_l$, given the constants in (39.18) and the choice of parameters above. Recall that

$$\mu_1 + \mu_2 = \underbrace{K_{km}C_{ml}}_{C}(M_1^{lk}\underbrace{K^{k\bar{m}}A_{\bar{m}\bar{k}}}_{A} + \frac{1}{2}M_2^{lk}) = C(M_1 A + \frac{1}{2}M_2)$$

and

$$\mu_3 = K_{\bar{k}\bar{m}}A_{\bar{m}\bar{k}} = \bar{A},$$

with $C, M_1, M_2, A, \bar{A}$ constants related to f and the smoothing operators. From here on we drop the superscripts from the bounds M_1 and M_2 on the first and second derivative of $f : C^k \to C^l$.

39.7 Sufficient convergence condition on initial value

Finally we examine the initial inequalities we need. For p_1 we need, since $u_0 = 0$, that

$$p_1 = |u_1|_k = |S_0|_{km}|L(0)|_{ml}|f(0)|_l \leq e^{(k-m)\beta}\underbrace{K_{km}C_{ml}}_{C}|f(0)|_l \leq e^{-\gamma\lambda},$$

while via

$$|u_1|_{\bar{k}} \leq |S(0)|_{\bar{k}m}|L(0)|_{ml}|f(0)|_l \leq K_{\bar{k}m}e^{(\bar{k}-m)\beta}C_{ml}|f(0)|_l \leq e^{\delta\lambda}$$

we need

$$1 + \underbrace{K_{\bar{k}m}C_{ml}}_{\bar{C}}e^{(\bar{k}-m)\beta}|f(0)|_l \leq e^{\delta\lambda}$$

for r_1. Thus

$$Cq_0 \leq e^{-(k-m)\beta}e^{-\gamma\lambda} \quad \text{and} \quad \bar{C}q_0 \leq e^{-(\bar{k}-m)\beta}(e^{\delta\lambda} - 1) \tag{39.37}$$

are sufficient conditions on

$$q_0 = |f(0|_l$$

to have a solution of $f(u) = 0$ with $|u|_k < 1$, once the parameters have been chosen according to Section 39.6 to make the induction steps work in the proof of the desired estimates (39.12) for p_n and r_n.

39.8 The optimal choice of parameters

At this point we compare (39.35) and (39.36) to (38.23). The strict inequalities in (39.34) are really strict in the sense that the gaps have to be taken sufficiently large large given the explicit constants related to f and $S(t)$. The other two inequalities are not strict. Recalling that $k-m=\bar{k}-\bar{m}$, the coefficient

$$\frac{\bar{m}-k}{k-m}=\frac{\bar{m}-m}{k-m}-1=\frac{\bar{m}-m}{\bar{k}-\bar{m}}-1=\frac{\bar{k}-k}{k-m}-1=N-1 \tag{39.38}$$

has to be sufficiently large for the set of allowable b, as defined by (39.21), to be nonempty. Note that in Nash' strategy to get around the ill-posedness of Newton's method, (39.1) is the natural definition of N as the ratio of the required increase of smoothness by $\bar{k}-k$ to the loss of smoothness by $m-k$ in $u\to L(u)f(u)$.

The minimal largeness condition on N is obtained by taking the right hand sides of the inequalities in (39.34) equal to one another, so as to maximize the allowable upper bound for β. Thus we choose $1<\lambda<2$ such that

$$\gamma(2-\lambda)=\delta(\lambda-1)\quad\text{whence}\quad\lambda=\frac{2\gamma+\delta}{\gamma+\delta} \tag{39.39}$$

and (39.33,39.34) become

$$\frac{4\gamma^2+3\gamma\delta+\delta^2}{\gamma+\delta}\le(N-1)b<(N-1)\frac{\gamma\delta}{\gamma+\delta} \tag{39.40}$$

for

$$b=(k-m)\beta=(\bar{k}-\bar{m})\beta$$

in terms of γ,δ,N, subject to (39.31,39.32) which reduce to

$$e^{\frac{2\gamma+\delta}{\gamma+\delta}\frac{\delta}{\gamma+\delta}}+1\le e^{\gamma\frac{2\gamma+\delta}{\gamma+\delta}}\quad\text{and}\quad\ln 2\le\delta\,\frac{2\gamma+\delta}{\gamma+\delta}\left(\frac{\gamma}{\gamma+\delta}\right)^2. \tag{39.41}$$

In particular (39.40) requires

$$N>\frac{4\gamma}{\delta}+4+\frac{\delta}{\gamma}\ge 8, \tag{39.42}$$

the minimum 8 being realised by

$$\delta=2\gamma. \tag{39.43}$$

The further choice of parameters depends on the constants which are as indicated in (39.18), at the end of Section 39.6 and in (39.37):

$$C = K_{km}C_{ml};\ \bar{C} = K_{\bar{k}m}C_{ml};\ A = K^{k\bar{m}}A_{\bar{m}\bar{k}};\ \bar{A} = K_{\bar{k}\bar{m}}A_{\bar{m}\bar{k}}; \tag{39.44}$$

$$\mu_1 + \mu_2 = C(M_1 A + \frac{1}{2}M_2); \quad \mu_3 = \bar{A}. \tag{39.45}$$

We collect these constants in one single constant Θ as

$$\Theta = \frac{3}{4}\max(\ln C + \ln(M_1 A + \frac{1}{2}M_2), \ln(1 + 3\bar{A})) \tag{39.46}$$

and, depending on N, the remaining parameters γ, b have to be chosen to control these constants via

$$\Theta \leq \frac{2\gamma}{3} - b \tag{39.47}$$

and

$$\frac{14\gamma}{3} \leq (N-1)b < \frac{2\gamma}{3}(N-1), \quad e^{\frac{8}{9}} + 1 \leq e^{\frac{4\gamma}{3}}, \quad \ln 2 \leq \frac{8\gamma}{27}, \tag{39.48}$$

which is (39.40,39.41) with $\delta = 2\gamma$. The last inequality now implies the one preceding it.

For the initial condition q_0 we arrive via (39.39) at

$$Cq_0 \leq e^{-\gamma\frac{2\gamma+\delta}{\gamma+\delta}}e^{-b} \quad \text{and} \quad \bar{C}q_0 \leq (e^{\delta\frac{2\gamma+\delta}{\gamma+\delta}} - 1)e^{-(\bar{k}-m)\beta} =$$

$$(e^{\delta\frac{2\gamma+\delta}{\gamma+\delta}} - 1)e^{-(\bar{k}-\bar{m})\beta}e^{-(\bar{m}-m)\beta} = (e^{\delta\frac{2\gamma+\delta}{\gamma+\delta}} - 1)e^{-(N+1)b},$$

so that with $\delta = 2\gamma$ the conditions on q_0 reduce to

$$Cq_0 \leq e^{-\frac{4\gamma}{3}}e^{-b} \quad \text{and} \quad \bar{C}q_0 \leq (e^{\frac{8\gamma}{3}} - 1)e^{-(N+1)b}. \tag{39.49}$$

Setting

$$\rho = \frac{2\gamma}{3}$$

we arrive at

$$\Theta \leq \rho - b, \quad 7\rho \leq (N-1)b, \quad \rho \geq \frac{9}{4}\ln 2,$$

$$\ln C + \ln q_0 \leq -2\rho - b, \quad \ln \bar{C} + \ln q_0 \leq \ln 80 - (N+1)b,$$

as sufficient conditions. Note that we have used the lower bound for ρ to relax the bound on $\bar{C}q_0$.

Choosing

$$N > 8 \quad \text{and} \quad \rho = \frac{N-1}{7} b$$

and using the last lower bound for ρ we arrive at

$$b \geq \max(\frac{63}{4}\frac{\ln 2}{N-1}, \frac{7\Theta}{N-8}) \quad \text{and} \quad q_0 \leq \min(\frac{1}{C} e^{-\frac{2N+5}{7N} b}, \frac{80}{\bar{C}} e^{-(N-1)b}) \tag{39.50}$$

as sufficient conditions, to be used as: given Θ choose $N > 8$ and $b = (k-m)\beta$ sufficiently large to make the condition on q_0 follow and thereby obtain a solution of $f(u) = 0$ with $|u|_k < 1$.

39.9 Continuity

Given the constants related to f and the smoothing operators we constructed a solution in the open unit k-ball, that is, with $|u|_k < 1$. We did not prove or state that the solution is unique, but it is well-defined as the limit of an explicitly constructed sequence shown to be convergent if $|f(0)|_k$ is sufficiently small. The following issue relates to the continuity of the inverse function of f, if it were to exist, since we should naturally also ask for a condition $|f(0)|_k$ guaranteeing the constructed solution to have $|u|_k \leq \varepsilon$. This only changes the condition on the sum of the increments and leads to

$$\gamma\lambda(2-\lambda) \leq \ln(e^{\gamma\lambda} - \frac{1}{\varepsilon})$$

leading to

$$\Theta \leq \frac{2\gamma}{3} - b, \quad \frac{14\gamma}{3} \leq (N-1)b < \frac{2\gamma}{3}(N-1), \quad e^{\frac{8}{9}} + \frac{1}{\varepsilon} \leq e^{\frac{4\gamma}{3}}, \quad \ln 2 \leq \frac{8\gamma}{27},$$

in stead of (39.47,39.48). The conditions on γ rewrite as

$$\gamma \geq \max(\frac{3}{4}\ln(\frac{1}{\varepsilon} + e^{\frac{8}{9}}), 2^{\frac{27}{8}}) \sim \varepsilon^{-\frac{3}{4}}$$

as $\varepsilon \to 0$. This forces a larger choice of b and thereby via (39.49) a smaller (exponentially small in terms of ε in fact) bound on q_0 for the Nash scheme to converge within the ball of k-radius ε, as was to be expected of course. The fact that the limit u is a solution of $f(u) = 0$ is immediate from (39.14).

Note that for the standard Newton method the constructed solution of $f(u) = 0$ will have $|u| < \varepsilon$ if we take equalities in (38.30) and replace the -1 by $-\frac{1}{\varepsilon}$. The upper bound $\bar{P}$ than has to be replaced by $\bar{P}_\varepsilon = \frac{\varepsilon}{1+\mu\varepsilon}$ and the condition on q_0 becomes $q_0 \leq C\bar{P}_\varepsilon$.

40 The Nash embedding theorem

The Schwartz's lecture notes contain a nice but nonconstructive argument to apply the above together with convexity arguments and the Hahn-Banach Theorem to prove that the n-dimensional torus with any nonstandard Riemannian metric embeds in some $\mathbb{R}^N$. To be explained here. See Chapter 37. Requires a deeper discussion of the smoothing operators used in the proof of Theroem 35.21 and the Fouriertransfrom.

41 Hartman-Grobman stelling

Some material prepared for this very enjoyable event (see also Section 15):

`www.universiteitleiden.nl/agenda/2017/04/nationaal-wiskunde-symposium`

In [HM] hebben we uitgebreid gekeken naar de Methode van Newton voor het oplossen van vergelijkingen, met als eerste voorbeeld het snel benaderen van algebraïsche getallen, bijvoorbeeld $\sqrt{2}$, dat een vast punt is van de afbeelding

$$x \xrightarrow{F} F(x) = \frac{1}{2}(x + \frac{2}{x}),$$

een afbeelding die ongeveer 3000 jaar oud is, en later herontdekt is via

$$f(x) = x^2 - 2 \quad \text{en} \quad F(x) = x - f'(x)^{-1} f(x) = x - \frac{f(x)}{f'(x)}.$$

De mooie eigenschappen van het discrete dynamisch systeem gedefinieerd door

$$x_n = F(x_{n-1}) \quad (n \in \mathbb{N})$$

worden deels verklaard door het feit

$$F'(x) = \frac{f(x)f''(x)}{f'(x)^2}$$

gelijk is aan 0 in nulpunten van $f(x)$ en

$$F(x) = x \iff f(x) = 0$$

voor elke x met $f'(x) \neq 0$. Een curieus voorbeeldje in Hoofdstuk 6 van The Beauty of Fractals van Peitgen en Richter is

$$f(x) = \frac{x}{1-x} \quad \text{en} \quad F(x) = x^2,$$

en dat is er eentje uit een curieuze familie, bijvoorbeeld

$$f(x) = \frac{x}{(1-x)^{\frac{1}{7}}(1 + x + x^2 + x^3 + x^4 + x^5 + x^6)^{\frac{1}{7}}} \quad \text{en} \quad F(x) = x^8,$$

maar dat terzijde. De afbeeldingen

$$x \to x^2 \quad \text{en} \quad x \to \frac{1}{2}(x + \frac{2}{x})$$

hebben vaste punten waar hun afgeleide 0 is en het is niet moeilijk jezelf ervan te overtuigen dat dit leidt tot snelle convergentie de rij x_n naar evenwicht, als je begint met x_0 in de buurt van een evenwicht.

Neem je zomaar een functie F om een dynamisch systeem te maken zoals hierboven, en is $F(0) = 0$, dan bepaalt de afgeleide $F'(0)$ in het algemeen of $x = 0$ een (lokaal) stabiel of onstabiel evenwicht is, zoals het voorbeeld

$$x \to \lambda x$$

met $\lambda \in \mathbb{R}$ bij inspectie meteen laat zien. Een voor de hand liggende vraag is dan of de dynamische systemen gedefinieerd door

$$x \to F(x) \quad \text{en} \quad \tilde{x} \to F'(0)\tilde{x}$$

niet eigenlijk hetzelfde zijn via een conjugatie:

$$\begin{array}{ccc} x & \stackrel{\phi}{\to} & \tilde{x} \\ F \downarrow & & \downarrow F'(0) \\ F(x) & \stackrel{\phi}{\to} & F'(0)\tilde{x} \end{array}$$

Dus is er een inverteerbare afbeelding ϕ waarmee

$$F'(0)\phi(x) = \phi(F(x))$$

voor x in een zo groot mogelijke buurt van $x = 0$? Als we $F(x)$ schrijven als

$$F(x) = \lambda x + a(x)$$

dan is de vraag dus of we gegeven $\lambda \in \mathbb{R}$ en $a : \mathbb{R} \to \mathbb{R}$ met $a'(0) = 0$ de functie $\phi : \mathbb{R} \to \mathbb{R}$ kunnen vinden zodanig dat

$$\lambda\phi(x) = \phi(\lambda x + a(x))$$

in de buurt van $x = 0$, en dit kunnen we proberen op te lossen door middel van

$$\phi_n(x) = \frac{\phi_{n-1}(\lambda x + a(x))}{\lambda} \quad \text{beginnend met} \quad \phi_0(x) = x.$$

Als we een a nemen met $a'(0) = 0$ en $a(x) = 0$ voor $|x|$ buiten een interval gedefinieerd door $|x| \leq \eta$ met η wellicht nog te kiezen, dan zien we dat de onbekende functie ϕ voor $|x| \geq \eta$ wel gegeven moet worden door $\phi(x) = x$. Hoewel? Is het duidelijk dat gegeven $\lambda \in \mathbb{R}$ uit

$$\lambda\phi(x) = \phi(\lambda x)$$

voor alle $x \in \mathbb{R}$ volgt dat $\phi(x) = x$ voor alle $x \in \mathbb{R}$? Niet meteen dus. Maar $\phi(x) = x$ doet het wel.

Terzijde, als ϕ differentieerbaar is volgt (als $\lambda \neq 0$) dat

$$\phi'(x) = \phi'(\lambda x)$$

voor alle $x \in \mathbb{R}$, en daarmee zijn heel veel ϕ' waarden gelijk aan elkaar, tenzij $|\lambda| = 1$. Als ϕ' continu is in 0 moet wel te bewijzen zijn dat $\phi'(x) = \phi'(0)$ voor alle x. En $\phi'(0) = 1$ ligt voor de hand als normaliserende voorwaarde.

Of we voor voor $a(x) \not\equiv 0$ zo'n differentieerbare ϕ wel maken is echter zeer de vraag. In het iteratieproces helpt de λ in de noemer wellicht als $|\lambda| > 1$ is. Weer terzijde, het voorbeeld met $a(x) = x^2$ laat zien dat zonder de aanname dat $a(x) \equiv 0$ voor $|x|$ groot er weinig hoop is, want we krijgen

$$\phi_1(x) = x + \frac{x^2}{\lambda},$$

$$\phi_2(x) = x + \left(\frac{1}{\lambda} + 1\right) x^2 + \frac{2x^3}{\lambda} + \frac{x^4}{\lambda^2},$$

$$\begin{aligned}\phi_3(x) = x &+ \left(\frac{1}{\lambda} + 1 + \lambda\right) x^2 + \left(\frac{2}{\lambda} + 2 + 2\,\lambda\right) x^3 + \left(\frac{1}{\lambda^2} + \frac{1}{\lambda} + 6 + \lambda\right) x^4 \\ &+ \left(\frac{6}{\lambda} + 4\right) x^5 + \left(\frac{2}{\lambda^2} + \frac{6}{\lambda}\right) x^6 + \frac{4x^7}{\lambda^2} + \frac{x^8}{\lambda^3},\end{aligned}$$

$$\begin{aligned}\phi_4(x) = x &+ \left(\frac{1}{\lambda} + 1 + \lambda + \lambda^2\right) x^2 + \left(\frac{2}{\lambda} + 2 + 4\,\lambda + 2\,\lambda^2 + 2\,\lambda^3\right) x^3 \\ &+ \left(\frac{1}{\lambda^2} + \frac{1}{\lambda} + 7 + 7\,\lambda + 6\,\lambda^3 + \lambda^4 + 7\,\lambda^2\right) x^4 + \cdots + \frac{x^{16}}{\lambda^4},\end{aligned}$$

enzovoorts.

De Stelling van Hartman Grobman gaat in het simpelste geval om de vraag of voor de afbeelding

$$(x, y) \xrightarrow{F} (\xi, \eta) = (\lambda x + a(x, y), \mu y + b(x, y))$$

het stelsel

$$\begin{aligned}\lambda\phi(x, y) &= \phi(\lambda x + a(x, y), \mu y + b(x, y)) \\ \mu\psi(x, y) &= \psi(\lambda x + a(x, y), \mu y + b(x, y))\end{aligned}$$

kunnen oplossen naar de functies ϕ, ψ onder de aanname dat

$$0 < |\mu| < 1 < |\lambda|,$$

teneinde de afbeelding F te conjugeren met de afbeelding

$$(\tilde{x}, \tilde{y}) \to (\tilde{\xi}, \tilde{\eta}) = (\lambda\tilde{x}, \mu\tilde{y}).$$

Dit zijn twee vergelijkingen, in de eerste is de onbekende de functie ϕ, in de tweede de functie ψ. Die voor ϕ lijkt op de vergelijking waarmee we begonnen en waarvoor de geschetste aanpak kans van slagen heeft als $|\lambda| > 1$. De aannamen op $a(x, y)$ en $b(x, y)$ zijn nu zoals die op $a(x)$ hierboven, dus

$$a_x(0, 0) = a_y(0, 0) = b_x(0, 0) = b_y(0, 0) = 0,$$

en $a(x, y)$ en $b(x, y)$ tenminste continu differentieerbaar. Met die conditie is het stelsel

$$\xi = \lambda x + a(x, y)$$

$$\eta = \mu y + b(x, y)$$

in de buurt van $(0, 0)$ op te lossen naar x, y in de vorm

$$x = \frac{1}{\lambda}\,\xi + \alpha(\xi, \eta)$$

$$y = \frac{1}{\mu}\,\eta + \beta(\xi, \eta)$$

met $\alpha(\xi, \eta)$ en $\beta(\xi, \eta)$ continu differentieerbaar in de buurt van $(0, 0)$ en

$$\alpha_\xi(0, 0) = \alpha_\eta(0, 0) = \beta_\xi(0, 0) = \beta_\eta(0, 0) = 0.$$

De eerste vergelijking houden we zoals die was, de tweede schrijven we in ξ, η. Beide vergelijkingen hebben dan dezelfde vorm:

$$\phi(x, y) = \frac{1}{\lambda}\phi(\lambda x + a(x, y), \mu y + b(x, y))$$

$$\psi(\xi, \eta) = \mu\psi(\frac{1}{\lambda}\xi + \alpha(\xi, \eta), \frac{1}{\mu}\eta + \beta(\xi, \eta))$$

Om deze vergelijkingen op te lossen moeten we dus eerst weten hoe we de eerdere vergelijking voor $\phi : \mathbb{R} \to \mathbb{R}$ oplossen.

Als $|a(x)| \leq \varepsilon|x|$ dan volgt

$$|\phi_1(x) - \phi_0(x)| = |\frac{1}{\lambda}(\lambda x + a(x)) - x| = |\frac{a(x)}{\lambda}| \leq \frac{\varepsilon}{\lambda}|x|,$$

en dan

$$|\phi_2(x) - \phi_1(x)| = |\frac{1}{\lambda}\phi_1(\lambda x + a(x)) - \frac{1}{\lambda}\phi_0(\lambda x + a(x))|$$

$$\leq \frac{\varepsilon}{|\lambda|^2}|\lambda x + a(x)| \leq \frac{\varepsilon(|\lambda| + \varepsilon)}{|\lambda|^2}|x|,$$

waarna

$$|\phi_3(x) - \phi_2(x)| = |\frac{1}{\lambda}\phi_2(\lambda x + a(x)) - \frac{1}{\lambda}\phi_1(\lambda x + a(x))|$$

$$\leq \frac{\varepsilon(|\lambda| + \varepsilon)}{|\lambda|^3}|\lambda x + a(x)| \leq \frac{\varepsilon(|\lambda| + \varepsilon)^2}{|\lambda|^3}|x|.$$

Zo wordt duidelijk dat

$$|\phi_n(x) - \phi_{n-1}(x)| \leq \frac{\varepsilon(\lambda + \varepsilon)^{n-1}}{|\lambda|^n}|x|,$$

niet genoeg om de rij $\phi_n(x)$ convergent te krijgen, maar als ook geldt dat $a(x) = 0$ voor $|x| \geq \eta$ dan kunnen we met $0 < \delta < 1$ de schattingen aanpassen als

$$|\phi_1(x) - \phi_0(x)| \leq \frac{\varepsilon}{|\lambda|}\eta^{1-\delta}|x|^\delta,$$

$$|\phi_2(x) - \phi_1(x)| \leq \frac{\varepsilon}{|\lambda|^2}\eta^{1-\delta}|\lambda x + a(x)|^\delta \leq \frac{\varepsilon}{|\lambda|^2}\eta^{1-\delta}(|\lambda| + \varepsilon)|x|^\delta,$$

en dan wordt duidelijk dat

$$|\phi_n(x) - \phi_{n-1}(x)| \leq \varepsilon\eta^{1-\delta}\frac{(|\lambda| + \varepsilon)^{\delta(n-1)}}{|\lambda|^n}|x|^\delta.$$

Uniforme convergentie volgt als

$$(|\lambda| + \varepsilon)^\delta < |\lambda|.$$

42 Airy functions

I did the calculations below while reading Chapter 8 in Peter Olver's new PDE book with a little help of E.J. Hinch's nice little Cambridge Applied Math textbook on perturbation methods. Just goes to show how beautiful (also applied) complex analysis is.

The Airy function is defined by

$$\mathrm{Ai}(\xi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i(\xi x + \frac{x^3}{3})} dx,$$

an integral barely convergent. The Airy function plays the same role in the theory for $u_t + u_{xxx} = 0$ as the Gaussian $e^{-\frac{1}{2}x^2}$ for $u_t = u_{xx}$. Both functions define the spatial profile of the fundamental solution.

Replacing $\xi \in \mathbb{R}$ by $\zeta \in \mathbb{C}$ the Airy function is a complex analytic function of

$$\zeta = \xi + i\eta = \rho e^{i\psi}.$$

Replacing also $x \in \mathbb{R}$ by

$$z = x + iy \in \mathbb{C}$$

one may deform the "real" contour C defined by $z = z(t) = t$ with $-\infty < t < \infty$ to another contour γ_ζ that connects two points at infinity. This can be done (without changing the outcome) as long as the integrals of

$$e^{i(\zeta z + \frac{z^3}{3})} = e^{\Phi(z;\zeta)}$$

over the connecting arcs $|z| = R$ between C and the new contour γ_ζ go to zero as $R \to \infty$.

To answer the question

$$\mathrm{Ai}(\rho)e^{i\psi} \sim \quad ? \quad \text{as} \quad \rho \to \infty \quad (\text{for } \psi \text{ fixed}),$$

one chooses the new contour to be one along which the absolute value of the integrand, $e^{\mathrm{Re}\,\Phi(z;\zeta)}$, is peaked and has fast decay as $|z| \to \infty$, and along which $\mathrm{Im}\,\Phi(z;\zeta) = \phi_\zeta$ is a ζ-dependent constant, so that

$$\mathrm{Ai}(\zeta) = \frac{1}{2\pi}\int_{\gamma_\zeta} e^{i(\zeta z + \frac{z^3}{3})} dz = \frac{1}{2\pi}\int_{\gamma_\zeta} \underbrace{e^{\mathrm{Re}\,\Phi(z;\zeta)}}_{\text{real, positive}} dz\, e^{i\phi_\zeta},$$

in which the integrand is real, although $dz = dx + idy$ will typically still make the integral complex. The factor $e^{i\phi_\zeta}$ contains the "stationary phase"

ϕ_ζ. If M_ζ is the maximum of $\operatorname{Re}\Phi(z;\zeta)$ along γ_ζ, realised in some $z = m_\zeta$, one may also factor out e^{M_ζ} and write

$$\mathrm{Ai}(\zeta) = \frac{e^{M_\zeta + i\phi_\zeta}}{2\pi} \underbrace{\int_{\gamma_\zeta} e^{-f_0(z;\zeta)} dz}_{\to\, ?\,\text{as}\,|\zeta|\to\infty},$$

in which $f_0(z;\zeta) \geq 0$ along γ_ζ. Typically $f_0(z;\zeta)$ has a unique global minimum zero along γ_ζ and $f_0(z;\zeta) \to +\infty$ as $|z| \to \infty$ along γ_ψ. Note though that the integrand is likely to be ill-behaved as $\rho = |\zeta| \to \infty$, also because the contour γ_ζ may disappear in the limit. The resolution of this latter complication may be prepared by scaling x *before* going to complex variables and making the optimal choice of γ_ζ.

Thus, returning to the definition of $\mathrm{Ai}(\zeta)$ one writes $\mathrm{Ai}(\rho e^{i\psi}) =$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \underbrace{e^{i(\rho e^{i\psi}x + \frac{x^3}{3})}}_{\text{scale}\, x = \rho^{\frac{1}{2}}u} dx = \frac{\rho^{\frac{1}{2}}}{2\pi}\int_{-\infty}^{\infty} e^{i\rho^{\frac{3}{2}}(e^{i\psi}u + \frac{u^3}{3})} du = \frac{\rho^{\frac{1}{2}}}{2\pi}\int_{-\infty}^{\infty} e^{\rho^{\frac{3}{2}}\Psi(u)} du,$$

in which you should now view u as $u = \operatorname{Re} w$ with $w = u + iv \in \mathbb{C}$. The effect of this scaling is that the level lines of $\operatorname{Im}\Psi(w)$ are independent of ρ.

One has

$$\Psi(w) = i(e^{i\psi}w + \frac{w^3}{3}) = f(u,v;\psi) + ig(u,v;\psi),$$

with

$$f(u,v;\psi) = -v\cos\psi - u\sin\psi + v(-u^2 + \frac{v^2}{3})$$

and

$$g(u,v;\psi) = u\cos\psi - v\sin\psi + u(\frac{u^2}{3} - v^2).$$

These harmonic functions have mutually perpendicular level curves. It is convenient to think of the level curves of the imaginary part $g(u,v;\psi)$ as orbits of

$$\dot{u} = \frac{du}{dt} = f_u = \frac{\partial f}{\partial u} = -\sin\psi - 2uv$$
$$\dot{v} = \frac{dv}{dt} = f_v = \frac{\partial f}{\partial v} = -\cos\psi - u^2 + v^2,$$

a system of ordinary differential equations for $u = u(t)$ and $v = v(t)$. In fact, Cauchy-Riemann gives

$$\frac{df}{dt} = f_u\dot{u} + f_v\dot{v} = f_u^2 + f_v^2 > 0\,, \quad \frac{dg}{dt} = g_u\dot{u} + g_v\dot{v} = g_uf_u + g_vf_v = 0.$$

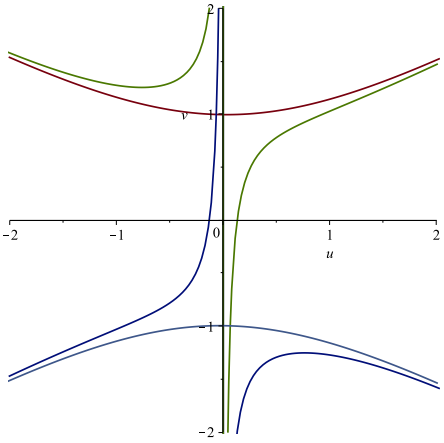


Figure 1: orbits for $\psi = \frac{\pi}{24}$

Thus the only orbits of interest as possible contours are the stable manifolds of saddle points, with the maximum of the real part f along the contour occurring in the saddle point.

This is illustrated for small positive ψ by Figure 1 which pictures the possibly relevant level curves of $g(u, v; \frac{\pi}{24})$. The red curve is the stable manifold of

$$m_\psi = (u_\psi, v_\psi) = (-\sin\frac{\psi}{2}, +\cos\frac{\psi}{2})$$

and asymptotes to $3v^2 = u^2$. In particular it has u ranging from $-\infty$ to $+\infty$, and may be written as the graph of a function $v = \varphi(u; \psi)$. The other stable manifold, that of

$$m_{\psi+2\pi} = (u_{\psi+2\pi}, v_{\psi+2\pi}) = (+\sin\frac{\psi}{2}, -\cos\frac{\psi}{2}),$$

the green curve on the right, fails this condition and has a vertical asymptote. The other branch of this level curve is the green curve on the left which is not a stable manifold of either two saddles. Neither of these two orbits is of direct use in relation to the Airy function, but this will change as ψ is taken larger. Note that although $\mathrm{Ai}(\rho e^{i\psi})$ is 2π-periodic in ψ, the parametrisation of the saddle point m_ψ is only 4π-periodic.

Deforming the contour as explained above,

$$\mathrm{Ai}(\rho e^{i\psi}) = \frac{\rho^{\frac{1}{2}}}{2\pi} \underbrace{\int_{-\infty}^{\infty} e^{\rho^{\frac{3}{2}}(f(u,\varphi(u;\psi);\psi)}(1 + i\varphi'(u;\psi))du}_{I(\rho,\psi)} e^{\rho^{\frac{3}{2}}(-\frac{2i}{3}(4\cos^2\frac{\psi}{2}-1)\sin\frac{\psi}{2})},$$

in which the phase factor has been made precise. It remains to examine the integral $I(\rho,\psi)$ in the limit $\rho\to\infty$. Clearly the most important information comes from the (second order) Taylor expansion

$$f = f(u,\varphi(u;\psi);\psi) = M_\psi - a_\psi^2(u-u_\psi)^2 + \cdots$$

with minor contributions coming from the higher order terms and the expansion of $\varphi'(u;\psi)$. Setting

$$u = u_\psi + p$$

one sees that to leading order the asymptotic expansion of the integral must be given by

$$I(\rho,\psi) \sim e^{M_\psi\rho^{\frac{3}{2}}}\int \underbrace{e^{-a_\psi^2\rho^{\frac{3}{2}}p^2}}_{\text{scale } s=a_\psi\rho^{\frac{3}{4}}p}\, dp \sim \frac{e^{M_\psi\rho^{\frac{3}{2}}}}{a_\psi\rho^{\frac{3}{4}}}\underbrace{\int e^{-s^2}ds}_{\sqrt{\pi}} + \cdots,$$

so that

$$\mathrm{Ai}(\rho e^{i\psi}) = \frac{1}{2\rho^{\frac{1}{4}}\sqrt{\pi}}\frac{e^{M_\psi\rho^{\frac{3}{2}}}}{a_\psi}\, e^{\rho^{\frac{3}{2}}(-\frac{2i}{3}(4\cos^2\frac{\psi}{2}-1)\sin\frac{\psi}{2})}(1+O(\rho^{-\frac{3}{2}}))$$

as $\rho\to\infty$. Notice the exponential decay combined with the increasingly rapid oscillations because of the phase factor.

At first sight you might expect an $O(\rho^{-\frac{3}{4}})$ error estimate but since the exponential function in the integrand expands as

$$e^{-s^2+b_3\frac{p^3}{\rho^{\frac{3}{4}}}+b_4\frac{p^4}{\rho^{\frac{6}{4}}}+b_5\frac{p^5}{\rho^{\frac{9}{4}}}+\cdots} = e^{-s^2}(1+(b_3\frac{p^3}{\rho^{\frac{3}{4}}}+\dots)+\frac{1}{2}(b_3\frac{p^3}{\rho^{\frac{3}{4}}}+\dots)^2+\cdots)$$

the higher order terms in the expansion of $\mathrm{Ai}(\rho e^{i\psi})$ involve the integrals

$$\int s^n e^{-s^2}ds \quad (n=3,4,\dots)$$

of which the odd ones vanish. Therefore a contribution of the first b_3-term appears only in combination with the first order term in the expansion of $\varphi'_\psi(u_\psi;\psi)$ (the second order term in the expansion of $\varphi_\psi(u_\psi;\psi)$). It is an exercise to make the expansion more precise.

One has

$$M_\psi = -\frac{2}{3}(4\cos^2\frac{\psi}{2}-3)\cos\frac{\psi}{2},$$

and by direct but tedious calculation the stable manifold is given by

$$v = \varphi_\psi(u) = \varphi_\psi(-s+p) = \frac{-sc + p\sqrt{1 - \frac{4}{3}sp + \frac{1}{3}p^2}}{-s+p} \quad (c = \cos\frac{\psi}{2},\ s = \sin\frac{\psi}{2}).$$

You should recognise a discriminant under the square root, which for the level curve going through the saddle has the property that it is everywhere positive, except in the saddle point m_ψ. Expansion gives

$$\varphi_\psi(-s+p) = c + \frac{-1+c}{s}p + \frac{1}{3}\frac{2c-1}{c+1}p^2 + \dots.$$

For $\psi = 0$ one has

$$v = \varphi_0(u) = 1 + \sqrt{1 + \frac{1}{3}u^2} = 1 + \frac{1}{6}u^2 - \frac{1}{72}u^4 + \cdots$$

and

$$f(u, \varphi_0(u)) = -2(1 + \frac{4}{9}u^2)\sqrt{1 + \frac{1}{3}u^2} = -\frac{2}{3} - u^2 - \frac{5}{36}u^4 + \cdots,$$

so that

$$a_0 = 1,\ M_0 = -\frac{2}{3},\ A_0 = 0,$$

and

$$\mathrm{Ai}(\xi) \sim \frac{e^{-\frac{2}{3}\xi^{\frac{3}{2}}}}{2\sqrt{\pi}\xi^{\frac{1}{4}}}$$

as $\xi \to +\infty$, give or take a mistake in the constants, without oscillations.

Increasing ψ there are changes as ψ crosses $\frac{\pi}{3}$ and $\frac{2\pi}{3}$. For all $0 \le \psi < \frac{2\pi}{3}$ it still holds that

$$\mathrm{Ai}(\rho e^{i\psi}) = \frac{\rho^{\frac{1}{2}}}{2\pi}\int_{-\infty}^{\infty} e^{\rho^{\frac{3}{2}}(f(u,\varphi(u;\psi);\psi)}(1 + i\varphi'(u;\psi))du\ e^{\rho^{\frac{3}{2}}(-\frac{2i}{3}(4\cos^2\frac{\psi}{2}-1)\sin\frac{\psi}{2})}.$$

Figure 2 shows the relevant orbits for $\psi = \frac{15\pi}{24}$, with the same stable manifold defining the contour, and the same asymptotics still valid, but with a different sign for M_ψ, as Figure 3 shows. The sign change occurs at $\frac{\pi}{3}$. Thus for $\frac{\pi}{3} < \psi < \frac{2\pi}{3}$ there is exponential growth of $\mathrm{Ai}(\rho e^{i\psi})$ as $\rho \to \infty$, while the nonzero phase factor accounts for increasingly rapid oscillations.

At $\psi = \frac{2\pi}{3}$, when the growth is maximal (and no oscillations, see Figure 8), the diagram (and the Maple automatic colour coding) changes. All orbits in Figure 4 are in the stable or unstable manifolds of the saddle points.

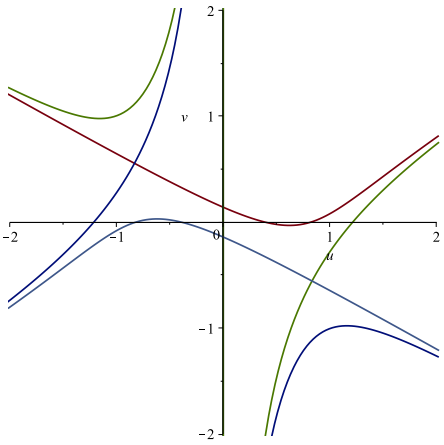


Figure 2: orbits for $\psi = \frac{15\pi}{24}$

The appropriate contour now consists of 3 orbits: the 2 orbits in the stable manifold of $m_{\frac{2\pi}{3}}$ (one of which is in the unstable manifold of $m_{\frac{8\pi}{3}}$), and one orbit in the stable manifold of $m_{\frac{8\pi}{3}}$. Can you see which one? You should convince yourself that $M_{\frac{8\pi}{3}}$ only enters the asymptotics beyond any relevant order.

Only as ψ is increased to $\psi = \pi$ both M_π and $M_{3\pi}$ are on par: $M_\pi = M_{3\pi} = 0$. The phases are then $\phi_\pi = \frac{2}{3}$ and $\phi_{3\pi} = -\frac{2}{3}$, and the two stable manifolds are given by

$$v = \frac{(u+1)\sqrt{u(u-2)}}{u\sqrt{3}} \quad (u < 0)\,, \quad v = \frac{(u-1)\sqrt{u(u+2)}}{u\sqrt{3}} \quad (u > 0).$$

You can now compute the expansion using both contours, with u running from $-\infty$ to 0 for the first integral and from 0 to ∞ for the second. Note the symmetry in Figure 7.

Observe that for $\frac{2\pi}{3} < \psi \leq \pi$ the contours are different. In Figures 5 you see the red curve turning blue after the turning point, and as it escapes to infinity along the negative v-axis it is joined by the green curve which is the stable manifold of the other saddle point. As in the case that $\psi = \pi$, the appropriate contour consists in fact of two contours: the sum of the integrals along both stable manifolds defines $\mathrm{Ai}(\rho e^{i\psi})$. For $\psi < \pi$ the main contribution comes from the contour on the left. Solving a cubic equation this contour can be written as a graph $u = \varphi(v)$, but the main contribution can be computed as above, still writing $v = \varphi(u)$ near the saddle point.

A similar program works for the solution of $u_t + \frac{1}{3}u_{xxx} = 0$ that starts

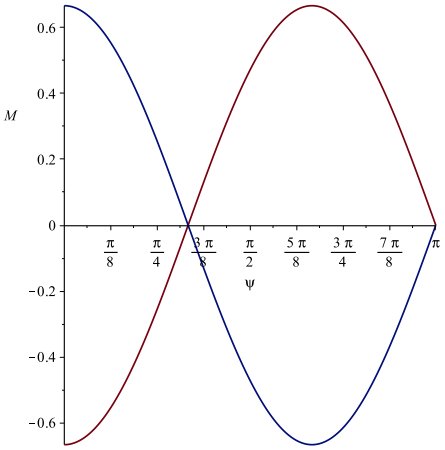


Figure 3: M_ψ and $M_{\psi+2\pi}$

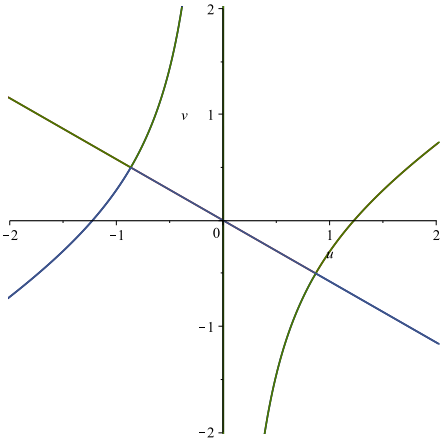


Figure 4: orbits for $\psi = \frac{16\pi}{24} = \frac{2\pi}{3}$

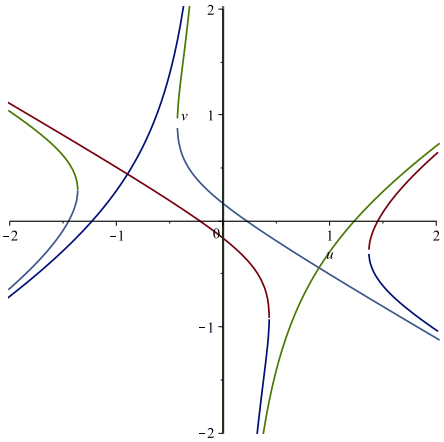


Figure 5: orbits for $\psi = \frac{17\pi}{24}$

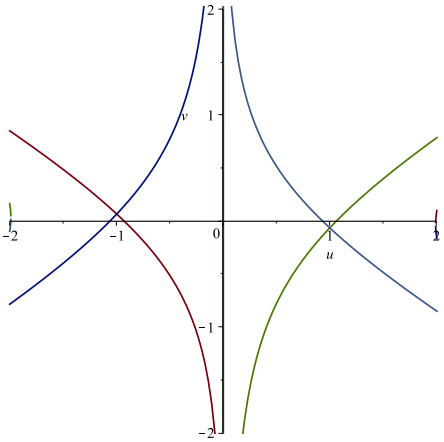


Figure 6: orbits for $\psi = \frac{23\pi}{24}$

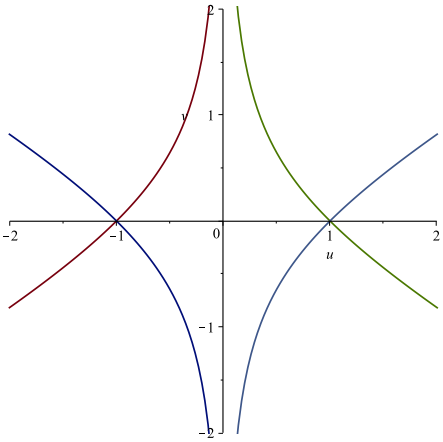


Figure 7: orbits (stable manifolds: blue curves) for $\psi = \frac{24\pi}{24} = \pi$

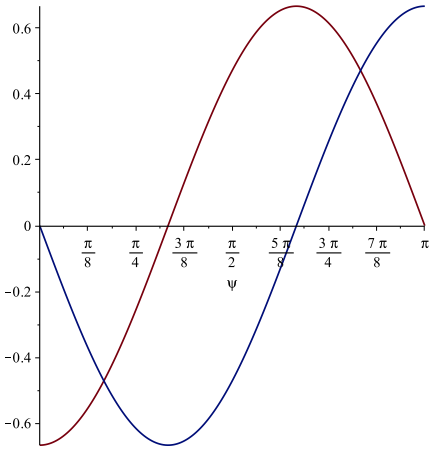


Figure 8: Amplitude M_ψ (red) and phase ϕ_ψ, changes at $\psi = 0, \frac{\pi}{3}, \frac{2\pi}{3}, \pi$.

from a "wave packet"

$$u_0(x) = e^{-\frac{x^2}{4a}} e^{ik_0 x},$$

along lines $x = ct + \xi$. One then has

$$u(t, ct+\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ik(c+\frac{1}{3}k^2)t + i\xi k - a(k-k_0)^2} dk = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t\Phi} dk.$$

Replace k by $z = x + iy$ (not the same x of course) and write

$$\Phi = \Phi(z) = \Phi(z;t) = \Phi(z;t,\xi,k_0,a) = f + ig$$

with

$$f = -ty(c + x^2 - \frac{1}{3}y^2) + a(-(x-k_0)^2 + y^2) - \xi y$$

and

$$g = tx(c + \frac{1}{3}x^2 - y^2) - 2a(x-k_0)y + \xi x.$$

Then as before on may rewrite

$$u(t, ct+\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\Phi(x;t)} dx = \frac{1}{\sqrt{2\pi}} \int_{\gamma} e^{\Phi(z;t)} dz$$

in which γ consists of orbits in stable manifolds of a suitable gradient flow of f, which is defined by

$$\dot{x} = \frac{dx}{d\tau} = \frac{1}{t}\frac{\partial f}{dx} = -2xy - \frac{2a}{t}(x-k_0),$$

$$\dot{y} = \frac{dy}{d\tau} = \frac{1}{t}\frac{\partial f}{dy} = -c - x^2 + y^2 + \frac{2ay - \xi}{t}.$$

Unlike in the analysis of the Airy function integral, there is now no need to scale x and y, because in the limit $t \to \infty$ the diagram in the x, y-plane is well defined. For $c = 1$ it is the same as in Figure 7 and for $c = -1$ it coincides with Figure 9 (with u, v replaced by x, y). Unlike the u, v-diagram the x, y-diagram varies with the parameter under consideration, as the role of ρ is now played by t. One computes the relevant unstable manifold(s) directly from solving $g = \phi$, which is a quadratic equation in y, asking that the discriminant

$$D = \frac{4}{3}x^2(x^2 + 3c)t^2 + 4x(\xi x - \phi)t + 4a^2(x-k_0)^2$$

of this equation is positive except in the saddle point, thus first determining simultaneously the saddle point and the phase ϕ by solving

$$D = \frac{dD}{dx} = 0.$$

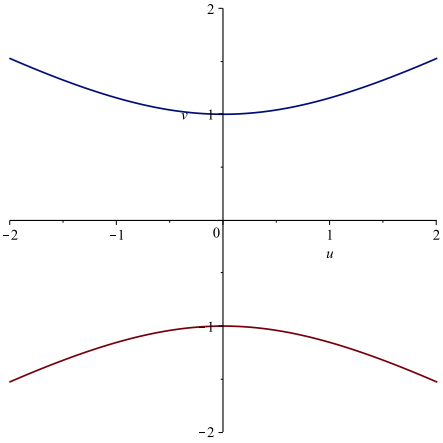


Figure 9: u, v-plane for $\psi = 0$, the vertical axis also consist of orbits

The phase ϕ then drops out of

$$x\frac{dD}{dx} - D = t^2x^4 + (a^2 + \xi t + ct^2)x^2 - k_0^2a^2 = 0,$$

which determines the square of the positive solution $x = x_c > 0$ uniquely in terms of the parameters t, c, ξ, a, the y-coordinate $y = y_c$. The phase ϕ_c and the value M_c of f in the saddle point (x_c, y_c) are then given by

$$y_c(t) = \frac{a(k_0 - x_c)}{x_c t}, \quad \phi_c(t) = -\frac{2tx_c^3}{3} - \frac{2a^2k_0(x_c - k_0)}{x_c t},$$

and

$$M_c(t) = -a(x_c - k_0)^2 - \frac{a^3(x_c + 2k_0)(x_c - k_0)^2}{3x_c^3t^2}.$$

For the values of y, ϕ, M in the other saddle point replace x_c by $-x_c$. Note that $x_c = x_c(t)$ and likewise for y_c, ϕ_c, M_c (the other dependencies are also suppressed in the notation). Observe the different behaviours as $t \to \infty$ for $c < 0$ and $c > 0$.

At this point I found it convenient to continue the calculations for the stable manifold with x_c implicitly defined by the quartic $x\frac{dD}{dx} - D$ and all other quantities explicitly in terms of x_c. With

$$x = x_c + u, \quad y = y_c + v,$$

the real and imaginary parts of Φ rewrite as

$$f = M_c + F_c, \quad F = F_c(t) = -t(2x_cuv + v(u^2 - \frac{v^2}{3})) - \frac{ak_0}{x_c}(u^2 - v^2),$$

$$g = \phi_c + G_c, \quad G = G_c(t) = t(x_c(u^2 - v^2) + u(\frac{u^2}{3} - v^2)) - \frac{2ak_0uv}{x_c},$$

the latter defining the stable manifold as the graph $v = \varphi_c(u) = \varphi_c(u;t)$ obtained from solving $G = 0$ for v, the discriminant having the desired behaviour: positive except for $u = 0$. For $c > 0$ this gives a globally defined function and deforming contours as before it follows that

$$u(t, ct + \xi) = \frac{e^{M_c(t) + i\phi_c(t)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{F_c(u,\varphi_c(u;t);t)}(1 + i\varphi_c'(u;t))du.$$

Note that c has disappeared completely from the formula's, except for the dependence through x_c.

The integral depends only on the formula's for F_c and G_c, φ_c being defined by solving $G = 0$, i.e.

$$(x_c + u)tv^2 + \frac{2ak_0u}{x_c}v - tu^2(x_c + \frac{u}{3}) = 0$$

and simplifying the discriminant using the quartic for x_c. This gives

$$v = \varphi_c(u) = \frac{u}{x_c(x_c + u)}\left(-\frac{ak_0}{t} + x_cR\right),$$

in which

$$R = \sqrt{\underbrace{c + 2x_c^2 + \frac{\xi}{t} + \frac{a^2}{t^2}}_{\text{positive}} + \frac{4x_cu}{3} + \frac{u^2}{3}}.$$

The derivative appears in the integral as

$$\varphi_c'(u) = -\frac{ak_0}{t(x_s + u)^2} + \frac{u(2x_c + u)}{3R(x_c + u)},$$

and the exponent in the integral rewrites as

$$F_c(u, \varphi_c(u;t);t) = -\frac{2}{9}\frac{tRu^2\,(3\,x_c + 2\,u)^2}{(x_c + u)^2} + \frac{2}{3}\frac{ak_0u^2\,(3\,x_c + 2\,u)}{(x_c + u)^2}$$

$$-\frac{2}{3}\frac{{k_0}^2a^2\,(Rx_ct - ak_0)\,u^2\,(3\,x_c + u)}{t^2{x_c}^3\,(x_c + u)^3}$$

Clearly these formula's suggest putting

$$u = x_cs, \quad k_0 = \frac{b}{a}, \quad t = b\tau, \quad \xi = b\eta$$

This scaling of u makes each term separate as far the integration variable and t are concerned, except for the R-terms. One now has to distinguish between $c < 0$ (the case discussed by Olver) when $u(t, ct + \xi)$ appears as the sum of 2 integrals involving the stable manifolds of both saddles, and $c > 0$, when $u(t, ct + \xi)$ appears as one single integral.

With x_c going to $\sqrt{-c}$ if $c < 0$, both integrals can be handled as in the Airy case, and the final expansion will depend on c. It may be handy to split the exponential in 3 separate exponentials before you proceed. From the c-dependence there should be a connection with the group velocity discussion by Olver, as we see below. On the other hand, when $c > 0$ (this case is not discussed by Olver) tx_c goes to a constant so $x_c \to 0$. Note that all 3 terms involve s^2, but with different signs. This is really an instructive example for understanding the method!

Now to back to WHY we did this analysis observe that the prefactor in the integral expression for

$$u(t, ct + \xi)$$

contains

$$e^{M_c}$$

which behaves very differently for $c < 0$ and $c > 0$.

For $c > 0$ it is the second term in the expression for $M_c(t)$ above that dominates and goes to infinity because tx_c goes to a constant, and this leading order term then goes to $-\infty$ linearly in t. Modulo the details of the analysis of the integral it follows that $u(t, ct + \xi) \to 0$ exponentially fast as $t \to \infty$.

On the other hand, if $c < 0$ then $x_c \to \sqrt{-c}$ and $M_c(t) \to -a(\sqrt{-c} - k_0)^2$ which is maximal and equal to zero for $c = -k_0^2$. Thus only for this value of c the solution is of order one along the line $x = ct + \xi$ as $t \to \infty$, with the more precise asymptotics following from a more detailed analysis of the integral, as in the Airy functions case, with contributions from both saddle points, and combining both phases and $\frac{2}{3}c^{\frac{3}{2}}t$ appearing in the imaginary part. Olver's point in the section about dispersion relations is that this *group velocity* $-c$ is 3 times larger as you would expect from looking at the single frequency solution with $a = 0$, and he did so by one single calculation starting from the dispersion relation. Read again what he did after the exam, and pay attention to the factor $\frac{1}{3}$ in the third order equation $u_t + \frac{1}{3}u_{xxx} = 0$ that I solved starting from a wave packet centered at $k = k_0$ rather than from a single wave with $k = k_0$.

1. This is an exercise about applying the Fourier transform to solve the equation $u_t + u_{xxx} = 0$ on the real line with initial data $u(0,x) = \delta(x)$, the Dirac δ-function, and to investigate the behaviour of the solution for $x \to -\infty$. The Fourier transform of a function $f = f(x)$ and the inverse transform are defined by

$$\hat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx} dx, \quad f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(k) e^{ikx} dk,$$

respectively whenever f and $\hat{f}$ are sufficiently nice. The improper integrals are to be understood in the principal value sense

$$\int_{-\infty}^{\infty} = \lim_{R \to \infty} \int_{-R}^{R}$$

and are often easiest evaluated using complex integration over appropriate contours.

   (a) Explain why $\hat{\delta}(k) = \frac{1}{\sqrt{2\pi}}$.

   (b) Show using integration by parts that $\widehat{(f')}(k) = ik\hat{f}(k)$.

   (c) Let u be a smooth solution of $u_t + u_{xxx} = 0$ which decays to zero sufficiently fast as $|x| \to \infty$ to have $(\hat{u})_t = \widehat{(u_t)}$. Here $\hat{u} = \hat{u}(t,k)$ denotes the Fourier transform of the function $x \to u(x,t)$. Denote the initial value of u by u_0, that is, $u_0(x) = u(0,x)$. Show that

$$\hat{u}(t,k) = \widehat{u}_0(k) e^{ik^3 t}$$

   (d) Show that the inversion formula formally applied to the case that $\widehat{u}_0(k) = \hat{\delta}(k) = \frac{1}{\sqrt{2\pi}}$ defines a solution formula

$$u(t,x) = \frac{1}{(3t)^{\frac{1}{3}}} \mathrm{Ai}\left(\frac{x}{(3t)^{\frac{1}{3}}}\right)$$

   in which

$$\mathrm{Ai}(\xi) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i(\xi x + \frac{x^3}{3})} dx.$$

   (e) Use the methods above to determine the asymptotic behaviour of $\mathrm{Ai}(\xi)$ for $\xi \to -\infty$.

43 Al of niet metrische topologie

Een aardig dictaatje is hier te vinden, uit de tijd dat Leiden de R nog in de naam had:

`http://www.few.vu.nl/~jhulshof/NOTES/anal.pdf`

Hieronder neem ik het over met wat aanvullingen en correcties:

43.1 Metrische ruimten; continue afbeeldingen

Aanvullend materiaal voor het college Analyse 1

J. Hulshof (destijds RUL)

Cursief wat opmerkingen van 24 jaar later ingevoegd tijdens het geven van het college Analyse 3 in het tweede jaar, met verwijzingen naar het boek Principles of Topology van Croom.

1. Inleiding. In deze syllabus behandelen we een aantal fundamentele onderwerpen uit de analyse. Uitgangspunt hierbij is de volgende algemene probleemstelling:

Laat X een puntverzameling zijn en A een niet-lege deelverzameling van X, eventueel X zelf. Zij $f: A \to \mathbb{R}$ een reëelwaardige functie. Hoe en onder wat voor veronderstellingen kunnen we dan concluderen dat de functie f op A een globaal maximum aanneemt, m.a.w. bestaat er een punt $x_0 \in A$, zo dat

$$f(x) \leq f(x_0) \quad \forall x \in A?$$

Om deze vraag te beantwoorden beschouwen we het supremum van f op A,

$$M = \sup\{f(x) : \ x \in A\} \ \in \ \mathbb{R} \cup \{+\infty\}.$$

Wat we van M nu willen weten is ten eerste of M eindig is, en zo ja, of de waarde M ook door de functie f wordt aangenomen. Omdat M het supremum is van alle door f aangenomen functiewaarden, kunnen we M benaderen met deze functiewaarden. We onderscheiden twee gevallen.

(i) $M < +\infty$. Dan bestaat er voor elk natuurlijk getal n een $x_n \in A$, zodat

$$f(x_n) > M - \frac{1}{n}.$$

(ii) $M = +\infty$. Dan bestaat er voor elk natuurlijk getal n een $x_n \in A$, zodat

$$f(x_n) > n.$$

In beide gevallen geeft dit ons een rij punten, genoteerd als $(x_n)_{n=1}^{\infty}$, in A. Als we nu kunnen concluderen dat, eventueel door een aantal van deze punten uit de rij weg te laten, deze punten voor grote waarden van $n \in \mathbb{N}$ steeds dichter komen te liggen bij een "limietpunt" dat zelf ook in A ligt, en dat de waarde van f in dat limietpunt gelijk is aan

$$\lim_{n\to\infty} f(x_n),$$

dan hebben we in één klap de beide bovenstaande vragen met ja beantwoord. In deze syllabus zullen we de voor bovenstaande probleemstelling relevante begrippen behandelen, met hier en daar een zijstapje.

In de vorige alinea staat de zinsnede "dichterbij". Aangezien we van X alleen maar hebben aangenomen dat X een puntverzameling is, heeft dit zonder verdere veronderstellingen geen betekenis. Een natuurlijke manier om dit de verhelpen is de invoering van een zogenaamde afstandsfunctie of metriek op X. Dit leidt dan tot de definitie van een metrische ruimte (Sectie 2). De deelverzamelingen waarvoor altijd een limietpunt van een rij bestaat blijken de zogenaamde rijkompakte verzamelingen te zijn (Sectie 3). Voor functies op (deelverzamelingen van) metrische ruimten kan het begrip "continu" worden gedefinieerd (Sectie 5), waarmee de bovenstaande limietovergang kan worden gerechtvaardigd. Als voorbeeld behandelen we de gevallen dat $X = \mathbb{R}$ en $X = \mathbb{R}^N$ (Sectie 6). In de appendix komen nog enige iets meer geavanceerde onderwerpen met betrekking tot kompaktheid aan de orde.

De schrijver van deze syllabus is van mening dat iedere student in de wiskunde of theoretische natuurkunde, onafhankelijk van wat hij/zij in de doctoraalfase als afstudeerrichting kiest, zich de basisstof in Sectie 1 tot en met 6 van deze syllabus moet eigen maken. In de meeste theoretische analyse boeken is deze stof, althans voor het geval $X = \mathbb{R}^N$, terug te vinden in de inleidende hoofdstukken. Zie bijvoorbeeld het boek "Mathematical Analysis, a modern approach to advanced calculus" van T.M. Apostol (Addison-Wesley 1957), waarin bijna alle analyse die in de eerste twee jaar van de studie aan de orde komt, is terug te vinden, of "Principles of Mathematical Analysis" van W. Rudin (McGraw Hill 1964). Voor meer algemene metrische (en topologische) ruimten zijn er o.a. de boeken "Topology and Normed Spaces" van G.J.O. Jameson (Wiley 1974) en "Introduction to Topology and Modern Analysis" van G.F. Simmons (McGraw Hill 1963).

Als voorkennis wordt verondersteld dat de lezer bekend is met de elementaire verzamelingsleer, begrippen als aftelbaar oneindig en overaftelbaar oneindig, en met de axioma's voor de natuurlijke getallen $\mathbb{N}$ en de reële getallen $\mathbb{R}$. Hoofdstuk 1 uit het boek "Calculus 1 2nd edition" van T.M. Apostol (Wiley 1967) is ruim voldoende.

2. Metrische ruimten. Laat X weer een puntverzameling zijn.

Definitie 2.1. Een functie $d: X \times X \to \mathbb{R}$ heet een *metriek* op X als
(i) $\forall x, y \in X$:

$$d(x,y) \geq 0 \text{ en } d(x,y) = 0 \iff x = y.$$

(ii) $\forall x, y \in X$:

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

(iii) $\forall x, y, z \in X$:

$$d(x,z) \leq d(x,y) + d(y,z) \quad \text{(driehoeksongelijkheid)}.$$

Als $d: X \times X \to \mathbb{R}$ een metriek is, dan heet het paar (X,d) een *metrische ruimte*.

Het is duidelijk dat op een verzameling X meerdere metrieken kunnen zijn gedefinieerd. Toch spreekt men vaak over de metrische ruimte X i.p.v. over de metrische ruimte (X,d). Aangenomen is dan dat over de stilzwijgend gemaakte keuze van de metriek d geen misverstand kan bestaan. In het vervolg is X nu steeds een metrische ruimte met metriek d.

Voorbeeld 2.2. $X = \mathbb{R}$ is de verzameling van de reële getallen, met $d(x,y) = |x-y|$.

Opmerking 2.3. Iedere deelverzameling van een metrische ruimte is met dezelfde metriek weer een metrische ruimte.

Definitie 2.4. Laat $A \subset X$.
(i) Een punt $a \in A$ heet een *inwendig punt* van A als

$$\varepsilon xists\delta > 0: \quad B_\delta(a) = \{x \in X: \ d(x,a) < \delta\} \ \subset \ A.$$

De verzameling $B_\delta(a)$ heet de open bol met straal δ en middelpunt a.
(ii) Een punt $a \in A$ heet een *geisoleerd punt* van A als

$$\exists \delta > 0: \quad B_\delta(a) \cap A = \{a\}.$$

(iii) Een punt $p \in X$ heet een *ophopingspunt* van A als

$$\forall \delta > 0 \ \exists a \in A: \quad a \neq p \text{ en } d(a,p) < \delta.$$

(iv) Als elk punt van A een inwendig punt van A is, dan heet A een *open deelverzameling* van X.
(v) Als het complement van A,

$$A^c = X - A = \{x \in X : x \notin A\},$$

open is, dan heet A een *gesloten deelverzameling* van X.

In het vervolg zullen we kortweg zeggen dat A open (gesloten) is als A een open (gesloten) deelverzameling van X is.

LET OP! *Definitie 2.4 benoemt eigenschappen van punten $a \in A$ en $p \in X$ met $A \subset X$ en X een metrische ruimte. Maar A is zelf ook weer een metrische ruimte en dus te zien in de rol van X hierboven. Je kunt de grotere X dan verder vergeten bij het doen van uitspraken over deelverzamelingen van A. Bijvoorbeeld over $A \subset A$, net zoals je uitspraken over X kunt doen, gezien als deelverzameling van de metrische ruimte X.*

Voorbeeld: *De gehele getallen vormen een verzameling waarvoor geldt dat $\mathbb{Z} \subset \mathbb{R}$. In $\mathbb{Z}$ is elk punt een geisoleerd punt en elke bol met straat $\frac{1}{2}$ een singleton, dus $\{0\}$ is een open deelverzameling van $\mathbb{Z}$ maar geen open deelverzameling van $\mathbb{R}$.*

Stelling 2.5. (i) Iedere open bol is een open verzameling.
(ii) Een verzameling A in X is open dan en slechts dan als A een vereniging van open bollen is.

Bewijs. Opgave.

Stelling 2.6. (i) X is open, de lege verzameling is open.
(ii) De vereniging van elke collectie open deelverzamelingen is open.
(iii) De doorsnede van elk eindig aantal open deelverzamelingen is open.

Bewijs. Opgave.

Stelling 2.7. (i) X is gesloten, de lege verzameling is gesloten.
(ii) De doorsnede van elke collectie gesloten deelverzamelingen is gesloten.
(iii) De vereniging van elk eindig aantal gesloten verzamelingen is gesloten.

Bewijs. Opgave.

Stelling 2.8. Laat $A \subset X$. Dan is A gesloten dan en slechts dan als A al zijn ophopingspunten bevat.

Bewijs. Neem eerst aan dat A gesloten is en laat p een ophopingspunt zijn van A. We moeten laten zien dat $p \in A$. Stel niet. Dan $p \in A^c$. Maar A^c is open, dus p is een inwendig punt van A^c. Zodoende is er een $\delta > 0$ waarvoor

$B_\delta(p) \subset A^c$, in tegenspraak met de veronderstelling dat p een ophopingspunt is van A. Dus p ligt wel in A.

Neem vervolgens aan dat A een verzameling is die al zijn ophopingspunten bevat. We tonen aan dat A gesloten is door te bewijzen dat A^c open is. Zij $p \in A^c$. Dan is p geen ophopingspunt van A, dus er is een $\delta > 0$ zo dat $B_\delta(p)$ geen punten van $A - \{p\}$ bevat, en wegens $p \in A^c$ betekent dit dat $B_\delta(p) \subset A^c$. m.a.w. p is een inwendig punt van A^c. Dit geldt voor elke $p \in A^c$, en dus is A^c open. Q.e.d.

Definitie 2.9. Een rij $(x_n)_{n=1}^\infty$ in X heet *convergent* als er een $x_0 \in X$ is zo dat

$$\forall \varepsilon > 0 \; \exists n_\varepsilon \in \mathbb{N} : \quad n \geq n_\varepsilon \implies d(x_n, x_0) < \varepsilon.$$

Het punt x_0 heet de *limiet* van de rij.

Stilzwijgend zijn n, n_ε, m en met andere letters in het midden van het alphabet genoteerde variabelen vaak elementen van $\mathbb{N}$. De definitie wordt vaak geschreven zonder de subindex ε:

$$\forall \varepsilon > 0 \; \exists N : \quad n \geq N \implies d(x_n, x_0) < \varepsilon.$$

Stelling 2.10. Als de rij $(x_n)_{n=1}^\infty$ in X convergent is, dan is de limiet x_0 eenduidig bepaald, notatie

$$\lim_{n\to\infty} x_n = x_0 \quad \text{of} \quad x_n \to x_0.$$

Bewijs. Neem aan dat de rij twee verschillende limieten heeft, zeg x_0 en x_0'. Omdat $x_0 \neq x_0'$, is $d(x_0, x_0') > 0$. Kies $0 < \varepsilon < \frac{1}{2} d(x_0, x_0')$. Dan is er een n_ε zo dat $d(x_n, x_0) < \varepsilon$ voor alle $n \geq n_\varepsilon$. Evenzo is er een n_ε' zo dat $d(x_n, x_0') < \varepsilon$ voor alle $n \geq n_\varepsilon'$. Met behulp van de driehoeksongelijkheid volgt nu

$$d(x_0, x_0') \leq d(x_0, x_n) + d(x_n, x_0') < \varepsilon + \varepsilon < d(x_0, x_0').$$

voor $n \geq \max(n_\varepsilon, n_e')$, tegenspraak. Q.e.d.

In plaats van "de rij $(x_n)_{n=1}^\infty$ is convergent", zegt men ook wel dat "$\lim_{n\to\infty} x_n$ bestaat".

Stelling 2.11. Als de rij $(x_n)_{n=1}^\infty$ in X convergent is, dan is de rij begrensd, d.w.z. bevat in een vaste (open) bol $B_\delta(a) \subset X$.

Bewijs. Opgave.

Stelling 2.12. Laat $A \subset X$ gesloten zijn. Als de rij $(x_n)_{n=1}^\infty$ in A convergent is in X, dan ligt de limiet in A.

Bewijs. Opgave.

Stelling 2.13. Laat $A \subset X$ en $p \in X$. De volgende drie uitspraken zijn equivalent:
(i) p is een ophopingspunt van A.
(ii) Er bestaat een rij $(a_n)_{n=1}^{\infty}$ in $A - \{p\} = \{x \in A : \ x \neq p\}$ die convergent is met p als limiet.
(iii) Iedere (niet-lege) open bol met middelpunt p bevat oneindig veel punten van A.

Bewijs. $(i \implies ii)$. Neem aan dat p een ophopingspunt van A is. We construeren de rij $(a_n)_{n=1}^{\infty}$ in A. Kies $\varepsilon_1 > 0$. Dan is er een $a_1 \in A - \{p\}$ met $d(a_1, p) < \varepsilon_1$. De rij $(a_n)_{n=1}^{\infty}$ wordt nu verder inductief gedefinieerd door voor $n = 1, 2, \ldots$, nadat a_n gekozen is, $\varepsilon_{n+1} = \frac{1}{2}d(a_n, p)$ te stellen, en $a_{n+1} \in A - \{p\}$ met $d(a_{n+1}, p) < \varepsilon_{n+1}$ te kiezen. Opgave: laat zien dat $\varepsilon_{n+1} < \frac{1}{2}\varepsilon_n$ en dat de rij naar p convergeert.
$(ii \implies iii)$. Opgave.
$(iii \implies i)$. Triviaal.

Correctie! *In de 1993-versie stond in Stelling 2.13 en in het bewijs daarvan hier en daar nog een a waar een p moest staan.*
Het gebruik van definities met de quantor voor alle ($\forall$). *Belangrijk is om te onthouden dat een definitie die begint met $\forall \delta > 0$ geldt "iets" waarbij δ een rol speelt, pas echt gebruikt is als voor een rij $\delta_n \downarrow 0$ dat "iets" gebruikt is. En in plaats van δ's kunnen natuurlijk ook ε's gebruikt worden.*
Variaties op het bewijs. *Het dalend kiezen van de rij $\varepsilon_n > 0$ in $(i \implies ii)$ kan natuurlijk op vele manieren. Met $\varepsilon_1 = 1$ en vervolgens*

$$\varepsilon_{n+1} = \min(d(a_n, p), \frac{1}{n+1})$$

voor $\varepsilon_2, \varepsilon_3, \ldots$ werkt het bewijs net zo goed. Met die constructie volgt dan meteen dat $\varepsilon_n \leq \frac{1}{n}$. In het als opgave te geven bewijs van $a_n \to p$ moet je voor alle $\varepsilon > 0$ een N vinden zoals onder Definitie 2.9. Kies daartoe N met $\frac{1}{N} < \varepsilon$ (waarom kan dat?) en gebruik de bijbehorende ε_N gedefinieerd zoals hier direct boven.
De Archimedische eigenschap van de verzameling van de reële getallen. *Waarom bestaat er voor elke ε eigenlijk een n met $\frac{1}{n} < \varepsilon$? Wel, indien niet dan zou er een $\varepsilon > 0$ zijn met $\frac{1}{n} \geq \varepsilon$ en dus $n \leq \frac{1}{\varepsilon}$ voor alle $n \in \mathbb{N}$. De (niet-lege) verzameling $\mathbb{N}$ zou dan naar boven begrensd zijn in $\mathbb{R}$ en in $\mathbb{R}$ een kleinste bovengrens S hebben. Dan is $S - \frac{1}{2}$ geen bovengrens en dus bestaat er een $N \in \mathbb{N}$ met $S - \frac{1}{2} < N \leq S$ en bijgevolg $N + 1 > S + \frac{1}{2}$. Om dat $N + 1 \in \mathbb{N}$ kan S dus geen bovengrens zijn van $\mathbb{N}$, laat staan de kleinste.*

Zonder epsilons kan het dus ook. *De nu bewezen uitspraak dat onder elke $\varepsilon > 0$ altijd een $\frac{1}{n}$ zit met $n \in \mathbb{N}$ wordt de Archimedische eigenschap van $\mathbb{R}$ genoemd en maakt dat iedere definitie die begint met $\forall \varepsilon > 0$ en eindigt met $< \varepsilon$ kan worden vervangen door een definitie die begint met $\forall n \in \mathbb{N}$ en eindigt met $< \frac{1}{n}$. Alleen komen we dan al snel letters in het midden van het alphabet te kort.*

Om welk axioma voor, of eigenschap van de verzameling van de reële getallen ging het? *De Archimedische eigenschap geldt dankzij het axioma over het bestaan van kleinste bovengrenzen in $\mathbb{R}$ voor naar bovengrensde niet-lege deelverzamelingen van $\mathbb{R}$. In het boek van Croom wordt dit besproken in Sectie 2.1.*

Croom *gebruikt in zijn Sectie 3.2 het woord limietpunt als ander woord voor ophopingspunt. Ik reserveer de term limietpunt voor het gebruik zoals in Definitie 3.1 hieronder: limiet van een convergente deelrij van een gegeven rij. Dus rijen kunnen limietpunten hebben en verzamelingen ophopingspunten. Een rij in (A of) X is strict genomen ook geen deelverzameling van (A of) X maar een functie of afbeelding van $\mathbb{N}$ naar (A of) X. Croom's index verwijst voor limit point naar pagina 66 waar Stelling 3.6 komt na de definitie onderaan pagina 65, en daar zie je dat limietpunt bij Croom een andere naam is voor ophopingspunt. Een naamgeving wellicht verdedigbaar door de uitspraak dat bij een ophopingspunt p van A in X altijd een rij $(a_n)_{n=1}^{\infty}$ te vinden is waarvoor $d(a_n, p)$ strict dalend is in n en convergeert naar 0. Merk op dat voor $n \to \infty$ de equivalentie*

$$a_n \to p \iff d(a_n, p) \to 0$$

vanuit de definitie van convergentie vanzelfsprekend is.

Definitie A-1. *Laat weer $A \subset X$ met X een metrische ruimte. De afsluiting van A is de vereniging van A met al zijn ophopingspunten, notatie $\bar{A}$. Dus $p \in \bar{A}$ betekent dat de implicatie*

$$p \notin A \implies p \text{ is een ophopingspunt van } A$$

moet gelden.

Opgave A-2. *Bewijs dat $\bar{A}$ gesloten is in X. Hint: bewijs dat een ophopingspunt van $\bar{A}$ ook een ophopingspunt van A is gebruik Stelling 2.8.*

Opgave A-3. *De sterkere uitspraak is dat $\bar{A}$ de kleinste gesloten verzameling in X is die A bevat: bewijs dat $\bar{A}$ de doorsnede is van alle gesloten deelverzamelingen F van X met $A \subset F$. Hint: die doorsnede D is van vanwege Stelling 2.7 gesloten dus vanwege Opgave A-2 is $D \subset \bar{A}$. Kan A ophopingspunten hebben die niet in die doorsnede liggen?*

Croom noemt de verzameling van alle ophopingspunten de afgeleide verzameling van A, zonder verdere notatie, en schrijft $cl(A)$ voor $\bar{A}$.

We merken nog eens op dat de bovenstaande begrippen in principe afhangen van de keuze van de metriek op X. Toch kunnen verschillende metrieken tot hetzelfde leiden.

Definitie 2.14. Laat X een metrische ruimte zijn met metriek d, en laat $\bar{d}$ een andere metriek op X zijn. Dan heten d en $\bar{d}$ *equivalent* op X als er een reëel getal $\lambda > 0$ bestaat zo dat

$$\frac{1}{\lambda} d(x,y) \leq \bar{d}(x,y) \leq \lambda d(x,y) \quad \forall x, y \in X.$$

Opgave 2.15. Laat zien dat bij overgang op een equivalente metriek op X de in deze Sectie geintroduceerde begrippen (open, gesloten, convergent, etc.) niet veranderen. Hetzelfde geldt voor de begrippen die in de volgende drie secties worden behandeld (rijkompaktheid, volledigheid en continuiteit).

Op grond van het voorafgaande is het duidelijk dat voor het welslagen van de in de inleiding geschetste bewijsmethode, de geslotenheid van A vereist is. Dit staat echter los van de vraag of de in de inleiding geconstrueerde rij een limietpunt heeft.

3. Rijkompakte verzamelingen. We gaan nu de deelverzamelingen A karakteriseren waarvoor de in de inleiding geschetste bewijsmethode zal slagen.

Definitie 3.1. Laat $(x_n)_{n=1}^{\infty}$ een rij zijn in X.
(i) Als $(n_k)_{k=1}^{\infty}$ een strict stijgende rij natuurlijke getallen is, dan heet de rij $(x_{n_k})_{k=1}^{\infty}$ een *deelrij* van $(x_n)_{n=1}^{\infty}$.
(ii) Als $x_0 \in X$ de limiet is van een convergente deelrij van $(x_n)_{n=1}^{\infty}$, dan heet x_0 een *limietpunt* (ook wel: rijophopingspunt) van $(x_n)_{n=1}^{\infty}$.

Stelling 3.2. Een rij $(x_n)_{n=1}^{\infty}$ in X heeft een convergente deelrij met limiet x_0 dan en slechts dan als

$$\forall \varepsilon > 0 \ \forall n \ \exists k \geq n : \quad d(x_k, x_0) < \varepsilon.$$

Bewijs. Opgave.

Definitie 3.3. Laat $A \subset X$. A heet *rijkompakt* als iedere rij in A een limietpunt in A heeft.

Er is nog een andere definitie van kompaktheid die niet uitgaat van de metriek op X, maar van de open deelverzamelingen van X. In de appendix

zullen we deze definitie behandelen, en laten zien dat de beide definities voor deelverzamelingen van metrische ruimten hetzelfde betekenen.

Stelling 3.4. Gesloten deelverzamelingen van rijkompakte verzamelingen zijn rijkompakt.

Bewijs. Merk eerst op: als $G \subset A \subset X$, dan kan het gesloten zijn van G op twee manieren worden opgevat: gesloten in A of gesloten in X. Opgave: laat zien dat, als A gesloten is in X, dan

$$G \text{ is gesloten in } A \iff G \text{ is gesloten in } X.$$

Stel dat $(a_n)_{n=1}^{\infty}$ een rij is in G. Omdat A rijkompakt is, bestaat er een convergente deelrij met limiet in A. Daar G gesloten is ligt de limiet in G. Conclusie: elke rij in G heeft een convergente deelrij met limiet in G. Q.e.d.

Stelling 3.5. (i) Laat (X_1, d_1) en (X_2, d_2) twee metrische ruimten zijn. Dan is het Cartesisch produkt X van X_1 en X_2,

$$X = X_1 \times X_2 = \{(x_1, x_2) : \quad x_1 \in X_1,\ x_2 \in X_2\},$$

weer een metrische ruimte t.a.v. de metriek gedefinieerd door

$$d(x, y) = d_1(x_1, y_1) + d_2(x_2, y_2) \ \ \forall x = (x_1, x_2),\ y = (y_1, y_2) \in X = X_1 \times X_2.$$

(ii) Als A_1 rijkompakt is in X_1 en A_2 rijkompakt is in X_2, dan is $A = A_1 \times A_2$ rijkompakt in $X = X_1 \times X_2$.

Bewijs. Opgave.

Stelling 3.6. Laat A een rijkompakte deelverzameling zijn van X. Dan is A gesloten en begrensd.

Bewijs. Opgave.

4. Volledigheid.

Definitie 4.1. Een rij $(x_n)_{n=1}^{\infty}$ in X heet *een Cauchyrij* als

$$\forall \varepsilon > 0\ \exists n_\varepsilon : \quad m, n \geq n_\varepsilon \implies d(x_m, x_n) < \varepsilon.$$

Stelling 4.2. Iedere convergente rij is een Cauchyrij.

Bewijs. Zij $(x_n)_{n=1}^{\infty}$ een convergente deelrij met limiet x_0. Laat $\varepsilon > 0$, en gebruik de definitie van convergentie met $\frac{1}{2}\varepsilon$. Dan,

$$\forall m, n \geq n_{\frac{1}{2}\varepsilon} : \quad d(x_m, x_n) \leq d(x_m, x_0) + d(x_0, x_n) < \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon.$$

$\varepsilon > 0$ was willekeurig, dus de rij is een Cauchyrij. Q.e.d.

Stelling 4.3. Iedere Cauchyrij is begrensd.

Bewijs. Opgave.

Stelling 4.4. Als een Cauchyrij een limietpunt heeft, dan is de Cauchyrij convergent.

Bewijs. Laat $(x_n)_{n=1}^{\infty}$ een Cauchrij zijn die een convergente deelrij heeft. Laat $\varepsilon > 0$ en gebruik de definitie van Cauchyrij. Dus

$$\forall m, n \geq n_\varepsilon : \quad d(x_m, x_n) < \frac{1}{2}\varepsilon.$$

Anderzijds, omdat de rij een convergente deelrij heeft, zeg met limiet x_0, bestaat er een $k_\varepsilon > n_\varepsilon$ zodat

$$d(x_{k_\varepsilon}, x_0) < \frac{1}{2}\varepsilon.$$

Beide ongelijkheden combinerend vinden we

$$\forall n \geq n_\varepsilon : \quad d(x_n, x_0) \leq d(x_n, x_{k_\varepsilon}) + d(x_{k_\varepsilon}, x_0) < \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon.$$

Definitie 4.5. Als iedere Cauchyrij in X convergent is, dan heet X *volledig*.

Om na te gaan of een rij $(x_n)_{n=1}^{\infty}$ in een volledige metrische ruimte X convergent is, hoeft men slechts na te gaan dat de rij een Cauchyrij is. Daarbij is het niet nodig om a priori de limietwaarde te kennen. Het bewijs van een belangrijke stelling in de analyse, de Banach contractie stelling, berust op dit principe:

Stelling 4.6. Laat X een volledige metrische ruimte zijn, en $T : X \to X$ een contractie, d.w.z. een afbeelding met de eigenschap dat

$$\exists \theta \in [0, 1) \ \forall x, y \in X \quad d(T(x), T(y)) \leq \theta d(x, y).$$

Dan is er precies één vast punt van T, d.w.z. een punt $\bar{x}$ in X met de eigenschap dat $T(\bar{x}) = \bar{x}$. Bovendien geldt voor elke $x \in X$ dat de rij gedefinieerd door

$$x_1 = T(x), \quad x_{n+1} = T(x_n) \ (n = 1, 2, \dots),$$

convergent is met limiet $\bar{x}$.

Bewijs. Merk eerst op dat er hoogstens een vast punt kan zijn, immers als $\bar{x}$ en $\bar{y}$ vaste punten zijn, dan

$$d(\bar{x}, \bar{y}) = d(T(\bar{x}), T(\bar{y})) \leq \theta d(\bar{x}, \bar{y}) \implies d(\bar{x}, \bar{y}) = 0 \implies \bar{x} = \bar{y}.$$

Laat nu $x \in X$ willekeurig en laat de rij $(x_n)_{n=1}^{\infty}$ de in de stelling gedefinieerde rij zijn. We gaan bewijzen dat dit een Cauchyrij is. Neem hiertoe twee natuurlijke getallen m, n met $m < n$. Dan, door herhaald toepassen van de driehoeksongelijkheid,

$$d(x_m, x_n) \leq d(x_m, x_{m+1}) + d(x_{m+1}, x_{m+2}) + \cdots + d(x_{n-1}, x_n).$$

We gebruiken de notatie

$$T^0(x) = x, \quad T^1(x) = T(x), \quad T^2(x) = T(T(x)), \quad T^3(x) = T(T(T(x))), \quad \text{etc.}$$

Omdat

$$d(x_k, x_{k+1}) = d(T^k(x), T^{k+1}(x)) \leq \theta^k d(x, T(x)),$$

volgt nu dat

$$d(x_m, x_n) \leq (\theta^m + \theta^{m+1} + \cdots + \theta^{n-1}) d(x, T(x)) \leq \frac{\theta^m}{1-\theta} d(x, T(x)) \to 0 \quad \text{als } m \to \infty.$$

Dus $(x_n)_{n=1}^{\infty}$ is een Cauchyrij. Omdat X volledig is heeft deze rij een limiet $\bar{x}$, dus $T^n(x) \to \bar{x}$. Maar dan geldt ook dat $T^{n+1}(x) \to \bar{x}$, terwijl

$$d(T^{n+1}(x), T(\bar{x})) = d(T(T^n(x)), T(\bar{x})) \leq \theta d(T^n(x), \bar{x}) \leq d(T^n(x), \bar{x}) \to 0.$$

Dit impliceert dat $T^{n+1}(x) \to T(\bar{x})$. Omdat de limiet van een convergente rij uniek bepaald is, kunnen we dus concluderen dat $\bar{x} = T(\bar{x})$. Q.e.d.

Er is nog een andere karakterisatie van het begrip volledigheid, die wordt gegeven door de volgende stelling.

Stelling 4.7. (Cantor) Een metrische ruimte X is volledig dan en slechts dan als voor elke dalende rij gesloten deelverzamelingen

$$F_1 \supset F_2 \supset F_3 \ldots,$$

met de eigenschap dat

$$\operatorname{diam}(F_n) = \sup_{x,y \in F_n} d(x, y) \to 0 \text{ als } n \to \infty,$$

geldt dat de doorsnede

$$\cap_{n=1}^{\infty} F_n$$

precies één punt bevat.

5. Continue afbeeldingen.

Definitie 5.1. Laat X en Y metrische ruimten zijn, $A \subset X$, en $f : A \to Y$ een afbeelding. (i) f heet *continu in* $a \in A$, als

$$\forall \varepsilon > 0 \; \exists \delta > 0 \; \forall x \in A : \quad d(x,a) < \delta \implies d(f(x), f(a)) < \varepsilon.$$

(ii) f heet continu in A als f continu is in elke $a \in A$.

Stelling 5.2. Laat X en Y metrische ruimten zijn, $A \subset X$, $f : A \to Y$ een afbeelding, en $a \in A$. Dan is f continu in a dan en slechts dan als voor elke rij $(a_n)_{n=1}^{\infty}$ in A met $a_n \to a$ in X geldt dat $f(a_n) \to f(a)$ in Y.

Bewijs. Neem aan dat f continu is in a en laat $(a_n)_{n=1}^{\infty}$ een rij zijn in A met $a_n \to a$. Kies $\varepsilon > 0$ willekeurig, en laat $\delta > 0$ de bijbehorende δ uit de definitie van continuiteit van f in a zijn. Gebruik deze δ nu als ε in de definitie van convergentie. Dan

$$n \geq n_\delta \implies d(a_n, a) < \delta \implies d(f(a_n), f(a)) < \varepsilon.$$

Dus de rij $(f(a_n))_{n=1}^{\infty}$ convergeert naar $f(a)$.

Anderzijds, als f niet continu is in a, dan

$$\exists \varepsilon > 0 \; \forall \delta > 0 \; \exists a_\delta \in A : \quad d(a_\delta, a) < \delta \text{ en } d(f(a_\delta), f(a)) \geq \varepsilon.$$

Kies nu $\delta = \frac{1}{n}$, en noem de bijbehorende a_δ nu a_n, dan is $(a_n)_{n=1}^{\infty}$ een rij in A met $a_n \to a$, terwijl de rij $(f(a_n))_{n=1}^{\infty}$ niet convergeert naar $f(a)$. Q.e.d.

Stelling 5.3. Laat X en Y metrische ruimten zijn, en $f : X \to Y$ een afbeelding. Dan is f continu op X dan en slechts dan als het inverse beeld onder f van iedere open deelverzameling van Y weer een open deelverzameling van X is.

Bewijs. Neem aan dat f continu is, en zij B een open verzameling in Y. We moeten bewijzen dat het inverse beeld onder f van B open is. We laten zien dat elk punt van $A = f^{-1}(B)$ een inwendig punt is. Laat hiertoe $a \in A$ en $b = f(a) \in B$. Omdat B open is, is b een inwendig punt van B, dus er bestaat een $\varepsilon > 0$ met $B_\varepsilon(b) \subset B$. Kies de bij ε horende δ uit de definitie van continuiteit. Dan $f(B_\delta(a)) \subset B_\varepsilon(b)$ waardoor $B_\delta(a) \subset A = f^{-1}(B)$. Dit geldt voor elke $a \in A$, dus A is open.

Omgekeerd, als het inverse beeld van elke open verzameling open is, is te bewijzen dat f continu is in elk punt van X. Laat hiertoe $a \in X$ en $b = f(a)$, en zij $\varepsilon > 0$ willekeurig. Dan is $B_\varepsilon(b)$ open in Y, dus $A = f^{-1}(B_\varepsilon(b))$ is open in X, en omdat $a \in A$, is er een $\delta > 0$ zo dat $B_\delta(a) \subset A$. Met deze δ is dan aan de uitspraak in de definitie van continuiteit voldaan. Q.e.d.

Tot nu toe hebben we gesproken over afbeeldingen $f : X \to Y$. Vaak worden afbeeldingen ook functies genoemd, met name in het geval dat $Y = \mathbb{R}$. We keren nu terug naar de vraagstelling in de inleiding.

Stelling 5.4. Laat $A \subset X$, en $f : A \to \mathbb{R}$ een continue functie. Als A rijkompakt is, dan heeft f een maximum in A.

Bewijs. Met dit bewijs waren we al begonnen in de inleiding. Dus laat $(x_n)_{n=1}^{\infty}$ de rij zijn waarvan de functiewaarden het supremum M van f op A benaderen, zoals precies gemaakt in de inleiding. Omdat A rijkompakt is, heeft deze rij een convergente deelrij $(x_{n_k})_{k=1}^{\infty}$ met limiet $a \in A$. Vanwege de continuiteit van f in a is de rij $(f(x_{n_k}))_{k=1}^{\infty}$ convergent met limiet $f(a)$. Omdat een convergente rij begrensd is sluit dit de mogelijkheid $M = +\infty$ uit, zo dat

$$M - \frac{1}{n} < f(x_n) \leq M.$$

Dit impliceert dat de rij $(f(x_n))_{n=1}^{\infty}$ convergeert naar M. Maar een deelrij convergeert naar $f(a)$. Dus $f(a) = M$. Q.e.d.

Stelling 5.5. Laat X en Y metrische ruimten zijn, $A \subset X$, en $f : A \to Y$ een continue afbeelding. Als A rijkompakt is in X, dan is het beeld van A onder f,

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

rijkompakt in Y.

Bewijs. Opgave.

Definitie 5.6. Laat X en Y metrische ruimten zijn, $A \subset X$, en $f : A \to Y$ een afbeelding. Dan heet f *uniform continu in* A, als

$$\forall \varepsilon > 0 \; \exists \delta > 0 \; \forall x, y \in A : \quad d(x, y) < \delta \implies d(f(x), f(y)) < \varepsilon.$$

Stelling 5.7. Laat X en Y metrische ruimten zijn, $A \subset X$, en $f : A \to Y$ een continue afbeelding. Als A rijkompakt is in X, dan is f uniform continu in A.

Bewijs. Stel niet, dan

$$\exists \varepsilon > 0 \; \forall \delta > 0 \; \exists x, y \in A : \quad d(x, y) < \delta \;\text{ en }\; d(f(x), f(y)) \geq \varepsilon.$$

Kies $\delta = \frac{1}{n}$ en laat x_n en y_n de bijbehorende x en y zijn zoals in de regel hierboven. Omdat A rijkompakt is heeft de rij $(x_n)_{n=1}^{\infty}$ een convergente deelrij $(x_{n_k})_{k=1}^{\infty}$, zeg met limiet $a \in A$. Dan is ook de rij $(y_{n_k})_{k=1}^{\infty}$ convergent met dezelfde limiet. (Waarom?) Maar nu geldt

$$\lim_{k \to \infty} f(x_{n_k}) = \lim_{k \to \infty} f(y_{n_k}) = f(a),$$

terwijl $d(f(x_{n_k}), f(y_{n_k})) \geq \varepsilon$, tegenspraak. Q.e.d.

6. Het geval $X = \mathbb{R}$ en $X = \mathbb{R}^N$. We hebben gezien dat rijkompakte deelverzamelingen van X altijd gesloten en begrensd zijn. Als $X = \mathbb{R}$, geldt ook het omgekeerde. We laten dit eerst zien voor een gesloten begrensd interval.

Stelling 6.1. (Heine-Borel) Laat $a, b \in \mathbb{R}$. Dan is het gesloten begrensde interval

$$[a, b] = \{x \in \mathbb{R} : \quad a \leq x \leq b\}$$

rijkompakt.

Bewijs. Laat $(x_n)_{n=1}^{\infty}$ een rij zijn in $[a, b]$. We moeten bewijzen dat deze rij een convergente deelrij heeft met limiet in $[a, b]$. We delen hiertoe het interval in twee gelijke stukken, d.w.z. in

$$[a, \frac{a+b}{2}] \text{ en } [\frac{a+b}{2}, b].$$

Dan bevat tenminste één van deze twee intervallen voor oneindig veel waarden van n het rijelement x_n. Als we dit interval $[a_1, b_1]$ noemen, kunnen we dus een deelrij van $(x_n)_{n=1}^{\infty}$ kiezen die volledig bevat is in $[a_1, b_1]$. Noteer deze rij als $(x_n^1)_{n=1}^{\infty}$. Dit argument herhalende, krijgen we een dalende rij intervallen

$$[a, b] \supset [a_1, b_1] \supset [a_2, b_2] \supset [a_3, b_3] \supset \ldots,$$

en bijbehorende steeds verdere deelrijen, genoteerd als $(x_n^j)_{n=1}^{\infty}$, met dezelfde eigenschap, namenlijk dat elke deelrij $(x_n^j)_{n=1}^{\infty}$ steeds volledig bevat is in $[a_j, b_j]$.

Vervolgens nemen we de zogenaamde diagonaalrij, dat is de rij $(x_n^n)_{n=1}^{\infty}$. Dit is zelf weer een deelrij van de oorspronkelijke rij $(x_n)_{n=1}^{\infty}$, en er geldt dat $x_n^n \in [a_n, b_n]$.

Nu is $(a_n)_{n=1}^{\infty}$ een begrensde niet-dalende rij getallen, en $(b_n)_{n=1}^{\infty}$ een begrensde niet-stijgende rij getallen. Dus bestaan

$$\alpha = \sup_n a_n \text{ en } \beta = \inf_n b_n,$$

$$a \leq \alpha \leq \beta \leq b, \text{ en } \alpha, \beta \in [a_n, b_n] \text{ voor elk natuurlijk getal } n.$$

Bovendien is $|\beta - \alpha| \leq b_n - a_n = 2^{-n}(b - a)$ voor elk natuurlijk getal n, dus $\alpha = \beta$.

Zij $\varepsilon > 0$. Dan is er een n_ε zodat

$$\alpha - \varepsilon < a_{n_\varepsilon} \leq \alpha = \beta \leq b_{n_\varepsilon} < \beta + \varepsilon.$$

Maar dan geldt voor elke $n \geq n_\varepsilon$ dat

$$\alpha - \varepsilon < a_{n_\varepsilon} \leq a_n \leq x_n^n \leq b_n \leq b_{n_\varepsilon} < \beta + \varepsilon.$$

Dus

$$x_n^n \in (\alpha - \varepsilon, \alpha + \varepsilon), \text{ m.a.w. } |x_n^n - \alpha| < \varepsilon.$$

Omdat $\varepsilon > 0$ willekeurig was betekent dit dat de (deel)rij $(x_n^n)_{n=1}^\infty$ convergent is met limiet $\alpha \in [a, b]$. Q.e.d.

Gevolg 6.2. (Bolzano-Weierstrass) Iedere begrensde rij in $\mathbb{R}$ heeft een convergente deelrij.

Stelling 6.3. Laat $A \subset \mathbb{R}$. Dan geldt

$$A \text{ is rijkompakt } \iff A \text{ is gesloten en begrensd.}$$

Bewijs. Opgave.

Stelling 6.4. Laat A een gesloten begrensde deelverzameling van $\mathbb{R}$ zijn. Als $f : A \to \mathbb{R}$ een continue functie is, dan heeft f een maximum in A.

Bewijs. Opgave.

Stelling 6.5. Iedere Cauchyrij in $\mathbb{R}$ is convergent (m.a.w. $\mathbb{R}$ is volledig).

Bewijs. Opgave.

Stelling 6.6. (absoluut convergente reeksen zijn convergent) Als $(a_n)_{n=1}^\infty$ een rij is in $\mathbb{R}$, en

$$\sum_{n=1}^{\infty} |a_n| = \lim_{k\to\infty} \sum_{n=1}^{k} |a_n|$$

bestaat, dan bestaat ook

$$\sum_{n=1}^{\infty} a_n = \lim_{k\to\infty} \sum_{n=1}^{k} a_n.$$

Bewijs. Opgave. Hint: laat zien dat de rij $(s_k)_{k=1}^\infty$, gedefinieerd door

$$s_k = \sum_{n=1}^{k} a_n,$$

een Cauchyrij is.

De laatste vijf stellingen gelden ook voor de metrische ruimte

$$\mathbb{R}^N = \{x = (x_1, x_2, \ldots, x_N) : \ x_1, x_2, \ldots, x_N \in \mathbb{R}\}, \quad (N \in \mathbb{N})$$

met de standaard Euclidische metriek

$$d(x,y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \cdots + (x_N-y_N)^2}.$$

Dit kunnen we inzien door $\mathbb{R}^N$ op te vatten als het herhaald Cartesisch produkt van $\mathbb{R}$ met zich zelf. Echter, de produktmetriek zoals gedefinieerd in Sectie 3, is niet de Euclidische metriek, maar

$$d_{prod}(x,y) = |x_1-y_1| + |x_2-y_2| + \cdots + |x_N-y_N|.$$

Opgave 6.7. Laat zien dat deze twee metrieken equivalent zijn, en bewijs Stelling 6.2 tot en met 6.6 voor $\mathbb{R}^N$.

Appendix. Kompaktheid en rijkompaktheid.

Definitie. Laat $A \subset X$.
(i) Een door een verzameling I geindexeerde collectie (open) deelverzamelingen $\{O_i : \ i \in I\}$ van X heet een *(open) overdekking* van A als

$$A \ \subset \ \cup_{i\in I}\, O_i.$$

(ii) Als iedere open overdekking $\{O_i : \ i \in I\}$ van A een eindige deeloverdekking van A bevat, d.w.z.

$$\exists i_1, i_2, \ldots, i_m \in I : \quad A \ \subset \ \cup_{j=1,2,\ldots,m}\, O_{i_j},$$

dan heet A *kompakt* .

Stelling. Laat $A \subset X$. Dan

$$A \text{ is kompakt } \iff \ A \text{ is rijkompakt.}$$

Bewijs. ($\Longrightarrow$). Stel A is kompakt en laat $(a_n)_{n=1}^{\infty}$ en rij zijn in A. We moeten bewijzen dat A een convergente deelrij met limiet in A heeft. Stel niet, dan is er voor elke $p \in A$ een $\varepsilon_p > 0$ en een n_p zodanig dat $a_n \notin B_{\varepsilon_p}(p)$ voor alle $n > n_p$. Neem nu als open overdekking van A de zojuist gevonden open bollen, dus

$$\{B_{\varepsilon_p}(p) : \ p \in A\}.$$

Omdat A kompakt is heeft deze overdekking een eindige deeloverdekking, dus er zijn $p_1, p_2, \ldots, p_m$ in A zodat

$$A \subset B_{\varepsilon_{p_1}}(p_1) \cup B_{\varepsilon_{p_2}}(p_2) \cup \cdots \cup B_{\varepsilon_{p_m}}(p_m),$$

waardoor A op zijn hoogst $n_1 + n_2 + \cdots + n_m$ punten van de rij kan bevatten, m.a.w. de rij bevat slechts eindig veel verschillende punten van A,

en tenminste een punt moet dus oneindig vaak voorkomen. Dus kunnen we een deelrij maken waarin dit punt alleen maar voorkomt, en dit is dan de gezochte deelrij.

($\Leftarrow$) Dit heeft wat meer voeten in de aarde. Daartoe eerst het volgende.

Definitie. Laat A een deelverzameling zijn van een metrische ruimte X. Dan heet A *aftelbaar kompakt* als voor elke rij open verzamelingen $(O_n)_{n=1}^{\infty}$ met

$$A \subset \cup_{n=1}^{\infty} O_n,$$

er een index k is zo dat

$$A \subset \cup_{n=1}^{k} O_n.$$

Stelling. Als $A \subset X$ rijkompakt is, dan is A aftelbaar kompakt.

Bewijs. Neem aan dat A rijkompakt is maar niet aftelbaar kompakt. Dan is er een rij open verzamelingen $(O_n)_{n=1}^{\infty}$ van A die A geheel bedekt, en die geen eindige deeloverdekking heeft. Dat betekent dat er voor elk natuurlijk getal k een punt p_k in A is met

$$p_k \notin \cup_{n=1}^{k} O_n.$$

Omdat A rijkompakt is heeft de rij $(p_n)_{n=1}^{\infty}$ een convergente deelrij met limiet p in A. Dus moet er een m zijn waarvoor $p \in O_m$. Omdat O_m open is zijn er dus oneindig veel punten van de rij $(p_n)_{n=1}^{\infty}$ die in O_m liggen. Voor $n > m$ is dit in tegenspraak met de keuze van p_n. Q.e.d.

Definitie. Laat A een deelverzameling zijn van een metrische ruimte X. Dan heet A *totaal begrensd* als er voor elke $\varepsilon > 0$ een eindige deelverzameling $\{p_1, p_2, \ldots, p_n\}$ van A bestaat met

$$A \subset B_\varepsilon(p_1) \cup B_\varepsilon(p_2) \cup \cdots \cup B_\varepsilon(p_n).$$

Stelling. Als $A \subset X$ rijkompakt is, dan is A totaal begrensd.

Bewijs. Stel A is niet totaal begrensd. Dan is er een $\varepsilon > 0$ waarvoor geen eindige verzameling als in de definitie kan worden gevonden. Kies nu $p_1 \in A$. Inductief kiezen we nu voor $n = 1, 2, \ldots$ een punt $p_{n+1} \in A$ met de eigenschap dat

$$p_{n+1} \notin B_\varepsilon(p_1) \cup B_\varepsilon(p_2) \cup \cdots \cup B_\varepsilon(p_n).$$

Maar dan is $d(p_i, p_j) \geq \varepsilon$ voor alle $i \neq j$. Dus de rij $(p_n)_{n=1}^{\infty}$ kan geen convergente deelrij hebben, in tegenspraak met de rijkompaktheid van A. Q.e.d.

Definitie. Een metrische ruimte X heet *separabel* als er een rij is in X zo dat elk punt van X een limietpunt is van deze rij.

Stelling. Als $A \subset X$ totaal begrensd is, dan is A separabel.

Bewijs. Voor elke k bestaat er een eindige deelverzameling A_k van A zo dat elk punt van A dichter dan $\frac{1}{k}$ bij een punt van A_k ligt. Maak nu een rij door eerst de elementen van A_1 te kiezen, dan de elementen van A_2, A_3, enzovoort. Dan is elk punt van A een limietpunt van de aldus verkregen rij. Q.e.d.

Stelling. Als $A \subset X$ separabel is, dan heeft elke open overdekking van A een aftelbare deeloverdekking.

Bewijs. Laat $(p_n)_{n=1}^{\infty}$ een rij zijn in A met de eigenschap dat elk punt van A limietpunt is van deze rij, en laat $\{O_i : \ i \in I\}$ een open overdekking zijn van A. Dan is elk punt p in A bevat in een O_i. Kies nu een natuurlijk getal m zo groot dat $B_{\frac{2}{m}}(p) \subset O_i$, en daarna een n zo dat $d(p_n, p) < \frac{1}{m}$. Dan is

$$p \in B_{\frac{1}{m}}(p_n) \subset O_i.$$

Laat J de verzameling zijn van de paren natuurlijke getallen (m, n) die we zo tegenkomen als p de verzameling A doorloopt. Dan is J aftelbaar en

$$A \subset \cup_{(m,n)\in J} B_{\frac{1}{m}}(p_n).$$

Maar voor elke $(m, n) \in J$ is er tenminste één O_i die $B_{\frac{1}{m}}(p_n)$ bevat. Kies er voor elke $(m, n) \in J$ precies één. Dit geeft een aftelbare deelcollectie die A overdekt. Q.e.d.

Uit bovenstaande stellingen volgt dat als $A \subset X$ rijkompakt is, dat A ook "gewoon" kompakt is.

Leiden, juni 1993.

43.2 Metrische ruimten

Onze genormeerde ruimten X, waaronder $\mathbb{R}, \mathbb{R}^2$ en ook $C([a,b])$ met de maximumnorm, maar helaas niet $R([a,b])$ met de 1-norm, zijn voorbeelden van metrische ruimten met het afstandsbegrip gedefinieerd door de metriek

$$(x,y) \xrightarrow{d} d(x,y) = |x-y|, \tag{43.1}$$

een afbeelding[1] d van $X \times X$ naar $[0,\infty)$ met de eigenschappen dat voor alle $x,y,z \in X$ geldt dat

$$(i) \quad d(x,y) = 0 \iff x = y; \quad (ii) \quad d(x,y) = d(y,x);$$

$$(iii) \quad d(x,y) \leq d(x,z) + d(z,x). \tag{43.2}$$

Iedere niet-lege deelverzameling A van X is zo een metrische ruimte, waarbij we de algebraische vectorruimte operaties nu vergeten.

Definition 43.1. *Een metrische ruimte is een niet-lege verzameling X met een afbeelding $d : X \times X \to [0,\infty)$ waarvoor (i),(ii) en (iii) uit (43.2) hierboven gelden voor alle $x,y,z \in X$.*

Exercise 43.2. De ε, N-definitie van $d(x_n, x_m) \to 0$ als $m,n \to \infty$ definieert wat een Cauchyrij in X is. Geef die definitie. Geef ook de definitie van het convergent zijn van de rij x_n in X.

We gebruiken hieronder de notatie $x_n \to x$ voor $x_1, x_2, x_3, \ldots, x \in X$ zonder er steeds $n \to \infty$ bij te zetten en spreken over ook een rij x_n zonder te vermelden dat $n \in \mathbb{N}$ (of een andere deelverzameling van $\mathbb{Z}$ van de vorm $m + \mathbb{N}$ met $m \in \mathbb{Z}$, bijvoorbeeld $\mathbb{N}_0$).

Exercise 43.3. Een flauwe opgave om aan de de notaties, definities en axioma's te wennen: laat zien dat als $x_n \to x$ en $x_n \to y$ (alles in X) voor de limieten x en y geldt dat $x = y$. De limiet van een convergente rij is dus uniek.

Met convergente rijen kunnen we voor metrische ruimten X en Y zeggen wat het voor een afbeelding

$$F : X \to Y$$

betekent om continu te zijn in $a \in X$.

[1] De d van distance, a van afstand doen we maar niet.

Definition 43.4. *Een afbeelding F van een metrische ruimte X naar een (niet per se andere) metrische ruimte Y heet continu in $a \in X$ als de implicatie*

$$x_n \to a \implies F(x_n) \to F(a)$$

geldt voor elke rij x_n in X. Als dit het geval is voor elke $a \in X$ dan zeggen we dat $F : X \to Y$ continu is.

Exercise 43.5. Als X, Y, Z metrische ruimten en

$$X \xrightarrow{F} Y \quad \text{en} \quad Y \xrightarrow{G} Z$$

afbeeldingen dan is de afbeelding

$$X \xrightarrow{G \circ F} Z \quad \text{gedefinieerd door} \quad X \xrightarrow{F} Y \xrightarrow{G} Z$$

continu in $a \in X$ als F continu is in a en G continu is in $b = F(a)$. Hint: triviaal, leg uit.

Definition 43.6. *Een metrische ruimte heet rijkompakt als elke rij in X een convergente deelrij heeft, en volledig als elke Cauchyrij in X convergent is (in beide gevallen met limiet in X dus).*

Exercise 43.7. Bewijs dat rijkompakte metrische ruimten volledig zijn.

Exercise 43.8. Als X en Y metrische ruimten zijn, met X rijkompakt, dan is iedere continue $F : X \to Y$ uniform continu, i.e.

$$\forall \varepsilon > 0 \, \exists \delta > 0 \, \forall x, a \in X : \, d(x, a) \leq \delta \implies d(F(x), F(a)) \leq \varepsilon.$$

Bewijs dit door een eerder bewijs over te schrijven.

Exercise 43.9. Als X een rijkompakte metrische ruimte is, dan heeft iedere continue $F : X \to \mathbb{R}$ een globaal maximum en een globaal minimum op X. Bewijs ook dit door een eerder bewijs over te schrijven.

43.3 Omgevingen, open en gesloten verzamelingen

Continuiteit kunnen we ook met open verzamelingen beschrijven. In het standaardjargon heet een deelverzameling $G \subset X$ van een metrische ruimte X gesloten als voor iedere rij x_n in G met $x_n \to x \in X$ de limiet x in G zit (je kan G niet uit door limieten te nemen). Een verzameling $O \subset X$ heet open[2] als zijn complement gesloten is.

Exercise 43.10. Bewijs dat in een Banachruimte iedere rijkompakte deelverzameling begrensd en gesloten is en dat in $\mathbb{R}^n$ ook de omgekeerde uitspraak geldt.

Uit Opgave 43.9 en Opgave 43.10 volgt dat de stelling over maxima en minima van continue functies op gesloten begrensde deelverzamelingen van $\mathbb{R}^{\mathrm{N}}$.

Theorem 43.11. *Laat $K \subset \mathbb{R}^{\mathrm{N}}$ begrensd en gesloten zijn en $F : K \to \mathbb{R}$ continu zijn. Dan zijn er $a, b \in K$ met $F(a) \leq F(x) \leq F(b)$ voor alle $x \in K$. De punten a en b heten de minimizer en de maximizer voor F, en de waarden $F(a)$ en $F(b)$ het minimum en het maximum van F.*

Exercise 43.12. De collectie $\mathcal{G}$ van alle gesloten deelverzamelingen van een metrische ruimte X heeft drie belangrijke eigenschappen:

$$(i) \quad \emptyset \in \mathcal{G},\, X \in \mathcal{G}; \quad (ii) \quad G_1, G_2 \in \mathcal{G} \implies G_1 \cup G_2 \in \mathcal{G};$$

en (voor elke indexverzameling I)

$$(iii) \quad G_i \in \mathcal{G}\ \forall i \in I \implies \cap_{i\in I} G_i \in \mathcal{G}.$$

Bewijs dit via de definitie dat $G \in \mathcal{G}$ als voor iedere rij x_n in G met $x_n \to x \in X$ voor de limiet geldt $x \in G$.

Exercise 43.13. De collectie $\mathcal{O}$ van alle open deelverzamelingen van een metrische ruimte X heeft de volgende eigenschappen:

$$\emptyset \in \mathcal{O},\, X \in \mathcal{O}; \quad O_1, O_2 \in \mathcal{O} \implies O_1 \cap O_2 \in \mathcal{O};$$

en (voor elke indexverzameling I)

$$O_i \in \mathcal{O}\ \forall i \in I \implies \cup_{i\in I} O_i \in \mathcal{O}.$$

[2]Minder gelukkige naamgeving, sorry, is niet anders.

Bewijs dit via de definitie dat $O \in \mathcal{O}$ als

$$O^c = \{x \in X : x \notin O\} \in \mathcal{G}.$$

Exercise 43.14. Laat zien dat in een metrische ruimte X een deelverzameling $O \subset X$ open is dan en slechts dan als voor elke $a \in O$ er een $r > 0$ is zo dat

$$\bar{B}_r(a) = \{x \in A : d(x, a) \leq r\} \subset O.$$

Bewijs ook dat $\bar{B}_r(a)$ gesloten is.

Om te weten welke verzamelingen open zijn moet je dus weten wat de gesloten bollen $\bar{B}_r(a)$ zijn maar niet eens dat. Heb je bijvoorbeeld twee normen en noemen we de bijbehorende bollen $\bar{B}_r(a)$ en $\bar{K}_s(a)$ dan krijgen we precies dezelfde open verzamelingen als elke $\bar{B}_r(a)$ met $r > 0$ altijd een $\bar{K}_s(a)$ bevat met $s > 0$ en omgekeerd. Is X een vectorruimte over $\mathbb{R}$ met twee normen dan noemen we die normen equivalent als ze dezelfde collectie $\mathcal{O}$ definieren. Via Opgave 43.12 leidt dat tot deze karakterisatie van equivalente normen op X.

Exercise 43.15. Als twee normen

$$x \to |x|_1 \quad \text{en} \quad x \to |x|_2$$

dezelfde collectie $\mathcal{O}$ van open verzamelingen definieren dan zijn er constanten A_1 en A_2 zo dat voor alle $x \in X$ geldt

$$|x|_1 \leq A_2|x|_2 \quad \text{en} \quad |x|_2 \leq A_1|x|_1.$$

Bewijs dit. Terzijde, omgekeerd geldt ook en is makkelijker.

Exercise 43.16. Laat zien dat in een metrische ruimte X een deelverzameling $O \subset X$ open is dan en slechts dan als voor elke $a \in O$ er een $r > 0$ is zo dat

$$B_r(a) = \{x \in A : d(x, a) < r\} \subset O.$$

Bewijs ook dat $B_r(a)$ open is.

Theorem 43.17. *Laat X en Y metrische ruimten zijn en $F : X \to Y$. Dan is F continu dan en slechts dan als alle inverse beelden van open verzamelingen in Y open zijn in X.*

Exercise 43.18. Wel een kluifje: bewijs Stelling 43.17. Triviaal daarna is dat als X, Y, Z metrische ruimten zijn en

$$X \xrightarrow{F} Y \quad \text{en} \quad Y \xrightarrow{G} Z$$

continue afbeeldingen, dat de afbeelding

$$X \xrightarrow{G\circ F} Z \quad \text{gedefinieerd door} \quad X \xrightarrow{F} Y \xrightarrow{G} Z$$

continu is. Waarom? Zie nog even Opgave 43.5.

In $\mathbb{R}^2$ hebben we behalve the standaardnorm

$$|x| = \sqrt{x_1^2 + x_2^2} = \sqrt{x \cdot x} \quad \text{voor} \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

afkomstig van het standaardinprodukt

$$x \cdot y = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = x_1 y_1 + x_2 y_2,$$

de normen

$$|x|_p = \sqrt[p]{|x_1|^p + |x_2|^p} \quad \text{voor} \quad p \geq 1 \quad \text{en} \quad |x|_\infty = \max(|x_1|, |x_2|).$$

Als deze normen zijn equivalent.

Exercise 43.19. Bewijs dat al deze p-normen equivalent zijn en teken in het x_1, x_2-vlak de gesloten eenheidsbollen $\bar{B}^p = \{x \in \mathbb{R}^2 : |x|_p \leq 1\}$ voor $p = 1, 2$ en $p = \infty$, en voor nog twee p's naar keuze. Blader nog even terug naar Opgave 43.15 en de karakterisatie daaronder en boven van open verzamelingen met behulp bollen, gesloten of open, zoals $B^p_\varepsilon(\xi) = \{x \in \mathbb{R}^2 : |x - x_0|_p < \varepsilon\}$ met $\xi \in \mathbb{R}^2$ en $\varepsilon > 0$.

Exercise 43.20. De bollen B^1 en B^∞ zijn ook te beschrijven als doorsnijdingen van open halfvlakken van de form $K = \{x \in \mathbb{R}^2 : f(x) < b\}$ met $f : \mathbb{R}^2 \to \mathbb{R}$ lineair gegeven door $f(x) = a_1 x_1 + a_2 x_2$ en $a_1, a_2, b \in \mathbb{R}$. Laat dat zien.

Exercise 43.21. Een alternatieve manier om te zeggen dat een $O \in \mathbb{R}^2$ open is te zeggen dat er voor elke $\xi \in O$ drie[3] open halfvlakken K_1, K_2, K_3 zijn zoals in Opgave 43.20, waarvoor geldt

$$\xi \in K_1 \cap K_2 \cap K_3 \subset O.$$

Waarom definieert dit dezelfde open verzamelingen? Geef ook zo'n definitie van open in $\mathbb{R}^3$.

Exercise 43.22. Een verzameling W in een genormeerde ruimte X heet zwak open als er voor elke $\xi \in W$ geldt dat er er eindig veel open halfvlakken zijn zo dat geldt

$$\xi \in K_1 \cap \cdots \cap K_n \subset W.$$

Bewijs dat voor deze zwak open verzamelingen W dezelfde eigenschappen gelden als in Opgave 43.13. Met eindige doorsnijdingen van open halfvlakken is dus een topologie te maken: een collectie van "open" verzamelingen die voldoet aan de "axioma's" in Opgave 43.13. In het geval dat $X = \mathbb{R}^n$ zijn alle normen op X en deze topologie equivalent.

`https://www.youtube.com/watch?v=fmTcSGuk04o`

Exercise 43.23. Bewijs dat iedere norm $x \to |x|$ op $\mathbb{R}^2$ equivalent is met de 2-norm. Hint: laat eerst zien dat $x \to |x|$ op $S = \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$ een positief minimum en maximum heeft.

Exercise 43.24. Laat X_1 en X_2 genormeerde ruimten zijn. Bewijs dat

$$X_1 \times X_2 = \{x = (x_1, x_2) : x_1 \in X_1, x_2 \in X_2\}$$

met de voor de hand liggende bewerkingen weer een genormeerde ruimte is met (equivalente) normen (voor $p \geq 1$)

$$x \to \sqrt[p]{|x_1|^p + |x_2|^p} \quad \text{en} \quad x \to \max(|x_1|, |x_2|).$$

[3] $3 = 2 + 1$.

Exercise 43.25. Laat X_1 en X_2 genormeerde ruimten zijn en $X = X_1 \times X_2$. Bewijs dat iedere $f \in X^*$ van de vorm

$$x = (x_1, x_2) \xrightarrow{f} f_1(x_1) + f_2(x_2)$$

is met $f_1 \in X_1^*, f_2 \in X_2^*$. Met andere woorden $X^* = X_1^* \times X_2^*$.

Exercise 43.26. Laat X_1 en X_2 genormeerde ruimten zijn en $f \in X^* = X_1^* \times X_2^*$. Bepaal de norm van f in X^* als voor de norm op $X = X_1 \times X_2$ de norm $x \to |x_1| + |x_2|$ genomen wordt. Zelfde vraag voor $x \to \max(|x_1|, |x_2|)$.

44 Welke fundamenten?

This introduction was intended for a different audience. Deze oude inleiding was bedoeld voor een breed publiek. De eerstejaars wiskunde student kan voor de lol lezen wat ik hier schrijf. Ik begin met de verzameling $\mathbb{R}$ van de *reële getallen* en aftelbare sommen van die getallen. De stijl is zoals in het *groene* boekje *Wiskunde in je vingers* met Ronald Meester. In dat boekje, dat vanaf nu [HM] heet, kwamen we vanuit getallenrepresentaties als

$$\frac{1}{3} = \frac{3}{10} + \frac{3}{100} + \frac{3}{1000} + \frac{3}{10000} + \frac{3}{100000} + \cdots = \sum_{n=1}^{\infty} \frac{3}{10^n} = 3\sum_{n=1}^{\infty} \frac{1}{10^n}$$

op natuurlijke wijze tot het inzicht dat ieder (reëel) getal van de vorm

$$k + \sum_{n=1}^{\infty} \frac{d_n}{10^n} \tag{44.1}$$

is. In (44.1) is $k \in \mathbb{Z}$, de verzameling van de *gehele getallen*. De decimalen zijn

$$d_n \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

waarbij n de verzameling $\mathbb{N}$ van de positieve[1] gehele getallen doorloopt, ook wel de *natuurlijke getallen* genoemd.

De verschillende notaties hierboven voor het rationale getal dan wel de breuk $\frac{1}{3}$ kunnen tot enige controverse leiden. De breuk $\frac{1}{3}$ heeft immers net als iedere andere breuk een teller en een noemer, in dit geval teller 1 en noemer 3. Evenzo heeft de breuk $\frac{2}{6}$ teller 2 en noemer 6. De breuken $\frac{1}{3}$ en $\frac{2}{6}$ zijn echter als rationale getallen gelijk aan elkaar. Mag je nu van het rationale getal $\frac{1}{3}$ zeggen dat zijn teller 1 en zijn noemer 3 is? Van mij wel, maar daar wordt soms anders over gedacht. Dus daarom hierbij de afspraak dat we stilzwijgend het rationale getal altijd als breuk met een minimale noemer[2] in $\mathbb{N}$ schrijven als we het over teller en noemer van het rationale getal hebben.

Ook de naam "reeks" voor de uitdrukking met het somteken Σ leidt tot controverses, alsmede het gebruik van het symbool ∞ boven op dat somteken. Wat het eerste betreft zou ik liever zoveel mogelijk over aftelbare sommen willen spreken, maar niet te vergeten dat de term "reeks" nu eenmaal door iedereen gebruikt wordt in zinsdelen als "de som van de reeks".

Het gebruik van het symbool ∞ is wellicht te vermijden door

$$\sum_{n \in \mathbb{N}} \quad \text{in plaats van} \quad \sum_{n=1}^{\infty}$$

[1] NB, 0 is niet positief, $\mathbb{N} = \{n \in \mathbb{Z} : n > 0\}$, $\mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$.

[2] Ontbind teller en noemer in priemfactoren en streep gemeenschappelijke factoren weg.

te schrijven, maar dan is de volgorde waarin de termen in de som bij elkaar op worden geteld niet meer zo eenduidig specificeerd als in de meer gebruikelijke notatie. Die wordt namelijk doorgaans uitgesproken als *de som van de termen in de reeks, waarbij n loopt vanaf het getal* 1 *tot (en niet tot en met) oneindig*[3].

Tenslotte merken we op dat de schrijfwijze in (44.1) niet altijd uniek is omdat getallen van de vorm

$$k + \sum_{n=1}^{m} \frac{d_n}{10^n} \tag{44.2}$$

nu eenmaal twee representaties hebben, bijvoorbeeld

$$1 = \frac{9}{10} + \frac{9}{100} + \frac{9}{1000} + \frac{9}{10000} + \frac{9}{100000} + \cdots = \sum_{n=1}^{\infty} \frac{9}{10^n}, \tag{44.3}$$

wellicht het eerste voorbeeld van een zogenaamde meetkundige reeks dat ieder kind in het basisonderwijs hopelijk wel eens te zien krijgt.

Het simpelste voorbeeld van zo'n meetkundige reeks betreft de rij breuken

$$\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64} \ldots,$$

met in de noemers de getallen uit de eerste rij getallen die ik ooit van mijn vader leerde, toen ik een jaar of 2^2 was. Als we die rij beschrijven met

$$a_n = \frac{1}{2^n}$$

met n de verzameling $\mathbb{N}$ doorlopend, dan is de bijbehorende som gelijk aan het getal 1. Nog altijd de mooiste som die er bestaat. Een eindeloze rij getallen die optellen tot 1. Wat wil je nog meer?

44.1 Academisch speelkwartier: kolomcijferen

We kiezen nu voor een wat basaler perspectief dan gebruikelijk om meer inzicht te krijgen in wat de reële getallen zijn. Deze inventariserende[4] subsectie kan overgeslagen worden bij eerste lezing[5], maar een opmerking van Jan Wiegerinck zette me aan het denken, ook in relatie tot Opgave 1.7 in [HM]. Hoe rekenen we met (eerste maar positieve) getallen in de vorm (44.1)? Een

[3] Rekenen met ∞ doen wij hier niet.

[4] We gaan niet recht op een doel af nu.

[5] En ook bij tweede lezing.

rekenvraag die we nu rekenend en redenerend willen beantwoorden zonder over verzamelingen van getallen te spreken.

Terzijde, helemaal consequent is de schrijfwijze in (44.1) niet. De k is duidelijk anders dan de rest van de termen in deze aftelbare som[6]. Als we ons beperken tot de positieve reële getallen, die we als verzameling[7] gezien aanduiden met $\mathbb{R}^+$, dan is het eleganter om het "Romeinse" perspectief van eenheden, tientallen, tienden, honderdtallen, hondersten, duizendtallen, duizendsten, etc alleen te combineren met de "Arabische" cijfers

$$0, 1, 2, 3, 4, 5, 6, 7, 8, 9.$$

Met

$$\text{tien} = 10, \quad \text{honderd} = 10 \times 10 = 100, \quad \text{duizend} = 10 \times 10 \times 10 = 1000,$$

enzovoorts, en op zijn kop

$$\text{een tiende} = \frac{1}{10}, \quad \text{een honderste} = \frac{1}{100}, \quad \text{een duizendste} = \frac{1}{1000},$$

und so weiter, is het niet heel raar om bijvoorbeeld

$$8 \times 1000 + 7 \times 100 + 6 \times 10 + 5 \times 1 + 4 \times \frac{1}{10} + 3 \times \frac{1}{100} + 2 \times \frac{1}{1000}$$

als corresponderend met een punt op een lijn van hier tot ginder te zien. Bij zo'n punt hoort een getal dat we noteren als

$$8765{,}4321 \tag{44.4}$$

in onze decimale notatie van vandaag de dag, met ook (hier 5 keer) de gewone eenheid 1, en een Nederlandse komma waarachter in dit geval vier cijfers staan.

Over getallen als (44.4) hoeven verder geen misverstanden te bestaan. Links van de komma tellen de decimalen van rechts af de 1-tallen, 10-tallen, 10×10-tallen, $10 \times 10 \times 10$-tallen, en rechts van de komma vanaf links de $\frac{1}{10}$-tallen, $\frac{1}{10}$-tallen, $\frac{1}{10} \times \frac{1}{10}$-tallen, $\frac{1}{10} \times \frac{1}{10} \times \frac{1}{10}$-tallen, $\frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10}$-tallen. Aan beide kanten van de komma breekt het af, maar rechts is dat niet nodig. In principe kunnen we er nog een rij cijfers achter plaatsen, en een groter getal maken, bijvoorbeeld

$$8765{,}432143214321 \quad \text{of} \quad 8765{,}432199999999999999999999999,$$

[6]Som van een eindeloze rij, zoals Marjolein Kool dat zo mooi noemt.

[7]Getallen in verzamelingen willen stoppen is een beetje een beroepsafwijking.

getallen die allebei groter zijn dan 8765,4321 en kleiner dan 8765,43211.

Omdat het in negen gelijke stukken verdelen van het lijnstuk tussen het beginpunt van diezelfde lijn van hier tot ginder, en het punt waar

$$0 \times 1000 + 0 \times 100 + 0 \times 10 + 1 \times 1 + 0 \times \frac{1}{10} + 0 \times \frac{1}{100} + 0 \times \frac{1}{1000} = 1$$

staat, negen lijnstukken geeft waarvan het eerste nog net niet loopt tot

$$0 \times 1000 + 0 \times 100 + 0 \times 10 + 1 \times 1 + 1 \times \frac{1}{10} + 1 \times \frac{1}{100} + 1 \times \frac{1}{1000} = 0{,}1111,$$

zien we dat er geen reden is waarom elk punt op de lijn een afbrekende getalrepresentatie zou moeten hebben. Wat heet, één negende correpondeert omherroepelijk met een representatie als in (44.4) waarbij er voor de komma alleen maar 0-en staan, en achter de komma alleen maar 1-en, zonder dat het rechts afbreekt[8].

Dat

$$9 \times 0{,}111111111\ldots = 9 \times 0{,}\underline{1} \quad \text{gelijk is aan} \quad 1,$$

is een conclusie die we willen trekken als resultaat van de herhaalde optelling

$$0{,}\underline{1} + 0{,}\underline{1} + 0{,}\underline{1} + 0{,}\underline{1} + 0{,}\underline{1} + 0{,}\underline{1} + 0{,}\underline{1} + 0{,}\underline{1} + 0{,}\underline{1} = 0{,}\underline{9} = 1.$$

Op dat optellen komen we zo terug, en op $0{,}\underline{9} = 1$ indirect ook. Bij een (met de nog te speciferen regels) decimaal geschreven getal $0{,}\underline{9}$ optellen verhoogt het gehele getal voor de komma met 1.

Verdelen we hetzelfde lijnstuk niet in negen maar in 99 gelijke stukken, dan zien we

$$0{,}010101010101010101010101\ldots \tag{44.5}$$

als getalrepresentatie voor één negenennegentigste verschijnen. De puntjes geven hier aan dat de decimale ontwikkeling niet eindigt. Tegenwoordig schrijven we

$$\frac{1}{9} = 0{,}\underline{1}, \quad \frac{1}{99} = 0{,}\underline{01}, \quad \frac{1}{999} = 0{,}\underline{001},$$

met links steeds een rationaal getal en rechts de decimale representatie van dat getal, dat we met liefde ook een breuk mogen noemen, een breuk met teller 1 en en een noemer met alleen maar 9-ens.

We zien in (44.5) dat de 0 als cijfer erg handig is, de 0 die correspondeert met nul vingers op de twee gebalde vuisten van je handen waar je geen ruzie mee wil krijgen. Het tellen zelf begint met 1, eindigt op de vingers bij tien = 10, en gaat daarna verder met 11, 12, 13, 14, 15, 16, 17, 18, 19, 20,

[8] En links eigenlijk ook niet, al schrijven we die nullen nooit op.

21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, en een kind kan al lerend voor het slapen gaan zien en begrijpen hoe dat zo altijd maar doorgaat als dromenland geen redding brengt. De eindeloze rij van de natuurlijke getallen, beginnend met 1, wordt zo ondubbelzinnig vastgelegd door de aftelling in het decimale stelsel.

Getallen uit die rij kunnen we optellen en vermenigvuldigen (in wezen herhaald optellen). Dat doen we cijferend met de getallen onder elkaar gezet, onhandig vanaf links of handig vanaf rechts per kolom. Die methodes[9] werken ook voor het rekenen met de positieve kommagetallen die we krijgen door voor de komma van het kommagetal het cijfer 0 of een natuurlijk getal te zetten, en achter de komma een natuurlijk getal opgevat als een rij cijfers, met voor dat getal al of niet nog een aantal nullen.

De natuurlijke getallen zelf zijn geen kommagetallen. Door getallen als 123456789 gelijk te zien aan een nepkommagetal 123456789,0 is dat snel verholpen, maar dit wordt in de natuurkunde[10] terecht als een minder gelukkige en te mijden conventie gezien. Afbrekende kommagetallen kunnen wellicht

[9] In het PO heeft de onhandige methode veelal de voorkeur gekregen.

[10] Waar de laatste decimaal meestal met meetnauwkeurigheid te maken heeft.

beter vermeden worden, en de niet afbrekende kommagetallen zijn nu net de kommagetallen die we nog missen, en waar we zeker ook mee willen rekenen en cijferen. Dat *cijferen* moet dan wel vanaf links als je er even over nadenkt.

Tellen we bijvoorbeeld het getal dat gerepresenteerd wordt door (44.5) achtennegentig keer bij zichzelf op dan zien we dat 99 keer $\frac{1}{99}$ niet alleen gelijk is aan 1 maar ook aan

$$0{,}\underline{9} = 0{,}999999999999999999999999\ldots,$$

en deze decimale representatie kan de meestal toch onbereikbare eenheid 1 best vervangen, als we afspreken dat decimale representaties van positieve reële getallen naar links altijd, maar naar rechts nooit doorlopen met alleen maar 0-en. Een onhandige conventie wellicht, maar omdat het *cijferen* hier toch alleen maar onhandig vanaf links kan niet eens zo gek eigenlijk.

Bovendien is elk natuurlijk getal nu op een natuurlijke manier ook een echt kommagetal. Met $0{,}\underline{9}$, $1{,}\underline{9}$, $2{,}\underline{9}$, $3{,}\underline{9}$, $4{,}\underline{9}$, $5{,}\underline{9}$, $6{,}\underline{9}$, $7{,}\underline{9}$, $8{,}\underline{9}$, $9{,}\underline{9}$, $10{,}\underline{9}$, $11{,}\underline{9}$, $12{,}\underline{9}$, $13{,}\underline{9}$, $14{,}\underline{9}$, $15{,}\underline{9}$, $16{,}\underline{9}$, $17{,}\underline{9}$, $18{,}\underline{9}$, $19{,}\underline{9}$, $20{,}\underline{9}$, $21{,}\underline{9}$, $22{,}\underline{9}$, $23{,}\underline{9}$, $24{,}\underline{9}$, $25{,}\underline{9}$, $26{,}\underline{9}$, $27{,}\underline{9}$, $28{,}\underline{9}$, $29{,}\underline{9}$, $30{,}\underline{9}$, $31{,}\underline{9}$, $32{,}\underline{9}$, $33{,}\underline{9}$, $34{,}\underline{9}$, $35{,}\underline{9}$, $36{,}\underline{9}$, $37{,}\underline{9}$, $38{,}\underline{9}$, $39{,}\underline{9}$, $40{,}\underline{9}$, $41{,}\underline{9}$, $42{,}\underline{9}$, komen we ook op weg in de rekenkunde. Ook al bekt het niet zo lekker, we gaan rekenen met deze niet afbrekende kommagetallen zoals ze ons gegeven worden.

44.1.1 Optellen

De vraag is nu of we met alle ons gegeven kommagetallen kunnen rekenen zoals je zou verwachten, en of rekenen dan (zoals tegenwoordig in het basisonderwijs) onhandig cijferen kan worden, zoals bijvoorbeeld in

$$\begin{array}{r}
0{,}9999999 \\
0{,}9999999 \\
\hline
+ \\
1{,}8000000 \\
0{,}1800000 \\
0{,}0180000 \\
0{,}0018000 \\
0{,}0001800 \\
0{,}0000180 \\
0{,}0000018 \\
\hline
+ \\
1{,}9999998
\end{array}$$

Immers, met doorlopende negens gaat dit onhandig[11] *cijferen* precies hetzelfde en duurt nauwelijks langer dan hierboven:

$$\begin{array}{l} 0,99999\ldots \\ 0,99999\ldots \\ \hline\hline + \\ \\ 1,80000\ldots \\ 0,18000\ldots \\ 0,01800\ldots \\ 0,00180\ldots \\ 0,000180.. \\ 0,000018.. \\ \ldots\ldots\ldots\ldots\ldots \\ \hline\hline + \\ \\ 1,99999\ldots \end{array} \tag{44.6}$$

Gelukkig: $1 + 1 = 2!$[12] Weliswaar schendt de realistisch tussenstap hier wel de regel dat we rechts geen doorlopende 0-en mogen hebben, de uitkomst van de som is duidelijk: de 8 combineert steeds met de 1 op de volgende rij tot een 9. De 18 op elke rij is de som van 9 en 9. Op de eerste rij betreft het $0{,}9 + 0{,}9$, op de tweede rij $0{,}09 + 0{,}09$, op de derde rij $0{,}009 + 0{,}009$, enzovoorts. Het is instructief[13] om zo'n sommetje als hierboven met twee andere doorlopende getallen met voor de komma alleen maar 0-en te doen. Dan vormen de twee cijfers op elke rij steeds een getal uit de rij

$$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,$$

en bij elk van deze getallen kan zowel van boven een getal uit de rij

$$0 = 00, 10, 20, 30, 40, 50, 60, 70, 80, 90$$

en daarna vanonder ook een getal uit de veel kortere rij $0, 1$ worden opgeteld. Op zijn hoogst krijgen we dus $90 + 18 + 1 = 109$.

Is de som groter dan 99 dan schuift er een 1 door naar links maar dat overhevelen blijft beperkt. Optellend per tweetal kolommen kan er een 1 naar links doorschuiven en een 1 van rechts binnenkomen. Die 1 kan van de 109 een 110 maken, maar ook die geeft nog steeds op zijn hoogst een 1

[11] Onhandig cijferen wordt ook wel kolomrekenen genoemd.
[12] Lees: één en één is twee uitroepteken.
[13] Wel doen!

naar links door. Dat optellen van twee getallen onhandig cijferend per twee kolommen tegelijk vanaf links gaat dus altijd wel lukken.

Hoe zit het met drie getallen? We nemen weer de moeilijkste som van dat type, met de cijfers zo groot mogelijk, dus

$$\begin{array}{l} 0,99999\ldots \\ 0,99999\ldots \\ 0,99999\ldots \\ \\ \text{————————} + \\ \\ 2,70000\ldots \\ 0,27000\ldots \\ 0,02700\ldots \\ 0,00270\ldots \\ 0,000270.. \\ 0,000027.. \\ \ldots\ldots\ldots\ldots\ldots \\ \text{————————} + \\ \\ 2,99999\ldots \end{array} \tag{44.7}$$

Gelukkig: $1 + 1 + 1 = 3$. Opnieuw is het instructief om zo'n sommetje als hierboven met drie andere doorlopende getallen met voor de komma alleen maar 0-en te doen. Dan vormen de twee cijfers op elke rij steeds een getal uit de rij

$$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27,$$

en bij elk van deze getallen kan zowel een getal uit de rij

$$0 = 00, 10, 20, 30, 40, 50, 60, 70, 80, 90$$

en daarna ook een getal uit de veel kortere rij

$$0, 1, 2$$

worden opgeteld. Op zijn hoogst krijgen we nu $90 + 27 + 2 = 119$. Het overhevelen naar links blijft weer beperkt. Met enig werk gaan we hier wel inzien dat drie zulke *nul komma nog minstens wat getallen* altijd cijferend bij elkaar opgeteld kunnen worden en dat het niet uitmaakt[14] of we er eerst

[14]Lees: $a + b + c = (a + b) + c = c + (a + b)$ met alle gepermuteerde variaties.

twee samen nemen en in welke volgorde we de getallen optellen, en dat in de som voor de komma ook een 1 of een 2 kan komen te staan.

Willen we positieve getallen met ook voor de komma decimalen hebben staan in de getallen die we bij elkaar optellen, dan doen we die apart. Bijvoorbeeld

$$99,\underline{9}+88,\underline{8}+77,\underline{7}=99+88+77+0,\underline{9}+0,\underline{8}+0,\underline{7},$$

waarbij

$$\begin{array}{r}99\\88\\77\\\hline 264\end{array}$$

nu ook (juist wel handig) van rechts af per kolom uitgecijferd[15] kan worden, en de som van de drie *nul komma nog minstens wat getallen* met de methode hierboven gelijk is aan 2,$\underline{6}$. Alles bij elkaar vinden we zo dat

$$99,\underline{9}+88,\underline{8}+77,\underline{7}=264+2,\underline{6}=266,\underline{6},$$

al had dat vast handiger gekund.

Is zo'n positief kommagetal p groter dan een ander positief kommagetal a, hetgeen betekent dat, na mogelijk een aantal gelijke decimalen van p en a, er een eerste decimaal is van p die groter is dan de overeenkomstige decimaal van a, dan kunnen we precies één (positieve) b vinden waarvoor geldt dat $p=a+b$. Het voorbeeldje

$$\begin{array}{r}0{,}909090909090909090...\\0{,}222222222222222222...\\\hline 0{,}686868686868686868...\end{array}$$

kan van linksaf kolomcijferend worden aangepakt. In eerste instantie is de eerste decimaal achter de komma dan gelijk aan 7 maar bij de volgende decimaal moet er van links 1 geleend worden om $10-2=8$ te krijgen, waarmee de 7 een 6 wordt. Al spelend zie je wel hoe het in het algemeen

[15] Ook kolomcijferen, maar door realistisch rekenen guru's mechanisch rekenen genoemd.

gaat, en ook dat $p > a$ gelijkwaardig is met $p + w > a + w$ voor ieder willekeurig ander positief getal w.

Nog een optelvoorbeeldje om het af te leren:

$$\begin{array}{l} 0,12345\ldots \\ 0,99999\ldots \\ \text{————————} + \\ \\ 1,00000\ldots \\ 0,11000\ldots \\ 0,01200\ldots \\ 0,00130\ldots \\ 0,000140.. \\ 0,000015.. \\ \ldots\ldots\ldots\ldots\ldots \\ \text{————————} + \\ \\ 1,12345\ldots \end{array} \tag{44.8}$$

Bij een doorlopend kommagetal het getal $0,\underline{9}$ optellen laat uiteindelijk alle cijfers achter de komma ongemoeid en telt een 1 op bij het getal voor de komma. En zo hoort dat ook. Na wat oefenen lukt dat ook wel in één keer en is het wellicht verstandig om nu verder te gaan met Sectie 44.1.4.

44.1.2 Vermenigvuldigen?

Kunnen we ook vermenigvuldigen? Dit gemene[16] sommetje bijvoorbeeld?

$$\begin{array}{l} 0,999999\ldots\ldots \\ 0,999999\ldots\ldots \\ \text{————————} \times \\ \\ ?????? \end{array}$$

In de vorige subsectie is het gelukt om de som van deze twee getallen kolomcijferend vanaf links zondere hogere wiskunde uit te werken. Kan dat met het produkt ook? We laten ons niet afschrikken en schrijven het produkt cijferend uit, waarbij we het *cijferen* symmetrisch houden in beide factoren, net zoals in (44.6) en (44.7) de uitwerking van de som symmetrisch in de bijdragen van de aparte termen was.

$$\begin{array}{l} 0,999999\ldots\ldots \\ 0,999999\ldots\ldots \end{array}$$

[16] Denk nog niet meteen aan $0,\underline{6} \times 0,\underline{6}$.

$$
\begin{array}{l}
\text{———————} \times \\
0,810000..... \\
0,081000..... \\
0,081000..... \\
0,008100..... \\
0,008100..... \\
0,008100..... \\
0,000810..... \\
0,000810..... \\
0,000810..... \\
0,000810..... \\
0,00008 10... \\
0,00008 10... \\
0,00008 10... \\
0,00008 10... \\
0,00008 10... \\
..... \\
\text{———————} \times \\
\\
????????
\end{array}
\tag{44.9}
$$

Dat ziet er een stuk ingewikkelder uit dan (44.6). Misschien is het wel geen goed idee het produkt van twee kommagetallen zo in één keer te willen doen. In deze doorlopende som zien we tussen de horizontale strepen de termen staan die we krijgen als we het produkt van de eerste decimaal van de eerste factor met de eerste decimaal van de tweede factor nemen (één term), van de eerste met de tweede en de tweede met de eerste (twee termen), van de eerste met de derde, de tweede met de tweede en de derde met de eerste (drie termen), enzovoorts. Gelukkig zien we links steeds meer nullen waardoor het lijkt of het blokje 81 naar rechts opschuift.

Ieder zulk blokje is het produkt van twee decimalen op steeds twee andere posities, decimalen die we hier toevallig allemaal gelijk aan 9 genomen hebben om de som[17] zo moeilijk mogelijk te maken. Het is de positie van het blokje dat opschuift, en op het blokje staat steeds het produkt van twee cijfers. Dus dit zijn de blokjes die voor kunnen komen:

$$00, 01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 12, 14, 15, 16, 18, 20, 24,$$

$$25, 27, 28, 30, 32, 35, 36, 40, 42, 45, 48, 49, 54, 56, 63, 64, 72, 81.$$

[17]Het betreft $1 \times 1 = 1$, maar dat terzijde.

Met alle blokjes gelijk aan 81 gaat het in (44.9) om de som van de getallen in het schema dat begint met

$$\begin{array}{cccccccccc}
 & & & & \frac{81}{10^2} & & & & & \\
 & & & \frac{81}{10^3} & & \frac{81}{10^3} & & & & \\
 & & \frac{81}{10^4} & & \frac{81}{10^4} & & \frac{81}{10^4} & & & \\
 & \frac{81}{10^5} & & \frac{81}{10^5} & & \frac{81}{10^5} & & \frac{81}{10^5} & & \\
 & \frac{81}{10^6} & & \frac{81}{10^6} & & \frac{81}{10^6} & & \frac{81}{10^6} & & \frac{81}{10^6} \\
\frac{81}{10^7} & & \frac{81}{10^7} & & \frac{81}{10^7} & & \frac{81}{10^7} & & \frac{81}{10^7} & & \frac{81}{10^7}
\end{array} \tag{44.10}$$

en naar beneden breder en breder doorloopt. Let wel, de volgorde waarin we cijferend optellen in (44.9) komt overeen met per regel optellen in (44.10) en leidt in de somnotatie tot

$$81 \times \sum_{n=1}^{\infty} \frac{n}{10^{n+1}} \tag{44.11}$$

als maximale uitkomst (vast wel gelijk[18] aan 1) van een produkt van twee *nul komma (minstens) nog wat getallen*.

En met drie zulke getallen gaat het om maximaal

$$\begin{array}{cccccccccc}
 & & & & \frac{729}{10^3} & & & & & \\
 & & & \frac{729}{10^4} & & \frac{729}{10^4} & & \frac{729}{10^4} & & \\
 & & \frac{729}{10^5} & & \frac{729}{10^5} & & \frac{729}{10^5} & & \frac{729}{10^5} & & \frac{729}{10^5} \\
\frac{729}{10^6} & \frac{729}{10^6} & & \frac{729}{10^6} & & \frac{729}{10^6} & & \frac{729}{10^6} & & \frac{729}{10^6} & & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6}
\end{array} \tag{44.12}$$

[18] Equivalent met

$$\sum_{n=1}^{\infty} \frac{n}{10^n} = \frac{10}{81}, \quad \text{wat zou} \quad \sum_{n=1}^{\infty} \frac{n^2}{10^n} \quad \text{zijn? Zie verder.}$$

enzovoorts, met $3, 6, 10, 15, 21, \ldots$ termen op elke regel, en maximaal

$$729 \times \sum_{n=1}^{\infty} \frac{n(n+1)}{2} 10^{n+2}$$

als maximale[19] uitkomst (vast wel gelijk aan 1) van een produkt van drie *nul komma (minstens) nog wat getallen*.

Het wordt er niet eenvoudiger op. We kijken nog een keer naar (44.9) waarmee we begonnen zijn. Het aantal niet-nullen is in de kolommen rechts van de komma achtereenvolgens $1, 3, 5, 7, 9, \ldots$, en

in kolom	1	2	3	4	5	6	7	8	$\ldots$
zien we	1	3	5	7	9	11	13	15	$\ldots$

niet-nullen. Per kolom gaan we bij het optellen dus onvermijdelijk over de 9 heen, en daarbij blijft het niet als we doorcijferen naar rechts, met een ruwe schatting

in kolom	1	2	3	4	5	6	7	8	$\ldots$
maximaal	1×9	3×9	5×9	7×9	9	11×9	13×9	15×9	$\ldots$

voor de kolomsommen, hetgeen leidt tot de vraag of

$$9 \times (\frac{1}{10} + \frac{3}{10^2} + \frac{5}{10^3} + \frac{7}{10^3} + \frac{9}{10^4} + \cdots) = 9 \times \sum_{n=1}^{\infty} \frac{2n-1}{10^n}$$

een decimaal ontwikkelbaar getal definieert waar *elke* eindige som van termen in (44.9) niet boven kan komen, een vraag vergelijkbaar met de minder ruw afgeleide vraag over (44.11). Maar het moge duidelijk zijn dat we opnieuw afdwalen van de basisschoolstof waar het hier toch om zou moeten gaan[20].

Het *cijferen* geeft wellicht meer begrip. Onhandig kolomcijferend zien we in (44.9) kolomsommen

$$8, 17, 26, 35, 44, 53, 62, 71, 80, 89, 98, 107, 116, 125, 134, 143, 152, 161, 170, 179,$$

enzovoorts verschijnen. Cijferend optellen geeft dat met weglating van de nul komma

8	7	6	5	4	3	2	1	0	9	8	7	6	5	4	3	2	1	0	9
1	2	3	4	5	6	7	8	8	9	0	1	2	3	4	5	6	7	7	8
									1	1	1	1	1	1	1	1	1	1

[19] Het betreft immers $1 \times 1 \times 1 = 1$.

[20] Voor een analysecursus zijn dit vragen om te onthouden!

en alles loopt niet alleen naar rechts maar ook naar beneden door.

Opnieuw optellen per kolom geeft

$$\begin{array}{cccccccccccccccccc} 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 8 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 8 & 8 \\ & & & & & & & & 1 & & & & & & & & 1 & 1 \end{array}$$

Enzovoorts. Zo te zien krijgen we op iedere plek inderdaad uiteindelijk een negen, maar alles loopt nog steeds (rechts) naar beneden door, al past het niet meer op de pagina.

De vraag is hoe we uitgaande van dit voorbeeld zien dat er voor twee willekeurige getallen zo altijd een decimale ontwikkeling van het produkt ontstaat, waarmee dan het produkt ondubbelzinnig vast ligt, en ook of er in het geval van het "maximale" voorbeeld alleen maar negens uitkomen. En ook voor produkten van drie getallen natuurlijk, met dezelfde overwegingen als bij optellen[21]. Als we dat wiskundig precies willen maken hebben we nodig dat volgordes en eerst samen nemen niet uit moet maken bij het optellen in doorlopende schema's beginnend als (44.12), als de breedte maar niet te snel toeneemt. Dat idee verkennen we in de volgende subsectie, waarin we opnieuw afdwalen van het cijferen.

44.1.3 Andere aftelbare sommen?

Een analysevraag om te stellen lijkt: voor welke rijen $a_1, a_2, a_3, \ldots$ gehele nietnegatieve getallen correspondeert een aftelbare maar niet eindige som

$$\sum_{n=1}^{\infty} \frac{a_n}{10^n} \tag{44.13}$$

ondubbelzinnig met een getal

$$\sum_{n=1}^{\infty} \frac{d_n}{10^n}$$

waarin alle d_n een cijfer zijn, i.e. $0, 1, 2, 3, 4, 5, 6, 7, 8$ of 9? Het liefst beantwoorden we die vraag zonder over andere uitdrukkingen dan die van de vorm (44.13) te praten.

Een noodzakelijke voorwaarde is dat de eindige sommen

$$A_1 = \frac{a_1}{10}, \quad A_2 = \frac{a_1}{10} + \frac{a_2}{10^2}, \quad A_3 = \frac{a_1}{10} + \frac{a_2}{10^2} + \frac{a_3}{10^3}, \quad \ldots \tag{44.14}$$

allemaal kleiner dan $1 = 0{,}\underline{9}$ zijn. Als $0{,}\underline{9}$ zo'n (strikte) bovengrens is dan is wellicht $0{,}8\underline{9}$ dat ook. Of niet. Kies het minimale cijfer d_1 waarvoor $0{,}d_1\underline{9}$

[21] Lees: $a \times b \times c = (a \times b) \times c = c \times (a \times b)$, weer met alle gepermuteerde variaties.

zo'n bovengrens is. Kies vervolgens het minimale cijfer d_2 waarvoor $0{,}d_1d_2\underline{9}$ een bovengrens is, enzovoorts. Dit proces definieert ondubbelzinnig een getal

$$0 = d_1d_2d_3\cdots = \sum_{n=1}^{\infty} \frac{d_n}{10^n}$$

dat kleiner is dan alle bovengrenzen $0{,}d_1\underline{9}$, $0{,}d_1d_2\underline{9}$, $0{,}d_1d_2d_3\underline{9}$, ..., en voor de bijbehorende

$$D_1 = \frac{d_1}{10}, \quad D_2 = \frac{d_1}{10} + \frac{d_2}{10^2}, \quad D_3 = \frac{d_1}{10} + \frac{d_2}{10^2} + \frac{d_3}{10^3}, \quad \ldots,$$

geldt dat

$$D_1 + \frac{1}{10}, \quad D_2 + \frac{1}{10^2}, \quad D_3 + \frac{1}{10^3}, \quad \ldots$$

bovengrenzen zijn. We kunnen niet uitsluiten dat na verloop van tijd alle d_n nul zijn, maar ze zijn zeker niet allemaal nul.

Kan het zo zijn dat de A_n-tjes niet boven de D_1 uitkomen? Wel, in dat geval zijn alle $A_n < D_1$ (want met $A_n = D_1$ komt een volgende A_n boven D_1), voor alle $n = 1, 2, 3, \ldots$, en was d_1 kennelijk niet minimaal gekozen om alle A_n onder $0{,}d_1\underline{9}$ te hebben. Dus A_n komt wel boven D_1 en blijft dan groter dan D_1. Hetzelfde geldt met het zelfde argument voor D_2, D_3, etcetera.

Als n_1 de eerste n is waarvoor $A_n > D_1$, n_2 de eerste n waarvoor $A_n > D_2$, n_3 de eerste n waarvoor $A_n > D_3$, etcetera, dan is n_2 minstens n_1, n_3 minstens n_2, n_4 minstens n_3, enzovoorts. We concluderen dat voor iedere k geldt dat

$$D_k < A_n < D_k + \frac{1}{10^k} \tag{44.15}$$

voor alle n vanaf $n = n_k$, en dat zou moeten betekenen dat

$$A_n \to D = 0{,}d_1d_2d_3d_4\ldots, \tag{44.16}$$

een nog niet precies gemaakte uitspraak voor een rij breuken A_n, breuken met noemers machten van 10 en $A_n \leq A_{n+1}$, met strikte ongelijkheid voor niet per se alle maar wel willekeurig grote n.

Elke A_n heeft decimalen genummerd door $j = 1, 2, 3, 4, \ldots$. De eerste decimaal kan niet kleiner worden met toenemende n. Dat betekent dat vanaf zekere $n = m_1$ de eerste decimaal van A_n niet meer verandert en gelijk is aan een vast cijfer α_1. Daarna geldt hetzelfde voor de tweede decimaal die vanaf zekere $n = m_2$ (waarbij we m_2 minstens gelijk aan m_1 kunnen nemen) niet meer verandert en gelijk is aan een vast cijfer α_2, enzovoorts.

Deze eigenschap moet voor de niet-dalende rij $A_,A_2, A_3, \ldots$ toch wel de enige zinvolle definitie van

$$A_n \to 0,\alpha_1\alpha_2\alpha_3\alpha_4\ldots$$

zijn. Graag zouden[22] we nu uit (44.15) concluderen dat

$$0,d_1d_2d_3d_4\cdots = 0,\alpha_1\alpha_2\alpha_3\alpha_4\ldots,$$

waarbij we opmerken dat de ontwikkeling in het rechterlid bij constructie niet af kan breken maar de ontwikkeling in het linkerlid wel. Het kan dus gebeuren dat de eerste zoveel α_n en d_n hetzelfde zijn, daarna één keer $\alpha_n + 1 = d_n$, en vervolgens alle $d_n = 0$ en alle $\alpha_n = 9$. Hoe het ook zij, de uitdrukking in (44.13) definieert dus ondubbelzinnig een *nul komma minstens nog wat getal*, mits we weten dat alle eindige sommen in (44.14) kleiner zijn dan $0,\underline{9}$. Maar wie voor (44.10) en (44.12) meteen ziet dat dat inderdaad zo is mag het zeggen. We zijn er dus nog niet uit wat betreft produkten van positieve kommagetallen.

44.1.4 Een cijfer keer een kommagetal

Terug naar het cijferen. We houden ons nog even aan de afspraak dat positieve kommagetallen de getallen zijn met een na de komma doorlopende rij cijfers waarin niet-nullen blijven voorkomen hoe ver je ook gaat in de decimale ontwikkeling. Zo'n positief kommagetallen heeft voor de komma een natuurlijke getal of een 0 staan. Het produkt van twee zulke getallen moet wel de som van vier bijdragen zijn: wat je krijgt van voor de komma keer voor de komma, van voor de komma keer achter de komma, van achter de komma keer voor de komma, en van achter de komma keer achter de komma.

De laatste lijkt het moeilijkst. Als we die kunnen dan kunnen we daarna ook alle produkten van positieve kommagetallen door eerst de komma's naar links te schuiven en in het antwoord de komma naar rechts te schuiven. Twee keer naar rechts eigenlijk, om beide verschuivingen naar links goed te maken. Helaas zijn we hierboven nog niet bevredigend uit produkten van zulke *nul komma nog wat getallen* gekomen.

De eerste van de vier bijdragen is het makkelijkst, hoe het daarmee zit is basisschoolstof. De volgende twee bijdragen zijn wat lastiger. Met een 1-cijferig natuurlijk getal 1, 2, 3, 4, 5, 6, 7, 8 of 9 is de moeilijkste $9 \times 0,\underline{9}$. Net zo moeilijk is $0,9 \times 0,\underline{9}$:

[22] Nog even nagaan dit dus.

$$
\begin{array}{l}
0,999999..... \\
0,9 \\
\hline\times \\
0,810000..... \\
0,081000..... \\
0,008100..... \\
0,000810..... \\
0,0000810... \\
..... \\
\hline\times \\
0,899999.....
\end{array}
$$

Daarna zijn produkten van cijfers met kommagetallen geen probleem meer. Met twee cijfers tegelijk in elke stap geeft een cijfers keer een blokje van twee maximaal $9 \times 99 = 891$. Cijferend per blokjes van twee vanaf links schuift er dus steeds maximaal een 8 naar links door. Bij het eerste blokje komt die gewoon voor het blokje te staan. Van het tweede blokje schuift er maximaal een 8 door naar links waarmee het blokje dat daar maximaal voor staat op zijn hoogst $91 + 9 = 99$ wordt. Enzovoorts. Het is weer instructief om een paar voorbeeldjes te doen en in één keer het antwoord op te schrijven op basis van de decimalen die je hebt in je voorbeeld.

44.1.5 Produkten van kommagetallen

Als het bovenstaande eenmaal in in één keer lukt als

$$
\begin{array}{l}
0,999999..... \\
0,9 \\
\hline\times \\
0,899999.....
\end{array}
$$

dan kan daarna

$$
\begin{array}{l}
0,999999..... \\
0,999999..... \\
\hline\times \\
0,899999..... \\
0,089999..... \\
0,008999..... \\
0,000899..... \\
0,000089.....
\end{array}
\qquad (44.17)
$$

enzovoort ook, en vervolgens kunnen we dan van boven af de kommagetallen term voor term optellen met wat we kolomcijferend geleerd hebben in sommetjes als (44.8).

De eerste stap is

$$
\begin{array}{l}
0,899999..... \\
0,089999..... \\
\text{————————} + \\
\\
0,899999..... \\
0,009999..... \\
0,080000..... \\
\text{————————} + \\
\\
0,909999..... \\
0,080000..... \\
\\
\text{————————} + \\
\\
0,989999.....
\end{array}
\qquad (44.18)
$$

In (44.18) hebben we de tweede rij negens afgesplitst. Opgeteld bij het kommagetal erboven verhogen die de 89 tot 90, en met de 8 eronder maken ze van de 89 een 98, waarbij de decimalen achter de 89 ongewijzigd blijven. Het resultaat is de som van de eerste twee kommagetallen in (44.17), waarbij in dit voorbeeld de 8 eentje opgeschoven is naar rechts.

Zo gaat dat verder. Nu we met (44.17) zijn gevorderd tot

$$
\begin{array}{l}
0,999999..... \\
0,999999..... \\
\text{————————} \times \\
\\
0,989999..... \\
0,008999..... \\
0,000899..... \\
0,000089..... \\
.....
\end{array}
\tag{44.19}
$$

zien we dat het patroon zich herhaalt in

$$
\begin{array}{l}
0,989999..... \\
0,008999..... \\
\text{————————} + \\
\\
0,998999.....
\end{array}
$$

met als resultaat de som van de eerste drie kommagetallen in (44.17). De 8 is weer eentje opgeschoven en dat gaat zo door. In de volgende stap zien we

$$
\begin{array}{l}
0,999999..... \\
0,999999..... \\
\text{————————} \times \\
\\
0,998999..... \\
0,000899..... \\
0,000089..... \\
.....
\end{array}
\tag{44.20}
$$

met nu boven de drie nullen na de komma in (44.20) alleen het derde cijfer dat nog zal veranderen bij verder cijferen. Zo vinden we al cijferend dat

$$0,\underline{9} \times 0,\underline{9} = 0,\underline{9},$$

hetgeen zoveel wil zeggen dat $1 \times 1 = 1$.

Is ieder tweetal kommagetallen zo cijferend met elkaar te vermenigvuldigen? Merk op dat een staartstuk in de ontwikkeling van de tweede factor steeds maximaal uit een rij negens bestaat en zo het cijfer in het antwoord op de positie waarna dat staartstuk begint maximaal met 1 verhoogt.

Om nog verder uit te werken dit alles, maar niet hier. Het idee is wel duidelijk nu. Zonder hier nu meteen Turing aan te roepen is het aardig om deze sectie te besluiten met de opmerking dat je in gedachten een machientje zou kunnen maken dat als input de doorlopende kommagetallen krijgt die als het ware van de ene kant cijfer voor cijfer naar binnen schuiven, en dan vervolgens aan de andere kant als output de som of produkt cijfer voor cijfer als doorlopend kommagetal uitspuugt, en het machientje daarmee tot het einde der tijden doorgaat.

44.2 Kleinste bovengrenzen

Net als de aftelbare som in (44.1) met $k \geq 0$ is (44.3) een mooi voorbeeld van

$$\sum_{n=0}^{\infty} a_n \tag{44.21}$$

met $a_n \geq 0$ voor alle $n \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Als de partiële sommen

$$S_N = \sum_{n=0}^{N} a_n$$

begrensd zijn dan is de *kleinste bovengrens* van de aftelbare vereniging

$$\cup_{N \in \mathbb{N}_0} \{S_N\} = \{S_0, S_1, S_2, \dots\}$$

per definitie de som van de *reeks* in (44.21), notatie

$$S = \sum_{n=0}^{\infty} a_n.$$

In het geval van (44.1) is $S_N \leq k + 1$ voor alle $N \in \mathbb{N}_0$ en kan deze uitspraak dus als *tautologie* gezien worden: het reële getal S is de limiet van zijn decimale ontwikkeling, een ontwikkeling waarin de decimalen d_n uit de cijfers 0 tot en met 9 gekozen worden.

Dat het überhaupt mogelijk is dat er uit een som met oneindig veel termen als (44.21) een eindig getal kan komen is zo vanuit (44.1) vanzelfsprekend, ook al dacht ene Zeno daar destijds anders over. Mooie voorbeelden waarbij er uit de som geen eindig getal komt zijn

$$S = \sum_{n=0}^{\infty} 1 \quad \text{met} \quad S_N = N, \quad \text{en} \quad S = \sum_{n=0}^{\infty} \frac{1}{n}. \tag{44.22}$$

Geen van deze twee definieert een $S \in \mathbb{R}^+$.

Waarom eigenlijk niet? Wel, de eerste S zou een kleinste bovengrens in $\mathbb{R}$ voor de verzameling $\mathbb{N}$ zijn. Maar dan is $S - \frac{1}{2}$ geen bovengrens voor $\mathbb{N}$. En dus is er een $N \in \mathbb{N}$ met $N > S - \frac{1}{2}$ en volgt dat $N + 1 > S + \frac{1}{2}$. Maar $N + 1 \in \mathbb{N}$ dus is S geen bovengrens voor $\mathbb{N}$, een tegenspraak[23]. Gelukkig maar, want het zou wel heel gek zijn als $\mathbb{N}$ wel begrensd is in $\mathbb{R}$. Komt meteen te pas bij het tweede voorbeeld in (44.22), waarover we opmerken dat

$$\underbrace{1 + \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{>\frac{1}{2}}}_{>1} + \underbrace{\underbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}_{>\frac{1}{2}} + \underbrace{\frac{1}{9} + \frac{1}{10} + \frac{1}{11} + \frac{1}{12} + \frac{1}{13} + \frac{1}{14} + \frac{1}{15} + \frac{1}{16}}_{>\frac{1}{2}}}_{>1},$$

enzovoorts, en zo komen de bijbehorende S_N boven elke $n \in \mathbb{N}$. Ook niet begrensd dus. Maar de som

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots$$

heeft wel een uitkomst[24], althans indien opgevat als

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n},$$

al zijn noch de positieve termen

$$1 + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \cdots,$$

noch de negatieve termen

$$-\frac{1}{2} - \frac{1}{4} - \frac{1}{6} - \cdots$$

op te tellen tot een eindige som. Sterker, gegeven een $S \in \mathbb{R}$ kun je de positieve en negatieve termen verweven[25] tot een rij a_n op zo'n manier dat

$$S = a_1 + a_2 + a_3 + a_4 + \cdots,$$

een goede reden om zoveel mogelijk alleen maar over reeksen zoals in Sectie 44.3 te spreken.

[23] Overtuigd?

[24] Ik meen $\ln 2$.

[25] Kies positieve termen om boven S te komen, dan negatieve om onder S, dan ...

44.3 Absoluut convergente reeksen

Als we hadden leren rekenen met $d_n \in \{-5,-4,-3,-2,-1,0,1,2,3,4\}$ en de tafels tot en met vijf, dan was (44.1) een voorbeeld geweest van (44.21) zonder de a priori informatie dat $a_n \geq 0$ maar *wel* met de eigenschap dat

$$\sum_{n=0}^{\infty} |a_n| < \infty, \tag{44.23}$$

omdat

$$\sum_{n=1}^{\infty} |\frac{d_n}{10^n}| \leq \sum_{n=0}^{\infty} \frac{5}{10^n} = \frac{5}{10} + \frac{5}{100} + \frac{5}{1000} + \frac{5}{10000} + \frac{5}{100000} + \cdots = \frac{5}{9}.$$

Ook nu geldt dat $S_N \to S$ voor een unieke $S \in \mathbb{R}$, dus

$$S = \sum_{n=0}^{\infty} a_n, \tag{44.24}$$

en hernummeren van de som verandert niets aan die uitkomst. Reeksen van de vorm (44.21) waarvoor (44.23) geldt heten *absoluut convergent* en zijn *onvoorwaardelijk convergent*: de volgorde van sommeren maakt niet uit voor de waarde S van de som en bovendien geldt dat

$$|S| = |\sum_{n=0}^{\infty} a_n| \leq \sum_{n=0}^{\infty} |a_n|. \tag{44.25}$$

Wat betreft het bewijs van (44.24) gegeven (44.23), de invariantie onder hernummeren en de aftelbare 3-hoeksongelijkheid (44.25): dat bewijs maakt gebruik van het feit dat in de reële getallen *Cauchyrijen*, dat zijn rijen waarvoor geldt dat

$$x_n - x_m \to 0 \quad \text{als} \quad m,n \to \infty,$$

een unieke limiet $\bar{x}$ hebben, een limiet $\bar{x}$ die bestaat als dan inderdaad het *enige* reële getal waarvoor

$$x_n \to \bar{x} \quad \text{als} \quad n \to \infty.$$

Zulke rijen heten *convergent*.

Uit Hoofdstuk 10 van [HM] of Hoofdstuk 8 van het *Basisboek Wiskunde* is de lezer wellicht al bekend met de wiskundige definitie van het begrip (limiet van een) convergente rij, waarin alleen [26] de "voor alle $p > 0$" nog door een

[26] Didactisch aardig in het basisboek is het gebruik van grote P naast kleine p.

"voor alle $\varepsilon > 0$" moet worden vervangen om tot het gebruikelijke jargon te komen, en later eventueel door $\forall \varepsilon > 0$. Wel is het in de analyse straks *praktischer* om met

$$|x_n - \bar{x}| \leq \varepsilon$$

te werken.

Wat ook elegant en praktisch is in het Basisboek Wiskunde is de zorgvuldige manier waarop gesproken wordt over *de rij waarvan het n-de element gelijk is aan* x_n, en het aan de lezer wordt overgelaten zich daarbij te realiseren dat n de getallen $1, 2, 3, 4, \ldots$ doorloopt, of een andere steeds met stap 1 oplopende rij gehele getallen. Wij zullen de notatie in het Basisboek Wiskunde afkorten tot simpelweg *de (door n genummerde) rij* x_n, vaak de rij reële getallen x_n. Evenzo spreken we over de rij rationale getallen q_n of de rij $q_n \in \mathbb{Q}$. De laatste notatie wordt hieronder nog gebruikt.

De ε-definities van uitspraken als hierboven komen in deze cursus aan de orde op het moment dat dat nodig is. Want ze zijn nodig, bijvoorbeeld om precies te maken dat sommen als (44.24) bestaan du moment dat je met één $M \in \mathbb{R}^+$ een schatting

$$\sum_{n=0}^{N} |a_n| \leq M$$

hebt voor alle partiële sommen *tegelijk*, en daaruit afleidt dat de door N genummerde rij S_N een Cauchyrij is. We merken hierbij op dat het in zogenaamde genormeerde ruimten dan om twee equivalente uitspraken gaat, uitspraken waarin noch de limiet $\bar{x}$ van de rij, noch de som S van de reeks waar het om gaat expliciet voorkomen:

absoluut convergente reeksen convergent $\iff$ Cauchyrijen convergent

Je kunt dus weten of $\bar{x}$ en S in $\mathbb{R}$ bestaan zonder ze eerst te hebben bepaald.

44.4 Verzamelingen in de praktijk

Voor sommige wiskundigen van de meer zuivere inclinatie zijn de uitspraken hierboven niet los te zien van een precieze maar voor de analyse zelf niet altijd even verhelderende wiskundige constructie van de reële getallen. Maar interessant zijn die constructies natuurlijk wel, en je moet ergens beginnen als je de wiskunde per se axiomatisch en wiskundig streng wil opzetten[27], vanuit wat men de leer van verzamelingen noemt.

Deze verzamelingenleer is iets waarover Paul Halmos in zijn mooie boekje Naive Set Theory[28] schreef: alle wiskundigen vinden dat je er wat van gezien

[27] Een vriendje van Einstein heeft helaas laten zien dat dat nooit bevredigend zal lukken.
[28] Vertaald ooit als Prisma pocket verkrijgbaar.

moet hebben, maar ze zijn het oneens over *wat* precies. Je kunt verzamelingenleer bijvoorbeeld bij het begin beginnen met het axioma dat de lege verzameling[29] bestaat.

Dat doen wij hier niet. Maar mocht je dat wel doen dan komen toch op enig moment ook de axioma's voor de verzameling van de natuurlijke getallen $\mathbb{N}$ voorbij, natuurlijke getallen die iedereen die op zijn vingers heeft leren tellen allang kent. En tellen begint natuurlijk bij 1[30], al is het handig om de verzameling

$$\mathbb{N}_0 = \mathbb{N} \cup \{0\} = \{0, 1, 2, 3, \dots\}$$

in te voeren, hier in een zuiver wiskundig gezien af te keuren maar wel zo begrijpelijke notatie met onfatsoenlijke stippeltjes, waarin we het hierboven al gebruikte verenigingssymbool $\cup$ weer terugzien[31].

Het is wel goed om één van die axioma's voor $\mathbb{N}$ te relateren aan de wiskundige praktijk van alledag. Want hoe bewijs je bijvoorbeeld dat voor iedere $N \in \mathbb{N} = \{1, 2, 3, \dots\}$ geldt dat de uitspraak

$$(P_N) \qquad 1^2 + 2^2 + \cdots + N^2 = \frac{N^3}{3} + \frac{N^2}{2} + \frac{N}{6}$$

waar is, *zonder* voor elke $N \in \mathbb{N}$ *apart* de uitspraak (P_N) te moeten controleren?

Binnen de zuivere wiskunde hoort daar een verhaal bij waarin voor al de puntjes hierboven eigenlijk geen plaats is. Dat verhaal eindigt met het principe van volledige inductie[32], dat er op neerkomt dat[33] als je voor $N = 1$ de uitspraak controleert, en je vervolgens laat zien dat de *implicatie*

$$(P_N) \implies (P_{N+1}) \tag{44.26}$$

geldt voor alle N waarvoor je hem nodig hebt, namelijk om via herhaald toepassen van de inductiestap (44.26) tot

$$(P_1) \implies (P_2) \implies (P_3) \implies (P_4) \implies (P_5) \implies (P_6) \implies (P_7) \implies \cdots$$

te komen, zover als je maar wil, de uitspraak inderdaad geldt voor alle $N \in \mathbb{N}$. De implicatie (44.26) moet daartoe voor alle $N \in \mathbb{N}$ worden aangetoond om beginnend met de juistheid voor $N = 1$ de keten hierboven zonder die stippeltjes in één keer af te maken.

[29] In LaTeχ: $\emptyset$. Op het schoolbord liever $\not\emptyset$.
[30] Tellend is $\mathbb{N}$ met nul een beetje flauwe kul, op 0^0 komen we nog terug.
[31] Gaat er dus eigenlijk om hoe je al die getallen zonder stippels tussen accolades vangt.
[32] Naamgeving volledig intimiderend, vriendelijker is: dominoprincipe.
[33] Nu komt een lange zin.

Dit soort oefeningen kunnen elders gedaan worden. Zie bijvoorbeeld in [HM] Sectie 6.1 en voetnoot 16. Relevant uit die sectie voor deze cursus zijn bewijzen voor rekenpartijtjes als in (P_N) hierboven, waarin niet alleen N maar ook $3 = p \in \mathbb{N}$ een parameter is, en de inductiestap van de vorm

$$(P_1) \wedge (P_2) \wedge \cdots \wedge (P_N) \implies (P_{N+1}) \tag{44.27}$$

is[34]. Zie verder ook Hoofdstuk IV [BE] .

Wat we hier precies met $(A) \implies (B)$ bedoelen moge duidelijk zijn: als de uitspraak (A) waar is dan is ook de uitspraak (B) waar. Hetgeen in onze wiskundige redenaties equivalent is met: als uitspraak (B) niet waar is dan kan uitspraak (A) ook niet waar zijn. Deze logica kan geformaliseerd worden met waarheidstabellen vol nullen en enen opgeleukt met bijzonder fraaie algebra, maar wat dat betreft laten we hier liever de Boole de Boole[35].

Du moment dat er over het bestaan van[36] $\mathbb{N}$ geen twijfel meer is, worden in de verzamelingsleer, bijvoorbeeld zoals in Hoofdstuk V en VI van [BE], $\mathbb{Z}$, $\mathbb{Q}$ en uiteindelijk $\mathbb{R}$ *wiskundig netjes* geconstrueerd. De constructie van $\mathbb{R}$ is in [BE] gebaseerd op de gedachte dat iedere manier om $\mathbb{Q}$ in twee stukken te knippen overeen zou moeten komen met een reëel getal, waarbij de rationale getallen dan wel met de schaar te maken krijgen en de overige getallen niet[37].

Het is instructief om de constructies van $\mathbb{Z}$ en $\mathbb{Q}$ uit $\mathbb{N}$ met elkaar te vergelijken zoals dat gebeurt in [BE]. Die van $\mathbb{Z}$ is inderdaad tamelijk kunstmatig. Die van $\mathbb{Q}$ is echter heel natuurlijk en gebaseerd op hoe je eigenlijk altijd al met de rationale getallen rekende, namelijk als breuken. Breuken met een teller en een noemer. Bijvoorbeeld

$$\frac{14}{333} = \frac{42}{999} = 0.\underline{042}$$

met een streep die aangeeft dat de decimale ontwikkeling van de breuk zich herhaalt. Anders dan gesuggereerd in de in

`http://www.few.vu.nl/~jhulshof/TAL.pdf`

besproken TAL-boekjes van het Freudenthal Instituut doe je echter het rekenen met rationale getallen bij voorkeur niet met zulke decimale ontwikkelingen, maar juist wel met de niet unieke representatie van rationale getallen als quotiënten van gehele getallen, dus in de vorm

$$q = \frac{t}{d}$$

[34] De $\wedge$ staat voor "en", dat is logisch. Denken aan dominosteentjes is nu lastiger.
[35] https://www.youtube.com/watch?v=DOzqUyW7jog
[36] Eventueel via Peano's axioma's.
[37] Want ze bestaan op dat moment nog niet.

met teller t en noemer d in $\mathbb{Z}$, de d niet gelijk aan 0, waarbij je moet afspreken dat

$$\frac{t_1}{d_1} = \frac{t_2}{d_2} \quad \text{als} \quad t_1 d_2 = t_2 d_1.$$

44.5 Equivalentierelaties

Wiskundigen noemen zo'n afspraak een equivalentierelatie. We komen nu in relatie tot $\mathbb{R}$ meer over dit belangrijke begrip te spreken, ook voor wie van $\mathbb{R}$ graag een inzichtelijke constructie wil zien. Een constructie waarvan de details overigens niet thuis horen in of voorafgaand aan een eerste vak Analyse. Ik meen dat ik zelf de constructie van $\mathbb{R}$ voor het eerst zag bij een college over de integraal van Lebesgue van Jan van de Craats in het vierde semester van wat toen de kandidaatsstudie wiskunde in Leiden was.

De onderliggende maattheorie voor dat vak over die andere integraal begint met de vraag wat de *oppervlakte* $|A|$ is van een willekeurige deelverzameling A van $\mathbb{R}^2$, en komt onvermijdelijk tot twee constateringen. Vroeger of later zijn dat respectievelijk

(i) het komt voor dat $A \subset B$ en $|A| = |B|$;

(ii) het zou kunnen voorkomen dat A eindige oppervlakte $|A|$ heeft maar opgeknipt kan worden in aftelbaar veel stukjes die allemaal dezelfde maat zouden moeten hebben[38],

en daar moet je mee omgaan. Leuk is dat (ii) ons dan later[39] weer terugvoert naar het boekje van Halmos. In een vroeger stadium doet (i) ons echter al het dringende verzoek om A en B in zekere[40] zin als hetzelfde te zien, en bijvoorbeeld ook hetzelfde als een C met $C \subset A$ en $|C| = |A|$, waarbij C geen deelverzameling van B hoeft te zijn of omgekeerd. Hoe formuleer je dan rechtstreeks dat B en C equivalent zijn?

Anders van aard is het gebruik van equivalentierelaties bij een inzichtelijke constructie van $\mathbb{R}$, waarbij je denkt aan reële getallen als denkbeeldige limieten van Cauchyrijtjes rationale getallen, zoals bijvoorbeeld de hierboven besproken decimale ontwikkelingen, maar dan moet je wel een goede afspraak maken over wat het betekent dat twee zulke Cauchijrijtjes hetzelfde reële getal (zouden moeten) definiëren. Denk bijvoorbeeld aan binaire benaderingen met alleen maar nullen en enen, of aan benaderingen met kettingbreuken, allebei erg fraai of juist minder[41] fraai, omdat ze afstand nemen

[38] Waarom is dat een paradox?

[39] Maar niet hier.

[40] Lees: in maatheoretische zin.

[41] Over gebrek aan smaak valt niet te twisten.

van de vingers waarin onze wiskunde zit. Kortom, een belangrijke vraag is hoe je van twee Cauchyrijen rationale getallen q_n en r_n zegt dat ze hetzelfde reële getal definiëren[42].

Als je er even over nadenkt is het logisch dat dit een definitie zou kunnen zijn:

$$q_n \sim r_n \quad \Longleftrightarrow \quad q_n - r_n \to 0 \quad \text{als} \quad n \to \infty$$

Deze tweezijdige equivalentiepijl definieert een *equivalentierelatie* op de verzameling van alle rijen rationale getallen. We walsen nu wellicht even over wat belangrijke details heen, maar een equivalentierelatie is niets anders dan een relatie met formeel dezelfde eigenschappen als de gelijkheidsrelatie voor elementen van een willekeurige verzameling A. Voor alle $a, b, c \in A$ geldt

$$a = a,$$

$$a = b \implies b = a,$$

$$a = b \wedge b = c \implies a = c$$

De relatie[43] gedefinieerd door het $=$ teken heet daarom reflexief, symmetrisch en transitief, en ook $\sim$ is zo'n equivalentierelatie, op de verzameling van alle rijen rationale getallen in dit geval. En die equivalentierelatie doet het!

Wat doet $\sim$ dan? De equivalentierelatie $\sim$ deelt de verzameling van alle rijen rationale getallen in. Waarin? In equivalentieklassen natuurlijk. Iedere rij $r_n \in \mathbb{Q}$ definieert een equivalentieklasse

$$[r_n] = \{q_n \in \mathbb{Q} : \, q_n \sim r_n\} \tag{44.28}$$

waar die rij zelf in zit, en een reëel getal is *per definitie* de *equivalentieklasse* van een Cauchyrij $r_n \in \mathbb{Q}$.

So much for the construction of the real numbers en we zullen het $\sim$ tekentje nu weer in laten leveren, omdat we dat symbool toch liever gebruiken als

$$x_n \sim y_n \quad \Longleftrightarrow \quad \frac{x_n}{y_n} \to 1 \quad \text{als} \quad n \to \infty$$

voor een andere en in de praktijk vaker gebruikte equivalentierelatie[44] op de verzameling van alle reële rijen. Een voorbeeld is

$$n! \sim (\frac{n}{e})^n \sqrt{2\pi n},$$

uitvoerig besproken in [HM].

[42] In je hoofd of op de getallenlijn.

[43] Ook dat woord heeft een *wiskundige* definitie natuurlijk.

[44] Wel een goede vraag hierboven is: wat is de beste representant?

44.6 Analyse in en van wat?

Of er nog andere verzamelingen zoals deze $\mathbb{R}$ zijn is niet een standaardvraag om hier te stellen. Wel belangrijk voor een eerste vak over analyse is dat de *rationale getallen* $\mathbb{Q}$, dat zijn de getallen die ontstaan als quotiënten van getallen in $\mathbb{Z}$ en getallen in $\mathbb{N}$, de "Cauchy eigenschap" niet hebben. Dat is de reden waarom we de analyse in $\mathbb{R}$ doen, het unieke geordende getallenlichaam waarin (alle) Cauchyrijen en absoluut convergente reeksen convergent zijn.

In het Basisboek Wiskunde worden deze getallen besproken in Hoofdstuk 24, en we gebruiken vrijwel dezelfde notaties, met de accolades ook. Lees ook Hoofdstuk 25 nog even door, we nemen de daar gebruikte input-output voorstelling voor functies[45] hier graag over als

$$x \xrightarrow{f} f(x) \quad \text{en} \quad D_f \xrightarrow{f} \mathbb{R}$$

met D_f het domein van f. Het bereik en de grafiek[46] van f zijn

$$B_f = \{f(x) : x \in D_f\} \quad \text{en} \quad G_f = \{(x, y) : x \in D_f,\, y = f(x)\}.$$

Soms zullen we liever over functies $f : \mathbb{R} \to \mathbb{R}$ spreken die op een bepaalde deelverzameling van $\mathbb{R}$ een bepaalde eigenschap hebben. Het *domein* D_f is dan de verzameling bestaande uit alle $x \in \mathbb{R}$ waarvoor $f(x)$ gedefinieerd is. Is het domein van f niet heel $\mathbb{R}$, dan kun je natuurlijk altijd $f(x)$ voor x-waarden buiten het domein een waarde geven die je toevallig goed uitkomt, nul bijvoorbeeld[47].

In deze cursus behandelen we ondermeer de analyse die de calculus onderbouwt voor functies $f : I \to \mathbb{R}$ met $I \subset \mathbb{R}$ een interval. Vaak, met $a, b \in \mathbb{R}$, is I daarbij een gesloten begrensd interval

$$I = [a, b] = \{x \in \mathbb{R} : a \leq x \leq b\},$$

of een open begrensd interval

$$I = (a, b) = \{x \in \mathbb{R} : a < x < b\}.$$

We beginnen met integraalrekening, eerst voor monotone functies, zonder over limieten te spreken, en daarna voor uniform continue functies $f : [a, b] \to \mathbb{R}$, waarbij we voor het eerst het limietbegrip tegenkomen en nodig hebben.

Voor zulke functies wordt

$$\int_a^b f(x)\, dx$$

via benaderende sommen gedefinieerd in relatie tot wat de oppervlakte van het gebied ingesloten door $x = a$, $x = b$, $y = 0$ en $y = f(x)$ in het x, y-vlak moet zijn in het geval dat f een positieve functie is. Je zou kunnen zeggen dat dit de eerste *probleemstelling* is in dit boek, geformuleerd in drie punten als:

Teken plaatje!

hoe definieer je de oppervlakte van niet meteen arbitraire verzamelingen;

en hoe reken je die vervolgens uit?

wat kun je vervolgens leren van de oplossing?

Dat laatste doe je dan wellicht zonder meteen een nieuw probleem te willen formuleren. Spelen met de verworven inzichten zonder een concreet doel op zich.

Bijvoorbeeld: met een variabele bovengrens in de integraal ontdekken we de opzet van de differentiaalrekening met behulp van lineaire benaderingen. Die werken we later uit voor *machtreeksen*

$$P(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \cdots ,$$

waarmee we een grote klasse van standaardfuncties tot onze beschikking krijgen, waarvoor "mag dat" vragen kort maar krachtig met "ja natuurlijk" te beantwoorden zijn. Binnen die klasse is de analyse namelijk ondergeschikt aan de algebra, en zodra je die algebra goed begrijpt, voor x^7 of zo, ben je wel klaar en daarmee dient een andere probleemstelling zich aan:

Hoe zit het met al die andere functies?

Als we buiten de klasse van machtreeksen treden verandert alles en moet er gewerkt worden. Dat werk halen we nu naar voren, waar we dat in [HM] zo lang mogelijk uitstelden.

De *middelwaardestelling* blijkt het belangrijkste hulpmiddel om ogenschijnlijk evidente uitspraken ook werkelijk te bewijzen. De uitspraak van die stelling is dat differentiequotiënten als

$$\frac{F(b) - F(a)}{b - a},$$

de richtingcoëfficiënt van

$$\text{de lijn door } (x, y) = (a, F(a) \text{ en } (x, y) = (b, F(b))$$

[45] Van het Latijnse fungor (deponens: een passieve vorm met actieve betekenis).

[46] Vaak slordig: de grafiek $y = f(x)$ in het x, y-vlak.

[47] Zoals wel eens voorgesteld in relatie tot $x \to \frac{1}{x}$ en het rekenonderwijs.

in het x, y-vlak, zelf doorgaans gelijk zijn aan de richtingscoëfficiënt van de raaklijn aan de grafiek van F in een punt met x-waarde tussen a en b, lees: aan de met de differentiaalrekening gedefinieerde

$$\text{afgeleide van } F'(x) \text{ van } F(x) \text{ in een tussenpunt.}$$

Is $F'(x)$ overal tussen a en b gelijk aan nul, dan is $F(x)$ kennelijk constant, aldus de NIET-TRIVIALE STELLING in [HM]. Die STELLING is niet zozeer de oplossing van een probleem, maar formuleert juist iets dat je zeker wil weten bij het oplossen van (bijvoorbeeld) differentiaalvergelijkingen. Het bewijs van de STELLING maakt essentieel gebruik van een fundamentele stelling over het bestaan van convergente deelrijen, die, triviaal[48] of niet, toch maar een apart hoofdstuk krijgt, waarin een wat minder bekende stelling die ik ken via Han Peters wordt geformuleerd.

Vergelijkingen oplossen is een belangrijke tak van niet alleen maar recreatieve *sport*[49] in de wiskunde. In de context van vergelijkingen van de vorm $F(x, y) = 0$, waarbij F een functie is van twee variabelen, introduceren we daarom ook meteen maar het begrip impliciete functie, met als speciaal geval het al behandelde begrip inverse functie. Het bewijs van de impliciete functiestelling draaien we binnenste buiten in een aparte sectie, gevolgd door twee secties waarin weer met verworven inzichten wordt gespeeld en een basis wordt gelegd voor alles dat later komt. Na een zijstapje over de methode van Newton wordt de basale theorie afgesloten met differentiaalrekening voor integralen met parameters, en partieel integreren en een stelling over Taylorbenaderingen met polynomen.

Daarna nemen we de tijd voor voorbeelden en meer voorbeelden, en herhalen de rekenregels nog een keer in de kale context van functies van één variabele zonder er functies van x en y bij te halen. We gaan uitvoerig in op de natuurlijke logaritme ln als inverse van de exponentiële functie exp en introduceren in die context ook zogenaamde asymptotische formules, waarvan de formule van Stirling[50] voor $n!$ als $n \to \infty$ een mooi voorbeeld[51] is.

In het tweede deel kunnen we de meeste van de in het eerste deel geformuleerde definities, stellingen en bewijzen uit de differentiaalrekening voor functies van $\mathbb{R}$ naar $\mathbb{R}$ vrijwel letterlijk overnemen. Alleen de notaties hoeven nog te worden uitgepakt. We beginnen daartoe met $\mathbb{C}$, de verzameling van de complexe getallen, en een in ons Leidse wat vergeten maar wel zo snel bewijs van de hoofdstelling van de algebra. Daarna komen functies van $\mathbb{C}$ naar $\mathbb{C}$ en afbeeldingen en functies met meerdere variabelen. Lineaire

[48] Denk ook aan valsspelen met meetwaarden.

[49] Geen sport zonder *techniek*.

[50] De voorbeeldformule met $\sim$ een paar pagina's terug, uit te spreken als "twiddles".

[51] En buitengewoon relevant voor probleemstellingen in de natuurkunde.

functies beschrijven we dan in matrixnotatie, en matrixrekening behandelen we daartoe zo kort door de bocht als hier mogelijk en voor het uitpakken voldoende is.

De kettingregel is een belangrijk voorbeeld en we laten zien hoe die regel op verschillende manieren wordt gebruikt, ook in de door fysici gebruikte manipulaties met afhankelijke en onafhankelijke grootheden bij het transformeren en oplossen van partiële differentiaalvergelijkingen. Integraalrekening in het vlak wordt nog wat kort behandeld, zowel in rechthoekige als in de uitvoerig besproken poolcoördinaten.

Nieuw is daarna de opzet van complexe functietheorie met lijnintegralen over alleen maar lijnstukjes en meteen de belangrijke hoofdstellingen, eerst zonder kromme poespas. Daarna bekijken we onderzoekend wat voor kromme krommen we na limietovergangen krijgen, en hoe we daarlangs kunnen integreren. De aanpak is zo precies tegenovergesteld aan de die van Conway, wiens fraaie opzet met equivalentieklassen van rectificeerbare krommen hier niet realiseerbaar is. Naast, voor of na de kromme aanpak, verkennen we de toepassingen van de hoofdstellingen bij het uitbreiden van de definitie van $f(z)$ met $z \in \mathbb{C}$ naar $f(A)$, eerst voor $A : \mathbb{R}^2 \to \mathbb{R}^2$ een lineaire afbeelding gegeven door een matrix, en daarna algemener, voor A van een (uiteindelijke complexe) Banachruimte X naar zichzelf. Een eerste kennismaking met Banachalgebra's[52] ligt hier voor de hand.

Gewone differentiaalvergelijkingen, al aan de orde geweest in de context van machtreeksen, motiveren de introductie van Banachruimten, met als belangrijkste voorbeeld $X = C([a, b])$, de ruimte van de continue $\mathbb{R}$-waardige functies op een interval $[a, b]$, waarin we de bijbehorende integraalvergelijkingen formuleren en oplossen.

De impliciete functiestelling kan dan weer worden overgeschreven. In Hoofdstuk 18 doen we dat al in één moeite door in combinatie de multiplicatorenmethode van Lagrange[53] voor stationaire punten van gewone functies van meer variabelen onder randvoorwaarden. Essentieel hier is het inzicht dat de oplossingsverzameling van een stelsel van bijvoorbeeld 3 vergelijkingen in $\mathbb{R}^{5=2+3}$, lokaal te schrijven is als de grafiek van een functie van $x \in \mathbb{R}^2$ naar $y \in \mathbb{R}^3$, tenzij er te veel nullen in de relevante berekeningen voorkomen.

De term onderdompeling wordt hier nog niet geintroduceerd[54]. De meer abstracte formulering van de methode van Lagrange in Hoofdstuk 16 is opgenomen for amusement. Ook wat pittiger is de behandeling van tweede orde afgeleiden die we pas in abstracte setting in meer detail doen. Het

[52] Door mijn medestudenten destijds ook wel Bananachalgebra's genoemd.

[53] De eerste stelling die ik ooit zelf aan anderen uitlegde, maar nu heel anders.

[54] Zie www.encyclo.nl/begrip/Submersie èn www.encyclo.nl/begrip/Immersie.

hoofdresultaat is het Lemma van Morse, Stelling 16.10, waarin met een coördinatentransformatie een functie waarvan de tweede afgeleide continu is, in de buurt van een stationair punt puur kwadratisch gemaakt wordt. Denk aan

$$F(x,y) = ax^2 + bxy + cy^2 + \cdots$$

en een transformatie die de puntjes wegwerkt als de discriminant niet gelijk is aan 0.

Zo'n transformatie is van de vorm

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} = A(x,y) \begin{pmatrix} x \\ y \end{pmatrix},$$

met $A(x,y)$ een van x en y afhankelijke matrix met

$$A(0,0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

die maakt dat

$$F(x,y) = a\xi^2 + b\xi\eta + c\eta^2.$$

We laten zien hoe $A = A(x,y)$ gevonden kan worden als oplossing van een kwadratische matrixvergelijking voor A die met worteltrekken kan worden opgelost.

45 Terug naar het platte vlak

Written for different audience, mathematics in the plane to prepare for Hilbert space. In dit hoofdstuk verzamelen we op informele wijze onze basiskennis over het platte vlak, met in ons achterhoofd de gedachte dat we later niet in twee maar in meer dimensies willen denken en werken: $3, 4, \ldots,$ tot en met aftelbaar oneindig. Bij het schrijven van dit hoofdstuk beginnen we in taal die hopelijk ook aansluit bij de schoolles, en nemen we soms ook dat perspectief als het gaat om wat we met inproducten van vectoren formuleren. Wie voor de klas staat of gaat staan heeft daar wellicht profijt van. De meeste opgaven zijn bedoeld als onderdeel van de uitleg. Convexe en gesloten deelverzamelingen, Cauchyrijen, en projecties zijn de belangrijkste begrippen die langskomen.

45.1 Punten en vectoren in het platte vlak

Exercise 45.1. Neem pen en blanco papier en teken een xy-vlak[1].

Zo, nu kunnen we aan de slag. Met en in een plat vlak waarin elk punt P gegeven is door 2 reële coördinaten, zeg $a \in \mathbb{R}$ en $b \in \mathbb{R}$. De assen labelen we met x en y. Het punt P is dus het punt met $x = a$ en $y = b$. We nummeren in deze notatie dus met het alfabet en zolang we in het vlak zitten is dat geen probleem. Ook in de 3-dimensionale ruimte kunnen we met 3 assen en $x = a, y = b, z = c$ prima uit de voeten maar vanaf dimensie 4 is het alfabet op als we beginnen bij x.

Op enig moment zullen we dus liever vanaf het begin met $x_1 = a_1$ en $x_2 = a_2$ willen werken. Een punt P gegeven door $x_1 = a_1$ en $x_2 = a_2$ kunnen we dan gewoon x noemen, soms dik gedrukt als $\mathbf{x}$, hetgeen met pen en papier weer vervelend is. Daarom ook vaak de notatie $\underline{x} = (x_1, x_2)$ voor een willekeurig, onbekend of variabel punt in het vlak, en vaak $\underline{a} = (a_1, a_2)$ voor een gegeven (vast) punt[2] in het vlak. De assen zijn dan de x_1-as en de x_2-as.

De punten $(1, 0)$ en $(0, 1)$ markeren we door er een 1 bij te zetten waarmee de schaalverdeling op de assen vast ligt. Beide punten zien we als liggend op afstand 1 tot de oorsprong $(0, 0)$, zonder fysische eenheid[3]. Het punt $(1, 1)$ heeft met Pythagoras dan afstand $\sqrt{2}$ tot $(0, 0)$.

[1] Suggestie: x-as horizontaal naar rechts, y-as verticaal omhoog.
[2] Dat we ook weer kunnen variëren natuurlijk.
[3] In de schoolpraktijk wordt vaak 1 cm als afstand tussen $(0, 0)$ en $(1, 0)$ aangehouden.

Van een punt kun je een vector maken. In de tekening door een lijntje te trekken van de oorsprong $O = (0, 0)$ naar een punt $\underline{a} = (a_1, a_2)$ met een pijlkopje in $\underline{a}$. Het pijltje associëren we met de vector

$$\vec{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix},$$

en de lengte van het pijltje is met Pythagoras weer gelijk aan $\sqrt{a_1^2 + a_2^2}$. Correspondentie met de tekening of niet, de (Euclidische) norm van $\underline{a}$ en $\vec{a}$ is bij afspraak gelijk aan en genoteerd als

$$|\underline{a}| = |\vec{a}| = \sqrt{a_1^2 + a_2^2},$$

en voldoet aan de driehoeksongelijkheid. Er geldt voor alle $\vec{a}, \vec{b} \in \mathbb{R}^2$ dat

$$|\vec{a} + \vec{b}| \leq |\vec{a}| + |\vec{b}|,$$

het derde axioma voor de eigenschappen waar normen aan moeten voldoen.

Exercise 45.2. De eerste twee norm-axioma's zijn $|\vec{a}| > 0$ als $\vec{a}$ niet de nulvector is en $|t\vec{a}| = |t||\vec{a}|$ voor $t \in \mathbb{R}$ en $\vec{a} \in \mathbb{R}^2$. Verifieer dat de Euclidische norm aan de norm-axioma's voldoet.

We *denken* aan $\vec{a}$ als een pijltje dat we op kunnen schuiven[4] zodat de staart in een ander punt komt te liggen. Bijvoorbeeld in het punt $\underline{b}$, zodat de kop van het pijltje in het punt

$$\underline{c} = \underline{a} + \underline{b} = (a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 + b_2)$$

komt te liggen, waarbij we dan de vector

$$\vec{c} = \vec{a} + \vec{b} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \end{pmatrix}$$

hebben. De vector $\vec{a}$ ligt dan met zijn staart in $\underline{b}$ en met zijn kop in $\underline{c}$. Dat kan natuurlijk ook andersom, met de staart van $\vec{b}$ in $\underline{a}$ en de kop van $\vec{b}$ in $\underline{c}$. De afstand tussen $\underline{c}$ en $\underline{b}$ is dus de lengte van het pijltje $\vec{a} = \vec{c} - \vec{b}$: de norm van de vector $\vec{a} = \vec{c} - \vec{b}$.

We switchen regelmatig heen en weer tussen rij- en kolomnotatie en tussen punten en vectoren, al naar gelang het zo uitkomt. Een in de tijd bewegend

[4] In het *platte* vlak geen probleem maar google op Gauss en kromming.

punt $\underline{x}$ heeft op elk moment een snelheid $\vec{v}$ die we ons vanwege de fysische interpretatie het liefst met de staart in $\underline{x}$ voorstellen. En als het handig is dan zien we $\underline{x}$ ook als $\vec{x}$. Bijvoorbeeld in

$$\vec{x} = \vec{s} + t\vec{v} = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} + t\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} + \begin{pmatrix} tv_1 \\ tv_2 \end{pmatrix} = \begin{pmatrix} s_1 + tv_1 \\ s_2 + tv_2 \end{pmatrix},$$

de formule[5] voor een punt dat beweegt over een rechte lijn l door het punt $\underline{s}$ met snelheidsvector $\vec{v}$.

Exercise 45.3. De lijn l door $\underline{s} \in \mathbb{R}^2$ met richtingsvector $\vec{v} \in \mathbb{R}^2$ kan ook gegeven worden door een vergelijking van de vorm

$$a_1x_1 + a_2x_2 = c$$

voor de punten $\underline{x} = (x_1, x_2)$ op de lijn l. Voor welke lijnen kan dat met $c = 1$? Bepaal voor die lijnen de bijbehorende a_1 en a_2.

Naast de vectoroptelling is in de vectorvoorstelling van een rechte lijn met steunvector $\vec{s}$ en richtingsvector $\vec{v}$ ook de scalaire vermenigvuldiging gebruikt. Voor iedere $t \in \mathbb{R}$ en $\vec{v} \in \mathbb{R}^2$ is $t\vec{v}$ gedefinieerd zoals je zou verwachten. De formule voor $\vec{c} = \vec{a} + \vec{b}$ gaat via $\vec{c} = \vec{x}$, $\vec{a} = \vec{s}$ en $\vec{b} = t\vec{v}$ over in de vectorvoorstelling van de lijn, waarin $\vec{x}$ de met t variërende vector is bij het punt $\underline{x}$.

In de formules mogen alle punten in het platte vlak voorkomen. En alle punten dat zijn alle punten van de vorm $\underline{x} = (x_1, x_2)$ met $x_1, x_2 \in \mathbb{R}$. Het platte vlak past daarmee weliswaar niet in ons universum maar gelukkig wel in ons hoofd, waar het de naam $\mathbb{R}^2$ gekregen heeft, met de 2 van 2-dimensionaal.

Ieder element uit de verzameling $\mathbb{R}^2$ wordt gegeven door een geordend reëel getallenpaar dat we aan kunnen geven met de letters die we willen, en met de notatie die we willen. Nummerend met het alfabet of met indices 1 en 2, achter elkaar of boven elkaar als

$$v_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = v_1\vec{e}_1 + v_2\vec{e}_2$$

geschreven, of eventueel ook als

$$v_1 + iv_2,$$

[5] Vectorvoorstelling van een lijn.

als maar duidelijk is dat v_1 de eerste, en v_2 de twee coördinaat is. De laatste twee vormen suggereren alvast de correspondentie

$$1 \quad \leftrightarrow \quad \vec{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$i \quad \leftrightarrow \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

en de representatie van de complexe getallen $\mathbb{C}$ als het (complexe) vlak $\mathbb{R}^2$ met een wat rare notatie[6].

45.2 Kortste afstanden

De kortste verbinding tussen twee punten in het vlak is de rechte lijn. In welk vlak? In het vlak dat we in ons hoofd hebben via de introductie van $\mathbb{R}^2$ in Sectie 45.1. Welke punten? Iedere $\underline{a}$ en $\underline{b}$ in die $\mathbb{R}^2$. Welke rechte lijn? Geen rechte lijn, maar het lijnstuk

$$\{t\underline{a} + (1-t)\underline{b} : \, 0 \leq t \leq 1\},$$

een stuk van de rechte lijn door steunvector $\underline{b}$ met richtingsvector $\vec{a} - \vec{b}$.

Er zijn geen andere paden van $\underline{b}$ naar $\underline{a}$ met een kortere afgelegde weg, een in het dagelijks leven op het Groningse platte land geboren uitspraak over *alle* paden van $\underline{b}$ naar $\underline{a}$, waarin twee begrippen voorkomen die wiskundig gezien hier nog niet eens gedefinieerd[7] zijn. Maar die kortste afgelegde weg moet natuurlijk wel gelijk zijn aan wat we de afstand tussen $\underline{a}$ en $\underline{b}$ noemen. Kortom, kortste afstanden gaan hier niet nog even niet over de weg van $\underline{a}$ naar $\underline{b}$. Er is maar een afstand tussen $\underline{a}$ en $\underline{b}$ en dat is

$$d(\underline{a}, \underline{b}) = |\underline{a} - \underline{b}| = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} = |\vec{a} - \vec{b}|,$$

de lengte van de vector $\vec{a} - \vec{b}$.

Over de kortste afstand tussen $\underline{a}$ en $\underline{b}$ hoeven we het dus in het platte vlak niet te hebben. Daar is een formule voor die we als vanzelfsprekend zien. En die formule definieert een afstandsbegrip dat voldoet aan axioma's: de axioma's van een metriek[8].

Maar wat is de kortste afstand tussen een niet-lege deelverzameling A van $\mathbb{R}^2$ en een punt $\underline{b}$? Met andere woorden, als de functie $f_b : \mathbb{R}^2 \to \mathbb{R}$ gedefinieerd wordt door

$$f_{\underline{b}}(\underline{x}) = d(\underline{x}, \underline{b}) = |\vec{x} - \vec{b}|,$$

[6] En extra algebra gebaseerd op de afspraak dat i keer i is $i^2 = -1$.

[7] Om welke twee begrippen gaat het?

[8] Wat is een metriek? Zoek op.

wat kun je dan zeggen over de waardenverzameling

$$W = \{f_{\underline{b}}(\underline{x}) : \underline{x} \in A\}\,?$$

Heeft deze deelverzameling van $\mathbb{R}$ een kleinste element?

Wel, de waardenverzameling W is niet leeg en naar beneden begrensd door 0. Op grond van de axioma's (of eigenschappen) van de reële getallen heeft W dus een grootste ondergrens[9] d die we vanaf nu de afstand van $\underline{b}$ to A noemen:

$$d = d(\underline{b}, A) = \inf W = \inf_{\underline{x} \in A} d(\underline{x}, \underline{b}).$$

Dus ook als de kleinste waarde niet bestaat, of als we dat niet a priori weten, is zo de afstand d tussen $\underline{b}$ en A wiskundig gedefinieerd. Of d nu wordt aangenomen door $d(\underline{x}, \underline{b})$ voor een x in W of niet.

De wiskundige definitie vertelt ons dat voor iedere[10] positieve gehele n er een $x_n \in A$ is met

$$d(\underline{b}, A) \leq d(\underline{b}, \underline{x}_n) < d(\underline{b}, A) + \frac{1}{n},$$

want iedere n waarvoor zo'n $\underline{x}$ niet bestaat zou een grotere ondergrens voor W zijn. Of je de wiskundige de afstand d ook echt kan vinden als horende bij een $\underline{a} \in A$ via $d = d(\underline{a}, \underline{b})$ is maar de vraag natuurlijk.

Een strategie om aan de kleinste waarde d te komen is om de rij $\underline{x}_n$ convergent te kiezen. Als dat kan dan heeft de rij een limiet $\underline{a}$. Als vervolgens blijkt dat $\underline{a}$ in A ligt volgt hopelijk ook dat $d(\underline{b}, A) = d(\underline{b}, \underline{a})$. En blijft vervolgens nog de vraag of het punt in A waarin de kleinste afstand aangenomen wordt uniek is. Het gaat dus om twee zaken. Het vinden van convergerende minimaliserende rijen in A en daarna de vraag om daar altijd dezelfde limiet bij hoort.

Maar soms kun je d meteen uitrekenen. Hoewel?

Exercise 45.4. Wat is de kortste afstand tussen $\underline{a} = (1, 1)$ en de lijn met vergelijking $3x_1 + x_2 = 1$?

Exercise 45.5. De kortste afstand tussen $\underline{a} = (1, 1)$ en de deelverzameling $E \subset \mathbb{R}^2$ gegeven door $9x_1^2 + x_2^2 \leq 1$ is niet zo eenvoudig uit te rekenen. Probeer het maar. Maar is het punt in E met minimale afstand tot $\underline{a}$ uniek denk je? Waarom? Maak een plaatje.

[9] Ander woord: infimum.

[10] We mijden hier de $\varepsilon > 0$, for all practical purposes is $\frac{1}{n}$ net zo goed.

Exercise 45.6. Reflecteer[11] op wat het begrip loodrecht met het begrip afstand te maken heeft.

Exercise 45.7. Teken voor verschillende (reële) waarden van a en b in je xy-vlak de vectoren[12]

$$\begin{pmatrix} a \\ b \end{pmatrix} \quad \text{en} \quad \begin{pmatrix} -b \\ a \end{pmatrix}$$

en reflecteer op het begrip loodrecht. Kun je andere paren vectoren in het vlak bedenken waarop het begrip loodrecht van toepassing is?

Exercise 45.8. Een deelverzameling $K \subset \mathbb{R}^2$ heet convex als met elk tweetal punten $\underline{a}$ en $\underline{b}$ in K ook het lijnstuk

$$\{t\underline{a} + (1-t)\underline{b} : 0 \leq t \leq 1\}$$

dat $\underline{a}$ en $\underline{b}$ verbindt in K ligt. Kunnen er twee punten in K zijn die $f_O(\underline{x}) = |\underline{x}|$ minimaliseren op K? Maak een plaatje dat je helpt om de vraag te beantwoorden.

45.3 Vlakke meetkunde met het inprodukt

Bij het maken van deze opgaven heb je ongetwijfeld rechte hoeken en driehoeken getekend en de (Stelling van) Pythagoras weer gebruikt, en wellicht al het inwendige produkt van vectoren gebruikt. Het *standaard inwendige produkt* in $\mathbb{R}^2$ wordt gedefinieerd door

$$\begin{pmatrix} a \\ b \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = ax + cy,$$

hetgeen voor elke keuze van de 2-vectoren

$$\begin{pmatrix} a \\ b \end{pmatrix} \quad \text{en} \quad \begin{pmatrix} x \\ y \end{pmatrix}$$

een reëel getal definieert, dat vastgelegd wordt door de vier reële getallen a, b, x, y. De opgaven hebben je overtuigd dat twee vectoren in $\mathbb{R}^2$ loodrecht op elkaar staan precies dan als hun inwendig produkt nul is.

[11] Minimum op de rand, denk ook aan multiplicatoren van Lagrange.

[12] Al of niet met de staart in de oorsprong O.

Loodrecht is hier een begrip dat je buiten de wiskunde kende en nu in de wiskunde van betekenis hebt voorzien, en wel in het abstracte platte vlak in je hoofd, en de meetkunde die je daarin hebt leren bedrijven, al of niet gebruikmakend van twee onderling loodrecht voorgestelde coördinaatassen, gemarkeerd met 0 en 1.

De afstand van $(0,0)$ tot $\underline{a} = (a_1, a_2)$ is met Pythagoras gelijk aan $\sqrt{\vec{a} \cdot \vec{a}}$, de wortel uit het inwendige produkt van de bijbehorende vector $\vec{a}$ met zichzelf. Zo hebben we de begrippen afstand en loodrecht die we uit de dagelijkse werkelijkheid kennen in verband gebracht met het standaard inwendig produkt in $\mathbb{R}^2$, ons model voor het platte vlak. Dit verband zit stevig tussen onze oren, wat het verder ook moge betekenen. Wiskundige uitspraken doen we vanaf nu in termen van $\mathbb{R}^2$ met zijn vectoroptelling en het standaard inwendige produkt.

Exercise 45.9. Bewijs dat $|\vec{a} \cdot \vec{b}| \leq |\vec{a}||\vec{b}|$, met andere woorden, dat

$$(a_1 b_1 + a_2 b_2)^2 \leq (a_1^2 + a_2^2)(b_1^2 + b_2^2).$$

Hint: breng alles naar de rechterkant, doe de algebra en herken het kwadraat. Doe vervolgens ook

$$(a_1 b_1 + a_2 b_2 + a_3 b_3)^2 \leq (a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2),$$

en overtuig jezelf ervan dat (even wat combinatoriek)

$$(\sum_{k=1}^{n} a_k^2)(\sum_{k=1}^{n} b_k^2) - (\sum_{k=1}^{n} a_k b_k)^2$$

de som is van $\frac{n(n-1)}{2}$ kwadraten.

Exercise 45.10. Teken twee vectoren $\vec{a}$ en $\vec{b}$ waarvoor $\vec{a} \cdot \vec{b} = 0$ en schuif een van de twee vectoren op en wel zó dat de kop van deze ene vector in de staart van de andere vector ligt (en een rechthoekige driehoek ontstaat). Werk $(\vec{a} + \vec{b}) \cdot (\vec{a} + \vec{b})$ uit tot de bekende formule voor $|\vec{a}|$, $|\vec{b}|$ en $|\vec{a} + \vec{b}|$.

Exercise 45.11. Leid met Opgave 45.9 en Opgave 45.10 nog een keer af dat de norm aan de driehoeksongelijkheid $|\vec{a} + \vec{b}| \leq |\vec{a}| + |\vec{b}|$ voldoet, ook voor $\vec{a} \cdot \vec{b} \neq 0$.

Exercise 45.12. Teken twee vectoren $\vec{a}$ en $\vec{b}$ waarvoor niet per se $\vec{a}\cdot\vec{b}=0$ en schuif een van de twee op zó dat de kop van deze ene in de staart van de ander vector ligt (en een driehoek ontstaat). Werk $(\vec{a}+\vec{b})\cdot(\vec{a}+\vec{b})$ en doe hetzelfde voor $\vec{a}$ en $-\vec{b}$. Beide uitdrukkingen bevatten $\vec{a}\cdot\vec{b}$ maar na sommatie vallen deze kruistermen weg. Formuleer wat bekend staat als de parallellogramwet.

Exercise 45.13. Een elegant bewijs van de Stelling van Pythagoras zonder vectoren maar met bijvoorbeeld vierkanten heeft iedereen wel eens gezien natuurlijk. Zie bijvoorbeeld

`http://www.few.vu.nl/~jhulshof/RYB.mov`

Is er ook zo'n elegant bewijs[13] van de parallellogramwet?

45.4 Projecteren op convexe verzamelingen

Vlakke en Euclidische meetkunde betreffen tamelijk expliciete zaken. Denk aan lijnen, vlakken etc. Teken een lijn in het vlak en doe wat. Het plaatje is altijd hetzelfde. Projecteren op een lijn, iedereen kan het. Bij projecteren op convexe verzamelingen gaat over een veel grotere klasse van verzamelingen maar met de algebra van het inprodukt is goed te begrijpen hoe dat gaat. Die algebra is niet beperkt tot het platte vlak. Maar nu eerst even wel.

Exercise 45.14. Als $\underline{b}\in\mathbb{R}^2$ en $K\subset\mathbb{R}^2$ niet leeg en convex is, dan heeft iedere minimaliserende rij $\underline{x}_n\in K$ met $d(\underline{x}_n,\underline{b})\to d$ de eigenschap dat

$$d(\underline{x}_n,\underline{x}_m)\to 0 \quad \text{as} \quad m,n\to\infty$$

en dat kun je algemeen bewijzen. Neem zonder beperking der algemeenheid $\underline{b}=O$ en $d(\underline{x}_n,O)$ dalend, en laat dit zien door voor $m>n$ met de parallellogramwet $|\underline{x}_n-\underline{x}_m|^2$ af te schatten op $\varepsilon_n=4(d+\frac{1}{n})^2-d^2)$. Hint: je hebt alleen nodig dat het midden van elk lijnstuk tussen twee punten in K weer in K zit ($t=\frac{1}{2}$ in de definitie).

Onze meetkundige kennis is in de opgaven hierboven in uitspraken over vectoren en inwendige produkten vertaald, met als opmerkelijk conclusie het resultaat in Opgave 45.14 dat zegt dat de minimaliserende rij een Cauchyrij[14]

[13] Vast wel, maar ik heb het zelf nog nooit gezien.

[14] Wat was dat ook al weer?

is. Net als in $\mathbb{R}$ zijn in $\mathbb{R}^2$ Cauchyrijen convergent. De limiet $\underline{a}$, waarvoor geldt dat

$$d(\underline{x}_n, \underline{a}) \to 0 \quad \text{as} \quad n \to \infty,$$

hoeft natuurlijk niet per se in A te liggen, maar doet dat wel als A gesloten is.

Exercise 45.15. $A \subset \mathbb{R}^2$ heet gesloten als iedere convergente rij x_n in A ook zijn limiet in A heeft. Als A niet gesloten is dan zijn er dus convergente rijen in A waarvan de limiet niet in A ligt. Bewijs dat de afsluiting $\overline{A}$, dat is A verenigd met al die limieten, altijd gesloten is.

Exercise 45.16. Voor iedere niet-lege convexe $K \subset \mathbb{R}^2$ en voor iedere $b \in \mathbb{R}^2$ bestaat er een $a \in \overline{K}$ met $d(\underline{b}, \underline{a}) = d(\underline{b}, K)$. Bewijs dit met de voorafgaande resultaten en laat zien dat $\underline{a}$ uniek is. Concludeer dat $\underline{b} \to \underline{a}$ een afbeelding $P_K : \mathbb{R}^2 \to \overline{K}$ definieert. Laat ook zien $(P_K(\vec{a}) - \vec{a}) \cdot (\vec{x} - P_K(\vec{a})) \geq 0$ voor alle $\underline{x} \in K$ en maak een plaatje om de betekenis van deze uitspraak meetkundig te begrijpen.

Exercise 45.17. Laat zien dat de afbeelding P_K een contractie is in de zin dat voor alle $\underline{x}, \underline{y} \in \mathbb{R}^2$ geldt dat $d(P_K(\underline{x}), P_K(\underline{y})) \leq d(\underline{x}, \underline{y})$. Hint: deze is lastig, spelen met het inprodukt, te leuk om voor te zeggen. Let op, voor variabele punten in K heb je nu een andere letter nodig.

Exercise 45.18. Pas de vorige opgave toe op het geval $K = l$, met l de lijn door $\underline{s}$ met richtingsvector $\vec{v}$ en geef een formule voor P_l. Hint: waarom wordt de ongelijkheid in Opgave 45.16 nu een gelijkheid voor alle $\underline{x} \in l$? Gebruik dit en reken $P_l(\underline{b})$ gewoon uit voor gegeven $\underline{b}$.

Exercise 45.19. Neem in de vorige opgave $\underline{s} = O$ en laat zien dat de nulverzameling

$$N(P_l) = \{\underline{x} \in \mathbb{R}^2 : \ P_l(\underline{x}) = \underline{0}\}$$

van P_l weer een lijn is, zeg lijn m, en dat m en l loodrecht op elkaar staan in dat vlak in je hoofd.

45.5 Andere inprodukten en bilineaire vormen

Het standaard inwendig produkt van

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{and} \quad \vec{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

is een voorbeeld van een bilineaire functie, ook wel bilineaire vorm genoemd. Zulke *bilineaire vormen* $B : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ zijn altijd te schrijven als

$$B(\vec{x}, \vec{y}) = a_{11}x_1y_1 + a_{12}x_1y_2 + a_{21}x_2y_1 + a_{22}x_2y_2,$$

dit vanwege wat je in de volgende opgave nu uitwerkt.

Exercise 45.20. Laat zien dat als $B : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ voldoet aan

$$B(\vec{x}_1 + \vec{x}_2, \vec{y}) = B(\vec{x}_1, \vec{y}) + B(\vec{x}_2, \vec{y});$$

$$B(\vec{x}, \vec{y}_1 + \vec{y}_2) = B(\vec{x}, \vec{y}_1) + B(\vec{x}, \vec{y}_2);$$

$$B(t\vec{x}, \vec{y}) = B(\vec{x}, t\vec{y}) = tB(\vec{x}, \vec{y}),$$

voor alle $t \in \mathbb{R}$ en $\vec{x}, \vec{x}_1, \vec{x}_2, \vec{y}, \vec{y}_1, \vec{y}_2 \in \mathbb{R}^2$, dat B gegeven wordt door[15]

$$B(\vec{x}, \vec{y}) = \sum_{i,j=1}^{2} a_{ij}x_iy_j,$$

en dat $B(\vec{x}, \vec{y}) = B(\vec{y}, \vec{x})$ voor alle $\vec{x}, \vec{y} \in \mathbb{R}^2$ gelijkwaardig is met $a_{ij} = a_{ji}$ voor alle $i, j \in \{1, 2\}$.

Kortom, $B(\vec{x}, \vec{y})$ is van de vorm

$$B(\vec{x}, \vec{y}) = A\vec{x} \cdot \vec{y},$$

waarbij $A : \mathbb{R}^2 \to \mathbb{R}^2$ de lineaire afbeelding is gegeven is door

$$A\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

en de symmetrie van B equivalent is met de symmetrie van de lineaire afbeelding A en de bijbehorende matrix (a_{ij}):

$$B(\vec{x}, \vec{y}) = B(\vec{y}, \vec{x}) \quad \Leftrightarrow \quad A\vec{x} \cdot \vec{y} = \vec{x} \cdot A\vec{y} \quad \Leftrightarrow \quad a_{ij} = a_{ji}$$

[15] Let op: x_i en y_j zijn nu componenten van $\vec{x}$ en $\vec{y}$.

Een symmetrische bilineaire vorm definieert een inwendig produkt als de bijbehorende kwadratische vorm positief definiet is, dat wil zeggen

$$A\vec{x}\cdot\vec{x}>0 \quad \text{as} \quad \vec{x}\neq\vec{0}=\begin{pmatrix}0\\0\end{pmatrix},$$

en in dat geval heet A zelf ook positief[16] definiet[17]. Voorlopig zullen we in de notatie geen onderscheid maken tussen A als lineaire afbeelding en A als matrix. We schrijven dus ook

$$A=\begin{pmatrix} a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix}$$

en spreken over ook positief definiete (symmetrische) matrices.

Kwadratische vormen zijn homogene poynomen van graad twee in de variabelen. Een kwadratische vorm $Q:\mathbb{R}^2\to\mathbb{R}$ wordt dus gegeven door

$$Q(\vec{x})=Q(\underline{x})=Q(x_1,x_2)=q_{11}x_1^2+q_{12}x_1x_2+q_{22}x_2^2=\sum_{1\leq i\leq j=2}^{2}q_{ij}x_ix_j$$

en is altijd te schrijven als $Q(x_1,x_2)=B(\vec{x},\vec{x})=A\vec{x}\cdot\vec{x}$, met

$$a_{ii}=q_{ii} \quad \text{and} \quad a_{ij}=a_{ji}=\frac{1}{2}q_{ij} \quad (i<j).$$

Omdat

$$m=\min_{|\underline{x}|\leq 1}Q(\underline{x})=\min_{|\underline{x}|=1}Q(\underline{x}) \quad \text{and} \quad M=\max_{|\underline{x}|\leq 1}Q(\underline{x})=\max_{|\underline{x}|=1}Q(\underline{x})$$

bestaan als (op de rand aangenomen[18]) minimum en maximum van Q op de gesloten disk gegeven door

$$x_1^2+x_2^2\leq 1,$$

definieert een symmetrische A dus een (niet-standaard) inwendig produkt als $m>0$.

Exercise 45.21. Neem aan dat $0\leq m\leq M$. Laat zien dat

$$m\,\vec{x}\cdot\vec{x}\leq A\vec{x}\cdot\vec{x}\leq M\,\vec{x}\cdot\vec{x}$$

voor alle $\vec{x}\in\mathbb{R}^2$. Wat kun je zeggen zonder de aanname op de tekens van m en M?

[16]Echt iets anders dan $a_{ij}>0$ voor $i,j=1,2$.

[17]Impliciet is A dus symmetrisch verondersteld.

[18]Mini- en maximaliserende rijen $\underline{x}_1,\underline{x}_2,\ldots$ kunnen convergent gekozen worden.

De rand van de disk is een cirkel die kunnen we parametriseren met

$$x_1 = \cos(t) \quad \text{and} \quad x_2 = \sin(t),$$

waarin de functies cos en sin uniek gedefinieerd zijn door bijvoorbeeld[19]

$$\cos t = \cos(t) = \sum_{n=0}^{\infty} \frac{(-t)^{2n}}{(2n)!} = 1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \frac{t^6}{6!} + \cdots$$

$$\sin t = \sin(t) = \sum_{n=0}^{\infty} \frac{(-t)^{2n+1}}{(2n+1)!} = t - \frac{t^3}{3!} + \frac{t^5}{5!} - \frac{t^7}{7!} + \cdots,$$

met[20] $\sin' = \cos$, $\cos' = -\sin$, $\cos(0) = 1$, $\sin(0) = 0$.

Exercise 45.22. Bereken het maximum M en het minimum m van de functie $q : \mathbb{R} \to \mathbb{R}$ gedefinieerd door

$$q(t) = Q(\cos t, \sin t) = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix} \cdot \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}$$

Hint, herschrijf als $q(t) = a\cos^2 t + b\cos t \sin t + c\sin^2 t$, neem eerst $b \neq 0$ en herleid $q'(t) = 0$ tot een vierkantsvergelijking voor $\tan t$. Verifieer dat in de *minimizers*

$$A\begin{pmatrix} \cos t \\ \sin t \end{pmatrix} = m\begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$$

geldt, en in de *maximizers*

$$A\begin{pmatrix} \cos t \\ \sin t \end{pmatrix} = M\begin{pmatrix} \cos t \\ \sin t \end{pmatrix}.$$

Deze opgave laat zien dat m en M de twee reële eigenwaarden zijn van de symmetische matrix A. In het geval dat A positief definiet is nummeren we deze eigenwaarden $\lambda_1 = M \geq \lambda_2 = m > 0$. Je ziet[21] dat de bijbehorende eigenvectoren loodrecht staan. In het geval dat $M = m$ zijn alle vectoren eigenvectoren en kunnen ze loodrecht gekozen worden, $\vec{e}_1$ en $\vec{e}_2$ bijvoorbeeld.

[19] Zie [HM, hoofdstuk 10].

[20] De twee differentiaalvergelijkingen en beginvoorwaarden definiëren sin en cos.

[21] Misschien niet meteen.

Exercise 45.23. Bewijs direct, dus zonder cosinussen en sinussen, dat voor elke symmetrische (hier nog twee bij twee) matrix A geldt dat het maximum μ van de *absolute waarde* van de bijbehorende kwadratische vorm Q op $|\vec{x}| = 1$ wordt aangenomen in een eigenvector, en dat iedere *maximizer* een eigenvector is, bij μ of bij $-\mu$ (of bij allebei in bijzonder gevallen).

Exercise 45.24. De eigenvector in Opgave 45.23 bij $\lambda_1 = \pm\mu$ noemen we $\vec{v}_1$. De lijn door O met richtingsvector $\vec{v}_1$ noemen we l_1. Pas nu Opgave 45.19 toe[22] op $l = l_1$ en noem $m = l_2$. Laat zien dat A deze l_2 op zichzelf afbeeldt.

45.6 Om te onthouden

Symmetrische twee bij twee matrices komen met paren onderling loodrechte lijnen die we, zo we willen, als nieuwe coördinaatassen kunnen gebruiken. Met in die lijnen (eigen)vectoren $\vec{v}_1$ en $\vec{v}_2$ die onderling loodrecht staan en lengte 1 hebben,

$$\vec{v}_1 \cdot \vec{v}_1 = \vec{v}_2 \cdot \vec{v}_2 = 1 \quad \text{and} \quad \vec{v}_1 \cdot \vec{v}_2 = 0,$$

bij eigenwaarden λ_1 en λ_2,

$$A\vec{v}_1 = \lambda_1\vec{v}_1 \quad \text{and} \quad A\vec{v}_2 = \lambda_2\vec{v}_2.$$

In het bijzondere geval dat A een diagonaalmatrix

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

is, krijgen we als eigenvectoren de standaardbasisvectoren $\vec{e}_1$ en $\vec{e}_2$.

Het andere belangrijke resultaat is dat we op (gesloten) convexe verzamelingen kunnen projecteren, Opgave 45.16. Niet benadrukt nog is wat de essentie was van het bewijs dat je in Opgave 45.23 hebt gegeven. Waar het resultaat in Opgave 45.16 via Opgave 45.14 en een convergente minimaliserende rij tot stand kwam, is in Opgave 45.23 een maximaliserende rij *niet* automatisch convergent en moet eerst een convergente deelrij genomen worden. En iedere begrensde rij in $\mathbb{R}^2$ heeft zo'n convergente deelrij. Alles wat we hier behandeld hebben gaat dus door voor $\mathbb{R}^3$, $\mathbb{R}^4$,..., met een kleine aanpassing bij Opgave 45.19. Pas in $\mathbb{R}^\infty$ gaat het een beetje anders.

[22] De notaties $\underline{x}$ en $\vec{x}$ liepen al door elkaar heen, liever $x = \underline{x} = \vec{x}$ vanaf nu?

45.7 Poolcoördinaten in het (complexe) vlak

We besluiten dit hoofdstuk met een korte herhaling van $\mathbb{R}^2$ gezien als de verzameling van complexe getallen $\mathbb{C}$. Het punt $(1,0)$ zien we als het getal 1 en het punt $(0,1)$ als het imaginaire getal i. We introduceren $\mathbb{C}$ door de correspondentie

$$(x,y) \in \mathbb{R}^2 \quad \leftrightarrow \quad z = x + yi = x + iy \in \mathbb{C}$$

met in $\mathbb{C}$ de gebruikelijke rekenoperaties: de complexe optelling en de complexe vermenigvuldiging. Die krijg je door te rekenen met uitdrukkingen als $z = x + iy$ en $c = a + bi$ alsof het eerstegraads polynomen in i zijn, met de afspraak dat $i^2 = -1$. De rollen van i en $-i$ zijn daarbij uitwisselbaar want ook $(-i)^2 = 1$. De coëfficiënten x, y, a, b zijn zelf reëel, en x en a heten de reële delen van respectievelijk z en c. De *imaginaire* delen zijn y en b en zijn net zo reëel als de reële delen.

We gaan ervan uit dat de lezer vertrouwd[23] is met deze complexe getallen en het waarom van de notatie en correspondentie

$$(\cos(t), \sin(t)) \quad \leftrightarrow \quad \exp(it) = \cos(t) + i\sin(t)$$

voor het over de eenheidscirkel bewegende punt $(\cos(t), \sin(t))$.

Die eenheidscirkel wordt gegeven door $|z| = 1$, waarbij de absolute waarde van $z = x + iy$ per definitie gelijk is aan

$$|z| = \sqrt{x^2 + y^2},$$

meestal r genoemd. Voor elke $r > 0$ doorloopt het punt

$$(r\cos(t), r\sin(t)) \quad \leftrightarrow \quad r\exp(it) = r(\cos(t) + i\sin(t))$$

een cirkel met straal r in het al of niet complexe vlak, en de (tijd) t is *per definitie* de hoek in radialen die de met dit punt corresponderende vector maakt met de positieve x-as. Ieder punt in het vlak wordt zo gegeven door een r en een t, en elke 2-vector is van de vorm

$$\vec{x} = \begin{pmatrix} r\cos(t) \\ r\sin(t) \end{pmatrix} = r\begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}, \quad \text{het scalaire product van } r \text{ en } \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}.$$

Behalve de oorsprong heeft ieder punt $\underline{x}$ en iedere vector $\vec{x}$ een unieke r en een unieke t, waarbij je moet afspreken dat de t-waarden module 2π worden gerekend. En 2π per definitie het reële getal is waarvoor deze laatste karakterisatie correct is. In (tijd) $t = 2\pi$ ga je de cirkel rond.

[23] Zie anderes eventueel [HM, hoofdstuk 11].

Met behulp van deze *poolcoördinaten* volgt voor

$$\vec{c} = p\begin{pmatrix}\cos(s)\\ \sin(s)\end{pmatrix} \quad \text{en} \quad \vec{x} = r\begin{pmatrix}\cos(t)\\ \sin(t)\end{pmatrix}$$

dat

$$\vec{c}\cdot\vec{x} = pr(\cos(s)\cos(t) + \sin(s)\sin(t)) = pr\cos(s-t),$$

het product van de twee lengten en de cosinus van wat de *ingesloten hoek* wordt genoemd. Is die hoek gelijk aan $\pm\frac{\pi}{2}$ dan is het inproduct nul en staan de vectoren loodrecht op elkaar.

Exercise 45.25. De complexe afbeelding $z \to \frac{1}{z}$ laat zich in rechthoekige coördinaten x, y en in poolcoördinaten r en t bestuderen. Verifieer dat deze afbeelding de samenstelling is van $z \to \overline{z}$, een spiegeling in de x-as, en een andere afbeelding die spiegeling in de eenheidscirkel wordt genoemd, gegeven door $r \to \frac{1}{r}$. Construeer gegeven een punt binnen de cirkel zijn spiegelbeeld in de cirkel met behulp van een bij het gegeven punt geschikt gekozen raaklijn aan de cirkel.

Exercise 45.26. Merk op dat de uitkomst voor het *inwendig* product te vergelijken is met het gewone *complexe* product van de met de vectoren $\vec{c}$ en $\vec{x}$ corresponderende c en z. Verifieer dat voor

$$c = p\exp(is) \quad \text{en} \quad z = r\exp(it)$$

geldt dat

$$cz = p(\cos(s) + i\sin(s))r(\cos(t) + i\sin(t)) = pr(\cos(s+t) + i\sin(s+t)),$$

en bepaal het reële deel van $c\overline{z}$, waarin $\overline{z} = x - iy$ de complex geconjugeerde is van $z = x + iy$.

46 Into Hilbert space

For a different audience, from Euclidean space to Hilbert space and applications. In $\mathbb{R}^3$, $\mathbb{R}^4$,.... we can do the same algebra as in Chapter 45 for $\mathbb{R}^2$. In $\mathbb{R}^3$ we have

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = ax + by + cz \quad \text{or} \quad \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \sum_{i=1}^{3} a_i x_i,$$

in $\mathbb{R}^{42}$

$$\vec{a} \cdot \vec{x} = \sum_{i=1}^{42} a_i x_i \quad \text{for} \quad \vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{42} \end{pmatrix} \quad \text{and} \quad \vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{42} \end{pmatrix},$$

in $\mathbb{R}^\infty$, dropping the arrows,

$$a \cdot x = \sum_{i=1}^{\infty} a_i x_i \quad \text{for} \quad a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{pmatrix}$$

Points or vectors, we maken het onderscheid in de notatie tussen x als $\vec{x}$ en $\underline{x}$ steeds vaker alleen als het echt nodig is[1].

De laatste uitdrukking definieert $a \cdot x$ soms wel en soms niet, want zonder restricties op $a, x \in \mathbb{R}^\infty$ kan het met

$$a \cdot x = \sum_{i=1}^{\infty} a_i x_i = \sum_{i \in \mathbb{N}} a_i x_i$$

alle kanten op. En het wordt nog spannender als we de $i \in \mathbb{N}$ in vervangen door bijvoorbeeld $t \in \mathbb{R} = (-\infty, \infty)$ of[2] $t \in [-\pi, \pi]$. In zulke gevallen is ook $\sum$ aan vervanging toe. Overaftelbare[3] sommen gaan niet werken en sommeren moet hier dus wel integreren worden, wat anders? Met de notatie $t \to a_t = a(t)$ en $t \to x_t = x(t)$ voor functies $a : \mathbb{R} \to \mathbb{R}$ en $x : \mathbb{R} \to \mathbb{R}$ wordt een voor de hand liggend inwendig produkt van de functies a en x nu gedefinieerd met behulp van de formule

$$a \cdot x = \int_{-\infty}^{\infty} a(t)x(t)dt,$$

[1] Als we niet meer recht kunnen praten wat krom is en rechte pijltjes niet passen.
[2] Denk ook aan de poolcoördinaten in het platte vlak.
[3] Waarom niet?

waarin *alle* $a(t)$ en $x(t)$ waarden gelijkwaardig voorkomen maar, paradoxaal wellicht, individueel geen invloed hebben op de uitkomst van de integraal die $a \cdot x$ definieert. Ook met die uitkomst kan het, bijvoorbeeld voor continue functies, alle kanten op, net als met $a \cdot x$ voor $a, x \in \mathbb{R}^\infty$.

Voor 2π-periodieke continue functies heeft deze integraalformule geen betekenis maar de formule

$$a \cdot x = \int_{-\pi}^{\pi} a(t)x(t)dt$$

vaak wel, het standaard inwendig produkt waarmee we werken in het geval van 2π-periodieke functies a en x, (goed) gedefinieerd voor continue functies als gewone Riemann integraal[4].

Exercise 46.1. Voor $n = 1, 2, 3, \ldots$ zijn de 2π-periodieke functies c_n en s_n gedefinieerd door $c_n(t) = \cos(nt)$ en $s_n(t) = \sin(nt)$. Bereken nog eens $c_n \cdot c_m$, $c_n \cdot s_m$, $s_n \cdot s_m$, voor $m, n = 1, 2, 3, \ldots$.

Je ziet het niet meteen, maar al deze cosinussen en sinussen staan "loodrecht" op elkaar, en ze hebben ook allemaal dezelfde "lengte", de wortel uit het inprodukt van de functie met zichzelf.

Exercise 46.2. Er is nog een functie die loodrecht staat op al deze cosinussen en sinussen. Welke functie?

46.1 Standaardassenkruizen

Tja[5], wat zijn dat? In het vlak waar we mee begonnen zijn wordt het assenkruis gevormd door 2 lijnen: de x-as door de oorsprong O en het punt $(1, 0)$ en de y-as door O en het punt $(0, 1)$, of wellicht liever de x_1-as en de x_2-as. Een punt dat zich over zo'n as beweegt heeft een lange weg te gaan en kwam van ver. De x-as wordt geparametriseerd door $(x, y) = (t, 0)$, en de y-as door $(x, y) = (0, t)$, met bijbehorende snelheidsvectoren[6]

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{en} \quad \vec{v}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

[4] En later via een subtiel proces voor nog veel meer functies.

[5] Een vraag voor de woensdagmiddag wellicht.

[6] In de wiskundeles meestal richtingsvectoren genoemd.

die samen de standaardbasis van $\mathbb{R}^2$ als vectorruimte vormen.

Evenzo bestaat in $\mathbb{R}^3$ het standaardassenkruis uit 3 lijnen, de x- of x_1-as, de y- of x_2-as, en de z- of x_3-as, met bijbehorende vectoren[7]

$$\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{v}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \vec{v}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

die samen de standaardbasis van $\mathbb{R}^3$ genoemd worden. Drie vectoren met lengte 1 die onderling loodrecht staan.

En, we zouden het bijna vergeten, standaard of niet, een basis vormen ze. Iedere vector $\vec{v} \in \mathbb{R}^3$ is vanzelfsprekend uniek te schrijven als

$$\vec{v} = v_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + v_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}.$$

Precies zoals in $\mathbb{R}^2$ waar iedere $\vec{v}$ van de vorm

$$\vec{v} = v_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

is, met een unieke correspondentie

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \quad \leftrightarrow \quad (v_1, v_2) = \underline{v},$$

waarin links en rechts v_1 en v_2 *hetzelfde* zijn.

De vectoren

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{en} \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

hebben lengte 1 en staan onderling loodrecht. In termen van het *standaard inwendig produkt*:

$$\vec{e}_1 \cdot \vec{e}_1 = \vec{e}_2 \cdot \vec{e}_2 = 1 \quad \text{en} \quad \vec{e}_1 \cdot \vec{e}_2 = \vec{e}_2 \cdot \vec{e}_1 = 0.$$

Met de gebruikelijke rekenregels volgt nu dat

$$\vec{v} \cdot \vec{e}_1 = (v_1\vec{e}_1 + v_2\vec{e}_2) \cdot \vec{e}_1 = v_1\vec{e}_1 \cdot \vec{e}_1 + v_2\vec{e}_2 \cdot \vec{e}_1 = v_1;$$

$$\vec{v} \cdot \vec{e}_2 = (v_1\vec{e}_1 + v_2\vec{e}_2) \cdot \vec{e}_2 = v_1\vec{e}_1 \cdot \vec{e}_2 + v_2\vec{e}_2 \cdot \vec{e}_2 = v_2,$$

[7] Snelheidsvectoren, richtingsvectoren, het zijn maar woorden.

en

$$\vec{v} = (\vec{v} \cdot \vec{e}_1)\vec{e}_1 + (\vec{v} \cdot \vec{e}_2)\vec{e}_2 = \sum_{i=1}^{2} (\vec{v} \cdot \vec{e}_i)\vec{e}_i$$

voor vectoren $\vec{v} \in \mathbb{R}^2$. In iedere $\mathbb{R}^n$ met n positief en geheel gaat het hetzelfde,

$$\vec{v} = \sum_{i=1}^{n} (\vec{v} \cdot \vec{e}_i)\vec{e}_i,$$

en pas in $\mathbb{R}^\infty$ wordt het wat lastiger.

46.2 Symmetrische matrices

Net als in de twee-dimensionale context heeft iedere symmetrische $n \times n$ matrix

$$A = (a_{ij})_{i,j=1,\ldots,n}$$

(een basis van) eigenvectoren $\vec{v}_1, \ldots, \vec{v}_n \in \mathbb{R}^n$ met

$$\vec{v}_i \cdot \vec{v}_j = \delta_{ij} = \begin{cases} 1 & \text{als } i = j \\ 0 & \text{als } i \neq j, \end{cases}$$

bij (reële) eigenwaarden

$$\lambda_1, \ldots, \lambda_n,$$

die gemaakt worden door Opgave 45.23 herhaald toe te passen. Dit geeft in ieder stap zowel een nieuwe λ_k als een nieuwe $\vec{v}_k$ via

$$|\lambda_k| = \max_{\substack{\vec{v} \cdot \vec{v} \leq 1 \\ \vec{v} \cdot \vec{v}_1 = \cdots = \vec{v} \cdot \vec{v}_{k-1} = 0}} |A\vec{v} \cdot \vec{v}|,$$

waarbij $k = 1, \ldots, n$.

Details van deze constructie komen aan de orde in de context van de standaard aftelbaar oneindig-dimensionale Hilbertruimte die we in $\mathbb{R}^\infty$ maken. *Daarvóór* komen we *voor het eerst* over abstracte Hilbertruimten[8] H te praten die we zo snel mogelijk gelijk[9] praten aan het standaardvoorbeeld in $\mathbb{R}^\infty$, onder de aanname van separabiliteit van H: het bestaan van een rij $x_1, x_2, x_3, \ldots$ in H die als limieten van zijn convergente deelrijen *alle* elementen in H heeft.

Kortom, in termen van de dimensie van onze ruimten maken we in één keer de stap van $n = 2$ en concreet ($\mathbb{R}^2$) naar $n = \infty$ en abstract (niet concreet). Let wel, dat kan *alleen* voor ruimten met een inwendig produkt.

[8] Inprodukt ruimten waarin Cauchy rijtjes convergent zijn.

[9] Lees: isomorf.

46.3 Reële Hilbertruimten

Een reële Hilbertruimte H is een vectorruimte over $\mathbb{R}$ die naast de vectoroptelling en scalaire vemenigvuldiging ook een inwendig produkt

$$(x, y) \in H \times H \to (x, y)_H = x \cdot y$$

heeft, met de standaardrekenregels, en de eigenschap dat alle Cauchyrijtjes (dat zijn rijtjes waarvoor

$$(x_n - x_m) \cdot (x_n - x_m) \to 0$$

als $m, n \to \infty$) in H ook convergent zijn met limiet $\overline{x} \in H$ (i.e.

$$(x_n - \overline{x}) \cdot (x_n - \overline{x}) \to 0$$

als $n \to \infty$).

De norm wordt gegeven door $|x|_H^2 = (x, x)_H = x \cdot x$ en de onderlinge afstand van bijvoorbeeld x_n en x_m is

$$d_H(x_n, x_m) = |x_n - x_m|_H = \sqrt{(x_n - x_m) \cdot (x_n - x_m)},$$

waarin $d_H : H \times H \to \mathbb{R}^+ = [0, \infty)$ de *metriek* is op H. De subscripts H laten we voortaan weg, tenzij dat verwarring geeft.

Exercise 46.3. Formuleer en bewijs de ongelijkheid van Cauchy-Schwarz[10] (inclusief de karakterisatie van het geval van gelijkheid), bewijs de driehoeksongelijkheid, en formuleer en bewijs nog een keer Pythagoras en de parallellogramwet. Hint: overschrijven uit willekeurig Lineaire Algebra boek. Formuleer ook de axioma's voor metrische ruimten en bewijs deze voor d.

Exercise 46.4. Laat H een Hilbertruimte zijn, $K \subset H$ een gesloten convexe verzameling, en $a \in H$. Bewijs dat er een unieke $p \in K$ is die de afstand $d(a, K)$ van a tot K realiseert middels

$$|p - a| = \inf_{x \in K} |x - a| = d(a, K)$$

en laat zien dat $(p - a) \cdot (x - p) \geq 0$ voor alle $x \in K$. Hint: geef eerst de definities van gesloten, convex en afstand, en gebruik daarna de parallellogramwet, net zoals in Opgave 45.14. Bewijs ook dat de afbeelding $P_K : H \to K$ gedefinieerd door $P_K(a) = p$ de eigenschap heeft dat $|P_K(a) - P_K(b)| \leq |a - b|$ voor alle $a, b \in H$.

[10] De ongelijkheid in Opgave 45.9.

Exercise 46.5. Laat H een Hilbertruimte zijn, $L \subset H$ een gesloten lineaire deelruimte. Bewijs dat $P_L : H \to L$ lineair is en dat

$$M = N(P_L) = \{x \in H : P_L(x) = 0\} = L^\perp = \{x \in H : x \cdot y = 0 \ \forall\, y \in L\}^{11},$$

de *kern* of *nulruimte* van P_L, ook een gesloten lineaire deelruimte is met $M \cap L = \{0\}$. Laat zien dat $M + L = H$ en concludeer dat $L \oplus M = H$: iedere $x \in H$ is uniek te schrijven als $x = p + q$ met $p \in L$ en $q \in M$.

De uitspraak over het bestaan van p in Opgave 46.4 is natuurlijk equivalent met de uitspraak over het bestaan van het minimum van

$$(x - a) \cdot (x - a) = x \cdot x - 2a \cdot x + a \cdot a,$$

en daarmee dus equivalent met een uitspraak over minima op K van wat je parabolische functies zou kunnen noemen:

Exercise 46.6. Laat H een Hilbertruimte zijn, $K \subset H$ een gesloten convexe verzameling. Dan neemt voor iedere $b \in H$ de kwadratische uitdrukking[12]

$$|x|^2 + b \cdot x$$

op K in precies één punt een minimum[13] aan.

Let op de eerste voeTnoot in Opgave 46.6. Het standaardinprodukt in $\mathbb{R}^2$ geeft via

$$\begin{pmatrix} a \\ b \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \underbrace{ax + cy = \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}}_{matrix\ notatie}$$

een representatie van de lineaire functie

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \underbrace{ax + cy = \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}}_{matrix\ notatie}$$

en omgekeerd is *iedere* lineaire[14] functie van deze vorm. De correspondentie

$$\begin{pmatrix} a \\ b \end{pmatrix} \quad \leftrightarrow \quad \begin{pmatrix} a & b \end{pmatrix}$$

[11] $x \cdot y = 0$ voor alle y in L wordt kort geschreven als: $x \cdot y = 0 \ \forall\, y \in L$.
[12] Het kwadratische stuk kan algemener, het lineaire stuk niet!
[13] K is i.h.a. *niet* begrensd, laat staan rijkompakt (iets met convergente deelrijen).
[14] De constante functie is NIET lineair, tenzij de constante nul is.

is evident bijectief en lineair. Links staat een 2-vector, rechts een 1 bij 2 matrix waarmee een lineaire afbeelding van $\mathbb{R}^2$ naar $\mathbb{R}$ gemaakt wordt.

In een willekeurige Hilbertruimte is er a priori geen matrixnotatie voor het maken van lineaire afbeeldingen. Welke lineaire afbeeldingen hebben we op zo'n Hilbertruimte H als van H verder niets gegeven is, behalve dan dat het een H is? Wel, in ieder geval is voor elke $y \in H$ de afbeelding $\phi_y : H \to \mathbb{R}$ gedefinieerd door[15]

$$x \to y \cdot x = \phi_y(x) = \phi_y x =< \phi_y, x > .$$

Kijk even goed, in de 1 na laatste notatie hebben we de haken weggelaten, zoals vaker bij lineaire afbeeldingen[16], en in de laatste staan ϕ_y en x zo te zien *gelijkwaardig* tussen strange brackets[17], waarbij net als in $y \cdot x$ de rollen van de tegenspelers verwisseld kunnen worden. Dualiteit heet dat met een mooi woord.

Voorlopig gebruiken we de notatie die het meest op de schoolnotatie lijkt. Een functie f van x, in dit geval ϕ_y, maak je expliciet[18] via $f(x)$, in dit geval $\phi_y(x) = y \cdot x$. Dat lijkt expliciet maar is het natuurlijk niet echt als we niet zeggen wat H is. Expliciet of niet, uit de ongelijkheid van Cauchy-Schwarz volgt nu dat

$$|\phi_y(x)| = |y \cdot x| \leq |y||x|$$

en dus ook, *vanwege de lineairiteit* , dat

$$|\phi_y(x_1) - \phi_y(x_2)| = |y \cdot (x_1 - x_2)| \leq L|x_1 - x_2| \quad \text{with} \quad L = |y|.$$

Een reëelwaardige functie f op een vectorruimte met een norm, die voldoet aan

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2|$$

voor alle x_1 en x_2 in die genormeerde vectorruimte, heet *Lipschitz continu*. Een mooi begrip, dat differentiaalrekening noch epsilons en delta's nodig heeft.

Exercise 46.7. Als zo'n (niet per se lineaire) functie een L heeft dan heeft hij ook een kleinste L. Bewijs dit. Hint: denk aan grootste ondergrenzen (infima).

[15] Meteen maar met drie notaties.

[16] en bij $\cos, \sin, \tan, \ldots$.

[17] Tussen bra en ket, zoals fysici soms zeggen.

[18] Of niet, en dat veroorzaakt vaak veel verwarring.

Die kleinste L is dus voor alle Lipschitz continue functies op onze genormeerde ruimte (laten we die X noemen) gedefinieerd. Daar hoort een zijstapje bij:

Exercise 46.8. Voor elke genormeerde ruimte X vormen de Lipschitz continue functies $f : X \to \mathbb{R}$ een vectorruimte $Lip(X)$ met de vectorbewerkingen gedefinieerd door

$$(f+g)(x) = f(x) + g(x) \quad \text{and} \quad (tf)(x) = tf(x).$$

Voor elke f is de kleinste L zoals boven per definitie een soort norm van f, die we noteren met $L = [f]_{Lip}$. Waarom definieert

$$f \to [f]_{Lip}$$

geen norm op $Lip(X)$? En waarom wel op

$$Lip_0(X) = \{f \in Lip(X) : \ f(0) = 0\}?$$

Bewijs dat met deze norm elke Cauchyrij $f_n \in Lip_0(X)$ convergent is. Hint: bewijs dit eerst voor $X = \mathbb{R}$ en schrijf je bewijs nog een keer over voor $X = X$.

Klein probleempje is natuurlijk dat er misschien maar weinig van die al of niet lineaire Lipschitz functies op X zijn, als je verder niks van X weet. Maar op H is dat probleempje er niet. Elke $y \in H$ geeft je een ϕ_y in $Lip_0(H)$ die nog lineair is ook, en je ziet meteen wat de kleinste L is: op zijn hoogst $|y|$ en kleiner kan niet, vul maar $x = y$ in. Dat betekent dat we met $y \to \phi_y$ een afbeelding

$$\Phi := H \to Lip_0(H)$$

hebben, en het beeld van Φ is bevat in H^*, de (genormeerde) ruimte van Lipschitz continue *lineaire* functies $f : H \to \mathbb{R}$, en Φ is zelf weer lineair[19]:

Exercise 46.9. Verifieer dat $\Phi : H \to H^*$ voldoet aan

$$\Phi(x_1 + x_2) = \Phi(x_1) + \Phi(x_2) \quad \text{and} \quad \Phi(tx) = t\Phi(x)$$

voor alle $t \in \mathbb{R}$ en $x, x_1, x_2 \in H$, en dat $[\Phi(x)]_{Lip} = |x|$.

De vraag nu is of Φ surjectief is: is elke $f \in H^*$ van de vorm ϕ_y? Bekijk daartoe[20]

$$N_f = \{x \in H : f(x) = 0\}.$$

[19] Nu maar eens de axioma's noemen en verifiëren.

[20] We schrijven nu N_f i.p.v. $N(f)$, t.b.v. het onderscheid tussen f en P_L.

Exercise 46.10. Bewijs dat $N_f \subset H$ een gesloten lineaire deelruimte is.

In het bijzonder bestaat nu dankzij Opgave 46.5 de projectie

$$P_{N_f} : H \to N,$$

ook weer een lineaire afbeelding, en in de volgende opgave gaat het om de nulruimte van deze projectie op de nulruimte van f.

Exercise 46.11. Bewijs dat $M = N(P_{N_f})$ een gesloten lineaire deelruimte is die gegeven wordt door $M = \{te : t \in \mathbb{R}\}$ waarin $e \in N_f^{\perp}$ met $|e| = 1$. Laat zien dat f een veelvoud is van ϕ_e: $f(x) = f(e)e \cdot x$.

Exercise 46.12. Leg uit waarom met het resultaat in Opgave 46.10 de afbeelding $\Phi : H \to H^*$ een lineaire isometrie is.

Lineaire isometriën zijn de mooiste continue afbeeldingen die er bestaan. De inverse van Φ wordt de Riesz representatie van H^* genoemd, en via deze isometrie erft H^* ook het inwendig produkt van H: de reële Hilbertruimten H en H^* zijn als Hilbertruimten hetzelfde, al is het in concrete situaties niet altijd even handig om hier de nadruk op te leggen.

Het resultaat geldt zonder enige verdere restrictie op H en het is ook niet nodig om aan te nemen dat H separabel is. We noteren de inverse van Φ als

$$R_H,$$

met de ruimte H als subscript aan $R = \Phi^{-1}$ gehangen. Het domein van R_H is zo de deelruimte

$$H^* \subsetneq Lip_0(H).$$

Exercise 46.13. Gebruik Opgave 46.4 om aan te tonen dat er plenty niet-lineaire functies in $Lip_0(H)$ zijn.

46.4 De standaard Hilbertruimte

De wat informeel geintroduceerde ruimte $\mathbb{R}^\infty$ bestaat uit alle functies $f, x, a : \mathbb{N} \to \mathbb{R}$, hoe je ze ook wil noemen[21]. We kunnen deze functies zien als kolomvectoren $\vec{f}$ met daarin de waarden van f, al protesteert LaTeχ daarbij zo te zien een beetje. Helemaal op dezelfde hoogte lukt typografisch niet,

$$f = \begin{pmatrix} f(1) \\ f(2) \\ f(3) \\ \vdots \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{pmatrix} = \vec{f},$$

en ook voor functies $f, x, a : \{1, 2, 3\} \to \mathbb{R}$ oogt

$$f = \begin{pmatrix} f(1) \\ f(2) \\ f(3) \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} = \vec{f}$$

niet echt lekker.

Wiskundig gezien praten we de facto over functies $f : A \to \mathbb{R}$, waarbij A hier een *discrete* verzameling is, en de verzameling van deze functies wordt ook wel genoteerd als $\mathbb{R}^A$. Als je $n \in \mathbb{N}$ gedefinieerd hebt als een wat rare verzameling, via[22]

$$1 = \{\emptyset\},\ 2 = \{\emptyset, \{\emptyset\}\},\ 3 = \{\emptyset, \{\emptyset, \{\emptyset\}\}\}, \ldots$$

of zoiets, dan is de notatie voor $\mathbb{R}^A$ consistent[23] met die voor $\mathbb{R}^n$.

Elke $f \in \mathbb{R}^A$ heeft als Pythagoras norm

$$|f| = \sqrt{\sum_{a \in A} |f(a)|^2},$$

hetgeen voor $A = \{1, 2, 3\}$ overeenkomt met de Euclidische lengte

$$\sqrt{f_1^2 + f_2^2 + f_3^3}$$

van de vector $\vec{f}$ hierboven.

Het ligt voor de hand om $\mathbb{R}^A$ en $\mathbb{R}^B$ als dezelfde ruimte te zien als er een bijectie bestaat tussen A en B. Voor eindige verzamelingen A is de Pythagoras norm natuurlijk op heel $\mathbb{R}^A$ gedefinieerd, maar als A oneindig veel elementen bevat[24] dan is dat niet meer het geval.

[21] Het zijn er meer dan 26.

[22] Ik schuif wat ik naief in Halmos las wellicht eentje op, boekje niet bij de hand.

[23] Let wel, $0 = \emptyset$ doet niet mee.

[24] We zeggen dan gemakshalve dat A oneindig is.

Exercise 46.14. Stel dat A overaftelbaar is en $f \in \mathbb{R}^A$ eindige Pythagorasnorm heeft. Bewijs dat de verzameling

$$\{a \in A : \, f(a) \neq 0\}$$

aftelbaar is en ga na dat het in de somnotatie dan niet nodig is de volgorde van sommeren vast te leggen[25].

In het licht van deze opgave beperken we de aandacht voor oneindige A tot aftelbare A en die zijn allemaal bijectief met $\mathbb{N}$. We kunnen de functiewaarden van elementen $x, f, \ldots \in \mathbb{R}^{\mathbb{N}}$ dan op wat voor manier dan ook weer (niet allemaal) opschrijven, genummerd als $f(n)$ of x_n met $n = 1, 2, \ldots$, in bijvoorbeeld een kolomvector of rijvector met puntjes.

Onze standaard aftelbaar oneindig-dimensionale Hilbertruimte is nu

$$l^{(2)} = \{x = (x_1, x_2, \ldots) \in \mathbb{R}^{\mathbb{N}} : \sum_{n=1}^{\infty} x_n^2 < \infty\},$$

spreek uit: *(kleine) el twee*. Er is ook een *grote el twee*, namelijk de verzameling van kwadratisch integreerbare meetbare functies op een maatruimte, bijvoorbeeld[26] $\mathbb{R}$, voorzien van de gewone (Lebesgue) lengtemaat[27]. Die *el twee* wordt genoteerd met

$$L^2(\mathbb{R}),$$

strict genomen geen functieruimte maar een ruimte van equivalentieklassen. We zeggen dat een (meetbare) functie f en een andere (meetbare) gunctie g equivalent zijn, notatie $f \sim g$, als de verzameling waarop ze verschillen (uitwendige) maat NUL heeft, en met f bedoelen we stiekem $[f]$, de equivalentieklasse van alle g waarvoor $g \sim f$.

De inwendige produkten zijn, respectievelijk,

$$x \cdot y = (x, y)_{l^{(2)}} = \sum_{n=1}^{\infty} x_n y_n \quad \text{and} \quad f \cdot g = (f, g)_{L^2(\mathbb{R})} = \int_{\infty}^{\infty} f(x) g(x) dx,$$

waarbij de integraalnotatie bij (niet ieders) voorkeur hetzelfde gekozen wordt als die van de Riemann integraal.

Exercise 46.15. Bewijs dat $l^{(2)}$ *volledig* is. Dat wil zeggen, laat zien dat Cauchy rijtjes in $l^{(2)}$ convergent zijn met limiet in $l^{(2)}$.

[25] Dit heet onvoorwaardelijke convergentie.

[26] Ander voorbeeld: $\mathbb{R}$ modulo 2π, de facto de eenheidscirkel in $\mathbb{R}^2$.

[27] Zie "Wiskunde in je vingers" van H&M voor snelle intro maattheorie.

Als we $\mathbb{N}$ zien als maatruimte voorzien van de telmaat dan wordt kleine l weer groot. En met recht, want iedere separabele Hilbertruimte H is met $l^{(2)}$ te identificeren[28]. Hoe gaat dat? Wel, neem een rijtje $a_1, a_2, a_2, \ldots$ in H dat als limietpunten alle elementen van H heeft. Zet

$$e_1 = \frac{1}{|a_1|} a_1$$

als $a_1 \neq 0$ maar gooi a_1 weg als $a_1 = 0$. Hernummer in dat geval de rij en herhaal deze stap, net zolang[29] tot je een $a_1 \neq 0$ hebt. Stel vervolgens

$$y_2 = a_2 - (a_2, e_1)e_1 \quad \text{and} \quad e_2 = \frac{1}{|y_2|} y_2$$

als $y_2 \neq 0$, maar gooi a_2 weg als $y_2 = 0$ en hernummer in dat geval weer de rij. Herhaal deze stap, net zolang tot je een $y_2 \neq 0$ hebt en daarmee ook een e_2. Stel vervolgens

$$y_3 = a_3 - (a_3, e_2)e_2 - (a_3, e_1)e_1 \quad \text{and} \quad e_3 = \frac{1}{|y_3|} y_3,$$

als $y_3 \neq 0$, maar $\ldots$, enzovoorts. Dit produceert een rij $e_1, e_2, e_3, \ldots$ van vectoren waarvoor

$$(e_i, e_j) = \delta_{ij},$$

en deze vectoren spannen een lineaire deelruimte op in H.

Exercise 46.16. Bewijs dat

$$H = \{x = \sum_{n=1}^{\infty} x_n e_n : \sum_{n=1}^{\infty} x_n^2 < \infty\},$$

waarmee H dus met de standaard Hilbertruimte $l^{(2)}$ geïdentificeerd *kan* worden.

[28] Indien gewenst.

[29] Nou ja, als er geen dubbelen in de rij voorkomen dan...

47 Fourier series

Het ligt voor de hand om de abstracte constructie van H uit $C(\mathbb{R}_{2\pi})$ te zien als gebeurende in het platte vlak $\mathbb{R}^2$, waarbij de grafiek G van een functie $f : \mathbb{R}_{2\pi} \to \mathbb{R}$ dus de eigenschap moet hebben dat

$$\frac{x_1 - x_2}{2\pi} \in \mathbb{Z} \implies f(x_1) = f(x_2),$$

hetgeen overeenkomt met het oprolbaar[1] zijn van het oneindige platte vlak tot een cylinder waarin de grafiek G keurig over zichzelf heen ligt.

Met of zonder voorstelling, twee verschillende 2-Cauchyrijtjes $f_1, f_2, \ldots$ en $g_1, g_2, \ldots$ in $C(\mathbb{R}_{2\pi})$ moeten hetzelfde element uit de te maken H zijn als geldt dat

$$|f_n - g_n| \to 0$$

voor $n \to \infty$. Opgave 47.10 laat bijvoorbeeld zien dat de zaagtandfunctie Z in de te maken H moet zitten, maar niet iedereen zal dezelfde rij $Z_1, Z_2, \ldots$ als grafieken getekend hebben. Het is goed om dat nog wat preciezer te bekijken.

Exercise 47.1. Maak Opgave 47.10 nog een keer maar anders. Teken de grafieken van een rij functies $\tilde{Z}_n \in C(\mathbb{R}_{2\pi})$ waarvoor geldt dat $(\tilde{Z}_n - Z, \tilde{Z}_n - Z) \to 0$ als $n \to \infty$. Kies de rij functies $\tilde{Z}_1, \tilde{Z}_2, \ldots$ nu zo dat voor alle $n \in \mathbb{N}$ geldt dat $\tilde{Z}_n(0) = 1$.

Exercise 47.2. Maak Opgave 47.1 maar nu met $\tilde{Z}_n(0) = 0$.

Exercise 47.3. Maak Opgave 47.2 maar nu met $\tilde{Z}_n(0) = 2$.

Exercise 47.4. Maak Opgave 47.2 maar nu met $\tilde{Z}_n(0) = n$.

Exercise 47.5. Laat in Opgaven 47.1,47.2,47.3,47.4 hierboven zien dat $|Z_n - \tilde{Z}_n| \to 0$ als $n \to \infty$, waarbij Z_n is als in Opgave 47.1.

[1] Stel je de problemen bij het oprollen even voor....

Wat deze opgaven laten zien is dat functies in H geen gewone functies kunnen zijn. Abstract gezien zouden alle benaderende rijen dezelfde $Z : \mathbb{R}_{2\pi} \to \mathbb{R}$ moeten maken, maar dat lijkt via de opgaven hierboven te leiden tot de conclusie dat

$$0 = Z(0) = 1$$

en meer verwarring. Soortgelijke spelletjes kunnen we spelen met de nulfunctie

$$x \xrightarrow{\mathbf{0}} 0$$

zelf.

Exercise 47.6. Maak een rij functies $f_1, f_2, \ldots$ in $C(\mathbb{R}_{2\pi})$ waarvoor geldt dat $f_n(0) = 1$ en $|f_n| = |f_n - \mathbf{0}| \to 0$.

Iedere rij echte functies f_n die we gebruiken om een f in H te maken kan veranderd worden in een rij $\tilde{f}_n$ die in een gegeven punt gek gedrag vertoont, zoals convergeren naar een 'verkeerde' limiet, maar wel de eigenschap heeft dat $|f_n - \tilde{f}_n| \to 0$. We moeten kennelijk af van het idee dat een functie in elk punt gedefinieerd is. In sommige punten is dat wellicht een artefact, zoals in Opgave 47.6, maar bij functies als Z is er echt een keuze die gemaakt moet worden. Of niet, als we afspreken dat functies niet per se in elk punt van hun definitiegebied gedefinieerd hoeven zijn. Let op, met Z zitten ook alle verschoven zaagtandfuncties Z_p (met $p \in \mathbb{R}$)

$$x \xrightarrow{Z_p} Z(x - p)$$

in H, en daarmee ook een grote klasse van functies van de vorm

$$S = \sum_{n=1}^{\infty} a_n Z_{p_n},$$

waarbij $p_1, p_2, \ldots$ een willekeurige rij punten in $\mathbb{R}$ mag zijn, en elke p_n een probleempunt is voor wat betreft de definitie van $S(p_n)$.

Exercise 47.7. Neem aan dat H geconstrueerd is zoals hierboven beschreven. Neem aan dat $p_1, p_2, \ldots$ en $a_1, a_2, \ldots$ rijen in $\mathbb{R}$ zijn, en dat

$$\sum_{n=1}^{\infty} |a_n| < \infty.$$

Waarom moet gelden dat $S \in H$? Hint: laat eerst zien dat S een begrensde functie is.

Kortom, behalve de mooie periodieke functies

$$x \xrightarrow{c_n} \cos nx$$

en

$$x \xrightarrow{s_n} \sin nx$$

($n \in \mathbb{N}$), waarvan we ook sommen van de vorm

$$\sum_{n=1}^{\infty} a_n c_n + \sum_{n=1}^{\infty} b_n s_n \tag{47.1}$$

kunnen nemen, met coefficiënten als in Opgave 47.7, moeten er in de H die we zoeken een heleboel lelijke functies zitten. Daarbij moeten evident verschillende functies soms (of vaak) als element van H als dezelfde functie gezien worden. Waarom? Omdat twee Cauchyrijen $f_1, f_2, \ldots$ en $\tilde{f}_1, \tilde{f}_2, \ldots$ in $C(\mathbb{R}_{2\pi})$ met de eigenschap dat $f_n - \tilde{f}_n \to 0$ dezelfde f in H moeten maken, en we in de voorbeelden gezien hebben dat bijvoorbeeld $f_n(0)$ en $\tilde{f}_n(0)$ verschillende of helemaal geen limieten kunnen hebben.

Exercise 47.8. Maak een Cauchyrij $f_1, f_2, \ldots$ die naar de nulfunctie $\mathbf{0}$ convergeert in de inproduktnorm maar waarvoor de rij $f_1(x), f_2(x), \ldots$ niet convergeert, welke $x \in \mathbb{R}_{2\pi}$ je ook kiest.

De vraag is dus niet alleen welke functie je kiest als de meest natuurlijke functie binnen een equivalentieklasse van functies die in H niet van elkaar te onderscheiden zijn, maar ook hoe je überhaupt aan zo'n functie komt als f in H gedefinieerd is via een Cauchyrij $f_1, f_2, \ldots$ in $C(\mathbb{R}_{2\pi})$.

47.1 Standaard Hilbertruimten voor 'functies'

In wat volgt maken we enerzijds precies welke functies f op te vatten zijn als $f \in H$ en anderzijds waarom we die functies nog wel als functies zien. Iedere $f \in H$ moet daartoe voor bijna[2] alle $x \in \mathbb{R}_{2\pi}$ een natuurlijke waarde hebben, waarbij het gedrag van f in de buurt van elk zulk een x leidend moet zijn[3]. Voor de zaagtand Z leidt dit bij het gelijkwegen van wat $Z(x)$ is voor $x < 0$ en $x > 0$ onherroepelijk tot $Z(0) = 0$ als de natuurlijke keuze voor $Z(0)$, het gemiddelde van de linker- en rechterlimiet. Maar of zulke limieten

[2]Wat *bijna* betekent is de hamvraag.

[3]Waarom eigenlijk? Wel, we zijn uitgegaan van continue functies.

voor iedere f in de H die we maken altijd in genoeg punten bestaan is (zeker a priori) niet zo duidelijk.

Hoe het ook zij, de waarde van $f \in H$ in $\pi = -\pi \in S$ doet er niet toe. Voor iedere $a \in \mathbb{R}$ en iedere functie $f : (a-\pi, a+\pi)$ die we toe willen laten in H na periodieke uitbreiding van f tot $\mathbb{R} \to \mathbb{R}$ is het niet belangrijk of en hoe $f(a-\pi)$ en $f(a+\pi)$ gedefinieerd zijn. In het bijzonder is de functie $\tilde{Z}$ gedefinieerd

$$x \in (0, 2\pi) \xrightarrow{\tilde{Z}} \pi - x$$

na periodieke uitbreiding tot $Z : \mathbb{R} \to \mathbb{R}$ in H gelijk aan de Z uit Opgave 47.10. Waar bij de functies c_n en s_n het periodiek uitbreiden vanzelf gaat, is het bij functies als Z vervelend om de formules überhaupt op te schrijven.

De functie Z heeft in ieder geheel veelvoud van 2π een sprong. De eveneens oneven blokfunctie $\mathit{blok} \in H$, gedefinieerd door

$$\mathit{blok}(x) = \begin{cases} 1 & \text{als } x \in (0, \pi) \ ; \\ -1 & \text{als } x \in [-\pi, 0), \end{cases}$$

heeft in ieder geheel veelvoud van π een sprong. De even *kartelrandfunctie* $\mathit{Ka} \in H$ daarentegen, gedefinieerd door

$$\mathit{Ka}(x) = \begin{cases} \frac{\pi}{2} - x & \text{als } x \in (0, \pi) \ ; \\ \frac{\pi}{2} + x & \text{als } x \in [-\pi, 0), \end{cases}$$

heeft geen sprongen als we de definitie van Ka uitbreiden met $\mathit{Ka}(2\pi n) = \frac{\pi}{2}$ in de gehele veelvouden $2\pi n$ van 2π ($n \in \mathbb{Z}$). Al deze functies zijn instructief als voorbeeld bij de vraag of ze te schrijven zijn als een oneindige som van de vorm (47.1). Met name de zaagtand is een bron van leerzaam vermaak zoals we zullen zien.

Exercise 47.9. Schets de grafieken van $Z, \mathit{blok}, \mathit{Ka}$, en ook van $c_1 = \cos$ en $s_1 = \sin$.

Ga nog eens na dat de functies

$$\frac{c_n}{\sqrt{\pi}}, \ \frac{s_n}{\sqrt{\pi}} \quad (n \in \mathbb{N}), \ \frac{1}{\sqrt{2\pi}}$$

een orthonormaal stelsel vormen, en dat H dus alle 'functies' f van de vorm

$$f = a_0 \frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} a_n \frac{c_n}{\sqrt{\pi}} + \sum_{n=1}^{\infty} b_n \frac{s_n}{\sqrt{\pi}} \tag{47.2}$$

zou moeten bevatten, meestal geschreven als

$$f = \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n c_n + b_n s_n),$$

als de (net iets andere) rijtjes van coëfficiënten a_n en b_n maar kwadratisch sommeerbaar zijn.

De vraag of je zo alle f in H krijgt kan beginnen met de vraag of de oneven functies Z en *blok* te schrijven zijn als

$$\sum_{n=1}^{\infty} b_n s_n,$$

en de even functie *Ka* als

$$\sum_{n=1}^{\infty} a_n c_n.$$

De sommen moeten hierbij convergent zijn in de 2-norm die hoort bij het standaard inproduct

$$f \cdot g = (f, g) = \int_{-\pi}^{\pi} f(x)g(x)\,dx.$$

Ons doel is te laten zien dat iedere f in H inderdaad van de vorm (47.2) met

$$\sum_{n=0}^{\infty} a_n^2 < \infty, \quad \sum_{n=1}^{\infty} b_n^2 < \infty,$$

en een karakterisatie van H als $L^2(\mathbb{R}_{2\pi})$ die los staat van de specifieke keuze die we met (47.2) maken.

47.2 Functies op de cirkel

Als we afspreken dat twee getallen in $\mathbb{R}$ eigenlijk hetzelfde zijn als ze een geheel veelvoud van 2π verschillen dan maken $\mathbb{R}$ en 2π de verzameling $\mathbb{R}_{2\pi}$, een verzameling waarin op natuurlijke manier de optelling is gedefinieerd. Dat gaat net als in $\mathbb{Z}_n$, de verzameling die we krijgen uit de verzameling $\mathbb{Z}$ van gehele getallen en een vast getal $n \in \mathbb{N}$, door af te spreken dat twee gehele getallen gelijk zijn als ze een geheel veelvoud van n verschillen. Zoals vaak

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

wordt geschreven, met $0 = n$, kunnen we ook

$$\mathbb{R}_{2\pi} = [0, 2\pi)$$

schrijven, maar we geven er de voorkeur om $\mathbb{R}_{2\pi}$ in de schrijfwijze te laten corresponderen met $[-\pi, \pi)$, waarbij $-\pi = \pi$. Deze π is hier een positief reëel getal, waarvoor we op enig moment de π die we van de cirkel kennen zullen nemen, *maar dat hoeft nu nog even niet*. Net als $\mathbb{Z}_n$ is $\mathbb{R}_{2\pi}$ met de voor de hand liggende optelling een commutatieve groep[4].

Functies $f : \mathbb{R} \to \mathbb{R}$ die 2π-periodiek zijn kunnen we ook opvatten als functies $f : \mathbb{R}_{2\pi} \to \mathbb{R}$, en omgekeerd. De verzameling van continue 2π-periodieke functies noemen we

$$C(\mathbb{R}_{2\pi}).$$

Ieder tweetal functies gedefinieerd op dezelfde verzameling, dus ook f en g in $C(\mathbb{R}_{2\pi})$, kunnen we bij elkaar optellen[5] middels

$$x \xrightarrow{f+g} f(x) + g(x)$$

als definitie van $f + g \in C(\mathbb{R}_{2\pi})$. Met

$$x \xrightarrow{tf} tf(x)$$

voor $t \in \mathbb{R}$ en $f \in C(\mathbb{R}_{2\pi})$ is ook de scalaire vermenigvuldiging gedefinieerd en zo is $C(\mathbb{R}_{2\pi})$ een vectorruimte[6] over $\mathbb{R}$, waarop

$$f \cdot g = (f, g) = \int_{-\pi}^{\pi} f(x)g(x)\,dx$$

een inwendig produkt[7] definieert, maar $C(\mathbb{R}_{2\pi})$ is met dit integraalinprodukt geen Hilbertruimte, zoals de volgende opgave laat zien.

Exercise 47.10. De zaagtandfunctie Z wordt gedefinieerd door

$$Z(x) = \begin{cases} \pi - x & \text{als } x \in (0, \pi] \ ; \\ -x - \pi & \text{als } x \in [-\pi, 0), \end{cases}$$

[4]Google: Abelian group.

[5]Evenzo is natuurlijk ook fg gedefinieerd via $x \xrightarrow{fg} f(x)g(x)$.

[6]En met de vermenigvulding een algebra.

[7]Let op, met $(f, g) \dot{\to} f \cdot g = (f, g)$ is de haakjesnotatie soms verwarrend.

en door $Z(0) = 0$. Met deze keuze voor $Z(0)$ behoort Z tot $\mathcal{G}_0(\mathbb{R}_{2\pi})$, de ruimte[8] van functies $f : \mathbb{R}_{2\pi} \to \mathbb{R}$ die continu zijn in iedere $x \in \mathbb{R}_{2\pi}$, behalve eventueel in $x = 0$, maar waarvoor wel geldt dat

$$f(0) = \frac{1}{2}\left(\lim_{x\downarrow 0} f(x) + \lim_{x\uparrow 0} f(x)\right),$$

waarbij linker- en rechterlimiet dus allebei bestaan. Op $\mathcal{G}_0(\mathbb{R}_{2\pi})$ is $(f, g) \to f \cdot g$ ook een inprodukt. Teken de grafieken van een rij functies $Z_1, Z_2, \ldots$ in $C(\mathbb{R}_{2\pi})$ waarvoor geldt dat $(Z_n - Z, Z_n - Z) \to 0$ als $n \to \infty$.

De rij $Z_1, Z_2, \ldots$ is convergent in $\mathcal{G}_0(\mathbb{R}_{2\pi})$ met betrekking tot de inproduktnorm

$$f \to |f| = \sqrt{(f, f)} = \left(\int_{-\pi}^{\pi} |f|^2\right)^{\frac{1}{2}}$$

omdat $|Z_n - Z| \to 0$ als $n \to \infty$, en dus is de rij $Z_1, Z_2, \ldots$ ook een Cauchyrij in $C(\mathbb{R}_{2\pi})$ met die inproduktnorm, die echter in $C(\mathbb{R}_{2\pi})$ geen limiet heeft[9]. Om van $C(\mathbb{R}_{2\pi})$ met de inproduktnorm een Hilbertruimte te maken, die we de naam

$$L^2(\mathbb{R}_{2\pi})$$

willen geven, moeten we alle limieten van Cauchyrijtjes aan $C(\mathbb{R}_{2\pi})$ toevoegen, maar hoe doe je dat?

47.3 Dat andere inprodukt met afgeleiden

De deelruimte V van H die de rol gaat spelen zoals in de eerdere voorbeelden met $H = l^{(2)}$ wordt gedefinieerd door het inprodukt

$$((f, g)) = (f', g'),$$

hetgeen niet voor alle f en g in H gedefinieerd is, net zoals het inprodukt in Opgave 34.3 niet voor alle x en y in $l^{(2)}$ gedefinieerd is. Informeel wordt V gegeven door

$$V = \{f \in H : f' \in L^2(-\pi, \pi)\},$$

waarbij met f ook f' steeds 2π-periodiek wordt uitgebreid tot een functie gedefinieerd op heel $\mathbb{R}$.

Dat uitbreiden is makkelijk, en komt in de opgaven hieronder eerst nog aan de orde, ook ter voorbereiding van wat een stuk lastiger is: *wat betekent het dat f' als meetbare en kwadratisch integreerbare functie bestaat?*

[8] De notatie $\mathcal{G}$ is alleen voor nu even.

[9] Waarom niet?

Exercise 47.11. Ga na dat (ook) voor functies f in $L^2(-\pi,\pi)$ met $f(-\pi) \neq f(\pi)$ er geen problemen zijn met de uitbreiding f naar een $f \in H$.

Exercise 47.12. Zijn er functies f in $L^2(-\pi,\pi)$ waarvoor aan $f(0)$ geen betekenis[10] kan worden gegeven?

Exercise 47.13. Er is maar één 2π-periodieke oneven[11] functie Ka die voldoet aan $Ka(x) = 1$ voor $0 < x < \pi$. Schets de grafiek van Ka en maak een rij 2π-periodieke oneven continue functies $Ka_1, Ka_2 \ldots$ waarvoor geldt dat $|Ka_n - Ka|_2 \to 0$ als $n \to \infty$. Hint: schets eerst de grafieken van Ka_n.

Exercise 47.14. Bewijs dat een oneven 2π-periodieke functie wordt vastgelegd door zijn functiewaarden op het interval $(0,\pi)$. Hint: gebruik de regels $f(-x) = -f(x)$ en $f(x) = f(x+2\pi)$. Wat is $f(0)$? En $f(\pi)$?

Leuke functies om over na te denken, maar zulke functies komen we niet tegen als we een zinvolle definitie van de uitspraak dat f' bestaat in bijvoorbeeld $L^2(0,\pi)$ kunnen geven. Wel is het zo f' best zelf zo'n functie kan zijn. Bijvoorbeeld als je f definieert als

$$f(x) = \int_0^x S(s)\,ds,$$

met een begrensde S zoals eerder gemaakt in Opgave 47.7. Iedere primitieve functie

$$F(x) = \int_0^x f(s)\,ds$$

van een f in H is natuurlijk in principe kandidaat om tot V te behoren.

[10] Lees: een betekenisvolle waarde kan worden toegekend?

[11] $Ka(-x) = -Ka(x)$ voor alle $x \in \mathbb{R}$.

Exercise 47.15. Verifieer dat zo'n F een begrensde (2π-periodieke) functie is als $f \in H$, en dat het essentieel is dat in de definitie van H is opgenomen dat voor $f \in H$ moet gelden dat[12]

$$\int_{-\pi}^{\pi} f(x)dx = 0\,!$$

De ruimte V krijgen we nu als bestaande uit de primitieve functies van functies in H, waarbij de spreekwoordelijke constante wel goed gekozen moet worden.

Exercise 47.16. Als $f \in H$ dan is F periodiek. Waarom? Ga na dat er voor elke $f \in H$ precies één constante C is waarvoor $x \to F(x) - C$ in H zit.

We weten nu dus wat V moet zijn. De ruimte

$$\{F \in L^2_{loc}(\mathbb{R}) : \ (\forall x \in \mathbb{R})\ F(x) = F(x + 2\pi),\ f = F' \in L^2_{loc}(\mathbb{R})\}$$

is gelijk aan

$$\{F \in L^2_{loc}(\mathbb{R}) : \ f = F' \in H\}$$

de ruimte van *alle* primitieven F van functies $f \in H$, en V krijgen we door voor iedere primitieve F precies die constante te nemen waarmee de primitieve gemiddeld nul wordt. Dus

$$V = \{F \in L^2_{loc}(\mathbb{R}) : \ f = F' \in H,\ \int_{-\pi}^{\pi} F(x)\,dx = 0\}$$

De kwadratisch integreerbare periodieke functies $f : \mathbb{R} \to \mathbb{R}$ vormen een nul-dimensionale vectorruimte waarover nog wel het een en ander te vertellen is. Dat zullen we hier niet doen. Periodieke functies kunnen natuurlijk wel *lokaal* kwadratisch integreerbaar zijn. We schrijven

$$L^2_{loc}(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} : \ f \in L^2(I) \text{ voor elke begrensd interval } I \subset \mathbb{R}\},$$

maar de periodieke functies in $L^2_{loc}(\mathbb{R})$ vormen geen vectorruimte[13]. En $L^2_{loc}(\mathbb{R})$ zelf is wel een vectorruimte maar geen genormeerde ruimte, althans niet met een natuurlijke Maar

$$H = \{f \in L^2_{loc}(\mathbb{R}) : \ (\forall x \in \mathbb{R})\ f(x) = f(x + 2\pi);\ \int_{-\pi}^{\pi} f(x)dx = 0\}$$

[12] $0! = 1$, maar hier roept het uitroepteken wel.

[13] Waarom niet?

wel, de ruimte van 2π-periodieke kwadratisch integreerbare[14] periodieke functies, met inprodukt

$$(f,g) = \int_{-\pi}^{\pi} f(x)g(x)\,dx,$$

waarbij we de ruimte nu beperken tot functies die gemiddeld nul zijn.

De constante functies zijn in deze H buitengesloten, omdat ze in het verhaal dat gaat volgen een vervelend buitenbeentje zijn. Bijgevolg van deze keuze zitten er in H ook geen positieve functies trouwens. Wel in H zitten de functies c_n en s_n uit Opgave 45.22 en net als elke functie in H zijn deze door beperking tot het interval $(-\pi,\pi)$ op te vatten als element van

$$\tilde{H} = \{f \in L^2(-\pi,\pi) : \int_{-\pi}^{\pi} f(x)\,dx = 0\},$$

een ruimte die we voor gemak met H identificeren door iedere $f \in \tilde{H}$ weer uit te breiden tot heel $\mathbb{R}$ middels $f(x) = f(x+2\pi)$ voor alle x.

47.4 Blipfuncties

Het formulevoorschrift

$$x \xrightarrow{\textit{blip}} \begin{cases} \exp(-\frac{1}{x}) & \text{for } x > 0 \\ 0 & \text{for } x \leq 0, \end{cases}$$

definieert de functie $\textit{blip} : \mathbb{R} \to [0,1)$ die met goed recht zowel oogverblindend mooi als gruwelijk lelijk genoemd[15] mag worden.

Exercise 47.17. Schets de grafiek van *blip* en onderzoek het gedrag van $\textit{blip}'(x)$ als $x \downarrow 0$. En van $\textit{blip}''(x)$. En van alle afgeleiden van *blip*. Concludeer dat *alle* afgeleiden van *blip* als continue functies van $\mathbb{R}$ naar $\mathbb{R}$ bestaan!

Exercise 47.18. Je kunt *blip* ook schalen. Definieer $\textit{blip}_n$ door

$$\textit{blip}_n(x) = \textit{blip}(nx) = \exp(-\frac{1}{nx})$$

en bepaal $\lim_{n\to\infty} \textit{blip}_n(x)$ voor elke $x \in \mathbb{R}$. De limietfunctie heet de Heaviside[16] functie, hier genoteerd als $He(x)$. Deze functie is niet continu in $x = 0$. De waarde

[14] D.w.z. f is meetbaar en $\int_{-\pi}^{\pi} f(x)^2 dx < \infty$.

[15] De naam *blip* heb ik gezien in een mooi boek, weet niet meer welk.

[16] Google Heaviside.

van $He(0)$ als limietwaarde van $blip_n(0)$ is 0, maar net zo vaak wordt $He(0) = \frac{1}{2}$ of $He(0) = 1$ genomen. Of zelfs $He(0) = [0, 1]$.

Exercise 47.19. De functies *blip* en *He* zitten niet in $L^2(\mathbb{R})$. Waarom niet? Maar $He - blip$ wel. Waarom? En $He - blip_n$ ook. De affiene ruimte

$$He + L^2(\mathbb{R}) = \{f = He + g : \, g \in L^2(\mathbb{R})\}$$

is voorzien van de 2-metriek

$$d(f, g) = \left(\int_{-\infty}^{\infty} |f(x) - g(x)|^2 \, dx \right)^{\frac{1}{2}}$$

een volledige metrische ruimte[17]. Laat zien zien dat $d(blip_n, He) \to 0$ als $n \to \infty$.

Exercise 47.20. Definieer de functies $blok_n$ door

$$blok_n i(x) = blip_n(x) blip_n(\pi - x)$$

en laat zien dat

$$\lim_{n \to \infty} blok_n(x) = \chi_{(0,\pi)}(x)$$

voor alle $x \in \mathbb{R}$. Waarom geldt dat $blok_n \to \chi_{(0,\pi)}$ in 2-norm?

Exercise 47.21. Dezelfde vragen als in Opgave 47.20 maar nu voor $blok_n$ gedefinieerd door

$$blok_n i(x) = blip_n(x - \frac{1}{n}) blip_n(\pi - \frac{1}{n} - x),$$

Opgave 47.21 laat zien dat $\chi_{(0,\pi)}$, opgevat als

Exercise 47.22. Het is goed om op een rijtje te zetten hoe je zeker weet dat elke $f \in L^2(-\pi, \pi)$ te benaderen is met een rij functies $f_1, f_2, \ldots$ in

$$C_c^\infty(-\pi, \pi),$$

[17] Wat is dat?

de ruimte van van functies $f : (-\pi, \pi) \to \mathbb{R}$ die oneindig vaak differentieerbaar zijn en identiek nul zijn in de buurt van $x = 0$ en $x = 2\pi$. Benaderen betekent hier dat $f_n \to f$ in de 2-norm. Het speciale geval om eerst te begrijpen is

$$f(x) = \chi_I(x) = \begin{cases} 1 & \text{als } x \in I \\ 0; & \text{als } x \notin I, \end{cases}$$

met I een interval.

47.5 Intermezzo: out of Hilbertspace

De 2-norm is een bijzonder geval van

$$f \to |f|_p = \left(\int_{-\pi}^{\pi} |f(x)|^p \, dx \right)^{\frac{1}{p}},$$

waarmee voor $1 \leq p < \infty$ de p-norm op $C[-\pi, \pi]$ wordt gedefinieerd, en

$$|f|_\infty = \max_{x \in [-\pi, \pi]} |f(x)|,$$

de maximumnorm van f. Deze p-normen ($1 \leq p \leq \infty$) zijn te vergelijken met

$$|x|_p = \left(\sum_{j=1}^{N} |x_j|^p \right)^{\frac{1}{p}},$$

de p-norm van $x = (x_1, \ldots, x_n) \in \mathbb{R}^{\mathrm{N}}$.

Exercise 47.23. Terug naar de overgeslagen calculussommetjes, bewijs (de ongelijkheid van Hölder)

$$|x \cdot y| \leq |x|_p |y|_q$$

voor $1 \leq p, q \leq \infty$ die voldoen aan

$$\frac{1}{p} + \frac{1}{q} = 1,$$

en $x = (x_1, \ldots, x_n)$ en $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^{\mathrm{N}}$. Hint: leg eerst uit waarom het *geen* beperking is om aan te nemen dat $|x|_p = |y|_q = 1$.

Exercise 47.24. Bewijs dat $x \to |x|_p$ een norm is op $\mathbb{R}^{\mathrm{N}}$.

Exercise 47.25. Bewijs dat $|x|_p \to |x|_\infty$ als $p \to \infty$.

Exercise 47.26. Verzin en maak de analoge opgaven voor

$$f \to \left(\int_{-\pi}^{\pi} |f(x)|^p \, dx \right)^{\frac{1}{p}},$$

de p-norm op $C[-\pi, \pi]$.