# Fundamentals of analysis and more

Not another $\varepsilon$-book

watch on you tube via

https://youtu.be/taKW6cNomJk

Chapters 1-11 are course notes for the first year bachelor course Mathematical Analysis (XB_0009). Chapters 4 and 5 conclude the first part of a course that can be based on $\varepsilon - N$ techniques and estimates. The language in these chapters is fundamental for Chapter 7, the basics of integration and integral equations. Chapter 7 is essential for what follows in Chapter 28 or elsewhere on Fourier theory. Chapter 4 is convenient, though not essential, for Chapter 9: differential calculus for power series. From the full blooded $\varepsilon - \delta$ machinery Chapter 8 only Theorem 8.1 is used for differentiation, for Theorem 10.10 only in fact.

NB the blue parts we won't be able to do in a first course

1. Archimedes and the real numbers: https://www.youtube.com/playlist?list=PLQgy2W8pIli9lL_Yp_4Tr_CG0grGUBdZa
   (first video of this one has course contents and all that, not the first one I made)
2. Epsilon-N-techniques: https://www.youtube.com/playlist?list=PLQgy2W8pIli8enQPliiLO8VW6XdQF8rkT
   (includes Banach fixed point theorem in complete metric spaces and $C([a,b])$ as a complete normed space)
3. Riemann integrals: https://www.youtube.com/playlist?list=PLQgy2W8pIli9sGfgTURtHzNwItqyiI3Tz
   (the first of my corona videos)
4. Integral equations: https://www.youtube.com/playlist?list=PLQgy2W8pIli9fO_Zc8A-TEN9RknR2ZMmq
5. Differential calculus for power series: https://www.youtube.com/playlist?list=PLQgy2W8pIli_IRJbP205fsUsBTIx1vNEk
   Differential calculus, a shortcut: https://www.youtube.com/playlist?list=PLQgy2W8pIli-mMFVTcMcKO5Zy2uJU193Y
6. Epsilons and deltas: https://www.youtube.com/playlist?list=PLQgy2W8pIli8e9Hm34hqKFtil6pZg4I6z
   includes the mean value theorem and the fundamental theorem of calculus
   with detour examples of continuity https://www.youtube.com/playlist?list=PLQgy2W8pIli9_T5budfPhqXqhnSVel_KM
7. The Mean Value Theorem in integral form: https://www.youtube.com/playlist?list=PLQgy2W8pIli9jyuYN76HM3YdXjwBZrx8L
   for Newton's method, adapted for the inverse function theorem next
8. Solving $f(x) = y$, Inverse Function Theorem: https://www.youtube.com/playlist?list=PLQgy2W8pIli-7124huziMvr6eTmFVOCuh

9abc. Figuring it out, compressed in 8 later, see caption of https://youtu.be/_2ZGd6HmhV4

10. Applications in machine learning: https://www.youtube.com/playlist?list=PLQgy2W8pIli8K9-hnT_cb_NkKGmOmvJbv

youtube hyperlinks and some new text throughout the notes, e.g. first page Chapter 14, begin and end of Section 5.2

index after Chapter 14, some low tex figures in the introduction



Eventuele auteursinkomsten komen ten goede aan de "geluksmachine":

http://orchestra18c.com

## About the videos for the course

Each playlist has a description of the playlist, and a list of the videos in it. The videos have informative names, and descriptions with corrections and links: to the notes, to every previous video and every next video. The first 6 playlists are for the course. Do sneak around in the others too. Whereas the course notes are organised by topic, the videos are organised in part by technique.

The short first playlist corresponds to the introductory Chapter 1 and sets the scene with YBC7289, Archimedes, and the (set of) real numbers. The long second playlist then covers what can be done with $\varepsilon - N$ arguments in the context of convergence of sequences of real numbers, the concept of continuity for functions, and sequences of continuous functions. The number sequences include sequences obtained from the iteration of contractive maps, such as in particular Heron's sequence for $\sqrt{2}$.

Section 4.4 summarises the results on sequences of numbers, continuity of functions, sequences of continuous functions, and spaces of continuous functions. It is followed by the first (and only) completely abstract chapter: Chapter 5 is about abstract metric spaces, includes an outlook on topology, and concludes the first part (period 4) of the course.

The next four playlists 3-6 are about integration and differentiation, Chapters 6-11 in the notes. These constitute part two (period 5) of the course, which starts from square one again: Chapter 6, on the integration of monotone functions, requires little more than Chapter 1.

Chapter 7 then builds on all of the first part of the course. It covers the general theory for Riemann integration and uses the Banach Fixed Point Theorem from Chapter 5 to solve integral equations in spaces of integrable functions.

Only thereafter, Chapter 8 and playlist 6 introduce and exploit $\varepsilon - \delta$ arguments, to establish in particular the integrability of continuous functions and addresses the issue of the existence of convergent subsequences in Section 4.4. For bounded sequences of real numbers this was settled in Chapter 3.

Chapter 9 in turn, on the algebraic approach to differentiation for power series, is largely independent of everything done before, although some of the language from Chapter 4 is used to formulate the results. It introduces the fundamental approach to differentiation with linear approximations and *no limits but* error estimates.

The general theory of differentiation in Chapters 10 and 11 uses the $\varepsilon - \delta$ machinery again, to establish the relation between integration and differentiation, and most of the facts for and from calculus. *The pointwise continuity statement in Theorem 8.1 is used in Section 10.2.*

# 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 pointwise convergent monotone functions[1];

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[2];

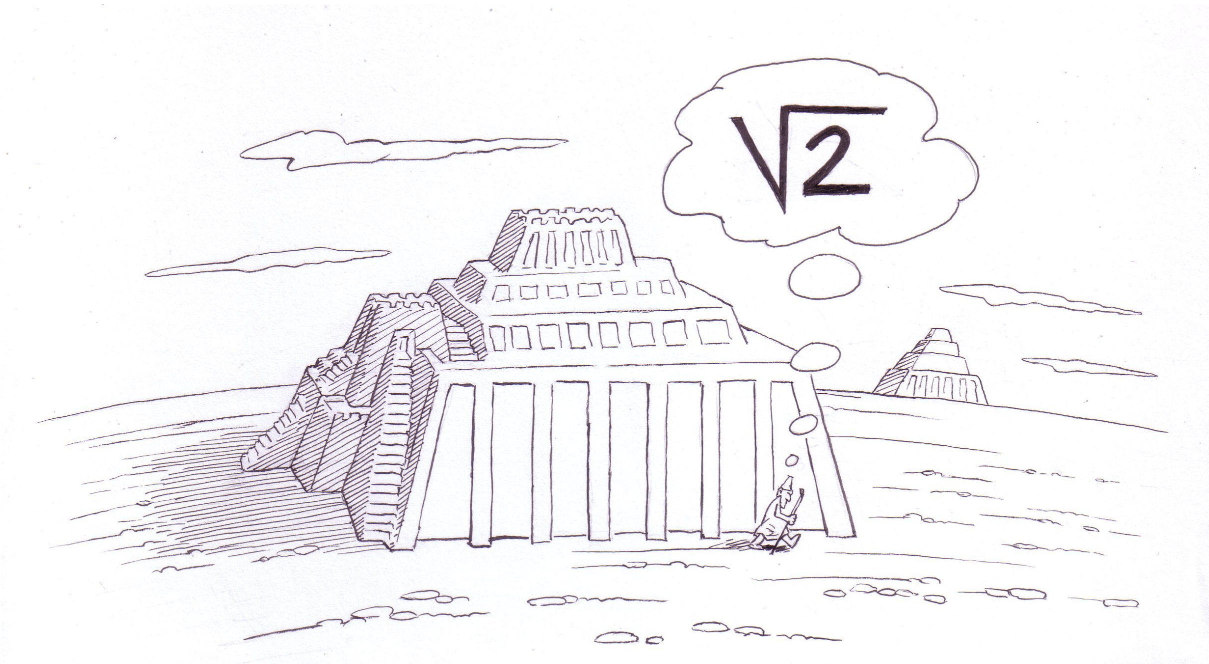
5. ... knows the statement of the Banach Fixed Point Theorem, and can apply this theorem to solve fixed point equations, in particular integral equations in $C([a,b])$ for solutions of differential equations.

**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, in relation to two if you like geometric numbers: $\sqrt{2}$ and $\frac{1}{3}$.

[1]This will be a not too hard part at the start of the second block.
[2]This will be a hard part at the end of the first block.


$\sqrt{2}$

*'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 are lecture notes for a first course in mathematical analysis and what can follow later. In these days of corona I made videos for most of the topics covered. The video `https://youtu.be/taKW6cNomJk` should get you started. Click MEER WEERGEVEN for an overview of the video playlists that I listed on the front page of these notes. To go the next video click

go to `https://youtu.be/4vj5LX_okSA` next

and likewise in every second line of the text under every next video. The link in every first line takes you back to the previous video.

Read the text with every video before you watch it, read the text with every playlist before you play it. I will still be editing them. Every video has a link to the playlist it belongs to. First course 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 once was *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 must have 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 Newton's method, the *scheme*

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

for solving the equation $f(x) = 0$ numerically. Starting with some $x_0$ and $n = 1$ this scheme produces a sequence $x_1, x_2, \ldots,$ which typically converges to a solution of $f(x) = 0$ very fast, see Chapter 12.2.

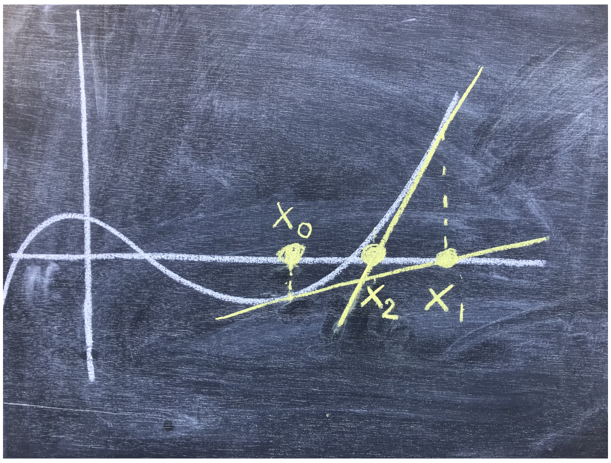



Figure 1: Newton's iterative method for solving $f(x) = 0$ pictured with the graph of $f$. To find the next iterate intersect the line tangent to the graph in the previous iterate with the horizontal axis.

**Exercise 1.1.** Consider the graph defined by $y = f(x)$. Use your highschool maths to write down a formula for the line tangent to the graph of $f$ in the point $(x, y) = (x_{n-1}, f(x_{n-1}))$. 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 $f$ is given by $f(x) = x^2 - 2$.

**Exercise 1.2.** Show for $f(x) = x^2 - 2$ 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

Have a look at `https://youtu.be/DyG1B3WYh-k`.

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.

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

[1] That I know of.

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

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.

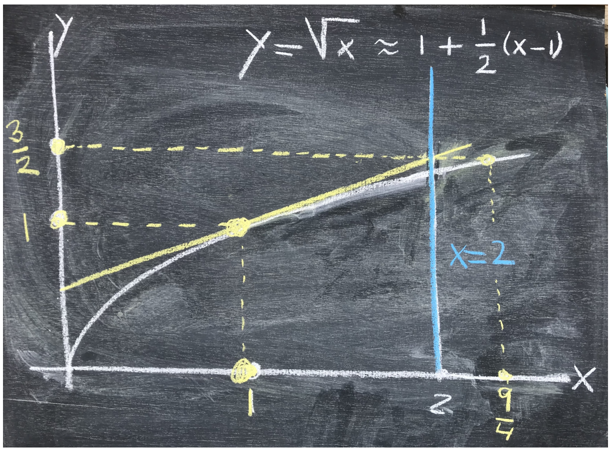



Figure 2: The first step in the Babylonian scheme. The line tangent to $y = \sqrt{x}$ in $(1, 1)$ intersects $x = 2$ in $y = \frac{3}{2}$. To find the next iterate intersect the line tangent in $(\frac{9}{4}, \frac{3}{2})$ with $x = 2$. And so on.

## 1.2 One third of what?

https://www.youtube.com/playlist?list=PLQgy2W8pIli9lL_Yp_4Tr_CGOgrGUBdZa

The playlist linked to in the previous line starts with https://youtu.be/taKW6cNomJk, and concerns another *geometric number*: the number $\frac{1}{3}$ that appears as the *volume $V$ of a pyramid* with unit base area and unit height.

To see how $\frac{1}{3}$ appears we divide the pyramid into say 6 horizontal layers of height $\frac{1}{6}$ and write $n$ for 6. The maximal width of each layer varies from 1 at the bottom to $\frac{1}{6} = \frac{1}{n}$ 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[3]

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

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[4] that therefore the difference of $V$ and $W$ must be zero, and who are we to dispute? As a consequence of what we

[3] This is the first time we use the symbol $<$ but you know what it means.
[4] https://youtu.be/4vj5LX_okSA

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[5]. 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

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

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

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, or domino principle. Think of the $n^{th}$ statement $(C_n)$ as being written on the $n^{th}$ 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 on the dominos still standing.

You may have noted that[6]

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

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[7]

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

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

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

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

[7] Frits Beukers showed me this neat trick.

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.

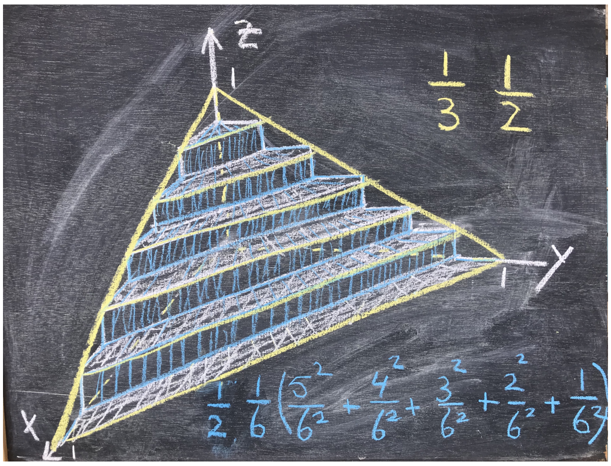



Figure 3: The volume of the tretahedron bounded by $x = 0$, $y = 0$, $z = 0$ and $x + y + z = 1$ in $xyz$-space is equal to the product of $\frac{1}{3}$ and $\frac{1}{2}$. The latter factor is the area of the triangular base, but how does $\frac{1}{3}$ appear? And how would you measure the object bounded by $x_1 = 0$, $x_2 = 0$, $x_3 = 0$, $x_4 = 0$ and $x_1 + x_2 + x_3 + x_4 = 1$ in 4-space?

## 1.3 Number sets and 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

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

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 $\pi$ 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 $1, 2, 3, \ldots$. 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

$$x_1, x_2, x_3, \ldots$$

is an enumeration of $\mathbb{R}$. Then $\mathbb{R}$ is completely covered by the intervals[8]

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

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

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[9] of the real number line should be larger than any positive number. Have we proved the following theorem?

**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[10] $d_{nj}$ in a block

[8]For numbers $a < b$ we denote by $(a, b)$ the set of all real numbers $x$ with $a < x < b$.
[9]Here we touch upon measure theory, see cartoon before Chapter 8 and Section 8.4.
[10]Which can be any of $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$.

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

Now choose $d_n$ with $d_n - d_{nn} = 2$ or $d_{nn} - d_n = 2$. Then 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

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

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 $\exists$, as well as the proverbial epsilon. Enjoy `https://youtu.be/CFGzPqAadEU`.

## 1.4 Enter the epsilons

The theorem below says that there is no positive 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[11].

**Theorem 1.5.** *The Archimedean Principle:*

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

**Exercise 1.6.** Maybe Archimedes thought of his principle as[12]

$$\forall_{\varepsilon>0} \, \underbrace{\exists_{N\in\mathbb{N}} : \frac{1}{N}}_{\text{whatever}} \leq \varepsilon.$$

Looks weaker as a statement but it's not[13], why?

**Exercise 1.7.** Explain why the Archimedean Principle with $\varepsilon = \frac{1}{x}$ and $x \in \mathbb{R}$ positive leads to the equivalent statement

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

Note that for $x \leq 0$ the inequality holds for all $N \in \mathbb{N}$.

**Exercise 1.8.** We used the symbol $N$ to exhibit that the statements in Theorem 1.5 and Exercise 1.7 concern the existence of a single $N$. For which $n$ other than $n = N$ do these Archimedean statements also hold?

## 1.5 The geometric series

See

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

---

[11] Don't we do *important work*?

[12] If he ever did. Note the first use of the symbol $\leq$ and that you know what it means.

[13] We could choose to write all future $\forall_{\varepsilon>0} \cdots < \varepsilon$ statements with $\leq \varepsilon$, but we won't.

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 1.9.** *For $x\in\mathbb{R}$ it holds that*[14]

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

**Exercise 1.10.** 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.

[14]We use the *absolute value* $|x|$ of $x$ for the first time here.

In particular it follows for all $x$ with $|x| < 1$ that[15]

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

which reduces to (1.8) for $x = \frac{1}{n}$, but is a far more general statement[16].

A mathematical proof of Theorem 1.9 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[17]

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

Making such and other limits statements mathematically sound is another important task for this course, but don't ignore the algebraic beauty in (1.10). The temptation to for now leave Archimedes and the epsilons for what they are and jump to Chapter 9 is hard to resist. Very hard.

## 1.6 Exercises

**Exercise 1.11.** 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 1.10: draw graphs to examine how well the partial sums do as approximations.

**Exercise 1.12.** We turn (1.10) around. Show that for every $x \neq 1$ it holds that

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

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

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

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

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[18] you about the line tangent to the graph defined by $y = x^n$ in $x = a$?

**Exercise 1.13.** Look at (1.1). Verify that

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

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

**Exercise 1.14.** 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 1.15.** Find the complete hexigesimal expansion[19] of

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

Hint: use hexigesimal *long division* to write the hexigesimal expansion of $\frac{12}{17}$, which should come out periodic. Then add the finite hexigesimal expansion of $\frac{17}{24}$.

**Exercise 1.16.** 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.

[18] In Chapter 9 this starts an approach to *differentiation* that *avoids the usual limits*.
[19] You need the multiplicative tables in base 60.

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

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

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

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

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.** Referring to Exercise 1.12 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 $n \in \mathbb{N}$?

**Exercise 1.19.** Use the Archimedean Principle in Theorem 1.5 to show that

$$\forall_{\varepsilon>0}\,\exists_{n\in\mathbb{N}} : \frac{2019}{n} \le \varepsilon.$$

**Exercise 1.20.** Show that

$$\forall_{\varepsilon>0}\,\exists_{n\in\mathbb{N}} : \frac{1}{n^2} < \varepsilon; \quad \forall_{\varepsilon>0}\,\exists_{n\in\mathbb{N}} : \frac{1}{\sqrt{n}} < \varepsilon; \quad \forall_{\varepsilon>0}\,\exists_{n\in\mathbb{N}} : \frac{n}{n^2+1} < \varepsilon.$$

**Exercise 1.21.** 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 1.22.** This exercise and (1.14) 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 1.21 to conclude that

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{m,n\in\mathbb{N}} : \quad m \geq n \geq N \implies \sum_{k=n}^{m} \frac{1}{2^k} < \varepsilon. \tag{1.14}$$

**Exercise 1.23.** 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 1.24.** This exercise relates to (1.12). Suppose that $0 < x < 1$. From (1.10) it follows that[20]

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

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

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

**Exercise 1.25.** 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.

**Exercise 1.26.** 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 1.27.** This exercise introduces a happy couple for later. Let $p > 1$ and $q > 1$ be real numbers. Show that

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

[20] Compare to Exercise 1.23.

## 1.7 Compound interest

Think of $x$ in Exercise 1.29 as a fixed and positive annual interest rate on your savings account, $n$ as the number of months in a year, and the bank multiplying your balance by $1+\frac{x}{n}$ every month. Wouldn't you like to have $n \to \infty$? And what if $x$ turns negative?

**Exercise 1.28.** Have a look at

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

and then use the domino principle to show that

$$(1+a)^n = 1 + na + \frac{n(n-1)}{2}a^2 + \frac{n(n-1)(n-2)}{3\cdot 2}a^3 + \cdots$$

for every $n \in \mathbb{N}$ and every $a \in \mathbb{R}$.

**Exercise 1.29.** (continued) Show that

$$\left(1+\frac{x}{n}\right)^n = 1+x+(1-\frac{1}{n})\frac{x^2}{2!}+(1-\frac{1}{n})(1-\frac{2}{n})\frac{x^3}{3!}+(1-\frac{1}{n})(1-\frac{2}{n})(1-\frac{3}{n})\frac{x^4}{4!}+\cdots$$

for every $n \in \mathbb{N}$ and every $x \in \mathbb{R}$.

**Exercise 1.30.** (continued) Write

$$s_n(x) = \left(1+\frac{x}{n}\right)^n$$

and show that $s_{n+1}(x) > s_n(x)$ for $x > 0$. What would you guess for $-n < x < 0$? See if you were right by verifying and using that

$$s_{n+1}(x) - s_n(x) = \left(1+\frac{x}{n+1}\right)^{n+1} - \left(1+\frac{x}{n}\right)^{n+1} + \left(1+\frac{x}{n}\right)^{n+1} - \left(1+\frac{x}{n}\right)^n$$

$$= (\frac{x}{n+1}-\frac{x}{n})\left(\left(1+\frac{x}{n+1}\right)^n + \cdots + \left(1+\frac{x}{n}\right)^n\right) + \left(1+\frac{x}{n}\right)^n\left(1+\frac{x}{n}-1\right)$$

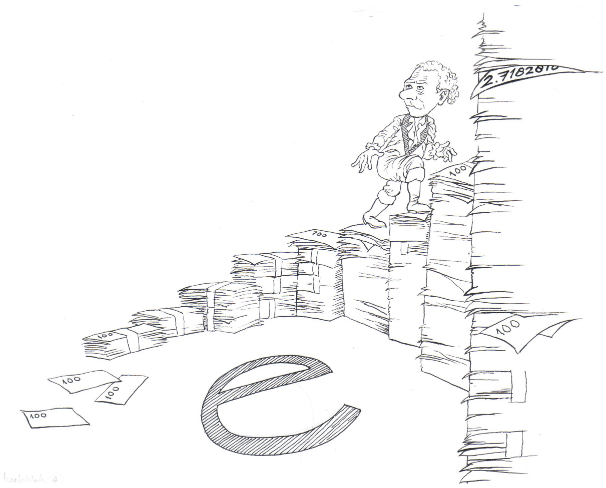
$$= \frac{-x}{n(n+1)}\left(\left(1+\frac{x}{n+1}\right)^n + \cdots + \left(1+\frac{x}{n}\right)^n\right) + \frac{x}{n}\left(1+\frac{x}{n}\right)^n$$

$$= \frac{x}{n(n+1)}\left((n+1)\left(1+\frac{x}{n}\right)^n - \left(1+\frac{x}{n+1}\right)^n - \cdots - \left(1+\frac{x}{n}\right)^n\right).$$

**Exercise 1.31.** Show that the supremum

$$S(x) = \sup_{n \geq -x} \left(1 + \frac{x}{n}\right)^n$$

exists as a positive number for every $x \in \mathbb{R}$ and think of a better name for $S$. Hint: how would $S(x+y)$, $S(x)$ and $S(y)$ be related?




## 1.8 Outlook: norms beyond the real numbers

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

$$\sum_{n=0}^{\infty} A^n = (I - A)^{-1}, \tag{1.15}$$

where $A$ is a square matrix, and in which $A^n$ is a *matrix product*[21], 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 (1.15), we will in fact need a condition of the form $|A| < 1$. A possible "absolute value" of a matrix is

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

[21] Matrix products are explained in Section 18.1.

the square root of the sum of all the squared entries of $A$. This is called the *Frobenius*[22] *norm* of $A$. The Frobenius *norm* has the remarkable properties that

$$|A\,B|_{\mathrm{Frob}} \leq |A|_{\mathrm{Frob}}\,|B|_{\mathrm{Frob}} \quad \text{and} \quad |A+B|_{\mathrm{Frob}} \leq |A|_{\mathrm{Frob}} + |B|_{\mathrm{Frob}} \tag{1.16}$$

for all square[23] 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 *maximum norm* is defined as

$$|f|_{\max} = \max_{a\leq x\leq b} |f(x)|, \tag{1.17}$$

if this maximum exists. For two functions[24] $f$ and $g$ we will have that

$$|fg|_{\max} \leq |f|_{\max}\,|g|_{\max} \quad \text{and} \quad |f+g|_{\max} \leq |f|_{\max} + |g|_{\max}, \tag{1.18}$$

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

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.

## 1.9 An exam question: inverting functions

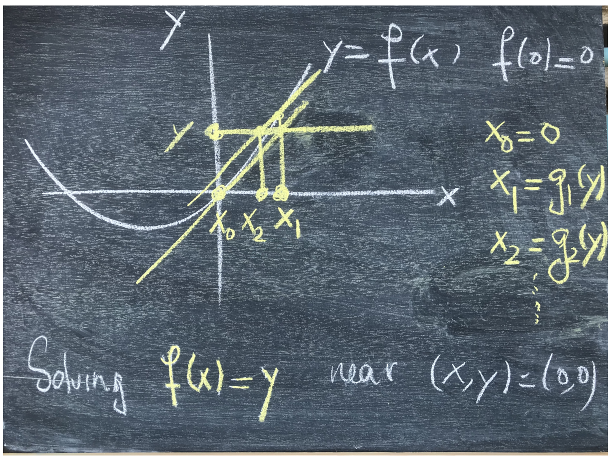



Figure 4: Using lines parallel to the tangent line in the starting point to define functions $g_1, g_2, \ldots$ of $y$. Exam Problem 4 in Section 11.6.

[22]To some dismay of Euclides and Pythagoras perhaps.
[23]In general $AB \neq BA$! In fact the estimates only require $AB$ or $A+B$ to be defined.
[24]With clearly $fg = gf$.

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

https://www.youtube.com/playlist?list=PLQgy2W8pIli8enQPliiLO8VW6XdQF8rkT

Before you watch the above playlist recall https://youtu.be/DyG1B3WYh-k and Exercise 1.2. Heron's scheme 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, a number sequence $x_n$ indexed by $n \in \mathbb{N}$, designed by Heron to solve the equation

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

In time we shall think of such sequences as actually *defining* real numbers.

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 initial value $x_0 = 1$.

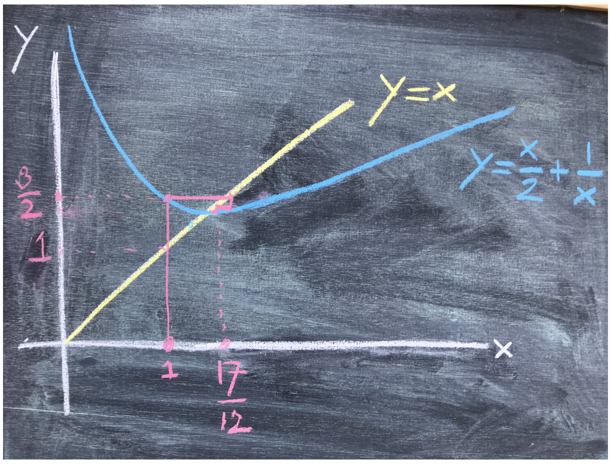



Figure 5: The $xy$-plane with the graph of $f$, the diagonal, and the first few iterates on the axes. Such pictures help you understand what's going on: (fast) convergence to the solution. But we don't need them to continue.

[1]Or the function $f$ if you like, but note we're not solving $f(x) = 0$ here, why?

Note that[2]

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

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.

**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[3]. The reasoning should then be that

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

[2] The trick to remember is that $(a+b)^2 - 4ab$ gives .....

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

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!

This section comes with `https://youtu.be/kqvlUrbURtk`. We saw that Heron's sequence is strictly decreasing and bounded from below. Sequences of numbers[4] $x_n$ with either

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

are called *monotone sequences*, and we shall *first restrict the attention to such monotone sequences*. There are two types of them: 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

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

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.

[4]For the moment rational numbers.

**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[5] 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[6] 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[7] 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 $S$ will turn out to be the unique *limit* of $x_n$. Likewise every nonincreasing bounded sequence has a largest lower bound (*infimum*)

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

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

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

[5] Counting 5 digits in 1.4142.

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

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

- 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*[8]. 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 $L_{\text{Heron}}$, 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.

[8] But see Section 3.8.

## 2.2 The epsilon-limit definition

This section comes with https://youtu.be/phCHnY6lCcI. The defining property of the infimum $L$ of a sequence $x_n$ is that $x_n \geq L$ for all $n \in \mathbb{N}$, and 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

$$\forall_{\varepsilon>0} \underbrace{\exists_{N\in\mathbb{N}} \forall_{n\geq N} \,:\, |x_n - L|}_{\text{See Exercise 1.6!}} < \varepsilon, \tag{2.8}$$

a statement to be pronounced as[9]: 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 $\varepsilon$. For now $d(x, y)$ is only a short hand notation for the distance between *two real numbers* $x$ *and* $y$, which we agree to be equal to

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

the absolute value of $x - y$. Here we use *algebra*[10] with real[11] numbers.

By the way, the statement in (2.8) makes sense for every real $L$ and every real sequence[12], not just for monotone sequences.

**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*

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

*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 $n \to \infty$".*

[9] Equivalent: $\forall_{\varepsilon>0} \exists_{N\in\mathbb{N}} \forall_{n\in\mathbb{N}} : n \geq N \implies |x_n - L| < \varepsilon$.

[10] For now: a human activity with the operations $+, -, \times, /$ and certain algebraic rules.

[11] So it's not really algebra....

[12] It does not matter that $n$ runs from 1 upwards, any other starting integer is fine.

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? Recalling the continuity statement (2.7) 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[13] convenient choice of $\varepsilon>0$, say $\varepsilon=1$, and we obtain[14]

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

[13]See also your proof Proposition 2.9 below.

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

The triangle inequality[15] 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[16] *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[17]. This will give a different value of $N$, say $\tilde{N}$, such that

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

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[18]

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

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[19] important items in this proof. $\square$

It's `https://youtu.be/CCApJK8xrdQ` next.

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

[16]The *subtract and add* the same term trick.

[17]Let's call this the *$M$-trick*.

[18]With a tilde on $N$ to distinguish from the earlier (also arbitrary) choice of $\varepsilon$.

[19]The other one being the trick with $M$.

**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[20] convenient choice of $\varepsilon$ and use the triangle inequality. Don't forget the $n$ with $n < N$.

**Proposition 2.11.** *The limit of a convergent sequence is unique.*

**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[21], the triangle inequality, and the specific choice[22]

$$\varepsilon = \frac{1}{2}|L_1 - L_2| > 0$$

allow you to derive a contradiction.

**Exercise 2.13.** Note that (2.8) is the first occurrence of an absolute[23] 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*[24]

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

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

$$||a| - |b|| \leq |a + b| \leq |a| + |b|, \tag{2.19}$$

which you may like to memorise.

[20] So you don't use the full strength of the definition!

[21] As in the proof of (2.12).

[22] This requires the full strength of the definition!

[23] Recall the *absolute value $|x|$ is also called the norm* of $x$.

[24] A nice statement about the map $x \to |x|$ from $\mathbb{R}$ to $[0, \infty)$.

**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

$$2x_n(x_{n+1} - x_n) = 2 - x_n^2. \tag{2.20}$$

**Exercise 2.18.** Recall that Heron's sequence is convergent. Use this to prove[25] that $x_{n+1} - x_n \to 0$.

**Exercise 2.19.** Prove that it holds for Heron's sequence that $x_n^2 \to 2$.
Hint: combine (2.20) and Exercise 2.18.

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

**Exercise 2.20.** Recall that

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

Prove that $L^2_{\textit{Heron}} = 2$.

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

**Exercise 2.23.** Exercise 2.3 produced a bounded nondecreasing[26] sequence which therefore has a supremum $S$. Prove that $S^2 = 2$. Thus $S = L_{\textit{Heron}} = \sqrt{2}$.

## 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.

**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 $A$.*

[26] Is that sequence strictly increasing?

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*[27] *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 $\infty$ nor $-\infty$ exists, for that matter.*

[27] 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 set* $\mathbb{N}$ *is not bounded from above in* $\mathbb{R}$ *and therefore (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 1.8 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[28] $m > 0$ 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, according to the first statement in the theorem. If it were, then the sequence $x_n = n$ would have a supremum $S \in \mathbb{R}$. With this lowest upper bound $S$ at our disposal[29], we then observe $S - \frac{1}{2}$ is not an

[28] Here $m \in \mathbb{R}$. Alternatively: reason from existence of $\varepsilon > 0$ with $\frac{1}{n} \geq \varepsilon$ for all $n \in \mathbb{N}$.

[29] To dispose of in fact.

upper bound. This means that there exists $n \in \mathbb{N}$ with $n > S - \frac{1}{2}$. Hence[30] the number $n + 1 \in \mathbb{N}$ satisfies $n + 1 > S + \frac{1}{2} > S$ and disqualifies $S$ as the supremum of the sequence $x_n = n$, since $S$ would not even be an upper bound. This completes the proof of Theorem 2.29. In particular we have

$$\inf_{n\in\mathbb{N}} \frac{1}{n} = 0, \tag{2.21}$$

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 (2.21) 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.** *Spelling it out again. 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*[31] *on $\varepsilon$. Observe that*

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

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

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

*and that*

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

*Thus the desired inequality is equivalent to*

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

*which certainly holds for all $n \in \mathbb{N}$ if $\varepsilon \geq 2$. For $\varepsilon < 2$ we invoke the Archimedean Principle again. Observe that it is more convenient to just use the restated form in Exercise 1.7, without making the distinction between $\varepsilon \geq 2$ and $\varepsilon > 0$. This gives the existence of an $N \in \mathbb{N}$ with*

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

*for otherwise the set $\mathbb{N}$ would be bounded from above. But then*[32] *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*[33] .

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

*This proves the claim that*[34]

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

□

## 2.6 Basic limit theorems for 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 $\leq$.*

[32] See the point made by Exercise 1.8.
[33] Was the careful reasoning with inequalities really necessary? See Exercise 1.6.
[34] More of the same in Exercise 2.52.

**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 $n$?

**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

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

**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:** `https://youtu.be/up8ET9go3FI`. 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

$$\underbrace{|(x_n + y_n) - (\bar{x} + \bar{y})|}_{d(x_n + y_n, \bar{x} + \bar{y})} = \underbrace{|(x_n - \bar{x}) + (y_n - \bar{y})|}_{\textit{reshuffled}} \le \underbrace{|x_n - \bar{x}|}_{d(x_n, \bar{x})} + \underbrace{|y_n - \bar{y}|}_{d(y_n, \bar{y})}. \tag{2.22}$$

The equality in (2.22) is a *reshuffle trick*. It uses the algebraic properties of addition and subtraction in $\mathbb{R}$.

With (2.22) we are in position to start up the proof with a default sentence.

$$\text{Let } \varepsilon > 0 \text{ be arbitrary.}$$

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

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

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 $\varepsilon$-*engine* run.

Our initial estimate (2.22) and the two $\varepsilon$-statements establish that

$$\forall_{n \geq N} : \underbrace{|(x_n + y_n) - (x + y)|}_{d(x_n+y_n,\bar{x}+\bar{y})} < \varepsilon + \varepsilon = 2\varepsilon. \tag{2.23}$$

Now we are not completely happy with $2\varepsilon$. 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

$$\tilde{\varepsilon} = \frac{\varepsilon}{2}, \tag{2.24}$$

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 (2.22) 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

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

$\square$

**Remark 2.37.** *In hindsight we might just as well start with (2.22), jump to (2.24) 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 (2.23), and maybe forget about (2.24) and what followed.*

**Proof of the product statement:** `https://youtu.be/up8ET9go3FI`.

The limit of the sequence $x_ny_n$ should be $\bar{x}\bar{y}$, so we have to show that the distance between $x_ny_n$ 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

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

it follows that

$$\underbrace{|x_ny_n - \bar{x}\bar{y}|}_{d(x_ny_n,\bar{x}\bar{y})} \leq \underbrace{|x_n - \bar{x}|}_{d(x_n,\bar{x})} |y_n| + |\bar{x}| \underbrace{|y_n - \bar{y}|}_{d(y_n,\bar{y})}. \tag{2.25}$$

Next we do copy-paste of what's between (2.22) and (2.23) 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 (2.25). We arrive, for *the same* $N \in \mathbb{N}$, at

$$\forall_{n\geq N} : |x_ny_n - \bar{x}\bar{y}| \leq \underbrace{|x_n - \bar{x}|}_{<\varepsilon} |y_n| + |\bar{x}| \underbrace{|y_n - \bar{y}|}_{<\varepsilon}. \tag{2.26}$$

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 (2.26) that

$$\forall_{n\geq N} : |x_ny_n - \bar{x}\bar{y}| < (M + |\bar{x}|)\varepsilon. \tag{2.27}$$

Now we are happy again, because with (2.27) 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 $x_ny_n \to \bar{x}\bar{y}$ and well deserves a remark next. $\square$

**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

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

You may like to sketch the graph defined by

$$y = \frac{1}{x}$$

in the $xy$-plane for what follows. We observe that

$$|x_n - \bar{x}| < \varepsilon \iff x_n \in (\bar{x} - \varepsilon, \bar{x} + \varepsilon) \tag{2.28}$$

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

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

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

$$|x_n| > \frac{1}{2}|\bar{x}| \quad \text{whence} \quad \left|\frac{1}{x_n}\right| < \frac{2}{|\bar{x}|} \tag{2.29}$$

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[35].

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

**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.** Recalling the continuity statement (2.7) you should note here again that Theorem 2.40 is in fact the same statement in every $\bar{x} \neq 0$ for $f$ defined by

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

in $x \neq 0$. Write out a complete proof of Theorem 2.40. And verify that, no matter what value we assign to $f(0)$, this statement is false in $\bar{x} = 0$.

## 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 $\mathbb{R}$.*

**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 because $k(k+1) > k^2$. If you use $k(k+1) < (k+1)^2$ however....

**Exercise 2.45.** Define the sequence $s_n$ by

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

Determine the limit as $n \to \infty$.

**Exercise 2.46.** Same question for

$$\sum_{k=1}^{n} \frac{1}{k(k+3)}.$$

**Exercise 2.47.** Same question for $3$ replaced by $4, 5, 6, \ldots$: find a nice sum formula for the limit of

$$\sum_{k=1}^{n} \frac{1}{k(k+m)}$$

as $n \to \infty$ if $m \in \mathbb{N}$.

**Exercise 2.48.** It's not easy to generalise the *telescope trick* in Exercise 2.44 for

$$S_n = \sum_{k=1}^{n} \frac{1}{k^2}$$

to other exponents in the denominator. Here's that neat trick[36] of Gauss again, for sums of the form

$$S_n = \sum_{k=1}^{n} b_k,$$

with $b_k$ nonnegative and nonincreasing. Play with

$$b_1 + (b_2 + b_3) + (\underbrace{b_4 + b_5 + b_6 + b_7}_{2^2 b_7 \leq \cdots \leq 2^2 b_3}) + (\underbrace{b_8 + b_9 + b_{10} + b_{11} + b_{12} + b_{13} + b_{14} + b_{15}}_{2^3 b_{15} \leq \cdots \leq 3^2 b_8}),$$

and so on. See what you conclude for

$$\sum_{k=1}^{n} \frac{1}{k^p}$$

as $n \to \infty$, depending on $p > 0$.

[36] See also Exercise 1.22.

**Exercise 2.49.** 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 2.50.** Does the sequence

$$\sqrt{2}, \sqrt{2+\sqrt{2}}, \sqrt{2+\sqrt{2+\sqrt{2}}}, \ldots$$

converge?

Hint: figure out[37] what's on the dots and determine a limit if you can.

**Exercise 2.51.** Same question for

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

and then, for both sequences, play with $p \geq 1$ instead of $2$, as we so often do.

**Exercise 2.52.** For each of the following sequences decide on convergence and prove your conclusion directly from Definition 2.6.

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

Determine, without proof, the suprema and infima of these sequences, if they exist.

**Exercise 2.53.** We define the integer part of $x \in \mathbb{R}$ by

$$[x] = \sup\{n \in \mathbb{N} : n \leq x\}.$$

For each of the sequences[38] $x_n$ in Exercise 2.52 decide on convergence of $[x_n]$.

[37] Which $f$ would generate the sequence?

[38] Call them $x_n$.

**Exercise 2.54.** 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 2.55.** 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 2.56.** 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 2.57.** 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 $n \to \infty$, the *quotient rule* for limits of sequences.

**Exercise 2.58.** Let $x_n$ be a convergent sequence of real numbers indexed by $n \in \mathbb{N}$, and let

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

be the sequence of averages. 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 2.59.** For a sequence $x_n$ we can define the sums

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

Find an example of a sequence $x_n \to 0$ with $S_n$ divergent en $\sigma_n$ convergent.

**Exercise 2.60.** For $a \in (0, \frac{1}{4})$ define the sequence $x_n$ by $x_0 = 1$ and

$$x_n = 1 - \frac{a}{x_{n-1}}.$$

a) Use induction to show that $x_n > \frac{1}{2}$ for all $n \in \mathbb{N}$.

b) Use induction to show that $x_n < x_{n-1}$ for all $n \in \mathbb{N}$.

c) Prove that the sequence $x_n$ is convergent. What is its limit?

**Exercise 2.61.** For $a \in (0, \frac{1}{4})$ define the sequence $x_n$ by $x_0 = 0$ and

$$x_n = a + x_{n-1}^2.$$

a) Use induction to show that $x_n < \frac{1}{2}$ for all $n \in \mathbb{N}$.

b) Use induction to show that $x_n > x_{n-1}$ for all $n \in \mathbb{N}$.

c) Prove that the sequence $x_n$ is convergent. What is its limit?

**Exercise 2.62.** Let $P$ be the second degree polynomial given by[39]

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

This is to show that the equation $P(x) = 1$ has a unique solution in $(0, 1)$ without actually solving it. Note that $P(0) = 0$ and $P(1) > 1$. Define the sequence $x_n$ by

$$x_n = x_{n-1} + 1 - P(x_{n-1}) \quad \text{for all} \quad n \in \mathbb{N} \quad \text{and} \quad x_0 = 1.$$

a) Show that in the $xy$-plane the point $(x_n, 1)$ is the intersection of the line $y = 1$ with the line through $(x_{n-1}, P(x_{n-1}))$ with slope $1$.

b) Show that
$$P(x) - P(y) > x - y \quad \text{if} \quad 0 \le y < x \le 1,$$

c) Prove that $x_n$ and $P(x_n)$ are strictly decreasing sequences with $P(x_n) > 1$.

d) Show that
$$\inf_{n \in \mathbb{N}} P(x_n) = 1.$$

[39]This will come back in Exercise 7.79, and the same ideas in Exercise 7.74.

e) Let
$$p = \inf_{n \in \mathbb{N}} x_n \geq 0.$$
Show that $P(p) \leq 1$. Hint: if not then $P(x_n) > P(p) > 1$.

f) Show that
$$P(x) - P(y) < 2(x - y) \quad \text{if} \quad 0 \leq y < x \leq 1.$$

g) Show that $P(p) = 1$. Hint: if not then $P(p) < 1$, use $P(x_n) < P(p) + 2(x_n - p)$.

h) Why is $x = p$ the *unique* solution of $P(x) = 1$ in $(0, 1)$?

**Exercise 2.63.** Show that
$$x + \frac{x^2}{2} - \frac{x^3}{6} - \frac{x^4}{12} = 1$$
has a unique solution in the interval $(0, 1)$.

Hint: get your inspiration from Exercise 2.62.

**Exercise 2.64.** Show that
$$x + \frac{x^2}{2!} - \frac{x^3}{3!} - \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} = 1$$
and
$$x + \frac{x^2}{2!} - \frac{x^3}{3!} - \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} - \frac{x^7}{7!} - \frac{x^8}{8!} = 1$$
have unique solutions in the interval $(0, 1)$.

**Exercise 2.65.** 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 2.66.** 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 2.67.** Same question for the suprema and infima of[40] $A+B$ and $A-B$ in terms of $\sup A$, $\sup B$, $\inf A$ and $\inf B$.

**Exercise 2.68.** Let $I_n = [a_n, b_n]$ be a sequence of closed intervals in $\mathbb{R}$ with $I_{n+1} \subset I_n$ for all $n \in \mathbb{N}$. Prove there exists $c \in \mathbb{R}$ such that $c \in I_n$ for every $n \in \mathbb{N}$.

## 2.8 Cliffhanger: limits and limit points

Recall that Remark 2.7 and (2.10) said convergence of $x_n \in \mathbb{R}$ means

$$\underbrace{\exists_{\bar{x}\in\mathbb{R}}}_{\text{limit exists}} \ \forall_{\varepsilon>0}\,\exists_{N\in\mathbb{N}}\,\forall_{n\geq N} \ : \ |x_n - \bar{x}| < \varepsilon. \tag{2.30}$$

Clearly one issue remains: a reformulation of (2.30) that does not involve the limit $\bar{x}$, which Proposition 2.11 said was in fact unique if (2.30) holds true, while Proposition 2.9 said that any such convergent sequence is bounded.

It is clearly not true that every bounded sequence is convergent, as the simple counterexample

$$x_n = (-1)^n$$

shows. The sequence

$$x_n = (-1)^n + \frac{1}{n}$$

provides another such example, but do note that $x_{2n} \to 1$ and $x_{2n-1} \to -1$, so 1 and $-1$ do seem to appear as limits. Since they are not, not of the sequence $x_n$ that is, we think of them as *limit points* instead.

We now announce a fundamental theorem that says that

every bounded sequence in $\mathbb{R}$ has a limit point[41],

and we will use this very limit point theorem to prove that a certain French engineer was right in proposing that a sequence $x_n$ should be convergent if and only if

$$\forall_{\varepsilon>0}\,\exists_{N\in\mathbb{N}}\,\forall_{m,n\geq N} \ : \ |x_n - x_m| < \varepsilon. \tag{2.31}$$

That is to say, if (2.31) holds, then

$$\forall_{\varepsilon>0}\,\exists_{N\in\mathbb{N}}\,\forall_{n\geq N} \ : \ |x_n - \bar{x}| < \varepsilon \tag{2.32}$$

holds, and $\bar{x} \in \mathbb{R}$ is then the unique real number for which this is the case.

[40] $A+B = \{a+b : \, a\in A, b\in B\}$, $A-B = \{a-b : \, a\in A, b\in B\}$.

[41] A point that satisfies $\forall_{\varepsilon>0}\,\forall_{N\in\mathbb{N}}\,\exists_{n\geq N} \ : \ |x_n - \bar{x}| < \varepsilon$, spot the difference with (2.32).

# 3 Contractions and non-monotone sequences

Until Section 3.3 this corresponds to `https://youtu.be/RgzhfiwrCkk` in the epsilon-N playlist. We present another approach to conclude that Heron's sequence has a limit, *the story of the first part of this course really*. 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 or steps can be used to reproduce $x_n$ from $x_0$ because

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

In $n$ steps[1] we get from $x_0$ to $x_n$. 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[2].

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 1.5 that Zeno never liked so much. 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 3.9, see Theorem 3.73.

## 3.1 Estimates for the increments

*This is the beginning of a main story line that will come back*, see e.g. (11.32) in Problem 4 of the 2020 exam. The scheme there is easier to analyse, but it contains a parameter, which by itself makes things more complicated[3]. Here we continue with Heron's scheme and use some algebra to estimate every $d(x_{n+1}, x_n) = |x_{n+1} - x_n| = |\xi_{n+1}|$ in terms of the previous distance

[1] We denote the steps by $\xi_1, \xi_2, \ldots$ here, in the exam by $s_1, s_2, \ldots$, sorry.
[2] By the way, variation of parameters helps in solving equations, see Exercise 3.35.
[3] Did you do Exercises 2.60 and 2.61?

$|x_n - x_{n-1}|$. This will get us started towards Theorem 5.14 at the end of this story, which will rely on an abstract version of Definition 3.3 below. Here we go.

**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

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

and that therefore

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

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

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

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[4]

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

and so on. It follows that

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

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

**Exercise 3.2.** Let the map[5] $f$ be defined by

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

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

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

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 $x_0 > 0$.

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*[6] *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* nonexpansive *if $L = 1$.*

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

**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:

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

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

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

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

Using (3.4) it follows that

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

Now go back to (1.14) and what we spelled out in Exercises 1.21 and 1.22 with the observation that

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

It follows that

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

in which the $\varepsilon$-estimate holds for all $m, n$ with $m > n \geq N$, provided $N$ is as in Exercise 1.21. We conclude that for all $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that[7]

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

which brings us to a crucial section next.

## 3.3 Cauchy sequences, monotone subsequences

https://youtu.be/T6pSZcV3Oqk and https://youtu.be/REITj_IigqE

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

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

[7] Also for $m = n$.

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 $m \geq N$ and $n \geq N$ the distance between $x_n$ and $x_m$ is smaller than $\varepsilon$.

**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, or equivalently*[8] *if*

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

*If so we say that $d(x_n, x_m) \to 0$ if $m, n \to \infty$.*

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?

**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*[9]*. In other words: every sequence has a monotone subsequence.*

[8]See Exercise 1.6 again!

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

**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

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

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 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 $n \in \mathbb{N}$.*

**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 $\theta^n$ is decreasing[10].

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

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

for $m > n \geq N$.

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

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

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

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 $n \geq N$. We then conclude from (3.11) that[11]

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

for all $n \geq N$.

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 $x \neq y$ we would have that

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

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

**Remark 3.19.** *You should carefully note that*

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

*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 is crucial for proving theorems[12] about continuous[13] functions later on.

[11] Note that with $n \geq N$ also $n + 1 \geq N$.

[12] Like the integral $\int_a^b f$ having a meaning for $f : [a, b] \to \mathbb{R}$ continuous.

[13] We used this term in relation to (2.7).

**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 the very end of Section 3.8. 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. Do note the mind boggling difference with

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

and $x_n \to \bar{x}$ stipulated next.

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

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

*This looks deceptively similar*[14] *to the convergence statement (2.10).*

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

**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.

[14] Both statements have an equivalent version with $\leq$ of course, see Exercise 1.6.

## 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[15] over to[16] 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*[17]

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

*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 $\xi$ 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[18] $N \in \mathbb{N}$ 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. $\square$

**Exercise 3.29.** Prove the opposite implication in Theorem 3.28: if such a sequence exists then its limit $\xi$ 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.*

[15] Only Theorem 3.20 *will not generalise* to the metric space context in Chapter 5.
[16] With $\mathbb{R}$ replaced by $X$, $X$ and $d$ as in definition 5.6.
[17] See Exercise 1.6, what's the obvious equivalent statement?
[18] This uses the Archimedean Principle in the form of Exercise 1.7 again.

**Proof**. Suppose $\xi$ 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. $\square$

**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*[19]

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

*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 $int(\mathcal{O}) = \mathcal{O}$.*

**Theorem 3.33.** *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.*

**Remark 3.34.** *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.*

## 3.7 Exercises

**Exercise 3.35.** 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 $y^2z$ and $yz^2$. Compare what you get to the obvious "solution" $q = x^3 + x$ for the parameter $q$.

**Exercise 3.36.** 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

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

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

**Exercise 3.37.** For each of the functions in Exercise 2.49 find a closed subset $A \subset \mathbb{R}$ such that $f : A \to A$ is a contraction.

**Exercise 3.38.** Which of the following sequences is a Cauchy sequence? Prove your conclusion directly from Definition 3.7.

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

**Exercise 3.39.** Let $x_n$ and $y_n$ be Cauchy sequences in $\mathbb{R}$. Prove that $x_n y_n$ is a Cauchy sequence.

**Exercise 3.40.** Let $x_n$ and $y_n$ be Cauchy sequences in $\mathbb{R}$. Assume that $y_n > 0$ for all $n \in \mathbb{N}$. Does it follow that

$$\frac{x_n}{y_n}$$

is a Cauchy sequence? Same question if $y_n > 1$ for all $n \in \mathbb{N}$. In both cases, give a proof or a counter example.

**Exercise 3.41.** Determine all limit points of the sequences defined by

$$x_n = (-1)^n, \quad x_n = (-1)^n + \frac{1}{n}, \quad x_n = (-1)^n + (-1)^{2n}.$$

**Exercise 3.42.** 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 3.43.** 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 3.44.** For $x \in \mathbb{R}$ with $x > 0$ let

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

a) Show that $f : [2, \infty) \to [2, \infty)$ is a contraction.

b) Define the sequence $x_n$ by $x_0 = 1$ and

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

Why is this sequence convergent? What is its limit?

**Exercise 3.45.** 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 3.46.** Same question for $0 < a < 1$.

**Exercise 3.47.** For $x \in \mathbb{R}$ let

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

and define the sequence $x_n$ by $x_0 = -\frac{1}{5}$ and

$$x_n = f(x_{n-1}).$$

a) Prove that $x_n \in [\frac{1}{4}, \frac{3}{4}]$ for all $n \in \mathbb{N}$.

b) Prove that $|x_{n+1} - x_n| \leq \frac{1}{2}|x_n - x_{n-1}|$ for all $n \geq 2$.

c) Let $N \geq 2$. Prove that

$$|x_n - x_m| < \frac{1}{2^{N-1}}$$

for all $m, n \geq N$.

d) Why is the sequence $x_n$ convergent? What is its limit?

**Exercise 3.48.** 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 3.49.** Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

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

Use your tricks from Exercise 3.48 to prove that $f$ is a contraction.

**Exercise 3.50.** Which of the functions defined by the following formulas is Lipschitz continuous on $[0,1]$? And on $[1,\infty)$? And on $[0,\infty)$?

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

**Exercise 3.51.** Prove that $B_\delta(x_0)$ in Exercise 3.31 is itself an open subset of $\mathbb{R}$.

Hint: use the *triangle inequality*.

**Exercise 3.52.** Let $a, b \in \mathbb{R}$ with $a < b$. Prove that the closed interval $[a, b]$ is a closed subset of $\mathbb{R}$.

**Exercise 3.53.** Let $a, b \in \mathbb{R}$ with $a < b$. Prove that the open interval $(a, b)$ is an open subset of $\mathbb{R}$.

**Exercise 3.54.** 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 3.55.** Prove Theorem 3.33.

**Exercise 3.56.** Let $A$ and $B$ be closed subsets of $\mathbb{R}$. Prove that $A \cup B$ and $A \cap B$ are closed.

**Exercise 3.57.** 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. Formulate and prove similar statements for open subsets.

**Exercise 3.58.** 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 3.59.** 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[20]. Prove that $C$ is a closed uncountable set with empty interior, and that for two such numbers

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

Hint: construct $C$ from nested closed sets $C_n$ of such numbers with $t_n \in \{0, 1, 2\}$.

**Exercise 3.60.** (*continued*) The representation of numbers in $C_n$ is not unique but in $C$ it is. The set $C$ is called *Cantor's discontinuum*. Describe $D = \{x \in [0, 1] : x \notin C\}$ as a countable disjoint union of open intervals indexed by a *binary tree*.

## 3.8 From the rational numbers to the real numbers

**Exercise 3.61.** Let $x_n \in \mathbb{Q}$ be the Heron sequence defined in Exercise 1.2 and let $y_n$ be the sequence defined by

$$y_n = \frac{y_{n-1}}{2} + \frac{1}{y_{n-1}} \quad \text{and} \quad y_0 = 2.$$

Prove that $y_n \in \mathbb{Q}$ is also a Cauchy sequence of rationals and that $|x_n - y_n| \to 0$ as $n \to \infty$. In particular the $\varepsilon$-statements[21] hold for all $\varepsilon$ of the form $\varepsilon = \frac{1}{k}$.

In everyday practice real numbers are represented by Cauchy sequences of rational numbers, e.g. as in Section 1.3. In the following exercises we first develop the algebra for such rational Cauchy sequences, taking the natural nonuniqueness[22] of such real number representing Cauchy sequences properly into account.

**Exercise 3.62.** Let $q_n \in \mathbb{Q}$ be a Cauchy sequence in the sense that

$$\forall_{k \in \mathbb{N}} \, \exists_{N \in \mathbb{N}} \, \forall_{m,n \geq N} \, : \, |q_n - q_m| < \frac{1}{k},$$

[20]Unlike in the context of (1.6), when expansions ending in only zero's were excluded.
[21]This is the true way of analysis, see that book.
[22]Not all populations in our universe can be expected to use base 10 for arithmetic.

and let $p_n \in \mathbb{Q}$ be a another such Cauchy sequence. Prove[23] that $p_n + q_n$, $p_n - q_n$, and $p_n q_n$ are also[24] Cauchy in this sense.

**Exercise 3.63.** When should we think of two such Cauchy sequences of rationals as being the same in relation to their task? Answer[25]: let $q_n \in \mathbb{Q}$ and $s_n \in \mathbb{Q}$ be *any* rational sequences. We think of $q_n$ and $s_n$ as being the same for our purposes and write $q_n \sim s_n$ if[26]

$$\forall_{k \in \mathbb{N}} \, \exists_{N \in \mathbb{N}} \, \forall_{n \geq N} \; : \; |q_n - s_n| < \frac{1}{k}.$$

Clearly $q_n \sim q_n$ for every sequence and $q_n \sim s_n$ if and only if $s_n \sim q_n$. Let $r_n \in \mathbb{Q}$ be another rational sequence and assume that $r_n \sim q_n$ and $r_n \sim s_n$. Prove that $q_n \sim s_n$.

**Exercise 3.64.** Let $q_n \in \mathbb{Q}$ and $r_n \in \mathbb{Q}$ be sequences, and assume $q_n$ is Cauchy in the sense of the $k$-statement in Exercise 3.62. Prove that so is $r_n$.

**Exercise 3.65.** Let $p_n, q_n, r_n, s_n \in \mathbb{Q}$ be Cauchy sequences with $p_n \sim q_n$ and $r_n \sim s_n$. Continue from Exercise 3.62 to show for the Cauchy sequences $p_n + r_n$, $q_n + s_n$, $p_n r_n$, $q_n s_n$ that $p_n + r_n \sim q_n + s_n$ and $p_n r_n \sim q_n s_n$.

**Exercise 3.66.** So we're fine with rational Cauchy sequences for adding and multiplying. Let $p_n, q_n \in \mathbb{Q}$ be such Cauchy sequences. Prove that $|p_n - q_n|$ is a Cauchy sequence. Show that $p_n \sim q_n$ if and only if $|p_n - q_n| \sim 0$, the Cauchy sequence $0, 0, \ldots$ encountered in Exercise 3.62. For which rational Cauchy sequences $q_n$ is $\frac{1}{q_n}$ also a rational Cauchy sequence? And what can you then say about $\frac{1}{p_n}$ if $p_n \sim q_n$?

**Exercise 3.67.** Now that we have a meaningful[27] algebra for Cauchy sequences of rationals we turn to analysis and say that a rational sequence $q^l$ indexed[28] by supercript

[23] See Exercise 3.39.

[24] In particular $p_n - p_n = 0$ is a Cauchy sequence, and so is every $q \in \mathbb{Q}$.

[25] Why is this the answer? Exercise 3.65.

[26] Beware this notation is also used with a quite *different* meaning, see (13.13).

[27] Whatever some people may think about meaning....

[28] A change of notation to sneakily prepare for dropping $l$ from the notation.

$l$ is *easily $k$-controled* if

$$\exists_{L\in\mathbb{N}}\forall_{l\geq L}\ :\ |q^l| < \frac{1}{k+1}.$$

Prove a rational sequence $q^l$ is easily $k$-controled for every $k \in \mathbb{N}$ if and only if $q_l \sim 0$.

**Definition 3.68.** *Let $q_n^l$ be a sequence of rational Cauchy sequences, with subscript $n$ numbering the sequences, and superscript $l$ numbering the rationals in every such Cauchy sequence. Dropping $l$ from the notation we say that $q_n \to 0$ as $n \to \infty$ if for every $k \in \mathbb{N}$ there exists $N$ such that the Cauchy sequence $q_n$ is easily $k$-controled for every $n \geq N$. Likewise we say that $|q_n - q_m| \to 0$ as $m, n \to \infty$ if for every $k \in \mathbb{N}$ there exists $N$ such that the Cauchy sequence $q_n - q_m$ is easily $k$-controled for all $m, n \geq N$.*

**Theorem 3.69.** *Let $q_n$ be a sequence of rational Cauchy sequences with $|q_n - q_m| \to 0$ as $m, n \to \infty$. Then there exists a rational Cauchy sequence $r$ with $q_n - r \to 0$ as $n \to \infty$.*

**Exercise 3.70.** Prove the theorem. Hint: For each $n$ choose a number $r_n \in \mathbb{Q}$ such that $q_n - r_n$ is easily $n$-controled. Then show that $r_n$ is a Cauchy sequence which does the desired job.

**Exercise 3.71.** Reflect on the upshot: we don't get anything new by considering Cauchy sequences of Cauchy sequences of rationals. We thus may recognise $\mathbb{R}$ as simply being the set of all equivalence classes of rational Cauchy sequences. Then we also introduce the *real* distance between two such Cauchy sequences, to quickly forget about the above notion of easy $k$-control. You may bring back the epsilons now, or restrict to $\varepsilon = \frac{1}{k}$, as in the book "The Way of Analysis".

We complete this section with another diagonal argument. Here is the standard proof[29] of Theorem 3.20. 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

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

[29] Similar arguments will be used in the proof of the Arzelà-Ascoli Theorem.

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[30] 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 3.72.** Finish this proof of Theorem 3.20.

Hints: $a_k \le x_{n_{kk}} \le b_k$, the sequences $a_k, b_k$ are monotone, $b_k - a_k = 2^{-k}$.

## 3.9 Absolute and unconditional convergence

Referring to (1.7) the expression

$$\sum_{k=1}^{\infty} x_k = x_1 + x_2 + \cdots$$

is called a series. Another example of such a series is

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots, \tag{3.14}$$

and calls for some debate which will be concluded in Section 6.6.

**Theorem 3.73.** *Let $x_n$ be a sequence of real numbers indexed by $n \in \mathbb{N}$. Suppose that*

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

[30] A *diagonal argument* was also used in the proof of Theorem 1.4.

*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[31] . □

**Exercise 3.74.** Prove the convergence of both sequences $M_n$ and $S_n$.

Hint: for $m, n \in \mathbb{N}$ with $m < n$ use

$$|S_n - S_m| = \left| \sum_{k=m+1}^{n} x_k \right| \leq \sum_{k=m+1}^{n} |x_k| = M_n - M_m.$$

**Exercise 3.75.** (*continued*) Show that the limit $S$ in Theorem3.73 satisfies

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

**Remark 3.76.** *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*

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

[31] See Exercise 5.26 for a more general statement about absolutely convergent series.

*It may of course happen that the sequence $M_n$ is not bounded. Then (3.15) has no meaning but (3.16) may still hold for some number $S \in \mathbb{R}$. In that case we say that the series is* convergent with sum *$S$, but* not absolutely convergent.

**Exercise 3.77.** Think about (3.14) and show that

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

defines a real number[32].

Hint: look at partial sums with even and odd numbers 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 a bounded sequence.** 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.

[32]Which number? See Exercise 6.26 and further.

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 $\phi$ 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 3.78.** Show that $S_n^\phi$ converges to the same sum $S \in \mathbb{R}$ if $M_n$ is bounded.

**Exercise 3.79.** The series in Exercise 3.77 has sum $\ln 2$. Show that the bijection $\phi : \mathbb{N} \to \mathbb{N}$ can be chosen to make the sequence $S_n^\phi$ converge to zero.

# 4 Continuous functions

This chapter is about functions $f : [a, b] \to \mathbb{R}$ which have the property that $f$ is *continuous*[1] *in every point* of $[a, b]$, and the *space*[2] $C([a, b])$ of all such functions. Here $[a, b]$ is a given bounded closed interval with $a < b$. Our tools will be

sequences of real numbers;
the equivalent[3] definitions of convergent and Cauchy sequences;
the elementary properties of convergent sequences;
the Bolzano-Weierstrass Theorem;
the suprema and infima of bounded sets of real numbers[4].

The existence of convergent subsequences of bounded sequences in $\mathbb{R}$ will be needed for the proof that (1.17) is indeed a proper definition of the "absolute value" of a function $f \in C([a, b])$ in Definition 4.5 below.

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 $\xi$ 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 $\xi$, then it is still possible that there exists $L \in \mathbb{R}$ such that*

$$f(x_n) \to L$$

[1]Continuity was mentioned in (2.7), Definition 4.1 below should not come unexpected.
[2]We call it a space and not just a set because of Theorem 4.7 below.
[3]In hindsight.
[4]See Section 2.4.

*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 $\xi$ is an accumulation point of $A$, and there's no need to assume that $\xi \in A$.*

## 4.1 Extrema and the maximum norm

This is `https://youtu.be/OhCBZUmSZGQ` in the playlist. 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* maximiser *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* minimiser *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 upper 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 from above 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 monotone convergent subsequence $x_{n_k}$ because of Theorem 3.10 and 3.20. The limit $\bar{x}$ is in $A$ because $A$ is closed[5]. 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 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 maximiser[6].

The argument for the global minimum is similar. This completes the proof of Theorem 4.4. $\square$

**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 for every $f \in C([a,b])$ the number*

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

*the maximum norm of $f$. This norm is to the function $f$ what the absolute value $|x|$ is to the real number $x$.*

**Exercise 4.6.** Let $f \in C([a,b])$ and $\varepsilon > 0$. Explain very carefully why[7]

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

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

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

**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*

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

[5] You showed this for $A = [a,b]$ in Exercise 3.52.

[6] Which need not be unique, see Exercise 4.37.

[7] Note that $|f|_{\mathsf{max}} \leq \varepsilon \iff \forall_{x \in [a,b]} : |f(x)| \leq \varepsilon$ is easier to prove.

**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 $\xi$ 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 $\xi$. 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 $\xi$. Moreover, the argument is valid for every $\xi \in [a,b]$. Thus $f+g, fg \in C([a,b])$.

Finally, let $\bar{x}, \bar{y}, \bar{z}, \bar{w} \in [a,b]$ be the maximisers for the continuous[8] 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

This is `https://youtu.be/J_KdlZ0beQI` in the playlist.

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

$$d(f,g) = |f-g|_{max} \tag{4.1}$$

*is called the* uniform distance *between $f$ and $g$. It is defined as the maximum norm of the difference of the functions $f$ and $g$, just as the distance between two real numbers $x$ and $y$ is defined as the absolute value of $x-y$.*

**Definition 4.9.** *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*

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

[8] See the hint in Exercise 4.6.

**Exercise 4.10.** Why do we need to assume that $f_n$ is in $C([a,b])$ in Definition 4.9?

**Exercise 4.11.** 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 4.12.** Show that there are bounded sequences in $C([0,1])$ which do not have any uniformly convergent subsequence.

Hint: $f_n(x) = x^n$. Argue by contradiction[9] with the pointwise limit.

**Proposition 4.13.** *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 4.14.** Prove Proposition 4.13. 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 4.15.** *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, $f \in C([a,b])$.*

[9] In Exercise 4.42 you use your calculus abilities for a more direct proof.

**Exercise 4.16.** Formulate and prove the counterpart for sequences in $C([a,b])$ of the other implication in Theorem 3.9.

**Proof of Theorem 4.15**. 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 $\varepsilon$. 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[10]. 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 $\xi$.

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[11] the proof.

In fact the continuity of $f$ is consequence of the statement in Theorem 4.17 below. With a proof of Theorem 4.17 the proof of Theorem 4.15 will thus be complete.

**Theorem 4.17.** *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 $C([a,b])$.*

[10] This part of the reasoning does not use that $f_n \in C([a,b])$!
[11] Explain how. NB in the limit $< \varepsilon$ became $\leq \varepsilon$.

**Proof of Theorem 4.17**. Let $\xi \in [a,b]$. To prove that $f$ is continuous in $\xi$ let $x_k$ be a sequence converging to $\xi$. 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

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

We indicated with underbraces that (4.9) can be applied to two of the terms. The inequalities hold for all $n \ge N$.

In particular it follows with $n = N$ that

$$|f(x_k) - f(\xi)| \le 2\varepsilon + |f_N(x_k) - f_N(\xi)| \tag{4.10}$$

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 $\xi$, 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

$$|f(x_k) - f(\xi)| \le 2\varepsilon + \underbrace{|f_N(x_k) - f_N(\xi)|}_{<\varepsilon} < 3\varepsilon$$

for all $k \ge K$.

Remark 2.38 with $M = 3$ now tells us that the proof of Theorem 4.17 is complete. Thus the proof of Theorem 4.15 is also complete. We don't forget to record the property of sequences formulated by Theorem 4.17 in a definition for functions that are not necessarily continuous. $\square$

**Definition 4.18.** *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*[12], *if*

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\ge N}\, \forall_{x\in[a,b]} \, : \, |f_n(x) - f(x)| < \varepsilon.$$

Theorem 4.17 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.

[12] We prefer with Thomas to write the definition with $< \varepsilon$.

**Exercise 4.19.** 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 $\xi \in [a,b]$. Adapt the proof of Theorem 4.17 to prove that if the functions $f_n$ are all continuous in $\xi$, then so is $f$.

## 4.3 Exercises

**Exercise 4.20.** Prove that in Definition 4.1 it is sufficient to verify the condition for monotone sequences $x_n \to \xi$.

**Exercise 4.21.** Let $f : [0,1] \to [0,1]$ be defined by $f(x) = \sqrt{x}$. Prove directly from Definition 4.1 that $f$ is continuous.

Hint: you may prefer to work with monotone sequences in Definition 4.1.

**Exercise 4.22.** 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 4.23.** Same question for $f$ defined by

$$f(x) = \frac{x}{1+x^2}, \quad f(x) = \begin{cases} 1+x^3 & \text{for } x > 0 \\ 1-x^2 & \text{for } x \leq 0 \end{cases}, \quad f(x) = \begin{cases} \frac{\sqrt{1+x}-1}{x} & \text{for } x > 0 \\ \frac{1}{2} - x^3 & \text{for } x \leq 0 \end{cases},$$

the last one requires a little calculus trick.

**Exercise 4.24.** In which points is the (Dirichlet) function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 1 & \text{for } x \notin \mathbb{Q} \\ 0 & \text{for } x \in \mathbb{Q} \end{cases}$$

continuous?

**Exercise 4.25.** 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 4.26.** 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 4.27.** 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 4.28.** 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 $\xi$ is an accumulation point of $A$.

**Exercise 4.29.** 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 $\xi$ if and only if

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

*exists* and is *equal to* $f(\xi)$, which by assumption *also exists*[13].

[13]Three statements.

**Exercise 4.30.** Let $A$ be a subset of $\mathbb{R}$, let $f : A \to \mathbb{R}$ and assume $\xi \in A$ is *not* an accumulation point of $A$. Why does Definition 4.1 imply $f$ is continuous in $\xi$?

**Exercise 4.31.** Let $I \subset \mathbb{R}$ be an nonempty open interval, let $f : I \to \mathbb{R}$, and $\xi$ 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 4.32.** 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 $\xi$ in $I$, and that $f(\xi^-) \leq f(\xi^+)$.

**Exercise 4.33.** (*continued*) Prove that

$$\{\xi \in I : \, f(\xi^-) < f(\xi^+)\}$$

is finite or countable.

Hint: consider open subintervals $(a, b) \subset I$ first.

**Exercise 4.34.** For $x \in \mathbb{R}$ with $x > 0$ let

$$f(x) = \frac{1}{x^2} + \frac{1}{x} + x^2 \quad \text{and let} \quad m = \inf\{f(x) : \, x > 0\}.$$

a) Prove there exists a sequence $x_n > 0$ with $f(x_n) \to m$ as $n \to \infty$ and show that $m \leq 3$.

b) Prove that $m$ is a global positive minimum of $f$.

**Exercise 4.35.** 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: apply[14] 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 4.36.** 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 4.37.** Let $C$ be the Cantor set in Exercise 3.59 and let $f : C \to \mathbb{R}$ defined by $f(x) = x(1-x)$. Explain why $f$ has a global maximum, then find its maximisers.

**Exercise 4.38.** 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 all $x \in \mathbb{R}$. Is the convergence uniform on $\mathbb{R}$? And on $[0, \infty)$? And on $[0, 1]$?

**Exercise 4.39.** Same question as in Exercise 4.38, but with

$$F(x) = \frac{|x|}{1+|x|}.$$

**Exercise 4.40.** Same question as in Exercise 4.38 and Exercise 4.39, but with

$$f_n(x) = F\left(\frac{x}{n}\right).$$

[14] Don't try to compute the minimiser.

**Exercise 4.41.** For $n \in \mathbb{N}$ define $f_n : \mathbb{R} \to \mathbb{R}$ by

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

a) Let $x \in \mathbb{R}$. Prove that $f_n(x) \to 1$ as $n \to \infty$.

b) Is the sequence $f_n$ uniformly convergent on $[0, 1]$?

**Exercise 4.42.** Let the sequence of functions $f_n : [0, 1] \to [0, 1]$ be defined by

$$f_n(x) = x^n$$

for all $x \in [0, 1]$.

a) Let $m > n$ and

$$x_{mn} = \left(\frac{n}{m}\right)^{\frac{1}{m-n}}.$$

Explain why[15]

$$d(f_m, f_n) \geq f_n(x_{mn}) - f_m(x_{mn}).$$

b) Show that $f_n$ is not a Cauchy sequence in $C([0, 1])$. Hint: consider $d(f_{2n}, f_n)$.

c) Verify that the pointwise limit function $f$ exists but is not continuous in $x = 1$.

**Exercise 4.43.** 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 3.60.

Hint: take the values on these intervals to be fractions with denominators equal to a power of 2.

**Exercise 4.44.** 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

$$H_q(x) = \begin{cases} 0 & \text{for } x < q \\ 1 & \text{for } x \geq q \end{cases}$$

[15] In fact the distance is equal to this difference: Exercise 10.21.

and numerate $\mathbb{Q}$ as a sequence $q_n$ to consider

$$\sum_{n=1}^{\infty} \frac{1}{n^2} H_{q_n}(x).$$

Use Exercise 2.44.

**Exercise 4.45.** 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.

$$\inf_{n \in \mathbb{N}} f_n(x) = 0. \tag{4.11}$$

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 (4.11).

## 4.4 Summary

$$\begin{array}{ccc}
\begin{array}{c}\text{The definition of convergence of}\\ \text{a sequence in } \mathbb{R} \text{ to a limit in } \mathbb{R}\end{array} & \xrightarrow{\text{used to define}} & \begin{array}{c}\text{continuity of } f : [a,b] \to \mathbb{R}\\ \text{in a point } \bar{x} \in [a,b]\end{array} \\[2ex]
\begin{array}{c}\text{bounded real sequences}\\ \text{have limit points}\end{array} & \xrightarrow{\text{used to prove that}} & \begin{array}{c} f \text{ continuous in every } \bar{x} \in [a,b]\\ \text{has } |f|_{\max} = \max_{a \le x \le b} |f(x)| < \infty\end{array} \\
\downarrow \text{ used to prove} & & \downarrow \text{ used to define} \\
\text{Cauchy's criterion for convergence in } \mathbb{R} & & \text{(uniform) convergence in } C([a,b])
\end{array}$$

Cauchy's criterion also holds in $C([a,b])$ for convergence in the maximum norm, *but a bounded sequence $f_n$ in $C([a,b])$ may not have a limit point.*

Counterexamples: $[a,b] = [0,1]$, $f_n(x) = \frac{nx}{1+nx}$, $g_n(x) = x^n$.

NB The sequence $n$ is not bounded in $\mathbb{R}$; this implies Archimedes' principle.

# 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

$$d(f, g) = \max_{a \le x \le b} |f(x) - g(x)| = |f - g|_{\max} = |f - g|_\infty$$

for the *uniform distance* between two functions $f$ and $g$ in $C([a, b])$. Referring to Definition 3.7 and Theorem 3.9 we reformulate Theorem 4.15 and look ahead with some remarks before we come to the topics of this chapter.

**Theorem 5.1.** *Let $f_n$ be a sequence in $C([a, b])$ for which $d(f_n, f_m) \to 0$ as $m, n \to \infty$. Then there exists $f \in C([a, b])$ such that $d(f_n, f) \to 0$ as $n \to \infty$.*

**Remark 5.2.** *The space $C([a, b])$ is a lot like $\mathbb{R}$ as far as multiplication, addition and norms are concerned. A minor difference*[1] *is that in general*

$$|fg|_{max} \le |f|_{max}\, |g|_{max}$$

*does not hold with equality*[2]*. Because of the properties in Theorem 4.7 and Theorem 4.15 we say that $C([a, b])$ is a* complete[3] normed algebra. *Such algebras are called* Banach algebras[4].

**Remark 5.3.** *The space $C([a, b])$ is commonly used for the construction of solutions of differential equations, via a transformation to so-called integral equations*[5]*. Via Theorem 5.14 below these will be solved in Section 7.5.*

**Remark 5.4.** *We note that $C([a, b])$ is 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.*

**Remark 5.5.** *The Banach algebra $C([a, b])$ is contained in $B([a, b])$, the Banach algebra of all* bounded functions *$f : [a, b] \to \mathbb{R}$, normed by*

$$|f|_\infty = \sup_{a \le x \le b} |f(x)|. \tag{5.1}$$

*Unfortunately most of the functions in $B([a, b])$ resist integration.*

[1] A major difference: there is no Theorem 3.20 for $C([a, b])$, see Exercise 4.12.
[2] Whereas $|xy| = |x|\,|y|$ holds for all $x, y \in \mathbb{R}$.
[3] The word "complete" will be explained in Definition 5.10.
[4] You may know the expression from *Flowers for Algernon*.
[5] The commonly used space in fact, but we'll have second thoughts in Section 7.5.

## 5.1 Complete metric spaces

Henceforth we shall call a map $d$ which assigns to every pair of elements of a set $X$ a number in $\mathbb{R}$ a metric if it has the following three properties:

$$d(x,x) = 0 \quad \text{for all} \quad x \in X; \tag{5.2}$$

$$d(x,y) = d(y,x) > 0 \quad \text{for all} \quad x,y \in X \quad \text{with} \quad x \neq y; \tag{5.3}$$

$$d(x,y) \leq d(x,z) + d(z,y) \quad \text{for all} \quad x,y,z \in X. \tag{5.4}$$

The examples $X = \mathbb{R}$ and $X = C([a,b])$ lead us to consider *metric spaces*[6].

**Definition 5.6.** *Let $X$ be a nonempty set. A function*

$$d : X \times X \to \mathbb{R}$$

*is called a* metric *if the properties (5.2),(5.3),(5.4) hold. The set $X$ is then called a* metric space *with metric $d$. The number $d(x,y)$ is also called the distance from $x$ to $y$.*

In particular $X = \mathbb{R}$ and $X = C([a,b])$ are examples of metric spaces. Every nonempty subset $A$ of a metric space $X$ is also a metric space, with its metric inherited from the metric on $X$.

**Remark 5.7.** *The completeness of $B([a,b])$ follows (much easier) along the lines of the proof of Theorem 4.15. For $f \in C([a,b])$ the* supremum norm *in (5.1) is just the maximum norm announced in (1.17) and defined in Definition 4.5.*

**Exercise 5.8.** Think about other examples. Subsets of $\mathbb{R}^2$ with the Pythagorean distance[7]. 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.

Have a look at Exercise 4.14 to extrapolate some terminology to the general case. The metric $d$ is called a *strictly positive symmetric function*, because axiom[8] (5.3) says that $d(x,y) = d(y,x) > 0$ for $x \neq y$. Axiom (5.4), the *triangle inequality*, was already hinted at in Exercise 2.14, in the absence of triangles. The first axiom (5.2) 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.

[6] Forgetting about algebra for now.

[7] Illustrate the triangle inequality with a picture of a triangle in this case!

[8] An axiom is a property that we assume.

**Remark 5.9.** *The axioms (5.2,5.3,5.4) 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 $x, y, z \in X$. Nonnegativity of $d(x,y)$ follows using symmetry and the triangle inequality. See Exercise 5.70.*

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 $X = C_0(\mathbb{R})$ in Exercise 5.13 below 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 5.10.** *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*[9]*. 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*[10]*.*

**Exercise 5.11.** Explain again why $\mathbb{R}$ is complete with $d(x,y) = |x-y|$, and that so is every closed subset of $\mathbb{R}$. Then explain again why $C([a,b])$ is complete with the metric defined by

$$d(f,g) = |f-g|_{\max}.$$

**Exercise 5.12.** Let $X$ be a complete metric space. Prove that the intersection of a sequence of dense[11] open subsets of $X$ is itself dense in $X$ (*Baire's Theorem*).

[9]With limit $\bar{x}$ in $X$, because there's nothing outside $X$ here.

[10]The completeness assumption is in fact equivalent to the statement in Theorem 2.5.

[11]A set $D \subset X$ is called dense in $X$ if every $x$ in $X$ is a limit of a sequence $x_n$ in $D$.

**Exercise 5.13.** Let $f : \mathbb{R} \to \mathbb{R}$ be a function. We say that $f$ vanishes at infinity if

$$\forall_{\varepsilon>0}\,\exists_{R>0}\,\forall_{x\in\mathbb{R}} : |x| \geq R \implies |f(x)| < \varepsilon. \tag{5.5}$$

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[12] with the (maximum-)norm well-defined by

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

## 5.2 The Banach Contraction Theorem

In view of Definition 5.10 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. In `https://youtu.be/g9-z76Bcrr4` we actually cut the story short proving the theorem for $X = \mathbb{R}$ first, rewrite the proof without those absolute values, and identify what we need of $X$ and $d$.

**Theorem 5.14.** *(Banach Contraction Theorem) Let $X$ be a complete metric space and let $f : X \to X$ be a contraction, i.e.*

$$\exists_{\theta\in(0,1)}\,\forall_{x,y\in X} : d(f(x), f(y)) \leq \theta\, d(x, y).$$

*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.** This is `https://youtu.be/A9CTKR3qwJg`. Differences $x_n - x_m$ have meaning nor part in the formulation of Theorem 5.14, so the proof of Theorem 3.16 cannot be copy-pasted as it is. Still, the proof remains largely the same, and in fact the small changes make the proof more transparent.

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

$$d(x_1, x_2) \leq \theta d(x_0, x_1), \quad d(x_2, x_3) \leq \theta d(x_1, x_2) \leq \theta^2 d(x_0, x_1),$$

$$d(x_3, x_4) \leq \theta d(x_2, x_3) \leq \theta^3 d(x_0, x_1), \quad d(x_4, x_{4+1}) \leq \theta^4 d(x_0, x_1),$$

and so on. Replacing 4 by $n$ in the last inequality we have

$$d(x_n, x_{n+1}) \leq \theta^n d(x_0, x_1), \tag{5.6}$$

[12] A bit less like $\mathbb{R}$ since it does not contain a neutral element for multiplication.

which holds[13] 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[14]

$$
\begin{aligned}
d(x_n, x_m) &\le d(x_n, x_{n+1}) + d(x_{n+1}, x_m) \\
&\le d(x_n, x_{n+1}) + \cdots + d(x_{m-1}, x_m) \\
&\le (\theta^n + \cdots + \theta^{m-1})\, d(x_0, x_1) \\
&< \frac{\theta^n}{1-\theta}\, d(x_0, x_1) \le \frac{\theta^N}{1-\theta}\, d(x_0, x_1)
\end{aligned}
$$

for $m > n \ge N$, with $N$ waiting for $\varepsilon > 0$ to show up. Here it is.

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. It follows that

$$d(x_n, x_m) < \varepsilon$$

for all $m > n \ge N$. We have thus proved that $x_n$ is a Cauchy sequence.

Since $X$ is complete the sequence $x_n$ is convergent[15]. Denote its limit by $\bar{x}$ and introduce $x_n$ as before in (3.11) by means of the triangle inequality. This yields

$$
\begin{aligned}
d(\bar{x}, f(\bar{x})) &\le d(\bar{x}, x_{n+1}) + d(x_{n+1}, f(\bar{x})) \\
&= d(\bar{x}, x_{n+1}) + d(f(x_n), f(\bar{x})) \\
&\le d(\bar{x}, x_{n+1}) + \theta d(x_n, \bar{x}) < (1+\theta)\varepsilon
\end{aligned}
$$

for all $n \ge N$, the $N$ that comes with $\varepsilon$ 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

$$0 < d(\bar{x}, \tilde{x}) = d(f(\bar{x}), f(\tilde{x})) \le \theta\, d(\bar{x}, \tilde{x}) < d(\bar{x}, \tilde{x}),$$

a contradiction. This completes a clean proof without algebra. □

Theorem 5.14 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.

[13] By induction if you insist.

[14] As in Exercise 3.18.

[15] The same conclusion trivially holds if $x_0$ is a fixed point of $f$.

**Definition 5.15.** *A subset $A$ of a metric space $X$ is called* closed *in $X$ 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 a small modification 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 reparation for subsets $A$ of $X$ flawing this property was not yet formulated[16].

**Theorem 5.16.** *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[17] $x_n$ 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 $A$.*

**Exercise 5.17.** Prove Theorem 5.16.

Hint: show that $\bar{A}$ is closed, and that there is no closed $\tilde{A} \neq \bar{A}$ with $A \subset \tilde{A} \subset \bar{A}$.

**Theorem 5.18.** *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 5.19.** Prove Theorem 5.18.

In particular the closure of the so-called *open ball*

$$B_r(x_0) = \{x \in X : d(x, x_0) < r\}$$

with *center* $x_0 \in X$ and *radius* $r > 0$ in $\mathbb{R}$ is given by the closed ball

$$\bar{B}_r(x_0) = \{x \in X : d(x, x_0) \leq r\}.$$

Here we like closed subsets of complete metric spaces over open subsets[18], because Theorem 5.14 applies to contractions of such closed sets. But there are good reasons to prefer open balls[19] over closed balls: `https://youtu.be/dtRw7Zq2kQU`.

---

[16]Which you should compare to constructions of $\mathbb{R}$ out of the rational numbers.

[17]Including sequences $a, a, a, a, \ldots$ with $a \in A$.

[18]Not defined yet, for lack of good reasons to do so.

[19]We like our subsets to be closed but our balls to be open.

## 5.3 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 5.20.** *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 $\xi$ 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 5.21.** 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 5.22.** *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*[20]

$$\exists_{\bar{x} \in \mathbb{R}} \forall_{\varepsilon > 0} \forall_{N \in \mathbb{N}} \exists_{n \geq N} : d(x_n, \bar{x}) < \varepsilon. \tag{5.7}$$

*Such metric spaces are called (sequentially) compact*[21]*. This leads to:*

**Theorem 5.23.** *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 every point of $X$ then $f$ has a global maximum and a global minimum. The real number*

$$|f|_{max} = \max_{x \in X} |f(x)| \tag{5.8}$$

*is thus well-defined and called the maximum norm of $f$.*

**Remark 5.24.** *In Theorem 5.14 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 5.25.** *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*

$$d_Y(f(x), f(y)) \leq L\, d_X(x, y). \tag{5.9}$$

[20] See the reformulation of the convergent subsequence property in Remark 3.23.
[21] See Exercise 4.36.

Examples[22] are $Y = X$, with

$$d(f(x), f(y)) \leq L\, d(x, y),$$

and $Y = \mathbb{R}$, with

$$|f(x) - f(y)| \leq L\, d(x, y).$$

## 5.4 Normed spaces and Lipschitz functions

You can jump to Theorem 5.39 for the main result of this section.

**Exercise 5.26.** 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

$$\sum_{n=1}^{\infty} |x_n|_X < \infty. \tag{5.10}$$

Prove that

$$|S|_X \leq \sum_{n=1}^{\infty} |x_n|_X.$$

**Exercise 5.27.** (*continued*) Prove that $X$ is complete as a metric space with the norm defined by $d(x, y) = |x - y|_X$ if

$$\sum_{n=1}^{\infty} |x_n|_X < \infty.$$

implies that the sequence

$$S_n = x_1 + \cdots + x_n$$

is convergent. Also formulate and prove the converse of this stament.

[22] Not to bore you with the example in Definition 3.3.

**Exercise 5.28.** *Classification of finite-dimensional normed spaces*[23]. Let $X$ be a normed (vector) space of finite dimension $N$, meaning that there are $e_1, \ldots, e_N$ in $X$ such that every $x \in X$ is uniquely written as

$$x = \xi_1 e_1 + \cdots + \xi_N e_N, \quad \text{with} \quad \xi_1, \ldots, \xi_N \in \mathrm{I\!R}.$$

Then the linear map $L : \mathrm{I\!R}^{\mathrm{N}} \to X$ defined by

$$L(\xi) = L(\xi_1, \ldots, \xi_N) = \xi_1 e_1 + \cdots + \xi_N e_N$$

is a bijection. Prove that the function $f : \mathrm{I\!R}^{\mathrm{N}} \to [0, \infty)$ defined by

$$f(\xi) = |L(\xi)|$$

is continuous.

**Exercise 5.29.** (*continued*) Prove there exist $\underline{\xi}$ and $\overline{\xi}$ in $\mathrm{I\!R}^{\mathrm{N}}$ with $|\underline{\xi}|_2 = |\overline{\xi}|_2 = 1$ such that $0 < m = f(\underline{\xi}) \le M = f(\overline{\xi})$ and

$$\forall_{\xi \in \mathrm{I\!R}^{\mathrm{N}}} : \quad |\xi|_2 = 1 \implies m \le f(\xi) \le M.$$

Hint: apply Theorem 5.23.

**Exercise 5.30.** (*continued*) Show $L$ is Lipschitz continuous with Lipschitz constant $M$, and show its inverse $L^{-1}$ is Lipschitz continuous with Lipschitz constant $\frac{1}{m}$.

**Exercise 5.31.** (*continued*) Prove that every bounded sequence in the finite-dimensional $X$ under consideration in Exercise 5.29 has a convergent subsequence, and explain why $X$ is complete.

**Exercise 5.32.** Let $X$ be a normed vector space, let $f : X \to \mathrm{I\!R}$ be a linear function, and let $\xi \in X$. Suppose there exist[24] one $\varepsilon > 0$ and a $\delta > 0$ such that

$$\forall_{x \in X} : \quad |x - \xi| < \delta \implies |f(x) - f(\xi)| < \varepsilon.$$

Prove that[25]

$$\forall_{x \in X} : \quad |x| < \delta \implies |f(x)| < \varepsilon. \tag{5.11}$$

Hint: $f(x) = f(x + \xi - \xi) = f(x + \xi) - f(\xi)$.

[23] You can jump to Chapter 29 from here if you like.
[24] Just one $\varepsilon$ is needed here.
[25] That is: this statement for $\xi$ also holds for $\xi = 0$, and vice versa in fact.

**Exercise 5.33.** (*continued*) Let $X$ be a normed vector space, let $f : X \to \mathbb{R}$ be a linear function, and suppose that (5.11) holds for some $\varepsilon > 0$ and $\delta > 0$. Let $L > 0$ be such that $\delta L = \varepsilon$. Prove that

$$\forall_{x \in X}: \quad |x| < 1 \implies |f(x)| < L. \tag{5.12}$$

Hint: $f(\delta x) = \delta f(x)$.

**Exercise 5.34.** (*continued*) Let $X$ be a normed vector space, let $f : X \to \mathbb{R}$ be a linear function, and suppose that (5.12) holds for some $L > 0$. Prove that[26]

$$\forall_{x \in X}: \quad |f(x)| \leq L|x|, \tag{5.13}$$

and that $f$ is Lipschitz continuous with Lipschitz constant $L$.

**Definition 5.35.** *Let $X$ be a metric space and $\xi \in X$. Then $Lip_\xi(X)$ is by definition the set of all[27] Lipschitz continuous functions $f : X \to \mathbb{R}$ with $f(\xi) = 0$. We define the Lipschitz norm $|f|_{Lip}$ of $f \in Lip_\xi(X)$ to be the smallest $L \geq 0$ such that*

$$\forall_{x,y \in X}: \quad |f(x) - f(y)| \leq Ld(x, y).$$

*Special case: $X$ a normed space, $\xi = 0$, $f : X \to \mathbb{R}$ linear satisfying (5.13).*

**Exercise 5.36.** Prove that the definitions

$$(f + g)(x) = f(x) + g(x), \quad (\lambda f)(x) = \lambda f(x)$$

make $Lip_\xi(X)$ a real vector space.

**Exercise 5.37.** Prove that the real vector space $Lip_\xi(X)$ is a normed space with the norm defined in Definition 5.35.

**Exercise 5.38.** Prove that the real normed space $Lip_\xi(X)$ is complete.
Hint:

$$|f_n(x) - f_m(x)| = |(f_n - f_m)(x) - (f_n - f_m)(\xi)| \leq |f_n - f_m|_{\text{Lip}}\, d(x, \xi)$$

allows to define[28] the pointwise limit of a Cauchy sequence $f_n$ in $Lip_\xi(X)$.

---

[26] This also called the boundedness of $f$, not on $X$, but on (all) balls.
[27] We cannot speak of linear here.
[28] Just as (4.7) did.

**Theorem 5.39.** *Let $X$ be a real normed space, and let $X^*$ be the set of all linear Lipschitz continuous functions $f : X \to \mathbb{R}$. Then $X^*$ is a closed linear subspace of $Lip_\xi(X)$ and thereby complete*[29]*. We write*

$$|f| = \sup_{0 \neq x \in X} \frac{|f(x)|}{|x|}$$

*for the norm of $f$ and call $X^*$ the dual space of $X$.*

## 5.5 Outlook: topology

There's more to be copy-pasted from Section 3.6 with $\mathbb{R}$ replaced by $X$. We refrase what we did with metric spaces in terms of open sets. This prepares for the generalisation from metric spaces to topological spaces.

**Definition 5.40.** *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*[30] *an* open ball *centered at $\xi$ with radius $r > 0$.*

**Exercise 5.41.** Prove that the set $B_r(\xi)$ in Definition 5.40 is open.

**Exercise 5.42.** 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 5.43.** *If we denote the collection of all open subsets of $X$ by $\mathcal{T}$, then Exercise 5.42 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 5.40.*

[29] Which does not exclude $X^*$ only containing the zero function, but see Exercise 29.11.

[30] Whatever meaning these words may have.

**Theorem 5.44.** *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[31] 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 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 $\xi$. 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 $\xi$. Therefore there exists $r > 0$, but let's call it $\delta$, 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

$$d_X(x, \xi) < \delta \implies d_Y(f(x), f(\xi)) < \varepsilon. \tag{5.14}$$

To finish we should not forget the sequence $x_n$ we started with, and its limit $\xi$. 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. $\square$

**Remark 5.45.** *The reformulation of continuity in every point in terms of open sets given in Theorem 5.44 involved the first $\varepsilon$-$\delta$-statement (5.14) in these lecture notes. Such statements are the subject of Chapter 8 first.*

[31] The reasoning is similar to that in the proof of Theorem 3.28.

## 5.6 Compactness with open coverings

Compactness via convergent subsequences can be reformulated in terms of open sets too. We first establish the reformulation as a consequence and then show that in turn it implies (sequential) compactness.

**Definition 5.46.** *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 5.47.** *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 called 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

$$p_n \notin O_1 \cup \cdots \cup O_n. \tag{5.15}$$

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 (5.15) 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*[32], 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[33] that thereby

$$p \in B_{\frac{1}{m}}(p_n) \subset O_i$$

[32] A separable metric space which is complete is called a Polish space.
[33] Prove this claim.

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 $p_1, \dots, p_n$ 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.

So suppose $A$ is sequentially compact but not totally bounded. Then there exists $\varepsilon > 0$ for which no $p_1, \dots, p_n$ as above exist. Choose $p_1 \in A$ and inductively for $n = 1, 2, \dots$ 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 5.48.** *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 $p_1, p_2, \dots, p_m$ 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. □

## 5.7 Exercises about the plane

This course is about $y = f(x)$ with $x, y \in \mathbb{R}$ and $C([a,b])$ is a Banach space consisting of some such $f$. We draw their graphs in the real *plane* $\mathbb{R}^2$, also a space.

**Exercise 5.49.** For $x = (x_1, x_2) \in \mathbb{R}^2$ define $|x|_{\text{max}} = \max(|x_1|, |x_2|)$. 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 $x_1 x_2$-plane.

**Exercise 5.50.** (*continued*) Same question for $|x|_1 = |x_1| + |x_2|$.

**Exercise 5.51.** (*continued*) By definition the Euclidean distance $d$ derives from the norm defined by $|x|_2 = \sqrt{x_1^2 + x_2^2}$. Prove the triangle inequality[34] for this norm and give the definition of $d(x,y)$ for $x, y \in \mathbb{R}^2$ in terms of this norm.

**Exercise 5.52.** (*continued*) The Euclidean distance also derives from the standard inner product defined by $x \cdot y = x_1 y_1 + x_2 y_2$, so

$$d(x,y) = \sqrt{(x-y)\cdot(x-y)}.$$

Prove the *parallelogram law* for the parallelogram with vertices $0$, $x$, $y$ and $x+y$. In case you don't know this law, find it. It relates the squares of all possible distances between the four vertices to one another by one simple formula, which we call the parallelogram law for $x$ and $y$.

**Exercise 5.53.** 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[35]

$$\xi \in K_1 \cap K_2 \cap K_3 \subset O,$$

with $K_1, K_2, K_3$ open half planes. An open half plane is a set of the form

$$K = \{x \in \mathbb{R}^2 : a_1 x_1 + a_2 x_2 < b\}$$

[34] No pictures allowed in the proof.

[35] The number of halfspaces needed is $3 = 2 + 1$, the dimension of $\mathbb{R}^2$ plus 1.

with $a_1, a_2, b \in \mathbb{R}$ and $a_1$, $a_2$ not both equal to zero. Prove this statement.

**Remark 5.54.** *We could start topology in $\mathbb{R}^2$ with the observation that $K$ as above should be open in any reasonable approach and take it from there*[36].

**Exercise 5.55.** Prove that every such $K$ is a convex set, meaning that with $x, y \in K$ also the closed segment[37]

$$[x, y] = \{tx + (1-t)y : 0 \le t \le 1\} \tag{5.16}$$

is in $K$, i.e. $K$ is *convex* if the implication

$$x, y \in K \implies [x, y] \in K \tag{5.17}$$

holds true.

**Exercise 5.56.** If $K_1$ and $K_2$ are convex then so is $K_1 \cap K_2$. Why? Prove that any intersection

$$\cap_{i \in I} K_i$$

of convex sets $K_i$ indexed by some index set $I$ is convex.

**Exercise 5.57.** Let $K$ be convex and nonempty. Then the function $f_0 : K \to \mathbb{R}$ defined by $f_0(x) = d(x, 0)^2 = x \cdot x$ is nonnegative on $K$. Thus there exists a (so-called minimizing) sequence $x_n \in K$ such that

$$f_0(x_n) \to I_0 = \inf_{x \in K} f_0(x)$$

as $n \to \infty$. Use the parallelogram law to show that $x_n$ is a Cauchy sequence.

Hint: $t = \frac{1}{2}$ in (5.16) implies

$$\frac{x_n + x_m}{2} \in K, \quad \text{so} \quad \frac{x_n + x_m}{2} \cdot \frac{x_n + x_m}{2} \ge I_0,$$

and you know what $x_n \cdot x_n$ and $x_m \cdot x_m$ do as $m, n \to \infty$. The parallelogram law then tells you what $(x_n - x_m) \cdot (x_n - x_m)$ must do.

**Exercise 5.58.** (*continued*) Thus $x_n$ converges to a limit $\xi_0 \in \mathbb{R}^2$. Prove that every minimizing sequence converges to $\xi_0$.

[36] And then nobody will call non-balls balls anymore.

[37] With this notation $[x, y] = [y, x]$.

## 5.8 Exercises

**Exercise 5.59.** Let $x_n$ be a sequence in a metric space $X$. Prove that $x_n \to \bar{x} \in X$ 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 5.60.** Prove that every compact metric space is complete.

**Exercise 5.61.** 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 5.21 and the negation of (5.7).

**Exercise 5.62.** (*Dini's theorem*) 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 $X$ .

Hint: Exercise 4.45.

**Exercise 5.63.** Let $X$ and $Y$ be metric spaces. Prove that $f : X \to Y$ is continuous if and only if $f^{-1}(G) = \{x \in X : f(x) \in G\}$ is closed in $X$ for every $G$ closed in $Y$.

**Exercise 5.64.** Let $X$ and $Y$ be metric spaces. Prove that Lipschitz continuity of $f : X \to Y$ implies pointwise continuity of $f$.

**Exercise 5.65.** 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. Is it Lipschitz continuous?

**Exercise 5.66.** Referring to Remark 5.2, prove that $B([a, b])$ is a complete metric space with the metric defined by $d(f, g) = |f - g|_\infty$.

**Exercise 5.67.** For $T > 0$ and $k \in \mathbb{N}$ let $B_k([a,b])$ be the space of functions $f : [a,b] \to \mathbb{R}$ for which the statement[38]

$$\exists_{M\geq 0} \forall_{x\in[a,b]} \quad |f(x)| \leq M|x|^k \tag{5.18}$$

holds true, and let $|f|_k$ be the smallest $M \geq 0$ for which (5.18) holds. Prove that this makes $B_k([a,b])$ a complete metric space with $d_k$ defined by $d_k(f,g) = |f-g|_k$.

**Exercise 5.68.** 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[39] by

$$g(y) = y^2 \quad \text{and} \quad g(y) = \frac{y}{1+y^2}.$$

**Exercise 5.69.** 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 5.70.** Suppose $X$ is a set and $d : X \times X \to \mathbb{R}$ 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 $x, y \in X$. If in addition

$$d(x,y) = 0 \implies x = y$$

for all $x, y \in X$, then $d$ is a metric, and $X$ is a metric space with metric $d$.

[38] With $[a,b] = [-T,T]$ this prepares for Exercise 7.57, note $k = 0$ is not of new interest.
[39] See Exercise 3.48.

**Exercise 5.71.** Let $X$ be a metric space with metric $d$. Prove that

$$\tilde{d}(x,y) = \frac{d(x,y)}{1+d(x,y)}$$

also defines a metric on $X$. Explain why this metric is bounded by $1$.

**Exercise 5.72.** (continued) Prove that $d$ and $\tilde{d}$ define the same collection of open balls in $X$ if we agree to call $X$ itself an open ball too.

**Exercise 5.73.** (continued) Prove that $d$ and $\tilde{d}$ give rise to the same continuous functions $f : X \to \mathbb{R}$.

**Exercise 5.74.** Suppose that a nonempty set $X$ has two metrics $d_1$ and $d_2$ defined on it. Prove that

$$\bar{d}(x,y) = \frac{1}{2}d_1(x,y) + \frac{1}{4}d_2(x,y)$$

also defines a metric on $X$.

**Exercise 5.75.** Suppose that a nonempty set $X$ has a sequence of metrics $d_n$ bounded by $1$ and indexed by $n \in \mathbb{N}$ defined on it. Prove that

$$d_\infty(x,y) = \sum_{n=1}^{\infty} \frac{d_n(x,y)}{2^n}$$

also defines a metric on $X$ bounded by $1$.

**Exercise 5.76.** (continued) What does it mean for a function $f : X \to \mathbb{R}$ to be continuous with respect to $d_\infty$?

The story so far:

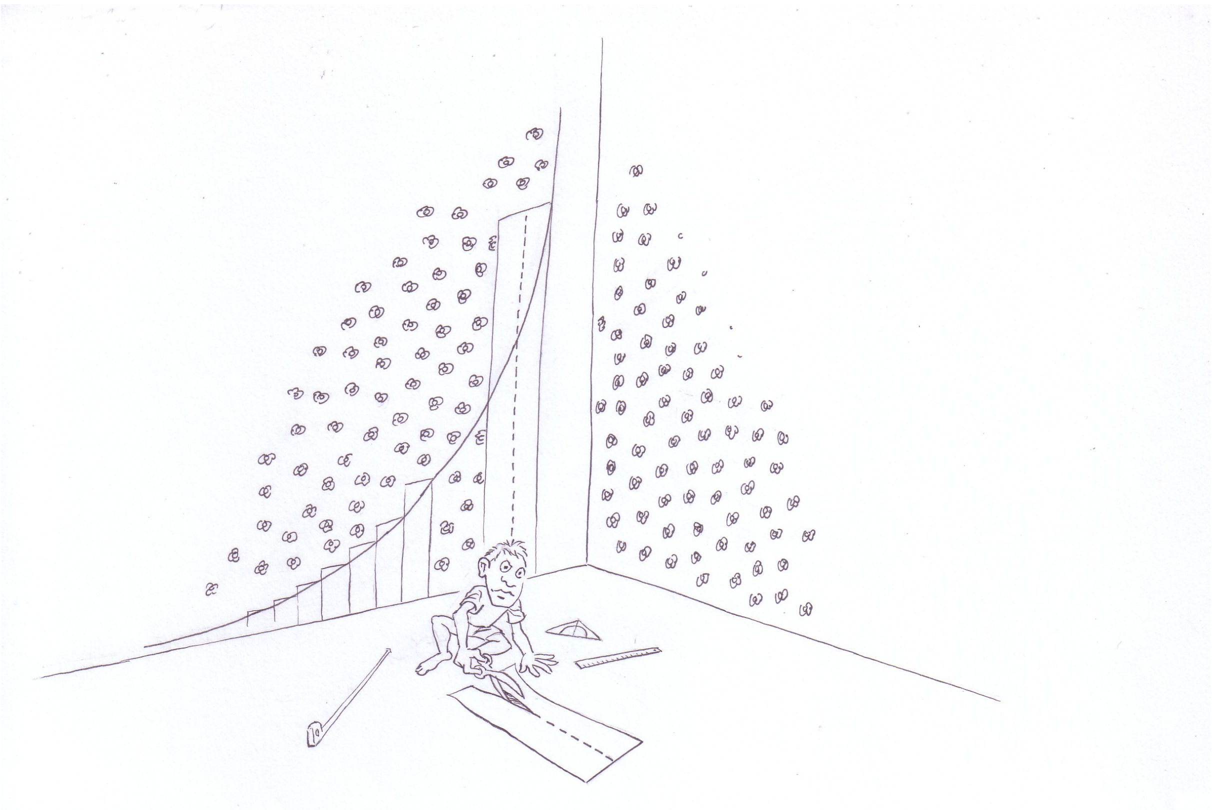
"In the beginning the Universe was created. This has made a lot of people very angry and been widely regarded as a bad move. Many races believe that it was created by some sort of God, though the Jatravartid people of Viltvodle VI believe that the entire Universe was in fact sneezed out of the nose of 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 the BCT thereby holds, its metric defined by $d(f,g) = |f-g|_{\max}$, 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: https://youtu.be/J_KdlZObeQI.

We continue with Part 2, revisit Archimedes and the pyramids, to first study integrals of functions $f : [a,b] \to \mathbb{R}$, ignoring our beloved $C([a,b])$ as long as we can. Back to square 1, with rectangles in fact.



# 6 Integration of monotone functions

https://www.youtube.com/playlist?list=PLQgy2W8pIli9sGfgTURtHzNwItqyiI3Tz

The above playlist combines this and the next chapter, but here we slow down the pace and consider monotone functions first. 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

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

[1] https://www.youtube.com/watch?v=2vcvh2K9wIk

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

$$S = \{(x, y) \in \mathbb{R}^2 : 0 < x < 1, 0 < y < 1\}$$

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.

For $p = 2, 3, 4, \ldots$ there is no such symmetry argument and the example $f(x) = x^2$ requires a new approach. We 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*

$$\frac{1}{10}\left(\frac{1}{100} + \frac{4}{100} + \frac{9}{100} + \frac{16}{100} + \frac{25}{100} + \frac{36}{100} + \frac{49}{100} + \frac{64}{100} + \frac{81}{100} + \frac{100}{100}\right),$$

[2]Don't bother about $0^0$ in $x = 0$ yet but note that we also agree that $|\bar{S}| = 1$.

but more than the *lower sum*

$$\frac{1}{10}\left(\frac{0}{100}+\frac{1}{100}+\frac{4}{100}+\frac{9}{100}+\frac{16}{100}+\frac{25}{100}+\frac{36}{100}+\frac{49}{100}+\frac{64}{100}+\frac{81}{100}\right).$$

Hint: look at the cartoon[3] preceding this chapter.

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 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 number we found for the volume of the pyramid in Section 1.2.

[3] Every nonzero term in the sums is the area of a rectangle with width $\frac{1}{10}$ in your sketch.
[4] Note the same reasoning applies to $\bar{A}_2$.

**Exercise 6.3.** Convince yourself that for all $p \in \mathbb{N}$ it must hold that

$$|A_p| = \frac{1}{p+1}.$$

Hint: use lower and upper sums, and Exercise 1.17.

**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 $x_1, x_2 \in [a, b]$.*

Such nonnegative nondecreasing functions can be pretty wild[5], but for the indicated approach with lower and upper sums we will now show that there are no problems in defining an integral.

[5]See Exercises 4.43 and 4.44.

**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 $f : [a,b] \to \mathrm{I\!R}$ we define the *left endpoint sums*[6]

$$L := \sum_{k=1}^{N} \underbrace{f(x_{k-1})}_{m_k}(x_k - x_{k-1}) \tag{6.9}$$

$$= f(x_0)(x_1 - x_0) + \cdots + f(x_{N-1})(x_N - x_{N-1}).$$

Each nonzero term in (6.9) is the area of an open rectangle[7]

$$(x_{k-1}, x_k) \times (0, f(x_{k-1})) = \{(x,y) \in \mathrm{I\!R}^2 :\ 0 < y < f(x_{k-1}),\ x_{k-1} < x < x_k\}$$

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 say that the left endpoint sum $L$ is a *Riemann lower 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[8]

$$[x_{k-1}, x_k] \times [0, f(x_k)] = \{(x,y) \in \mathrm{I\!R}^2 :\ 0 \le y \le f(x_k),\ x_{k-1} \le x \le x_k\},$$

with $k$ running from 1 to $N$, cover $A$ completely, because we recognize $x_k$ as *maximizer* for $f$ on $I_k$:

$$M_k = \max_{x \in I_k} f(x) = f(x_k). \tag{6.11}$$

We thus say that the *right endpoint sum*

$$R = \sum_{k=1}^{N} \underbrace{f(x_k)}_{M_k}(x_k - x_{k-1}) \tag{6.12}$$

[6]Make a sketch in which you see what these sums are.
[7]Possibly empty, if $x_{k-1} = x_k$ or $f(x_{k-1}) = 0$.
[8]Possibly reducing to line segments or points with zero area.

is an *Riemann upper sum* for the integral (6.6) that we want to define. In particular,

$$J \le R$$

if $J$ exists. We are ready to give a definition of integrability for nondecreasing functions.

**Definition 6.7.** *A nondecreasing*[9] *function $f : [a, b] \to \mathbb{R}$ is called* Riemann integrable *if there is a unique number $J$ such that*

$$L \le J \le R \tag{6.13}$$

*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[10].

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[11]

$$0 \le R_N - L_N = \sum_{k=1}^{N} (f(x_k) - f(x_{k-1}))\,\frac{b-a}{N} = (f(b) - f(a))\,\frac{b-a}{N}. \tag{6.14}$$

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.

**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 $J = \int_a^b f$.*

[9]Not necessarily nonnegative.
[10]Preferably not 1, 2, $a$, $b$, $d$ or $f$.
[11]We say that this finite sum is telescoping, see your answer to Exercise 2.44.

**Proof**. Restricting to $N = 2^n$ we obtain equidistant partitions with the property that

$$L_1 \leq L_2 \leq L_4 \leq L_8 \leq \cdots \leq R_8 \leq R_4 \leq R_2 \leq R_1. \tag{6.15}$$

You will prove this in Exercise 6.9 below. This by itself[12] implies that

$$\sup_{n\in\mathbb{N}} L_{2^n} \leq \inf_{n\in\mathbb{N}} R_{2^n},$$

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

$$L_{2^n} \leq J \leq R_{2^n}$$

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[13] $f$ is integrable. To show that (6.13) does indeed hold we need the following theorem.

**Theorem 6.11.** *Let $f : [a,b] \to \mathbb{R}$ be nondecreasing, let $P$ be a partition given by*

$$a = x_0 \leq x_1 \leq \cdots \leq x_N = b,$$

[12] In particular every such lower sum is less than or equal to every such upper sum.

[13] We did not specify $f$ so we cannot compute $J$ like we did for $f(x) = x^p$ with $p \in \mathbb{N}$.

*and let $Q$ be another partition given by*

$$a = y_0 \le y_1 \le \cdots \le y_M = b.$$

*Define the upper sum*[14]

$$\bar{S} = \sum_{k=1}^{N} f(x_k)(x_k - x_{k-1})$$

*and the lower sum*

$$\underline{S} = \sum_{l=1}^{M} f(y_{l-1})(y_l - y_{l-1}).$$

*Then* $\underline{S} \le \bar{S}$.

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

$$a = z_0 \le z_1 \le \cdots \le z_K = b, \tag{6.16}$$

*is the partition that is obtained by simultaneously putting the numbers*

$$x_1 \le \cdots \le x_{N-1} \quad \textit{and} \quad y_1 \le \cdots \le y_{M-1}$$

*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

$$\sum_{l=1}^{M} m_l(y_l - y_{l-1}) \le \sum_{i=1}^{K} \tilde{m}_i(z_i - z_{i-1}) \le \sum_{i=1}^{K} \tilde{M}_l(z_i - z_{i-1}) \le \sum_{k=1}^{N} M_k(x_k - x_{k-1})$$

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 $\underline{S} \le \bar{S}$. $\square$

[14] 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*[15]*. 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 $S_{\text{lower}}$. Then $\underline{S} \le J$ for every $\underline{S}$ because $J$ is an upper bound of $S_{\text{lower}}$. Since $\bar{S}$ is also an upper bound of $S_{\text{lower}}$ it must then be that $J \le \bar{S}$ because $J$ is the *lowest* upper bound of $S_{\text{lower}}$. 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. $\square$

**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 direct 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

$$\int_a^b (f+g) = \int_a^b f + \int_a^b g.$$

[15] The statement for nondecreasing functions is similar.

## 6.4 A limit theorem for monotone functions

This is `https://youtu.be/5dZ54_e5AKk` in the playlist.

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 nonincreasing $f_n : [a,b] \to \mathbb{R}$.*

**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

$$L_n \leq J_n \leq R_n \quad \text{and} \quad L \leq J \leq R.$$

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 $k = 0, \ldots, N$. In particular it follows that there is an $N \in \mathbb{N}$ such that for all $n \geq N$ we have

$$L - \varepsilon < L_n \leq J_n \leq R_n < R + \varepsilon.$$

But we also have that

$$L - \varepsilon < L \leq J \leq R < R + \varepsilon.$$

Thus[16]

$$|J_n - J| < R - L + 2\varepsilon < 3\varepsilon.$$

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

$$\{(x, y) \in \mathbb{R}^2 : 0 \leq y \leq x^p \leq b^p\}$$

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.

[16]With a bit more care we get $|J_n - J| < 2\varepsilon$ but so what?

**Exercise 6.20.** Likewise, for $0 \leq a < b$ and $p \in \mathbb{N}$, the area of

$$\{(x, y) \in \mathbb{R}^2 : a \leq x \leq b,\, 0 \leq y \leq x^p\}$$

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

$$e_n(x) = 1 + \int_0^x e_{n-1} \quad \text{and} \quad e_0(x) = 1 \quad \text{for every} \quad x \geq 0.$$

Then $e_n(0) = 1$ for every $n \in \mathbb{N}$. Use Exercise 6.19 and Exercise 6.16 to 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 $x \geq 0$. See also `https://youtu.be/RVJ996OxrHE`, this defines exp on the positive real axis.

Hint: establish that $e_n(x)$ is bounded from above for fixed[17] $x > 0$.

[17] Of course we don't want to keep this restriction to $x \geq 0$, see Exercise 7.54.

**Exercise 6.24.** (continued) Also show that[18]

$$\exp(\mu x) = 1 + \mu \int_0^x \exp(\mu s)\, ds$$

for every $\mu > 0$ and every $x \geq 0$. 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[19]

$$\ln n \geq \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{>\frac{1}{2}} + \underbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}_{>\frac{1}{2}} + \cdots + \frac{1}{n} = \sum_{k=2}^{n} \frac{1}{k}$$

to show[20] 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

$$f(x) = \begin{cases} g(x) & \text{for } 0 \leq x < 1 \\ 0 & \text{for } x = 1 \end{cases} \quad \text{and} \quad f_n(x) = \frac{1 - x^{2n}}{1+x}.$$

Combine Exercise 6.26 and Theorem 6.17 with $[a,b] = [0,1]$ to prove[21] that

$$\left(\frac{1}{1} - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{5} - \frac{1}{6}\right) + \left(\frac{1}{7} - \frac{1}{8}\right) + \cdots = \ln 2.$$

[18] This is for Exercise 7.30 and further.
[19] See also Exercise 2.44 and further.
[20] Give a definition first, in the spirit of Exercise 2.56.
[21] So $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots = \ln 2$.

Hint: for all $x$ with $0 \leq x < 1$ it follows from Theorem 1.9 that[22]

$$g(x) = \frac{1}{1+x} = \underbrace{1 - x + x^2 - x^3 + x^4 - x^5 + x^6 - x^7}_{f_4(x)} + \cdots = \lim_{n\to\infty} f_n(x).$$

Don't watch `https://youtu.be/D5Lc8PO6spO` yet.

**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 2.56 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 $\lambda$.

[22] Plot some graphs to see what's going on.

**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 $V_{n-1}$.

**Exercise 6.33.** What about nonincreasing functions

$$f_n : [0,\infty) \to [0,\infty)$$

with the property that $f_n(x)$ is a *converging* sequence for every $x \in [0,\infty)$ with limit $f(x)$? Consider[23]

$$J_n(x) = \int_0^x f_n$$

and see what statements you would like to make and can prove about

$$\int_0^\infty f_n \quad \text{and} \quad \int_0^\infty f.$$

[23] We may prefer to write $F_n(x) = \int_0^x f_n$ instead.

# 7 Integration of bounded functions?

https://www.youtube.com/playlist?list=PLQgy2W8pIli9sGfgTURtHzNwItqyiI3Tz

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 integrate them[2]?

## 7.1 Bounded integrable functions

This is https://youtu.be/EhCo23A57J4 in the playlist. 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

$$\begin{matrix} m_k = \inf\{f(x) : x \in I_k\} \\ M_k = \sup\{f(x) : x \in I_k\} \end{matrix} \quad \text{in which} \quad I_k = [x_{k-1}, x_k]. \tag{7.2}$$

These numbers exist because[3] 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* Riemann integrable *if there exists a unique number $J \in \mathbb{R}$ such that*

$$\underline{S} = \sum_{k=1}^{N} m_k(x_k - x_{k-1}) \leq J \leq \sum_{k=1}^{N} M_k(x_k - x_{k-1}) = \bar{S}$$

[1] We have used this notation before in Section 2.4 and Theorem 7.5.
[2] We already put them in a space, $B([a, b])$, see Section 5.3.
[3] See again Section 2.4.

*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*

$$J = \int_a^b f(x)\,dx.$$

*The following criterion 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 $\varepsilon$.*

**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. So much for $B([a, b])$ as a space to consider as a domain for

$$f \xrightarrow{\int_a^b} \int_a^b f \in \mathbb{R}.$$

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[4].

[4]And also the integrability of $F \circ f$ with $F$ Lipschitz continuous and $f$ monotone.

**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 $[a,b]$.*

$$\textstyle\int_a^b (f(x)\,dx \text{ exists} \implies \int_a^b F(f(x))\,dx \text{ exists}$$

*In particular every Lipschitz continuous $F : [a,b] \to \mathbb{R}$ is integrable.*

**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[5]

$$|F(y) - F(\tilde{y})| \leq L|y - \tilde{y}| \quad \text{for all} \quad y, \tilde{y} \in R_f,$$

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[6] 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. $\square$

[5]See (3.7) in Definition 3.3.

[6]See (2.18).

**Exercise 7.6.** Prove (7.3). It suffices to consider the case that $N = 1$ and show for

$$m = \inf_{a \le x \le b} f(x), \quad m^* = \inf_{a \le x \le b} F(f(x)),$$

$$M = \sup_{a \le x \le b} f(x), \quad M^* = \sup_{a \le x \le b} F(f(x))$$

that

$$M^* - m^* \le L(M - m).$$

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 *elementary properties* of the integral without proof.

**Exercise 7.7.** Let the bounded function $f : [a, b] \to \mathbb{R}$ be integrable. Prove that

$$|\int_a^b f| \le (b - a) \, |f|_\infty ,$$

in which the norm is the supremum norm defined in (5.1).

**Exercise 7.8.** 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 7.9.** *Let $f : [a, b] \to \mathbb{R}$ be bounded and integrable. Then*[7]

$$\int_b^a f(x)\, dx := -\int_a^b f(x)\, dx.$$

[7]Consistent with the intuition that $dx$ in $\int_a^b f(x)\, dx$ is negative if $a > b$.

**Exercise 7.10.** 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[8].

**Exercise 7.11.** 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 7.12.** 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 The fundamental limit theorem

This is `https://youtu.be/8FQad90pDGs` in the playlist. 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 7.13.** *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*

$$\int_a^b f_n(x)\,dx \to \int_a^b f(x)\,dx \quad \textit{as} \quad n \to \infty.$$

[8] As integrals of bounded functions of course.

**Proof of Theorem 7.13**. 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 4.18 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[9] prove that $J = \int_0^1 f$ exists. This is the first statement in the theorem, and in particular

$$J \in [\underline{S}, \bar{S}],$$

an interval of length less than $3\varepsilon$.

Let us also examine how $P$ does for the functions $f_n$. For $n \geq N$ we have from (7.5) that[10] $|\underline{S}_n - \underline{S}| \leq \varepsilon$ and $|\bar{S}_n - \bar{S}| \leq \varepsilon$. Therefore $J_n = \int_0^1 f_n$ has the property that

$$\underline{S} - \varepsilon \leq \underline{S}_n \leq J_n \leq \bar{S}_n \leq \bar{S} + \varepsilon.$$

[9] See (2.18).

[10] Is it clear why?

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[11] completes the proof that $J_n \to J$ as $n \to \infty$, which is the second statement in the theorem. □

## 7.4 Integrals are continuous linear functionals

The title of this section is explained by the convention of calling maps from function spaces to $\mathbb{R}$ *functionals*[12], `https://youtu.be/xFBOkvOBe_k` sums up what we have and need to continue with integral equations in the next section.

**Theorem 7.14.** *Let*
$$\mathrm{RI}([a,b]) = \{f : [a,b] \to \mathbb{R} : \; f \textit{ is bounded and integrable}\} \tag{7.6}$$
*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 7.14**. First we reformulate Exercise 5.7 as a separate result in Theorem 7.16 below. Recall that in Section 5.3 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.

**Definition 7.15.** *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*
$$d(f,g) = \sup_{a \leq x \leq b} |f(x) - g(x)|$$
*for all $f, g \in B([a,b])$, just as[13] in (7.7).*

[11] Not another footnote.
[12] So functionals are functions of functions.
[13] Check that this indeed defines a metric.

**Theorem 7.16.** *The space $B([a,b])$ is a complete metric space*[14].

**Proof of Theorem 7.16**. We only have to show that $B([a,b])$ is complete[15]. We note that for $\varepsilon > 0$ and[16] $f, g \in B([a,b])$

$$d(f,g) \leq \varepsilon \iff \forall_{x\in[a,b]} : |f(x) - g(x)| \leq \varepsilon \tag{7.9}$$

holds by the definition of supremum[17].

Now let $f_n$ be a Cauchy sequence in $B([a,b])$. This means that

$$\forall_{\varepsilon>0} \, \exists_{N\in\mathbb{N}} \, \forall_{m,n\geq N} \, \forall_{x\in[a,b]} \, : \, |f_n(x) - f_m(x)| < \varepsilon.$$

Just like in the proof of Theorem 4.15 it then follows that

$$f(x) = \lim_{m\to\infty} f_m(x)$$

exists for every $x \in [a,b]$, and that

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

In other words

$$\forall_{\varepsilon>0} \, \exists_{N\in\mathbb{N}} \, \forall_{n\geq N} \, : d(f_n, f) \leq \varepsilon.$$

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 7.16. □

We now complete the proof of Theorem 7.14. Recall that by (7.9) convergence in $B([a,b])$ is equivalent to uniform convergence. The first part of Theorem 7.13 says that the space $\mathrm{RI}([a,b])$ is a closed subset[18] of the complete metric space $B([a,b])$. Theorem 5.18 then implies that $\mathrm{RI}([a,b])$ is complete and so then is the proof of Theorem 7.14. □

**Theorem 7.17.** *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.*

[14] A Banach algebra in fact, see Remark 5.2; $[a,b]$ may be replaced by any set $A \neq \emptyset$.
[15] Note that its metric “extends” the metric defined in the smaller metric space $\mathrm{RI}([a,b])$.
[16] In fact only $f - g \in B([a,b])$ is needed to define $d(f,g)$.
[17] Note again that it does not matter whether we write $\leq \varepsilon$ or $< \varepsilon$ in $\forall_{\varepsilon>0}$-statements.
[18] See Definition 5.15.

**Proof of Theorem 7.17.** By the definition of continuity[19] this is now just a reformulation of the second part of Theorem 7.13. □

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 7.18.** *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 $\phi$ defined by (7.10) is linear. Moreover, the function $|f|$ defined by $|f|(x) = |f(x)|$ is also in $\mathrm{RI}([a, b])$, and*

$$\left|\int_a^b f(x)\,dx\right| \le \int_a^b |f(x)|\,dx. \tag{7.11}$$

*Thus the functional defined by (7.10) in Theorem 7.17 satisfies*

$$|\phi(f) - \phi(g)| \le (b - a)\, d(f, g)$$

*for all $f, g \in \mathrm{RI}([a, b])$ and is thereby Lipschitz continuous.*

**Remark 7.19.** *Summing up, the space* $\mathrm{RI}([a, b])$ is a complete normed vector space[20]*, and the map*

$$\phi : \mathrm{RI}([a, b]) \to \mathbb{R}$$

*defined by*

$$\phi(f) = \int_a^b f$$

*is linear and Lipschitz continuous[21] with Lipschitz constant $L = b - a$.*

**Exercise 7.20.** Prove the statements about $f + g$ in Theorem 7.18.
Hint: reason directly from Definition 7.1.

[19] See Definition 5.20.
[20] Complete normed vector spaces are called Banach spaces.
[21] So $\int_a^b$ is in the dual of $\mathrm{RI}([a, b])$.

**Exercise 7.21.** Easy: prove the statements about $\lambda f$ in Theorem 7.18.

**Exercise 7.22.** 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 7.23.** Prove the *Lipschitz continuity of $\phi$. Hint: use Exercise 7.7.*

## 7.5 Integral equations and weighted norms

https://www.youtube.com/playlist?list=PLQgy2W8pIli9f0_Zc8A-TEN9RknR2ZMmq

Exercise 6.23 provided us with a function[22] $f : [0,\infty) \to \mathbb{R}$ satisfying

$$f(x) = 1 + \int_0^x f = 1 + \int_0^x f(s)\,ds \tag{7.12}$$

for every $x > 0$. The playlist starts with https://youtu.be/LabyCbWhnuc and this integral equation for exp. 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 7.24.** 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[23]. 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 $L$.

[22]For good reasons denoted by exp, restriction to $x \geq 0$ for the sake of presentation.
[23]Designed to solve $f'(x) = F(f(x))$ with "initial" condition $f(0) = f_0$, Remark 7.29.

**Theorem 7.25.** *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 $L \max(|a|,|b|)$.*

**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 $\Phi$ 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[24]

$$|\Phi(f)(x) - \Phi(f)(y)| = \left| \int_y^x F(f(s))\, ds \right| \le \underbrace{\sup_{a \le s \le b} |F(f(s))|}_{=|F\circ f|_\infty < \infty} (x-y).$$

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[25]

$$|((\Phi(f_1) - \Phi(f_2))(x)| = \left| \int_0^x F(f_1(s)) - F(f_2(s))\, ds \right|$$

$$\le \left| \int_0^x |F(f_1(s)) - F(f_2(s))|\, ds \right| \le \left| \int_0^x L\, |f_1(s) - f_2(s)|\, ds \right|$$

$$= L \left| \int_0^x |f_1(s) - f_2(s)|\, ds \right| \le L \underbrace{\sup_{a \le x \le b} |f_1(s) - f_2(s)|}_{d(f_1,f_2)} |x|$$

---

[24] Recall (5.1).

[25] Using the inequality in (7.11).

for every $x \in [a,b]$. But $|x| \leq \max(|a|,|b|)$ so taking the supremum we get

$$d(\Phi(f_1),\Phi(f_2)) \leq L\max(|a|,|b|)\, d(f_1,f_2) \quad \text{for all} \quad f_1,f_2 \in \mathrm{RI}([a,b]).$$

This completes the proof. □

**Theorem 7.26.** *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 $L\max(|a|,|b|) < 1$. 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.** Note that in https://youtu.be/WansNjzVFRw this theorem is established with $f$ replaced by $x$, $x$ by $t$, and $[a,b]$ by $[-T,T]$. By Theorem 7.25 the Banach Contraction Theorem applies to $f = \Phi(f)$ in $\mathrm{RI}([a,b])$. □

**Remark 7.27.** *So we solve integral equations in $\mathrm{RI}([a,b])$, a closed subspace of $B([a,b])$. Recall from (5.1) that the supremum norm*[26] *in the vector space $B([a,b])$ was defined by*

$$|f|_\infty = \sup\{|f(x)| : x \in [a,b]\}.$$

*In Exercise 5.7 you showed that this norm*[27] *makes $B([a,b])$ complete.*

**Exercise 7.28.** 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[28] hold:

$$|f|_\infty > 0, \quad |\lambda f|_\infty = |\lambda|\,|f|_\infty, \quad |f+g|_\infty \leq |f|_\infty + |g|_\infty. \tag{7.16}$$

The zero element in $B([a,b])$ is the function defined by $f(x) = 0$ for all $x \in [a,b]$). Explain that $f \in B([a,b])$ is not equal to the zero element in $B([a,b])$ if and only if

$$\exists_{x\in[a,b]} : f(x) \neq 0.$$

[26] The use of the subscript $\infty$ is related to the limit of $\left(\int_a^b |f|^p\right)^{\frac{1}{p}}$ as $p \to \infty$ for nice $f$.
[27] Exercise 5.67 introduced variants that we will use starting from Exercise 7.56.
[28] These axioms may have been mentioned in Linear Algebra, see also Exercise 5.26.

**Remark 7.29.** *It turns out that Theorem 10.10 implies the unique solution $f$ of the integral equation (7.15) 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. The exercises below get rid of the restrictions on the interval on which the solution is constructed, but fall out of the scope of what we can do in a first course.*

**Exercise 7.30.** For $\mu > 0$ let $B_\mu([0,\infty))$ be the space of functions $f : [0,\infty) \to \mathbb{R}$ for which the *weighted norm*

$$|f|_\mu = \sup_{x \geq 0} \frac{|f(x)|}{\exp(\mu x)}$$

is finite. Show that $d_\mu(f,g) = |f-g|_\mu$ defines a metric $d_\mu$ on $B_\mu([0,\infty))$.

**Exercise 7.31.** (continued) 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 7.32.** Consider the integral equation

$$f(x) = f_0 + \int_0^x F(f(s))\,ds = f_0 + \underbrace{\int_0^x F \circ f}_{\Phi(f)(x)}, \tag{7.17}$$

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

$$|\Phi(f)(x) - \Phi(g)(x)| \leq L \int_0^x |f(s) - g(s)|\,ds$$

$$= \frac{L}{\mu} \exp(\mu x)\, \underbrace{|f-g|_\mu}_{d_\mu(f,g)}$$

for all $f, g \in \mathrm{RI}_\mu([0,\infty))$ and conclude that for the metric $d_\mu$ we have

$$d_\mu(\Phi(f), \Phi(g)) \leq \frac{L}{\mu} d_\mu(f, g).$$

**Exercise 7.33.** (continued) Then prove that there exists a $\mu > 0$ such that (7.17) has a unique solution in $\mathrm{RI}_\mu([0,\infty))$ for every $f_0 \in \mathbb{R}$.

Hint: use the Banach Contraction Theorem.

**Exercise 7.34.** (continued) Show that the integral equation (7.17) 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 (7.17) holds for all $x \in \mathbb{R}$.

Hint: put $x = -\xi$ to handle negative $x$.

**Exercise 7.35.** (concluded) We write $f(x; f_0)$ to indicate the dependence of the solution on $f_0$. We also write

$$S(x)(f_0) = f(x; f_0). \tag{7.18}$$

This defines a family of functions $S(x) : \mathbb{R} \to \mathbb{R}$. Prove that[29]

$$S(x_1 + x_2) = S(x_2) \circ S(x_1) = S(x_1) \circ S(x_2)$$

for every $x_1, x_2 \in \mathbb{R}$.

Hint: derive an integral equation for $g$ defined by $g(s) = f(s + x_1)$ from

$$f(x_1 + x_2) = f_0 + \int_0^{x_1} F(f(s))\, ds + \int_{x_1}^{x_1+x_2} F(f(s))\, ds.$$

**Exercise 7.36.** Consider the integral equation

$$f(x) = \int_0^x \int_0^t F(f(s))\, ds\, dt$$

[29] You may have guessed that this will imply that $\exp(x_1 + x_2) = \exp(x_1)\exp(x_2)$.

for $x \in [0,T]$, $T > 0$. Assume that $F : \mathbb{R} \to \mathbb{R}$ 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 7.37.** Let $f$ be the solution in Exercise 7.36. 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 7.38.** Prove that the integral equation in Exercise 7.36 has a unique solution in $\mathrm{RI}([0,T])$ for every $T > 0$.

Hint: reason as in Exercise 7.32.

**Exercise 7.39.** Prove that the integral equation in Exercise 7.36 has a unique solution in $\mathrm{RI}([-T,T])$ for every $T > 0$.

## 7.6 Exercises

**Exercise 7.40.** Show that[30]

$$f, g \in \mathrm{RI}([a,b]) \implies fg \in \mathrm{RI}([a,b]).$$

Hint: take a partition refining partitions chosen for $f$ and $g$ via[31]

$$\sup_I fg - \inf_I fg = \sup_{x,y \in I} |f(x)g(x) - f(y)g(y)|$$

and

$$f(x)g(x) - f(y)g(y) = (f(x) - f(y))g(x) + f(y)(g(x) - g(y)).$$

[30] This implies $\mathrm{RI}([a,b])$ is a Banach algebra, see Remark 5.2.

[31] https://youtu.be/pdE6tVig_FY

**Exercise 7.41.** Let $p > 1$ and $q > 1$ be as in Exercise 1.27, i.e.

$$\frac{1}{p} + \frac{1}{q} = 1, \tag{7.19}$$

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

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \quad \text{(Young's ineqality)}. \tag{7.20}$$

For amusement: give a direct proof using only algebra.

**Exercise 7.42.** Let $p > 1$ and $q > 1$ be as in Exercise 7.41, 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: use Exercise 7.41, it is sufficient to prove the inequality for the case that

$$\sum_{k=1}^{n} a_k^p = \sum_{k=1}^{n} b_k^q = 1.$$

**Exercise 7.43.** (continued) Prove *Hölders inequality*

$$\left|\int_a^b f(x)g(x)\, dx\right| \leq \left( \int_a^b |f(x)|^p\, dx \right)^{\frac{1}{p}} \left( \int_a^b |g(x)|^q\, dx \right)^{\frac{1}{q}}$$

for such $p$ and $q$ and $f, g \in \mathbb{R}\mathrm{I}([a,b])$.

Hint: Exercise 7.40 and the definition of integrability via finite sums.

**Exercise 7.44.** 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 7.45.** 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 $\int_{-1}^{1} f = 2\int_{0}^{1} f$.

**Exercise 7.46.** Exhibit a function $f : [a, b] \to \mathbb{R}$ not in $\mathrm{RI}([a, b])$ for which $|f|$ is.

**Exercise 7.47.** Use (7.16) to show the zero element in $B([a, b])$ has norm zero.

**Exercise 7.48.** Prove $|f + g|_\infty \leq |f|_\infty + |g|_\infty$ for $f, g \in B([a, b])$.

**Exercise 7.49.** For $f \in \mathrm{RI}([0, 1])$ define the function $\Phi(f) : [0, 1] \to \mathbb{R}$ by[32]

$$\Phi(f)(x) = \int_0^x (1 + f(s))\, ds$$

for all $x \in [0, 1]$. Show that this defines a Lipschitz continuous map[33]

$$\Phi : \mathrm{RI}([0, 1]) \to \mathrm{RI}([0, 1]).$$

Is[34] $\Phi$ a contraction?

**Exercise 7.50.** Same questions as in Exercise 7.49 for $\Phi$ defined by

$$\Phi(f)(x) = \int_0^x \frac{1}{1 + f(s)}\, ds,$$

but restricted to $\mathrm{RI}_+([0, 1]) = \{f \in \mathrm{RI}([0, 1]) : f(x) \geq 0 \text{ for all } x \in [0, 1]\}$. Why does $\Phi$ have a fixed point?

Hint[35]: consider $\mathrm{RI}_+([0, T])$ with $T < 1$.

**Exercise 7.51.** Let $g : \mathbb{R} \to [0, 1]$ defined by

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

[32] More on this integral equation for exp in Section 7.7, maybe skip what follows here.
[33] For a nonlinear variant see Exercise 7.50 and further.
[34] The exercises after Remark 7.29 deal with the disappointing answer.
[35] Don't use the exercises after Remark 7.29.

a) Show that $g$ is Lipschitz continuous with Lipschitz constant $L = 1$.

Hint: factorise $g(x) - g(y)$ and use

$$-\frac{1}{2} \leq \frac{x}{1+x^2} \leq \frac{1}{2}.$$

b) Prove that the integral equation

$$f(x) = \int_0^x \frac{1}{(1+s)(1+f(s)^2)}\,ds \quad \text{for all} \quad x \in [0,1]$$

has a unique solution $f$ in $\mathrm{RI}([0,1])$.

**Exercise 7.52.** 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 7.53.** 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 $x \in \mathbb{R}$.

## 7.7 Exercises on the integral equation for exp

This section continues from Exercise 7.49. So it's back to *linear integral equations*, for which we can do explicit calculations. The novelty is a trick[36] that is similar to what followed after Remark 7.29. It uses Definition 7.57 below rather than Exercise 7.31.

[36] See `https://youtu.be/H2nOB_SPEm8` and further in the playlist.

**Exercise 7.54.** As in Exercise 6.23 define $e_n(x)$ for all $x \in \mathbb{R}$ by

$$e_n(x) = 1 + \int_0^x e_{n-1}(s)\,ds$$

for every $n \in \mathbb{N}$, starting from $e_0(x) = 1$. Determine $e_1(x)$, $e_2(x)$, $e_3(x)$, $e_4(x)$. Is there any reason to restrict the values of $x$ as we increase $n$?

**Exercise 7.55.** Let $T > 0$. Show that

$$\Phi : \mathrm{RI}([-T,T]) \to \mathrm{RI}([-T,T])$$

defined by

$$\Phi(f)(x) = 1 + \int_0^x f(s)\,ds \tag{7.21}$$

is Lipschitz continuous. For which $T$ is $\Phi$ a contraction[37]?

Hint: estimate

$$(\Phi(f_1) - \Phi(f_2))(x) = \int_0^x (f_1(s) - f_2(s))\,ds.$$

**Exercise 7.56.** Write (7.21) as

$$\Phi_0(f) = \tilde{f}, \quad \tilde{f}(x) = 1 + \int_0^x f(s)\,ds,$$

and introduce $g$ and $\tilde{g}$ by setting

$$f(x) = 1 + g(x) \quad \text{and} \quad \tilde{f}(x) = 1 + \tilde{g}(x).$$

Which map $\Phi_1$ is defined by $\Phi_1(g) = \tilde{g}$? Explain why solving $\Phi_0(f) = f$ is equivalent[38] to solving $\Phi_1(g) = g$.

**Definition 7.57.** *Let $T > 0$ and $n \in \mathbb{N}$. Then $\mathrm{RI}_n([-T,T])$ is the space of Riemann integrable functions $f$ for which*[39]

$$\exists_{M \geq 0} \forall_{x \in [-T,T]} \quad |f(x)| \leq M|x|^n. \tag{7.22}$$

*As in Exercise 5.18 the smallest such $M$ is the norm $|f|_n$ of $f$ in $\mathrm{RI}_n([-T,T])$.*

[37]Only for such $T$ we can solve $\Phi(f) = f$ in $\mathrm{RI}([-T,T])$ using Theorem 5.14.

[38]Below we introduce spaces which then allow larger $T$ when invoking Theorem 5.14.

[39]This is perhaps simpler than what we did in Exercise 7.30.

**Exercise 7.58.** Why does the metric defined by $d_n(f,g) = |f-g|_n$ make $\mathrm{RI}_n([-T,T])$ a complete metric space?

**Exercise 7.59.** Take $n = 1$ in Definition 7.57. Show that $\Phi_1$ defined in Exercise 7.59 is also a map from $\mathrm{RI}_1([-T,T])$ to $\mathrm{RI}_1([-T,T])$.

Hint: use

$$|\int_0^x (f(s))\, ds| \le |\int_0^x |f|_1 \, |s|\, ds|.$$

**Exercise 7.60.** (continued) For which $T$ is $\Phi_1$ a contraction on $\mathrm{RI}_1([-T,T])$? How does your answer compare to that in Exercise 7.55?

Hint: note the $\frac{1}{2}$ in

$$|\int_0^x (f_1(s) - f_2(s))\, ds| \le |\int_0^x |f_1 - f_2|_1 \, s\, ds| = |f_1 - f_2|_1 \underbrace{\frac{1}{2} |x|^2}_{\le \frac{T}{2}|x|}.$$

**Exercise 7.61.** Referring to Exercise 7.59 and Exercise 7.54 we now choose to set

$$f(x) = \underbrace{1 + x}_{e_1(x)} + g(x) = f(x) \quad \text{and} \quad \tilde{f}(x) = e_1(x) + \tilde{g}(x).$$

Which map $\Phi_2$ is defined by $\Phi_2(g) = \tilde{g}$?

Hint: don't look at Exercise 7.63 yet.

**Exercise 7.62.** (continued) Take $n = 2$ in Definition 7.57. Show that $\Phi_2$ is also a map from $\mathrm{RI}_2([-T,T])$ to itself. For which $T$ is $\Phi_2$ a contraction on $\mathrm{RI}_2([-T,T])$? Compare your answer to your answers in Exercises 7.55 and 7.60.

**Exercise 7.63.** (continued) What flew for $e_1(x)$ in Exercise 7.61 also flies for $e_2(x)$, with $\Phi_3$ defined by

$$g(x) = f(x) - e_2(x), \quad \tilde{g}(x) = \tilde{f}(x) - e_2(x), \quad \Phi_3(g) = \tilde{g},$$

and so on. Use the maps $\Phi_n$ defined by

$$\Phi_n(g)(x) = \frac{x^n}{n!} + \int_0^x g$$

to convince yourself that the integral equation

$$f(x) = 1 + \int_0^x f(s)\, ds \quad \text{for all} \quad x \in \mathbb{R} \tag{7.23}$$

has a unique solution defined on the whole of $\mathbb{R}$. This solution is called $\exp$.

**Exercise 7.64.** For given $a, A \in \mathbb{R}$ show that the integral equation

$$g(x) = A + \int_a^x g(s)\, ds \quad \text{for all} \quad x \in \mathbb{R}$$

has a unique solution defined on the whole of $\mathbb{R}$ by transforming it into (7.23).

**Exercise 7.65.** (continued) Then take $A = \exp(a)$ and substitute $x = a + b$ to show that

$$\exp(a + b) = \exp(a) \exp(b) \tag{7.24}$$

for all $a, b \in \mathbb{R}$.

**Exercise 7.66.** People write

$$e^x = \exp(x).$$

Why would that be justified?

## 7.8 Exercises about and with sin and cos

Watch https://youtu.be/rfoNjVg7qZ0.

**Exercise 7.67.** Let $c_0 : \mathbb{R} \to \mathbb{R}$ be defined by $c_0(x) = 1$ for all $x \in \mathbb{R}$. Define the functions $s_n, c_{n+1} : \mathbb{R} \to \mathbb{R}$ by

$$s_n(x) = \int_0^x c_{n-1}, \quad c_{n+1}(x) = 1 - \int_0^x s_n,$$

for every odd $n \in \mathbb{N}$. Determine $s_1, c_2, s_3, c_4, s_5, c_6$. Write a formula with double integrals for $s_n$ in terms of $s_{n-2}$ and for $c_{n+1}$ in terms of $c_{n-1}$.

**Exercise 7.68.** Let $T > 0$. Show that

$$\Psi : \mathrm{RI}([-T,T]) \to \mathrm{RI}([-T,T])$$

defined by

$$\Psi(f)(x) = x - \int_0^x \int_0^t f(s)\, ds\, dt \tag{7.25}$$

is Lipschitz continuous. For which $T$ is $\Psi$ a contraction? If you like you can now jump to Remark 7.71, and skip the game beginning with $RI_1([-T,T])$ directly below.

**Exercise 7.69.** Write $\tilde{f} = \Psi_1(f) = \Psi(f)$ and introduce $g, \tilde{g}, \Psi_3$ as in Exercise 7.56, but now with $f(x) = x + g(x)$, $\tilde{f}(x) = x + \tilde{g}(x)$. Which $\mathrm{RI}_n([-T,T])$ would you take for $\Psi_3$? And which $\Psi_5$ in the next step? Convince yourself that the integral equation

$$f(x) = \underbrace{x - \int_0^x \int_0^t f(s)\, ds\, dt}_{\Psi_1(f)(x)} \tag{7.26}$$

has a unique solution defined for all $x \in \mathbb{R}$. It's called $\sin$. Note that $\Psi_1$ is contractive on $\mathrm{RI}_1([-T,T])$ if $T^2 < 6$. On $\mathrm{RI}_0([-T,T])$ it's only contractive if $T^2 < 2$.

**Exercise 7.70.** Play the same game as before to define $\cos$ as the unique solution of

$$g(x) = \underbrace{1 - \int_0^x \int_0^t g(s)\, ds\, dt}_{\Phi_0(g)(x)} \tag{7.27}$$

defined for all $x \in \mathbb{R}$.

Hint: start with $\mathrm{RI}_0([-T,T])$ and $\Phi_0$ contractive on $\mathrm{RI}_0([-T,T])$ if $T^2 < 2$.

**Remark 7.71.** *If we restrict the maps $\Phi$ and $\Psi$ defined by*

$$\Psi(f)(x) = x - \int_0^x \int_0^t f(s)\, ds\, dt \quad \textit{and} \quad \Phi(g)(x) = 1 - \int_0^x \int_0^t f(g)\, ds\, dt$$

*to $\mathrm{RI}([0,1])$ then both $\Phi$ and $\Psi$ are contractive with contraction factor $\frac{1}{2}$, and provide us with fixed points $f = \sin$ and $g = \cos$. Clearly $\sin$ extends as an odd function and $\cos$ as an even function to $-[1,1]$, and for $x \in [-1,1]$ we have that*

$$\cos x = 1 - \int_0^x \sin \quad \textit{and} \quad \sin x = \int_0^x \cos \tag{7.28}$$

*holds. To prove that* $\sin$ *and* $\cos$ *are the periodic functions with the properties that you know, it suffices below to show that*

$$\cos x = \sin x$$

*has a unique solution* $x = p$, *and that the picture with the two graphs of* $\cos$ *and* $\sin$ *is symmetric in the line* $x = p$.

**Exercise 7.72.** Prove (7.28).

Hint for the first one: put $f = \sin$ in (7.26) and rewrite it as

$$\sin(y) = \int_0^y \underbrace{(1 - \int_0^t f(s)\,ds)}_{g(t)}\,dt$$

to define $g$ and integrate from $y = 0$ to $y = x$.

**Exercise 7.73.** Use $\Phi$ and $\Psi$ in Remark 7.71 defined on $\mathbb{RI}([0,1])$ and assume that

$$c_2 \leq g \leq c_0 \quad \text{and} \quad s_3 \leq f \leq s_1.$$

Show that

$$c_0 \geq c_4 \geq \Phi(g) \geq c_2 \quad \text{and} \quad s_1 \geq s_5 \geq \Psi(f) \geq s_3,$$

and explain why

$$1 - \frac{x^2}{2} \leq \cos x \leq 1 \quad \text{and} \quad x - \frac{x^3}{6} \leq \sin x \leq x \quad \text{if} \quad 0 \leq x \leq 1.$$

**Exercise 7.74.** Use Exercise 7.73 to show that

$$1 + x\underbrace{(1 - \frac{x}{2} - \frac{x^2}{6})}_{>0} \leq \cos x + \sin x \leq 1 + x \quad \text{if} \quad 0 \leq x \leq 1.$$

and then (7.28) to conclude that $F$ defined by

$$F(x) = \sin x - \cos x = -1 + \int_0^x (\cos + \sin)$$

satisfies

$$x - y < F(x) - F(y) < 2(x - y) \quad \text{if} \quad 0 \leq y < x \leq 1,$$

and has $F(0) = -1$. Reason as in Exercise 2.62 to prove what we announced in Remark 7.71: in $(0, 1)$ the equation $\cos x = \sin x$ has a unique solution $x = p$.

**Exercise 7.75.** From Remark 7.71 and Exercise 7.74 we have $f = \sin$ and $g = \cos$ defined as strictly monotone functions on $[0,1]$, with $f(0) = 0$, $g(0) = 1$, and $p \in (0,1)$ defining $q = f(p) = g(p) \in (0,1)$. Verify for $f = \sin$ and $g = \cos$ that

$$f(x) = f(p) + \int_p^x g \quad \text{and} \quad g(x) = g(p) - \int_p^x f \quad \text{for all} \quad x \in [0,1].$$

Explain why the graphs of $\cos$ and $\sin$ are each others' mirror images in $x = p$. Then use even and odd extensions and symmetry arguments to complete the picture you know for all $x \in \mathbb{R}$ for the eventually globally defined solutions of (7.28).

**Exercise 7.76.** Thus, considered as a system of integral equations for the functions $\cos$ and $\sin$, (7.28) has a unique global solution. That is, the unique solution of

$$g(x) = 1 - \int_0^x f \quad \text{and} \quad f(x) = \int_0^x g \quad \text{for all} \quad x \in \mathbb{R}$$

is $(f,g) = (\sin, \cos)$. For $a, A, B \in \mathbb{R}$, $A, B$ not both zero, set $y = x + a$ and define functions $\phi$ and $\psi$ by

$$f(x) = A\phi(y) - B\psi(y) \quad \text{and} \quad g(x) = B\phi(y) + A\psi(y)$$

to derive

$$\psi(y) = \alpha - \int_a^y \phi \quad \text{and} \quad \phi(y) = \beta + \int_a^y \psi \quad \text{for all} \quad y \in \mathbb{R}, \tag{7.29}$$

with $\alpha, \beta$ depending on $A$ and $B$. Find the expressions for $\alpha$ and $\beta$.

**Exercise 7.77.** (continued) Explain why for every $\alpha, \beta \in \mathbb{R}$ the system (7.29) has a unique solution $(\phi, \psi)$ defined on the whole of $\mathbb{R}$, and why

$$\phi(y) = \alpha \sin(y-a) + \beta \cos(y-a),$$
$$\psi(y) = \alpha \cos(y-a) - \beta \sin(y-a)$$

for all $y \in \mathbb{R}$.

**Exercise 7.78.** (continued) Then put $\alpha = \cos a$, $\beta = \sin a$, $y = a + b$ to conclude that

$$\sin(a+b) = \cos a \sin b + \sin a \cos b,$$
$$\cos(a+b) = \cos a \cos b - \sin a \sin b$$

for all $a$ and $b$ in $\mathbb{R}$, whence also $\cos^2 + \sin^2 = 1$.

**Exercise 7.79.** We can do better than just saying that $\cos x = \sin x$ has a unique solution $p \in (0,1)$. Compute the unique solution $x = p_2$ of the quadratic equation

$$1 - \frac{x^2}{2} = x$$

in $(0,1)$, and show that

$$x - \frac{x^3}{6} = 1 - \frac{x^2}{2} + \frac{x^4}{24}$$

has a unique solution $x = p_4$ in $(0,1)$. Explain why

$$\cos p_2 > \sin p_2 \quad \text{and} \quad \cos p_4 < \sin p_4, \quad \text{and therefore} \quad p_2 < p < p_4.$$

Hint: use Exercise 2.62 and

$$P_4(x) = \underbrace{x + \frac{x^2}{2}}_{P_2(x)} - \frac{x^3}{6} - \frac{x^4}{24} = \int_0^x (1 + s - \frac{s^2}{2} - \frac{s^3}{6})\, ds$$

$$= x + \int_0^x s\underbrace{(1 - \frac{s}{2} - \frac{s^2}{6})}_{>0})\, ds.$$

Note that $x = p_2$ is the solution of $P_2(x) = 1$.

**Exercise 7.80.** Show that

$$s_n(x) = c_{n+1}(x)$$

has a unique solution $x = p_{n+1}$ in $(0,1)$ for every odd $n \in \mathbb{N}$, and explain why $p$ is the unique number with

$$p_2 < p_6 < p_{10} \leq \cdots < p < \cdots < p_{12} < p_8 < p_4.$$

Hint: use Exercise 2.62 again, and

$$P_6(x) = P_4(x) + \frac{x^5}{5!} + \frac{x^6}{6!} = \int_0^x (1 + s - \frac{s^2}{2} - \frac{s^3}{6} + \frac{s^4}{4!} + \frac{s^5}{5!})\, ds,$$

$$P_8(x) = \int_0^x (1 + s - \frac{s^2}{2} - \frac{s^3}{6} + \underbrace{\frac{s^4}{4!} + \frac{s^5}{5!} - \frac{s^6}{6!} - \frac{s^7}{7!}}_{\geq 0})\, ds, \quad \text{and so on.}$$

**Exercise 7.81.** You may like to verify that

$$1 \underbrace{\geq}_{\text{if } x^2 \leq 12} \underbrace{1 - \frac{x^2}{2!} + \frac{x^4}{4!}}_{c_4(x)} \underbrace{\geq}_{\text{if } x^2 \leq 56} \underbrace{1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!}}_{c_8(x)} \underbrace{\geq}_{\text{if } x^2 \leq 132} \cdots$$

$$\cdots \underbrace{\geq}_{\text{if } x^2 \leq 90} \underbrace{1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!}}_{c_6(x)} \underbrace{\geq}_{\text{if } x^2 \leq 30} \underbrace{1 - \frac{x^2}{2!}}_{c_2(x)} \underbrace{\geq}_{\text{if } x^2 \leq 2} 0,$$

$$x \underbrace{\geq}_{\text{if } 0 \leq x \leq \sqrt{20}} \underbrace{x - \frac{x^3}{3!} + \frac{x^5}{5!}}_{s_5(x)} \underbrace{\geq}_{\text{if } 0 \leq x \leq \sqrt{72}} \underbrace{x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!}}_{s_9(x)} \underbrace{\geq}_{\text{if } 0 \leq x \leq \sqrt{156}} \cdots$$

$$\cdots \underbrace{\geq}_{\text{if } 0 \leq x \leq \sqrt{110}} \underbrace{x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}}_{s_7(x)} \underbrace{\geq}_{\text{if } 0 \leq x \leq \sqrt{42}} \underbrace{x - \frac{x^3}{3!}}_{s_3(x)} \underbrace{\geq}_{\text{if } 0 \leq x \leq \sqrt{6}} 0,$$

and likewise for $x \leq 0$. The intervals on which the inequalities hold are consistent with the bounds on $T$ in the contraction arguments for (7.27) and (7.26) in $RI_n([-T, T])$.

**Exercise 7.82.** Let $f : [0,1] \to [-1,1]$ be given by

$$f(x) = \begin{cases} 0 & \text{for } x = 0 \\ \sin \frac{1}{x} & \text{for } x \neq 0 \end{cases}.$$

Recall that we write, for a partition $0 \leq x_0 \leq x_1 \leq \cdots \leq x_N = 1$,

$$I_k = [x_{k-1}, x_k], \quad m_k = \inf_{I_k} f, \quad M_k = \sup_{I_k} f,$$

$$\overline{S} = \sum_{k=1}^{N} M_k\,(x_k - x_{k-1}), \quad \underline{S} = \sum_{k=1}^{N} m_k\,(x_k - x_{k-1}).$$

a) Let $\varepsilon > 0$ and $a \in (0,1]$. Prove the existence of such a partition with $x_0 = a$ for which $\overline{S} - \underline{S} < \varepsilon$.

b) Prove that $f$ is Riemann integrable on $[0,1]$.

Hint: choose $x_0 = 0 < x_1 = a < \varepsilon$.

**Exercise 7.83.** Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} 0 & \text{for } x = 0 \\ \cos \frac{1}{x} & \text{for } x \neq 0 \end{cases}.$$

Prove that $f$ is Riemann integrable on $[0, 1]$.

**Exercise 7.84.** Referring to Definition 7.57 show that $f : [0, 1] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} x \sin \frac{1}{x} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

is in $\mathbb{RI}_1([0, 1])$.

**Exercise 7.85.** Sketch $y = x$ and $y = \cos x$ in the $x, y$-plane with $0 \leq x, y \leq 1$.

a) Use $\cos : [0, \cos 1] \to [0, \cos 1]$ to Prove that $x = \cos x$ has a unique solution.

b) Prove that the integral equation

$$f(x) = \int_0^x \cos f(s)\, ds \quad \text{for all} \quad x \in [0, 1]$$

has a unique solution in $\mathbb{RI}([0, 1])$.

## 7.9 Exercises on improper integrals and convolutions

This section prepares for a full treatment of Fourier integrals in Section 28.4 and further, without the use of Lebesgue integration theory. The new parts in red started with some questions from Daniele[40] about certain integral equations. You may like to give Exercise 6.33 another try before you read on. 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

$$\int_a^b f = \int_a^b f(x)\, dx = \lim_{\alpha \downarrow a} \lim_{\beta \uparrow b} \int_\alpha^\beta f(x)\, dx = \lim_{\beta \uparrow b} \lim_{\alpha \downarrow a} \int_\alpha^\beta f(x)\, dx \tag{7.30}$$

[40] https://twitter.com/AvitabileD

if the double limits exist. It's not hard to show that if one of 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.

**Exercise 7.86.** Show that

$$\int_{-\infty}^{\infty} f = \int_{-\infty}^{0} f + \int_{0}^{\infty} f$$

if both integrals on the right exist, and verify that

$$\int_{-\infty}^{0} f(x)\, dx = \int_{0}^{\infty} f(-x)\, dx.$$

Discuss why[41]

$$\int_{-\infty}^{\infty} \frac{\sin x}{x}\, dx = 2\int_{0}^{\infty} \frac{\sin x}{x}\, dx$$

makes sense and why the integral in (7.31) below exists. Hint: show that it is equal to the limit $S$ of the sequence defined by

$$S_n = \int_{0}^{n\pi} \frac{\sin x}{x}\, dx,$$

and that

$$S = \sum_{n=0}^{\infty} (-1)^n \int_{n\pi}^{(n+1)\pi} \frac{|\sin x|}{x}\, dx.$$

The integral

$$\int_{0}^{\infty} \frac{\sin x}{x}\, dx \tag{7.31}$$

will return[42]. You just noticed that we may not have a direct way to compute the limits in (7.30). For nonnegative functions things are usually simpler.

**Theorem 7.87.** *Suppose that $f$ and $g$ are functions from $\mathbb{R}$ to $\mathbb{R}$, integrable on every bounded interval, with $|f(x)| \leq g(x)$ for all $x \in \mathbb{R}$.*

$$\textit{If} \quad \int_{-\infty}^{\infty} g \quad \textit{exists then} \quad \int_{-\infty}^{\infty} f \quad \textit{exists.}$$

[41] Use what you know about sin, see Remark 7.71 and further.

[42] See Section 28.1 in Chapter 28.

**Exercise 7.88.** Prove Theorem 7.87 and use it with

$$g(x) = \max(1, \frac{1}{x^2})$$

to conclude that

$$\int_{-\infty}^{\infty} \frac{\sin^2 x}{x^2}\,dx$$

exists.

**Exercise 7.89.** Let $f : \mathbb{R} \to \mathbb{R}$ and suppose that

$$\int_{-\infty}^{\infty} |f| = \int_{-\infty}^{\infty} |f(x)|\,dx$$

exists, and let $g : \mathbb{R} \to \mathbb{R}$ be a bounded function which is integrable on every bounded interval. Use Exercise 7.40 to prove that

$$\int_{-\infty}^{\infty} fg = \int_{-\infty}^{\infty} f(x)g(x)\,dx$$

exists, and that

$$|\int_{-\infty}^{\infty} f(x)g(x)\,dx| \le \underbrace{\sup_{x\in\mathbb{R}} |g(x)|}_{|g|_\infty} \int_{-\infty}^{\infty} |f(x)|\,dx$$

**Exercise 7.90.** Let $f$ and $g$ be as in Exercise 7.89. For every $x \in \mathbb{R}$ we define $F(x)$ by

$$F(x) = \int_{-\infty}^{\infty} f(x-y)g(y)\,dy.$$

Explain why this integral is defined and prove that $F : \mathbb{R} \to \mathbb{R}$ is bounded. We say that $F = f * g$ is the convolution of $f$ and $g$. These convolutions will come back in Section 28.8, see (28.49). Prove that $f * g = g * f$.

**Exercise 7.91.** (continued) Assume that $f$ is continuous. Is $F = f * g$ continuous? Say in $0$? To get started take $x \in \mathbb{R}$ with $|x| \le 1$. Then let $R > 1$ and split the integral in

$$F(x) - F(0) = (f*g)(x) - (f*g)(0) = \int_{-\infty}^{\infty} (f(x-y) - f(-y))\,g(y)\,dy$$

as

$$\int_{-\infty}^{\infty} = \int_{-\infty}^{-R} + \int_{-R}^{R} + \int_{R}^{\infty}$$

to first show that

$$| \int_{R}^{\infty} (f(x-y) - f(-y))\, g(y)\, dy| \le |g|_\infty \left( \int_{R}^{\infty} |f(x-y)|\, dy + \int_{R}^{\infty} |f(-y)|\, dy \right)$$

$$\le |g|_\infty \left( \int_{-\infty}^{1-R} |f(y)|\, dy + \int_{-\infty}^{-R} |f(y)|\, dy \right) \le 2\, |g|_\infty \int_{-\infty}^{1-R} |f(y)|\, dy < \frac{\varepsilon}{3},$$

provided $R$ is sufficiently large. Hint: use that

$$\int_{-\infty}^{\infty} |f(x)|\, dx < \infty$$

and show that $R > 1$ can be chosen to have also

$$| \int_{-\infty}^{-R} (f(x-y) - f(-y))\, g(y)\, dy| < \frac{\varepsilon}{3}.$$

**Exercise 7.92.** (continued) For such a fixed $R$ use the uniform continuity of $f$ on $[-R-1, R+1]$ to show that there exists $\delta > 0$ such that

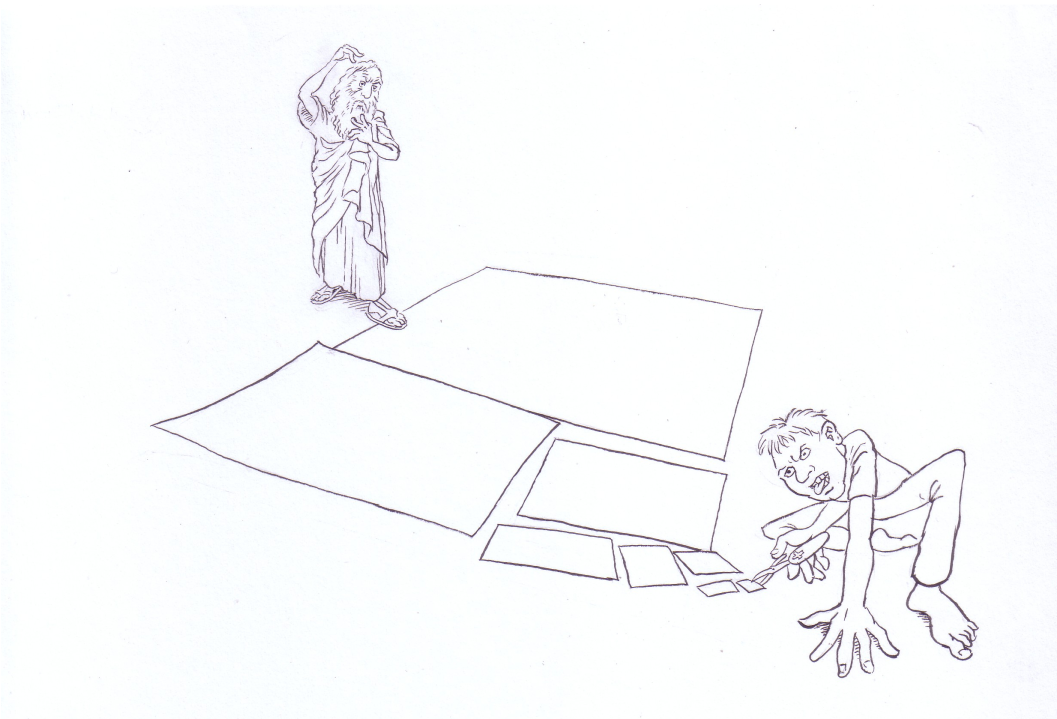
$$| \int_{-R}^{R} (f(x-y) - f(-y))\, g(y)\, dy| < \frac{\varepsilon}{3}$$

for all $x \in \mathbb{R}$ with $|x| < \delta$, and hence

$$|F(x) - F(0)| = |(f * g)(x) - (f * g)(0)| < \varepsilon.$$

Then write down and prove a theorem that says that $f * g$ is continuous specifying the minimal assumptions used above. Note the special case that $g(x) = \cos(ax)$ and have a look at Theorem 28.21. Prove that the continuity of $F$ is uniform on $\mathbb{R}$ if $f$ has the additional property that $f(x) \to 0$ if $|x| \to \infty$. Hint: show first that $f$ is uniformly continuous on $\mathbb{R}$.

**Exercise 7.93.** (continued) Give an additional condition on $g$ that ensures that also $F(x) \to 0$ if $|x| \to \infty$.



Also of interest: measure theory in $\mathbb{R}^2$

# 8 Epsilons and deltas

https://www.youtube.com/playlist?list=PLQgy2W8pIli8e9Hm34hqKFtil6pZg4I6z

Most of this chapter may be postponed till after the chapters on differentiation. Theorem 8.1 below is used in the proof of Theorem 10.10. Theorem 8.6, which says that every $f \in C([a,b])$ is integrable, merely simplifies the formulation of Theorem 10.12. The main other result in this chapter is Theorem 8.13, which formulates a condition that deals with the counterexamples in Section 4.4. We conclude this chapter with Theorem 8.15 which generalises Theorem 8.6 to the integrability of continuous functions from $[a,b]$ to a Banach space $X$.

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 5.20 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 $\xi$ 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

[1]Actually the proof of Theorem 5.44 already contained this statement, namely

$$\forall_{\varepsilon>0}\, \exists_{\delta>0} : d_X(x,\xi) < \delta \implies d_Y(f(x), f(\xi)) < \varepsilon.$$

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 $\varepsilon$ replaced by $\delta$. 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. □

## 8.1 Uniform continuity and integrability

The first video `https://youtu.be/PagygvVNsP8` in the epsilon-delta playlist states the first theorem in these notes that really requires epsilons and deltas. The other essential epsilon-delta statement[2] in analyis is discussed in `https://youtu.be/SmwYBRxNgyI`.

**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 $f(x) = \frac{1}{x}$. Hint: suppose there is an $\varepsilon > 0$ for which $\delta > 0$ can be chosen independent of $\xi \in A$. Can $f$ be unbounded?

In the above exercises we saw that sometimes $\delta$ depending on $\varepsilon$ can be chosen independent of $\xi$ for all $\varepsilon$, and sometimes it cannot. Such independence of

[2] See Definition 10.1.

$\delta$ on $\xi$ is needed to prove a theorem that we have postponed so far, namely that every continuous function $f : [a, b] \to \mathbb{R}$ is integrable, i.e.

$$C([a, b]) \subset \mathrm{RI}([a, b]). \tag{8.3}$$

**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. Now observe that for each $k$ the supremum $M_k$ is actually the global maximum of $f$ on $[x_{k-1}, x_k]$. Likewise $m_k$ is the global minimum of $f$ on $[x_{k-1}, x_k]$. The following definition establishes that given $\varepsilon > 0$ we can get $M_k - m_k < \varepsilon$ for all $k$ simultaneously by choosing a partition with $x_k - x_{k-1} < \delta$ for all $k$, $\delta > 0$ depending on $\varepsilon$.

**Definition 8.7.** *Let $A \subset \mathbb{R}$ be nonempty. A function $f : A \to \mathbb{R}$ is called* uniformly continuous on $A$ *if*

$$\forall_{\varepsilon>0} \, \exists_{\delta>0} \, \forall_{x,\xi \in A} : \underbrace{|x - \xi|}_{d(x,\xi)} < \delta \implies \underbrace{|f(x) - f(\xi)|}_{d(f(x),f(\xi))} < \varepsilon.$$

The choice for $x$ and $\xi$ in the notation is to help you compare carefully the statement in Definition 8.7 to the statement in Theorem 8.1, as well as to the statement in Definition 8.12. The continuity statement in Theorem 8.1 is clearly implied by the uniform continuity statement in Definition 8.7. According to Theorem 8.10 below, *both statements are equivalent if* $A$ is $[a, b]$ or any other closed bounded nonempty set in $\mathbb{R}$.

**Remark 8.8.** *The statement that $f$ is continuous in every $\xi \in A$ rewrites*[3] *as the non-uniform statement that*

$$\forall_{\varepsilon>0} \, \forall_{\xi \in A} \, \underbrace{\exists_{\delta>0} \, \forall_{x \in A}}_{pointwise} : |x - \xi| < \delta \implies |f(x) - f(\xi)| < \varepsilon,$$

*and differs by one $\forall_{\xi \in A}$-$\exists_{\delta>0}$ swap from the uniform statement refrased from Definition 8.7 as*

$$\forall_{\varepsilon>0} \, \underbrace{\exists_{\delta>0} \, \forall_{\xi \in A} \, \forall_{x \in A}}_{uniform} : |x - \xi| < \delta \implies |f(x) - f(\xi)| < \varepsilon.$$

[3] No difference between $\forall_{\varepsilon>0} \, \forall_{\xi \in A}$ and $\forall_{\xi \in A} \, \forall_{\varepsilon>0}$.

**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[4] 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

$$|x_n - \xi_n| < \frac{1}{n} \quad \text{and} \quad |f(x_n) - f(\xi_n)| \geq \varepsilon. \tag{8.5}$$

Both sequences are bounded. As in the proof of Theorem 4.4 it is the Bolzano-Weierstrass Theorem[5] 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. $\square$

**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[6] partition with

$$\frac{b-a}{N} < \delta,$$

[4] This nice explanation of *uniform* continuity I got from Thomas Rot.
[5] Theorem 3.20.
[6] Or any other partition with $x_k - x_{k-1} < \delta$ for all $k = 1, \ldots, N$.

it follows for $M_k$ and $m_k$ in (8.4) that $M_k - m_k < \varepsilon$ for all $k = 1, \dots, N$. This is because $m_k$ and $M_k$ as defined in (7.2) are realised[7] as values of $f$ in $I_k$, and $I_k$ has length smaller than $\delta$. 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[8]. □

## 8.2 The adjective uniform

It is instructive to have another look at the use of the adjective *uniform*[9]. Definition 4.18 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

$$\forall_{\varepsilon>0} \, \exists_{N\in\mathbb{N}} \, \forall_{n\geq N} \, \underbrace{\forall_{x\in A}}_{\text{uniform}} \; : \; |f_n(x) - f(x)| < \varepsilon \tag{8.6}$$

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).*

---

[7] Theorem 4.4 provides us with min- and maximizers.
[8] Should we mention the $(b-a)$-trick? Apply Definition 8.7 with $\frac{\varepsilon}{b-a}$.
[9] https://youtu.be/zQCCUp6cCh0
[10] Here $A$ could be any non-empty set!
[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 $\forall_{x\in A} \forall_{\varepsilon>0}$.

You should compare Remark 8.11 to Remark 8.8. Recalling (7.9) we emphasise again[13] 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*[14] *to uniform convergence on $A$.*

## 8.3 Uniform convergence and equicontinuity

In `https://youtu.be/pV1LDaEYv7Q` I discuss Theorem 8.13 below as a convergent subsequence theorem for $C([a,b])$. 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 5.2 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 4.12. The sequence $f_n$ is bounded in $C([0,1])$ but does not have a uniformly convergent subsequence.

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 $\xi$, but also[15] of continuity which is simultaneously uniform is $\xi$ 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*

$$\forall_{\varepsilon>0}\,\underbrace{\exists_{\delta>0}\,\forall_{n\in\mathbb{N}}\,\forall_{x,\xi\in A}}_{\textit{uniformly equi-}}\,:\,\underbrace{|x-\xi|}_{d(x,\xi)}<\delta\implies\underbrace{|f_n(x)-f_n(\xi)|}_{d(f_n(x),f_n(\xi))}<\varepsilon.$$

**Theorem 8.13.** *(Arzelà-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 $f\in C([a,b])$.*

[13]Recall Exercise 1.6, and note the similar statement for $\exists_{\delta>0}\cdots<\delta$.
[14]Note again that only $f_n-f\in B(A)$ is needed to have $d(f_n,f)$ well defined.
[15]We won't consider pointwise equicontinuity.

**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 $f_{n_k}(0)$ by the Bolzano-Weierstrass Theorem 3.20. Denote the limit by $f(0)$. By the same argument the subsequence $f_{n_k}$ of the sequence $f_n$ contains a further subsequence which converges in $x = 1$ as well. Denote the limit by $f(1)$. Along a further subsequence the values of $f_n$ in $x = \frac{1}{2}$ converge. The limit defines $f(\frac{1}{2})$.

Repeating the argument we define the values of our 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. Now let us denote the indices[16] of all these subsequences by

| | | | | | | |
|---|---|---|---|---|---|---|
| $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{14}$ | $n_{15}$ | $\dots$ | for the subsequence convergent in 0, |
| $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{24}$ | $n_{25}$ | $\dots$ | for the subsequence convergent also in 1, |
| $n_{31}$ | $n_{32}$ | $n_{33}$ | $n_{34}$ | $n_{35}$ | $\dots$ | for the subsequence convergent also in $\frac{1}{2}$, |
| $n_{41}$ | $n_{42}$ | $n_{43}$ | $n_{44}$ | $n_{45}$ | $\dots$ | for the subsequence convergent also in $\frac{1}{4}$, |
| $n_{51}$ | $n_{52}$ | $n_{53}$ | $n_{54}$ | $n_{55}$ | $\dots$ | for the subsequence convergent also in $\frac{3}{4}$, |

und so weiter. 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 $\mathbb{R}$, which is all we need to conclude the proof. In view of the completeness[17] of $C([0, 1])$ it 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

$$\exists_{\delta>0} \forall_{x,a\in A} \forall_{k\in\mathbb{N}} : |x - a| < \delta \implies |F_k(x) - F_k(a)| < \varepsilon.$$

[16]Have a look at the proof of Theorem 1.4.
[17]See Theorem 4.15.

We now choose $l \in \mathbb{N}$ with

$$\frac{1}{2^l} < \delta$$

and estimate the difference of $F_k(x)$ and $F_m(x)$ for arbitrary $x \in [0,1]$ by

$$|F_k(x) - F_m(x)| \leq \underbrace{|F_k(x) - F_k(a)|}_{<\varepsilon} + |F_k(a) - F_m(a)| + \underbrace{|F_m(a) - F_m(x)|}_{<\varepsilon},$$

in which for every $x \in [0,1]$ a number

$$a \in \mathcal{D}_l = \{0, \frac{1}{2^l}, \frac{2}{2^l}, \frac{3}{2^l}, \ldots, 1\}$$

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 $k, m \geq N$, 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 $k, m \geq N$. 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 4.15, namely that it is a uniform Cauchy sequence. Theorem 4.15, which stated the completeness of $C([a,b])$, then completes the proof. □

## 8.4 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 $\xi_k$ with*

$$a = x_0 \leq \xi_1 \leq x_1 \leq \xi_2 \cdots \leq \xi_N \leq x_N = b,$$

*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*

$$|J| \leq \int_a^b |f(x)|\, dx.$$

*In particular the statements above apply to the case that $X = \mathbb{C}$.*

**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.

[18] Why?

The playlist

https://www.youtube.com/playlist?list=PLQgy2W8pIli9jyuYN76HM3YdXjwBZrx8L

ends with `https://youtu.be/A5yDujcQyYQ` that you could watch in the context of Theorem 8.15.

**Exercise 8.17.** Not so difficult and useful in Section 28.2 and further: explain why the conclusions in Theorem 8.15 holds for all Riemann integrable real valued functions. Hint: use upper and lower sums to control the sums with the intermediate points. Then prove that the conclusions also holds for $\mathbb{R}^n$-valued functions if the components are Riemann integrable.

## 8.5 A 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 8.20 for open intervals $I$ and $J$ only, the proof relies crucially on Theorem 8.19 below, which has the simple[19] but important statement in Theorem 8.18 as a special case.

**Theorem 8.18.** *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 8.18 can be restated as the *intermediate value theorem*:

**Theorem 8.19.** *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*

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

[19]By now obvious?

*be the image of* $[a,b]$ *under* $f$. *Then*

$$f(a) < f(b) \implies [f(a), f(b)] \subset f([a,b]),$$

*and*

$$f(a) > f(b) \implies [f(b), f(a)] \subset f([a,b]).$$

**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 $\varepsilon$-$\delta$ 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 $\xi$ 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 $\xi$ 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. $\square$

**Theorem 8.20.** *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 8.19. 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 $\varepsilon < \delta_0$, $\delta_0$ 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. $\square$

## 8.6 Exercises

https://www.youtube.com/playlist?list=PLQgy2W8pIli9_T5budfPhqXqhnSVel_KM

NB I update the course notes regularly. The agreement before the course started was NOT to update the file in Canvas. But I suggest you use http://www.few.vu.nl/~jhulshof/2020new.pdf at all times now.

The above playlist discusses some easy examples in which you have to use epsilon-delta arguments. In the exercises below you can of course also use Definition 4.1 and Theorem 4.7. In particular products and sums of continuous functions defined on the same subset of $\mathbb{R}$ or any other metric space are continuous. But it's instructive to do the epsilon-delta proofs. And are the same statements true about uniformly continuous functions? What about $\frac{1}{f}$ if f is such a continuous or uniformly continuous function?

Do have a look at Theorem 5.44 and Remark 5.45 by the way.

**Exercise 8.21.** Let $f(x) = 2x + 1$. Prove directly from the definition that $f$ is uniformly continuous on $\mathbb{R}$.

**Exercise 8.22.** 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 8.23.** 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 8.24.** Let $f : A \to \mathbb{R}$ be Lipschitz continous. Prove that $f$ is uniformly continuous.

Iris asked Sophia about how to do uniform and Lipschitz continuity in the exercises below. Well, you can make use of the properties of the functions exp, cos, sin established in Sections 7.7 and 7.8. In particular you can write $\exp(x) - \exp(a)$, $\cos(x) - \cos(a)$, $\sin(x) - \sin(a)$ as integrals that you can estimate in terms of the maximum of the absolute value of the integrand on the integration interval times the length $|x - a|$ of that interval. But actually these special functions are examples of the power series $p(x)$ treated in Chapter 9.

If I move Chapter 9 to before the present chapter[20] you can estimate differences $p(x) - p(a)$ by factoring out $(x - a)$, first for monomials, e.g.

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

whence

$$|x^7 - a^7| \leq 7r^6\,|x - a| \quad \text{for all} \quad x, a \in [-r, r],$$

Lipschitz continuity of $x \to x^7$ on $[-r, r]$ with Lipschitz constant $7r^6$. Thus Theorem 9.3 is easily extended with the statement that

$$|p(x) - p(a)| \leq \sum_{n=1}^{\infty} n|\alpha_n|r^{n-1}\,|x - a| \quad \text{for all} \quad x, a \in [-r, r], \quad 0 < r < R.$$

Likewise for the Laurent series in Remark 9.9.

**Exercise 8.25.** Let $f : (0, 1] \to \mathbb{R}$ be defined by

$$f(x) = \sin\frac{1}{x}.$$

Show that $f$ is continuous but not uniformly continuous.

**Exercise 8.26.** Is the function $\exp : \mathbb{R} \to \mathbb{R}_+$ defined as the unique integrable solution $f = \exp$ of (7.23) uniformly continuous? Show that $\exp$ is Lipschitz continuous with Lipschitz constant $1$ on $(-\infty, 0]$.

**Exercise 8.27.** Recall (7.28) and $\cos^2 + \sin^2 = 1$. Prove that $\cos$ and $\sin$ are Lipschitz continuous with Lipschitz constant $1$.

**Exercise 8.28.** Show that the function $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} x\sin\frac{1}{x} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

is continuous. Is it uniformly continuous? Is it Lipschitz continuous?

[20]As I already did in the ordering of the YouTube playlists.

**Exercise 8.29.** Is the function $f : \mathbb{R} \to \mathbb{R}$ defined by[21]

$$f(x) = \begin{cases} \frac{\sin x}{x} & \text{if } x > 0 \\ \exp(x) & \text{if } x \leq 0 \end{cases}$$

continuous? Is it uniformly continuous? Is it Lipschitz continuous?

**Exercise 8.30.** Is the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} \sin x \cos \frac{1}{x} & \text{if } x < 0 \\ \exp(x) - 1 & \text{if } x \geq 0 \end{cases}$$

continuous? Is it uniformly continuous? Is it Lipschitz continuous?

**Exercise 8.31.** See `https://youtu.be/xGLEFpSuMqY`. Let $f : [0,1] \to \mathbb{R}$ be defined by $f(x) = \sqrt{x}$. For $n \in \mathbb{N}$ we define $f_n : [0,1] \to \mathbb{R}$ by

$$f_n(x_j) = f(x_j) \quad \text{for} \quad x_j = \frac{j}{n}, \quad j = 0, 1, 2, \ldots, n,$$

and by $f_n$ being linear on every interval $I_j = [\frac{j-1}{n}, \frac{j}{n}]$.

a) Explain why the functions $f_n$ are all Lipschitz continuous.

b) What is the Lipschitz constant of $f_2$? Hint: make a sketch of the graph of $f_2$.

c) Why is $f$ uniformly continuous? Is $f$ Lipschitz continuous?

d) Show that $f_n \to f$ uniformly on $[0,1]$.

e) For $\varepsilon > 0$ let $\delta > 0$ be given by the definition of uniform continuity of $f$, i.e.

$$\forall_{x,y \in [0,1]} : |x - y| < \delta \implies |f(x) - f(y)| < \varepsilon.$$

Let $n \in \mathbb{N}$ satisfy $n > \frac{1}{\delta}$. Prove that

$$|f_n(x) - f(x)| < 2\varepsilon$$

for all $x \in [0,1]$. Hint: given $x \in [0,1]$ use the inequality

$$|f_n(x) - f(x)| \leq |f_n(x) - f(x_j)| + |f(x_j) - f(x)|$$

and choose $j$ such that $x \in I_j$.

[21] See above Exercise 8.25 Iris: $\frac{\sin x}{x} = 1 - \frac{1}{2}x^2 + \cdots$ is a power series just as $\exp(x)$.

**Exercise 8.32.** 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 8.33.** 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 8.32 to define $\bar{f}(a)$ and $\bar{f}(b)$.

**Exercise 8.34.** Recall that Theorem 7.13 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 8.35.** 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 8.20? Specify $J$ and compute $g(y)$. What is $J$ if $I = (0, \infty)$?

**Exercise 8.36.** Formulate Theorem 8.20 for strictly decreasing functions.

**Exercise 8.37.** 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: approximate $F$ uniformly with a sequence of Lipschitz continuous functions and then use both Theorem 7.5 and Theorem 7.13.

**Exercise 8.38.** (continued, an alternative) 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 $\delta^2$. 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[22].

[22] Harold drew my attention to this proof due to Rudin, it relies on Theorem 7.2 only.

**Exercise 8.39.** 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 8.37.

**Exercise 8.40.** 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 8.41.** (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 8.40. 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 7.13.

**Exercise 8.42.** (continued) Let $F_n$ and $F$ be as in Exercise 8.41. 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: in Exercise 7.34 you showed 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, *not* an assumption on $F$ here.

**Exercise 8.43.** (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 8.44.** (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 8.43.

# 9 Differential calculus for power series

https://www.youtube.com/playlist?list=PLQgy2W8pIli_IRJbP205fsUsBTIx1vNEk

If you came here from Section 1.5 and skipped the epsilons: don't worry. You most likely will be familiar with the formula

$$f'(a) = \lim_{h\to 0} \frac{f(a+h)-f(a)}{h} = \lim_{x\to a} \frac{f(x)-f(a)}{x-a}, \tag{9.1}$$

which you may have taken for granted and ignored. Good. It's just the usual default 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.

In case the *limit* of the *difference quotient* in (9.1) exists it is called the *differential quotient* of $f$ in $x = a$. Such differential quotients are sometimes formally denoted as fractions with 'numerator' $df$ and 'denominator' $dx$, just like difference quotients are denoted as[1]

$$\frac{f(x+\Delta x)-f(x)}{\Delta x} = \frac{\Delta f}{\Delta x}$$

with $\Delta x = h \neq 0$. Notations to be handled with care or simply avoided perhaps, just like them limits. But do note that for the simplest examples to consider first, *monomials* such as

$$f_{42}(x) = x^{42}$$

for instance, the difference quotient is naturally defined for $x = a$ as well.

Indeed, for $x \neq a$ it holds that

$$\frac{x^{42}-a^{42}}{x-a} = x^{41} \underbrace{+ \quad \cdots\cdots \quad +}_{\text{Exercise 1.12!}} a^{41},$$

but the right hand side is clearly equal to $42a^{41}$ for $x = a$. Whatever the value of $a$, it must[2] thus follow that $f_{42}$ is differentiable in $a$, with

$$f_{42}'(a) = 42a^{41}.$$

In what follows we will do better[3] to avoid limits of difference quotients altogether and think of *differentiation as a method to best*[4] *approximate a given (nonlinear) function* $f$ by a linear one, i.e. to write

$$f(x) \approx f(a) + A(x-a) = Ax + B,$$

[1]Used to convince you to write $df = f'(x)dx$.
[2]Parafrasing Jaap Murre: who in his right mind would need a limit concept here?
[3]Meaning: let us see how far we can get without them epsilons.
[4]In some appropriate sense.

and choose $A$ and $B$ to make

$$R_a(x) = f(x) - Ax - B$$

as small as possible near $x = a$.

Since the obvious[5] choice for $B$ is

$$B = f(a) - Aa,$$

it remains to identify the best number $A$ to choose in

$$f(x) = f(a) + A(x-a) + R_a(x). \tag{9.2}$$

Thus we look for the *unique value* of $A$ for which the remainder term $R_a(x)$ has a suitably formulated *smallness property* that fails for $A$.

## 9.1 Linear approximations of monomials

Consider a *difference quotient* for the function $f_7$ defined by $f_7(x) = x^7$. A little algebra[6] 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[7]

$$x^7 = a^7 + (x^6 + ax^5 + a^2x^4 + a^3x^3 + a^4x^2 + a^5x + a^6)(x-a) = \tag{9.3}$$

$$\underbrace{a^7 + 7a^6\,(x-a)}_{Ax+B} + \underbrace{(x^5 + 2ax^4 + 3a^2x^3 + 4a^3x^2 + 5a^4x + 6a^5)\,(x-a)^2}_{\text{remainder term}}.$$

The *particular choice*

$$A = 7a^6, \quad B = -6a^7 \tag{9.4}$$

followed from putting $x = a$ in the 7 terms of the[8] prefactor in the second term on the right hand side of (9.3). 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

$$Ax + B = 7a^6x - 6a^7$$

[5] Why? Takes $x = a$ to convince to yourself.

[6] Long division for instance.

[7] See Exercise 1.18.

[8] 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*[9] which makes the *resulting remainder term*[10] contain a factor $(x-a)^2$.

Moreover, the prefactor in the remainder term under (9.3) is easily estimated if we assume that $x$ and $a$ are contained in a fixed interval $[-r, r]$. For example, if

$$|x| \le r \quad \text{and} \quad |a| \le r,$$

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 (9.3) 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*

$$|R_{an}(x)| \le \underbrace{\frac{n(n-1)}{2} r^{n-2}}_{r\text{-dependent constant}} (x-a)^2$$

*for all $x, a \in [-r, r]$.*

**Exercise 9.2.** You may guess a nice expression for $R_{an}(x)$ from (9.3). 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

$$p_k(x) = \sum_{n=0}^{k} \alpha_n x^n = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_k x^k$$

[9] Of course both $A$ and $B$ depend on $a$.

[10] 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 $\alpha_n$ and take the sum over $n$. With some care for $n = 0, 1, 2$ it follows that

$$p_k(x) = p_k(a) + \underbrace{\sum_{n=1}^{k} n\alpha_n a^{n-1}\,(x-a)}_{\text{linear approximation}} + \underbrace{\sum_{n=2}^{k} \alpha_n R_{an}(x)}_{\text{remainder term}}, \tag{9.5}$$

in which for all $x, a \in [-r, r]$ the remainder term satisfies

$$\underbrace{|\sum_{n=2}^{k} \alpha_n R_{an}(x)\,|}_{\text{remainder term}} \leq \underbrace{\sum_{n=2}^{k} |\alpha_n| \frac{n(n-1)}{2} r^{n-2}}_{r\text{-dependent constant}} (x-a)^2. \tag{9.6}$$

As before

$$p_k(a) + \underbrace{\sum_{n=1}^{k} n\alpha_n a^{n-1}}_{p_k'(a)}\,(x-a)$$

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

$$p(x) = 1 + 2x + 3x^2 + \cdots \tag{9.7}$$

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* $\{x \in \mathbb{R} : |x| \le r\}$ *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 (9.6). If for some $r > 0$ it holds that

$$C_r := \sum_{n=2}^{\infty} |\alpha_n| \frac{n(n-1)}{2} r^{n-2} < \infty, \tag{9.8}$$

we can let $k \to \infty$ in (9.5). Indeed, it then follows from Exercises 3.73 and 3.75 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

$$1 \le n \le \frac{n(n-1)}{2}$$

for $n \ge 2$, and so does the sum

$$R_a(x) = \sum_{n=2}^{\infty} \alpha_n R_{an}(x).$$

Thus (9.8) allows to take the limit $k \to \infty$ in (9.5) and (9.6) to obtain[11]

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n = p(a) + \underbrace{\sum_{n=1}^{\infty} n\alpha_n a^{n-1}}_{A} (x - a) + R_a(x) \tag{9.9}$$

[11]The convergence is in fact uniform on $[-r, r]$, why?

with

$$|R_a(x)| \le C_r(x-a)^2 \tag{9.10}$$

for all $x, a \in [-r, r]$. As before we observe that

$$A = \sum_{n=1}^{\infty} n\alpha_n a^{n-1} \tag{9.11}$$

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 (9.10) 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 (9.10) suffices to conclude from (9.8,9.9) that

$$\lim_{x\to a} \frac{p(x)-p(a)}{x-a} = A \tag{9.12}$$

as given by (9.11).

To conclude we note that the $r$-values for which (9.8) holds form an interval

$$\{r \ge 0 : \sum_{n=1}^{\infty} n^2 |\alpha_n| r^n < \infty\}$$

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 (9.12) is equivalent to saying that*

$$\lim_{x \to a} \frac{R_a(x)}{x-a} = 0. \tag{9.13}$$

*This means that for every $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$|R_a(x)| < \varepsilon |x-a| \quad \text{if} \quad |x-a| < \delta, \tag{9.14}$$

*a statement much weaker than the statement in (9.10). 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 Taylor series for power series

We substitute $x = x_0 + h$ in

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n. \tag{9.15}$$

Changing the order of summation[12] we find

$$\begin{aligned} p(x_0 + h) &= \sum_{n=0}^{\infty} \alpha_n (x_0 + h)^n = \sum_{n=0}^{\infty} \alpha_n \sum_{k=0}^{n} \binom{n}{k} x_0^{n-k} h^k \\ &= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \alpha_n \frac{n(n-1)\dots(n-k+1)}{k!} x_0^{n-k} h^k \end{aligned}$$

[12] This section relies on Section 3.9 but we will not expand on this here.

$$= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \alpha_n \frac{n(n-1)\dots(n-k+1)}{k!} x_0^{n-k} h^k$$

$$= \sum_{k=0}^{\infty} \frac{1}{k!} \underbrace{\sum_{n=k}^{\infty} \alpha_n n(n-1)\dots(n-k+1)\, x_0^{n-k}\, h^k}_{p^{(k)}(x_0)}$$

$$= \sum_{k=0}^{\infty} \frac{p^{(k)}(x_0)}{k!} h^k,$$

i.e.

$$p(x) = p(x_0 + h) = \sum_{n=0}^{\infty} \frac{p^{(n)}(x_0)}{n!} h^n = \sum_{n=0}^{\infty} \frac{p^{(n)}(x_0)}{n!} (x - x_0)^n. \tag{9.16}$$

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 (9.16) 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 3.9.

**Remark 9.9.** *Everything we did for the differentiation of power series in (9.17) also works for (Laurent series)*

$$L(x) = \sum_{n=-\infty}^{\infty} \alpha_n x^n = \dots + \frac{\alpha_{-2}}{x^2} + \frac{\alpha_{-1}}{x} + \alpha_0 + \alpha x + \alpha_2 x^2 + \cdots,$$

*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 $R_a(x)$ out and you're in business:* `https://youtu.be/_3TSrRd8ils`[13].

[13] Forgot to put this one in Playlist 5.

## 9.5 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 7.9 this restriction on $a$ and $b$ disappears:

**Exercise 9.10.** Verify that (9.18) holds for all $n \in \mathbb{N}$ and any $a, b \in \mathbb{R}$.

Theorem 7.18 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
$$|p_k(x) - p(x)| = |\sum_{n=k+1}^{\infty} \alpha_n x^n| \leq \sum_{n=k+1}^{\infty} |\alpha_n x^n| \leq \sum_{n=k+1}^{\infty} |\alpha_n| r^n,$$

provided $[a,b] \subset [-r,r]$. It follows that $p_k \to p$ in $C([a,b])$ and thus by Theorem 7.13 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[14] for integration variable $x \in [a,b] \subset (-R,R)$.

**Theorem 9.11.** *If $\alpha_n$ 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 $[a,b] \subset (-R,R)$.*

**Exercise 9.12.** Finish the proof of Theorem 7.13. 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 $R \in \mathbb{R}_+$.

## 9.6 Power series solutions of differential equations

We can solve linear *ordinary differential equations* (ODEs) for *power series* (9.15), for instance

$$p'(x) = p(x), \tag{9.26}$$

with boundary condition $p(0) = 1$. Let us try to find a solution of the form

$$p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \cdots,$$

[14] This is really Theorem 9.3 if you think about it.

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

$$\alpha_1 = 1,\ \alpha_2 = \frac{1}{2},\ \alpha_3 = \frac{1}{2}\frac{1}{3},\ \alpha_4 = \frac{1}{2}\frac{1}{3}\frac{1}{4},\ \cdots, \alpha_n = \frac{1}{n!},$$

so we encounter the exp you knew from Exercise 6.23 and Section 7.7, if you did not come here directly from the end of Section 1.5. Ask yourself what is really needed for the following theorem.

**Theorem 9.13.** *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 9.14.** Prove that $\exp(x)$ has $R = \infty$ and you have solved your first differential equation[15]. Hint: show for $N \in \mathbb{N}$ that

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^N}{N!} + R_N(x),$$

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

$$|R_N(x)| \leq \frac{|x|^N}{N!}\left(\frac{|x|}{N+1} + \left(\frac{|x|}{N+1}\right)^2 + \left(\frac{|x|}{N+1}\right)^3 + \cdots\right) = \frac{|x|^N}{N!}\frac{|x|}{N+1-|x|}$$

if $N + 1 > |x|$.

[15]No other functions can satisfy $f(0) = 1$ and $f'(x) = f(x)$, why is not clear yet.

**Definition 9.15.** *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 9.16.** 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 $x \to \infty$.

**Exercise 9.17.** 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 $\alpha \in \mathbb{N}_0$, that is, you should discover that

$$(1 + x)^\alpha = 1 + \alpha x + \alpha(\alpha - 1)\frac{x^2}{2} + \alpha(\alpha - 1)(\alpha - 2)\frac{x^2}{2 \times 3} + \cdots , \tag{9.27}$$

**Exercise 9.18.** (continued) Explain why this holds for all $x$ with $|x| < 1$ if the series does not terminate, which it only does if $\alpha \in \mathbb{N}_0$. Write the first terms of the series out for $\alpha = \frac{1}{2}, -\frac{1}{2}, -1$, also for $x$ replaced by $-x$.

**Exercise 9.19.** Use (9.27) with $\alpha = \frac{1}{2}$ to improve the Babylonian trick

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

in Section 1.1. Take $r = \frac{3}{2}$.

**Theorem 9.20.** *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*

$$\begin{cases} p''(x) + p(x) = 0 & \textit{for all } x \in \mathbb{R}; \\ p(0) = 0 \ \textit{ and } \ p'(0) = 1 & \textit{in } x = 0. \end{cases}$$

**Exercise 9.21.** Write $p(x)$ using the sum notation and prove Theorem 9.20. Let $\cos x = p'(x)$. What is the derivative[16] of cos?

**Exercise 9.22.** Use power series to compute the limits for $x \to 0$ of

$$\frac{\exp(x) - 1}{\sin x}; \quad \frac{1 - \cos x^3}{(x - \sin x)^2}; \quad \frac{\exp(x^2) - \cos x}{\sin x^2}; \quad \frac{\sqrt{1-x} - 1}{\sin x}.$$

**Exercise 9.23.** 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 $q(x)$. 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 the minimum of $R$ and $S$.
Hint: use Section 3.9.

[16] Do compare Theorem 9.20 to what we found in Section 7.8.

# 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.

The presentation in

`https://www.youtube.com/playlist?list=PLQgy2W8pIli-mMFVTcMcKO5Zy2uJU193Y`

combines the algebraic and the general approach: in that playlist I first do differentiation of monomials and power series in $x = a$ and then the general approach in $x = 0$ only, assuming $f : \mathbb{R} \to \mathbb{R}$ for the latter.

In `https://youtu.be/SmwYBRxNgyI` I then first explain why we cannot restrict to remainder terms which are quadratic, such as for instance the remainder term in Theorem 9.1. Thus it's analysis again in this chapter. *The smallness statement for the remainder term next may very well be the first essential example of an epsilon-delta statement*[2]*. To compare to (8.1) let $\xi$ be a fixed interior point of $D_f$, $f : D_f \to \mathbb{R}$, $A \in \mathbb{R}$, and write*

$$f(x) = f(\xi) + A(x - \xi) + R(x, \xi)$$

*for all $x \in D_f$. If the $\varepsilon - \delta$ statement*

$$\forall_{\varepsilon>0}\,\exists_{\delta>0} : |x - \xi| < \delta \implies |R(x,\xi)| < \varepsilon\,|x - \xi|$$

*holds, then the function $f$ is called* differentiable *in $\xi$ with derivative $f'(\xi) = A$. Recall that the continuity statement (8.1) for $f$ in $\xi$ is*

$$\forall_{\varepsilon>0}\,\exists_{\delta>0} : |x - \xi| < \delta \implies |f(x) - f(\xi)| < \varepsilon$$

In what follows the point in which we consider differentiability is called $x_0$ and $x$ is written as $x = x_0 + h$, the most common notation perhaps. This notation was used in Section 9.4, whereas the first part of Chapter 9 was with $x$ and $x_0 = a$. The proofs in Chapter 11 are perhaps most transparent with $x_0 = 0$, $h = x$, and also $f(0) = 0$.

[1] We kick the habit of writing $A$ for $D_f$ that started in Definition 3.3.

[2] Except perhaps for `https://youtu.be/NvVFG53Wq7Q`, uniform continuity!

**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[3]. *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 $A_0 = B_0$. $\square$

**Exercise 10.3.** Give the $\varepsilon$-$\delta$ 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 $x_0$.

[3] Not to be confused with big O of $h$.

## 10.1 Critical points and the mean value theorem

This is https://youtu.be/rmnsPnwVmhs. A critical point[4] 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 $\xi$.*

**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 $\xi$ 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 $\xi$ strictly between $a$ and $b$.*

**Proof**. In the special case that $f(a) = f(b)$ the point $\xi$ 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 $\xi$ 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 $f(a) = f(b)$. 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[5]. You will complete the proof of Theorem 10.7 in Exercise 10.8 by reduction of the general case to this special case. $\square$

[4] Also: a stationary point.

[5] Read about Rolle and his theorem in wikipidia.

**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

You may like to watch the last 5 videos of the epsilon-delta playlist

https://www.youtube.com/playlist?list=PLQgy2W8pIli8e9Hm34hqKFtil6pZg4I6z

from https://youtu.be/y54etUS4HZI to https://youtu.be/3_Jub7iFnxc

for the full story completed here. 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 4.44.

**Exercise 10.9.** Let $f$ be the bounded nondecreasing function in Exercise 4.44 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 Lipschitz continuous on $[a, b]$ with Lipschitz constant $|f|_\infty$, and differentiable in every $x_0 \in [a, b]$ where $f$ is continuous, with derivative $F'(x_0) = f(x_0)$.*

Note that $x_0$ is also allowed to be one of the boundary points, for which case the obvious one-sided statement[6] that $F$ is differentiable was not given yet.

[6] Formulate this statement for $x_0 = a$ and $x_0 = b$.

**Proof**. Lipschitz continuity is immediate from Exercise 7.7. Next we 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

$$|R_0(h)| \le |\int_{x_0}^{x_0+h} |f(s) - f(x_0)|\,ds| < \varepsilon|h| \quad \text{if} \quad 0 < |h| \le \delta \tag{10.4}$$

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 $x \in [a, b]$, then $f(x) = F'(x)$ defines a function $f : [a, b] \to \mathbb{R}$ called the derivative of the function $F$, and $F$ is called a* primitive function[7] *of $f$.*

Once we know Theorem 8.6, Theorem 10.10 says that every continuous function $f : [a, b] \to \mathbb{R}$ has a primitive on $[a, b]$. For this particular primitive we have that[8]

$$\int_a^b f(x)\,dx = F(b) - F(a), \tag{10.5}$$

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.

[7] Or anti-derivative.

[8] Have a look at Exercise 6.20 again.

**Theorem 10.12.** *The fundamental theorem of integral calculus: for every* $f \in C([a,b]) \cap \mathrm{RI}([a,b]) = C([a,b])$ *it holds that*[9]

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

*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 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 $t \in [0,T]$.

**Definition 10.14.** *The set of functions* $F$ *as in Theorem 10.12 is denoted by* $C^1([a,b])$*. The equivalence*

$$F \in C^1([a,b]) \iff F' \in C([a,b])$$

*characterises this function space. In a similar fashion we define* $C^2([a,b])$ *by*

$$F \in C^2([a,b]) \iff F' \in C^1([a,b])$$

*and so on for* $C^3([a,b])$, $C^4([a,b])$, $\ldots$, *and thus any* $C^k([a,b])$ *with* $k \in \mathbb{N}$*.*

**Exercise 10.15.** Prove that $C^1([a,b])$ is complete with the metric that comes from the norm defined by $|F| = \max(|F|_{\max}, |F'|_{\max})$. Which norm[10] makes $C^2([a,b])$ complete? And $C^k([a,b])$ with $k \in \mathbb{N}$? The intersection of all $C^k([a,b])$ is denoted by $C^\infty([a,b])$. What goes wrong when you try to define a norm on $C^\infty([a,b])$?

[9] Ignore the red part if you know Theorem 8.6.

[10] We can replace $[a,b]$ by any interval $I$ but then we lose the maximum norm.

## 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

$$F(x)\Big|_a^b = [F(x)]_a^b = F(b) - F(a).$$

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[11], and so on.

The algebra with forms will be defined later to mimic natural operations in multivariate integral calculus, and will be based on the formal rules[12] $dxdy + dydx = 0$, $ddx = 0$, and a Leibniz type rule, see Chapter 27 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

$$\Big|_a^b \quad \text{acting on} \quad F(x),$$

an interaction between "integrals" and differential forms.

## 10.4 Exercises

**Exercise 10.16.** Recall that $\cos$ and $\sin$ are integrable functions satisfying (7.28) and $\cos^2 + \sin^2 = 1$. Use Theorem 10.10 to prove that $\cos$ and $\sin$ are differentiable, and that $\sin' = \cos$, $\cos' = -\sin$.

**Exercise 10.17.** Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} 0 & \text{for } x = 0 \\ x^2 \sin \frac{1}{x} & \text{for } x \neq 0 \end{cases}$$

[11] Usually witten as $u(x,y)dx \wedge dy$.

[12] Recall from Definition 7.9 that we think of $dx$ and thus also $dy$ as having a sign.

Prove that $f$ is differentiable in $x = 0$. That is: guess the linear approximation of $f(x)$ near $x = 0$ and verify the $\varepsilon$-$\delta$ statement for the remainder term. Specify $\delta > 0$ for given $\varepsilon > 0$.

**Exercise 10.18.** Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} 0 & \text{for } x = 0 \\ x(1 + \sqrt{|x|} \sin \frac{1}{x}) & \text{for } x \neq 0 \end{cases}$$

Prove that $f$ is differentiable in $x = 0$.

HInt: first guess the linear approximation of $f(x)$ near $x = 0$.

**Exercise 10.19.** Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} 0 & \text{for } x = 0 \\ x(42 + |x|^{\frac{1}{41}} \sin \frac{1}{x^{43}}) & \text{for } x \neq 0 \end{cases}$$

Prove that $f$ is differentiable in $x = 0$.

**Exercise 10.20.** Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} 1 & \text{for } x = \frac{1}{2} \\ 2x + (2x-1)^2 \cos \frac{x}{2x-1} & \text{for } x \neq \frac{1}{2} \end{cases}$$

Prove that $f$ is differentiable in $x = \frac{1}{2}$.

Hint: what would the value of $f'(\frac{1}{2})$ be?

**Exercise 10.21.** In Exercise 4.42 prove equality in Part (a).

**Exercise 10.22.** 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.

**Exercise 10.23.** Show there exists $f : \mathbb{R} \to \mathbb{R}$ discontinuous in every $x \neq 0$ but differentiable in $x = 0$.

# 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 13 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

This is `https://youtu.be/kOxUBrTVpZg`, where I restrict to $x_0 = 0$ and $f(x) = ax + R(x)$. 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 11.4 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[3], 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**[4]. 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

$$f(x) = f(x_0) + \underbrace{A_0}_{f'(x_0)}(x - x_0) + R_0(x).$$

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 (11.27) for $f$,

$$\forall_{\varepsilon>0}\, \exists_{\delta>0} \quad 0 < |x - x_0| < \delta \implies |R_0(x)| < \varepsilon |x - x_0|,$$

[3] In https://youtu.be/c6QqDazAoAs I restrict to $x_0 = 0$ and $f(x) = ax + R(x)$.

[4] 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 $\varepsilon$ we started with.

If by chance $y=y_0$ then $S(y)=0$ and we're done. If not, then we need

$$0<|y-y_0|<\delta$$

if we want to conclude via (11.9). We actually have

$$\begin{aligned}|y-y_0|=|f(x)-f(x_0)|=|A_0(x-x_0)+R_0(x)|&\le|A_0|\,|x-x_0|+|R_0(x)|\\&<(|A_0|+1)\,|x-x_0| \quad\text{if}\quad 0<|x-x_0|<\delta_R,\end{aligned}$$

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. $\square$

## 11.3 Differentiability of inverse functions

Consider the functions $f$ and $g$ in Theorem 8.20. 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[5] for $-\delta < x < 0$.

[5] With the same $\delta$ 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 8.20 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 9.13 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 $x^\alpha$ 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.** On $(0, \pi)$ consider

$$x \to \cos x.$$

Show for the inverse that $\arccos y + \arcsin y$ is constant on $(-1, 1)$. Which constant?

**Exercise 11.10.** Solve the differential equation in Exercise 9.17 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 11.11.** 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 11.12.** Write out the first few terms of

$$\sqrt[n]{1+x} = 1 + \frac{x}{n} + \cdots \quad \text{and} \quad \frac{1}{\sqrt[n]{1+x}} = 1 - \frac{x}{n} + \cdots$$

## 11.4 Differentiation in normed spaces

In fact we may just as well speak about $D_f \subset X$, $X$ a normed space[6], $x_0$ 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[7] continuous[8]. The $\varepsilon$-$\delta$ statement (10.2) then becomes

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\,\forall_{h\in X} : \quad 0 < |h|_X < \delta \implies |R(h)| < \varepsilon|h|.$$

[6] In https://youtu.be/SmwYBRxNgyI I restrict to $x_0 = 0$ and $f(x) = ax + R(x)$.
[7] If $\phi : X \to \mathbb{R}$ is linear an continuous in 0 then it is Lipschitz continuous.
[8] In order to have $f$ differentiable in $x_0$ imply that $f$ is continuous in $x_0$, explain!

It implies that such $\phi_0$, 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 $\Phi$ 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[9] of $A_0(h)$ for $A_0 : X \to Y$ linear and Lipschitz[10] continuous. Definition 10.1 and Theorem 10.2 are just special cases of the following definition and theorem.

**Definition 11.13.** *Let $X, Y$ be real normed spaces, $D_\Phi \subset X$, $\Phi : D_\Phi \to Y$, $x_0$ an interior point of $D_\Phi$ and let $A_0 : X \to Y$ be* Lipschitz continuous. *Then*

$$\Phi(x_0 + h) = \Phi(x_0) + A_0 h + R_0(h) \tag{11.21}$$

*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*

$$|R_0(h)|_Y < \varepsilon |h|_X \quad \text{if} \quad 0 < |h|_X < \delta. \tag{11.22}$$

*If so then the map $\Phi$ is called* differentiable *in $x_0$.*

**Theorem 11.14.** *Let $X, Y$ be real normed spaces, $D_\Phi \subset X$, $\Phi : D_\Phi \to Y$, $x_0$ an interior point of $D_\Phi$, 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 11.13 holds, and $\Phi'(x_0) = A_0$ is called* the derivative *of $\Phi$ in $x_0$.*

**Remark 11.15.** *The space of all Lipschitz continuous linear maps $A$ from $X$ to $Y$ that qualify to be used in Definition 11.13 is denoted by $L(X, Y)$. We shall write*

$$|A|_{L(X,Y)} \tag{11.23}$$

*for the best (smallest) Lipschitz constant of such an $A$. It is commonly called the (operator) norm of $A$.*

**Theorem 11.16.** *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 $f : \mathcal{O} \to \mathrm{I\!R}$ differentiable. Then there exists*

$$\xi \in (x, y) = \{(1 - t)x + ty;\ 0 < t < 1\}$$

*such that*

$$f(y) - f(x) = f'(\xi)(y - x).$$

[9] Another standard notation is $\langle A_0, h\rangle$, see (17.59).

[10] Again: if $A_0 : X \to Y$ is linear an continuous in 0 then it is Lipschitz continuous.

**Exercise 11.17.** Give a direct proof that[11]

$$t \to f(\xi(t)) \tag{11.24}$$

is differentiable on $[0,1]$. Then use Theorem 10.7 to prove Theorem 11.16. 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 11.13 and Theorem 11.14. For the general Leibniz rule we suppose $\Phi$ and $\Psi$ map to a normed algebra $Y$ and are as in Definition 11.13 and Theorem 11.14. 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

$$(\Phi\Psi)'(x_0) = \Psi(x_0)A_0 + \Phi(x_0)B_0 = \Psi(x_0)\Phi'(x_0) + \Phi(x_0)\Psi'(x_0). \tag{11.25}$$

If multiplication in $Y$ is not commutative we have that

$$((\Phi\Psi)'(x_0))(h) = (\Phi'(x_0)(h))\Psi(x_0) + \Phi(x_0)(\Psi'(x_0)(h)) \tag{11.26}$$

defines $(\Phi\Psi)'(x_0)$. It is Lipschitz continuous because, using $|yz|_Y \le |y|_Y\,|z|_Y$ and recalling (11.23), 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

$$R_0(x)S_0(x).$$

[11] You really don't need the general chain rule in Theorem 11.4 to do so.

**Exercise 11.18.** Prove in the general setting of normed spaces $X$ and $Y$ that (11.6) holds. That is, use

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\,\forall_{x\in X}\quad 0<|x-x_0|_X<\delta \implies |R_0(x)|_Y<\varepsilon|x-x_0|_X \tag{11.27}$$

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 11.19.** 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 11.20.** *Exercise 11.19 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 11.21 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 $\Phi$ in $x_0$ and $\Psi$ in $y_0$, namely (11.27) for $\Phi$,

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\quad 0<|x-x_0|_X<\delta \implies |R_0(x)|_Y<\varepsilon|x-x_0|_X,$$

and

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\,0<|y-y_0|_Y<\delta \implies |S_0(y)|_Z<\varepsilon|y-y_0|_Y \tag{11.28}$$

for $\Psi$. 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

$$\Psi(\Phi(x)) = \Psi(y) = \Psi(y_0) + B_0(\underbrace{\Phi(x) - \Phi(x_0)}_{y-y_0}) + S_0(y) =$$

$$\Psi(y_0) + B_0A_0(x - x_0) + \underbrace{B_0R_0(x) + S_0(y)}_{T_0(x)},$$

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*

$$|B_0R_0(x)|_Z \leq |B_0|_{L(Y,Z)} |R_0(x)|_Y .$$

The rest of the proof is copy-paste from the proof for $X = Y = Z = \mathbb{R}$, with $f$ and $g$ replaced by $\Phi$ and $\Psi$, 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 11.21.** *Let $x_0$ be an interior point of the domain of $\Phi$, assume $\Phi$ differentiable in $x_0$, let $y_0 = \Phi(x_0)$ be an interior point of the domain of $\Psi$, and assume that $\Psi$ 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*

$$(\Psi \circ \Phi)'(x_0) = \Psi'(y_0)\Phi'(x_0). \tag{11.29}$$

## 11.5 Exercises

**Exercise 11.22.** Derive and prove the differentiation rules for $fg$ and $\frac{g}{f}$ if $f$ and $g$ are real valued functions from Exercise 11.19 and Theorem 11.4. Hint: use also $y \to \frac{1}{y}$.

**Exercise 11.23.** Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(0) = 0$ and

$$f(x) = x^2 \sin \frac{1}{x^2} \quad \text{for} \quad x \neq 0.$$

Show that $f$ is differentiable in every $x \in \mathbb{R}$. Is $f$ Lipschitz continuous on $[0, 1]$?

Hint: is $f'(x)$ bounded on $[0, 1]$?

**Exercise 11.24.** Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(0) = 0$ and

$$f(x) = x \sin \frac{1}{x}$$

for $x \neq 0$. Is $f$ is differentiable? Is $f$ Lipschitz continuous?

**Exercise 11.25.** Prove that the function $\exp$ defined as the unique integrable solution $f = \exp$ of (7.23) is differentiable in the sense of Definition 10.1. Verify that the value $f'(x_0)$ of the derivative in Theorem 10.2 is $f(x_0)$, to conclude that $f' = f$.

Hint: use $\exp$ is continuous[12].

**Exercise 11.26.** 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. We say that $f$ belongs to $C^\infty(\mathbb{R})$.

---

[12] Exercise 8.26.

## 11.6 Exam May 29, 2020

**Problem 1.** Let $F : \mathbb{R} \to \mathbb{R}$ be defined by

$$F(x) = \frac{x}{1+x}$$

for all $x \neq -1$ and $F(-1) = 1$. This $F$ can next earn you

**a + b + c + d = 1 + 2 + 3 + 3 = 9 points.**

a) *Sketch the graph of* $F$. Make sure you got it right for $x \geq 0$. Which interval is $\{F(x) : x \geq 0\}$?

b) Factorise $F(x) - F(y)$ and *prove* that $F$ is Lipschitz continuous on $[0, \infty)$ with Lipschitz constant 1.

Consider the integral equation

$$f(x) = 1 + \int_0^x \frac{f(s)}{(1+s)^2(1+f(s))}\, ds \quad \text{posed for all} \quad x \in [0,1] \quad (11.30)$$

and denote the right hand side of (11.30) by $(\Phi(f))(x)$. This defines a map $\Phi : A \to A$, where

$$A = \{f \in C([0,1]) : f(x) \geq 0 \text{ for all } x \in [0,1]\}$$

is the subset of nonnegative functions in $C([0,1])$.

c) *Prove that* $\Phi$ *is a contraction.* Hint: use b), in your estimates you may use that

$$\int_0^1 \frac{1}{(1+s)^2}\, ds < \int_0^\infty \frac{1}{(1+s)^2}\, ds = 1.$$

d) *Explain why it follows* that (11.30) has a unique *positive* solution $f$.

Answers NB This is the exercise in which you had to apply Theorem 5.14 to (a closed subset of) a complete metric space (of continuous functions) to solve an integral equation.

a) Near $x = 0$ the graph looks like $y = x$, for $|x|$ large like $y = 1$, and $\{F(x) : x \geq 0\} = [0, 1)$.

b) For $x, y \geq 0$ we have

$$F(x) - F(y) = \frac{x}{1+x} - \frac{y}{1+y} = \frac{x(1+y)-(1+x)y}{(1+x)(1+y)} = \frac{x-y}{(1+x)(1+y)},$$

with denominator $\geq 1$ for $x, y \geq 0$, so $|F(x) - F(y)| \leq |x - y|$.

c) Let $f \in A$. Then

$$(\Phi(f))(x) = 1 + \int_0^x \frac{f(s)}{(1+s)^2(1+f(s))}\,ds$$

exists for every $x \in [0,1]$ as one plus the nonnegative integral of a continuous nonnegative function.

As a function of $x$ the new function $\Phi(f) : [0,1] \to [1,\infty)$ is continuous because $\int_0^x \phi$ is (Lipschitz) continuous in $x$ for every integrable $\phi : [0,1] \to \mathbb{R}$. Thus $\Phi$ maps $A$ to $A$.

To see if $\Phi$ is a contraction let $f, g \in A$ and write

$$(\Phi(f) - \Phi(g))(x) = (\Phi(f))(x) - (\Phi(g))(x) =$$

$$1 + \int_0^x \frac{f(s)}{(1+s)^2(1+f(s))}\,ds - 1 - \int_0^x \frac{g(s)}{(1+s)^2(1+g(s))}\,ds =$$

$$\int_0^x \frac{1}{(1+s)^2} \underbrace{\left(\frac{f(s)}{1+f(s)} - \frac{g(s)}{1+g(s)}\right)}_{F(f(s))-F(g(s))} ds.$$

Using b) we estimate

$$|(\Phi(f) - \Phi(g))(x)| \leq \int_0^x \frac{1}{(1+s)^2}|F(f(s)) - F(g(s))|\,ds$$

$$\leq \int_0^x \frac{1}{(1+s)^2}|f(s) - g(s)|\,ds \leq \int_0^x \frac{1}{(1+s)^2}\,ds\,||f-g||_{max}$$

$$\leq \int_0^1 \frac{1}{(1+s)^2}\,ds\,||f-g||_{max}$$

for all $x \in [0,1]$, so

$$||\Phi(f) - \Phi(g)||_{max} \leq \underbrace{\int_0^1 \frac{1}{(1+s)^2}\,ds}_{=\frac{1}{2}}\,||f-g||_{max}.$$

This holds for all $f, g$ in $A$, so indeed $\Phi : A \to A$ is a contraction, with contraction factor $\frac{1}{2}$.

d) By definition of $\Phi$ the nonnegative solutions of (11.30) are precisely the fixed points of $\Phi : A \to A$.

The set $A$ is closed in $C([0,1])$, $C([0,1])$ is complete with metric defined by $d(f,g) = ||f-g||_{max}$, and $\Phi : A \to A$ is a contraction.

So $\Phi$ has a unique fixed point in $A$, and thereby there exists a unique solution in $A$.

**Problem 2.** Let $F$ be the function defined in Problem 1. This same $F$ can next earn you another

$$\mathbf{a + b + c = 2 + 3 + 3 = 8 \ points}$$

a) *Prove* that $F$ is discontinuous in $x = -1$.

b) Let $\mathbb{R}_+ = \{x \in \mathbb{R} : x > 0\}$. For every $n \in \mathbb{N}$ we define $f_n : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$f_n(x) = \frac{nx}{1+nx} \quad \text{for all} \quad x \in \mathbb{R}_+.$$

*Prove* that $f_n(x) \to 1$ as $n \to \infty$ for every $x \in \mathbb{R}_+$.

c) *Prove* that the convergence in b) is uniform on $[1, \infty)$.
Hint: estimate $|f_n(x) - 1|$ for all $x \in [1, \infty)$ simultaneously.

Answers NB This is the exercise in which you had to use epsilon arguments.

a) For $x \neq -1$ we have

$$F(x) - F(-1) = \frac{x}{1+x} - 1 = -\frac{1}{1+x}$$

which has a denominator that is small for $x$ close to $-1$, while for continuity of $F$ on $-1$ it should be smaller than $\varepsilon$ on an $\varepsilon$-dependent interval around $-1$. That's not going to work. For instance, we have

$$|x+1| < \frac{1}{2} \implies \frac{1}{|x+1|} > 2.$$

It follows that the epsilon-delta statement for continuity in $x = -1$ fails with $\varepsilon = 2$, because for every $\delta > 0$ we can choose $x$ with $|x - -1| < \delta$ as well as $|x - -1| < \frac{1}{2}$, and for such $x$ the inequality $|F(x) - F(-1)| < 2$ fails.

b) Fix $x > 0$. The limit 1 is given, so let $\varepsilon > 0$. We have to prove the existence of an $N \in \mathbb{N}$ such that

$$|f_n(x) - 1| = |\frac{nx}{1+nx} - 1| = \underbrace{\frac{1}{1+nx} < \varepsilon}_{\text{needed for convergence}}$$

for all $n \geq N$. The first reasoning I discussed was to rewrite the inequality needed for convergence. It is equivalent to

$$1 + nx > \frac{1}{\varepsilon} \iff n > \frac{1-\varepsilon}{\varepsilon x},$$

and once this holds for some $n = N$ then it also holds for all $n \geq N$. If $\varepsilon \geq 1$ this always holds, and we can take $N = 1$. For $\varepsilon < 1$

$$\frac{1-\varepsilon}{\varepsilon x}$$

is a positive real number, so call on Archimedes to choose $N \in \mathbb{N}$ such that

$$N > \frac{1-\varepsilon}{\varepsilon x}.$$

Then

$$n \geq N > \frac{1-\varepsilon}{\varepsilon x} \quad \text{and thereby} \quad |f_n(x) - 1| < \varepsilon$$

for all $n \geq N$. This completes the proof the way I did the first examples in the course (Exercise 2.32). But quicker is what we did in the tutorials. First do a convenient simplifying estimate

$$|f_n(x) - 1| = |\frac{nx}{1+nx} - 1| = \frac{1}{1+nx} < \frac{1}{nx} < \varepsilon,$$

and then take

$$N > \frac{1}{\varepsilon x}.$$

c) The above choices of $N$ rely on $x$. Depending on the set $A$ on which we consider the convergence, there may be a worst but still OK case for this choice of $N$. For $A = [1, \infty)$ we see that this worst case is $x = 1$. For all $x \geq 1$ we now choose this worst case $N$.

$$N > \frac{1-\varepsilon}{\varepsilon x}$$

as in b) for all $x \geq 1$ simultaneously, by choosing

$$N > \frac{1-\varepsilon}{\varepsilon}.$$

Then we see we're still OK:

$$\begin{matrix} n \geq N \\ x \geq 1 \end{matrix} \implies n \geq N \geq \frac{1-\varepsilon}{\varepsilon} \geq \frac{1-\varepsilon}{\varepsilon x},$$

whence

$$n > \frac{1-\varepsilon}{\varepsilon x},$$

equivalent to the desired inequality $|f_n(x) - 1| < \varepsilon$ as before. This completes the proof that the convergence is uniform on $A = [1, \infty)$. Here too we might have taken the quicker approach choosing $N > \frac{1}{\varepsilon}$, as in b).

d) Not asked: Can the convergence of $f_n(x)$ to 1 be uniform on $\mathbb{R}_+$ ? To disprove this statement you have to exhibit a sequence $x_n \in \mathbb{R}_+$ such that $|x_n - 1| \geq \varepsilon$ for all $n$.

**Problem 3.** Consider the differential equation $f''(x) = f(x)$.

a) Use a power series solution of the form

$$a_0 + a_2 x^2 + a_4 x^4 + a_6 x^6 + a_8 x^8 + \cdots$$

*to find an even solution with* $f(0) = 1$. You may guess the expression for $a_{2n}$ from your calculations.

b) The power series that you (should) have found converges for all $x \in \mathbb{R}$. You don't need your answer to a) to continue. The derivative of $f$ is an odd solution denoted by $g$. *Explain in detail why*

$$(f(x))^2 - (g(x))^2$$

*is constant. Which constant?*

**a + b = 2 + 4 = 6 points**

Answers

a) Write

$$f(x) = a_0 + a_2 x^2 + a_4 x^4 + a_6 x^6 + a_8 x^8 + \cdots ,$$

$$f'(x) = 2a_2 x + 4a_4 x^3 + 6a_6 x^5 + 8a_8 x^7 + \cdots ,$$

$$f''(x) = 2a_2 + 3 \cdot 4a_4 x^2 + 5 \cdot 6a_6 x^4 + 7 \cdot 8a_8 x^6 + \cdots .$$

Use $f(0) = 1$ conclude $a_0 = 1$ and then $f''(x) = f(x)$ to choose $a_2, a_4, \ldots$ with

$$2a_2 = a_0 = 1,\ 3 \cdot 4a_4 = a_2,\ 5 \cdot 6a_6 = a_4, \ldots,$$

whence

$$a_2 = \frac{1}{2}, \quad a_4 = \frac{1}{4 \cdot 3 \cdot 2}, \quad a_6 = \frac{1}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2},$$

and recognise the general expression $a_{2n} = \frac{1}{(2n)!}$, consistent also with $a_0$ and $a_2$.

b) It's given that the power series are valid for all $x$. Define $g = f'$. Then $g' = f'' = f$ so by the chain rule the derivative of $f^2 - g^2$ is equal to $2ff' - 2gg' = 2fg - 2gf = 0$ on the whole of $\mathbb{R}$. The mean value theorem now implies that $f(a)^2 - g(a)^2 = f(b)^2 - g(b)^2$ for all $a, b \in \mathbb{R}$, and thus that

$$f(x)^2 - g(x)^2 = f(0)^2 - g(0)^2 = 1 \quad \text{for all} \quad x \in \mathbb{R}.$$

**Problem 4.** Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = x(1 + x).$$

This $f$ can next earn you

**a + b + c + d + e = 3 + 1 + 3 + 3 + 5 = 15 points**

a) The linear approximation of $f(x)$ near $x = 0$ is given by $x$. *Verify the epsilon-delta statement for the remainder term* that implies that $f$ is differentiable in $x = 0$ with $f'(0) = 1$.

b) Now consider for $y \in \mathbb{R}$ the equation

$$f(x) = x(1 + x) = y \tag{11.31}$$

and the *modified* Newton scheme $x_n = x_{n-1} + f'(0)^{-1}(y - f(x_{n-1}))$ to solve $f(x) = y$. *Verify that*

$$x_n = y - x_{n-1}^2. \tag{11.32}$$

c) Starting from $x_0 = 0$ the scheme (11.32) defines a sequence $x_n$ that depends on $y$, and thereby a sequence of functions $g_n$ by defining $g_n(y) = x_n$. So $g_0(y) = 0$ and $g_1(y) = y$. *Evaluate* $g_2(y)$ *and* $g_3(y)$.

Next it's about finding an (inverse) function $g$ as the uniform limit of the sequence $g_n$.

d) Following the notation in (3.1) in Chapter 3 and avoiding the Greek letter $\xi$ we write $s_n = x_n - x_{n-1}$, with $s$ for *step*. Now suppose that for some $n \in \mathbb{N}$ it holds that

$$|x_{n-1}| \leq \frac{1}{4} \quad \text{and} \quad |x_n| \leq \frac{1}{4}\,.$$

*Use (11.32) to prove that then*

$$|s_{n+1}| \leq \frac{1}{2}\,|s_n|\,. \tag{11.33}$$

This allows to continue the story-line in Chapter 3 for all $y \in [-\frac{1}{8}, \frac{1}{8}]$ simultaneously.

e) Use d) to show that $g_n$ is a uniform Cauchy sequence in $C([-\frac{1}{8}, \frac{1}{8}])$, see Definition 4.2.
Hint: recall that $x_n = g_n(y)$ and *show first that*

$$|y| \leq \frac{1}{8} \implies |x_n| \leq 2\,|y| \leq \frac{1}{4} \quad \text{for all} \quad n \in \mathbb{N}.$$

Answers NB This is the exercise in which you had to *apply* the definitions and techniques that were introduced in Chapter 3 to prove Theorem 5.14 to an example with a parameter, and choose the final $N$ independent of this parameter. Considering the iterates as a function of this parameter, this is about uniform convergence then.

a) Since

$$f(x) = f(0) + 1\,x + R(x), \quad R(x) = x^2,$$

we see that $|R(x)| = |x|^2 = |x|\,|x| < \varepsilon|x|$ if $0 < |x| < \varepsilon$. It follows that for all $\varepsilon > 0$ there exists $\delta > 0$, namely $\delta = \varepsilon$, such that $0 < |x| < \delta = \varepsilon$ implies $|R(x)| < \varepsilon|x|$. Thus $f$ is differentiable in $x = 0$ with $f'(0) = 1$.

b) The scheme is of the form $x_n = F(x_n)$ with $F(x) = x + f'(0)^{-1}(y - f(x))$. For the given $f$ this gives $F(x) = x + 1^{-1}(y - x(1+x)) = x + y - x - x^2 = y - x^2$, so $x_n = y - x_{n-1}^2$.

c) With $x_1 = g_1(y) = y$ we have $g_2(y) = x_2 = y - y^2$ and then $g_3(y) = x_2 = y - (y - y^2)^2$.

d) We have

$$s_{n+1} = x_{n+1} - x_n = y - x_n^2 - y + x_{n-1}^2 = -(x_n + x_{n-1})(x_n - x_{n-1}) = (x_n + x_{n-1})s_n,$$

and thereby

$$|s_{n+1}| \leq |x_n + x_{n-1}|\,|s_n| \leq (|x_n| + |x_{n-1}|)\,|s_n| \leq (\frac{1}{4} + \frac{1}{4})|s_n| = \frac{1}{2}\,|s_n|.$$

Note that just like $x_n$ the $s_n$ are $y$-dependent.

e) We have to get $|x_m - x_n| = |g_m(y) - g_n(y)| < \varepsilon$ for all $y$ with $|y| \leq \frac{1}{8}$ simultaneously, provided $m > n \geq N$, $N \in \mathbb{N}$ to be found. Following the hint we start from

$$|x_1| = |s_1| = |y| \leq 2|y| \leq \frac{2}{8} = \frac{1}{4}$$

to get

$$|x_2| \leq |x_1| + |s_2| \leq |s_1| + \frac{1}{2}\,|s_1| = (1 + \frac{1}{2})|s_1| \leq 2|y| \leq \frac{1}{4},$$

whence

$$|x_3| \leq |x_2| + |s_3| \leq (1 + \frac{1}{2})|s_1| + \frac{1}{2}|s_2| \leq (1 + \frac{1}{2} + \frac{1}{4})|s_1| \leq 2|y| \leq \frac{1}{4}.$$

In the next step we have

$$|x_4| \leq |x_3| + |s_4| \leq (1 + \frac{1}{2} + \frac{1}{4})|s_1| + |s_4| \leq (1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8})|s_1| \leq 2|y| \leq \frac{1}{4},$$

because $|s_4| \leq \frac{1}{2}|s_3| \leq \frac{1}{4}|s_2| \leq \frac{1}{8}|s_1|$.

And so on. We conclude that all $|x_n| = |g_n(y)|$ are bounded by $\frac{1}{4}$, whence

$$|s_{n+1}| \leq \frac{1}{2^n}|s_1| = \frac{1}{2^n}|y| \leq \frac{1}{2^{n+3}}$$

for all $y$ with $|y| \leq \frac{1}{8}$. But then we have for all such $y$ and $m > n \geq N$ that

$$|g_m(y) - g_n(y)| = |x_m - x_n| \leq |s_{n+1}| + \cdots + |s_m| \leq |s_{n+1}| \underbrace{(1 + \frac{1}{2} + \frac{1}{4} + \cdots)}_{\text{finitely many terms}}$$

$$\leq 2|s_{n+1}| \leq \frac{1}{2^{n+2}} \leq \frac{1}{2^{N+2}}.$$

Choosing $N$ so large that $2^{N+2} > \frac{1}{\varepsilon}$ the proof that $g_n$ is uniformly Cauchy on the $y$-interval $[-\frac{1}{8}, \frac{1}{8}]$ is now complete.

## 11.7 An earlier version of Exercise 4

This relates to the first page of Chapter 14. Inverse functions, an example to explain theorem, method and proof.

**Problem 5.** Let $f : \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = x(1 + x).$$

a) Explain why the linear approximation of $f(x)$ near $x = 0$ is given by $x$. Verify the epsilon-delta statement for the remainder term that implies that $f$ is indeed differentiable in $x = 0$ with $f'(0) = 1$.

b) Now consider for $y \in \mathbb{R}$ the equation

$$f(x) = x(1 + x) = y \tag{11.34}$$

and the *modified* Newton scheme $x_n = x_{n-1} + f'(0)^{-1}(y - f(x))$ to solve $f(x) = y$. Verify that

$$x_n = y - x_{n-1}^2. \tag{11.35}$$

c) Starting from $x_0 = 0$ the scheme (11.35) defines a sequence $x_n$ that depends on $y$, and thereby a sequence of functions $g_n$ by writing $x_n = g_n(y)$. So $g_0(y) = 0$ and $g_1(y) = y$.

*Evaluate $g_2(y)$ and $g_3(y)$. Sketch the graph $x = g_2(y)$ in the $x, y$-plane.*

d) Use induction (*domino principle*) to prove that

$$g_n(y) = y + R_n(y)$$

with $R_n(y)$ a polynomial of degree $2^{n-1}$. Hint: verify that

$$R_n(y) = -y^2 - \underbrace{\cdots\cdots\cdots\cdots\cdots\cdots}_{\text{terms with exponents between 2 and } 2^{n-1}} - y^{2^{n-1}}.$$

e) Use d) to explain why $g_n'(0) = 1$ for all $n \in \mathbb{N}$.

f) Following the notation in (3.1) in Chapter 3 and avoiding the Greek letter $\xi$ we write $s_n = x_n - x_{n-1}$, with $s$ for *step*. Suppose that for some $n \in \mathbb{N}$ it holds that

$$|x_{n-1}| \leq \frac{1}{4} \quad \text{and} \quad |x_n| \leq \frac{1}{4}\,.$$

*Use (11.35) to prove that then*

$$|s_{n+1}| \leq \frac{1}{2}\,|s_n|\,. \tag{11.36}$$

This allows to continue the story-line in Chapter 3 for a whole range of $y$ simultaneously next.

g) Use f) to show that $g_n$ is a uniform Cauchy sequence in $C([-\frac{1}{8}, \frac{1}{8}])$, see Definition 4.2.
Hint: recall that $x_n = g_n(y)$ and show that

$$|y| \leq \frac{1}{8} \implies |x_n| < 2\,|y| \leq \frac{1}{2} \quad \text{for all} \quad n \in \mathbb{N}.$$

h) (continued) Which theorem now says that the limit of $g_n(y)$ defines a function $g \in C([-\frac{1}{8}, \frac{1}{8}])$?

i) Consider the scheme in (11.35) starting from $x_0 = 0$ for two different values of $y$, say $y$ and $\tilde{y}$, both in $[-\frac{1}{8}, \frac{1}{8}]$. Write $x_n = g_n(y)$, $\tilde{x}_n = g_n(\tilde{y})$ and let $d_n = |x_n - \tilde{x}_n| = |g_n(y) - g_n(\tilde{y})|$. Show that

$$d_n \leq |y - \tilde{y}| + \frac{1}{2}\,d_{n-1} \quad \text{for all} \quad n \in \mathbb{N},$$

and use this to show that all $g_n$ and $g$ are Lipschitz continuous on $[-\frac{1}{8}, \frac{1}{8}]$ with Lipschitz constant 2.

j) Let $g$ be the function obtained in h). Then $f(g(y)) = y$ for all $y \in [-\frac{1}{8}, \frac{1}{8}]$. Substitute $x = g(y)$ in (11.34) and re-arrange to prove that $g$ is differentiable in $y = 0$ with $g'(0) = 1$.

# 12 Newton's method revisited

For the analysis of Newton's method we need the mean value theorem in integral form.

**Exercise 12.1.** 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

$$F(b) - F(a) = \int_0^1 F'((1-t)a+tb)(b-a)\,dt = \int_0^1 F'((1-t)a+tb)\,dt\,(b-a), \quad (12.1)$$

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 (12.1) 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

$$F(x) = F(a) + F'(a)(x-a) + R(x; a) \quad (12.2)$$

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

$$|R(x; a)| \leq \int_0^1 Lt|x-a|^2\,dt = \frac{L}{2}|x-a|^2. \quad (12.3)$$

In (12.2) we have a linear approximation with a remainder term estimated in (12.3) 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 (9.10). 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]We change from subscript $a$ on $R(x)$ to $R(x; a)$.

## 12.1 The generalised mean value formula

https://www.youtube.com/playlist?list=PLQgy2W8pIli9jyuYN76HM3YdXjwBZrx8L

**Theorem 12.2.** *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 12.2 says that $F$ is a primitive of $f$, and that for this primitive

$$\int_a^b f(s)\, ds = F(b) - F(a), \tag{12.4}$$

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 12.3.** 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 12.4.** *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 12.2.*

Summing up, the mean value integral formula (12.1) 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 11.16, we apply Theorem 12.4 with $a = 0$ and $b = 1$ to the function defined by (11.24), and conclude that

$$F(y) - F(x) = \int_0^1 F'((1-t)x + ty)(y-x)\,dt, \tag{12.5}$$

as a $Y$-valued integral, which we can write as

$$F(y) - F(x) = \int_0^1 F'((1-t)x + ty)\,dt(y-x), \tag{12.6}$$

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 14.4.

## 12.2 Convergence of Newton's method

https://www.youtube.com/playlist?list=PLQgy2W8pIli-xh3hDqLh7u_npItPP2RCF

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 $B_r$, (12.2) rewrites as

$$f(x_n) = \underbrace{f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1})}_{\text{linear approximation}} + R(x_n; x_{n-1}), \tag{12.7}$$

[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

$$p_n = |x_n - x_{n-1}| \quad \text{and} \quad q_n = |f(x_n)|, \tag{12.8}$$

and assume that $x_n$ is defined by

$$x_n = x_{n-1} - (f'(x_{n-1}))^{-1} f(x_{n-1}) \quad (n \in \mathbb{N}), \tag{12.9}$$

with $x_0 = 0$. Then $x_n \in B_r$ as long as

$$p_1 + p_2 + \cdots + p_n < r, \tag{12.10}$$

in which case it follows that

$$p_n \leq C q_{n-1} \quad \text{and} \quad q_n \leq \frac{1}{2} L p_n^2, \tag{12.11}$$

because (12.9) puts the linear approximation in (12.7) equal to zero.

The inequalities in (12.11) can now be used beginning with

$$q_0 = |f(0)| \quad \text{and} \quad p_1 \leq C q_0 = C|f(0)|. \tag{12.12}$$

Combining (12.11) and (12.12) it follows that

$$p_n \leq \mu p_n^2 \quad \text{with} \quad \mu = \frac{1}{2} LC \quad \text{and} \quad p_1 \leq C|f(0)|. \tag{12.13}$$

The question then is for which $P$ we can conclude that the implication

$$p_1 \leq C|f(0)| < P \implies \sum_{n=1}^{\infty} p_n < r \tag{12.14}$$

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 (12.9) starting from $x_0 = 0$. If we take equalities in (12.13) and (12.14) then

$$p_n = \mu p_{n-1}^2 \quad \text{for} \quad n \in \mathbb{N}; \quad p_1 = P\,; \quad \sum_{n=1}^{\infty} p_n = r. \tag{12.15}$$

Putting $\xi_n = \mu p_n$ so that $\xi_n = \xi_{n-1}^2$, this is equivalent to

$$G(\mu P) = \mu r \quad \text{with} \quad G(\xi) = \xi + \xi^2 + \xi^4 + \xi^8 + \xi^{16} + \cdots . \tag{12.16}$$

This defines $P$ as a function of $\mu$ and $r$.

**Exercise 12.5.** Use

$$G(\xi) < \frac{\xi}{1-\xi}$$

to show that

$$|f(0)| \leq \frac{2r}{(2+rLC)C} = \frac{1}{(\frac{1}{r}+\frac{LC}{2})C}$$

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 12.6.** Put $y = 1 + z$ and see what you get for the sequence $z_n$ to understand why the convergence is so fast.

**Exercise 12.7.** Put $e = x^2 - 2$, rewrite (2.2) in terms of $e$ and $\tilde{e}$, examine the sequence $e_n$, and compare to Exercise 12.6.

# 13 Back to calculus

Most of this chapter could be part of any calculus course, except for Section 13.2 perhaps. Calculus is so much fun:

`https://en.wikipedia.org/wiki/Proof_that_22/7_exceeds_%CF%80`

## 13.1 More on exp and ln

**Exercise 13.1.** 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 13.2.** Same question as in Exercise 13.1 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 13.3.** 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 13.2. 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 $F_\alpha(x)$?

## 13.2 Early treatment of integrals with parameters

The results in this section are often needed and perhaps justly postponed till after the introduction of integral calculus for functions of multiple variables. Still, let's consider here

$$J(t) = \int_0^1 f(x, t)\, dx$$

[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? We examine a follow your nose argument for what (the one-sided) derivative $J'(0)$ should be and see what we need to prove it. You may want to jump to Theorem 13.5 for a simpler statement under stronger assumptions and a much simpler proof in Section 27.2.

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

$$f(t,x) = f(0,x) + f_t(0,x)t + \underbrace{f_t(\tau(x),x) - f_t(0,x)}_{R(t,x)}. \tag{13.1}$$

This defines $R(t,x)$. Continuity of $f$ and $f_t$ in $x$ allows to write

$$\begin{aligned} J(t) = \int_0^1 f(t,x)\,dx &= \int_0^1 (f(0,x) + f_t(0,x)t + R(t,x))\,dx \\ &= J(0) + t\int_0^1 f_t(x,0)\,dx + \int_0^t R(t,x)\,dx \qquad (13.2)\\ &= J(0) + t\int_0^1 f_t(x,0)\,dx + r(t). \end{aligned}$$

Here

$$r(t) = \int_0^t R(t,x)\,dx \quad \text{with} \quad R(t,x) = (\underbrace{f_t(\tau(x),x) - f_t(0,x)}_{<\varepsilon\,?})t$$

should have the usual property indicated by the question mark in the underbrace. Indeed, if we assume that

$$x \to f(t,x)$$

and

$$x \to f_t(0,x) = g(t,x)$$

[2]The integral form would require double integrals, see Section 27.2.

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 (13.2) is then also continuous. The second expression with $\tau(x) \in (0,t)$ above (13.1) can now be used to establish $r(t) = o(t)$ as $t \to 0$, since all we need for the remainder term $r(t)$ is that $|r(t)| < \varepsilon t$ for $t$ sufficiently small. Thus, if for $f_t(t,x) = g(t,x)$ it holds that

$$|g(t,x) - g(0,x)| < \varepsilon \tag{13.3}$$

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 $\varepsilon$-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 13.4.** *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 13.5.** *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 13.6.** To prove the continuity of the derivative you need to prove that

$$t \to j(t) = \int_a^b g(t,x)\,dx$$

is continuous on $I$ if $(t,x) \to g(t,x)$ is continuous on $I \times [a,b]$. Hint: use a uniform $\varepsilon$-argument and state a version of Theorem 8.10 needed to do so. NB. This continuity allows for a much quicker proof, see Section 27.2.

## 13.3 Partial integration and Taylor polynomials

**Theorem 13.7.** *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 13.10 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 13.7 is not really essential for the analysis of Newton's method in Chapter 12.2, but it is for the proof of Morse' Lemma in Chapter 15.

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

$$\int_\alpha^\beta x(t) y'(t)\,dt = [x(t)y(t)]_\alpha^\beta - \int_\alpha^\beta x'(t) y(t)\,dt. \tag{13.4}$$

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

$$-u''(x) = f(x) \quad \text{for all} \quad 0 \le x \le 1, \quad \text{and} \quad u(0) = u(1) = 0. \tag{13.5}$$

[3]And in a much more general setting in fact.

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

$$u(x) = u'(0)x - \int_0^x F(s)\,ds,$$

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

$$\begin{aligned}\underbrace{1\,F(s)}_{G'(s)F(s)} = \underbrace{(s-a)'}_{G'(s)}F(s) &= (\underbrace{(s-a)}_{G(s)}F(s))' - \underbrace{(s-a)}_{G(s)}F'(s) \\ &= \underbrace{((s-a)F(s))'}_{(G(s)F(s))'} - \underbrace{(s-a)f(s)}_{G(s)F'(s)},\end{aligned}$$

in which $1 = G'(s)$ with $G(s) = s - a$ and $a$ free to choose.

The primitive of $F(x)$ then rewrites as

$$\int_0^x F(s)\,ds = [(s-a)F(s)]_0^x - \int_0^x (s-a)f(s)\,ds = \int_0^x (x-s)f(s)\,ds. \quad (13.6)$$

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

$$\begin{aligned}u(x) &= \int_0^1 (1-s)f(s)\,ds\,x - \int_0^x (x-s)f(s)\,ds \\ &= x\int_x^1 (1-s)f(s)\,ds + (1-x)\int_0^x sf(s)\,ds = \int_0^1 A(x,s)f(s)\,ds.\end{aligned}$$

The expression

$$A(x,s) = \begin{cases}(1-x)s & \text{for } 0 \le s \le x \\ (1-s)x & \text{for } x \le s \le 1\end{cases} \quad (13.7)$$

is called the kernel for the solution operator, which gives $u$ in terms of $f$ as

$$u(x) = \int_0^1 A(x,s)f(s)\,ds. \tag{13.8}$$

You may prefer to memorize the integration by parts formula as

$$\int_a^b F(x)G'(x)\,dx = [F(x)G(x)]_a^b - \int_a^b F'(x)G(x)\,dx. \tag{13.9}$$

It's handy for computing integrals, but also for taking primitives of primitives, as we just saw and see again below.

**Exercise 13.8.** 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 (13.9) 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 13.9.** In the context of Exercise 13.8 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 13.10.** Modify the scheme in Exercise 13.9 as

$$F_0(x) = f(x), \quad F_n(x) = b_n + \int_a^x F_{n-1}(s)\,ds \quad (n = 1, 2, 3\dots), \tag{13.10}$$

and give a similar formula for $F_n(x)$ with more terms. By construction $F_n(a) = b_n$, $F_n'(a) = b_{n-1}$, $F_n''(a) = b_{n-2}$, ..., 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

$$f(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}(0)}{k!}(x-a)^k + \underbrace{\frac{f^{(n)}(\sigma)}{n!}(x-a)^n}_{\frac{1}{(n-1)!}\int_a^x (x-s)^{n-1} f^{(n)}(s)\,ds.} \tag{13.11}$$

for some $\sigma$ between $a$ and $x$.

The result in (13.11) holds in fact without the assumption that $f^{(n)}$ is continuous, with $\sigma$ strictly between $a$ and $x$, as a clever application of Theorem 10.7 shows. The case $n = 1$ reduces to Theorem 10.7.

## 13.4 Asymptotic formulas

This is not part of every standard calculus course. The notation

$$f(x) \sim g(x) \quad \text{for} \quad x \to a \tag{13.12}$$

means that

$$\frac{f(x)}{g(x)} \to 1 \quad \text{if} \quad x \to a,$$

in which often $a$ is 0 or $\infty$. Similarly the statement

$$n! \sim (\frac{n}{e})^n \sqrt{2\pi n} \quad \text{as} \quad n \to \infty \tag{13.13}$$

means that the limit of the quotient of the terms on both sides of the *twiddle* is 1.

**Exercise 13.11.** 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 13.12.** 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 $y \to \infty$.

## 13.5 Exercises

**Exercise 13.13.** Discuss the following formulas.

$$\int_\alpha^\beta x(t)\,\underbrace{y'(t)\,dt}_{dy} = [x(t)y(t)]_\alpha^\beta - \int_\alpha^\beta y(t)\,\underbrace{x'(t)\,dt}_{dx},$$

$$\int_\alpha^\beta \underbrace{F'(x(t))}_{f(x(t))}\,x'(t)\,dt = F(x(\beta)) - F(x(\alpha)) = \int_{x(\alpha)}^{x(\beta)} F'(x)\,dx = \int_a^b \underbrace{F'(x)}_{f(x)}\,dx,$$

$$\int_a^b f(x)\,dx = \int_\alpha^\beta f(x(t))x'(t)\,dt. \tag{13.14}$$

**Exercise 13.14.** 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 13.15.** Sketch the graph $y = x^n e^{-x}$ (for $n$ not too large) in the $x, y$-plane. Where's the top of the mountain?

**Exercise 13.16.** 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 13.17.** Write

$$g_n(x) = e^{-\psi_n(x)} \quad \text{met} \quad \psi_n(x) = -\ln(g_n(x)),$$

and verify that

$$\psi_n(x) = x - n\ln(1 + \frac{x}{n}) = n(\frac{x}{n} - \ln(1 + \frac{x}{n})) = n\psi_1(\frac{x}{n}).$$

**Exercise 13.18.** Put $x = s\sqrt{n}$ to conclude that[4]

$$n! = (\frac{n}{e})^n \sqrt{n} \int_{-\sqrt{n}}^\infty e^{-n\Psi(\frac{s}{\sqrt{n}})} ds \tag{13.15}$$

and show that

$$\int_{-\sqrt{n}}^\infty e^{-n\Psi(\frac{s}{\sqrt{n}})} ds \to \int_{-\infty}^\infty e^{-\frac{1}{2}s^2} ds \tag{13.16}$$

as $n \to \infty$. *Google Stirling's formula.*

[4] https://youtu.be/8sn5ekwvXSQ

## 13.6 Exercises about entropy

Some exercises that correspond to Section 1.6 of Bishop's Machine learning book. See if you have enough machinery at your disposal to do them. I'm updating this section while reading Finn's master thesis.

**Exercise 13.19.** Let $n > 1$ be an integer and let $p_1, \ldots, p_n$ be positive real numbers with $p_1 + \cdots + p_n = 1$. You may think of $p_1, \ldots, p_n$ as probabilities, each $p_k$ being the probability that out of $n$ events it is event $k$ that occurs.

a) Then the *entropy*[5]

$$H(p) = \sum_{i=1}^{n} p_i \ln \frac{1}{p_i}$$

is defined as a positive number. If $q_1, \ldots, q_m$ is another such finite sequence of positive probabilities, then also

$$H(q) = \sum_{j=1}^{m} q_j \ln \frac{1}{q_j} > 0,$$

and for the joint probabilities

$$\{p_i q_j : i = 1, \ldots, n, \, j = 1, \ldots, m\}$$

the entropy is defined likewise. Write $r = p \otimes q$ and $r_{ij} = p_i q_j$, define what $H(r) = H(p \otimes q)$ should be, and show that

$$H(p \otimes q) = H(p) + H(q).$$

What do you get without the assumption that $p_1 + \cdots + p_n = q_1 + \cdots + q_m = 1$?

b) Prove that, upto a multiplicative factor,

$$h(x) = \ln \frac{1}{x}$$

defines the only continuous function on $\mathbb{R}_+$ for which

$$H(p) = \sum_{i=1}^{n} p_i \, h(p_i) > 0$$

has this additivity property. For now this is a purely analytical or if you like algebraic statement. Note that $-H$ is strictly convex.

[5] The expected "improbability", if the improbability of event $k$ is defined by $\ln \frac{1}{p_k}$.

c) Examine limits of $H(p) = H(p_1, \ldots, p_n)$ as $p_1$ or some other $p_i$ goes to zero. Conclude that $H(p_1, \ldots, p_n)$ is well defined as a nonnegative number for all nonnegative $p_1, \ldots, p_n$.

The set of all such $(p_1, \ldots, p_n)$ with $p_1 + \cdots + p_n = 1$ is denoted by $\Sigma_n$. The global minimum of $H$ on $\Sigma_n$ is zero, attained in $(1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$. The global maximum of $H$ on $\Sigma_n$ is $\ln n$, attained in $(\frac{1}{n}, \ldots, \frac{1}{n})$. The union of all $\Sigma_n$ over $n \in \mathbb{N}$ may be denoted by $\Sigma_\infty$, considering $p = (p_1, \ldots, p_m)$ as $p = (p_1, \ldots, p_m, 0, 0, \ldots)$. Details in the items below. Connection with information theory in Exercises 13.20 and 13.21.

d) Show there exists a sequence $p_i \geq 0$ with

$$\sum_{i=1}^{\infty} p_i = 1 \quad \text{but} \quad \sum_{i=1}^{\infty} p_i \ln \frac{1}{p_i} = \infty,$$

so $H(p_1, p_2, \ldots)$ is not defined for all nonnegative sequences $p_1, p_2, \ldots$ with $p_1 + p_2 + \cdots = 1$. Hint: you cannot take $p_n = \frac{1}{n}$, but a logaritmic correction will do the job. It follows that we cannot define the entropy function $H$ on $\Sigma_\infty$.

e) Fix $n \in \mathbb{N}$ and consider

$$H_n(p_1, \ldots, p_n) = H_n(p) = \sum_{i=1}^{n} p_i \ln \frac{1}{p_i}$$

for variable $p_1, \ldots, p_n \geq 0$ with $p_1 + \cdots + p_n = 1$. Let $\Sigma_n$ be the set[6] of all such $p = (p_1, \ldots, p_n) \in \mathbb{R}^n$. Show that $H_n$ has a global maximum on $\Sigma_n$. Hint: prove that $H_n$ is continuous on $\Sigma_n$.

f) Let $p$ be a global maximizer for $H_n : \Sigma_n \to \mathbb{R}$. Suppose that $p_1 \cdots p_n > 0$. Prove that $p_1 = \cdots = p_n = \frac{1}{n}$. Hint: argue by contradiction and use the linear approximations of all terms in the sum that defines $H_n(p)$.

g) Prove that $H_n : \Sigma_n \to \mathbb{R}$ has a positive global maximum $\ln n$, and that the unique maximizer is given by $p_1 = \cdots = p_n = \frac{1}{n}$.

h) Determine all global minimizers of $H_n : \Sigma_n \to \mathbb{R}$.

**Exercise 13.20.** Consider the following $9$ binary words $w$ with probabilities $p$ as indicated.

| $w$ | 00 | 010 | 011 | 100 | 101 | 1100 | 1101 | 1110 | 1111 |
|---|---|---|---|---|---|---|---|---|---|
| $p$ | $\frac{1}{4}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{16}$ | $\frac{1}{16}$ | $\frac{1}{16}$ | $\frac{1}{16}$ |

Denote the set of these words by $W$.

[6] $\Sigma$ for simplex.

a) Verify that the probabilities $p = p(w)$ add up to $1$, that is,

$$\sum_{w \in W} p(w) = 1.$$

b) Explain how these probabilities are obtained from grouping $16$ equally probable words $0000, \ldots, 1111$. Hint:

$$\underbrace{0000, 0001, 0010, 0011}, \underbrace{0100, 0101}, \underbrace{0110, 0111}, \underbrace{1000, 1001}, \underbrace{1010, 1011}, 1100, 1101, 1110, 1111.$$

c) Each $w$ has $p = p(w) = 2^{-n}$ with $n$ the number of bits in $w$. Thus for the average (or expected) number of bits we have

$$\begin{aligned} \ln 2 \sum_{w \in W} p(w)\, {}^2\!\log \frac{1}{p(w)} &= \sum_{w \in W} p(w) \ln \frac{1}{p(w)} \\ &= H(\frac{1}{4}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}). \end{aligned}$$

Verify this $H$-value is smaller than $\ln 9$.

**Exercise 13.21.** Bishop uses the example

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $s$ | $0$ | $10$ | $110$ | $1110$ | $111100$ | $111101$ | $111110$ | $111111$ |
| $p$ | $\frac{1}{2}$ | $\frac{1}{4}$ | $\frac{1}{8}$ | $\frac{1}{16}$ | $\frac{1}{64}$ | $\frac{1}{64}$ | $\frac{1}{64}$ | $\frac{1}{64}$ |

to explain the relation with information theory. These 8 bit strings of varying length may be considered as representing the letters from an 8 letter alphabet, with different propabilities of occurring in words of a particular language that uses that alphabet. If these probabilities are as indicated the average number of bits used per letter would be

$$1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 4 \times \frac{1}{16} + 6 \times \frac{1}{64} + 6 \times \frac{1}{64} + 6 \times \frac{1}{64} + 6 \times \frac{1}{64} = 2.$$

Equal probilities with encodings

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $s$ | $000$ | $001$ | $010$ | $011$ | $100$ | $101$ | $110$ | $111$ |
| $p$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{8}$ |

would have 3 bits on average. Think of binary coding Dutch sentences consisting of the 32 symbols

$$abcdefghijklmnopqrstuvwxyz.,;:?!$$

with fewer bits for more likely and more bits for more unlikely letters and punctuations.

We saw that we cannot define the entropy function $H$ on

$$\Sigma_\infty = \{(p_1, p_2, p_3, \dots) : p_n \geq 0 \text{ for all } n \in \mathbb{N}, \sum_{n\in\mathbb{N}} p_n = 1\}.$$

How about the entropy of a continuous distribution? Jump to d) in Exercise 13.22 below for a shortcut to how the Kullback-Leiber divergence between two such distributions $f$ and $g$ appears when trying to answer this question. What comes first in Exercise 13.22 is for just one distribution $f$ and is not going to work for distributions on $\mathbb{R}$.

**Exercise 13.22.** Let $f \in C([a,b])$ be a nonnegative function[7] with $\int_a^b f = 1$, let $P$ be a partition

$$a = x_0 \leq x_1 \leq \cdots \leq x_n = b$$

of $[a,b]$ as in Definition 6.6, and define

$$p_i = \int_{x_{i-1}}^{x_i} f(x)\, dx \geq 0 \quad \text{for} \quad i = 1, \dots, n.$$

Then $p_1 + \cdots + p_n = 1$. Below you may like to take all $p_i > 0$ first.

a) Show that for ever $i \in \{1, \dots, n\}$ there exists $\xi_i \in (x_{i-1}, x_i)$ such that

$$p_i = f(\xi_i)(x_i - x_{i-1}).$$

b) Take the partition $P$ equidistant. Show that

$$H_n(p_1, \dots, p_n) = (b-a)\ln n + \sum_{i=1}^n f(\xi_i) \ln \frac{1}{f(\xi_i)} (x_i - x_{i-1})$$

to conclude that

$$(b-a)\, \ln n - H_n(p_1, \dots, p_n) \to \int_a^b f(x) \ln f(x)\, dx$$

as $n \to \infty$ for every nonnegative $f \in C([a,b])$ with $\int_a^b f = 1$.

c) The above derivation implies the nonnegativity property

$$\int_a^b f(x) \ln f(x)\, dx \geq 0,$$

[7]Think of $f$ as a probability distribution on $[a,b]$.

and equality holds if $f(x) \equiv 1$. If equality holds for some other nonnegative $f \in C([a,b])$ then $f$ must have values large and smaller than $1$. Show that such an $f$ with

$$\int_a^b f = 1$$

can then be modified to make the integral of $f \ln f$ smaller than zero and still have integral of $f$ equal to $1$, contradicting the above nonnegativity property.

Hint: use the strict positivity of the second derivative of $y \to y \ln y$.

d) This is to modify the above in such a way that we can also handle probability distributions on $\mathbb{R}$. But first we stick to $[a,b]$. Let $g \in C([a,b])$ be another function with $\int_a^b g = 1$ and strictly positive. Choose any partition as in a). We use

$$y = \int_a^x g, \quad y_i = \int_a^{x_i} g, \quad \tilde{f}(y) = f(x), \quad \tilde{g}(y) = g(x)$$

to transform[8]

$$p_i = \int_{x_{i-1}}^{x_i} f(x)\,dx$$

and see what we get for $H(p_1, \ldots, p_n)$ as $n \to \infty$. Verify that

$$p_i = \int_{x_{i-1}}^{x_i} \frac{f(x)}{g(x)}\, g(x)\,dx = \int_{y_{i-1}}^{y_i} \frac{\tilde{f}(y)}{\tilde{g}(y)}\,dy = \frac{\tilde{f}(\eta_i)}{\tilde{g}(\eta_i)}(y_i - y_{i-1}) = \frac{f(\xi_i)}{g(\xi_i)} \int_{x_{i-1}}^{x_i} g$$

for some $\eta_i \in (y_{i-1}, y_i)$ provided by the mean value theorem, and $\xi_i$ then defined by $\eta_i = \int_a^{\xi_i} g$.

e) But then we can define

$$q_i = \int_{x_{i-1}}^{x_i} g(x), \quad \text{just like we defined} \quad p_i = \int_{x_{i-1}}^{x_i} f(x)\,dx,$$

and conclude that

$$p_i\, g(\xi_i) = q_i\, f(\xi_i). \tag{13.17}$$

If $f$ is positive it should follow that

$$\sum_{i=1}^n p_i \ln \frac{q_i}{p_i} = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} f(x)\,dx\, \ln \frac{g(\xi_i)}{f(\xi_i)} \to \int_a^b f(x) \ln \frac{f(x)}{g(x)}\,dx \tag{13.18}$$

as $n \to \infty$. But under which conditions? For later worries. Playing around we have derived the formula for the so-called Kullback-Leiber divergence

$$\mathrm{KL}(f||g) = \int_a^b f(x) \ln \frac{f(x)}{g(x)}\,dx$$

between probability densities $f$ and $g$, and on the left we see $\mathrm{KL}(p||q)$.

[8] Using the rule you discover from (13.14), Theorem 10.12 and Theorem 11.4.

**Exercise 13.23.** Generalise the statements above to functions $f, g : \mathbb{R} \to \mathbb{R}^+$ with

$$\int_{\infty}^{\infty} f(x)\,dx = \int_{\infty}^{\infty} g(x)\,dx = 1,$$

for which

$$\int_{\infty}^{\infty} f(x) \ln \frac{f(x)}{g(x)}\,dx$$

has a meaning, not necessarily via

$$\int_{\infty}^{\infty} f(x) \ln f(x)\,dx - \int_{\infty}^{\infty} f(x) \ln g(x)\,dx.$$

The first integral above is called minus the differential entropy of $f$, and denoted here and elsewhere by

$$S(f) = \int_{\infty}^{\infty} f(x) \ln f(x)\,dx$$

when it exists. If so then $S(f_\lambda)$, $f_\lambda$ defined by

$$f_\lambda(x) = \frac{1}{\lambda} f(\frac{x}{\lambda}),$$

also exists.

a) Verify that

$$S(f_\lambda) = S(f) - \ln \lambda.$$

b) Verify that $S(f)$ is invariant under translation.

c) Given constraints of $f$, for instance

$$\int_{\infty}^{\infty} f(x)\,dx = 1, \quad \int_{\infty}^{\infty} x f(x)\,dx = \mu, \quad \int_{\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \sigma^2,$$

the functional $S$ may or may not have a minimum. Explain why it is no restriction to restrict to the case that $\mu = 0$ and $\sigma = 1$.

d) Try and find this minimum and its minimizer.

**Exercise 13.24.** In analogy with (13.18) consider

$$H(p||q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$$

for positive $p_1, q_1, \ldots, p_n, q_n$ with $p_1 + \cdots + p_n = q_1 + \cdots + q_n = 1$. Use

$$u_i = \frac{p_i}{q_i}, \quad F(x) = x \ln x,$$

to show $H(p||q) > 0$ unless $p_i = q_i$ for all $i = 1, \ldots, n$. Hint: $F$ is strictly convex.

**Exercise 13.25.** (continued) Think of $q_i \in (0,1)$ as probabilities for $n$ events, and assign to each event a value $u_i \in J$, with $J \subset \mathbb{R}$ an interval. This defines a stochastic variable $U$ with

$$P(U = u_i) = q_i.$$

If $F : J \to \mathbb{R}$, then $V = F(U)$ is also a stochastic variable with

$$P(V = v_i) = q_i, \quad v_i = F(u_i),$$

and expectations

$$EU = \sum_{i=1}^{n} u_i q_i, \quad EV = \sum_{i=1}^{n} v_i q_i = \sum_{i=1}^{n} f(u_i) q_i = EF(U)$$

satisfying $EF(U) < F(EU)$ if $F$ is strictly convex. Verify that $F : [0,\infty) \to \mathbb{R}$ is strictly convex and that for $p_1, \ldots, p_n \geq 0$ and

$$u_i = \frac{p_i}{q_i} \quad \text{it holds that} \quad F(EU) = F(\sum_{i=1}^{n} p_i) = F(1) = 0$$

if $p_1 + \cdots + p_n = 1$ and

$$E(F(U)) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}.$$

Thus $p_i > 0$ is not required to conclude $H(p||q) > 0$ for $(p_1, \ldots, p_n) \neq (q_1, \ldots, q_n)$ in Exercise 13.24.

**Exercise 13.26.** Let $J \subset \mathbb{R}$ be an interval and $g : J \to \mathbb{R}_+$ be continuous with

$$\int_J g(x)\, dx = 1,$$

and let $f$ be another nonnegative continuous function on $J$. Assume there exists $M > 0$ such that $f(x) \leq Mg(x)$ for all $x \in J$, and let

$$H(f||g) = \int_J f(x) \ln \frac{f(x)}{g(x)}\, dx.$$

a) Show that $H(f||g)$ is well defined.

b) Define $u : J \to [0, M]$ by

$$u(x) = \frac{f(x)}{g(x)}.$$

Let $F(x) = x \ln x$ as before and assume also $\int_J f(x)\, dx = 1$. Show that

$$0 = F(1) = F(\int_J f(x)\, dx) = F(\int_J u(x) g(x)\, dx) < \int_J F(u(x)) g(x)\, dx = H(f||g).$$

Not sure if this of any use. Let's examine what we really need to assume on $f$ and $g$ here. If $f$ and $g$ are bounded Riemann integrable functions on some bounded interval $(a,b)$ and $\frac{1}{g}$ is bounded then it is also integrable by Theorem 7.5, making also $\frac{f}{g}$ integrable via Exercise 7.40. In that case we can introduce $y = \int_a^x g$ as new variable and verify that

$$\int_a^b \frac{f(x)}{g(x)}\,dx = \int_0^I \tilde{f}(y)\,dy, \quad I = \int_a^b g(x)\,dx,$$

directly from the definitions. To have an $\eta \in (0,I)$ with

$$\int_0^I \tilde{f}(y)\,dy = \eta I$$

we then need the continuity of $\tilde{f}$, and thereby of $f$. So with continuous functions on $[a,b]$ we're good if one the two is positive. What about continous integrable functions on $(-\infty,0]$? Same results it seems. So we can generalise to partitions of $\mathbb{R}$ and nonnegative continuous integrable functions $f$ and $g$ for which one the two is positive on each open interval of the partition

$$-\infty = x_0 < x_1 < \cdots < x_n = \infty.$$

# 14 Implicit functions

https://www.youtube.com/playlist?list=PLQgy2W8pIli-7124huziMvr6eTmFVOCuh

In the playlist the inverse function theorem comes first. I solve $f(x) = y$, assuming $f(0) = 0$, $f'$ continuous and invertible in $x = 0$, via the scheme

$$g_n(y) = x_n = x_{n-1} + f'(0)^{-1}(y - f(x), \quad x_0 = 0.$$

The iterates are shown to be convergent in an open ball with radius $\rho > 0$, $\rho$ chosen to have

$$|f'(x) - f'(0)| < \eta, \quad \gamma\eta < 1, \quad \gamma = |f'(0)^{-1}| > 0,$$

provided

$$|y| < (\frac{1}{\gamma} - \eta)\rho.$$

For each such $y$ the limit $x = g(y)$ of the sequence $x_n$ is the unique solution of $f(x) = y$ in the open ball with radius $\rho$. On this ball the inverses of $f'(x)$ exist[1].

The approximating functions $g_n$ satisfy the uniform Lipschitz condition[2]

$$|g_n(y) - g_n(\tilde{y})| \leq \frac{|y - \tilde{y}|}{\frac{1}{\gamma} - \eta},$$

en so does the limit $g$, which is continuously differentiable, with derivative

$$g'(y) = (f'(g(y))^{-1}.$$

The playlist concludes with the same result for $F(x, y)$, $F(0, 0) = 0$, proved by the same method, under the assumption that $\lambda \to F(x, y)$ is uniformly Lipschitz continuous and $(x, y) \to F_x(x, y)$ continuous. In the case that the corresponding Lipschitz constant is 1, the result is identical to the one above, except for

$$F(g(y), y) = 0, \quad \text{and} \quad g'(y) = -(F_x(g(y), y))^{-1}(F_y(g(y), y)),$$

valid if $F_y(g(y), y)$ exists. In the playlist I have $y = \lambda$ and $F(g(\lambda), \lambda) = 0$, in the notes below I have the equation $g(y) = x$ as the special case of $F(x, y) = 0$ solved for $y$, and the implicit function $f$ satisfying $F(x, f(x)) = 0$.

[1]This is new, the geometric series argument needs no further restrictions on the radii.
[2]Also new, below $\eta = \tilde{\varepsilon}$ and $M = \gamma$ do appear in the same combination.

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

$$F(x,y) = 0 \tag{14.1}$$

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

$$F(x,y) = g(y) - x, \tag{14.2}$$

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[3]. 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[4]. The next $y_n$ is uniquely defined by

$$F(x, y_{n-1}) + F_y(x, y_{n-1})(y_n - y_{n-1}) = 0.$$

[3]Fast convergence of this method will be shown in Section 12.2.

[4]Think of $F_y(x, y_{n-1})$ as the map $h \to F_y(x, y_{n-1})h$ .

For $n = 1, 2, \ldots$ we have

$$y_n = y_{n-1} - F_y(x, y_{n-1})^{-1} F(x, y_{n-1}), \quad \text{starting from} \quad y_0 = 0. \tag{14.3}$$

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

$$x \to y = f(x). \tag{14.4}$$

We then expect/hope that

$$F(x, f(x)) = 0, \tag{14.5}$$

and that $y = f(x)$ is the only solution of (14.1) near $y = 0$. If so we also ask which conditions will make $f$ continuous and differentiable in $x = 0$.

## 14.1 A simpler version of Newton's method

A direct proof of (fast) convergence of the sequence $y_n$ defined by (14.3) was given in Chapter 12.2 via an estimate of the form

$$|y_{n+1} - y_n| \leq C|y_n - y_{n-1}|^2 \tag{14.6}$$

and required a condition on the second derivative[5] 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

$$y_n = y_{n-1} - F_y(0, 0)^{-1} F(x, y_{n-1}), \tag{14.7}$$

and we look for an estimate which is very much like the estimate (3.6) for Heron's sequence: we lose the square in (14.6) 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.

**Theorem 14.1.** *Let* $\bar{\delta} > 0$, $\bar{\varepsilon} > 0$,

$$B = \{x \in \mathbb{R} : |x| < \bar{\delta}\}, \quad C = \{y \in \mathbb{R} : |y| < \bar{\varepsilon}\},$$

[5] In fact Lipschitz continuity of $y \to F_y(x, y)$ will suffice, see Section 12.2.

*and suppose that* $F : B \times C \to \mathrm{I\!R}$ *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 (14.7) 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 (14.7) 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 $F(x,y) = 0$.

## 14.2 Estimating the steps: convergence

How large can $y_1$ be if $F(x,y_0) = F(x,0) \ne 0$? If we set

$$M_0 = |F_y(0,0)^{-1}| > 0. \tag{14.8}$$

then[6]

$$|y_1| = |F_y(0,0)^{-1}F(x,0)| \le M_0\,|F(x,0)|. \tag{14.9}$$

If $F(x,y_1)$ is defined we can estimate the next step by

$$|y_2 - y_1| = |F_y(0,0)^{-1}F(x,y_1)| \le M_0\,|F(x,y_1)|$$

using (14.7) with $n = 2$. The trick however is to use (14.7) with both $n = 1$ and $n = 2$ via

$$y_2 - y_1 = y_1 - F_y(0,0)^{-1}F(x,y_1) - y_0 + F_y(0,0)^{-1}F(x,y_0)$$

[6]For future purposes we only use $|F_y(0,0)^{-1}k| \le M_0\,|k|$.

$$= F_y(0,0)^{-1}\left(F(x,y_0) - F(x,y_1) + F_y(0,0)y_1 - F_y(0,0)y_0\right),$$

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 (12.1), the mean value theorem in integral form[7], 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[8]

$$\int_0^1 \left(F_y(x, ty_0 + (1-t)y_1) - F_y(0,0)\right)\, dt\, (y_0 - y_1),$$

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[9]

$$[y_0, y_1] = \{ty_0 + (1-t)y_1 : 0 \le t \le 1\} \tag{14.10}$$

for fixed $x$. Therefore

$$|y_2 - y_1| \le M_0 \int_0^1 |(F_y(x, ty_0 + (1-t)y_1) - F_y(0,0)|\, dt\, |y_0 - y_1|. \tag{14.11}$$

We now ask that $(x,y) \to F_y(x,y)$ is continuous[10] 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

$$|x| \le \varepsilon \text{ en } |y| \le \varepsilon \implies |F_y(x,y) - F_y(0,0)| < \eta \tag{14.12}$$

holds. Note that instead of an $\varepsilon$-$\delta$-statement we used an $\eta,\varepsilon$-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 $|x| \le \delta$, for some $\delta > 0$ depending

[7]Which will also do for $F : X \times Y \to Y$.

[8]Look at (12.6), this argument is not restricted to $F : \mathbb{R}^2 \to \mathbb{R}$!

[9]This notation for $[y_0, y_1]$ does not require $y_0 < y_1$.

[10]For $F(x,y) = g(y) - x$ this means $g'$ continuous in 0.

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 (14.11) and (14.12) we have that $|x|, |y_0|, |y_1| \le \varepsilon$ 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

$$|y_3 - y_2| \le \theta\,|y_2 - y_1|,$$

provided $|y_2| < \varepsilon$, and so on.

Any $\theta < 1$ is now fine for our purposes[11]: as long as $|y_n| < \varepsilon$ it holds that[12]

$$|y_{n+1}| = |y_{n+1} - y_0| \le \underbrace{|y_{n+1} - y_n|}_{\le\theta|y_n - y_{n-1}|} + \cdots + \underbrace{|y_2 - y_1|}_{<\theta|y_1|} + |y_1| <$$

$$(\theta^n + \cdots + 1)\,|y_1| < \frac{|y_1|}{1-\theta} \le \frac{M_0|F(x,0)|}{1-\theta},$$

so

$$|y_{n+1}| < \frac{M_0|F(x,0)|}{1-\theta} < \frac{M_0\tilde{\varepsilon}}{1-\theta} = \frac{M_0\tilde{\varepsilon}}{1-M_0\eta} \tag{14.13}$$

if $|x| \le \tilde{\delta}$. Here $\tilde{\varepsilon} > 0$ is still to be chosen and $\tilde{\delta} > 0$ corresponds to $\tilde{\varepsilon}$ via the definition[13] of continuity of $x \to F(x,0)$ in $x = 0$.

Now choose

$$\eta_0 < \frac{1}{M_0}, \tag{14.14}$$

and then, given the corresponding $\varepsilon_0$ as in (14.12), a positive $\tilde{\varepsilon}_0$ such that

$$\frac{M_0\tilde{\varepsilon}_0}{1-M_0\eta_0} < \varepsilon_0, \quad \text{i.e.} \quad \tilde{\varepsilon}_0 < (\frac{1}{M_0} - \eta_0)\varepsilon_0.$$

Then let $\tilde{\delta}_0 > 0$ correspond to $\tilde{\varepsilon}_0 > 0$ via the definition[14] of continuity of $x \to F(x,0)$ in $x = 0$.

---

[11] In (3.6) we chose $\theta = \frac{1}{2}$ for the sake of simplicity only.

[12] In view of $1 + \theta + \theta^2 + \cdots = \frac{1}{1-\theta}$, see Section 1.5.

[13] With $\le \tilde{\delta}$ instead of $< \tilde{\delta}$.

[14] With $\le \tilde{\delta}_0$ instead of $< \tilde{\delta}_0$.

Thus the chain of alternating choices and continuity arguments is

$$M_0 = |F_y(0,0)^{-1}| \quad \xrightarrow{\textit{choose}} \quad \eta_0 < \frac{1}{M_0} \quad \xrightarrow[\textit{continuous in } (0,0)]{(x,y)\to F_y(x,y)} \quad \varepsilon_0$$

$$\xrightarrow{\textit{choose}} \quad \tilde{\varepsilon}_0 < (\frac{1}{M_0} - \eta_0)\varepsilon_0 \quad \xrightarrow[\textit{continuous in } 0]{x\to F(x,0)} \quad \tilde{\delta}_0$$

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

$$y \xrightarrow{\Phi} y - F_y(0,0)^{-1}F(x,y), \tag{14.15}$$

and the estimate

$$|\Phi(x,y) - \Phi(x,\tilde{y})| \le \theta\,|y - \tilde{y}| \tag{14.16}$$

with $\theta < 1$ and strict inequality if $y \ne \tilde{y}$. Equation $F(x,y) = 0$ is via (14.15) 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[15] 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

$$\forall (x,y) \in \bar{B}_{\delta_0} \times \bar{B}_{\varepsilon_0} : \quad F(x,y) = 0 \iff y = f(x), \tag{14.17}$$

and Theorem 14.1 is proved.

## 14.3 Differentiable implicit functions

The implicit function in Theorem 14.1 satisfies

$$|f(x)| \le \frac{M_0|F(x,0)|}{1 - M_0\eta_0}, \tag{14.18}$$

in which $\eta_0$ was chosen at the beginning of Section 14.2, see (14.14). Estimate (14.18) 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?

[15]Continuity follows from differentiability.

Use (12.1) to write

$$0 = F(x, f(x)) = F(x, 0) + F(x, f(x)) - F(x, 0)$$

$$= F(x, 0) + \int_0^1 F_y(x, tf(x))f(x)\,dt = F(x, 0) + F_y(0, 0)f(x) + R(x),$$

$$\text{with} \quad R(x) = \int_0^1 (F_y(x, tf(x)) - F_y(0, 0))f(x)\,dt. \tag{14.19}$$

Clearly $x \to F(x, 0)$ differentiable in $x = 0$ is the natural additional assumption, because then

$$0 = F(x, f(x)) = F_x(0, 0)x + r(x) + F_y(0, 0)f(x) + R(x), \tag{14.20}$$

with $r(x) = o(|x|)$ as $x \to 0$.

**Theorem 14.2.** *Let $f$ be as in Theorem 14.1. 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, although using the Lipschitz continuity of $f$ would simplify it, watch `https://youtu.be/n92I8ua6K-8` and further. Isolating $f(x)$ in (14.20) we have

$$f(x) = \underbrace{-F_y(0, 0)^{-1}F_x(0, 0)}_{f'(0)?}\, x \underbrace{-F_y(0, 0)^{-1}r(x) - F_y(0, 0)^{-1}R(x)}_{remainder}. \tag{14.21}$$

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[16] $\varepsilon > 0$ 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 (14.19) we use (14.12) again to conclude that

$$|R(x)| < \tilde{\eta}\,|f(x)| \quad \text{if} \quad |x| < \tilde{\varepsilon} \quad \text{and} \quad |f(x)| < \tilde{\varepsilon}. \tag{14.22}$$

[16]Earlier we only took one fixed $\varepsilon_0$ corresponding to one fixed $\eta_0$ as in (14.14).

The latter inequality will hold if $|x| < \tilde{\delta}$, $\tilde{\delta}$ corresponding to $\tilde{\varepsilon}$ in the established statement, via the construction and (14.18), 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 $\delta_r$ 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 (14.22). 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 (14.21) is $o(|x|)$ as $x \to 0$. This then completes the proof of Theorem 14.2.

**Exercise 14.3.** Actually the continuity of $f$ in $x = 0$ follows directly from (14.21) and (14.19) if we assume that $|y| = |f(x)| \le \varepsilon_0$ with $\varepsilon_0$ chosen via (14.12) for (14.14). Use (14.20) in the form

$$0 = F(x,y) = F(x,0) + F_y(0,0)y + \underbrace{\int_0^1 (F_y(x,y) - F_y(0,0))y\,dt}_{\text{in norm less than } \eta_0|y| \quad \text{if } |x|,|y| \le \varepsilon_0}\ , \tag{14.23}$$

and derive that for solutions $(x,y)$ of $F(x,y) = 0$ it holds that

$$|y| \le \frac{M_0|F(x,0)|}{1 - M_0\eta_0} \quad \text{if} \quad |x| \le \varepsilon_0 \text{ and } |y| \le \varepsilon_0. \tag{14.24}$$

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 $\varepsilon_0$

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

$$f'(x) = -F_y(x, f(x))^{-1} F_x(x, f(x)). \tag{14.25}$$

For every $x \in \bar{B}_{\delta_0}$ the validity of (14.25) 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 14.4.** *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 14.1 and 14.2, 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 $\delta_0$ and $\varepsilon_0$. 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.

## 14.4 Application to integral equations

This concerns smooth dependence of the solution of (7.15) on $\xi$, 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 (14.26). Then

$$\xi \to x(\xi)$$

is continuously differentiable, and $x_\xi$ 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[17]. The first steps are sketched below.

For $a, b \in \mathbb{R}$ met $0 \in [a, b]$ and $\xi \in \mathbb{R}$ introduce

$$x = \xi + \Phi(x) \quad \text{with} \quad (\Phi(x))(t) = \int_0^t f(x(s))\, ds, \tag{14.26}$$

defining a new $\Phi(x) \in C([a, b])$ given and ("old") function $x \in C([a, b])$. Theorem 14.4 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

$$(\Phi(x+h))(t) = \int_0^t f(x(s) + h(s))\, ds = \int_0^t [f(x(s) + \tau h(s))]_0^1\, ds$$

$$= \int_0^t \int_0^1 f'(x(s) + \tau h(s))h(s)\, d\tau\, ds$$

$$= \int_0^t \int_0^1 f'(x(s))h(s)\, d\tau\, ds + \underbrace{\int_0^t \int_0^1 (f'(x(s) + \tau h(s)) - f(x(s))h(s)\, d\tau\, ds}_{R(h;x)(t)}$$

$$= (\Phi'(x)h)(t) + R(h; x)(t),$$

[17] We shall also deal with pararameters in $f$, e.g. $f(x, \mu, \varepsilon)$ or so, see Section 16.5.

in which

$$h \xrightarrow{\Phi'(x)} \Phi'(x)h \quad \text{with} \quad (\Phi'(x)h)(t) = \int_0^t f'(x(s))h(s)\,ds, \tag{14.27}$$

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 $f'$.

## 14.5 For later: partial differentiability $\implies$ ?

Exercise 11.19 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 $\varepsilon$.

**Exercise 14.5.** 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 11.16 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

$$F(x, y) = F(x_0, y_0) + F_x(x_0, y_0)(x - x_0) + F_y(x_0, y_0)(y - y_0) + R_0 \quad (14.28)$$

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 14.6.** *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 14.7.** 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 $\Phi$ is differentiable in $(x_0, y_0)$. Use (12.5) to prove this statement.

**Exercise 14.8.** If $X$, $Y$, $Z$ are normed spaces, $Z$ complete, and $\Phi : X \times Y \to Z$ has partial functions with partial derivatives $\Phi_x$ and $\Phi_y$ continuous on an open set $O$ in $X \times Y$, then $\Phi$ is differentiable in every point of $O$ and $\Phi' : O \to L(X \times Y, Z)$ is continuous and defined in every $(x_0, y_0) \in O$.

## 14.6 Stationary under a constraint

Suppose $\Phi$ 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

$$F_x(0, 0) + F_y(0, 0)f'(0) = 0. \tag{14.29}$$

In practice, $f$ is the implicit function in Theorems 14.1 and 14.2. Then $y = f(x)$ describes the solution set of $F(x, y) = 0$ near $(0, 0)$, and we are interested in the restriction of $\Phi$ to the zero set of $F$. Clearly

$$x \xrightarrow{\phi} \phi(x) = \Phi(x, f(x))$$

is differentiable in $x = 0$, with

$$\phi'(0) = \Phi_x(0,0) + \Phi_y(0,0)f'(0). \tag{14.30}$$

If $F_y(0,0)$ is invertible it follows from (14.29) and (14.30) that

$$\phi'(0) = 0 \iff \Phi_x(0,0) = \Phi_y(0,0)F_y(0,0)^{-1}F_x(0,0). \tag{14.31}$$

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 the Lagrange multiplier theorem which will be discussed in Chapter 17, starting from (14.31). An instructive example, which looks useless at first sight, is $F(x,y) = g(y)$ with $\Phi(x,y) = y - f(x)$. With a parameter $\theta$ and $\Phi(x,y) = y - f(x,\theta)$ it's even more instructive, see Section 17.4.

# 15 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 15.10 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. Its dual space $X^*$ is by definition 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 $\phi$. 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

$$X \ni x \xrightarrow{Q} (Sx)(x) = \langle Sx, x \rangle \in \mathbb{R} \tag{15.1}$$

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 (11.15) introduced in Chapter 11.4.

[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 15.1.** 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 30.4. 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 15.2.** 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 (15.1)? If so we say that $F$ and $Q$ are conjugate functions.

## 15.1 Intermezzo: second order partial derivatives

**Theorem 15.3.** *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 $\eta$ 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

$$g(x,y) - g(0,y) - g(x,0) + g(0,0) = g_{yx}(0,0)xy + y\,o(r)$$

$$= g_{yx}(0,0)xy + o(r^2) = g_{xy}(0,0)xy + o(r^2) \tag{15.2}$$

for $r = \sqrt{x^2+y^2} \to 0$. The second version under (15.2) follows by interchanging the roles of $x$ and $y$ and implies $g_{yx}(0,0) = g_{xy}(0,0)$.

## 15.2 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 $t$,

$$f'(t) = F'(tx)(x) = \langle F'(tx), x\rangle, \tag{15.3}$$

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

$$F(x) = f(1) = \int_0^1 (1-t) f''(t)\,dt, \tag{15.4}$$

see also Theorem 13.7.

**Exercise 15.4.** Give a direct proof of (15.3).

The differentiability of $t \to F'(tx)x = f'(t)$ will follow from differentiability of

$$x \to F'(x) \in X^*$$

in points $\xi$ near 0, which means that

$$F'(x) = F'(\xi) + F''(\xi)(x-\xi) + R(x;\xi), \tag{15.5}$$

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 (15.5) this becomes

$$\begin{aligned} F'(tx) &= F'(t_0x) + F''(t_0x)(tx - t_0x) + R(tx; t_0x) \\ &= F'(t_0x) + (t-t_0)F''(t_0x)x + R(tx;t_0x) \end{aligned}$$

in $X^*$, and (15.3) 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

$$f''(t) = \langle F''(tx)x, x\rangle. \tag{15.6}$$

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 $\mathcal{O}$.

## 15.3 The second derivative as symmetric bilinear form

**Theorem 15.5.** *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*

$$(h,k) \xrightarrow{F''(\xi)} (F''(\xi)h)k = \langle F''(\xi)h, k\rangle \in \mathbb{R} \tag{15.7}$$

*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 15.5 is proved by Exercise 15.6 and Theorem 15.3.

**Exercise 15.6.** 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

$$S \in S(X, X^*) \subset L(X, X^*). \tag{15.8}$$

With also $F''(tx) \in S(X, X^*)$ we have from (15.4) 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

$$F(x) = \langle \int_0^1 (1-t)F''(tx)\, dt\, x, x\rangle = \langle \Phi_x x, x\rangle, \tag{15.9}$$

in which

$$\Phi_x = \int_0^1 (1-t)F''(tx)\, dt \in S(X, X^*). \tag{15.10}$$

Here we use a subscript to denote the $x$-dependence of the operator $\Phi_x$ which acts on $X$.

It follows that

$$F(x) = \frac{1}{2}\langle F''(0)x, x\rangle + \langle \int_0^1 (1-t)(F''(tx) - F''(0))\, dt\, x, x\rangle$$

$$= \langle \Phi_0 x, x\rangle + o(|x|_X^2), \tag{15.11}$$

as $|x|_X \to 0$ if $F''$ is continuous in $x = 0$. The quadratic function defined by

$$Q_0(x, x) = \langle \Phi_0 x, x\rangle = \frac{1}{2}\langle F''(0)x, x\rangle \tag{15.12}$$

=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 15.7.** Check that continuity 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 15.8.** 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 15.9.** 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 $|x| < r$.

## 15.4 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

$$\langle \Phi_x x, x\rangle = (\Phi_x x)\, x = (\Phi_0 y)\, y = \langle \Phi_0 y, y\rangle \tag{15.13}$$

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 (15.13) holds if the map

$$h \to \Phi h \quad \text{is equal to the map} \quad h \to \Phi_0 T h \circ T = \kappa_0(T,T)\, h. \tag{15.14}$$

This is an $L(X, X^*)$-valued “quadratic” equation for $T \in L(X, X)$.

Abstractly we may write (15.14) as

$$\kappa_0(T, T) = \Phi, \tag{15.15}$$

in which

$$X \times X \xrightarrow{\kappa_0} L(X, X^*)$$

is the bilinear form defined by

$$h \to \kappa_0(T, U)\, h = \Phi_0 Th \circ U.$$

Clearly $T = I$ is a solution of (15.14) when $\Phi = \Phi_0$. We want a solution $T = T_x$ for $\Phi = \Phi_x$ given by (15.10) close to $\Phi_0$. If you like you can skip Section 15.5 and jump to (15.26), or even Exercise 15.13. Just put $T = I + H$ in (15.14) and see what you can get[5].

## 15.5 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 (15.15) rewrites as

$$\underbrace{\Phi_0 Hh + \Phi_0 h \circ H + \Phi_0 Hh \circ H}_{\chi_0(H)h} = (\Phi_x - \Phi_0)h \tag{15.16}$$

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 $\Phi_0$ in the “coefficients” of the two linear terms and one quadratic term. Writing (15.16) as

$$\chi_0(H) = \Phi_x - \Phi_0, \tag{15.17}$$

the right hand side is in $S(X, X^*)$.

[5]But that’s not how I found equation (15.27).

Look at (15.16). Clearly the derivative of $\chi_0$ 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

$$\langle \chi_0'(0)Hh, k\rangle = \langle \Phi_0 Hh, k\rangle + \langle \Phi_0 Hk, h\rangle, \tag{15.18}$$

we have that $\chi_0'(0)H \in S(X, X^*)$. Thus the invertibility condition cannot be that

$$\forall_{h\in X} :\ \chi_0'(I)H = \Phi_0 Hh + \Phi_0 h \circ H = Ch \tag{15.19}$$

is solvable for every $C \in L(X, X^*)$, while (15.19) 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 (15.18) reduces to

$$\langle \chi_0'(0)Hh, k\rangle = 2\langle \Phi_0 Hh, k\rangle, \tag{15.20}$$

and the invertibility condition (15.19) becomes

$$2\Phi_0 H = C, \tag{15.21}$$

which is solvable for $H$ as

$$H = \frac{1}{2}\Phi_0^{-1}C \tag{15.22}$$

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]

$$2Kh + Kh \circ \Phi_0^{-1}K) = (\Phi_x - \Phi_0)h \tag{15.23}$$

for all $h \in X$, which amounts to the equation

$$2K + T_0(K) = C_x = \Phi_x - \Phi_0 \tag{15.24}$$

for $K \in S(X, X^*)$, in which the quadratic term is given by

$$T_0 : S(X, X^*) \to S(X, X^*), \quad T_0(K)h = Kh \circ (\Phi_0^{-1}K) \tag{15.25}$$

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 (15.23) follows directly from (15.16).

## 15.6 Yes, but main result via power series instead

**Theorem 15.10.** *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 15.11.** Prove Theorem 15.10 by applying the implicit function theorem to (15.23).

**Remark 15.12.** *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 30.4.*

---

[8] See Chapter 30.

[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 (15.15) in the space

$$\{T \in L(X, X^*) : \Phi_0 T \in S(X, X^*)\}. \tag{15.26}$$

Via $T = I + H$ and (15.16) equation (15.15) was equivalent to (15.23) for

$$K = \Phi_0 H \in S(X, X^*).$$

We now return to an equation for $H$. Write (15.23) 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 $\Phi_0^{-1}$, the equation to solve for $H$, still under the assumption that $\Phi_0 H \in S(X, X^*)$, is

$$2H + HH = \Phi_0^{-1}\Phi - I = P, \tag{15.27}$$

in which $P \in L(X, X^*)$ also has $\Phi_0 P \in S(X, X^*)$.

**Exercise 15.13.** Derive (15.27) directly from (15.15), 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

$$I + 2H + H^2 = I + P \quad \text{in} \quad L(X) = L(X, X). \tag{15.28}$$

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 11.11. The same power series tricks[10] give

$$T = I + H = I + \frac{1}{2}\,P - \frac{1}{2!}\,\frac{1}{2}\,\frac{1}{2}\,P^2 + \frac{1}{3!}\,\frac{1}{2}\,\frac{1}{2}\,\frac{3}{2}\,P^3 - \frac{1}{4!}\,\frac{1}{2}\,\frac{1}{2}\,\frac{3}{2}\,\frac{5}{2}\,P^4 + \cdots \tag{15.29}$$

if $|P| < 1$, and so $y = T_x x$ with

$$T_x = I + E_x - \frac{1}{2!}\,E_x^2 + \frac{1 \cdot 3}{3!}\,E_x^3 - \frac{1 \cdot 3 \cdot 5}{4!}\,E_x^4 + \cdots \tag{15.30}$$

and

$$E_x = \int_0^1 (1 - t)(F''(0)^{-1}F''(tx) - I)\,dt, \tag{15.31}$$

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 15.14.** See if you can give a direct derivation of (15.30) and (15.31) 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 $L(X)$.

# 16 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 $r, \theta$ 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 $\theta$ 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 $2\pi$ 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 16.1.** 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 $\gamma$ the map

$$z \to \gamma z \tag{16.1}$$

is a rotation[1] around 0 followed by a point multiplication with 0 as fixed point, see (16.8) and Exercise 16.5.

## 16.1 Intermezzo: algebra's main theorem

The set $\mathbb{C}$ is algebraically closed: every polynomial

$$P(z) = \sum_{k=0}^{n-1} \alpha_k z^k + z^n \tag{16.2}$$

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

$$P(z) = (z - z_1) \cdots (z - z_n) \quad \text{with} \quad z_1, \ldots, z_n \in \mathbb{C}. \tag{16.3}$$

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

$$Q(w) = 1 + \sum_{k=1}^{n} \gamma_k w^k \tag{16.4}$$

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 16.1 we have

$$Q(w) = 1 + \sum_{k=1}^{n} c_k r^k \exp(i(\phi_k + k\theta)), \tag{16.5}$$

an expression[2] in which the $\phi_k$ are parameters and $r > 0$ can be taken as small as we want. Exercise 16.2 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 (16.2).

**Exercise 16.2.** 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 (16.6) 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

$$|P(z)| = |\sum_{k=0}^{n-1} \alpha_k z^k + z^n| \geq |z^n| - |\sum_{k=0}^{n-1} \alpha_k z^k| \geq r^n - \sum_{k=0}^{n-1} |\alpha_k| r^k. \tag{16.6}$$

[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 $\infty$ as $r$ increases from $r_0$ to $\infty$. 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.

## 16.2 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

$$w = H(z) = H(z_0) + \gamma(z - z_0) + T(z; z_0) \tag{16.7}$$
$$= H(z_0) + H'(z_0)(z - z_0) + o(|z - z_0|)$$

as $z \to z_0$.

If we unpack (16.7), 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 (16.7) unpacks as

$$\gamma(z - z_0) = (\alpha + i\beta)(h + ik) = \alpha h - \beta k + i(\beta h + \alpha k).$$

This corresponds to

$$\begin{pmatrix} \alpha h - \beta k \\ \beta h + \alpha k \end{pmatrix} = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} h \\ k \end{pmatrix}, \tag{16.8}$$

in which the matrix describes the map (16.1).

The complex expansion (16.7) 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 $2 \times 2$ 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 16.3.** *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 16.4.** Prove Theorem 16.3.

**Exercise 16.5.** Examine (16.1) using (16.8).

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

$$Ah = \begin{pmatrix} (Ah)_1 \\ (Ah)_2 \end{pmatrix} = \begin{pmatrix} A_{11}h_1 + A_{12}h_2 \\ A_{21}h_1 + A_{22}h_2 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \tag{16.9}$$

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]

$$|Ah|_2 \leq |A|_2 \, |h|_2 . \tag{16.10}$$

In (16.10) 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 (18.6).
[6]Actually this 2-norm of $A$ is called the Frobenius norm of $A$.
[7]If not note that (16.10) 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

$$F(x) = F(a) + A(x-a) + R(x;a) \quad \text{and} \quad |R(x;a)|_2 = o(|x-a|_2) \quad (16.11)$$

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

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(a_1, a_2) & \frac{\partial F_1}{\partial x_2}(a_1, a_2) \\ \frac{\partial F_2}{\partial x_1}(a_1, a_2) & \frac{\partial F_2}{\partial x_2}(a_1, a_2) \end{pmatrix} = F'(a) = DF(a) \quad (16.12)$$

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 14.5 for $F : \mathbb{R}^2 \to \mathbb{R}$ and $F : X \times Y \to \mathbb{R}$.

**Exercise 16.6.** State and prove a theorem for $F : \mathbb{R}^2 \to \mathbb{R}$ by specializing Theorem 14.6 to $X = Y = \mathbb{R}$ and generalise to $F : \mathbb{R}^m \to \mathbb{R}$ and $F : \mathbb{R}^m \to \mathbb{R}^n$.

## 16.3 Cauchy-Riemann equations, harmonic functions

Have another look at Theorem 16.3 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 16.7.** Show that $\alpha$ and $\beta$ are then given by

$$\alpha = \frac{\partial u}{\partial x} \quad \text{and} \quad \beta = -\frac{\partial u}{\partial y}, \tag{16.13}$$

evaluated in $(x,y) = (x_0, y_0)$.

Thus Theorem 16.3 says that $u$ and $v$, as functions of $x$ and $y$, must satisfy the so-called Cauchy-Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y} \tag{16.14}$$

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

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = \Delta v, \tag{16.15}$$

in which the differential operator $\Delta$, the *Laplacian*, occurs. This $\Delta$ is a feast to study, but not now. Here we want to be sure under what conditions (16.15) makes sense. We copy Theorem 15.3 from Section 15.1.

**Theorem 16.8.** *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 (16.15) 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 15.3, 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

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \tag{16.16}$$

without any $v$ interfering[10].

There are many non-constant solutions of (16.16). Indeed, you should have noticed

$$x, y, x^2 - y^2, 2xy, x^3 - 3xy^2, 3x^2y - y^3, x^4 - 6x^2y^2 + y^4, 4x^3y - 4xy^3, \ldots \tag{16.17}$$

above.

**Exercise 16.9.** 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

$$w \to \log w = \ln |w| + i \arg w.$$

.

## 16.4 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 $|y|, |b| \leq s$, we have

$$y^n = b^n + nb^{n-1}(y-b) + R_{b,n}(y), \quad |R_{b,n}(y)| \leq \frac{n(n-1)s^m}{2} K^2, \quad K = \frac{|y-b|}{s}.$$

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

$$|R_{2,2}| \leq \frac{1}{4} m^2 n^2 r^m s^n H^2 K^2,$$

$$|R_{21}| \leq \frac{1}{2} m^2 n r^m s^n H^2 K \leq \frac{1}{2} m^2 n^2 r^m s^n H^2 K,$$
$$|R_{12}| \leq \frac{1}{2} mn^2 r^m s^n H K^2 \leq \frac{1}{2} m^2 n^2 r^m s^n H K^2,$$

and also, a little less obvious maybe,

$$|R_2| \leq \frac{1}{4}(m^2 + n^2) r^m s^n (H^2 + K^2),$$

we conclude that

$$x^m y^n = a^m b^n + ma^{m-1}b^n(x-a) + na^m b^{n-1}(y-b) + \underbrace{R_{a,b,m,n}(x,y)}_{R}, \quad (16.18)$$

in which

$$|R| \leq \frac{r^m s^n}{4} \left( m^2 n^2 HK(HK + 2H + 2K) + (m^2 + n^2)(H^2 + K^2) \right). \quad (16.19)$$

The perhaps less obvious estimate for $R_2$ follows via

$$|R_2 \leq |s^n R_{a,m}(x)| + |mnr^{m-1}s^{n-1}(x-a)(y-b)| + |r^m R_{b,n}(y)| \leq$$

[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 (16.18) by coefficients $\alpha_{mn}$ 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

$$\sum_{m,n \in \mathbb{N}_0} (m^2 + n^2)\, |\alpha_{mn}| r^m s^n \quad \text{and} \quad \sum_{m,n \in \mathbb{N}_0} m^2 n^2\, |\alpha_{mn}| r^m s^n \tag{16.20}$$

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 (16.20).

**Exercise 16.10.** Fill in the details of the above proof. Show in addition that the convergence of

$$\sum_{m,n \in \mathbb{N}_0} (m^2 + n^2)\, |\alpha_{mn}| R^{m+n} \quad \text{and} \quad \sum_{m,n \in \mathbb{N}_0} m^2 n^2\, |\alpha_{mn}| R^{m+n} \tag{16.21}$$

suffices to have $P(x, y)$ exist as a differentiable function on the disk

$$\{(x, y) \in \mathbb{R}^2 : \, x^2 + y^2 < R\}.$$

## 16.5 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 $\mu$-systems.

In the special case that all the coefficients are zero the $\mu$-systems reduce to

$$\frac{dx}{dt} = \mu x - y;$$

$$\frac{dy}{dt} = x + \mu y.$$

The reduced $\mu$-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 $\mu xy$-space together form the set of all bounded solution orbits of the reduced $\mu$-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 $\mu$-system.

[13] Namely $x = \varepsilon \cos t, y = \varepsilon \sin t$, in which $\varepsilon > 0$ is not necessarily small.

**Exercise 16.11.** Assume that the coefficients $a_{mn}$ and $b_{mn}$ are bounded. Use Section 16.4 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

$$|a_{mn}| \le 1 \quad \text{and} \quad |b_{mn}| \le 1 \quad \text{for all} \quad m, n \in \mathbb{N} \quad \text{with} \quad m + n \ge 2, \tag{16.22}$$

and introduce polar coordinates $x = r\cos\theta, y = r\sin\theta$ to transform solutions of the $\mu$-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 16.12.** Use the chain rule[14] and Section 11.3 to determine the expressions for $\alpha_n$ and $\beta_n$ expressed in terms of $c = \cos\theta, s = \sin\theta, p_n(c,s), q_n(c,s)$. Show that

$$|\alpha_n| \le n \quad \text{and} \quad |\beta_n| \le n \quad \text{for all} \quad n \in \mathbb{N}_2,$$

and denoting the $r$-dependent part of the right hand side of the $\theta$-equation by

$$-\rho = \beta_2(\theta)r + \beta_3(\theta)r^2 + \cdots$$

that

$$|\rho| \le 2r + 3r^2 + 4r^3 + \cdots = \frac{r(2-r)}{(1-r)^2} < 1,$$

if $0 < r < 2 - \sqrt{2}$.

**Exercise 16.13.** 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 $\theta$, and that

$$\frac{dr}{d\theta} = r_\theta = \mu r + A_3(\theta,\mu)r^2 + A_4(\theta,\mu)r^3 + A_5(\theta,\mu)r^4 + \cdots , \tag{16.23}$$

[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 $\mu$ appears. Hint:

$$\frac{1}{1-\rho} = 1 + \rho + \rho^2 + \rho^3 + \cdots = \sum_{n=0}^{\infty} \rho^n.$$

**Exercise 16.14.** 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 16.15.** Show that

$$\int_0^{2\pi} A_3(\theta, \mu) d\theta = 0.$$

**Exercise 16.16.** 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 $\theta$. Why can this equation have no $2\pi$-periodic solutions? Hint: you should get an equation in which only $\frac{dw}{d\theta}$, $w$ and $A_3$ appear. Integrate from $0$ to $2\pi$ to derive a contradiction if $w(\theta)$ is a (positive) $2\pi$-periodic solution.

Consider (16.23) with $r(0) = \varepsilon > 0$ as initial value. For the original $\mu$-system this corresponds to the solution with $x(0) = \varepsilon, y(0) = 0$. Now scale $r$ by setting $r = \varepsilon R$. Then (16.23) becomes

$$\frac{dR}{d\theta} = R_\theta = \mu R + \varepsilon A_3(\theta, \mu) R^2 + \varepsilon^2 A_4(\theta, \mu) R^3 + \varepsilon^3 A_5(\theta, \mu) R^4 + \cdots, \quad (16.24)$$

and we look for solutions with $R(0) = 1$. Note that the explicit estimate in Exercise 16.14 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 $\mu$ and small $\varepsilon$, 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 14.1 and 14.2? The answer is yes, via what we already started in Section 14.4.

# 17 Stationary under constraints

This topic was started in Section 14.6 with the remarkable formula

$$\Phi_x = \Phi_y (F_y)^{-1} F_x \tag{17.25}$$

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 14.2 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

$$S = \{(x, y) : F(x, y) = F(0, 0)\} \tag{17.26}$$

is described as the graph of an implicitly defined 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 $2 \times 1$ 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 invertibility of the $1 \times 1$ matrix

$$A = F_y \tag{17.27}$$

in $(0, 0)$. Differentiability of

$$(x, y) \to \Phi(x, y)$$

sufficed to have (17.25) as both necessary and sufficient for $\phi'(x) = 0$, not only in $x = 0$ but as long as $F_y(x, f(x))$ is invertible on a whole neighbourhood of $x = 0$ in which $f(x)$ was constructed.

## 17.1 The method of Lagrange

This abstract section is part of a story line that started for the simplest concrete case in Section 14.6 and continues in Section 17.2 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 (14.31) 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 17.2, for now we write it as (17.25), 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

$$\Phi' = \Lambda \circ F'. \tag{17.28}$$

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 30.4. More details to follow perhaps, but not needed for the next section.

## 17.2 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 14 are essentially unchanged, beginning with (17.25) as the characterisation for

$$x \xrightarrow{\phi} \Phi(x, f(x))$$

being stationary, see (14.29) in Section 14.6.

Let's see how all this unpacks to give the method of Lagrange multipliers when we read (17.25) as a statement for Jacobi matrices and the corresponding linear maps. We write (17.25) in transposed form as

$$\nabla_x F \, (\nabla_y F)^{-1} \, \nabla \Phi_y = \nabla_x \Phi, \tag{17.29}$$

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 $F_x, F_y, \Phi_x, \Phi_y$.

Unpacking[1] the notation we have

$$\nabla_x F = (\nabla_x F_1 \;\; \nabla_x F_2 \;\; \nabla_x F_3) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_2}{\partial x_1} & \frac{\partial F_3}{\partial x_1} \\ \frac{\partial F_1}{\partial x_2} & \frac{\partial F_2}{\partial x_2} & \frac{\partial F_3}{\partial x_2} \end{pmatrix}$$

and likewise for $\nabla_y F$, which is a square $3 \times 3$ 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]

$$\nabla_y \Phi = \lambda_1 \nabla_y F_1 + \lambda_2 \nabla_y F_2 + \lambda_3 \nabla_y F_3 \tag{17.30}$$

with $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$. It follows that $\nabla_x F \, (\nabla_y F)^{-1}$ in the left hand side of (17.29) acts on (17.30) 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 30.28.

[4] This is possible in view of the invertibility condition imposed on $F_y$ in $(0,0,0,0,0)$.

by (17.29) again. With (17.30) this combines as

$$\nabla\Phi = \lambda_1\nabla F_1 + \lambda_2\nabla F_2 + \lambda_3\nabla F_3, \tag{17.31}$$

simply[5] because it holds for $\nabla_x$ and $\nabla_y$ 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 (17.31) holds in $(0,0,0,0,0)$. Note that (17.31) rewrites as

$$\nabla\Psi = 0 \quad \text{with} \quad \Psi = \Phi - \lambda_1 F_1 - \lambda_2 F_2 - \lambda_3 F_3,$$

which we may consider as a function of $x \in \mathbb{R}^5$ and $\lambda \in \mathbb{R}^3$ with the nice property that all $x$- and $\lambda$-derivative equal zero in the point we just characterised.

## 17.3 Application: Hölder's inequality

In (16.10) we had

$$|Ah|_2 \le |A|_2 \, |h|_2$$

as a special case of

$$|AB|_2 \le |A|_2 \, |B|_2 .$$

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

$$|\sum_{i=1}^n a_i b_i| \le \left(\sum_{i=1}^n |a_i|^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^n |b_i|^2\right)^{\frac{1}{2}} .$$

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

$$|\sum_{i=1}^n a_i b_i| \le |a|_p \, |b|_q , \tag{17.32}$$

if $|a|_p$ and $|b|_q$ are defined by

$$|a|_p^p = \sum_{i=1}^n |a_i|^p \quad \text{and} \quad |b|_q^q = \sum_{i=1}^n |b_i|^q . \tag{17.33}$$

Note that (17.32) is the Cauchy-Schwarz inequality of $p = q = 2$.

[5] No $3 \times 3$ matrix inverted here.

**Exercise 17.1.** Since (17.32) scales with $a$ and $b$ we can restrict the attention to vectors $a$ and $b$ for which $|a|_p = |b|_q = 1$. Explain!

Thus we introduce two boundary conditions

$$\phi(a_1, \dots, a_n) = |a_1|^p + \cdots + |a_n|^p = 1;$$

$$\psi(b_1, \dots, b_n) = |b_1|^q + \cdots + |b_n|^q = 1,$$

and max- and minimise

$$(a_1, \dots, a_n, b_1, \dots, b_n) \xrightarrow{F} a_1 b_1 + \cdots + a_n b_n.$$

**Exercise 17.2.** Explain why the maximum and the minimum of $F$ under the restriction $|a|_p = |b|_q = 1$ exist.

**Exercise 17.3.** Show that the functions $\phi$ and $\psi$ 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 $x \to p x^{p-1}$.

With two Lagrange multipliers $\lambda$ en $\mu$ we arrive at $2n$ equations

$$b_i = \lambda p a_i^{p-1}; \quad a_i = \mu q b_i^{q-1} \quad (i = 1, \dots, n)$$

to solve, together with

$$\sum_{i=1}^{n} |a_i|^p = \sum_{i=1}^{n} |b_i|^q = 1.$$

**Exercise 17.4.** Assume that $(p-1)(q-1) \neq 1$. Show that solutions have all $|a_i|$ equal and all $|b_i|$ equal, and therefore

$$\sum_{i=1}^{n} |a_i b_i| = n(\frac{1}{n})^{\frac{1}{p}+\frac{1}{q}} = n^{1-\frac{1}{p}-\frac{1}{q}}. \tag{17.34}$$

Deduce that (17.32) holds for $p > 1$ and $q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$.

## 17.4 Applications in machine learning

This is something Lotte told me, the original source seems to be the 1988 paper "A theoretical framework for back-propagation" by LeCun, Touresky, Hinton, and Sejnowski. These calculations continue from that discussion before me reading that paper. She was reading it to understand what was going on in `https://arxiv.org/abs/1806.07366`, the Neural ODE preprint of Chen et al. Have a look at the end of Section 14.6 before you consider for

$$x \in \mathbb{R}^{n_0}, \quad z_1 \in \mathbb{R}^{n_1}, \quad z_2 \in \mathbb{R}^{n_2}, \quad \dots \quad , \quad z_m \in \mathbb{R}^{n_m}$$

the system of transformations[6]

$$z_1 = f_1(x, \theta_1), \quad z_2 = f_2(z_1, \theta_2), \quad \dots \quad , \quad z_m = f_m(z_{m-1}, \theta_m)$$

with parameters

$$\theta_1 \in \mathbb{R}^{k_1}, \quad \dots \quad , \quad \theta_m \in \mathbb{R}^{k_m}.$$

These define

$$z_m = Z(x, \theta)$$

in $m$ steps as a vector valued function of $x$, with the $\theta$-variables as parameters.

### 17.4.1 Minimising some loss function

Suppose that given $x \in \mathbb{R}^{n_0}$ and $y \in \mathbb{R}^p$, a quantity

$$L(z_m, y)$$

has to be minimised by *learning* the $\theta$-parameters in the transformations $f_1, \dots, f_m$. This

$$L : \mathbb{R}^{n_m + p} \to \mathbb{R}$$

is called a *loss* function, which for now we assume to be as smooth as needed, and nonnegative. We then write

$$\tilde{L}(\theta, x, y) = L(Z(x, \theta), y),$$

but note that in *machine learning* practice[7] one considers growing sets of inputs $x$ and corresponding outputs $y$ for which one choice of all parameters has to do the job for all pairs $(x, y)$. The explanation below easily adapts to that setting.

[6] Assume all $f_i$ are smooth. This is Lotte's (4.5) and (4.12) with $m = N$.

[7] If my understanding is correct.

### 17.4.2 Lagrange multipliers to ...

To minimise $\tilde{L}$ we look for its stationary points and interpret the $z$-transformations in (17.38) as constraints

$$f_i(z_{i-1}, \theta_i) - z_i = 0, \quad i = 1, \ldots, m,$$

in which $z_0 = x$, and introduce[8]

$$\mathcal{L}(x, y, z, \lambda, \theta) = L(z_m, y) + \sum_{i=1}^{m} \lambda_i \cdot (f_i(z_{i-1}, \theta_i) - z_i)$$

as[9] a function of $z_1, \ldots, z_m$, $\lambda_1, \ldots, \lambda_m$, $x$, $y$ and $\theta_1, \ldots, \theta_m$. Under the constraints defined by (17.38) this expression reduces to

$$\tilde{L}(\theta, x, y) = \tilde{L}(\theta_1, \ldots, \theta_m, x, y) = L(z_m, y), \tag{17.35}$$

the quantity we want to have zero derivatives with respect to $\theta = (\theta_1, \ldots, \theta_m)$ for given fixed $x$ and $y$.

This trick allows a calculation of the $\theta$-derivatives of $\tilde{L}(\theta, x, y)$ which, as in Section 17.2, uses only transposes of Jacobian matrices, which I denote by gradients, namely

$$\nabla_{z_{m-1}} f_m, \ldots, \nabla_{z_0} f_1,$$

and begins by putting the $z$-gradients of $\mathcal{L}$ equal to zero. Compare this to the closing remarks in Section 17.2. The $z$-gradients are[10]

$$\nabla_{z_m} \mathcal{L} = \nabla_{z_m} L - \lambda_m \quad \text{and} \quad \nabla_{z_{i-1}} \mathcal{L} = \nabla_{z_{i-1}} (\lambda_i \cdot f_i) - \lambda_{i-1} \tag{17.36}$$

for $i = 2, \ldots, m$. Putting them equal to zero the resulting equations for the stationarity of

$$z \to L(z_m, y)$$

under the constraints (17.38) read[11]

$$\lambda_m = \nabla_{z_m} L, \quad \lambda_{m-1} = \nabla_{z_{m-1}} (\lambda_m \cdot f_m), \quad \ldots, \quad \lambda_1 = \nabla_{z_1} (\lambda_2 \cdot f_2),$$

with all $\lambda$'s en gradients column vectors.

---

[8] Lotte's (4.6), inner product dots, note the choice in the constraints with $f_i - z_i$.

[9] Or with $L(z_0, \ldots, z_m, y)$, or $L(z_m, y) + \hat{L}(z_0, \ldots, z_{m-1})$.

[10] Lotte's (4.7-8), (4.9) corresponds to what I mention at the end of Section 17.2.

[11] Lotte's (4.10-11), $\lambda_i = \lambda_l$ column vectors, *backwards passes*, $\lambda_m$ is like Chen's (34).

### 17.4.3 ... compute the gradient

This backwards scheme defines all $\lambda$-values introduced in the application of the Lagrange method, in terms of $y$ and the $z$-values already computed from $x$ and a particular choice of $\theta$-values used to compute $z_m$ from $x$ in $m$ steps. Note that with each[12]

$$\lambda_i = \nabla_{z_i}(\lambda_{i+1} \cdot f_{i+1})$$

we also have[13]

$$\nabla_{\theta_i}\tilde{L} = \nabla_{\theta_i}\mathcal{L} = \nabla_{\theta_i}(\lambda_i \cdot f_i) = \underbrace{\lambda_i \cdot \nabla_{\theta_i} f_i}_{\text{bad notation?}} = \underbrace{\nabla_{\theta_i} f_i \, \lambda_i}_{\text{fine with me}}$$

before we continue with[14]

$$\lambda_{i-1} = \nabla_{z_{i-1}}(\lambda_i \cdot f_i) = (\nabla_{z_{i-1}} f_i)\, \lambda_i \tag{17.37}$$

to compute $\nabla_{\theta_{i-1}}\tilde{L}$, and so on. Recalling that

$$z_i = f_i(z_{i-1}, \theta_i) \tag{17.38}$$

we see that we run forward and backward in

$$x = z_0 \to \cdots\cdots \to z_{i-1} \longrightarrow f_i(z_{i-1}, \theta_i) = z_i \longrightarrow \cdots\cdots\cdots \longrightarrow z_m$$

$$\lambda_1 \leftarrow \cdots \leftarrow \lambda_{i-1} = (\nabla_{z_{i-1}} f_i(z_{i-1}, \theta_i))\, \lambda_i \leftarrow \lambda_i \leftarrow \cdots \leftarrow \lambda_m = \nabla_{z_m} L(z_m, y)$$

$$\nabla_{\theta_i}\tilde{L} = \nabla_{\theta_i}\, \lambda_i \cdot f_i(z_{i-1}, \theta_i)$$

to systematically compute

$$\nabla_\theta \tilde{L},$$

using the transposed Jacobians which can be computed and stored on the way up from $x = z_0$ to $z_m$. In the last step also the partial derivatives of $\tilde{L}$ with respect to the coordinates of $x$ can be computed of course, simply by including these in $\theta_1$. The method also works for

$$L(z_m, y) = L(z_m) = \mathbf{e} \cdot z_m,$$

with $\mathbf{e}$ some unit vector, so we can compute the partial derivatives of the coordinates $Z(x, \theta)$ in a similar fashion. The only difference is that we start

[12] Lotte's (4.7), I keep the dots inside the sum for now.

[13] Lotte's (4.13), transposed Jacobians, see Section 17.2, acting on column vectors $\lambda_i$.

[14] Or $\lambda_{i-1} = (\nabla_{z_{i-1}} f_i)\lambda_i + \nabla_{z_{i-1}} L$, for the so-called (intermediate) adjoint states.

from $\lambda_m = \mathbf{e}$ then. Writing $u = Z(x,\theta)$ we have that the $k$-th component $u_k$ of $u$ has

$$\nabla_{\theta_j} u_k = \nabla_{\theta_j}(\lambda_j^k \cdot f_j(z_{j-1},\theta_j)), \quad \lambda_j^k = \nabla_{z_j}(\lambda_{j+1}^k \cdot f_{j+1}), \quad \lambda_m^k = \mathbf{e}_k, \quad (17.39)$$

with $\mathbf{e}_k$ the $k$-th unit vector. Thus we can evaluate $\nabla_{\theta_j}\tilde{L}$ via the chain rule for $\tilde{L}(Z(x,\theta),\theta)$ with $u = Z(x,\theta)$. I'll use this in Subsection 17.4.5, and after (17.61). For the standard loss function

$$L(z_m, y) = \frac{1}{2}|u-y|^2$$

we start from $\lambda_m = u - y$.

### 17.4.4 A poor man's tensor notation and the chain rule

We rewrite $\mathcal{L}$ as

$$\mathcal{L}(x,y,z,\lambda,\theta) = L(z_m,y) + \lambda_{k_i}^i\,(f_i^{k_i}(z_{i-1},\theta_i) - z_i^{k_i}),$$

summation over $i = 1,\ldots,m$, and, for each $i$, over the components[15] numbered by $k_i$. If you consider the $z_i$ as column vectors than the $\lambda^i$ are now row vectors[16]. We differentiate with respect to every $z_j$, each of which has components indexed by some index $q$ which I choose not to denote by $q_j$. But below I do use subscripts on indices used in a summation convention.

We compute and solve

$$\frac{\partial\mathcal{L}}{\partial z_{j-1}^q} = \lambda_{k_j}^j \frac{\partial f_j^{k_j}}{\partial z_{j-1}^q} - \lambda_q^{j-1} = 0$$

for $j = m,\ldots,2$, summation over $k_j$ only, starting from

$$\frac{\partial\mathcal{L}}{\partial z_m^q} = \frac{\partial L}{\partial z_m^q} - \lambda_q^m = 0,$$

no summation. This gives

$$\lambda_q^m = \frac{\partial L}{\partial z_m^q} \quad \text{and then} \quad \lambda_q^{j-1} = \lambda_{k_j}^j \frac{\partial f_j^{k_j}}{\partial z_{j-1}^q}, \quad j = m,\ldots,2, \qquad (17.40)$$

summation over $k_j$ for each $j$. Simultaneously we have

$$\frac{\partial\tilde{L}}{\partial\theta_j^p} = \lambda_{k_j}^j \frac{\partial f_j^{k_j}}{\partial\theta_j^p}, \quad j = m,\ldots,1, \qquad (17.41)$$

[15] So in $\lambda_k^i z_i^k$ we sum over $k$ first, and write $k = k_i$.

[16] You may prefer $\lambda_i^k z_k^i$, with $z_i$ row vectors and $\lambda^i$ column vectors.

with superscript $p$ for the components of each $\theta_j$. Note that (17.40) leads to

$$\lambda_q^m = \frac{\partial L}{\partial z_m^q}, \quad \lambda_q^{m-1} = \frac{\partial L}{\partial z_m^{k_m}} \frac{\partial f_m^{k_m}}{\partial z_{m-1}^q}, \quad \lambda_l^{m-2} = \frac{\partial L}{\partial z_m^{k_m}} \frac{\partial f_m^{k_m}}{\partial z_{m-1}^{k_{m-1}}} \frac{\partial f_{m-1}^{k_{m-1}}}{\partial z_{m-2}^q}, \dots,$$

and finally the $x$-derivatives of $\tilde{L}(x, \theta, y)$ with $j = 1$. This also follows from the chain rule, which is what the end of Section 14.6 alludes to. The less obvious expression is (17.41), which provides us with all components of $\nabla_\theta \tilde{L}$, layer by layer.

### 17.4.5 Along the gradient flow

This was written after reading a part in Feddrick's thesis about `https://arxiv.org/abs/1811.03804`. Let's denote the $z$-variable computed in the final step by[17]

$$u = z_m = f_m(z_{m-1}, \theta_m).$$

Then $u = u(x, \theta)$ is a function of all the $\theta$-parameters and $x = z_0$, and $\nabla_\theta \tilde{L}$ is the full $\theta$-gradient of $L(u(x, \theta), y)$. The gradient flow is the flow of the dynamical system

$$\dot{\theta} = -\nabla_\theta \tilde{L}, \tag{17.42}$$

and along that *weight updating* flow[18] we have, denoting the components of both $u$ and $\theta$ by superscripts, that[19] the equations

$$\dot{u}^j = \frac{du^j}{dt} = -\frac{\partial u^j}{\partial \theta^k} \frac{\partial \tilde{L}}{\partial \theta^k} = -\frac{\partial u^j}{\partial \theta^k} \frac{\partial L}{\partial u^i} \frac{\partial u^i}{\partial \theta^k} = -\underbrace{\nabla_\theta u^j \cdot \nabla_\theta u^i}_{G^{ij}} \frac{\partial L}{\partial u^i}$$

may be appended to (17.42). In conclusion, along the gradient flow we have that

$$\frac{\partial u}{dt} = -\nabla_\theta u \cdot \nabla_\theta \tilde{L}$$

rewrites as

$$\dot{u} = -G\, \nabla_u L, \tag{17.43}$$

in which $G$ is the $t$-dependent symmetric matrix consisting of the inner products of the full $\theta$-gradients of the components of $u$. Writing

$$A = \frac{\partial u}{\partial \theta}$$

[17] Feddrick used this notation, what follows only applies to $z_m$.
[18] Solved with Euler's forward method, stepsize $\alpha$ called learning rate, Lotte's (4.14).
[19] Summation convention for $k$ and $i$.

for the Jacobian matrix[20], the symmetric matrix $G$ is of the form

$$A\,A^T$$

and its eigenvalues[21] are the (nonnegative) singular values of $A$. Note that

$$G^{ij} = \nabla_\theta u^j \cdot \nabla_\theta u^i = \sum_{l=1}^{m} G_l^{ij}, \quad G_l^{ij} = \nabla_{\theta_l} u^j \cdot \nabla_{\theta_l} u^i, \tag{17.44}$$

is a sum of such matrices. If we use $\nabla_{\theta_l} u^k = \nabla_{\theta_l}(\lambda_l^k \cdot f_l(z_{l-1}, \theta_l))$ from (17.39) with $k = i, j$, then

$$G_l^{ij} = (\nabla_{\theta_l}(\lambda_l^i \cdot f_l(z_{l-1}, \theta_l))) \cdot (\nabla_{\theta_l}(\lambda_l^j \cdot f_l(z_{l-1}, \theta_l))) = \lambda_{lp}^i \lambda_{lq}^j \, \nabla_{\theta_l} f_l^p \cdot \nabla_{\theta_l} f_l^q,$$

in which we sum over $p, q$. Summing up and combining with (17.39) we have for $G_l^{ij}$ in (17.44) that

$$G_l^{ij} = \lambda_{lp}^i \lambda_{lq}^j \, \nabla_{\theta_l} f_l^p \cdot \nabla_{\theta_l} f_l^q, \quad \lambda_l^k = \nabla_{z_l}(\lambda_{l+1}^k \cdot f_{l+1}), \quad \lambda_m^k = \mathbf{e}_k, \tag{17.45}$$

in which the $\lambda^k$ are a basis for the solutions of the backward pass. Thus (17.43) becomes

$$\dot{u}^i = -\nabla_{\theta_l} f_l^p \cdot \nabla_{\theta_l} f_l^q \; \lambda_{lp}^i \lambda_{lq}^j \, \frac{\partial L}{\partial u^j},$$

summation over repeated indices $l, p, q, j$.

In the case that the loss function $L$ is given by

$$L(u, y) = \frac{1}{2}|u - y|^2,$$

the system (17.43) is linear:

$$\dot{u} = -G\,(u - y).$$

### 17.4.6 Along the neural network gradient flow

This is to evaluate (17.43) for neural networks. Needs to be double checked. Using column vectors $z_l$ for the layers, the maps $f_l$ in neural networks are of the form

$$z_l = \sigma_l(\underbrace{W_l z_{l-1} + b_l}_{x_l}),$$

[20]Which has typically many more columns than rows.
[21]See Chapter 18, Theorem 18.8.

in which the sigmiodal functions $\sigma_l$ act on $x_l$ componentwise to give $z_l$, via

$$z_l^p = \sigma_l(x_l^p), \quad x_l^p = w_{lq}^p z_{l-1}^q + b_l^p = \theta_{lq}^p z_{l-1}^q + \theta_{l0}^p,$$

with $\theta_{lq}^p$ collecting the parameters in the $l^{th}$ step, $q = 0, \ldots, n_{l-1}$, $p = 1, \ldots, n_l$. Changing to row vectors for the Lagrange multipliers we have that (17.39) rewrites with upper-lower summation convention as

$$\begin{gathered} \lambda_{l-1,i}^k = \lambda_{l,p}^k \tfrac{\partial f_l^p}{\partial z_{l-1}^i} = \lambda_{l,p}^k (\sigma_l'(x_l^p) w_{li}^p), \quad \lambda_{mi}^k = \delta_i^k, \\ x_l^p = w_{lq}^p z_{l-1}^q + b_l^p \end{gathered} \tag{17.46}$$

for the $\lambda$'s and subsequently

$$\frac{\partial \tilde{L}}{\partial w_{lj}^q} = \frac{\partial \tilde{L}}{\partial u^k} \frac{\partial u^k}{\partial w_{lj}^q} = \frac{\partial \tilde{L}}{\partial u^k} \lambda_{lp}^k \frac{\partial f_l^p}{\partial w_{lj}^q} = \underbrace{\lambda_{lp}^k \, \delta_{qp}}_{\lambda_{lq}^k} \, \sigma_l'(x_l^q) \, z_{l-1}^j \, \frac{\partial \tilde{L}}{\partial u^k},$$

and likewise

$$\frac{\partial \tilde{L}}{\partial b_l^q} = \frac{\partial \tilde{L}}{\partial u^k} \frac{\partial u^k}{\partial b_l^q} = \frac{\partial \tilde{L}}{\partial u^k} \lambda_{lp}^k \frac{\partial f_l^p}{\partial b_l^q} = \lambda_{lq}^k \, \sigma_l'(x_l^q) \, z_{l-1}^j \, \frac{\partial \tilde{L}}{\partial u^k}.$$

The gradient flow is given by

$$\dot{w}_{lj}^q = -\underbrace{\lambda_{lq}^k \, \sigma_l'(x_l^q)} \, z_{l-1}^j \, \frac{\partial \tilde{L}}{\partial u^k}, \quad \dot{b}_l^q = -\underbrace{\lambda_{lq}^k \, \sigma_l'(x_l^q)} \, \frac{\partial \tilde{L}}{\partial u^k}, \tag{17.47}$$

which we combine with

$$\frac{\partial u^i}{\partial w_{lj}^q} = \underbrace{\lambda_{lq}^i} \, \sigma_l'(x_l^q) \, z_{l-1}^j, \quad \frac{\partial u^i}{\partial b_l^q} = \lambda_{lq}^i \, \sigma_l'(x_l^q)$$

to get

$$\dot{u}^i + (\underbrace{\lambda_{lq}^i \, \sigma_l'(x_l^q) \, z_{l-1}^j \lambda_{lq}^k \, \sigma_l'(x_l^q) \, z_{l-1}^j}_{\text{sum over } j,\, q,\, l} + \underbrace{\lambda_{lq}^i \, \sigma_l'(x_l^q) \lambda_{lq}^k \, \sigma_l'(x_l^q)}_{\text{sum over } q,\, l}) \frac{\partial \tilde{L}}{\partial u^k} = 0.$$

This simplifies to

$$\dot{u}^i + (\underbrace{\sigma_l'(x_l^q) \sigma_l'(x_l^q) \, \lambda_{lq}^i \, \lambda_{lq}^k}_{\text{sum over } q,\, l} (1 + \underbrace{z_{l-1}^j \, z_{l-1}^j}_{\text{sum over } j})) \frac{\partial \tilde{L}}{\partial u^k} = 0, \tag{17.48}$$

from which we infer this truly amazing formula

$$G_l^{ik} = (1 + |z_{l-1}|^2)\, \underbrace{\sigma_l'(x_l^q)\, \lambda_{lq}^k\, \sigma_l'(x_l^q)\, \lambda_{lq}^i}_{\text{sum over } q},$$

which we should have seen quicker from (17.45). I kept the two identical factors $\sigma_l'(x_l^q)$, which we may want to absorb[22] in the $\lambda$'s by putting

$$\sigma_l'(x_l^q)\, \lambda_{lq}^k = \mu_{lq}^k, \tag{17.49}$$

which changes (17.47) into

$$\dot{w}_{lj}^q = -\mu_{lq}^k\, z_{l-1}^j \frac{\partial \tilde{L}}{\partial u^k}, \quad \dot{b}_l^q = -\mu_{lq}^k \frac{\partial \tilde{L}}{\partial u^k}, \tag{17.50}$$

and gives

$$\dot{u}^i = -G^{ik} \frac{\partial L}{\partial u^k}, \tag{17.51}$$

with

$$G^{ik} = \sum_{l=1}^m G_l^{ik}, \quad G_l^{ik} = (1 + |z_{l-1}|^2)\, \mu_{lq}^i\, \mu_{lq}^k \tag{17.52}$$

### 17.4.7 The space of neural network functions

Let's start from (17.38),

$$z_l = f_l(z_{l-1}, \theta_l), \quad l = 1, \ldots, m,$$

write $u = z_m \in \mathrm{IR}^{\mathrm{n}}$ as in Section 17.4.5, consider $u$ as a function of $x \in \mathrm{IR}^{\mathrm{N}}$ and all $\theta$, with the function $f_i$ as in Section 17.4.6, that is

$$z_l = \sigma_l(\underbrace{W_l z_{l-1} + b_l}_{x_l}).$$

We assume that all $\sigma_l$ are continuous. Then for every $a \in \mathrm{IR}^{\mathrm{n}}$ the function

$$x \to a \cdot z_m$$

is a function with parameters all $W$-, $a$- and $b$-components. *Varying $a$ over* $\mathrm{IR}^{\mathrm{n}}$ we get a *linear* space of (continuous) functions of $x$, in which we can restrict $x$ to[23] some closed bounded box $I$ in $\mathrm{IR}^{\mathrm{N}}$. Thus we get a linear

[22] Which perhaps leads to rewriting (17.46).
[23] To avoid complicated limit behaviour for large $x$.

subspace $\mathcal{N}$ of $C(I)$, the space of continuous functions on $I$. A natural question to ask is whether the closure of $\mathcal{N}$ in the maximum norm is $C(I)$.

If not then there is a nontrivial continuous linear functional $\Phi : C(I) \to \mathbb{R}$ such that $\Phi(f) = 0$ for all $f \in \mathcal{N}$. Feddrick formulates an already old result that says this is impossible if $m = 1$, that is, if the functions in $\mathcal{N}$ are all functions

$$f(x) = \sum_{i=1}^{n} a_i \sigma(A_i(x)),$$

with $A_i : \mathbb{R}^{\mathrm{N}} \to \mathbb{R}$ affine for $i = 1$ to some arbitrary $n$, $a_1, \ldots, a_n \in \mathbb{R}$, and $\sigma$ a fixed continuous function that is *discriminatory* for the signed Borel measures $\mu$ on $I$. Via the Riesz Representation[24]

$$\Phi(f) = \int_I f d\mu$$

these are exactly the continuous linear functionals on $C(I)$. Note that $\Phi(f) = 0$ for all $f \in \mathcal{N}$ is equivalent to $\Phi(f) = 0$ for all $f$ of the form

$$f = \sigma \circ A$$

with $A : \mathbb{R}^{\mathrm{N}} \to \mathbb{R}$ affine. This is because $\mathcal{N}$ is the linear space spanned by all such $\sigma \circ A$.

We may just as well say that $\sigma$ is discriminatory for all $\Phi$ in the dual space of the Banach space[25] $C(I)$ if $\Phi(\sigma \circ A) = 0$ for all affine $A : \mathbb{R}^{\mathrm{N}} \to \mathbb{R}$ implies that $\Phi$ is the zero functional. For such discriminatory $\sigma$ it is then immediate that the closure of $\mathcal{N}$ in the maximum norm is $C(I)$. Recall $\Phi(\sigma \circ A) = 0$ is equivalent to $\int_I \sigma \circ A \, d\mu = 0$ for the representing signed Borel measure $\mu$.

Now assume that $\sigma$ is continuous and nondecreasing, with $\sigma(-\infty) = 0$ and $\sigma(\infty) = 1$, let $H$ be a half space in $\mathbb{R}^{\mathrm{n}}$, and $\alpha \in (0, 1)$. Taking a suitable sequence $A_n$ we can make $\sigma \circ A_n$ converge pointwise to a function $S_{H,\alpha}$ which is 1 on $H$, $\alpha$ on $\partial H$, and 0 elsewhere (on the other side of $H$). It follows by the dominated convergence theorem that

$$\int_I S_{H,\alpha} \, d\mu = 0$$

for all $H$ and $\alpha$. In particular

$$\int_I s_\alpha \circ A \, d\mu = 0$$

[24]This is the Riesz-Markov-Kakutani representation Theorem, not treated in the notes.
[25]See Chapter 4 and further.

for all step functions which have values 0, $\alpha \in (0,1)$ and 1 on the left, in 0, on the right, and for all linear functions $A$. Such step functions can be linearly combined to obtain in suitable limits both cos and sin. It thus follows that the Fourier transform[26] of $\mu$ is zero, a contradiction.

### 17.4.8 Residual networks

In the special case that

$$f_i(z_{i-1}, \theta_i) = z_{i-1} + hg_i(z_{i-1}, \theta_i)$$

we have

$$\mathcal{L}(x, y, z, \lambda, \theta) = L(z_m, y) + \sum_{i=1}^{m} \lambda_i \cdot (z_{i-1} + hg_i(z_{i-1}, \theta_i) - z_i)$$

$$= L(z_m, y) + h \sum_{i=1}^{m} \lambda_i \cdot (g_i(z_{i-1}, \theta_i) - \frac{z_i - z_{i-1}}{h}), \tag{17.53}$$

in which we recognise a discretisation of (17.60) below if $mh = 1$. We see that (17.36) becomes

$$\nabla_{z_m} \mathcal{L} = \nabla_{z_m} L - \lambda_m, \quad \nabla_{z_{i-1}} \mathcal{L} = h \left( \nabla_{z_{i-1}} (\lambda_i \cdot g_i) + \frac{\lambda_i - \lambda_{i-1}}{h} \right)$$

In machine learning the resulting explicit Euler forward scheme with $h = 1$ is called a residual network. Such schemes allow calculations in which $L(z_m)$ is replaced by $L(x + g_1(x, \theta_1) + g_2(z_1, \theta_2) + \cdots + g_m(z_{m-1}, \theta_m))$, but we choose not to[27] and continue with $L(z_m)$. The Euler scheme

$$z_i = z_{i-1} + hg_i(z_{i-1}, \theta_i) \tag{17.54}$$

comes with

$$\lambda_{i-1} = \lambda_i + h(\nabla_{z_{i-1}} g_i(z_{i-1}, \theta_i))\, \lambda_i, \tag{17.55}$$

an explicit Euler backward scheme[28] from $\lambda_m = \nabla_{z_m} L$, supplemented by

$$\nabla_{\theta_i} \tilde{L} = h(\nabla_{\theta_i} g_i(z_{i-1}, \theta_i))\, \lambda_i.$$

[26] Which is really a Fourier series because we restrict to $x \in I$.

[27] This helps to follow Chen et al in their Neural Ordinary Differential Equations paper.

[28] Or $\lambda_{i-1} = h(\nabla_{z_{i-1}} g_i)\, \lambda_i - \nabla_{z_{i-1}} L$, but the last term requires an $h$ to modify (17.58).

By definition[29] this says that

$$\tilde{L}(\theta+\phi,x,y)=\tilde{L}(\theta,x,y)+\underbrace{h\sum_{i=1}^{m}\left(\left(\nabla_{\theta_i}g_i(z_{i-1},\theta_i)\right)\lambda_i\right)\cdot\phi_i}_{\nabla_\theta\tilde{L}(\theta,x,y)\cdot\phi}+o(|\phi|) \tag{17.56}$$

for

$$|\phi|^2=\sum_{i=1}^{m}|\phi_i|^2\to 0.$$

### 17.4.9 The continuous limit: ODE's

The $z$-scheme (17.54) is a numerical solver for the ODE[30]

$$\dot{z}(t)=g(t,z(t),\theta(t)) \tag{17.57}$$

on the interval $[0,1]$, given initial data $z(0)=x$ and a varying parameter $\theta(t)$. This ODE defines a solution $z(t)$ that depends on $x$ and the *function* $\theta$ in (17.57), and $L(z(1),y)$ is thereby a nonlinear functional $\tilde{L}$ of $\theta$, with parameters $x$ and $y$. We denote it[31] again[32] by $\tilde{L}(\theta,x,y)$.

The $\lambda$-scheme (17.55) is a numerical solver for[33]

$$\dot{\lambda}(t)+\left(\nabla_z g(t,z(t),\theta(t))\right)\lambda(t)=0, \tag{17.58}$$

with final data[34]

$$\lambda(1)=\nabla_z L(z(1),y),$$

and varying parameters $z(t)$ and $\theta(t)$. Finally the underbraced term in (17.56) is a Riemann sum for[35]

$$\int_0^1 \underbrace{\nabla_\theta g(t,z(t),\theta(t))\,\lambda(t)}\cdot\phi(t)\,dt, \tag{17.59}$$

in which we recognise the underbraced term as[36] the variational[37] $\theta$-derivative[38] of

$$\int_0^1 \lambda(t)\cdot\left(g(z(t),\theta(t))-\dot{z}(t)\right)dt \tag{17.60}$$

[29] See Exercise 16.6.
[30] Lotte's equation in (4.15), $f=g$, no $t$ in $\theta$, as in Chen's (3), but I take $\theta(t)$ here.
[31] This $\tilde{L}$ is the $L$ in (4.7) of Lotte's Lemma 4.1.
[32] Add $\int_0^1 \hat{L}(t,z(t))\,dt$ to $\tilde{L}$ to get $\lambda_{i-1}=\lambda_i+h(\nabla_{z_{i-1}}g_i(z_{i-1},\theta_i)\,\lambda_i-\nabla_{z_{i-1}}\hat{L}_i(z_{i-1},\theta_i))$.
[33] With a minus, Lotte's (4.19), Chen's (35) adjoint state $a$ is Lagrange multiplier $\lambda$.
[34] And likewise for $\tilde{L}(\theta,x,T)=L(z(T))$, compare to (17.35), and to Chen's (34).
[35] Chen's (5) without the minus.
[36] *Acting on constant functions only*, this correspond to Lotte's (4.17), up to a sign.
[37] Variational derivatives, unlike Lotte's (4.17) and Chen's (5).
[38] See (11.21) and the first footnote on that page, and also Section 14.4.

acting on the function $\phi = d\theta$. This integral is the continuous limit of the sum in (17.53). Of course (17.59) may be derived directly from (17.57) via (17.58), see https://youtu.be/fcv25Fwi7Pc and further. There I take the ODE to be defined for $z(s)$, $s$ running from $\tau$ to $t$, $z(\tau) = x$, and compute the derivatives of $L(z(t,\tau,x,\theta)$, two ordinary derivatives for $t$ and $\tau$, one gradient for $x$, and one variational derivative for $\theta$. My $s$ corresponds to $t$ in their (34), but note that their $L(z(t)$ also depends on the final time. It seems to me that Chen's method for the derivative with respect to $T$ and a constant parameter $\theta$ generalises to general time-dependent $\theta$. I think there's a factor $-f$ missing in their second formula in (52).

The derivative of $L(z(1), y)$ with respect to $\theta$ is computed via $\lambda(1) = \nabla_z L(z(1), y)$, the initial condition for the backwards equation. As in (17.39) we may write $u(x,\theta) = z(1; x, \theta)$ and find that the variational derivative of the $k^{th}$ component of $u(x,\theta)$ is given by

$$\langle u_\theta^k, d\theta \rangle = \int_0^1 \nabla_\theta g(t, z(t), \theta(t))\, \lambda^k(t) \cdot d\theta(t)\, dt, \tag{17.61}$$

$\lambda^k$ being a solution of (17.58) with $\lambda^k(1) = \mathbf{e}^k$. We then arrive at (17.43) with

$$G^{ij} = \int_0^1 \lambda_p^i \lambda_q^j \, \nabla_\theta g^p \cdot \nabla_\theta g^q \tag{17.62}$$

along the gradient flow

$$\frac{d\theta}{ds} = -\nabla_\theta(\lambda(t) \cdot g(t, z(t), \theta(t))) \tag{17.63}$$

if I'm not mistaken.

For a more general cost function

$$L(z(1), y) + \int_0^1 \hat{L}(t, z(t))\, dt$$

the equation for $\lambda(t)$ comes out as

$$\dot{\lambda}(t) + (\nabla_z g(t, z(t), \theta(t)))\, \lambda(t) = \nabla_z \hat{L}(t, z(t)).$$

See[39] Fleming&Rishel1975 and Banks&Kunisch1989. In the continuous setting the starting point is usually the sum of this cost function and (17.60), namely[40]

$$\mathcal{L} = L(z(1), y) + \int_0^1 \hat{L}(t, z(t))\, dt + \int_0^1 \lambda(t) \cdot (g(z(t), \theta(t)) - \dot{z}(t))\, dt,$$

[39] Thanks for the references to Joris Bierkens and Jan van Schuppen.

[40] Put $\hat{L} = 0$ and $t_0 = 0$, $t_1 = 1$ to get Lotte's (4.16).

## 17.5 Applications in optimal transport

This is part of what Finn's doing in his master project and me understanding what's going on from scratch. It concerns the Kantorovich version of the discrete Monge problem. I'm slightly changing his notation and replace $f$ by $\lambda$, $g$ by $\mu$, $C$ by $c$.

For $a_i \geq 0$, $b_j \geq 0$, $c_{ij} \geq 0$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, minimise[41]

$$\langle c, p \rangle = c_{ij} p_{ij} \tag{17.64}$$

over the set $U(a, b)$ of $m$ times $n$ matrices

$$p_{ij} \geq 0 \quad \text{with} \quad p_{ij} \mathit{1}_j = a_i \quad \text{and} \quad \mathit{1}_i p_{ij} = b_j, \tag{17.65}$$

$mn$ inequalities and $m + n$ equations, to obtain $M(a, b, c)$. So the column sums of $p$ are $a_1, \ldots, a_m$ and the row sums are $b_1, \ldots, b_n$. This minimum of $\langle c, p \rangle$ exists because $U(a, b)$ is compact.

Without loss of generality we may assume that

$$\mathit{1}_i a_i = 1 = \mathit{1}_j b_j, \quad \text{whence} \quad \mathit{1}_{ij} p_{ij} = 1, \tag{17.66}$$

and think of $a_i, b_j, p_{ij}$ as probabilities. We write $a \in \Sigma_m$, $b \in \Sigma_n$, $p \in \Sigma_{mn}$. The matrix $c_{ij}$ is called the cost matrix. Note that adding $\varepsilon$ to every $c_{ij}$ only means adding $\varepsilon$ to $\langle c, p \rangle$ and $M(a, b, c)$. Thus we may just as well consider arbitrary matrices $c_{ij}$, and then it is no restriction to assume that

$$c_{ij} \mathit{1}_i \mathit{1}_j$$

is zero from the beginning. But below we stick to nonnegative cost matrices first, as the average value $\bar{c}$ of the cost matrix values will play a role in the calculations. So if we like we can assume that in addition to $c_{ij} \geq 0$ for all $i, j$ also $\mathit{1}_{ij} c_{ij} = 1$.

The story below should perhaps start after a discussion of the function $L$ defined by

$$L(p) = \begin{cases} \langle c, p \rangle & \text{if } p \geq 0 \\ \infty & \text{else} \end{cases}$$

as a (not strictly) convex function $L : \mathbb{R}^{m \times n} \to \mathbb{R} \cup \infty$. If $p^* \in \mathbb{R}^{m \times n}$ is considered as being in the dual space for $p$ then

$$L^*(p^*) = \sup_{p \in \mathbb{R}^{m \times n}} \langle p^*, p \rangle - L(p)$$

[41]Summation convention repeated indices, $\mathit{1}_i = \mathit{1}_j = 1$ for all $i, j$, Frobenius notation.

defines the Legendre transform of $L$, and $L^{**} = L$, by a duality theorem that I first saw in the Masson edition of the Brezis book. In what follows below and in Section 17.5.1 we recognise the use of $p^* = \lambda \oplus \mu$ as

$$L^*(\lambda \oplus \mu) = \sup_{p \in \mathbb{R}^{m \times n}} (\lambda_i + \mu_j)p_{ij} - L(p) = \sup_{\forall i,j\, p_{ij} \geq 0} (\lambda_i + \mu_j - c_{ij})p_{ij}$$

$$= - \inf_{\forall i,j\, p_{ij} \geq 0} (c_{ij} - \lambda_i - \mu_j)p_{ij}.$$

Note that

$$L_{ab}(p) = \begin{cases} \langle c, p \rangle & \text{if } p \in U(a,b) \\ \infty & \text{if } p \notin U(a,b) \end{cases}$$

also defines a (not strictly) convex function, and

$$L^*(p^*) = \underbrace{\sup_{p \geq 0} \langle p^* - c, p \rangle}_{<\infty \iff p^* \leq c} \geq \underbrace{\max_{p \in U(a,b)} \langle p^* - c, p \rangle}_{-M(a,b,c-p^*)} = L^*_{ab}(p^*).$$

Introducing

$$\mathcal{L} = \mathcal{L}(p, \lambda, \mu, a, b, c) = c_{ij}p_{ij} + \lambda_i(a_i - p_{ij}1_j) + (b_j - 1_i p_{ij})\mu_j \tag{17.67}$$

$$= (c_{ij} - \lambda_i - \mu_j)p_{ij} + a_i\lambda_i + b_j\mu_j,$$

we have that

$$M(a,b,c) = \min_{p \geq 0} \sup_{\lambda,\mu} \mathcal{L} = \min_{\substack{p \geq 0 \\ p1=a,\, 1p=b}} c_{ij}p_{ij} \tag{17.68}$$

because the supremum over $\lambda, \mu$ of (17.67) is either $\langle c, p \rangle$, realised with $\lambda = 0, \mu = 0$, or $+\infty$, depending on whether the $m+n$ equalities in (17.65) are satisfied.

We note that the matrix

$$R = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \tag{17.69}$$

can be added to the matrix $p$ at the entries with indices $ij$, $il$, $kj$, $kl$ without changing the column and row sums. Now assume that for all such $i \neq k$ and $j \neq l$ it holds that

$$c_{ij} + c_{kl} \neq c_{il} + c_{kj}.$$

Then we can restrict to $ij$ and $kl$ with

$$c_{ij} + c_{kl} > c_{il} + c_{kj} \tag{17.70}$$

and conclude that the maximizing matrix $p$ has $p_{ij}p_{kl} = 0$.

### 17.5.1 Dual tricks

Let's follow[42][43] a certain Tom Waits fan in French politics and swap infimum and supremum. Clearly

$$\sup_{\lambda,\mu} \inf_{p\geq 0} \mathcal{L} \leq \inf_{p\geq 0} \sup_{\lambda,\mu} \mathcal{L} = \min_{\substack{p\geq 0 \\ p1\geq a,\, 1p\geq b}} c_{ij}p_{ij} \tag{17.71}$$

since

$$\inf_{p\geq 0} \mathcal{L} \leq \sup_{\lambda,\mu} \mathcal{L},$$

and the infimum *evaluates* as

$$\begin{aligned}
\inf_{p\geq 0} \mathcal{L} &= \inf_{p\geq 0} (\lambda_i a_i + \mu_j b_j + c_{ij}p_{ij} - \lambda_i p_{ij} 1_j - \mu_j 1_i p_{ij}) \\
&= \lambda_i a_i + \mu_j b_j + \inf_{p\geq 0} (c_{ij} - \lambda_i 1_j - 1_i \mu_j) p_{ij} \\
&= \underbrace{\langle a, \lambda\rangle + \langle b, \mu\rangle}_{\substack{\text{dual functional} \\ \langle a\oplus b, \lambda\oplus\mu\rangle}} + \underbrace{\inf_{p\geq 0} \langle c - \lambda\otimes 1 - 1\otimes\mu, p\rangle}_{-L^*(\lambda\oplus\mu)},
\end{aligned}$$

in which we introduced the dual functional

$$\langle \lambda, a\rangle + \langle \mu, b\rangle = a_i\lambda_i + b_j\mu_j = a_i b_j(\lambda_i + \mu_j) = \langle a\otimes b, \lambda\oplus\mu\rangle.$$

We have indicated how this infimum relates to the Legendre transform of $L$, but note it only involves $L^*$ acting on $p^*$ of the form $\lambda\oplus\mu$, so taking the supremum over all $\lambda,\mu$ we do not define $(L^*)^*(a\otimes b)$: the inequality in

$$\sup_{\lambda,\mu} \underbrace{(\langle \lambda\oplus\mu, a\otimes b\rangle - L^*(\lambda\oplus\mu))}_{\inf_{p\geq 0} \mathcal{L}}$$

$$\leq \sup_{p^*} (\langle p^*, a\otimes b\rangle - L^*(p^*)) = (L^*)^*(a\otimes b) = L(a\otimes b) = a_i c_{ij} b_j$$

is in general strict in view of the sharper estimate in (17.71).

The infimum is $-\infty$ unless

$$\forall i,j:\ \lambda_i + \mu_j \leq c_{ij}, \quad \text{i.e.} \quad \lambda\oplus\mu = \lambda\otimes 1 + 1\otimes\mu \leq c.$$

Thus the left hand side of (17.71) *evaluates* as

$$\sup_{\lambda,\mu} \inf_{p\geq 0} \mathcal{L}(p,\lambda,\mu,c,a,b) = \sup_{\lambda\oplus\mu\leq c} (\langle \lambda, a\rangle + \langle \mu, b\rangle),$$

[42] Justified with equality by Section 37 in Rockafellars Convex Analysis book, to check
[43] And also in Bertsimas and Tsitsiklis, Theorem 4.4.

whence

$$\sup_{\forall i,j:\,\lambda_i+\mu_j\le c_{ij}} (\lambda_i a_i + b_j\mu_j) \le \min_{\substack{\forall i,j:\,p_{ij}\ge 0\\ p_{ij}1_j=a_i,\,1_ip_{ij}=b_j}} c_{ij}p_{ij}. \tag{17.72}$$

Is this latter inequality an equality? It seems that the positive answer is given by Section 37 in Rockafellar. To show so by direct methods solve

$$\begin{aligned} \forall i,j:\ \lambda_i+\mu_j \le c_{ij} \quad &\text{and} \quad p_{ij}\ge 0;\\ \lambda_i a_i + b_j\mu_j = c_{ij}p_{ij}; \quad p_{ij}1_j &= a_i, \quad 1_ip_{ij}=b_j, \end{aligned} \tag{17.73}$$

$2mn$ inequalities and $1+m+n$ equations for $m+n+mn$ unknowns $\lambda_i,\mu_j,p_{ij}$. We will do this for an example in Section 17.5.5.

The existence of one such solution implies that $p_{ij}$ must be a minimiser for (17.68), and that (17.72) holds with equality and sup = max. Note that if $c_{ij}$ happens to be of the form $c_{ij}=\lambda_i+\mu_j$ then $p_{ij}=a_ib_j$ does the job, but in general $c_{ij}$ is not of this form.

Setting $p_{ij}=a_ib_j-q_{ij}$ we can rewrite (17.72) as

$$\sup_{\forall i,j:\,\lambda_i+\mu_j\le c_{ij}} (\lambda_i a_i + b_j\mu_j) + \max_{\substack{\forall i:\,q_{ij}1_j=0,\,\forall j:\,1_iq_{ij}=0\\ \forall i,j:\,q_{ij}\le a_ib_j}} c_{ij}q_{ij} \le a_ic_{ij}b_j, \tag{17.74}$$

whence we consider

$$a_i\lambda_i + b_j\mu_j + c_{ij}q_{ij} \tag{17.75}$$

subject to the constraints[44]

$$\begin{aligned} \forall i,j:\ \lambda_i+\mu_j \le c_{ij} \quad &\text{and} \quad q_{ij}\le a_ib_j;\\ q_{ij}1_j = 0 = 1_iq_{ij} \quad &\Longleftrightarrow \quad q\in T, \end{aligned} \tag{17.76}$$

and see if we can make (17.75) equal to its upper bound $a_ic_{ij}b_j$ in (17.74), which is equivalent to solving (17.73). The lower bound $a_ib_j-1\le q_{ij}$ is automatic[45] from $q_{ij}1_j=0=1_iq_{ij}$ and corresponds to $p_{ij}\le 1$.

Adding $\varepsilon$ to all $\lambda_i$ and substracting $\varepsilon$ from all $\mu_j$ does not change the dual functional $\lambda_ia_i+b_j\mu_j$ in (17.75), because of (17.66), it does not change the constraints in the supremum, and it does not change

$$\bar\lambda+\bar\mu = \frac{1}{m}1_i\lambda_i + \frac{1}{n}1_j\lambda_j,$$

the sum of the separate averages of the $\lambda_i$ and the $\mu_j$. Moreover, a maximising sequence $\lambda^k,\mu^k$ for the supremum in (17.74) can be chosen with all $\lambda_i^k\ge 0$, forcing $\mu^k$ to be bounded, and since $a_i\lambda_i+b_j\mu_j\le\lambda_i+\mu_j\le c_{ij}$, a convergent

[44] It's convenient to introduce a set $T$ for $q$ here, and we will project $c$ on $T$ shortly.

[45] In fact $p_{ij}\le\min(a_i,b_j)$ gives $a_ib_j-\min(a_i,b_j)\le q_{ij}$.

subsequence argument establishes that the supremum in (17.74) is in fact a maximum. We conclude that in

$$\max_{\forall i,j\colon \lambda_i+\mu_j\le c_{ij}} (\lambda_i a_i + b_j\mu_j) + \max_{\substack{\forall i\colon q_{ij}1_j=0,\,\forall j\colon 1_i q_{ij}=0\\ \forall i,j\colon q_{ij}\le a_i b_j}} c_{ij}q_{ij} \le a_i c_{ij} b_j \tag{17.77}$$

both maxima exist.

### 17.5.2 Reduction to zero column and row sums

The equalities in (17.76) can be written as the $m+n$ equations

$$\langle M^i, q\rangle = 0 = \langle N^j, q\rangle, \tag{17.78}$$

in which $M^i$ is a matrix with ones on the $i^{th}$ row and zeros elsewhere, and $N^j$ is a matrix with ones on the $j^{th}$ column and zeros elsewhere. Between the $m+n$ equations in (17.78) there is one linear dependence given by

$$1_i M^i = 1_j N^j,$$

and the space $T$ of admissible $q$ defined by (17.78) has dimension

$$\dim_T = mn - m - n + 1 = (m-1)(n-1).$$

It is spanned by the $(m-1)(n-1)$ matrices obtained by putting a $2\times 2$ matrix $R$ as in (17.69) in a zero $m\times n$ matrix.

The matrices normal to $T$ in the Frobenius sense are given by

$$\lambda_i M^i + \mu_j N^j \quad \text{with} \quad \bar\lambda = \bar\mu,$$

in which the second relation makes the representation unique. It follows that the matrix $c$ decomposes as

$$c = \lambda_k M^k + \mu_l N^l + \gamma, \quad \gamma \in T.$$

To find $\lambda_i$ and $\mu_j$ we introduce and set

$$\begin{aligned} x_i = n\xi_i = \langle c, M^i\rangle = \langle \gamma, M^i\rangle + \lambda_k\langle M^k, M^i\rangle + \mu_l\langle N^l, M^i\rangle = n\lambda_i + 1_l\mu_l;\\ y_j = m\eta_j = \langle c, N^j\rangle = \langle \gamma, N^j\rangle + \lambda_k\langle M^k, N^j\rangle + \mu_l\langle N^l, N^j\rangle = 1_k\lambda_k + m\mu_j, \end{aligned}$$

from which we infer

$$\xi_i = \lambda_i + \bar\mu; \quad \eta_j = \mu_j + \bar\lambda, \tag{17.79}$$

whence

$$\bar\xi = \bar\lambda + \bar\mu = \bar\eta$$

must hold for the averages. Note that

$$\bar{\xi} = \bar{\eta} = \bar{c}$$

follows from

$$1_i x_i = 1_{ij} c_{ij} = 1_j y_j.$$

Thus

$$c^{\perp} = \lambda_k M^k + \mu_l N^l = (\xi_k - \bar{\mu}) M^k + (\eta_l - \bar{\lambda}) N^l,$$

whence

$$c_{ij}^{\perp} = \lambda_k M_{ij}^k + \mu_l N_{ij}^l = (\xi_k - \bar{\mu}) M_{ij}^k + (\eta_l - \bar{\lambda}) N_{ij}^l = \xi_i + \eta_j - \bar{c}.$$

Summing up, we found that if we define

$$\xi_i = \frac{1}{n} \langle c, M^i \rangle = \frac{c_{ij} 1_j}{n}, \quad \eta_j = \frac{1}{m} \langle c, N^j \rangle = \frac{1_i c_{ij}}{m}, \quad \bar{c} = \frac{1}{mn} \langle 1, c \rangle$$

and

$$c^{\perp} = (\xi_i - \bar{c}) M^i + (\eta_j - \bar{c}) N^j, \quad \gamma = c - c^{\perp}$$

then

$$c = \gamma + c^{\perp}, \quad \gamma \in T, \quad c^{\perp} \in T^{\perp}.$$

We have

$$c_{kl}^{\perp} = \xi_k + \eta_l - \bar{c}, \qquad \gamma_{kl} = c_{kl} - \xi_k - \eta_l + \bar{c},$$

in which[46]

$$\xi_k = \frac{c_{kl} 1_l}{n} = \bar{c}_k^{row}, \quad \eta_l = \frac{1_k c_{kl}}{m} = \bar{c}_l^{column}, \quad \bar{c} = \frac{\langle 1, c \rangle}{mn}$$

are the average $c$-values in row $k$, column $l$, and in the whole matrix. Note that for each $i$ fixed $\xi_i M^i$ is the projection of $c$ on the matrix line spanned by $M^i$ and likewise each $\eta_j N^j$ is the projection on the line spanned by $N^j$.

### 17.5.3 Quadratic costs

With

$$c_{ij} = (i - j)^2, \quad i, j = 1, \ldots, m = n = N + 1,$$

and denoting

$$P_k(N) = 1^k + 2^k + \cdots + N^k$$

[46]Maybe write $\bar{c}_k^{row} 1_l = \bar{c}_{kl}^{row}$, $1_k \bar{c}_l^{column} = \bar{c}_{kl}^{column}$, $1_k 1_l \bar{c} = \bar{c}_{kl}$.

we get

$$\langle 1, c\rangle = 1_{ij}c_{ij} = 2\left(N + (N-1)\,2^2 + (N-2)\,3^2 + \cdots + (N-(N-1))\,N^2\right)$$

$$= 2NP_2(N) - 2P_3(N) + 2P_2(N) = 2(N+1)P_2(N) - 2P_3(N)$$

$$= 2(N+1)P_2(N) - 2P_1(N)^2 = \frac{2}{3}(N+1)^2(N+\frac{1}{2})N - \frac{1}{2}N^2(N+1)^2,$$

$$= \frac{1}{6}(N+2)(N+1)^2N, \quad \text{so} \quad \bar{c} = \frac{1}{6}N(N+2). \tag{17.80}$$

Numbering the rows and columns with $l = i-1$ and $l = j-1$ we

$$c_{1j}1_j = 1^2 + 2^2 + \cdots + N^2 = P_2(N) \quad (l=0),$$

$$c_{2j}1_j = 1^2 + 1^2 + 2^2 + \cdots + (N-1)^2 = 1 + P_2(N-1) \quad (l=1),$$

$$c_{3j}1_j = 1^2 + 2^2 + P_2(N-2) \quad (l=2),$$

so

$$c_{l+1\,j}1_j = P_2(l) + P_2(N-l) = \frac{1}{3}N^3 - N^2l + Nl^2 + \frac{1}{2}N^2 - Nl + l^2 + \frac{1}{6}N$$

$$= \frac{1}{3}(N+1)(N+\frac{1}{2})N - (N+1)l(N-l).$$

It follows that the column and row averages are

$$\bar{c}_{l+1\,.} = \bar{c}_{.\,l+1} = \frac{1}{3}N(N+\frac{1}{2}) - l(N-l), \quad l = 0, \ldots, N. \tag{17.81}$$

We thus find that

$$\gamma_{k+1\,l+1} = (k-l)^2 - \frac{1}{3}N(N+\frac{1}{2}) + k(N-k) - \frac{1}{3}N(N+\frac{1}{2}) + l(N-l) + \frac{1}{6}N(N+2)$$

$$= (k+l)N - 2kl - \frac{2}{3}N(N+\frac{1}{2}) + \frac{1}{6}N(N+2)$$

$$= (k+l)N - 2kl - \frac{1}{2}N^2,$$

so

$$\gamma_{ij} = (n+1)(i+j) - 2ij - \frac{1}{2}(n+1)(n-1). \tag{17.82}$$

Only the second term contributes to $\langle \gamma, q\rangle$ if $q \in T$.

### 17.5.4 General cost matrix

Recalling (17.77) we note that in

$$\max_{\forall i,j:\,\lambda_i+\mu_j\le c_{ij}}(\lambda_i a_i + b_j\mu_j) + \max_{\substack{\forall i:\,q_{ij}1_j=0,\,\forall j:\,1_i q_{ij}=0\\ \forall i,j:\,q_{ij}\le a_i b_j}} c_{ij}q_{ij} \le a_i c_{ij} b_j$$

we have

$$c_{ij}q_{ij} = \gamma_{ij}q_{ij}$$

and

$$c_{ij}a_i b_j = \gamma_{ij}a_i b_j + \underbrace{\xi_i 1_j a_i b_j + 1_i\eta_j a_i b_j}_{a_i\xi_i+b_j\eta_j} = \gamma_{ij}a_i b_j + \frac{a_i c_{ij}1_j}{n} + \frac{1_i c_{ij} b_j}{m}.$$

Assuming $\bar{c} = 0$ we see that (17.77) rewrites as

$$\max_{\forall i,j:\,\lambda_i+\mu_j\le\gamma_{ij}}(a_i\lambda_i + b_j\mu_j) + \max_{\substack{\forall i:\,q_{ij}1_j=0,\,\forall j:\,1_i q_{ij}=0\\ \forall i,j:\,q_{ij}\le a_i b_j}} \gamma_{ij}q_{ij} \le \gamma_{ij}a_i b_j.$$

Given $\gamma \in T$ and $0 < a \in \Sigma_m$, $0 < b \in \Sigma_n$, we're now down to asking about equality in

$$\underbrace{a_i\lambda_i + b_j\mu_j}_{\text{maximize}} + \underbrace{\gamma_{ij}q_{ij}}_{\text{maximize}} \le \gamma_{ij}a_i b_j \tag{17.83}$$

subject to $\lambda, \mu$ with

$$\lambda_i + \mu_j \le \gamma_{ij} \quad \text{and} \quad q \in T \quad \text{with} \quad q_{ij} \le a_i b_j. \tag{17.84}$$

If so then $\lambda, \mu$ realise the first maximum and $q$ realises the second maximum.

We can decompose $a \otimes b$ just as we did with $c$, by

$$a \otimes b = \underbrace{\alpha\otimes\beta}_{\in T} + (a\otimes b)^{\perp}, \quad \alpha_i = a_i - \frac{1}{m}, \quad \beta_j = b_j - \frac{1}{n},$$

and then write $q = \alpha \otimes \beta + Q$ to transform (17.83,17.84) in

$$\underbrace{a_i\lambda_i + b_j\mu_j}_{\text{maximize}} + \underbrace{\gamma_{ij}Q_{ij}}_{\text{maximize}} \le 0,$$

subject to

$$\lambda_i + \mu_j \le \gamma_{ij} \quad \text{and} \quad Q \in T \quad \text{with} \quad Q_{ij} \le \frac{\alpha_i}{n} + \frac{\beta_j}{m} + \frac{1}{mn}.$$

Section 37 in Rockafellar should explain again why the red inequalitity can be achieved.

### 17.5.5 The simplest nontrivial example

Can we solve (17.73) with $c$ replaced by $\gamma$? Let's take a matrix $\gamma$ with $\gamma_{11} = \gamma_{22} = -1$, $\gamma_{12} = \gamma_{21} = 1$, and all other entries zero. We first maximise $a_i\lambda_i + b_j\mu_j$ under the constraints on $\lambda, \mu$.

For fixed $i$ the constraints $\lambda_i + \mu_j \leq 0$ are satisfied for all $j \geq 3$ if and only if $\lambda_i + \bar{\mu}_3 \leq 0$, in which $\bar{\mu}_3$ is the maximum of all $\mu_3, \ldots, \mu_n$, and likewise for fixed $j$ and $\bar{\lambda}_3$ the maximum of all $\lambda_3, \ldots, \lambda_m$. The constraints are therefore equivalent to

$$\lambda_1 + \mu_1 \leq -1, \quad \lambda_1 + \mu_2 \leq 1, \quad \lambda_1 + \bar{\mu}_3 \leq 0,$$

$$\lambda_2 + \mu_1 \leq 1, \quad \lambda_2 + \mu_2 \leq -1, \quad \lambda_2 + \bar{\mu}_3 \leq 0,$$

$$\bar{\lambda}_3 + \mu_1 \leq 0, \quad \bar{\lambda}_3 + \mu_2 \leq 0, \quad \bar{\lambda}_3 + \bar{\mu}_3 \leq 0.$$

Dropping the bars each $\mu_j$ can be chosen maximal so as to have an equality in each row[47].

Suppose we have equalities in the inequalities with $-1$, that is

$$\lambda_1 + \mu_1 = \lambda_2 + \mu_2 = -1.$$

Then

$$\begin{aligned} &\mu_1 = -1 - \lambda_1, \quad \lambda_1 - \lambda_2 \leq 2, \quad \lambda_1 + \mu_3 \leq 0, \\ &\lambda_2 - \lambda_1 \leq 2, \quad \mu_2 = -1 - \lambda_2, \quad \lambda_2 + \mu_3 \leq 0, \\ &\lambda_3 - \lambda_1 \leq 1, \quad \lambda_3 - \lambda_2 \leq 1, \quad \lambda_3 + \mu_3 \leq 0, \end{aligned} \tag{17.85}$$

and under the constraints

$$\lambda_3 - 1 \leq \lambda_1, \lambda_2, \lambda_3 \leq -\mu_3, \quad |\lambda_1 - \lambda_2| \leq 2, \tag{17.86}$$

we must maximise

$$a_i\lambda_i + b_j\mu_j = -b_1 - b_2 + (a_1 - b_1)\lambda_1 + (a_2 - b_2)\lambda_2 + a_3\lambda_3 + b_3\mu_3. \tag{17.87}$$

If $a_1 \geq b_1$ and $a_2 \geq b_2$ then $\lambda_1 = \lambda_2 = \lambda_3 = -\mu_3$ maximise

$$a_i\lambda_i + b_j\mu_j = -b_1 - b_2 - (a_1 - b_1)\mu_3 - (a_2 - b_2)\mu_3 - a_3\mu_3 + b_3\mu_3 = -b_1 - b_2,$$

and

$$p_{11} = b_1, \quad p_{12} = 0, \quad p_{12} = a_1 - b_1,$$

$$p_{21} = 0, \quad p_{22} = b_2, \quad p_{23} = a_2 - b_2,$$

47 Then repeat with the columns, in the Sinkhorn spirit?

$$p_{31} = 0, \quad p_{32} = 0, \quad p_{33} = a_3$$

uniquely realise the minimum $M(a, b, \gamma) = -b_1 - b_2$. So the min and the max coincide. This deals with the case that

$$a_1 \geq b_1 \quad \text{and} \quad a_2 \geq b_2.$$

The case $a_1 \leq b_1$ and $a_2 \leq b_2$ follows by interchanging $i$ and $j$.

If $a_1 \geq b_1$ and $a_2 \leq b_2$ then $\lambda_1 = \lambda_3 = -\mu_3$ and $\lambda_2 = \lambda_3 - 1 = -\mu_3 - 1$ are allowed in view of $|\lambda_1 - \lambda_2| = \lambda_1 - \lambda_2 = 1 \leq 2$ to maximise

$$a_i\lambda_i + b_j\mu_j = -b_1 - b_2 - (a_1 - b_1)\mu_3 - (a_2 - b_2)(\mu_3 + 1) - a_3\mu_3 + b_3\mu_3$$

$$= -b_1 - b_2 - (a_2 - b_2) = -b_1 - a_2,$$

but only if $a_2 + a_3 \geq b_2$ we have

$$p_{11} = b_1, \quad p_{12} = 0, \quad p_{12} = a_1 - b_1,$$

$$p_{21} = 0, \quad p_{22} = a_2, \quad p_{23} = 0,$$

$$p_{31} = 0, \quad p_{32} = b_2 - a_2, \quad p_{33} = b_3 - a_1 + b_1 = a_3 - b_2 + a_1 = a_2 + a_3 - b_2$$

uniquely realising the minimum $M(a, b, \gamma) = -b_1 - a_2$. This deals with the case that

$$a_2 + a_3 \geq b_2 \geq a_2 \quad \text{and} \quad a_1 \geq b_1,$$

for which we see the min and the max coincide again.

Finally we consider the case

$$a_1 \geq b_1 \quad \text{and} \quad b_2 > a_2 + a_3$$

when the optimal choice

$$p_{11} = b_1, \quad p_{12} = p_{21} = 0, \quad p_{22} = a_2$$

is not realisable under the constraints becaude $p_{33}$ would be negative. Note that

$$b_2 > a_2 + a_3 \iff a_1 > b_1 + b_3.$$

Let's now first look at the minimisation of $\langle \gamma, p \rangle$ on $U(a, b)$. If we have to take $p_{11}$ maximal to minimise $\langle \gamma, p \rangle$ then

$$p_{11} = b_1, \quad p_{21} = 0, \quad p_{31} = 0,$$

then

$$p_{12} = (1 - s)b_2, \quad p_{22} = ta_2, \quad p_{32} = sb_2 - ta_2$$

gives

$$p_{13} = a_1 - b_1 - (1-s)b_2, \quad p_{23} = (1-t)a_2, \quad p_{33} = a_3 + ta_2 - sb_2.$$

The constraints on $s, t \in [0,1]$ are then

$$ta_2 \le sb_2 \le ta_2 + a_3, \quad a_1 + sb_2 \ge b_1 + b_2.$$

Fixing $t$ we take $s$ maximal by $sb_2 = ta_2 + a_3$ and then see if $t = 1$ maximal is allowed. We find

$$p_{11} = b_1, \quad p_{12} = b_2 - a_2 - a_3, \quad p_{13} = b_3,$$

$$p_{21} = 0, \quad p_{22} = a_2, \quad p_{23} = 0,$$

$$p_{31} = 0, \quad p_{32} = a_3, \quad p_{33} = 0,$$

which makes $\langle \gamma, p \rangle = b_2 - b_1 - 2a_2 - a_3$, but requires $a_2 + a_3 \le b_2$, which by itself implies $a_2 \le b_2$. Note that for $b_2 > a_2 + a_3$ we have $p_{12} > 0$.

If we have to take $p_{22}$ maximal to minimise $\langle \gamma, p \rangle$ then

$$p_{21} = 0, \quad p_{22} = a_2, \quad p_{23} = 0,$$

and

$$p_{11} = tb_1, \quad p_{12} = (1-s)a_1, \quad p_{13} = sa_1 - tb_1,$$

gives

$$p_{31} = (1-t)b_1, \quad p_{32} = b_2 - a_2 - (1-s)a_1, \quad p_{33} = b_3 + tb_1 - sa_1,$$

with constraints

$$tb_1 \le sa_1 \le tb_1 + b_3, \quad b_2 + sa_1 \ge a_1 + a_2.$$

Fixing $t \in [0,1]$ we make $s$ maximal by $sa_1 = tb_1 + b_3$ whence $t = 1$ maximal gives $tb_1 = b_1$ and $sa_1 = b_1 + b_3$, and we get the same matrix for $p$ as with $p_{11}$ maximal, and the same value $\langle \gamma, p \rangle = b_2 - b_1 - 2a_2 - a_3$. To check is that this is the minimum $M(a, b, \gamma)$.

### 17.5.6 Entropy modification

Recall (13.18) and let $\varepsilon > 0$. We replace (17.64) by

$$C_\varepsilon(p) = \langle c, p \rangle + \varepsilon \mathrm{KL}(p || a \otimes b), \tag{17.88}$$

in which[48]

$$\mathrm{KL}(p||a\otimes b) = p_{ij}\ln\frac{p_{ij}}{a_ib_j} + 1_i(a_ib_j - p_{ij})1_j, \tag{17.89}$$

and minimise $C_\varepsilon(p)$ over the set $U(a,b)$ of $m$ times $n$ matrices with

$$p_{ij}\geq 0 \quad \text{with} \quad p_{ij}1_j = a_i \quad \text{and} \quad 1_ip_{ij} = b_j \tag{17.90}$$

as before. Note that the second term in (17.89) vanishes on $U(a,b)$. If $M_\varepsilon(a,b,c)$ denotes the minimum of[49]

$$C_\varepsilon(p) = (c_{ij} + \varepsilon\ln\frac{p_{ij}}{a_ib_j} - \varepsilon)p_{ij} + \varepsilon 1_ia_ib_j1_j \tag{17.91}$$

then the minimiser is unique and has all $p_{ij}>0$. This is because the function $C_\varepsilon$ is *strictly* convex[50] and

$$\frac{\partial C_\varepsilon}{\partial p_{ij}} \to -\infty \quad \text{as} \quad p_{ij}\to 0.$$

As a consequence of (the reasoning that produced) the Lagrange multiplier statement in (17.31), the minimiser must be a solution of the system of $mn+m+n$ equations

$$c_{ij} + \varepsilon\ln\frac{p_{ij}}{a_ib_j} = \lambda_i + \mu_j, \quad p_{il}1_l = a_i, \quad 1_kp_{kj} = b_j \tag{17.92}$$

for $p_{ij}>0$, $\lambda_i,\mu_j$, $i=1,\dots,m$, $j=1,\dots,n$. These say that (17.88) is stationary in $p$ along the $m+n$ linear constraints in (17.90).

Now (17.92) simplifies (17.91) to

$$(\lambda_i+\mu_j-\varepsilon)p_{ij} + \varepsilon 1_ia_ib_j1_j = (\lambda_i+\mu_j)a_ib_j\exp(\frac{\lambda_i+\mu_j-c_{ij}}{\varepsilon})$$

because the first $mn$ equations in (17.92) are uniquely solved by[51]

$$p_{ij} = a_ib_j\exp(\frac{\lambda_i+\mu_j-c_{ij}}{\varepsilon}), \tag{17.93}$$

[48] I first did my calculations without the second term in the KL-definition 17.89.

[49] With $c = c^\perp + \gamma$ as at the end of Section 17.5.2 we have $\langle c^\perp, p\rangle = a_k\bar{c}_k^{row} + b_l\bar{c}_l^{column} - \bar{c}$.

[50] As the sum of all $p_{ij}\ln p_{ij}$ and a linear function of $p$.

[51] Writing $p_{ij} = a_ib_j - q_{ij}$ as before it follows that

$$q_{ij} = a_ib_j\exp(\frac{c_{ij}-\lambda_i-\mu_j}{\varepsilon}).$$

which we then plug into the remaining $m+n$ equations in (17.92) to get

$$a_i b_j \exp(\frac{\lambda_i + \mu_j - c_{ij}}{\varepsilon}) 1_j = a_i, \quad 1_i a_i b_j \exp(\frac{\lambda_i + \mu_j - c_{ij}}{\varepsilon}) = b_j, \tag{17.94}$$

a system of $m+n$ equations for $\lambda_i, \mu_j$, $i = 1, \ldots, m$, $j = 1, \ldots, n$. In the first we sum over $j$, better use $l$ later, in the second over $i$, better use $k$ later. But these we recognise[52] as the formula's for stationarity of the concave function

$$\Phi_\varepsilon(\lambda, \mu) = \varepsilon 1_i a_i b_j 1_j \underbrace{-\varepsilon a_i b_j \exp(\frac{\lambda_i + \mu_j - c_{ij}}{\varepsilon}) + a_i \lambda_i + b_j \mu_j}_{\text{defines } \varepsilon F_\varepsilon(u,v) \text{ setting } \lambda_i = \varepsilon \ln u_i,\, \mu_j = \varepsilon \ln v_j} . \tag{17.95}$$

Via (17.93) the stationary points of (17.95) thus define the points $p$ where (17.88) is stationary along the $m+n$ linear constraints in (17.90). Equivalently, the stationary points of $F_\varepsilon(u, v)$ define $p$ by

$$p_{ij} = a_i b_j u_i v_j \exp(-\frac{c_{ij}}{\varepsilon}).$$

We already know that the minimiser $p$ is unique in view of the strict convexity of $C_\varepsilon$ makes[53] that this $p$ is also the unique stationary point of $C_\varepsilon$ along $U(a, b)$. For the function $F_\varepsilon$ announced in (17.95) defined by

$$F_\varepsilon(u, v) = a_i \ln u_i + b_j \ln v_j - a_i b_j \exp(-\frac{c_{ij}}{\varepsilon}) u_i v_j$$

things are less obvious, but one of its stationary points must produce the maximiser $p$, and is thereby itself the global maximiser for $F_\varepsilon$.

So is what follows[54] still needed? If we modify (17.67) as

$$(c_{ij} + \varepsilon \ln \frac{p_{ij}}{a_i b_j} - \varepsilon) p_{ij} + \varepsilon 1_i a_i b_j 1_j + \lambda_i (a_i - p_{ij} 1_j) + (b_j - 1_i p_{ij}) \mu_j =$$

$$\mathcal{L}_\varepsilon = (c_{ij} - \lambda_i - \mu_j + \varepsilon \ln \frac{p_{ij}}{a_i b_j} - \varepsilon) p_{ij} + a_i \lambda_i + b_j \mu_j + \varepsilon 1_i a_i b_j 1_j, \tag{17.96}$$

then $\mathcal{L}_\varepsilon$ is again a strictly convex function of $p$. Note that in the $p$-dependent part it's only the cost matrix[55] in (17.88) that has been changed, but now we take

$$p_{ij} \geq 0, \quad i = 1, \ldots, m, \quad i = 1, \ldots, n$$

[52] To expand on.

[53] All stationary points being *strict* local minimisers allows for no mountain passes.

[54] The min = inf sup = max min reasoning from the $\varepsilon = 0$ case in Section 17.5.1.

[55] See the comment below (17.73) in this respect.

as the only restrictions when minimising with respect to $p$. Again the minimiser has all $p_{ij} > 0$. These are found by equating the $p_{ij}$-derivatives of $\mathcal{L}_\varepsilon$ to zero, which reproduces the first system of $mn$ equations in (17.92). If we then use

$$M_\varepsilon(a,b,c) = \min_{p\geq 0} \sup_{\lambda,\mu} \mathcal{L}_\varepsilon = \max_{\lambda,\mu} \inf_{p\geq 0} \mathcal{L}_\varepsilon,$$

we see that

$$\inf_{p\geq 0} \mathcal{L}_\varepsilon = \min_{p\geq 0} \mathcal{L}_\varepsilon$$

is realised by (17.93) and equal to

$$-\varepsilon 1_i 1_j p_{ij} + a_i\lambda_i + b_j\mu_j$$

with $p_{ij}$ given by (17.93). Thus

$$\min_{p\geq 0} \mathcal{L}_\varepsilon = \varepsilon a_i b_j (1 - \exp(\frac{\lambda_i + \mu_j - c_{ij}}{\varepsilon})) + a_i\lambda_i + b_j\mu_j = \Phi_\varepsilon(\lambda,\mu),$$

our strictly concave function, of which we determined the maximiser as the unique stationary point via the Lagrange multiplier method and (17.94) above. This maximum coincides with the minimum of (17.91) over $U(a,b)$.

If $a \in \Sigma_m$,$b \in \Sigma_n$ then

$$\Phi_\varepsilon(\lambda,\mu) = \varepsilon - \varepsilon a_i b_j \exp(\frac{\lambda_i + \mu_j - c_{ij}}{\varepsilon}) + a_i\lambda_i + b_j\mu_j.$$

### 17.5.7 Sinkhorn method

This is what Augusto[56] suggested Finn to do in his last chapter. Note that (17.94) rewrites as

$$\begin{aligned} b_j \exp(\tfrac{\lambda_i+\mu_j-c_{ij}}{\varepsilon}) &= 1, \quad i = 1,\ldots,m, \quad \text{sum over} \quad j; \\ a_i \exp(\tfrac{\lambda_i+\mu_j-c_{ij}}{\varepsilon}) &= 1, \quad j = 1,\ldots,n, \quad \text{sum over} \quad i, \end{aligned}$$

and we know a priori that this system must have a solution $\lambda,\mu$ that defines a *unique* minimiser $p \in U(a,b)$ for $M_\varepsilon(a,b,c)$ via

$$p_{ij} = a_i b_j \exp(\frac{\lambda_i + \mu_j - c_{ij}}{\varepsilon}). \tag{17.97}$$

We may write the above system as[57]

$$\begin{aligned} \lambda_i &= -\varepsilon \ln\left(b_j \exp(\tfrac{\mu_j - c_{ij}}{\varepsilon})\right) = \mu_i^{C,\varepsilon}; \\ \mu_j &= -\varepsilon \ln\left(a_i \exp(\tfrac{\lambda_i - c_{ij}}{\varepsilon})\right) = \lambda_j^{C,\varepsilon}, \end{aligned} \tag{17.98}$$

[56] https://arxiv.org/pdf/1911.06850.pdf, see also https://arxiv.org/pdf/2006.06033.pdf

[57] On the right Finn's notation, with $\lambda$ for $f$, $\mu$ for $g$, and $C$-transform terminology.

with summation convention in the logarithms. The right hand sides are called the $(c, \varepsilon)$-transforms of $\mu$ and $\lambda$, and the functions

$$\varepsilon \ln \left( b_j \exp(\frac{\mu_j}{\varepsilon}) \right), \quad \varepsilon \ln \left( a_i \exp(\frac{\lambda_i}{\varepsilon}) \right)$$

are used in estimates.

Since the $\lambda_i, \mu_j$ are only a means to an end we may just as well work with

$$u_i = \exp(\frac{\lambda_i}{\varepsilon}), \quad v_j = \exp(\frac{\mu_j}{\varepsilon})$$

for that matter. Recalling (17.95) we have

$$\Phi_\varepsilon(\lambda, \mu) = \varepsilon \Psi_\varepsilon(u, v),$$

$$\Psi_\varepsilon(u, v) = \mathit{1}_i a_i b_j \mathit{1}_j + \underbrace{a_i \ln u_i + b_j \ln v_j - u_i K^\varepsilon_{ij} v_j}_{F_\varepsilon(u,v)}. \tag{17.99}$$

Rewriting (17.98) as

$$\begin{aligned} b_j \exp(-\tfrac{c_{ij}}{\varepsilon}) v_j &= \tfrac{1}{u_i}, \quad i = 1, \ldots, m; \\ a_i \exp(-\tfrac{c_{ij}}{\varepsilon}) u_i &= \tfrac{1}{v_j}, \quad j = 1, \ldots, n, \end{aligned}$$

the Gibbs kernel

$$K^\varepsilon_{ij} = a_i \exp(-\frac{c_{ij}}{\varepsilon}) b_j$$

makes these equations read[58]

$$u_i = \frac{a_i}{K^\varepsilon_{il} v_l}; \quad v_j = \frac{b_j}{u_k K^\varepsilon_{kj}}; \quad p_{ij} = u_i K^\varepsilon_{ij} v_j = a_i u_i \exp(-\frac{c_{ij}}{\varepsilon}) b_j v_j,$$

with summation over the repeated index in both denominators. The solution $u, v$ may be *nonunique*, but we know $p$ is unique.

Now the Sinkhorn method is solving this system with one of the first two dropped, and then in the next step the other one. This should define a scheme for $(u, v)$ that converges, with $p$ to follow. Note that using the first in the third we have

$$\mathit{1}_i p_{ij} = \frac{a_i}{K^\varepsilon_{ij} v_j} K^\varepsilon_{ij} v_j = a_i,$$

[58] Do the method with $c$ replaced by $\gamma \in T$? Very different Gibbs kernel!

and likewise the second in the third gives

$$p_{ij}1_j = u_i K^\varepsilon_{ij} \frac{b_j}{u_i K^\varepsilon_{ij}} = b_j.$$

So start from

$$v_j = 1_j$$

and then repeat the commands

$$\underbrace{u_i := \frac{a_i}{K^\varepsilon_{il} v_l}}_{\text{new } u} = \underbrace{\frac{1}{\exp(-\frac{c_{il}}{\varepsilon}) b_l v_l}}_{\text{sum over } l}; \quad \tilde{p}_{ij} := \underbrace{u_i K^\varepsilon_{ij} v_j}_{\Longrightarrow \tilde{p}_{ij}1_j = a_i} = \underbrace{\frac{a_i \exp(-\frac{c_{ij}}{\varepsilon}) b_j v_j}{\exp(-\frac{c_{il}}{\varepsilon}) b_l v_l}}_{\text{sum over } l};$$

$$\underbrace{v_j := \frac{b_j}{u_k K^\varepsilon_{kj}}}_{\text{new } v} = \underbrace{\frac{1}{a_k u_k \exp(-\frac{c_{kj}}{\varepsilon})}}_{\text{sum over } k}; \quad \hat{p}_{ij} := \underbrace{u_i K^\varepsilon_{ij} v_j}_{\Longrightarrow 1_i \hat{p}_{ij} = b_j} = \underbrace{\frac{a_i u_i \exp(-\frac{c_{ij}}{\varepsilon}) b_j}{a_k u_k \exp(-\frac{c_{kj}}{\varepsilon})}}_{\text{sum over } k},$$

as long as some non-stopping condition is satisfied. In every step the new $u$ and $\tilde{p}$ are defined in terms of the old $v$, and the new $v$ and $\hat{p}$ in terms of the new $u$ just computed. The rows of $\tilde{p}$ sum up to the $a_i$ and the colums of $\hat{p}$ sum up to $b_j$. This is because of the structure

$$u_i := \frac{1}{\text{sum over terms indexed by } l}, \quad \tilde{p}_{ij} := a_i \frac{j^{th} \text{ term}}{\text{sum over all terms}}$$

and likewise for $v$, $\hat{p}$ in the loop above.

Convergence of these iterations relies[59] on the transformation

$$(u, v) \to (\tilde{u}, \tilde{v}) \quad \text{defined by} \quad \tilde{u}_i = \frac{a_i}{K^\varepsilon_{ij} v_j}, \quad \tilde{v}_j = \frac{b_j}{\tilde{u}_i K^\varepsilon_{ij}}$$

increasing

$$F_\varepsilon(u, v) = a_i \ln u_i + b_j \ln v_j - \underbrace{a_i b_j \exp(-\frac{c_{ij}}{\varepsilon})}_{K^\varepsilon_{ij}} u_i v_j,$$

see (17.99). Finn formulates and proves this for the scheme derived from (17.98) and $\Phi_\varepsilon(\lambda, \mu)$. His statements correspond to[60]

$$F_\varepsilon(u, v) \le F_\varepsilon(\tilde{u}, v) \le F_\varepsilon(\tilde{u}, \tilde{v}),$$

and imply that along the sequence of Sinkhorn iterates the value of $F_\varepsilon$ is nondecreasing. A convergent subsequence argument should then provide a limit point in which $F_\varepsilon$ is stationary in both $u$ and $v$ and thereby maximal, in view of its properties in its separate $u$ and $v$ variables.

[59] Still have to check this.
[60] This only uses that $f(x) \le f(\tilde{x})$ if $\tilde{x}$ is defined by $f'(\tilde{x}) = 0$ and $f$ is concave.

**Exercise 17.5.** Why is the maximum thus found equal to $M(a, b, c)$?

The compactness needed here is provided in Section 17.5.9, as I realised continuing reading Finn's master thesis.

### 17.5.8 Sinkhorn for the first nontrivial example

Put
$$\delta = \exp(-\frac{1}{\varepsilon}),$$
start with $u_1 = u_2 = u_3 = 1$ and iterate, with
$$v_1 := \frac{1}{a_1u_1/\delta + \delta a_2u_2 + a_3u_3}; \quad v_2 := \frac{1}{\delta a_1u_1 + a_2u_2/\delta + a_3u_3};$$
$$v_3 := \frac{1}{a_1u_1 + a_2u_2 + a_3u_3}$$
as intermediate step, the scheme
$$\frac{1}{u_1} := \frac{b_1/\delta}{a_1u_1/\delta + \delta a_2u_2 + a_3u_3} + \frac{\delta b_2}{\delta a_1u_1 + a_2u_2/\delta + a_3u_3} + \frac{b_3}{a_1u_1 + a_2u_2 + a_3u_3};$$
$$\frac{1}{u_2} := \frac{\delta b_1}{a_1u_1/\delta + \delta a_2u_2 + a_3u_3} + \frac{b_2/\delta}{\delta a_1u_1 + a_2u_2/\delta + a_3u_3} + \frac{b_3}{a_1u_1 + a_2u_2 + a_3u_3};$$
$$\frac{1}{u_3} := \frac{b_1}{a_1u_1/\delta + \delta a_2u_2 + a_3u_3} + \frac{b_2}{\delta a_1u_1 + a_2u_2/\delta + a_3u_3} + \frac{b_3}{a_1u_1 + a_2u_2 + a_3u_3}.$$
The scheme rewrites as
$$\frac{1}{u_1} := \frac{b_1}{a_1u_1 + \delta^2 a_2u_2 + \delta a_3u_3} + \frac{\delta^2 b_2}{\delta^2 a_1u_1 + a_2u_2 + \delta a_3u_3} + \frac{b_3}{a_1u_1 + a_2u_2 + a_3u_3};$$
$$\frac{1}{u_2} := \frac{\delta^2 b_1}{a_1u_1 + \delta^2 a_2u_2 + \delta a_3u_3} + \frac{b_2}{\delta^2 a_1u_1 + a_2u_2 + \delta a_3u_3} + \frac{b_3}{a_1u_1 + a_2u_2 + a_3u_3};$$
$$\frac{1}{u_3} := \frac{\delta b_1}{a_1u_1 + \delta^2 a_2u_2 + \delta a_3u_3} + \frac{\delta b_2}{\delta^2 a_1u_1 + a_2u_2 + \delta a_3u_3} + \frac{b_3}{a_1u_1 + a_2u_2 + a_3u_3},$$
and putting $\delta = 0$ we obtain[61] the scheme
$$\frac{1}{u_1} := \frac{b_1}{a_1u_1} + \frac{b_3}{a_1u_1 + a_2u_2 + a_3u_3}; \quad \frac{1}{u_2} := \frac{b_2}{a_2u_2} + \frac{b_3}{a_1u_1 + a_2u_2 + a_3u_3},$$
$$\frac{1}{u_3} := \frac{b_3}{a_1u_1 + a_2u_2 + a_3u_3},$$
It would seem that the Sinkhorn method works for $\varepsilon = 0$ as well. Via the system above for the example[62], but not only for the example.

[61] Use reciprocals to continue? Or write this projectively? Notation $u_1 : u_2 : u_3$?
[62] Re-examine what we did and compare!

### 17.5.9 The Hilbert metric and Birkhoff tricks

We're continuing with the $u$-scheme, but then in the general case. In view of the scaling we measure the difference between $u$ and another $u$ called $v$, not $v$ as above, by transforming

$$x_i = \ln u_i, \quad y_i = \ln v_i,$$

$$z_i = x_i - y_i, \quad d_{\mathcal{H}}(u,v) = \max_{i,j}(z_i - z_j) = \max_{i,j}|z_i - z_j| = d_\infty(x,y),$$

which is like the maximum norm of $z$ modulo a constant. That is, if we choose $c$ such that $\bar{z}_i = z_i + c$ has its minimal coordinate equal to 0 then $|[z]|_{\max} = |\bar{z}|_{\max}$, in which $[z]$ is the equivalence class of $z$ for the equivalence relation $x \sim y \iff x_i - y_i \equiv c$ for some $c$. In terms of $u$ and $v$ this reads $u \sim v$ if and only $v$ is a scalar multiple of $v$, and with

$$\ln w_i = z_i = x_i - y_i = \ln u_i - \ln v_i = \ln \frac{u_i}{v_i}$$

we have

$$d_{\mathcal{H}}(u,v) = \max_{i,j}(z_i - z_j) = \max_{i,j}(\ln w_i - \ln w_j) = \max_{i,j}(\ln \frac{w_i}{w_j}) = \max_{i,j}(\ln \frac{u_i v_j}{u_j v_i})$$

**Exercise 17.6.** Write $\mathit{1} = (1,\ldots,1)$ for the $n$-vector with only ones in $\mathbb{R}_n$. We can consider $\mathbb{R}^n_{\mathit{1}}$ is $\mathbb{R}_n$ modulo $\mathit{1}$ with the metric $d_\infty$. This is a nice space. Why?

If $K$ is a matrix with positive entries $K_{ij}$ then

$$X_i = \ln(K_{il}\exp(x_l)) = \Phi(x)_i$$

defines a map $\Phi$ that commutes with adding a constant $c$ to all coordinates. We need to estimate $d_\infty(X,Y) = d_\infty(\Phi(x),\Phi(y))$ in terms of $d_\infty(x,y)$. This involves the ln of the maximum of

$$\frac{K_{ik}\exp(x_k)K_{jl}\exp(y_l)}{K_{jk}\exp(x_k)K_{il}\exp(y_l)} = \frac{K_{ik}K_{jl}\exp(x_k+y_l)}{K_{jk}K_{il}\exp(x_k+y_l)},$$

with summation over $k$ and $l$ in numerator and denominator, over $i$ and $j$. This can be estimated in terms of

$$\max_{ijkl}\frac{K_{ik}K_{jl}}{K_{jk}K_{il}} \quad \text{and} \quad \exp(d_\infty(x,y)).$$

We write $v = U$, $x_i = \ln u_i$, $X_j = \ln U_j$, and estimate

$$d_\infty(X, Y) = \max_{ij}(X_i + Y_j - X_j - Y_i) = \ln(\max_{ij} \frac{U_i V_j}{U_j V_i})$$

in terms of

$$d_\infty(x, y) = \max_{ij}(x_i + y_j - x_j - y_i) = \ln(\max_{ij} \frac{u_i v_j}{u_j v_i}).$$

We have for $i \neq j$ fixed that

$$\frac{U_i V_j}{U_j V_i} = \frac{K_{jl} u_l}{K_{ik} u_k} \frac{K_{ik} v_k}{K_{jl} v_l} = \frac{K_{ik} v_k}{K_{ik} u_k} \frac{K_{jl} u_l}{K_{jl} v_l} = \underbrace{\frac{K_{ik} K_{jl} u_l v_k}{K_{ik} K_{jl} u_k v_l}}_{\frac{A^{ij}_{kl} z_{kl}}{A^{ij}_{kl} w_{kl}}} = \underbrace{\frac{K_{jk} K_{il} u_k v_l}{K_{ik} K_{jl} u_k v_l}}_{\frac{A^{ij}_{kl} z_{kl}}{B^{ij}_{kl} z_{kl}}},$$

with summation over $k, l$ in numerators and denominators. We conclude that

$$\frac{U_i V_j}{U_j V_i} \leq \max_{kl} \frac{u_l v_k}{u_k v_l} \quad \text{and} \quad \frac{U_i V_j}{U_j V_i} \leq \max_{kl} \frac{K_{ik} K_{jl}}{K_{il} K_{jk}} = \eta_{ij}(K),$$

and that $K$ is an isometry if $K$ is a diagonal matrix. In particular the diameter of the range of $\Phi$ is bounded by

$$\eta(K) = \max_{ij} \eta_{ij}(K) \geq 1.$$

Garrett Birkhoff[63] showed that in fact it holds that $\Phi$ is contractive with contraction factor

$$\lambda(K) = \frac{\sqrt{\eta(K)} - 1}{\sqrt{\eta(K)} + 1} < 1.$$

He first rewrites

$$u = u_1 : u_2 \xrightarrow{K} U = U_1 : U_2 = (K_{11} u_1 + K_{12} u_2) : (K_{21} u_1 + K_{22} u_2)$$

for

$$e^\xi = x = \frac{u_2}{u_1} \quad \text{and} \quad e^\eta = y = \frac{U_2}{U_1}$$

as

$$x \xrightarrow{K} y = \frac{K_{21} u_1 + K_{22} u_2}{K_{11} u_1 + K_{12} u_2} = \frac{K_{21} + K_{22}\, x}{K_{11} + K_{12}\, x} = \frac{ax + b}{cx + d}$$

[63] Google Extensions of Jentzsch's Theorem.

to effectively compute

$$\frac{\partial\eta}{\partial\xi}=\frac{\partial\eta}{\partial y}\frac{\partial y}{\partial x}\frac{\partial x}{\partial\xi}=\frac{1}{y}\frac{ad-bc}{(cx+d)^2}x=\frac{(ad-bc)x}{(ax+b)(cx+d)},$$

which is maximised in absolute value by

$$x=\sqrt{\frac{bd}{ac}}$$

to give, with

$$\mu^2=\frac{ad}{bc},$$

$$\frac{\partial\eta}{\partial\xi}=\frac{ad-bc}{ad+bc+2\sqrt{abcd}}=\frac{\mu^2-1}{1+2\mu+\mu^2}=\frac{\sqrt{\frac{ad}{bc}}-1}{\sqrt{\frac{ad}{bc}}+1}=\tanh\frac{1}{4}\ln\frac{ad}{bc}.$$

Let's see how this should generalise. For a positive matrix $K$ indexed with $0,1,2$ we consider

$$u_0:u_1:u_2\xrightarrow{K}U_0:U_1:U_2=$$

$$(K_{00}u_0+K_{01}u_1+K_{02}u_2):(K_{10}u_0+K_{11}u_1+K_{12}u_2):(K_{20}u_0+K_{21}u_1+K_{22}u_2).$$

Then

$$e^{\xi_i}=x_i=\frac{u_i}{u_0},\quad e^{\eta_i}=y_i=\frac{U_i}{U_0},\quad i=1,2$$

leads to

$$y_1=\frac{K_{10}+K_{11}x_1+K_{12}x_2}{K_{00}+K_{01}x_1+K_{02}x_2},\quad y_2=\frac{K_{20}+K_{21}x_1+K_{22}x_2}{K_{00}+K_{01}x_1+K_{02}x_2}.$$

More generally we can write

$$u_0:u_j\xrightarrow{K}U_0:U_j$$

for

$$u_0:u_1:\cdots:u_n\xrightarrow{K}U_0:U_1:\cdots:U_n$$

with

$$U_0=K_{00}u_0+K_{0i}u_i,\quad U_j=K_{j0}u_0+K_{ji}u_i,$$

summation over $i=1,\ldots,n$, so

$$y_i=\frac{K_{i0}+K_{ik}x_k}{K_{00}+K_{0l}x_l},$$

summation over $k = 1, \dots, n$, $l = 1, \dots, n$, whence

$$\frac{\partial y_i}{\partial x_j} = \frac{K_{ij}K_{00} - K_{0j}K_{i0} + (K_{ij}K_{0k} - K_{ik}K_{0j})x_k}{(K_{00} + K_{0l}x_l)^2},$$

summation over $k \neq 0, j$ in the numerator, over $l \neq 0$ in the denominator. Thus

$$\frac{\partial \eta_i}{\partial \xi_j} = \frac{x_j}{y_i}\frac{\partial y_i}{\partial x_j} = x_j \frac{K_{ij}K_{00} - K_{0j}K_{i0} + (K_{ij}K_{0k} - K_{ik}K_{0j})x_k}{(K_{i0} + K_{ik}x_k)(K_{00} + K_{0l}x_l)},$$

But this is not what GB did. For $u_0 : u_1 : \cdots : u_n$ and $v_0 : v_1 : \cdots : v_n$ he would choose representations $u = (u_0, u_1, \dots, u_n)$ and $v = (v_0, v_1, \dots, v_n)$ with $u_0 + u_1 + \cdots + u_n = v_0 + v_1 + \cdots + v_n = 1$, and intersect the line $l$ through $u$ and $v$ with the closed cone of all points with nonnegative coordinates, thus obtaining a closed line segment $[a, b]$, $a$ and $b$ chosen such that $u \in [a, v]$, $v \in [u, b]$, and map the triangle $Oab$ to $\mathbb{R}^2$, $O \in \mathbb{R}^{n+1}$ to $O \in \mathbb{R}^2$, $a$ to $(1, 0)$, $b$ to $(0, 1)$, denote the images of $u$ and $v$ by $f$ and $g$, and observe[64] that

$$d_{\mathcal{H}}(u, v) = d_{\mathcal{H}}(f, g).$$

Since diagonal matrices are isometries, $a$ can in fact be mapped to any point on the first positive axis, and $b$ to any point on the second positive axis.

We should also look at the Jacobian matrix

$$\frac{\partial X_i}{\partial x_k},$$

for

$$\ln(U_i) = \underbrace{X_i = \ln(\exp(\Gamma_{ij})e^{x_j})} = \ln(\exp(\Gamma_{ij})u_j),$$

and observe that

$$\frac{\partial X_i}{\partial x_k} = \exp(\Gamma_{ik} + x_k - X_i) = \kappa_{ik} = \frac{\exp(\Gamma_{ik})u_k}{U_i}$$

acts in the tangent space and should have a norm there which derives from $d_\infty$.

[64]NB. In fact he uses this as a definition, why does it reproduce what we have?

# 18 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 $\mathbf{a}$ and $\mathbf{b}$ 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 $\mathbf{a}$ and $\mathbf{b}$ 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 $\mathbf{a}$ and $\mathbf{b}$? The answer involves the matrix product

$$S = A^T A, \tag{18.1}$$

a symmetric matrix to which Section 18.4 applies.

[1]Essentially Theorem 18.6.
[2]We momentarily surrender to the boldface vector notation in physics......

## 18.1 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

$$\sum_{j=1}^{n} a_{ij} b_{jk}, \tag{18.2}$$

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 (18.2) 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

$$\sum_{k=1}^{p}(\sum_{j=1}^{n} a_{ij}b_{jk})c_{kl} = \sum_{k=1}^{p}\sum_{j=1}^{n} a_{ij}b_{jk}c_{kl}, \tag{18.3}$$

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 17.2 for (17.25).

## 18.2 Matrix norms

The series

$$I + A + A^2 + A^3 + \cdots , \tag{18.4}$$

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 23.1. You should also compare (18.4) to (15.29), 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 $2 \times 2$ matrix as in (16.9) and estimates for $A : \mathbb{R}^2 \to \mathbb{R}^2$ of the form (16.10).

Indeed you easily check[8] that for every $n \times n$ matrix and every real $n$-vector $h$ it is true that

$$|Ah|_2 \leq M|h|_2, \tag{18.5}$$

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

$$|A + B|_2 \leq |A|_2 + |B|_2 \quad \text{and} \quad |AB|_2 \leq |A|_2 \, |B|_2. \tag{18.6}$$

[7]A $2 \times 2$ matrix as in (16.9) for instance.

[8]Using proof by induction if you like.

holds[9].

If you like all of the above is just algebra with matrices. The smallest $M$ for which (18.5) 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 (18.6). 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 18.1.** 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 1.15.

[13] Hardy would dislike this proof and prefer an explicit construction of the $\mu_n$.

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 18.2.** If $A \in L(\mathbb{R}^n)$ has $|A|_{op} < 1$ then it holds for the series in (18.4) 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 18.3.** *It should by now be clear that the whole machinery of power series carries over to Banach algebra's.*

## 18.3 Quadratic forms and operator norms

In (18.2) 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]

$$S = AA^T \in L(\mathbb{R}^m, \mathbb{R}^m) \quad \text{with entries} \quad s_{ik} = \sum_{j=1}^{n} a_{ij} a_{kj} = s_{ki}. \tag{18.7}$$

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

$$|A^T|_{op}^2 = \max_{|z|_2 = 1} \underbrace{|A^T z|_2^2}_{A^T z \cdot A^T z} = \max_{|z|_2 = 1} AA^T z \cdot z = \max_{|z|_2 = 1} Sz \cdot z = \max_{0 \neq z \in \mathbb{R}^m} \frac{Sz \cdot z}{z \cdot z}, \tag{18.8}$$

and we note that the bilinear mapping

$$(z, w) \to Sz \cdot w$$

[14]Don't let (18.1) 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 18.4.** 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

$$S(z + w) \cdot (z + w) + S(z - w) \cdot (z - w) = 2\,Sz \cdot z + 2\,Sw \cdot w, \tag{18.9}$$

an identity to play with, with $S = AA^T$as above, but also with $S = I$ the identity:

**Exercise 18.5.** 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 $M$, $z$ and $w$ 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 18.5 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.

## 18.4 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

$$|S|_{op} = \sup_{0 \neq z \in H} \frac{|Sz|}{|z|} = \sup_{0 \neq z \in H} \sqrt{\frac{Sz \cdot Sz}{z \cdot z}} = \sup_{z \cdot z = 1} \sqrt{Sz \cdot Sz}, \tag{18.10}$$

and

$$M = \sup_{z \cdot z = 1} Sz \cdot z, \tag{18.11}$$

it suffices to have that $S$ is bounded on the unit ball in $H$ to have

$$M = |S|_{op} < \infty. \tag{18.12}$$

Ignoring the trivial case that $M = 0$ we now observe that the Cauchy-Schwarz inequality in Exercise 18.4 also holds with $S$ replaced by $M - S = MI - S$, $I$ being the identity map, and it thus holds that

$$|(M - S)z \cdot w| \leq \sqrt{(M - S)z \cdot z}\,\sqrt{(M - S)w \cdot w}, \tag{18.13}$$

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 18.6.** *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 $\lambda_1$ is an eigenvalue of $S$ whose eigenvectors are precisely the maximisers*[15] *of the quotient under consideration.*

**Remark 18.7.** *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 $\lambda_1$ exists, and is an eigenvalue of $S$ whose eigenvectors are the maximisers. 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 (18.13) still holding[18] while the version in Exercise 18.4 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.

## 18.5 Singular values and measures of parallelotopes

In the case that $H = \mathbb{R}^m$ the subsequence argument is not needed as the maximiser $w$ for the maximum in Theorem 18.6 exists in view of the compactness of the unit ball in $\mathbb{R}^m$. Now consider the matrix $A$ defined by

$$A^T = \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} \tag{18.14}$$

and[19]

$$S = A^T A = \begin{pmatrix} a_1^2 + a_2^2 + a_3^2 & a_1 b_1 + a_2 b_2 + a_3 b_3 \\ b_1 a_1 + b_2 a_2 + b_3 a_3 & b_1^2 + b_2^2 + b_3^2 \end{pmatrix} = \begin{pmatrix} a \cdot a & a \cdot b \\ b \cdot a & b \cdot b \end{pmatrix} \tag{18.15}$$

[18] I first saw this Cauchy-Schwarz trick in the appendix of the PDE book of Craig Evans.
[19] Compared to (18.7) we switch from $A$ to $A^T$, back to (18.1) 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

$$|a \times b|^2 = (a_2b_3 - a_3b_2)^2 + (a_3b_1 - a_1b_3)^2 + (a_1b_2 - a_2b_1)^2 = \det(\mathrm{A}^\mathrm{T}\mathrm{A}),$$

as you should verify. That is to say, $\det(\mathrm{A}^\mathrm{T}\mathrm{A})$ is the sum of all the squares of all the $2 \times 2$-determinants of $2 \times 2$ submatrices of $A$. Here we count these $2 \times 2$ submatrices modulo the column permutations in (18.14).

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 $2 \times 2$-determinants in (18.14). 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 $\mathrm{I\!R}^\mathrm{N}$.

**Theorem 18.8.** *Let $1 \leq n \leq N$. Consider the parallelotope $P$ spanned by the vectors $x_1, \ldots, x_n$ in $\mathrm{I\!R}^\mathrm{N}$. 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 (nonnegative) singular values of $A$.*

Let us sketch a proof of this statement, first for (18.14), 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 $t \in \mathrm{I\!R}$ 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 (18.7) 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 $\Lambda$ 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 $B = [B_1, \dots, B_n]$ with $B_1, \dots, B_n$ the column vectors of $B$ and using product notation, that

$$\det(\mathrm{A}^{\mathrm{T}}\mathrm{A}) = \underbrace{\Pi_j \lambda_j^2}_{\lambda_j^1 \cdots \lambda_j^n} + \mathrm{s}_{11} \underbrace{\Pi_{j\neq 1}\lambda_j^2}_{\lambda_j^2 \cdots \lambda_j^n} + \cdots + \mathrm{s_{nn}}\Pi_{\mathrm{j}\neq\mathrm{n}}\lambda_{\mathrm{j}}^2$$

$$+\det\begin{pmatrix} \mathrm{s}_{11} & \mathrm{s}_{12} \\ \mathrm{s}_{12} & \mathrm{s}_{22} \end{pmatrix} \Pi_{\mathrm{j}\neq 1,2}\lambda_{\mathrm{j}}^2 + \cdots + \det \mathrm{S} =$$

$$\Pi_j \lambda_j^2 + (B_1 \cdot B_1)\Pi_{j\neq 1}\lambda_j^2 + \cdots + \det\begin{pmatrix} \mathrm{B}_1 \cdot \mathrm{B}_1 & \mathrm{B}_1 \cdot \mathrm{B}_2 \\ \mathrm{B}_1 \cdot \mathrm{B}_2 & \mathrm{B}_2 \cdot \mathrm{B2} \end{pmatrix} \Pi_{\mathrm{j}\neq 1,2}\lambda_{\mathrm{j}}^2 + \cdots\ ,$$

in which we wrote the term of degree $n$ and only the first terms of degree $2n-2$ and degree $2n-4$ in $\lambda_1, \dots, \lambda_n$. 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 now look at the dimensions to observe that we can systematically reduce the statement we want to prove to lower dimensions of the matrix under consideration, until we reach the easy case that $m = 1$.

# 19 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 (14.2) 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 $\Phi$.

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 16 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

$$x \to \Phi(x, y_j) \quad \text{en} \quad y \to \Phi(x_i, y) \quad (i = 0, \dots, \mathrm{M},\ j = 0 \dots, N),$$

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

$$\frac{|S_{ij}|}{|R_{ij}|} \sim |\mathrm{det} J(x_i, y_i)| \tag{19.1}$$

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

$$\iint_R f \, |\mathrm{det} J| = \iint_A g. \tag{19.2}$$

[2]To do, note that $J$ is constant if $\Phi$ is linear.

[3]See e.g. Section 5 of Chapter III in the Advanced Calculus book of Edwards.

# 20 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.

## 20.1 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

$$\iint_{\mathbb{R}^2} u = \lim_{R\to\infty} \underbrace{\iint_{[-R,R]\times[-R,R]} u(x,y)d(x,y)}_{J(R)} = \lim_{R\to\infty} J(R) \tag{20.1}$$

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

$$\iint_{x^2+y^2\leq R^2} u(x,y)d(x,y) = \int_0^{2\pi}\int_0^R u(r\cos\theta, r\sin\theta)\,rdr\,d\theta$$

$$= \int_0^R\int_0^{2\pi} u(r\cos\theta, r\sin\theta)\,d\theta\,rdr \tag{20.2}$$

gebruiken we

$$0 = r_0 \leq r_1 \leq \cdots \leq r_M = R \quad \text{met} \quad M \in \mathbb{N} \tag{20.3}$$

en

$$0 = \theta_0 \leq \theta_1 \leq \cdots \leq \theta_N = 2\pi \quad \text{met} \quad N \in \mathbb{N}, \tag{20.4}$$

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 (20.2) in de volgende stelling over harmonische functies.

**Exercise 20.1.** 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 13.5 als je de integraal van $\Delta u$ 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

$$\iint_{\mathbb{R}^2} u = \lim_{R\to\infty}\iint_{x^2+y^2\le R^2} u(x,y)d(x,y) \tag{20.5}$$

nu net zo natuurlijk gedefinieerd is in $[0, \infty]$ als door (20.1). Alleen een wiskundige vraagt zich dan af dit consistent is. Dat moet en dat mag hoor:

**Exercise 20.2.** 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 20.2 en de functie

$$(x,y) \xrightarrow{u} e^{-\frac{1}{2}(x^2+y^2)}$$

Kort door de bocht opgeschreven concluderen we dat

$$\left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx\right)^2 = \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\,dy = \iint_{\mathbb{R}^2} e^{-\frac{1}{2}(x^2+y^2)}\,d(x,y)$$

$$= \int_0^{\infty}\int_0^{2\pi} e^{-\frac{1}{2}r^2}\,r d\theta dr = \int_0^{\infty} 2\pi e^{-\frac{1}{2}r^2}\,rdr = 2\pi[-e^{-\frac{1}{2}r^2}]_0^{\infty} = 2\pi.$$

**Exercise 20.3.** Laat met Opgave 20.2 zien dat

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\, dx = \sqrt{2\pi}.$$

Bijgevolg hebben

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \text{en} \quad u(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} \tag{20.6}$$

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 (20.6) worden de 1-en 2-dimensionale standaard *normale verdeling* genoemd. Is de functie $g$ hetzelfde als de functie $f$ in (20.6), 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 $\theta$ 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 20.4.** 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.

## 20.2 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

$$x = r\cos\theta \quad \text{en} \quad y = r\sin\theta \tag{20.7}$$

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 $\theta$. In wat volgt zien we (20.7) 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.

### 20.2.1 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

$$F(x_0 + h, y_0 + k) = F(x_0, y_0) + ah + bk + R(h, k; x_0, y_0), \tag{20.8}$$

met $a, b \in \mathbb{R}$ en

$$\frac{R(h, k; x_0, y_0)}{\sqrt{h^2 + k^2}} \to 0 \quad \text{als} \quad \sqrt{h^2 + k^2} \to 0, \tag{20.9}$$

vergelijk met het eerdere uitpakken. De volgende opgave is misschien nu wat dubbelop, maar dat kan geen enkel kwaad.

**Exercise 20.5.** 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 (20.8) geldt met (20.9). 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

$$a = F_x(x_0, y_0) = \frac{\partial F}{\partial x}(x_0, y_0) = (\delta_x F)(x_0, y_0) = (D_1 F)(x_0, y_0); \tag{20.10}$$

$$b = F_y(x_0, y_0) = \frac{\partial F}{\partial y}(x_0, y_0) = (\delta_y F)(x_0, y_0) = (D_2 F)(x_0, y_0), \tag{20.11}$$

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

$$e_x = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{en} \quad e_y = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{20.12}$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \nabla F = \frac{\partial F}{\partial x} e_x + \frac{\partial F}{\partial y} e_y = e_x \frac{\partial F}{\partial x} + e_y \frac{\partial F}{\partial y}, \tag{20.13}$$

de *gradient* van $F$ in $(x_0, y_0)$, geschreven zonder $(x_0, y_0)$. Merk op dat het lineaire gedeelte in (20.8) te schrijven is als

$$ah + bk = \begin{pmatrix} a \\ b \end{pmatrix} \cdot \begin{pmatrix} h \\ k \end{pmatrix} = \begin{pmatrix} h \\ k \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = h \frac{\partial F}{\partial x} + k \frac{\partial F}{\partial y}, \tag{20.14}$$

het inprodukt[2] van $\nabla F$ 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 (20.14) 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 $\nabla$ als

$$\nabla = e_x \frac{\partial}{\partial x} + e_y \frac{\partial}{\partial y} \quad \text{werkend op} \quad F \quad \text{geeft} \quad \nabla F, \tag{20.15}$$

een vectorwaardige differentiaaloperator.

Middels het inprodukt kan $\nabla$ 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

$$\nabla \cdot V = (e_x \frac{\partial}{\partial x} + e_y \frac{\partial}{\partial y}) \cdot (V_x e_x + V_y e_y) = \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y}, \tag{20.16}$$

[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 (20.12) 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 $\nabla$ werkt nu alleen op $V_x$ en $V_y$, en omdat

$$e_x \cdot e_x = e_y \cdot e_y = 1 \quad \text{en} \quad e_x \cdot e_y = e_y \cdot e_x = 0, \tag{20.17}$$

blijven er maar twee termen over in (20.16). 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 (20.16) 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 (20.17) en de andere twee vereenvoudigen en blijven in die vorm over in (20.16).

Als $V = \nabla F$ differentieerbaar is dan volgt zo dat

$$\nabla \cdot \nabla F = (e_x \frac{\partial}{\partial x} + e_y \frac{\partial}{\partial y}) \cdot (\frac{\partial F}{\partial x} e_x + \frac{\partial F}{\partial y} e_y) = \frac{\partial}{\partial x} \frac{\partial F}{\partial x} + \frac{\partial}{\partial y} \frac{\partial F}{\partial y} = \Delta F, \tag{20.18}$$

de Laplaciaan van $F$, die weer gezien kan worden als

$$\Delta F \quad \text{is} \quad \Delta = \frac{\partial}{\partial x} \frac{\partial}{\partial x} + \frac{\partial}{\partial y} \frac{\partial}{\partial y} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \quad \text{werkend op} \quad F. \tag{20.19}$$

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!

### 20.2.2 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 $(r_0,\theta_0)$ 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 $(r_0,\theta_0)$ 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 (20.14). 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 (20.14) 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

$$\frac{\partial G}{\partial r} = \frac{\partial X}{\partial r}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial r}\frac{\partial F}{\partial y} \tag{20.20}$$

$$\frac{\partial G}{\partial \theta} = \frac{\partial X}{\partial \theta}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial \theta}\frac{\partial F}{\partial y}, \tag{20.21}$$

in vector-matrixnotatie te schrijven als

$$\begin{pmatrix} \frac{\partial G}{\partial r} \\ \frac{\partial G}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \frac{\partial X}{\partial r}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial r}\frac{\partial F}{\partial y} \\ \frac{\partial X}{\partial \theta}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial \theta}\frac{\partial F}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{\partial X}{\partial r} & \frac{\partial Y}{\partial r} \\ \frac{\partial X}{\partial \theta} & \frac{\partial Y}{\partial \theta} \end{pmatrix} \begin{pmatrix} \frac{\partial F}{\partial x} \\ \frac{\partial F}{\partial y} \end{pmatrix}, \tag{20.22}$$

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 (20.7) 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 (20.14) hebben gebruikt.

De kolomvectoren in (20.22) zien er uit als gradiënten, maar dat is slechts misleidende schijn, zoals we in Sectie 20.2.4 zullen zien.

### 20.2.3 Kettingregel met Jacobimatrices

Mooie voorbeelden van matrixprodukten als in (18.3) zien we als we in (20.22) aan beide kanten links $(\tilde{h}\ \tilde{k})$ erbij zetten. Dan is

$$(\tilde{h}\ \tilde{k})\begin{pmatrix} \frac{\partial G}{\partial r} \\ \frac{\partial G}{\partial \theta} \end{pmatrix} = (\tilde{h}\ \tilde{k})\begin{pmatrix} \frac{\partial X}{\partial r} & \frac{\partial Y}{\partial r} \\ \frac{\partial X}{\partial \theta} & \frac{\partial Y}{\partial \theta} \end{pmatrix} \begin{pmatrix} \frac{\partial F}{\partial x} \\ \frac{\partial F}{\partial y} \end{pmatrix}, \tag{20.23}$$

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 (20.20) en (20.21) te komen.

Via links en rechts transponeren is (20.23) equivalent met

$$(\frac{\partial G}{\partial r}\ \frac{\partial G}{\partial \theta})\begin{pmatrix} \tilde{h} \\ \tilde{k} \end{pmatrix} = (\frac{\partial F}{\partial x}\ \frac{\partial F}{\partial y})\begin{pmatrix} \frac{\partial X}{\partial r} & \frac{\partial X}{\partial \theta} \\ \frac{\partial Y}{\partial r} & \frac{\partial Y}{\partial \theta} \end{pmatrix} \begin{pmatrix} \tilde{h} \\ \tilde{k} \end{pmatrix}, \tag{20.24}$$

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 (20.24) 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

$$\begin{pmatrix} \frac{\partial G_1}{\partial r} & \frac{\partial G_1}{\partial \theta} \\ \frac{\partial G_2}{\partial r} & \frac{\partial G_2}{\partial \theta} \end{pmatrix} \begin{pmatrix} \tilde{h} \\ \tilde{k} \end{pmatrix} = \begin{pmatrix} \frac{\partial F_1}{\partial x} & \frac{\partial F_1}{\partial y} \\ \frac{\partial F_2}{\partial x} & \frac{\partial F_2}{\partial y} \end{pmatrix} \begin{pmatrix} \frac{\partial X}{\partial r} & \frac{\partial X}{\partial \theta} \\ \frac{\partial Y}{\partial r} & \frac{\partial Y}{\partial \theta} \end{pmatrix} \begin{pmatrix} \tilde{h} \\ \tilde{k} \end{pmatrix}, \tag{20.25}$$

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 $r, \theta$ 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.

### 20.2.4 Omschrijven van differentiaaloperatoren

De notatie (20.23) 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 (20.22) en de kettingregel voor coördinatentransformaties te schrijven als

$$\begin{pmatrix} \frac{\partial}{\partial r} \\ \frac{\partial}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \frac{\partial X}{\partial r} & \frac{\partial Y}{\partial r} \\ \frac{\partial X}{\partial \theta} & \frac{\partial Y}{\partial \theta} \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix}, \tag{20.26}$$

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 20.6.** 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 20.6 zijn we nog niet klaar als we in (20.15)

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix} = e_x\frac{\partial}{\partial x} + e_y\frac{\partial}{\partial y}$$

willen omschrijven naar $r$ en $\theta$. De vraag is ook hoe we $e_x$ en $e_y$ omschrijven naar $e_r$ en $e_\theta$, en daarvoor komt de vraag wat $e_r$ en $e_\theta$ 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

$$\nabla = e_x\frac{\partial}{\partial x} + e_y\frac{\partial}{\partial y} = e_r\frac{\partial}{\partial r} + \frac{1}{r}e_\theta\frac{\partial}{\partial \theta} \tag{20.27}$$

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 $e_\theta$ 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 $\theta$ varieert. De voorfactor $\frac{1}{r}$ compenseert de met $r$ evenredige snelheid bij gelijkmatige toename van $\theta$.

**Exercise 20.7.** In (20.27) staan twee representaties van dezelfde operator. Door $e_x$ en $e_y$ in $e_r$ en $e_\theta$ uit te drukken en Opgave 20.6 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 20.8.** 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 $e_\theta$ van $\theta$ afhangen werkt met de produktregel van Leibniz de $e_\theta \frac{\partial}{\partial \theta}$ in de factor links nu ook op $e_r$ en $e_\theta$ in de factor rechts, en één van die twee geeft na inprodukt met de voorfactor $e_\theta$ een bijdrage.

**Exercise 20.9.** Pas de regel in Opgave 20.8 nu toe op $\nabla$ 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 20.8 toepassen op $\nabla$ 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 20.9 zien we

$$\Delta = \Delta_r + \frac{1}{r^2}\Delta_S, \tag{20.28}$$

[5]Wiskundig is dit per definitie en consistent met wat je ziet als je pijltjes tekent.

waarin $\Delta_r$ de radiële Laplaciaan is, die ook werkt op functies $R = R(r)$, en $\Delta_S$ de Laplace-Beltrami operator op

$$S = \{(x, y) \in \mathbb{R}^2 : \ x^2 + y^2 = 1\},$$

uitgedrukt in de hoekvariabele $\theta$ 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.

## 20.3 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 20.9 los te laten op

$$u(x, y) = R(r)\Theta(\theta), \tag{20.29}$$

en het resultaat gelijk aan nul te stellen. Dit geeft

$$\begin{aligned} 0 &= \left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}\right)R(r)\Theta(\theta) \\ &= \Theta(\theta)\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}\right)R(r) + \frac{R(r)}{r^2}\frac{\partial^2}{\partial \theta^2}\Theta(\theta) \\ &= \Theta(\theta)(R''(r) + \frac{1}{r}R'(r)) + \frac{R(r)}{r^2}\Theta''(\theta). \end{aligned}$$

Als $\Theta''$ een veelvoud is van $\Theta$, zeg

$$-\Theta'' = \mu\Theta \tag{20.30}$$

dan volgt Euler's vergelijking

$$R''(r) + \frac{1}{r}R'(r) = \mu\frac{R(r)}{r^2} \tag{20.31}$$

voor $R(r)$.

Merk op dat (20.30) gezien kan worden als een (eigenwaarde)probleem voor

$$-\Delta_S = -\frac{d^2}{d\theta}$$

op de eenheidscirkel waar bij $\Theta$ een $2\pi$-periodieke functie moet zijn om een functie op de cirkel

$$S = \{(x,y) : x^2 + y^2 = 1\}$$

te definiëren.

**Exercise 20.10.** Welke $\mu$ zijn toegestaan in (20.30) voor oplossingen (20.29) 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 (20.30) te kennen. Schrijf die harmonische polynomen in gescheiden variabelen $r$ en $\theta$ als $R(r)\Theta(\theta)$ en verifieer dat $R(r)$ een oplossing is van (20.31) met de bijbehorende $\mu$.

**Exercise 20.11.** 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 20.11 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

$$u(\cos\theta, \sin\theta) = f(\theta) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos k\theta + b_k \sin k\theta), \tag{20.32}$$

een zogenaamde *Fourierreeks*[7] voor een $2\pi$-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 20.12.** 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 $e_r, e_\theta, e_\phi$ 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

$$\nabla = e_r\frac{\partial}{\partial r} + \frac{1}{r}e_\theta\frac{\partial}{\partial\theta} + \frac{1}{r\sin\theta}e_\phi\frac{\partial}{\partial\phi}, \tag{20.33}$$

en gebruik (20.33) om voor

$$V = V_r e_r + V_\theta e_\theta + V_\phi e_\phi$$

eerst af te leiden dat

$$\nabla \cdot V = \frac{\partial V_r}{\partial r} + \frac{2}{r}V_r + \frac{1}{r}\left(\frac{\partial V_\theta}{\partial\theta} + \frac{\cos\theta}{\sin\theta}V_\theta + \frac{1}{\sin\theta}\frac{\partial V_\phi}{\partial\phi}\right),$$

en vervolgens via $V = \nabla F$ dat

$$\Delta = \underbrace{\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}}_{\Delta_r} + \frac{1}{r^2}\underbrace{\left(\frac{\partial^2}{\partial\theta^2} + \frac{\cos\theta}{\sin\theta}\frac{\partial}{\partial\theta} + \frac{1}{\sin\theta}\frac{\partial^2}{\partial\phi^2}\right)}_{\Delta_S}.$$

Wederom zien we hier (20.9), maar nu met $\Delta_S$ 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 $\Delta_S$ 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 20.13.** 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 $e_r, e_\theta, e_\phi$ 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

$$\nabla = e_r\frac{\partial}{\partial r} + \frac{1}{r}e_\theta\frac{\partial}{\partial\theta} + \frac{1}{r\sin\theta}e_\theta\frac{\partial}{\partial\phi} \tag{20.34}$$

en gebruik (20.34) 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 (20.9), maar nu met $\Delta_S$ 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 $\Delta_S$ 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.

## 20.4 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 $\nu$ on $\partial B$ 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 $\varepsilon$, basically as in Section 13.2. This leads to

$$\int_B (\varepsilon_t + \nabla \cdot w) = 0$$

for every such $B$ and therefore[10] to

$$\varepsilon_t + \nabla \cdot w = 0 \tag{20.35}$$

in $\Omega$.

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

$$w = -\kappa\nabla u, \quad \kappa = \kappa(x) > 0 \text{ thermal conductivity.} \tag{20.36}$$

If $\nu$ 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 (20.35) and (20.36), and a possible heat source $h = h(t,x)$, lead to the linear partial differential equation

$$(\chi\rho u)_t = \nabla \cdot \kappa\nabla u + h \tag{20.37}$$

[8] This section relates to Section 4.1 in Olver's PDE book.
[9] See Chapter 27.5.
[10] Certainly if the integrand is continuous.

for the temperature. *Only* if $\chi, \rho, \kappa$ are independent of $x$ this reduces, in the absence of heat sources, to

$$u_t = \gamma \Delta u, \quad \gamma = \frac{\kappa}{\chi \rho}. \tag{20.38}$$

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 (20.37). The standard homogeneous boundary conditions are therefore

$$u = 0 \quad \text{on} \quad \partial\Omega \qquad \text{(Dirichlet)} \tag{20.39}$$

and

$$\nu \cdot \nabla u = 0 \quad \text{on} \quad \partial\Omega \qquad \text{(Neumann)} \tag{20.40}$$

for $t > 0$.

The Robin boundary condition prescribes the flux in terms of also the temperature, e.g. the homogenous boundary condition

$$\nu \cdot \kappa \nabla u + \beta u = 0 \tag{20.41}$$

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 (20.40). 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 (20.38). 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

$$-\Delta v = \lambda v. \tag{20.42}$$

There are now two natural boundary conditions to choose from,

$$u = 0 \qquad \text{(Dirichlet) and} \qquad \nu \cdot \nabla u + \beta u = 0 \qquad \text{(Robin).} \tag{20.43}$$

The Neumann boundary condition corresponds to $\beta = 0$.

## 20.5 Intermezzo: het waterstofatoom

Met

$$V(r) = -\frac{e^2}{r}$$

is de stationaire Schrödinger vergelijking voor het waterstofatoom

$$\frac{\hbar^2}{2m}\Delta\psi - \frac{e^2}{r}\psi = E\psi, \tag{20.44}$$

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 (20.44) 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

$$\Delta = \underbrace{\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}}_{\Delta_r} + \frac{1}{r^2}\underbrace{(\frac{\partial^2}{\partial \theta^2} + \frac{\cos\theta}{\sin\theta}\frac{\partial}{\partial\theta} + \frac{1}{\sin\theta}\frac{\partial^2}{\partial\phi^2})}_{\Delta_S}.$$

Via

$$\psi(x,y,z) = R(r)P_l(\frac{x}{r},\frac{y}{r},\frac{z}{r}),$$

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

$$\frac{d^2u}{dx^2} + (\frac{4l}{x} - 2)\frac{du}{dx} = 2\frac{n-l-1}{x}u.$$

**Exercise 20.14.** Corrigeer eventuele typo's hierboven. De machtreeksoplossing[11]

$$u(x) = 1 + a_1x + a_2x^2 + a_3x^3 + \cdots$$

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.

# 21 Differential forms

Have a look at Section 10.3 and then look at the first part of the proof of Theorem 27.7. 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

$$\int_\Omega v_x = -\int_{\partial\Omega} vdy \quad \text{and} \quad \int_\Omega v_y = \int_{\partial\Omega} vdx, \tag{21.1}$$

in which the right hand sides are evaluated using the parameterisation[1]

$$x = x(t) = t \quad \text{and} \quad y = y(t) = f(t). \tag{21.2}$$

$$\underbrace{x(t)}_{x}, \quad \underbrace{y = f(t)}_{y}, \quad \underbrace{dx = x'(t)\,dt = dt}_{dx} \quad \text{and} \quad \underbrace{dy = y'(t)\,dt = f'(t)\,dt}_{dy}.$$

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 (21.1) as

$$\int_{\partial\Omega} \quad \text{acting on the 1-forms} \quad vdy = v(x, y)dy \quad \text{and} \quad vdx = v(x, y)dx?$$

If so, how should we see the (double) integrals on the left hand sides then? Recall that in Theorem 27.1 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

$$\int_c^d \underbrace{\int_a^b u(x,y)\,dx}_{\text{function of } y}\,dy = \underbrace{\int_{[a,b]\times[c,d]} u}_{J \text{ as in Theorem 27.1}} = \underbrace{\int_{[a,b]\times[c,d]}}_{\text{integral acting on}} \underbrace{udxdy}_{\text{2-form}},$$

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 (21.1) 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.

## 21.1 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

$$du = u_x dx + u_y dy = \frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy \tag{21.3}$$

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 7.9 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:

$$d(fdx + gdy) = d(fdx) + d(gdy) \quad \textit{(a sum rule for d)}$$

$$= \underbrace{dfdx + fddx}_{d(fdx)} + \underbrace{dgdy + gddy}_{d(gdy)} \quad \textit{(twice a Leibniz rule for d)}$$

$$= \underbrace{(f_x dx + f_y dy)}_{\textit{definition of df}} dx + fddx + \underbrace{(g_x dx + g_y dy)}_{\textit{definition of dg}} dy + gddy.$$

$$= f_x dxdx + f_y dydx + g_x dxdy + g_y dydy + fddx + gddy \quad \textit{(bye bye brackets)}$$

$$= f_x dxdx - f_y dxdy + g_x dxdy + g_y dydy + fddx + gddy \quad \textit{(if we use } dydx = -dxdy\textit{)}$$

$$= (g_x - f_y)dxdy + fddx + gddy \quad \textit{(if we use } dxdx = 0 = dydy\textit{)}.$$

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

$$u(x,y) \xrightarrow{d} u_x dx + u_y dy \xrightarrow{d} (u_{yx} - u_{xy})dxdy = 0$$

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

$$\begin{gathered} f \xrightarrow{d} \tfrac{\partial f}{\partial x}dx + \tfrac{\partial f}{\partial y}dy, \quad fdx + gdy \xrightarrow{d} (\tfrac{\partial g}{\partial x} - \tfrac{\partial f}{\partial y})dxdy, \\ fdxdy \xrightarrow{d} 0, \quad gdxdy \xrightarrow{d} 0, \end{gathered} \tag{21.4}$$

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 21.1.** Look at the forms in (21.1) and see how they are related by the $d$-algebra just developed.

We will be looking for a formulation in which the result is

$$\int_{\Omega} d\omega = \int_{\delta\Omega} d\omega \tag{21.5}$$

for a 1-form $\omega$ and a bounded domain $\Omega$ 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 27.8 in Section 27.17.

**Exercise 21.2.** 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 21.3.** Do it again for $F \xrightarrow{d} F'(x)dx$ and $f(x)dx \xrightarrow{d} 0$ with $f(x)$ and $F(x)$.

**Remark 21.4.** *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 $\frac{\partial}{\partial y}$.

## 21.2 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

$$dx = x'(t)dt \tag{21.6}$$

we recognise the d-algebra from Section 21.1, and we see that a 1-form $f(x)dx$ in $x$ is *pulled back* by $t \to x(t)$ to a 1-form

$$f(x)dx = f(x(t))x'(t)dt \tag{21.7}$$

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

$$f(x,y)dx + g(x,y)dy = (f(x(t),y(t))x'(t) + g(x(t),y(t))y'(t))dt \tag{21.8}$$

in $t$. So $t \to (x(t),y(t))$ leads to

$$f(x,y)dx + g(x,y)dy \xrightarrow{\textit{pull back}} (f(x(t),y(t))x'(t) + g(x(t),y(t))y'(t))dt$$

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

$$dxdy \to (\frac{\partial x}{\partial r}dr + \frac{\partial x}{\partial \theta}d\theta)(\frac{\partial y}{\partial r}dr + \frac{\partial y}{\partial \theta}d\theta) = (\frac{\partial x}{\partial r}\frac{\partial y}{\partial \theta} - \frac{\partial y}{\partial r}\frac{\partial x}{\partial \theta})drd\theta, \tag{21.9}$$

with the determinant[7] of the Jacobi matrix.

[7]Plus or minus the area spanned by the two vectors, compare to (19.2) in Chapter 27.10.

In case of $x = r\cos\theta, y = r\sin\theta$ this reads

$$dxdy \to rdrd\theta.$$

Note that (21.8) does not correspond to a coordinate transformation but (21.9) does. Have another look at the derivation of (21.8) and replace $t$ by $\phi$ in $[0, 2\pi]$. With $x(0) = x(2\pi)$ and $y(0) = y(2\pi)$ this compares to (21.9) with $r$ fixed, and you discover how the pull back algebra works for

$$(\theta, \phi) \to (x(\theta, \phi), y(\theta, \phi), z(\theta, \phi)) \tag{21.10}$$

and 2-forms in $x, y, z$.

**Exercise 21.5.** Pull $f(x, y, z)dx + g(x, y, z)dy + h(x, y, z)dz$ back to a 2-form in $\theta$ and $\phi$.

**Remark 21.6.** *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 21 introduced notation with $dx$ and $dy$ not separated and $dxdy = -dydx$.

# 22 Parameterisations and integrals

Part of this chapter was already done in Chapter 27. Let us recall the main result, which concerns a bounded open set $\Omega \subset \mathbb{R}^{\mathrm{N}}$ with $\partial\Omega \in C^1$ and a function $v \in C^1(\Omega) \cap C(\overline{\Omega})$. In the proof of Theorem 27.17 we explained how

$$\int_\Omega v_{x_i} = \int_{\partial\Omega} \nu_i\, v \tag{22.1}$$

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 27.3.

## 22.1 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 $\gamma : [a,b] \to \mathbb{R}^{\mathrm{N}}$. For any such $\gamma$ 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 $\gamma$ 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 $\gamma : [a,b] \to \mathbb{R}^{\mathrm{N}}$ 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

$$\Phi :\; u \to \gamma(\phi^{-1}(u)) \tag{22.2}$$

[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

$$\int_\gamma f = \int_\gamma f ds = \int_a^b f(\gamma(t))|\gamma'(t)|_2 \, dt = \int_c^d f(\Phi(u))|\Phi'(u)|_2 \, du = \int_\Phi f ds. \tag{22.3}$$

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.

## 22.2 Line integrals of vector fields along curves

Besides (22.3) as a 1-dimensional example of what is to come in (22.13) we can also define an integral for $F = F(x) \in \mathbb{R}^n$ continuous on $\gamma([a,b])$, namely

$$\int_\gamma F \cdot ds = \int_a^b F(\gamma(t)) \cdot \gamma'(t) \, dt = \int_a^b F(\gamma(t)) \cdot \underbrace{\frac{\gamma'(t)}{|\gamma'(t)|_2}}_{T(t)} |\gamma'(t)|_2 \, dt, \tag{22.4}$$

but Edwards avoids the commonly used notation in the left hand side of (22.4). 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 $\Phi$ defined as in (22.2) 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

$$\int_\gamma F \cdot ds = \int_\gamma F_1 dx_1 + \cdots + \int_\gamma F_N dx_N = \int_\gamma F_1 dx_1 + \cdots + F_N dx_N. \quad (22.5)$$

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]

$$df = \frac{\partial f}{\partial x_1}dx_1 + \cdots + \frac{\partial f}{\partial x_N}dx_N. \quad (22.6)$$

These 1-forms act on vectors. Whereas the $x$-dependent vector[7]

$$F(x) = F_1(x)e_1 + \cdots + F_N(x)e_N \quad (22.7)$$

The vector

$$v = v_1e_1 + \cdots + v_Ne_N$$

have and $x$-dependent inner product

$$F(x) \cdot v = F_1(x)v_1 + \cdots + F_N(x)v_N.$$

the 1-form

$$\omega = F_1(x)dx_1 + \cdots + F_N(x)dx_N \quad (22.8)$$

[5]Writing $f$ instead of $F$ again.

[6]Compare to (21.3) in Section 21.1.

[7]As in Section 17.2 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, $\omega$ 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 $\gamma$ with $u = \phi(t)$ and $\phi'(t) < 0$ changes the sign of the integral.

The notation for $\omega$ 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 $\omega = F_1(x)dx_1 + \cdots + F(x)dx_N$.

## 22.3 Surface area

We need some linear algebra[8] for integrals over more general surface patches then the ones encountered in Section 27.5. *We now understand a surface patch* to be a set in $\mathbb{R}^3$ parameterised by a continuously differentiable injective map

$$\Phi : [0,1] \times [0,1] \to \mathbb{R}^3, \tag{22.9}$$

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 $\Phi_1, \Phi_2, \Phi_3$ of $\Phi$ 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 (17).

Momentarily switching to a notation with $\Phi_1, \Phi_2, \Phi_3$ 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

$$\Phi_i(u, v) = a_i u + b_i v \tag{22.10}$$

[8] Theorem 18.8.

[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 (18.14) starting the discussion in Section 18.5 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 $\Phi$ defined by (22.10), and its area is then equal to

$$\int_0^1 \int_0^1 \mathcal{M}_2(\frac{\partial\Phi}{\partial u}, \frac{\partial\Phi}{\partial v})\, du\, dv, \tag{22.11}$$

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 (22.10).

It will be no surprise that (22.11) will also be used to define the area of the surface patch defined by $\Phi$ if $\Phi$ is not a linear map from $[0,1]^2$ to $\mathbb{R}^3$, and that everything generalises to $\Phi : [0,1]^n \to \mathbb{R}^{\mathrm{N}}$ with $1 \leq n < N$. We expand on the linear case of this generalisation next.

## 22.4 Surface integrals

I now return to (22.11). Generalising to $1 \leq n \leq N$ we consider

$$\int_{[0,1]^n} \mathcal{M}_n(\frac{\partial\Phi}{\partial u})\, du = \int_0^1 \cdots \int_0^1 \mathcal{M}_n(\frac{\partial\Phi}{\partial u_1}, \ldots, \frac{\partial\Phi}{\partial u_n})\, du_1 \cdots du_n \tag{22.12}$$

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 18.8. 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

$$dS_n = \mathcal{M}_n(\frac{\partial \Phi}{\partial u})\, du = \mathcal{M}_n(\Phi_{u_1}, \ldots, \Phi_{u_n})\, du_1 \cdots du_n$$

in (22.12), i.e.

$$\int_\Phi dS_n = \int_0^1 \cdots \int_0^1 \mathcal{M}_n(\frac{\partial \Phi}{\partial u_1}, \ldots, \frac{\partial \Phi}{\partial u_n})\, du_1 \cdots du_n.$$

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

$$\int_\Phi f dS_n = \int_{[0,1]^n} f(\Phi(u)) \mathcal{M}_n(\frac{\partial \Phi}{\partial u})\, du = \tag{22.13}$$

$$\int_0^1 \cdots \int_0^1 f(\Phi(u_1, \ldots, u_n)) \mathcal{M}_n(\frac{\partial \Phi}{\partial u_1}, \ldots, \frac{\partial \Phi}{\partial u_n})\, du_1 \cdots du_n.$$

The subscript $\Phi$ on the integral is consistent with the case $n = 1$ and $ds = dS_1$, and coincides with the notation in the second part of (22.3). 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 $[0, 1]^n$.

# 23 Varieties in Euclidean space

In this chapter we think of manifolds as solution sets of systems of equations in $\mathbb{R}^N$. In Chapter 24 this will bother us a 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 $\mathbb{R}^{\mathrm{N}}$ 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]

$$y = Ax + b, \tag{23.1}$$

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 23.1.** 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 (23.1).

If we see $x$ and $y$ as column vectors then (23.1) reads as

$$(A \quad -I)\begin{pmatrix} x \\ y \end{pmatrix} = b \in \mathbb{R}^m,$$

with $C = (A \quad -I)$ 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 $\mathbb{R}^{\mathrm{N}}$. Thus (23.1) is a system of $m$ linear equation for $N$unknowns $z_1, z_2, \ldots, z_N$:

$$C_{11}z_1 + C_{12}z_2 + \cdots + C_{1N}z_N = b_1;$$

$$C_{21}z_1 + C_{22}z_2 + \cdots + C_{2N}z_N = b_2;$$

$$\vdots$$

[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 (23.1). We have

$$Cz = b \iff y = Ax + b \tag{23.2}$$

as equivalent descriptions of non-trivial linear varieties in $\mathbb{R}^{\mathrm{N}}$, under the assumption that $C$ hax maximal rank.

## 23.1 Implicit function theorem in Euclidean spaces

Referring to Theorem 14.4 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 (14.25) 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 23.2.** *(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 14.1 and 14.2. 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 $0 < \varepsilon_0 \le r$ 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 23.2 can be read as stating the existence of $0 < \delta_0 \leq \varepsilon_0 \leq r$ for which the assertions hold.

Applying Theorem 23.2 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 23.3.** *(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 23.2 gives $f : \bar{B}_{\delta_0}(a) \to B_{\varepsilon_0}(b)$ in Theorem 23.3 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 23.3 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

$$\mathbb{R}^n \supset \bar{B}_{\delta_0}(a) \xrightarrow{f} \mathbb{R}^n \tag{23.3}$$

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 23.4.** Derive 23.2 from Theorem 23.3. 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.

## 23.2 General subvarieties

For in general nonlinear subvarieties[5] we ask about an equivalence simliar to (23.2), starting from the nonlinear version $Cz = b$, written in Theorem 23.2 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 (23.1) to agree what we mean by a subvariety $\mathcal{M} \subset \mathbb{R}^{\mathrm{N}}$:

**Definition 23.5.** *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 23.2: 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 23.6.** Let $F : \mathrm{I\!R}^{\mathrm{N}} \to \mathrm{I\!R}^m$ be continuously differentiable. Assume that for all $z \in \mathrm{I\!R}^{\mathrm{N}}$ with $F(z) = 0$ the derivative $F'(z)$, seen as matrix, has maximal rank. Prove that $\{z \in \mathrm{I\!R}^{\mathrm{N}} : F(z) = 0\}$ is an $n$-dimensional subvariety of $\mathrm{I\!R}^{\mathrm{N}}$, with $n + m = N$.

**Exercise 23.7.** Give an example of an $n$-dimensional subvariety $\mathcal{M} \subset \mathrm{I\!R}^{\mathrm{N}}$ which is not given by a function $F$ as in Exercise 23.6.

The standard example for Exercise 23.6 is the boundary of a ball in $\mathrm{I\!R}^n$ with center $(a_1, a_2, \dots, a_n)$ and radius $\delta > 0$:

$$(x_1 - a_1)^2 + \cdots + (x_n - a_n)^2 - \delta^2 = 0. \tag{23.4}$$

There are three equivalent ways to say that $\mathcal{M} \subset \mathrm{I\!R}^{\mathrm{N}}$ is an $n$-dimensional subvariety:

(A) $\mathcal{M}$ is locally the graph of a continuously differentiable function

$$f : \mathrm{I\!R}^n \to \mathrm{I\!R}^m \quad (n + m = N),$$

given by $y = f(x)$ after renumbering $z = (x, y)$.

(B) $\mathcal{M}$ is locally the zero level set of

$$F : \mathrm{I\!R}^{\mathrm{N}} \to \mathrm{I\!R}^m \quad (n + m = N),$$

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

$$\Phi : \mathrm{I\!R}^n \to \mathrm{I\!R}^{\mathrm{N}},$$

which is injective and has $\Phi'$ of maximal rank.

Theorem 23.2 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 $\Phi$ is called a chart on $M$.

$C$ with $\Phi(x) = (x, f(x))$. To complete the circle with a proof that (C) $\implies$ (A) we use Theorem 23.3 and the chain rule.

To wit, consider $\Phi$ 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 23.3, applied to $g = \Psi$ with $y = x$, provided us with a continuously differentiable injective function $f$ renamed here as $\phi$, $\phi : \bar{B}_{\delta_0}(\Psi(a)) \to \mathbb{R}^n$, 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

$$\text{(A)} \iff \text{(B)} \iff \text{(C)}$$

is now complete.

**Exercise 23.8.** 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 23.9.** As Exercise 23.8, 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.

## 23.3 Images of ball boundaries

With $0 < \varepsilon_1 \leq \delta_1 \leq \delta_0 \leq \varepsilon_0$ Theorem 23.3 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 $\delta_1$ and $\varepsilon_1$ 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

$$S_h = \{h(x)x : x \in S^{n-1}\}. \tag{23.5}$$

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 23.10.** 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 23.10 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 23.3 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 23.2, to be expanded on.

**Exercise 23.11.** Prove the statement in Exercise 23.10 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.

## 23.4 Coordinate transformations

If a point $P$ on an $n$-dimensional subvariety $\mathcal{M}$ of $\mathrm{IR}^{\mathrm{N}}$ lies in the image of a $\Phi$ and a $\Psi$ as in (C) in Section 23.2, say with $\Phi(\xi)$ and $\Psi(\eta)$, and $P = \Phi(0) = \Psi(0)$, with 0 an interior point of the domains of $\Phi$ and $\Psi$, then $\xi$ are $\eta$ are related by statements as in Theorem 23.3 in a neighbourhood of 0.

## 23.5 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.

# 24 Integration over manifolds

Section 23.2 and Section 25.2 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 25.2 above[1], such that

$$M = \Phi_1((a_1, b_1)) \cup \cdots \cup \Phi_m((a_n, b_n)) = \Phi_1([a_1, b_1]) \cup \cdots \cup \Phi_m([a_n, b_n]), \tag{24.1}$$

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 $\zeta_i$ 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 $\zeta_i$ leave the whole of $M$ as *bright* as it was before. Such *fading* functions $\zeta_i$ 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

$$\int_M f dS_n = \int_{\Phi_1} f\zeta_1 dS_n + \cdots + \int_{\Phi_m} f\zeta_m dS_n, \tag{24.2}$$

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 23.4. 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]

$$M = \Phi_1([a_1, b_1]) \cup \cdots \cup \Phi_m([a_n, b_n]) \quad \text{with} \quad \Phi_i((a_i, b_i)) \cap \Phi_j((a_j, b_j)) = \emptyset \tag{24.3}$$

for $i \neq j$, then

$$\int_M f dS_n = \int_{\Phi_1} f\, dS_n + \cdots + \int_{\Phi_m} f dS_n \tag{24.4}$$

is the obvious definition which Edwards uses, and which is what you do in examples. Usually there are many ways to choose the patches.

## 24.1 More integration of differential forms

We look again at the right hand side of (27.17) with $N = n+1$, evaluated for $\tilde{v}_i = \zeta v_i$ with $\zeta$ 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

$$\Phi(u_1, \ldots, u_n) = (u_1, \ldots, u_n, f(u_1, \ldots, u_n)) \tag{24.5}$$

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

$$\begin{vmatrix} \frac{\partial \Phi_1}{\partial u_1} & \frac{\partial \Phi_2}{\partial u_1} & \frac{\partial \Phi_3}{\partial u_1} \\ \frac{\partial \Phi_1}{\partial u_2} & \frac{\partial \Phi_2}{\partial u_2} & \frac{\partial \Phi_3}{\partial u_2} \\ e_1 & e_2 & e_3 \end{vmatrix} = \begin{vmatrix} \frac{\partial \Phi_1}{\partial u_1} & \frac{\partial \Phi_2}{\partial u_1} \\ \frac{\partial \Phi_1}{\partial u_2} & \frac{\partial \Phi_2}{\partial u_2} \end{vmatrix} e_3 + \begin{vmatrix} \frac{\partial \Phi_2}{\partial u_1} & \frac{\partial \Phi_3}{\partial u_1} \\ \frac{\partial \Phi_2}{\partial u_2} & \frac{\partial \Phi_3}{\partial u_2} \end{vmatrix} e_1 + \begin{vmatrix} \frac{\partial \Phi_3}{\partial u_1} & \frac{\partial \Phi_1}{\partial u_1} \\ \frac{\partial \Phi_3}{\partial u_2} & \frac{\partial \Phi_1}{\partial u_2} \end{vmatrix} e_2 \tag{24.6}$$

[4] Edwards: a hard theorem says this can be done.

[5] Like in Section 27.5.

is commonly called the cross product of the vectors $\Phi_{u_1}$ and $\Phi_{u_2}$, and for $\Phi(u_1,u_2) = (u_1,u_2,f(u_1,u_2))$ it evaluates as[6]

$$e_3 - \frac{\partial f}{\partial u_1}e_1 - \frac{\partial f}{\partial u_2}e_2 = -\frac{\partial f}{\partial u_1}e_1 - \frac{\partial f}{\partial u_2}e_2 + e_3, \tag{24.7}$$

which is a positive multiple of the unit vector $\nu$ 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 $\Phi_{u_1}$ and $\Phi_{u_2}$ linearly independent, the vector defined by (24.6) is perpendicular to the plane spanned by $\Phi_{u_1}$ and $\Phi_{u_2}$, 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 18.8. If we call this normalised vector $\nu$, which in case of (24.7) is simply[7]

$$\nu = \frac{1}{\sqrt{1+f_{u_1}^2+f_{u_2}^2}}\left(-\frac{\partial f}{\partial u_1}e_1 - \frac{\partial f}{\partial u_2}e_2 + e_3\right), \tag{24.8}$$

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

$$\int_\Phi \tilde{v}_1\, dx_2dx_3 + \tilde{v}_2\, dx_3dx_1 + \tilde{v}_3\, dx_1dx_2 = \int_\Phi \omega, \tag{24.9}$$

with

$$\omega = \tilde{v}_1\, dx_2dx_3 + \tilde{v}_2\, dx_3dx_1 + \tilde{v}_3\, dx_1dx_2,$$

[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 (24.9) 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

$$\int_\Omega \left( \frac{\partial \tilde{v}_1}{\partial x_1} + \frac{\partial \tilde{v}_2}{\partial x_2} + \frac{\partial \tilde{v}_3}{\partial x_3} + \frac{\partial \tilde{v}_4}{\partial x_4} \right) dx \tag{24.10}$$

$$= \int_\Phi \tilde{v}_1\, dx_2dx_3dx_4 + \cdots \text{ (cyclicly permutated terms) } \cdots = \int_\Phi \omega,$$

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 (24.10) 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 21, 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 (21.9) in Section 21.2.

[9] This is why we put the unit vectors in the last row of the determinant in (24.6).

We conclude with the simplest but slightly confusing case, $n = 1$ and $N = 2$, when (24.6) should be replaced by

$$\begin{vmatrix} \frac{\partial \Phi_1}{\partial u} & \frac{\partial \Phi_2}{\partial u} \\ e_1 & e_2 \end{vmatrix} = \frac{\partial \Phi_2}{\partial u}\, e_1 - \frac{\partial \Phi_1}{\partial u}\, e_2, \tag{24.11}$$

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

$$\int_{\partial\Omega} p(x,y)dx + q(x,y)dy = \int_\Omega (q_x - p_y)dxdy, \tag{24.12}$$

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 22.2. Section 24.2 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 $\Omega$ 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.

## 24.2 From Green's to Stokes' curl theorem

Now consider (24.5) as a local description of a manifold $M$ and forget about $\Omega$ as being a domain with $M = \partial\Omega$. Instead let $\Omega$ be as in (22.1) with $N = 2$ and let $M$ be the graph of $f : \Omega \to \mathrm{I\!R}$. 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

$$t \xrightarrow{\gamma} (u_1(t), u_2(t), f(u_1(t), u_2(t))) \tag{24.1}$$

parameterises the "boundary"

$$\partial M = \{\underbrace{(u, f(u))}_{\Phi(u)} : \ u \in \partial\Omega\},$$

and

$$u \xrightarrow{\Phi} (u, f(u)) \tag{24.2}$$

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 (22.7) and (22.8) and consider the integral

$$\int_{\partial M} \omega$$

as in (22.5). It evaluates as

$$\int_{\partial M} \omega = \int_0^1 (F_1(\gamma(t))\gamma_1'(t) + F_2(\gamma(t))\gamma_2'(t)) + F_3(\gamma(t))\gamma_3'(t)))\, dt$$

$$= \int_0^1 (F_1(u(t), f(u(t))u_1'(t) + F_3(u(t), f(u(t)))f_{u_1}(u(t)u_1'(t))\, dt$$

$$+ \int_0^1 (F_2(u(t), f(u(t))u_2'(t) + F_3(u(t), f(u(t)))f_{u_2}(u(t)u_2'(t))\, dt =$$

$$\int_{\partial\Omega} \zeta = \int_\Omega d\zeta, \tag{24.3}$$

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

$$d\zeta = \phi(u_1, u_2) du_1 du_2 \tag{24.4}$$

with $\phi(u_1, u_2)$ given by

$$\phi = -\underbrace{\left(\frac{\partial F_3}{\partial x_2} - \frac{\partial F_2}{\partial x_3}\right)}_{G_1}\frac{\partial f}{\partial u_1} - \underbrace{\left(\frac{\partial F_1}{\partial x_3} - \frac{\partial F_3}{\partial x_1}\right)}_{G_2}\frac{\partial f}{\partial u_2} + \underbrace{\left(\frac{\partial F_2}{\partial x_1} - \frac{\partial F_1}{\partial x_2}\right)}_{G_3} \tag{24.5}$$

$$= -G_1\frac{\partial f}{\partial u_1} - G_2\frac{\partial f}{\partial u_2} + G_3.$$

You should note that the *second order derivatives* of (24.2) are dropouts in the calculations that lead to (24.5).

Now compare (24.5) to $\nu$ in (24.8) and recall that for $\Phi$ given by (24.2) 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 (24.5), and commonly denoted as $G = \nabla \times F$, i.e.

$$\int_{\partial M} (F \cdot \tau)\, dS_1 = \int_{M} (G \cdot \nu)\, dS_2 \quad \text{with} \quad G = \nabla \times F, \tag{24.6}$$

using the parameterisations as indicated[11]. But don't say goodbye:

## 24.3 Pullbacks and the action of d

We already saw in the reasoning from (22.5) to (22.6) that $d$ acting on a $C^1$-function $f = f(x_1, \ldots, x_N)$ produces a 1-form

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_N} dx_N = \frac{\partial f}{\partial x_i} dx_i, \tag{24.7}$$

using the convention that we sum over repeated indices. With $f(x_1, \ldots, x_N)$ replaced by $u(x, y)$ this is (21.3) in Section 21.1. There I played with the $d$-algebra that emerges whenever you do integration using formal notations such as (10.6), which is just (24.7) 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 23.2. We use $\Phi$ 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 (24.9) and the formal rules that emerge in the calculations to do so. Thus we certainly want to deal with

$$f(x) = \phi(u) \quad \text{via} \quad x = \Phi(u). \tag{24.8}$$

A mathematician's way to do so is to introduce

$$\phi = \Phi^*(f) = f \circ \Phi, \tag{24.9}$$

the pullback of $f$ via $\Phi$, which then also provides us with

$$d\phi = \frac{\partial \phi}{\partial u_1} du_1 + \cdots + \frac{\partial \phi}{\partial u_n} du_n. \tag{24.10}$$

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

$$\Phi^*(f_i dx_i) = \underbrace{\Phi^*(f_i)}_{\phi_i} \Phi^*(dx_i), \tag{24.11}$$

[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

$$\Phi^*(df) = \underbrace{\frac{\partial f}{\partial x_i}((\Phi(u))}_{\Phi^*(D_i f)(u)} \Phi^*(dx_i). \tag{24.12}$$

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 (24.10) implies via the chain rule that

$$d(\Phi^*(f)) = \frac{\partial}{\partial u_j}(f(\Phi(u))\, du_j = \frac{\partial f}{\partial x_i}(\Phi(u))\frac{\partial \Phi_i}{\partial u_j}\, du_j, \tag{24.13}$$

and comparing to (24.12) we see that, if we define the pullback of $dx_i$ under $\Phi$ to be

$$\Phi^*(dx_i) = \frac{\partial \Phi_i}{\partial u_j}\, du_j, \tag{24.14}$$

it follows that

$$\Phi^*(df) = \Phi^*(df). \tag{24.15}$$

The definition of $\Phi^*(dx_i)$ by (24.14) 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 (24.8) is formalised by (24.9). It implies that the pullback of the 1-form in (24.11) evaluates as

$$\underbrace{\Phi^*(f_i dx_i) = \phi_i \frac{\partial \Phi_i}{\partial u_j}\, du_j}_{with\ \phi_i(u)=f_i(\Phi(u))} = f_i(\Phi(u))\, \frac{\partial \Phi_i}{\partial u_j}\, du_j = f_i(\Phi(u))\, D_j\Phi_i(u)\, du_j. \tag{24.16}$$

Next we observe that $d$ acting on the resulting 1-form in (24.16) may be evaluated, using the chain rule and $du_k du_j = -du_j du_k$, as

$$d(\Phi^*(f_i dx_i)) = d(f_i(\Phi(u))\, \frac{\partial \Phi_i}{\partial u_j}\, du_j) = \frac{\partial}{\partial u_k}(f_i(\Phi(u))\, \frac{\partial \Phi_i}{\partial u_j})\, du_k du_j$$

$$= (\frac{\partial}{\partial u_k}(f_i(\Phi(u)))\, \frac{\partial \Phi_i}{\partial u_j}\, du_k du_j + f_i(\Phi(u))\, \underbrace{\frac{\partial^2 \Phi_i}{\partial u_k \partial u_j}\, du_k du_j}_{zero\ the\ hero!} =$$

$$\frac{\partial f_i}{\partial x_k}(\Phi(u))\, \frac{\partial \Phi_k}{\partial u_k}\, \frac{\partial \Phi_i}{\partial u_j}\, du_k\, du_j = \Phi^*(D_k f_i)\, \frac{\partial \Phi_k}{\partial u_k}\, \frac{\partial \Phi_i}{\partial u_j}\, du_k\, du_j, \tag{24.17}$$

in which we used

$$d(f_i dx_i) = \frac{\partial f_i}{\partial x_k}\, dx_k\, dx_i \tag{24.18}$$

in the $u$-variables. Recall that this was the definition[12] in Section 21.1 of the action of $d$ on 1-forms. With

$$\Phi^*(f_{ij} dx_i dx_j) = \underbrace{\Phi^*(f_{ij})}_{\phi_{ij}}\, \Phi^*(dx_i dx_j) \tag{24.19}$$

as the obvious defining analog of (24.11), we have that

$$\Phi^*(d(f_i dx_i)) = \Phi^*(\frac{\partial f_i}{\partial x_k}\, dx_k dx_i) = \Phi^*(D_k f_i)\, \Phi^*(dx_k dx_i). \tag{24.20}$$

Comparing to (24.20) to (24.17) it follows that

$$\Phi^*(d(f_i dx_i)) = d(\Phi^*(f_i dx_i)), \tag{24.21}$$

provided we define

$$\Phi^*(dx_k dx_i) = \underbrace{\frac{\partial \Phi_k}{\partial u_k}\frac{\partial \Phi_i}{\partial u_j}\, du_k\, du_j}_{sum\ over\ 1\leq k,j\leq n} = \underbrace{(\frac{\partial \Phi_k}{\partial u_k}\frac{\partial \Phi_i}{\partial u_j} - \frac{\partial \Phi_k}{\partial u_k}\frac{\partial \Phi_i}{\partial u_j})\, \underline{du_k\, du_j}}_{sum\ over\ 1\leq k<j\leq n}$$

$$= \frac{\partial(\Phi_k, \Phi_i)}{\partial u_k \partial u_j}\, \underline{du_k\, du_j}, \tag{24.22}$$

in which the underline indicates that we sum over all $k, j$ with $1 \leq k < j \leq n$. Just as in (24.15) we see that the actions of $d$ and $\Phi^*$ commute.

Note that the second order derivatives have disapperared in (24.17). 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 $\Phi$ is only in $C^1$, because the determinants in (24.22) are exactly the determinants that showed up in (24.6) and the subsequent derivation of (24.9), 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 $\Phi^*$ on $k$-forms of any order $k$ is easily made once the reasoning above is understood. For any $k$-form

$$\omega = f_{i_1,\ldots,i_k} dx_{i_1} \cdots dx_{i_k}$$

[12] Recall the choice to set $ddx_i = 0$, leading to $dd\omega = 0$ for any form $\omega$.

[13] Using approximation arguments.

we have

$$\Phi^*(d\omega) = d(\Phi^*(\omega)) \tag{24.23}$$

Every such form may be written as

$$\omega = f_{i_1,\dots,i_k} dx_{i_1} \cdots dx_{i_k} = \tilde{f}_{i_1,\dots,i_k} \, \underline{dx_{i_1} \cdots dx_{i_k}}, \tag{24.24}$$

where in the second expression we sum only over those $i_1, \dots, i_k$ for which $1 \le i_1 < \cdots < i_k \le N$. 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

$$d\zeta = (\frac{\partial g_1}{\partial x_1} + \frac{\partial g_2}{\partial x_2} + \frac{\partial g_3}{\partial x_3}) \, dx dy dz$$

in Section 24.4 show.

Finally we observe that if we put the coefficients $f_1, f_2, f_3$ of this $\omega$ 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 $\eta$ 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 24.4 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:

$$\nabla \cdot \nabla \times F = 0.$$

## 24.4 From Gauss' to general Stokes' Theorem

From Section 24.1 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 $\Omega$ in every patch as above. For every continuously differentiable $V : \Omega \to \mathbb{R}^{N}$ it now holds that

$$\int_{\Omega} \nabla \cdot V = \int_{\partial\Omega} \nu \cdot V \, dS_{N-1}, \tag{24.25}$$

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

$$\omega = f_1(x)dx_1 + f_2(x)dx_2 + f_3(x)dx_3 \tag{24.26}$$

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 (24.12) applies, whence

$$\int_{\partial\Omega} \zeta = \int_{\partial\Omega} p_1(u_1, u_2)du_1 + p_2(u_1, u_2)du_2 = \int_{\Omega} (\frac{\partial p_2}{\partial u_1} - \frac{\partial p_1}{\partial u_2})du_1 du_2 = \int_{\Omega} d\zeta. \tag{24.27}$$

Observe that the second equality in (24.27) holds in view of (24.12), which is a rewritten version of (24.25) 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 (24.18).

We need

$$\int_{\partial\Omega} \zeta = \int_{\partial\Omega} \Phi^* \omega = \int_{\Phi(\partial\Omega)} \omega, \tag{24.28}$$

and

$$\int_{\Omega} d\zeta = \int_{\Omega} d\,\Phi^* \omega = \int_{\Omega} \Phi^*(d\omega) = \int_{\phi(\Omega)} d\omega \tag{24.29}$$

to conclude for $\omega$ given by (24.26) that

$$\int_{dS} \omega = \int_{dS} f_1 dx_1 + f_2 dx_2 + f_3 dx_3 = \int_S d\omega, \tag{24.30}$$

in which $S = \Phi(\Omega)$. It is the last equality in each of (24.28) and (24.29) that has to be checked, the other equalities follow from our $d$-algebra and the commutation of $d$ and $\Phi^*$.

Let us once more spell out the $d$-algebra by which (24.18) evaluates as

$$d\omega = (\frac{\partial f_1}{\partial x_1} dx_1 + \frac{\partial f_1}{\partial x_2} dx_2 + \frac{\partial f_1}{\partial x_3} dx_3) dx_1$$

$$+(\frac{\partial f_2}{\partial x_1} dx_1 + \frac{\partial f_2}{\partial x_2} dx_2 + \frac{\partial f_2}{\partial x_3} dx_3) dx_2$$

$$+(\frac{\partial f_3}{\partial x_1} dx_1 + \frac{\partial f_3}{\partial x_2} dx_2 + \frac{\partial f_3}{\partial x_3} dx_3) dx_3 =$$

$$\frac{\partial f_1}{\partial x_2} dx_2 dx_1 + \frac{\partial f_2}{\partial x_1} dx_1 dx_2 + \frac{\partial f_1}{\partial x_3} dx_3 dx_1 + \frac{\partial f_3}{\partial x_1} dx_1 dx_3 + \frac{\partial f_2}{\partial x_3} dx_2 + \frac{\partial f_3}{\partial x_2} dx_2 dx_3$$

$$= (\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2}) dx_1 dx_2 + (\frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3}) dx_2 dx_3 + (\frac{\partial f_1}{\partial x_3} - \frac{\partial f_3}{\partial x_1}) dx_3 dx_1$$

$$(\frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3}) dx_2 dx_3 + (\frac{\partial f_1}{\partial x_3} - \frac{\partial f_3}{\partial x_1}) dx_3 dx_1 + (\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2}) dx_1 dx_2$$

$$= g_1 dx_2 dx_3 + g_2 dx_3 dx_1 + g_3 dx_1 dx_2.$$

Comparing to (24.9) we recognise for $F(x) = f_1(x)e_1 + f_2(x)e_2 + f_3(x)e_3$ that

$$\int_{dS} f_1 dx_1 + f_2 dx_2 + f_3 dx_3 =$$

$$\int_S (\frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3}) dx_2 dx_3 + (\frac{\partial f_1}{\partial x_3} - \frac{\partial f_3}{\partial x_1}) dx_3 dx_1 + (\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2}) dx_1 dx_2$$

$$= \int_S G \cdot \nu = \int_S (\nabla \times F) \cdot \nu, \tag{24.31}$$

in which $g_1(x)e_1 + g_2(x)e_2 + g_3(x)e_3 = G(x) = \nabla \times F$ and $\nu$ is the normal vector on $S = \Phi(\Omega)$ defined by (24.6).

It thus remains to check the two analytical statements

$$\int_{\partial\Omega} \Phi^*\omega = \int_{\Phi(\partial\Omega)} \omega \quad \text{and} \quad \int_{\Omega} \Phi^*(d\omega) = \int_{\Phi(\Omega)} d\omega, \tag{24.32}$$

which complement the $d$-algebra presented above, and which are both of the form

$$\int_M \Phi^*\omega = \int_{\Phi(M)} \omega, \tag{24.33}$$

with respectively $M = \partial\Omega$ and $M = \Omega$. For this we need again Section 19 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 $\Phi$ and patches of $M$, see the remark at the end of Section 25.3. Also, we still have to deal with integrals over manifolds with boundaries, to obtain

$$\int_{\Phi(\partial\Omega)} \omega = \int_{\Phi(\Omega)} d\omega, \tag{24.34}$$

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, $\Phi$ a continuously differentiable injective map from $\Omega$ to $\mathbb{R}^N$ with Jacobian matrix of rank $n$ throughout $\Omega$, and $\omega$ an $n$-form with continuously differentiable coefficients. Generalisations to piecewise $C^1$-boundaries then still have to be discussed.

## 24.5 More exercises

Let $\Omega$ be the open unit disk. Then its boundary $\partial\Omega$ is the circle defined by

$$x^2 + y^2 = 1.$$

Graph parameterisations such as

$$x \to (x, \sqrt{1-x^2}), \quad x \to (x, -\sqrt{1-x^2}), \tag{24.35}$$

$$y \to (\sqrt{1-y^2}, y), \quad y \to (-\sqrt{1-y^2}, y)$$

are ugly for calculations. Much nicer and more in common is of course

$$\phi \to (\cos\phi, \sin\phi), \tag{24.36}$$

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

$$t = \frac{s}{1-s} \tag{24.37}$$

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 24.1.** Use the $t$-parameterisations above to calculate the area of the unit disk via integrals such as $\int xdy$ of $\int ydx$ 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 24.2.** 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 (24.11) defines a vector pointing out of $\Omega$.

**Exercise 24.3.** Same as Exercise 24.2 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

$$F(x, y) = x^3 + y^3 - 3xy = 0, \tag{24.38}$$

in the form

$$x = x(t) = \frac{3t}{1+t^3};\ y = y(t) = \frac{3t^2}{1+t^3}, \tag{24.39}$$

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 (24.39) with $t \in (-1, 1)$, the other by (24.39) with $x$ and $y$ interchanged. Exercise 2.3 in Chapter V of Edwards is about

$$\Omega = \{(x, y) \in \mathbb{R}^2 : x > 0,\ y > 0,\ F(x, y) = x^3 + y^3 - 3xy < 0\}. \tag{24.40}$$

with $\partial\Omega$ given by (24.39) and $t \in [0, \infty)$. You should examine the graphs of $x$ and $y$ as functions of $t$ in (24.39). You can get the area of $\Omega$ as

$$-\int_0^\infty y(t)x'(t)\,dt = \int_0^\infty x(t)y'(t)\,dt, \tag{24.41}$$

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 (24.39) 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 24.4.** Substitute $y = t^{\frac{1}{3}}x$ in the equation for the folium to get $x$ and $y$ in terms of $t$ and evaluate (24.41) 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 24.5.** As Exercise 24.4 but use (24.37) to get the boundary parameterised with $0 \le s \le 1$.

In the last exercise you see that the boundary of (24.40) 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 23.2 fails. The same issue occurs in the trivial case of Exercise 27.11.

Note that (24.40) 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

$$x = s^{\frac{1}{p(p-2)}} \, (1-s)^{\frac{p-1}{p(p-2)}} \; ; \; y = s^{\frac{p-1}{p(p-2)}} \, (1-s)^{\frac{1}{p(p-2)}} \, , \tag{24.42}$$

parameterises the loop in the solution set of $x^p + y^p = xy$, with

$$x\frac{dy}{ds} - y\frac{dx}{ds} = \frac{1}{p}\, s^{\frac{1}{p-2}-1}(1-s)^{\frac{1}{p-2}-1}, \tag{24.43}$$

which looks much better than the individual terms $x\frac{dy}{ds}$ and $y\frac{dx}{ds}$. With the $\beta$-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

$$x^p + y^p = xy \tag{24.44}$$

is thus equal to

$$A_p = \frac{1}{2p}\, B\left(\frac{1}{p-2}, \frac{1}{p-2}\right), \tag{24.45}$$

which gives $\frac{1}{6}$ for $p = 3$ and differs from Exercise 24.4 by a factor $3^2$, consistent with (24.39).

Note that in deriving (24.43) from (24.42) 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 $\alpha$ and $\beta$. I also suggest to write derivatives such as

$$\frac{d}{ds}\, s^\alpha (1-s)^\beta = \left(\frac{\alpha}{s} - \frac{\beta}{1-s}\right) s^\alpha (1-s)^\beta = (\alpha - (\alpha+\beta)s)\, s^{\alpha-1}(1-s)^{\beta-1},$$

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 24.12.

[17] More on the $\beta$-function in [HM].

**Exercise 24.6.** 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

$$xdydz = x\left(\frac{\partial y}{\partial s}\frac{\partial z}{\partial t} - \frac{\partial y}{\partial t}\frac{\partial z}{\partial s}\right)$$

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 $\beta$-functions again.

Just in case, I arrived via

$$xyz = \frac{(st)^{\frac{1}{p-3}}}{(1+s+t)^{\frac{3}{p-3}}}$$

and

$$\frac{1}{yz}\left(\frac{\partial y}{\partial s}\frac{\partial z}{\partial t} - \frac{\partial y}{\partial t}\frac{\partial z}{\partial s}\right) = \frac{1}{p^2(p-3)st}\left(\frac{p}{1+s+t} - 1\right)$$

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 $\beta$-functions. It should lead to what we get in Exercise 24.11, which is really nice[20].

[18] Via $s = \frac{t}{1-t}$, a substitution I avoided for (24.6).

[19] Via $t = (1+s)\tau$.

[20] There were mistakes in an earlier version and then it did not, but now it does.

**Exercise 24.7.** 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 24.8.** Verify that computing the area of (24.40) using polar coordinates is even a bigger pain than using the $y = tx$-trick.

**Exercise 24.9.** In Exercise 24.7 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 24.4 from the formula for $x(t)$ in (24.39).

**Exercise 24.10.** Verify (24.45) by putting $y = tx$ in (24.44), solve for $x$, and set $t^p = s$ in the integral you get from Exercise 24.9 and convert to $\beta$-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 24.11.** See Exercise 24.9. 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 24.12.** 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....

# 25 Partitions of compact manifolds and.....

We apply the techniques in Section 27.3 to a non-empty compact set $M \subset \mathrm{IR}^{\mathrm{N}}$ for which (C) in Section 23.2 applies in a sense we make more precise in Section 25.2. In that section we specify blocks $[\tilde{a}, \tilde{b}] \subset \mathrm{IR}^{\mathrm{N}}$ in which the description (A) of Section 23.2 can be given, see (25.8). Below we rather choose blocks $[\tilde{a}_i, \tilde{b}_i] \subset \mathrm{IR}^n$, given $\Phi_i$ as in (24.1).

Thus for each $p \in M$ there exists a continuously differentiable injective

$$\Phi_i : [a, b] = [a_1, b_1] \times \cdots \times [a_n, b_n] \to \mathrm{IR}^{\mathrm{N}}$$

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

$$x \in M \iff x \in \Phi((a, b)) \tag{25.1}$$

We would now like to consider the sets $\Phi((\tilde{a}, \tilde{b}))$ as open sets covering $M$, so that by compactness

$$M \subset \Phi_1((\tilde{a}_1, \tilde{b}_1)) \cup \cdots \cup \Phi_m((\tilde{a}_m, \tilde{b}_m)), \tag{25.2}$$

for some finite collection $\Phi_j$, but clearly the sets $\Phi((\tilde{a}, \tilde{b}))$ are not open[24] in $\mathrm{IR}^{\mathrm{N}}$, 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 (25.1) hold for all $x \in \underline{O}$ as well, and such that $\underline{O}$ does not intersect the (compact) image under $\Phi$ 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 $\Phi$ is injective.

Varying $p \in M$ the open sets $\underline{O}$ cover $M$ and by compactness there exists a finite collection $O_1, \ldots, O_m$ 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 (25.2) 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 (25.1) characterises the elements of $M$. The description following (24.1) in Section 24 with unit blocks then results from Section 27.3.

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 (25.2), but it is then slightly more complicated to formulate (27.11), because each

[24] Of course they should be open in $M$.

$\chi_j$ 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 $\Omega$ and $M = \partial\Omega$ as in Chapter 24 allow a choice of functions $\zeta_1, \ldots, \zeta_n \in C_c^\infty((a_i, b_i))$ with $0 \leq \zeta_i \leq 1$ and $\zeta_1 + \cdots + \zeta_n \equiv 1$ on a neighbourhood of $\Omega$ 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 Chapter 24.

## 25.1 Changing partitions

We still have to check that the integrals do not depend on the choice of the partitioning functions $\zeta_1, \ldots, \zeta_n$. We observe that (24.2) defines a linear map

$$f \xrightarrow{L} \int_M f \, dS_n \tag{25.3}$$

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 $|Lf| \leq C|f|_\infty$, just as in Section 7.4, but we will not be using this below[25].

The partition naturally defines linear subspaces

$$X_i = \{\zeta_i f : f \in C(M)\},$$

and the same holds for any other partition of $M$, given by say $\eta_1, \ldots, \eta_J$, which also defines a linear map

$$f \xrightarrow{K} \int_M f \, dS_n \tag{25.4}$$

via (24.2), 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

$$f = f \sum_{i=1}^{I} \zeta_i = \sum_{i=1}^{I} \zeta_i f = \sum_{i=1}^{I} \zeta_i f \sum_{j=1}^{J} \eta_j = \sum_{i=1}^{I} \sum_{j=1}^{J} \zeta_i \eta_j f, \tag{25.5}$$

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

$$Kf = \sum_{j=1}^{J} \sum_{i=1}^{I} \int_{\Psi_j} \eta_j \zeta_i f \, dS_n,$$

[25] But we will need it to get rid of the annoying assumption $\Phi \in C^2$ in Section 24.3.

and thus it remains to show that

$$\int_{\Phi_i} \zeta_i \eta_j f \, dS_n = \int_{\Psi_j} \eta_j \zeta_i f \, dS_n \tag{25.6}$$

The integral on the left is defined via (22.13) 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 (22.13) 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 25.3. 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[26] before we can apply Section 19 to conclude that the two integrals are the same.

## 25.2 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 23.2, 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

$$[a,b] = [a_1,b_1] \times \cdots \times [a_n,b_n] \times [a_N,b_N] \times \cdots \times [a_N,b_N] \subset \mathrm{I\!R}^{\mathrm{N}}$$

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 23.2 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

$$F^{-1}(p) \cap [\tilde{a},\tilde{b}] = \{(x,f(x)) : x \in [\tilde{a}_x,\tilde{b}_x]\} \subset [\tilde{a}_x,\tilde{b}_x] \times (\tilde{a}_y,\tilde{b}_y), \tag{25.7}$$

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

[26] Try this one by yourself.

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 $\Phi$, 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 23.3 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[27], 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 $\Phi_x$.

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 $\Phi$ is injective on $[a,b]$ then the image under $\Phi$ 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

$$\Phi([a,b]\backslash\phi((\underline{a}_x,\underline{b}_x)))\cap[\tilde{a},\tilde{b}]=\emptyset.$$

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

$$\Phi([a,b]\cap[\tilde{a},\tilde{b}]=\{(x,f(x)):\,x\in[\tilde{a}_x,\tilde{b}_x]\}\subset[\tilde{a}_x,\tilde{b}_x]\times(\tilde{a}_y,\tilde{b}_y). \tag{25.8}$$

The two statements (25.7) and (25.8) should be compared to the definition Edwards gives in Section 4 of his Chapter III for $M\subset\mathbb{R}^{\mathrm{N}}$ 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

[27]See the discussion after Theorem 23.3.

$O \supset [\tilde{a}, \tilde{b}] \supset (\tilde{a}, \tilde{b})$ for some $(\tilde{a}, \tilde{b}) \ni p$, and thus it is completely equivalent to ask that $p \in (\tilde{a}, \tilde{b})$ and

$$p \in M \cap [\tilde{a}, \tilde{b}] = \{(x, f(x)) : x \in [\tilde{a}_x, \tilde{b}_x]\} \subset [\tilde{a}_x, \tilde{b}_x] \times (\tilde{a}_y, \tilde{b}_y) \tag{25.9}$$

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 (25.7,25.8), the patch being

$$M \cap (\tilde{a}, \tilde{b}) = \{(x, f(x)) : x \in (\tilde{a}_x, \tilde{b}_x)\} \ni p = (p_x, f(p_x)). \tag{25.10}$$

In the closed $N$-block $[\tilde{a}, \tilde{b}]$ there are no other points of $M$ than the points on the graph of $f : [\tilde{a}_x, \tilde{b}_x] \to (\tilde{a}_y, \tilde{b}_y)$.

## 25.3 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

$$u \to \Phi(u) = (u, f(u)) \quad \text{and} \quad v \to \Psi(v) = (v, g(v)) \tag{25.11}$$

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 $\mathbb{R}^{\mathrm{N}}$ 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 $\Psi$ in the form $\Phi$ we need the invertibility of respectively

$$\frac{\partial g_2}{\partial v_3} \quad \text{and} \quad \frac{\partial g_1}{\partial v_2}, \tag{25.12}$$

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 $\Psi$-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 (25.11) 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 (25.12) 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[28] of $\Phi(u) = \Psi(v)$ via

$$v_1 = u_1,\ v_3 = f_3(u_1, u_2) \quad \text{and} \quad u_1 = v_1,\ u_2 = g_2(v_1, v_3) \tag{25.13}$$

in the first case, and

$$u_1 = g_1(v_2) \quad \text{and} \quad v_2 = f_2(u_1) \tag{25.14}$$

in the second case. The invertiblity of the partial Jacobian matrices in (25.12) in $p$ follows because otherwise the $\Psi$-patch cannot achieve all respectively $(u_1, u_2)$-directions and $u_1$-directions that occur in the $\Phi$-patch, contradicting the assumption that the $\Psi$-patch covers all of $M$ in its defining neighbourhood.

The restriction to patches of the form (25.10) looks like an obvious choice for simplicity, but may bother us later when dealing with (24.33), we'll see.

[28] All equations but the last one, which then requires some argument to hold as well.

# 26 Functional calculus

This chapter aims to present precisely that part of complex function theory needed for operational calculus. We also play a bit with the definition of line integrals. To be translated and modified.

## 26.1 Lijnintegralen over polygonen en Goursat

We beginnen met formule (12.1) 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 $z, z(t), z_0, z_1, f(z)$,

$$t \to z(t) = (1-t)z_0 + tz_1 \tag{26.1}$$

het interval $[z_0, z_1]$ parametriseert, en

$$\int_0^1 f(z(t))z'(t)dt =$$

$$\int_0^1 f((1-t)z_0 + tz_1)dt\,(z_1 - z_0) = \int_{z_0}^{z_1} f(z)\, dz. \tag{26.2}$$

Dit is een formule die we, zonder dat (12.1) 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 (26.1), 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.

[1]Quod non.

**Exercise 26.1.** 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 $n \in \mathbb{N}$.

**Exercise 26.2.** 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 26.3.** Laat zien dat

$$|\int_{z_0}^{z_1} f(z)\,dz| \le |z_0 - z_1| \max_{z \in [z_0, z_1]} |f(z)|.$$

Integralen gedefinieerd als hierboven door (26.2) kunnen we rijgen tot een integraal over een *polygonaal pad*

$$z_0 \to z_1 \to \cdots \to z_n$$

middels

$$\int_{z_0}^{z_1} f(z)\,dz + \int_{z_1}^{z_2} f(z)\,dz + \cdots + \int_{z_{n-1}}^{z_n} f(z)\,dz = \int_{z_0,\ldots,z_n} f(z)\,dz, \quad (26.3)$$

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 26.4.** Ga dit na. Hint: neem eerst $z_0 = z_2, z_1 \in \mathbb{R}$ om te zien hoe het moet werken.

[2] De eersten zullen de laatsten zijn.

Wat deze opgave zegt is dat heen en weer lopen geen bijdrage aan een keten als in (26.3) 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.

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 $\Delta$ 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.

[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.

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

$$|\int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} f(z)\,dz| \geq \frac{1}{4^k}$$

voor $k = 0, 1, 2, 3, \ldots$.

**Exercise 26.5.** Bewijs dat de rijen $z_0^{(k)}, z_1^{(k)}, z_2^{(k)}$ convergeren naar een limiet in $\Delta_{z_0,z_1,z_2}$ 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

$$\begin{aligned}\int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} f(z)\,dz &= \int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} f'(0)z\,dz + \int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} R(z)\,dz \\ &= \int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} R(z)\,dz,\end{aligned}$$

omdat de eerste integraal nul is vanwege (26.1) 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

$$\frac{1}{4^k} \leq |\int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} R(z)\,dz| \leq \frac{|z_0 - z_1| + |z_1 - z_2| + |z_2 - z_0|}{2^k} \max_{z \in \delta\Delta^{(k)}} |R(z)|,$$

[6] Door $f(z)$ te delen door zijn integraal over de rand van $\Delta_{z_0,z_1,z_2}$.

[7] Schuif de boel anders op.

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 26.6.** Gebruik (de definitie van) $|R(z)| = o(|z|)$ als $|z| \to 0$ om met de laatste twee ongelijkheden een tegenspraak te forceren.

**Theorem 26.7.** *Voor iedere $f : \Delta_{z_0,z_1,z_2} \to \mathbb{C}$ die complex differentieerbaar is op de gesloten*[8] *driehoek $\Delta_{z_0,z_1,z_2}$ 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 (26.3) is gemaakt de zijden zijn van zo'n door

$$z_n = z_0, \ldots, z_{n-1} \tag{26.4}$$

[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.

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

$$\oint_{z_n=z_0,\ldots,z_n} f(z)\,dz = 0. \tag{26.5}$$

Bovendien kan ieder punt $z_k$ dan met de andere punten vastgehouden naar binnen geschoven worden naar een $\tilde{z}_k$ in de door (26.4) gemaakte veelhoek, waarbij (26.5) niet verandert met eenzelfde argument waarin driehoeken met hoekpunten $z_{k-1}, z_k, \tilde{z}_k, z_{k+1}$ voorkomen. Kortom, (26.5) 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 (26.4), ook als de $n$-hoek niet gemaakt is zoals boven (26.4) 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 (26.5) 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 $\zeta_j$ echt in ligt, als die allemaal maar met dezelfde orientatie genomen worden. In dat geval is dus

$$\oint_{z_n=z_0,\ldots,z_n} f(z)\,dz = \sum_{j=1}^{p} \oint_{z_2^{(j)}=z_0^{(j)},z_1^{(j)},z_2^{(j)},} f(z)\,dz, \tag{26.6}$$

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 (26.6) een verrassend simpele vorm krijgen.

[10] De veelhoek is dan convex.

## 26.2 De 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 26.7 toe te passen op integralen van

$$z \to \frac{f(z) - f(\zeta)}{z - \zeta}$$

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 $\zeta$ 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

$$\oint_{z_{0-n}} g(z)\, dz =$$

$$\int_{z_0, z_1, \ldots, z_n = z_0} g(z)\, dz = \int_{w_0, w_1, \ldots, w_n = w_0} g(z)\, dz \tag{26.7}$$

[11] Denk aan deze contouren: http://www.fi.uu.nl/publicaties/literatuur/7244.pdf

[12] Nou ja, rond...

$$= \oint_{w_{0-n}} g(z)\, dz,$$

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 26.8.** Bewijs (26.7) 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.

Nu nemen we voor $g(z)$ het differentiaalquotiënt

$$\frac{f(z) - f(\zeta)}{z - \zeta}$$

en concluderen dat

$$\oint_{w_{0-n}} \frac{f(z) - f(\zeta)}{z - \zeta}\, dz = \oint_{z_{0-n}} \frac{f(z) - f(\zeta)}{z - \zeta}\, dz. \tag{26.8}$$

**Exercise 26.9.** De integraal in het linkerlid van (26.8) hangt van $\rho$ af. Gebruik de differentieerbaarheid van $f$ in $\zeta$ 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 (26.8) is gelijk aan de integraal in het rechterlid en hing niet van $\rho$ af als $\rho < r - |\zeta|$. Hij ging[15] dus niet naar 0 want hij was al 0. Zo reduceert (26.8) tot

$$0 = \oint_{z_{0-n}} \frac{f(z)}{z - \zeta}\, dz - \oint_{z_{0-n}} \frac{f(\zeta)}{z - \zeta}\, dz,$$

en volgt

$$f(\zeta) \oint_{z_{0-n}} \frac{1}{z - \zeta}\, dz = \oint_{z_{0-n}} \frac{f(z)}{z - \zeta}\, dz.$$

[13] Voor grote $n$ is de veelhoek bijna een rondje.

[14] Dat is maar een woord hier, de punten bepalen de richting.

[15] Letterlijk gesproken.

**Exercise 26.10.** 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 26.8 en de tweede integraal hangt niet van $\zeta$ of $\rho$ 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 16.9 die je als het goed is in Opgave 26.2 als laatste hebt gebruikt.

**Theorem 26.11.** *Als $\zeta$ 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

$$f(\zeta) = \frac{1}{2\pi i}\oint_{z_{0-n}} f(z)\left(\frac{1}{z} + \frac{\zeta}{z^2} + \frac{\zeta^2}{z^3} + \frac{\zeta^3}{z^4} + \cdots\right)dz =$$

$$\frac{1}{2\pi i}\oint_{z_{0-n}} \frac{f(z)}{z}\,dz + \frac{1}{2\pi i}\oint_{z_{0-n}} \frac{f(z)}{z^2}\,dz\,\zeta + \frac{1}{2\pi i}\oint_{z_{0-n}} \frac{f(z)}{z^3}\,dz\,\zeta^2 + \cdots,$$

met

$$a_j = \frac{1}{2\pi i}\oint_{z_{0-n}} \frac{f(z)}{z^{j+1}}\,dz \tag{26.9}$$

voor alle $j \in \mathbb{N}_0$.

**Theorem 26.12.** *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 (26.9), 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.*

[16] We prefereren weer de notatie met puntjes natuurlijk.

Het rechtvaardigen van het verwisselen van $\oint$ en $\sum$ is wezen niets anders dan opmerken dat voor elke $\alpha$ en $\beta$ 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.

Voor hier is het nog de vraag of we de limiet $n \to \infty$ willen nemen in Stelling 26.11 en (26.9) teneinde de $\oint$ te nemen over de cirkel geparametriseerd door

$$z = r\exp(i\theta) \quad \text{met} \quad dz = ir\exp(i\theta)d\theta \tag{26.10}$$

und so weiter.

Dat laatste kan komen na de observatie dat voor $0 \le \rho < R$ 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.

$$f(z) = \sum_{j=-\infty}^{\infty} a_j z^j = \sum_{j=0}^{\infty} a_j z^j + \sum_{j=1}^{\infty} \frac{a_{-j}}{z^j}, \tag{26.11}$$

met $a_j$ gegeven door (26.9) voor *alle* $j \in \mathbb{Z}$, maar $n \ge 3$ 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 $\zeta$ 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 (26.10) 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.

[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.

**Exercise 26.13.** Natuurlijk geldt Stelling 26.12 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.

De formule in Stelling 26.11 herschrijven we nu met $1 = I$ en $\zeta = A$ als

$$f(A) = \frac{1}{2\pi i} \oint_{z_{0-n}} f(z)(zI - A)^{-1} \, dz, \tag{26.12}$$

nu voor een willekeurig polygon waar $A$ binnen ligt en waarop

$$z \to (zI - A)^{-1} \tag{26.13}$$

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 (26.12) 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 26.4.

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 (26.12) zomaar overschrijven met $\zeta$ vervangen door een lineaire operator $A : X \to X$. Voor de hand ligt dat $A$ zo moet zijn

[20] Via een zigzagintegraal als in Opgave 26.8 volgt de geldigheid van (26.12).

[21] Een Banachalgebra zelfs, zie Charlie's teleurstelling in Flowers for Algernon.

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.

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 (26.13) op het polygon en zijn buitengebied moet bestaan in de complexe versie van $L(X)$. Lees wat dit betreft verder in Sectie 26.4.

## 26.3 Kromme lijnintegralen

Met behulp van (26.2) is in (26.3)

$$\int_{z_0,\ldots,z_n} f(z)\,dz = \sum_{k=1}^{n} \int_{z_{k-1}}^{z_k} f(z)\,dz \tag{26.14}$$

gedefinieerd voor een rij punten die we (nog) niet als partitie zien, waarvoor we ook (nog) niet Riemanntussensommen als

$$\sum_{k=1}^{n} f(\zeta_k)(z_k - z_{k-1}) \quad \text{met} \quad \zeta_k \in [z_{k-1}, z_k] \tag{26.15}$$

hebben ingevoerd. Maar als de "incrementen" $z_k - z_{k-1}$ klein zijn ligt gezien (26.2) iedere term in (26.15) voor de hand als benadering voor de overeenkomstige term in het rechterlid van (26.14) 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

$$z_k = z_k^{(n)}$$

[22] Denk hier even over na.

zinvol afhankelijk van $n$ gekozen, maar wat is zinvol? Hieronder wat overwegingen en een aanzet tot een uitgewerkt antwoord.

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

$$\int_{P_n} f(z)\,dz = \sum_{k=1}^{n} \int_{z_{k-1}^{(n)}}^{z_k^{(n)}} f(z)\,dz \quad \text{en} \quad \sum_{k=1}^{n} f(\zeta_k^{(n)})(z_k^{(n)} - z_{k-1}^{(n)}) \tag{26.16}$$

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

$$\mu_n = \max_{k=1,\ldots,n} |z_k^{(n)} - z_{k-1}^{(n)}| \to 0 \tag{26.17}$$

en

$$L_n = \sum_{k=1}^{n} |z_k^{(n)} - z_{k-1}^{(n)}| \quad \text{begrensd} \tag{26.18}$$

is als $n \to \infty$ voldoende om het verschil tussen de termen in (26.16) 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 (26.17) en (26.18) gelden, en

$$z_0^{(n)} = a \quad \text{en} \quad z_n^{(n)} = b \tag{26.19}$$

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

$$|z(t_1) - z(t_2)| \leq |t_1 - t_2| L, \tag{26.20}$$

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 (26.1) 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 $z \to f(z)$.

## 26.4 Calculus in Banachalgebras van operatoren

Deze sectie is nog wat schetsmatig maar niettemin precies. We willen (26.12) uitwerken voor $A \in L(X)$ en schrijven met $z$ vervangen door $\lambda$

$$f(A) = \frac{1}{2\pi i} \oint_P f(\lambda)(\lambda - A)^{-1}\, d\lambda, \tag{26.21}$$

nu voor een willekeurig polygon[25] met hoekpunten $\lambda_1, \dots, \lambda_n = \lambda_0$, waarop en waarbuiten[26]

$$\lambda \to (\lambda - A)^{-1} = (\lambda I - A)^{-1} \tag{26.22}$$

gedefinieerd is. Het complement van het domein van (26.22) 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 (26.22) heet de resolvente van $A$.

**Exercise 26.14.** 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 (26.22) complex differentieerbaar is op $\rho(A)$. Wat is de afgeleide?

**Exercise 26.15.** 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 (26.22) 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 26.13.

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

[25] Of een vereniging daarvan.
[26] Wat bedoelen we daarmee?

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 $2 \times 2$ matrix, met complexe of reële entries. In dat geval bestaat $\sigma(A)$ meestal uit 2 punten, en met twee disjuncte driehoekjes $\Delta_1$ en $\Delta_2$ om die punten heen kunnen we al aan de slag met ieder paar complex differentieerbare functies

$$f_1 : \Delta_1 \to \mathbb{C} \quad \text{en} \quad f_1 : \Delta_1 \to \mathbb{C}$$

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 (13.8) 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

$$\sigma(A) \subset K_{int} \subset K = C_0 \cap U_1 \cap \cdots \cap U_n, \tag{26.23}$$

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

$$f(A) = \frac{1}{2\pi i} \oint_{\delta K} f(\lambda)(\lambda - A)^{-1}\, d\lambda =$$

$$\frac{1}{2\pi i} \left( \oint_{P_0} f(\lambda)(\lambda - A)^{-1}\, d\lambda - \sum_{j=1}^{n} \oint_{P_j} f(\lambda)(\lambda - A)^{-1}\, d\lambda \right) \tag{26.24}$$

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

$$f(A) = \frac{1}{2\pi i} \sum_{j=1}^{m} \oint_{\delta K_j} f(\lambda)(\lambda - A)^{-1}\, d\lambda \tag{26.25}$$

met iedere term in de som gedefinieerd als in (26.24) 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 $\Delta_j$ 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 (26.24) en (26.25) 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.

[27]Als we alle zijden van alle poygonen dicht bij $\sigma(A)$ willen hebben.

**Exercise 26.16.** 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 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 (26.25) niet, mits de Stelling van Goursat geldt voor driehoekjes waarop en waarbinnen (26.22) 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 (26.24) 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 $\mu_l$ een klein beetje naar binnen geschoven hebben, niet omdat het moet, maar omdat het kan, iets

[28] Op detail nog te bespreken.

minder ver naar binnen dan de hoekpunten $\lambda_k$. Het $\mu$-polygon komt zo binnen het $\lambda$-polygon te liggen.

Omdat

$$f(A) = \frac{1}{2\pi i} \sum_{k=1}^{n} \int_{\lambda_{k-1}}^{\lambda_k} f(\lambda)(\lambda - A)^{-1}\, d\lambda,$$

$$g(A) = \frac{1}{2\pi i} \sum_{l=1}^{n} \int_{\mu_{l-1}}^{\mu_l} g(\mu)(\mu - A)^{-1}\, d\mu,$$

wordt $f(A)g(A)$ afgezien van de voorfactoren dankzij overwegingen als bij (27.4) 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

$$(fg)(A) = \frac{1}{2\pi i}\oint_{\lambda_{0-n}} f(\lambda)g(\lambda)(\lambda - A)^{-1}\,d\lambda = f(A)g(A) = g(A)f(A), \tag{26.26}$$

en daar is nog veel mee te spelen.

**Exercise 26.17.** Ga na dat in het algemene geval (26.25), 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 (26.26) 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 (26.24) en (26.25) 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 26.18.** *Laat voor $A \in L(X)$ en een complexwaardige $f$ de operator $f(A)$ gedefinieerd zijn via (26.24) en (26.25). 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 (26.25).

[29] Lees: linksom, maar met een min voor het integraalteken.

De driehoekjes in $I_0$ zijn niet relevant voor (26.25), 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 $\Delta$ 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

$$\Delta \cup \tilde{\Delta},$$

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 (26.25).

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 $\Delta$ 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 (26.23). 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 $\rho$ klein te kiezen. Hoe dichter bij $\sigma(A)$ hoe meer je er nodig hebt en hoe wilder de standjes kunnen worden.

[30]Waarvan er drie een zijde met $\Delta$ gemeen hebben en de rest alleen een hoekpunt.

[31]Ik pas, maar dat is voor even.

We zijn nu klaar voor het bewijs van Stelling 26.18. 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

$$g(A)(\mu - f(A) = (\mu - f(A)g(A) = I,$$

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 $\lambda$ kunnen we zonder beperking der algemeenheid wel aannemen dat $\lambda_0 = 0 = \mu_0$ en dat de machtreeks van $f$ begint met $\lambda^n$ 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 26.19.** 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 26.20.** Laat zien dat $P_i P_j = \delta_{ij} P_j$, $AP_j = P_j A$, $\sigma(AP_j) = \sigma(A) \cap K_j$,

$$I = \sum_{j=1}^{N} P_j \quad \text{en} \quad A = \sum_{j=1}^{N} AP_j = \sum_{j=1}^{N} P_j A.$$

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.

# 27 Elementary multi-variate integral calculus

We want to integrate continuous functions and partial derivatives of continuously differentiable functions over bounded sufficiently nice domains $\Omega$. The goal is an early version of Green's Theorem (and thereby the Gauss Divergence Theorem), Theorem 27.7 in Section 27.5. 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 also return to Section 13.2 and the issue of differentiation under the integral.

We first integrate over closed rectangles $[a, b] \times [c, d]$, and over sets such as

$$\{(x, y) \in [a, b] \times [c, d] : y \leq f(x)\}, \quad \text{in which} \quad f \in C^1([a, b]) \tag{27.1}$$

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 $\mathbb{R}^{\mathrm{N}} = \mathbb{R}^{n+1}$ is then completely similar, as well as integral calculus over sets described by

$$a_N \leq x_N \leq f(x_1, \ldots, x_{N-1}) < b_N \tag{27.2}$$

or

$$a_N \leq f(x_1, \ldots, x_{N-1}) \leq x_N < b_N, \tag{27.3}$$

and similar sets obtained by permutation of the variables.

To also integrate over closures of bounded open sets $\Omega$ in $\mathbb{R}^{\mathrm{N}} = \mathbb{R}^{n+1}$ 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 (27.2) or (27.3). If so we say that $\Omega$ 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 24.

[1] Thus windows are open.

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,

$$M \subset P_1 \cup \cdots \cup P_k, \tag{27.4}$$

but in general $\bar{\Omega}$ is not a subset of $W_1 \cup \cdots \cup W_k$. However, there are then[3] 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 $\Omega$ not yet covered by $W_1, \ldots, W_k$. Thus

$$\bar{\Omega} \subset W_1 \cup \cdots \cup W_m. \tag{27.5}$$

This covering of $\Omega$ will allow us to integrate continuous functions $u : \bar{\Omega} \to \mathrm{I\!R}$ over $\bar{\Omega}$, using what we may call *fading*[4] *functions*. To do so we choose closed blocks $K_i \subset W_i$ such that

$$\bar{\Omega} \subset K_1 \cup \cdots \cup K_m.$$

## 27.1 Integrals over blocks

To integrate continuous functions

$$u : [a, b] \times [c, d] \to \mathrm{I\!R}$$

we use partitions $P$ as in (6.8) for $[a, b]$, and partitions

$$c = y_0 \leq y_1 \leq \cdots \leq y_M = b \tag{27.6}$$

for $[c, d]$. Lower- and undersums, or better, sums of the form[5]

$$S = \sum_{k=1}^{N} \sum_{l=1}^{M} u(\xi_k, \eta_l)(x_k - x_{k-1})(y_l - y_{l-1}), \tag{27.7}$$

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.

[2]This follows from the compactness of $M$.
[3]This follows from the compactness of $\bar{\Omega}$.
[4]The less friendly terms is cut-off functions.
[5]See Theorem 8.15.

**Theorem 27.1.** *Let $u : [a,b] \times [c,d] \to \mathrm{I\!R}$ 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 (27.7) 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 integral of $f$ over $[a,b] \times [c,d]$ by*

$$\int_{[a,b]\times[c,d]} u = J,$$

*and we have*

$$J = \int_a^b \underbrace{\int_c^d u(x,y)\,dy}_{\textit{continous function of } x} \quad dx = \int_c^d \underbrace{\int_a^b u(x,y)\,dx}_{\textit{continous function of } y} \quad dy,$$

*with*

$$|J| = \left|\int_{[a,b]\times[c,d]} u\right| \leq \int_{[a,b]\times[c,d]} |u|.$$

The repeated integrals are handled by the integration techniques for continuous functions on closed bounded intervals, see Theorems 8.6 and 8.15. Theorem 27.1 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 $\mathrm{I\!R}^N$, and $X$ a complete metric vector space.

**Exercise 27.2.** This is more or less Exercises 8.16 continued. Prove Theorem 27.1 without using lower- and upper sums for $X = \mathrm{I\!R}$. Then explain why it is also a proof for $X$.

## 27.2 Differentation under the integral

We use Theorem 10.10, which is part of the fundamental theorem of calculus, for a simple proof of a theorem that was already stated[6] in Section 13.2:

[6] And in Section 28.5 we give a version needed for integrals over the whole real line.

**Theorem 27.3.** *Let $f$ and $f_t$ exist as continuous functions on $I \times [a,b]$, with $I$ some $t$-interval. Then*

$$J(t) = \int_a^b f(t,x)\,dx$$

*exists and $J : I \to \mathbb{R}$ is continuously differentiable with derivative*

$$J'(t) = \int_a^b f_t(t,x)\,dx.$$

**Proof**. Let $g(t,x) = f_t(t,x)$, the continuous partial derivative of $f$. In Exercise 13.6 we already observed that

$$t \to j(t) = \int_a^b g(t,x)\,dx$$

is continuous on $I$ if $(t,x) \to g(t,x)$ is continuous on $I \times [a,b]$. This is a consequence of the uniform continuity of $g$ on blocks of the form $[a,b] \times [\alpha,\beta]$ with $[\alpha,\beta] \subset I$. It gives

$$|j(t) - j(s)| = \Big| \int_a^b g(t,x)\,dx - \int_a^b g(s,x)\,dx \Big| = \Big| \int_a^b (g(t,x) - g(s,x))\,dx \Big|$$

$$\leq \int_a^b |g(t,x) - g(s,x)|\,dx < \varepsilon(b-a) \quad \text{if} \quad |t-s| < \delta$$

for $t, s \in [\alpha,\beta]$, with $\delta$ provided by the uniform continuity of $g$ on $[a,b] \times [\alpha,\beta]$. And then

$$J(s) - J(\alpha) = \int_a^b f(s,x)\,dx - \int_a^b f(\alpha,x)\,dx = \int_a^b (f(s,x) - f(\alpha,x))\,dx$$

$$= \int_a^b \int_\alpha^s g(t,x)\,dt\,dx = \int_\alpha^s \int_a^b g(t,x)\,dx\,dt = \int_\alpha^s j(t)\,dt$$

in which we have used the fundamental theorem of calculus for every $x$ fixed with $g = f_t$, and Theorem 27.1 to change the order of integration. But now we use the fundamental theorem of calculus again to conclude that $J$ is a primitive of the continuous function $j$ and thereby differentiable with derivative $f$. $\square$

## 27.3 Cut-off functions and partitions of unity

Green's Theorem 27.7 in Section 27.4 requires the introduction of integrals over smooth bounded domains. To define these integrals we use partitions of unity. We now first explain this basic tool for cutting up functions in smaller parts which are localised. It 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 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 compactness of $K$ implies[7] 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 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[8]

$$1 - \chi(x) = (1 - \eta_1(x)) \cdots (1 - \eta_k(x)). \tag{27.8}$$

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 (27.8) 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 (27.8) 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

$$\chi \in C_c^\infty(O), \quad \forall x \in \Omega \quad \chi(x) \in [0, 1], \quad \forall x \in K \quad \chi(x) = 1, \tag{27.9}$$

and this is why $\chi$ is called a cut-off function for $K$ in $O$.

The second trick applies the first trick to a finite collection of such sets

$$\emptyset \neq K_1 \subset O_1, \ldots, \emptyset \neq K_m \subset O_m \quad \text{with} \quad \eta_j \in C_c^\infty(O_j) \tag{27.10}$$

[7] See Section 5.6.

[8] I first saw this elegant trick in Folland's Real Analysis book.

cut-off functions as in (27.9). We define $\zeta_j \in C_c^\infty(O_j)$ by

$$\zeta_j(x) = \frac{\chi_j(x)}{\chi_1(x) + \cdots + \chi_m(x)} \tag{27.11}$$

and extend $\zeta_j$ to $\mathbb{R}^{\mathrm{N}}$ via $\zeta_j(x) \equiv 0$ outside $O_i$. Note that below we don't really use the last part of (27.9) 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

$$\zeta_1(x) + \cdots + \zeta_m(x) = 1. \tag{27.12}$$

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 $\zeta_j$ not harming any smoothness the original function $f$ may enjoy. Adding more $K_j$ to the collection changes the functions $\zeta_j$ only via (27.11), with (27.12) remaining valid. This allows to have countable families of $K_i$ with nice properties. The first application however is to (27.5), when we choose closed blocks $K_i \subset W_i$ such that

$$\bar{\Omega} \subset K_1 \cup \cdots \cup K_m, \tag{27.13}$$

and open blocks $O_i \subset W_i$ such that

$$K_i \subset O_i \subset \bar{O}_i \subset W_i.$$

## 27.4 Integrals over bounded smooth domains

Next we consider windows such as used in (27.4), for example (27.1). The following theorem ties the obvious outcome to a definition via a partition of unity. Here we only need continuity of the function $f$ that describes the upper boundary.

**Theorem 27.4.** *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*

$$\int_a^b \underbrace{\int_c^{f(x)} u(x,y)\,dy}_{\textit{continuous function of } x} \, dx$$

*is equal to*

$$J = \int_A u.$$

*This integral $J$ is uniquely defined as in Theorem 27.1, with approximating sums (27.7) 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 (27.5). It is then possible to choose[9] 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

$$\zeta_1 + \cdots + \zeta_m \equiv 1 \quad \text{on a neighbourhood of } \bar{\Omega}. \tag{27.14}$$

We use each $\zeta_i$ 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[10]

$$\int_\Omega u_i = \int_{\bar{\Omega}\cap\overline{W}_i} u_i.$$

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 $\Omega$ is

$$\int_\Omega u = \int_\Omega u_1 + \cdots + \int_\Omega u_m = \sum_{i=1}^m \int_{\bar{\Omega}\cap\overline{W}_i} \zeta_i u. \tag{27.15}$$

[9]Use Section 27.3, note that $\zeta_i \in C_c^\infty(W_i)$.

[10]In accordance with $\int_a^b = \int_{[a,b]} = \int_{(a,b)}$ we just put $\Omega$ as a subscript on $\int$.

**Exercise 27.5.** Show the outcome in (27.15) 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 $\chi_j$, $j = 1, \ldots, r$, 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.

**Remark 27.6.** *It is also possible to give such a definition if we only assume $\partial\Omega \in C$, meaning that, possibly*[11] *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.*

## 27.5 Green's Theorem

We now *integrate partial derivatives to discover a theorem*, and in particular the right hand side in (27.17) below. It involves the outwards pointing unit normal vector $\nu$ on $\partial\Omega$, as we will see from the local calculations we do in the separate boundary windows. In particular we discover[12]

$$dS_n = \sqrt{1 + \left(\frac{\partial f}{\partial x_1}\right)^2 + \cdots + \left(\frac{\partial f}{\partial x_n}\right)^2}\; dx_1 \cdots dx_n \tag{27.16}$$

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 $f \in C^1([a, b])$.

[11] This makes a bit more technical.
[12] More on this later: Chapter 22.

**Theorem 27.7.** *Let $\Omega$ be a bounded open set in $\mathrm{I\!R}^N = \mathrm{I\!R}^{n+1}$ with $\partial\Omega \in C^1$, let $v : \bar{\Omega} \to \mathrm{I\!R}$ be continuously differentiable. Then the integral of every partial derivative $v_{x_j}$ of $v$ evaluates as*

$$\int_\Omega v_{x_j} = \int_{\partial\Omega} \nu_j v \, dS_n. \tag{27.17}$$

*The integral on the right hand side will be defined in the proof, as well as $\nu_j$, the $j^{th}$ component of the outwards pointing unit normal vector $\nu$ 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

$$\tilde{\Omega} = \Omega \cap ([a,b] \times [c,d]) = \{(x,y) : \, a \leq x \leq b, \, f(x) < y \leq d\}, \tag{27.18}$$

and multiply $v$ by a function $\zeta \in C^1(\mathrm{I\!R}^2)$ 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 27.4 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 $\Phi$ indicates that we use the *parameterisation*

$$\Phi(x) = (x, f(x))$$

for the boundary integral. There are of course many other[13] 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

$$\nu = \frac{1}{\sqrt{1 + f'(x)^2}} \begin{pmatrix} -f'(x) \\ 1 \end{pmatrix}$$

[13] With issues for later worries.

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 $\xi, \eta$ 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

$$\phi(\xi,\eta) = \tilde{v}(x,y) \quad \text{we have} \quad \tilde{v}_x(x,y) = \phi_\xi(\xi,\eta) - f'(\xi)\phi_\eta(\xi,\eta)$$

via the chain rule, whence[14]

$$\int_{\tilde{\Omega}} \tilde{v}_x = \int_D (\phi_\xi - f'\phi_\eta) = \int_D \phi_\xi - \int_D f'\phi_\eta$$

$$= \int_0 \left(\int_a^b \phi_\xi(\xi,\eta)\,d\xi\right) d\eta - \int_a^b \left(\int_0 f'(\xi)\phi_\eta(\xi,\eta)\,d\eta\right) d\xi$$

$$= \int_0 (\phi(b,\eta) - \phi(a,\eta))\,d\eta - \int_a^b f'(\xi)\int_0 \phi_\eta(\xi,\eta)\,d\eta\,d\xi$$

$$= \int_a^b f'(\xi)\phi(\xi,0)\,d\xi = \int_a^b v(x,f(x))f'(x)\,dx$$

$$= \int_a^b \underbrace{\frac{f'(x)}{\sqrt{1+f'(x)^2}}}_{\nu_x}\,\tilde{v}(x,f(x))\,\underbrace{\sqrt{1+f'(x)^2}dx}_{ds=dS_1} = \int_\Phi \nu_x\tilde{v}\,ds,$$

after inserting $\sqrt{1+f'(x)^2}$ to get $ds = dS_1$ and recognising the $x$-component of the normal vector $\nu$. The subscript $\Phi$ 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 $\Phi$ as definition of the boundary integrals over $\partial\Omega$.

[14] We drop a conveniently chosen fixed upper bound in the $\eta$-integrals from the notation.

Likewise we have, for the general case with $n \geq 1$, that

$$\int_{\Omega} (\zeta_i v)_{x_j} = \int_{\partial\Omega} \nu_j \zeta_i v \, dS_n, \tag{27.19}$$

for all $j = 1, \ldots, m$ and all $i = 1, \ldots, N = n+1$, with expressions like (27.16) and

$$\nu_1 = \frac{1}{\sqrt{1 + |\nabla f|^2}} \frac{\partial f}{\partial x_1}, \ldots, \nu_N = \frac{1}{\sqrt{1 + |\nabla f|^2}} \frac{\partial f}{\partial x_n},$$

$$\nu_N = \nu_{n+1} = \frac{-1}{\sqrt{1 + |\nabla f|^2}}$$

for the normal unit vector $\nu$. Note that in (27.19) the integrals with $i = k+1, \ldots, m$ all vanish.

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 (27.15). We can then conclude that

$$\int_{\Omega} v_{x_j} = \int_{\partial\Omega} \nu_j v \, dS_n,$$

which is (27.17) in Theorem 27.7.

**Remark 27.8.** *If we put a subscript $j$ on $v$, and view $v_j$ as the coordinates of a vector field $V$, we obtain*

$$\int_{\Omega} \nabla \cdot V = \int_{\partial\Omega} \nu \cdot V \, dS_n, \tag{27.20}$$

*the statement of the Gauss Divergence Theorem.*

**Remark 27.9.** *Applying (27.17) to the product of $v$ and some other function $\zeta \in C^1(\Omega)$ we obtain the integration by parts formula*

$$\int_{\Omega} v_{x_i} \zeta = \int_{\partial\Omega} \zeta v \, \nu_i dS_{N-1} - \int_{\Omega} \zeta_{x_i} v. \tag{27.21}$$

*For $\zeta$ we may take a function such as one of the $\zeta_i$ in (27.14) to have integrals of functions supported in one single block $[a, b]$.*

**Remark 27.10.** *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 24. Of course we need these later too, which requires Chapters 19 and 23.*

**Exercise 27.11.** In physics results like (27.7) 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[15]. Verify directly that (27.20) holds for $v : [a, b] \to \mathbb{R}^2$ continuously differentiable.

**Exercise 27.12.** 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 $\Omega$. Show that

$$\iint_\Omega \left(\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y}\right) dxdy = 0.$$

## 27.6 Some integral equations in two variables

Have a look at Section 7.5 before you read on. The integral equations in this section relate to *partial differential equations* (PDE's).

**Exercise 27.13.** Let $F : \mathbb{R} \to \mathbb{R}$ be continuous and suppose that $u \in C^2(\mathbb{R}^2)$ is a solution of

$$u_{xy} = F(u) \tag{27.22}$$

with $u = 0$ of the both the axes. Show that

$$u(x, y) = \underbrace{\int_0^x \int_0^y F(u(\xi, \eta))\, d\eta\, d\xi}_{\Phi(u)(x,y)} \quad (x, y \in \mathbb{R}). \tag{27.23}$$

This is like a two variable version of (7.17) with $u_0 = 0$ which we solved in Exercise 7.32 using weighted norms.

[15] https://www.quora.com/What-is-the-plural-of-axis

**Exercise 27.14.** For $F : \mathbb{R} \to \mathbb{R}$ Lipschitz continuous we solve (27.23) 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 $\Phi$ is a contraction. Use this to show that (27.23) has a unique solution in $C(\mathbb{R}^2)$.

**Remark 27.15.** *Did we solve (27.22)? The solution of (27.23) does have some differentiability properties, but it is not so clear whether it is in $C^2(\mathbb{R}^2)$. Note that (27.22) is the nonlinear one-dimensional wave equation*

$$v_{tt} = v_{xx} + G(v) \tag{27.24}$$

*in disguise. For the linear inhomogeneous wave equation*

$$v_{tt} = v_{xx} + F(t,x) \tag{27.25}$$

*there exists the d'Alembert solution formula*

$$v(t,x) = \frac{1}{2}(f(x-t) + f(x+t)) + \frac{1}{2}\int_{x-t}^{x+t} g + \frac{1}{2}\iint_{C(t,x)} F \tag{27.26}$$

*for the solution with initial data*

$$v(0,x) = f(x), \quad v_t(0,x) = g(x) \quad (x \in \mathbb{R}). \tag{27.27}$$

*In (27.26)*

$$C(t,x) = \{(\tau,\xi) : 0 \le \tau \le t, x + \tau - t \le \xi \le x + t - \tau\}$$

*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 (27.26) depends on the smoothness of $f, g, F$.*

**Exercise 27.16.** Consider the integral equation

$$v(t,x) = \frac{1}{2}(f(x-t) + f(x+t)) + \frac{1}{2}\int_{x-t}^{x+t} g + \frac{1}{2}\iint_{C(t,x)} G(v), \tag{27.28}$$

which would correspond to the solution of (27.24) with initial data given by (27.27). 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 (27.28) has a unique solution in $C_T$ for every $T$ with $LT^2 < 1$.

**Exercise 27.17.** (continued) Modify the argument in the spirit of Exercise 27.13 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 (27.28) has a unique continuous solution $v : \mathbb{R} \times [0, \infty) \to \mathbb{R}$ which is in every $C_T$.

**Exercise 27.18.** Let $F : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous. Rewrite

$$u_{xyz} = F(u) \tag{27.29}$$

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 $\mathbb{R}^n$.

# 28 Fourier theory

Consider 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 $2\pi$. On the interval $(-\pi, \pi)$ the graph[1] 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

$$x = 2\sum_{k=1}^{\infty}(-1)^{k+1}\frac{\sin kx}{k} \tag{28.1}$$

if $|x| < \pi$. Section 28.2 cuts that long Hilbert space story short in order to quickly procede to Fourier integrals in Section 28.4. The novelty in this chapter is Section 28.3 which prepares for Fourier theory in more variables along the same lines. The two exercises below provide a link with power series, but you may want to jump to (28.2) in Section 28.1, the *sawtooth*.

**Exercise 28.1.** Connection with power series: The right hand side of (28.1) 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 $\zeta$, take the sum and then the primitive.

**Exercise 28.2.** The complex version of the Leibniz criterion says that the series in Exercise 28.1 converges for all $\zeta$ with $|\zeta| = 1$ except $\zeta = -1$. Assume that the sum you found in Exercise 28.1 is valid for all such $\zeta$. Verify (28.1).

## 28.1 The sawtooth function

In the spirit of Exercise 28.2 consider

$$1 + \zeta + \zeta^2 + \cdots + \zeta^{N-1} \quad \text{and its primitive} \quad \zeta + \frac{1}{2}\zeta^2 + \cdots + \frac{1}{N}\zeta^N,$$

[1]Use some package.

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

$$Z_N(x) = \sum_{n=1}^{N} \frac{2}{n} \sin nx, \tag{28.2}$$

which looks a bit like the example that lead to (28.1), but is slightly better as a first example for our purposes in view of its direct connection to the Dirichlet kernel $D_N$ below.

**Exercise 28.3.** What you see is what you get. Plot some graphs of $Z_N$ for small and large values of $N$, to examine how $Z_N$ converges to a limit function called the sawtooth. The remaining exercises in this section lead you through a nice proof of what you see, including the overshoot behaviour near $x = 0$, but you may want to jump to Remark 28.10 which wraps it all up.

**Exercise 28.4.** Use $e^{ix} = \cos x + i \sin x$ to show that

$$Z_N(x) = x - \int_0^x D_N(s)ds = \pi - x + \int_x^\pi D_N(s)ds,$$

in which[2]

$$D_N(s) = \frac{\sin(N + \frac{1}{2})s}{\sin \frac{s}{2}} = \sum_{n=-N}^{N} e^{inx}.$$

**Exercise 28.5.** (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 $x \in \mathbb{R}$.

[2]See (28.11), this is the Dirichlet kernel.

**Exercise 28.6.** (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
$$M_N = \int_0^{\frac{\pi}{N+\frac{1}{2}}} \frac{\sin(N + \frac{1}{2})s}{\sin \frac{s}{2}} ds = 2 \int_0^{\pi} \frac{\sin t}{t} \frac{\frac{t}{2N+1}}{\sin \frac{t}{2N+1}} dt.$$

**Exercise 28.7.** (continued) Show that
$$\frac{\frac{t}{2N+1}}{\sin \frac{t}{2N+1}} \to 1,$$
uniformly on $t \in [0, \pi]$.

**Exercise 28.8.** (continued) Show that
$$M_N \to 2 \int_0^{\pi} \frac{\sin t}{t} dt$$
as $N \to \infty$.

**Exercise 28.9.** (continued) You must have seen[3] 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 28.10.** *Conclusion: the function sequence $Z_N$ converges pointwise to the sawtooth function $Z$ which is defined by being $2\pi$-periodic, odd[4], and $Z(x) = \pi - x$ for $x$ between $0$ and[5] $\pi$, but its maxima and minima near $0$*

[3] Computed using the complex function $\frac{e^{iz}}{z}$.
[4] So odd, draw a sawtooth picture of its graph.
[5] And thus also for $0 < x < 2\pi$.

*and multiples of* $2\pi$ *over- and undershoot the values* $Z(0^{\pm}) = \pm\pi$ *by a factor of about* 1.178979744*.*

## 28.2 Fourier series

We consider complex valued $2\pi$-periodic continuous functions. The (complex) vector space of all such functions is denoted by $C_{2\pi}$. Piecewise continuous functions as usually considered in this context are de facto functions of the form

$$g(x) = f(x) + \sum_{k=1}^{m} A_k Z(x - \xi_k) \quad \text{with} \quad f \in C_{2\pi} \quad \text{and} \quad A_k \in \mathbb{C},\, \xi_k \in (0, \pi),$$

and since we already understand $Z$ and its Fourier series we may just as well restrict our attention to $f \in C_{2\pi}$.

**Theorem 28.11.** *The space* $C_{2\pi}$ *of all* $2\pi$*-periodic continuous* $f : \mathbb{R} \to \mathbb{C}$ *is a complete metric (complex vector) space with respect to the metric*[6]

$$d(f, g) = \max_{x \in \mathbb{R}} |f(x) - g(x)|.$$

For $f \in C_{2\pi}$ we consider the *Fourier series* of $f$, namely the right hand side of the $\sim$ symbol in

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{inx} = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx), \tag{28.3}$$

in which

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx \, dx, \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx \, dx, \tag{28.4}$$

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} \, dx \tag{28.5}$$

are the *Fourier coefficients* of $f$. In (28.3) we use the symbol $\sim$ because it is hard to say in which sense the left and the right hand side are equal to one another. We sometimes write

$$f(x) \sim \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{inx}, \quad \hat{f}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} \, dx, \tag{28.6}$$

[6] Just like in Section 4.2.

and we choose *not to modify* this notation.

The formulas for the coefficients contain integrals of complex valued functions. At this point we may refer to Theorem 8.15 and the special case that $X = \mathbb{C}$, but it is more effective to write $f(x) = u(x) + iv(x)$, $u, v$ real valued and define

$$\underbrace{\int_{-\pi}^{\pi} f}_{F} = \underbrace{\int_{-\pi}^{\pi} u}_{U} + i \underbrace{\int_{-\pi}^{\pi} v}_{V}, \tag{28.7}$$

in which we momentarily use capitals for the values of the integrals. It is easily checked that

$$\int_{-\pi}^{\pi} (f+g) = \int_{-\pi}^{\pi} f + \int_{-\pi}^{\pi} g \quad \text{and} \quad \int_{-\pi}^{\pi} \lambda f = \lambda \int_{-\pi}^{\pi} f$$

if the $f$- and $g$-integrals exist and $\lambda = \alpha + i\beta \in \mathbb{C}$. In particular we can choose $\lambda \in \mathbb{C}$ with $|\lambda| = 1$ such that

$$\int_{-\pi}^{\pi} \lambda f = \lambda \int_{-\pi}^{\pi} f = \lambda F = (\alpha + i\beta)(U + iV) \in \mathbb{R}$$

and therefore write

$$\lambda F = (\alpha + i\beta)(U + iV) = \int_{-\pi}^{\pi} (\alpha u - \beta v).$$

Then

$$\Big| \int_{-\pi}^{\pi} f \Big| = \Big| \int_{-\pi}^{\pi} (\alpha u - \beta v) \Big| \leq \int_{-\pi}^{\pi} |\alpha u - \beta v| \leq \int_{-\pi}^{\pi} |\lambda f| = \int_{-\pi}^{\pi} |f| \tag{28.8}$$

establishes the familiar triangle inequality also for complex valued integrals. Of course you can also use $n = 2$ in Exercise 8.17 to come to the same conclusion.

From here on we only use (28.5), which may be derived from considerations involving complex version of the $L^2$-inner product, but in what follows we choose to take (28.5) for granted. Thus we forget about the Hilbert space perspective and see what we can say about the $N$-th partial sum

$$S_N f(x) = \sum_{n=-N}^{N} c_n e^{inx} = \frac{a_0}{2} + \sum_{n=1}^{N} (a_n \cos nx + b_n \sin nx) \tag{28.9}$$

of the Fourier series of $f$ in (28.3). A calculation[7] with complex exponential geometric series then first tells us that

$$S_N f(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} D_N(y) f(x-y)\, dy = \frac{1}{2\pi} \int_{-\pi}^{\pi} D_N(y) f(x+y)\, dy, \tag{28.10}$$

[7]You did it in Exercise 28.4, we'll do it below for Fourier series of functions $f(x, y)$.

in which

$$D_N(x) = \frac{\sin(N+\frac{1}{2})x}{\sin\frac{1}{2}x} = \sum_{k=-N}^{N} e^{ikx} \tag{28.11}$$

is called the Dirichlet kernel. We say that $2\pi\, S_N f$ is the convolution of $D_N$ and $f$ because of the first equality in (28.9), notation

$$2\pi\, S_N f = D_N * f.$$

The $2\pi$-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 $2\pi$. 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 allow $S_N f(x)$ converge to $f(x)$ if $f$ is nicer than just being in the space $C_{2\pi}$.

The average of $D_0, \ldots, D_N$ however, which via another miraculous calculation with complex exponentials is equal to

$$F_N(x) = \frac{1}{N+1}\left(D_0(x) + \cdots + D_N(x)\right) = \frac{1}{N+1}\frac{\sin^2\frac{(N+1)x}{2}}{\sin^2\frac{x}{2}}, \tag{28.12}$$

the $2\pi$-periodic Féjer kernel, is much nicer. It is nonnegative, has integral $2\pi$, and concentrates in 0 as $N$ gets large, thereby forcing it to be small away from multiples of $2\pi$. Such functions are called *good kernels*. The following not so very hard theorem explains why.

**Theorem 28.12.** *Define*

$$\sigma_N f = \frac{1}{N+1}(S_0 f + S_1 f + \cdots + S_N f) = \frac{1}{N+1}\sum_{n=0}^{N} S_n(f),$$

*the so-called Cesàro sums of $S_n f$. Then*

$$\sigma_N f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_N(\xi) f(x-\xi)d\xi, \quad \textit{i.e.} \quad 2\pi\sigma_N f = F_N * f, \tag{28.13}$$

*and $\sigma_N f \to f$ in $C_{2\pi}$ if $f \in C_{2\pi}$. That is, the convergence is uniform.*

**Exercise 28.13.** Let $f \in C_{2\pi}$, let $M = |f|_{\mathsf{max}}$ 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 28.12.

**Exercise 28.14.** Let $\xi \in (0, \pi)$. Let $Z(x)$ as in Exercise 28.5 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 $N \to \infty$.

## 28.3 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 (28.11) and (28.12) in (28.14) and (28.16) below. To see how let $f$ be in $C_{2\pi}(\mathrm{I\!R}^2)$, i.e. $f(x, y)$ is continuous in $(x, y)$, and $2\pi$-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 $\xi$ and $\eta$ 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

$$S_{MN} f(x,y) = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} f(x+\xi, y+\eta) D_M(\xi) D_N(\eta)\, d\xi\, d\eta, \tag{28.14}$$

in which

$$D_M(\xi) = \sum_{m=-M}^{M} e^{im\xi} = \frac{e^{-iM\xi} - e^{i(M+1)\xi}}{1-e^{i\xi}} = \frac{\sin(M+\frac{1}{2})\xi}{\sin\frac{1}{2}\xi}, \tag{28.15}$$

and likewise

$$D_N(\eta) = \frac{\sin(N+\frac{1}{2})\eta}{\sin\frac{1}{2}\eta}.$$

Thus (28.14) generalises (28.10) to the 2-variable case, but it gets nicer than just that. 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

$$\sigma_{MN} f(x,y) = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} F_M(\xi) F_N(\eta) f(x+\xi, y+\eta)\, d\xi\, d\eta, \tag{28.16}$$

in which

$$F_M(\xi) = \frac{1}{M+1} \sum_{m=0}^{M} D_m(\xi) = \frac{1}{M+1} \frac{\sin^2 \frac{(M+1)\xi}{2}}{\sin^2 \frac{\xi}{2}}, \tag{28.17}$$

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[8] 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] An easy variant of Exercise 2.58 is needed here.

**Theorem 28.15.** *Let $f(x) = f(x_1, \ldots, x_n)$ be complex valued and continuous in $x = (x_1, \ldots, x_n)$, $2\pi$-periodic in every $x_i$. If the Fourier coeffients*

$$c_m = c_{m_1 \ldots m_n} = \frac{1}{(2\pi)^n} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} f(x) e^{-i(m \cdot x)}\, dx_1 \cdots dx_n$$

*have the property that*

$$\sum_m |c_m| < \infty,$$

*then*

$$f(x) = \sum_m c_m e^{i(m \cdot x)}.$$

*The convergence is uniform in $x$.*

## 28.4 Derivation of the integral Fourier transform

I first discuss Fourier transform for functions of one variable. This starts from the intuitive presentation in §7.1 of Olver's PDE book, which I slightly modify and then merge with the rigorous approach in Folland's 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. The arguments above and below rely on the theory of Riemann integrals[9] only.

So 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

$$F(y) \sim \sum_{n=-\infty}^{\infty} C_n e^{iny} \tag{28.18}$$

just as in (28.3) 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 $\eta$-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.

[9]And may come before measure theory and Lebesgue integrals.

The Fourier series of $f$ on the interval $[-l,l]$ is obtained via scaling from the uniformly convergent Fourier series of the $2\pi$-periodic smooth extension of the smooth function $F$. This gives

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \frac{\pi}{l} \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-l}^{l} f(\xi) e^{-i\frac{n\pi}{l}\xi}\, d\xi}_{\text{this is } \widehat{f}(\frac{n\pi}{l}) \text{ if supp} f \subset [-l,l].} \; e^{i\frac{n\pi}{l}x} \tag{28.19}$$

for $x \in [-l,l]$. The underbraced term is de facto equal to[10]

$$\tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\xi) e^{-i\underset{\underset{n\Delta k}{\uparrow}}{\frac{n\pi}{l}}\xi}\, d\xi$$

for $l$ sufficiently large. Here $\xi$ is just a convenient dummy variable, and as indicated we recognised

$$\frac{n\pi}{l} = n\Delta k \quad \text{as an integer multiple of} \quad \frac{\pi}{l} = \Delta k.$$

Introducing the *Fourier integral transform*[11] $\widehat{f}$ of $f$ by

$$\widehat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\xi) e^{-ikx}\, d\xi, \tag{28.20}$$

the terms in the sum on the right hand side of (28.19) are

$$\Delta k\, \widehat{f}(n\Delta k)\, e^{ixn\Delta k}.$$

We now *see* that

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \widehat{f}(n\Delta k) e^{ixn\Delta k}\, \Delta k, \tag{28.21}$$

in which the sum looks like a Riemann sum[12] for

$$\int_{-\infty}^{\infty} \widehat{f}(k) e^{ikx}\, dk.$$

Note that we have changed the prefactor in order to have

$$\widehat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx}\, dx \quad \text{and} \quad f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \widehat{f}(k) e^{ikx}\, dk$$

[10] We dealt with integrals over the whole line in Section 7.9. See Section 28.5 below.

[11] Note the notational difference between $\widehat{f}(k)$ here and $\hat{f}(n)$ for Fourier coeffients.

[12] A series really.

as the outcome of a limit argument for $\Delta k \to 0$.

Likewise it follows for $l$ sufficiently large that

$$\int_{-\infty}^{\infty} |f(x)|^2 \, dx = \sum_{n=-\infty}^{\infty} |\widehat{f}(n\Delta k)|^2 \, \Delta k, \tag{28.22}$$

which looks like a Riemann sum for

$$\int_{-\infty}^{\infty} |\widehat{f}(k)|^2 \, dk,$$

and the identities (28.21) and (28.22) remain valid if we increase $l$ and thereby decrease the step size $\Delta k$. Both Riemann sums are independent of $\Delta k$ in the limit $\Delta k \to 0$. Can we conclude that then both

$$f(x) \quad \text{and} \quad \int_{-\infty}^{\infty} |f(x)|^2 \, dx$$

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 arguments[13] 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[14]

$$\widehat{f}(k) = \frac{\widehat{f'}(k)}{ik} = \frac{\widehat{f''}(k)}{(ik)^2}, \tag{28.23}$$

and so on for the Fourier transforms of the derivatives of $f$. Therefore

$$|\widehat{f}(k)| \leq \frac{1}{\sqrt{2\pi}k^2} \int_{-\infty}^{\infty} |f''(x)| \, dx. \tag{28.24}$$

For the limit $\Delta k \to 0$ in both (28.19) and (28.22) this suffices. The continuity of $f''$ and the compact support of $f$ thus imply all statements in the following theorem, except for the one about the smoothness of $\widehat{f}$, which follows from Theorem 13.5 or Theorem 27.3.

[13] We rely on Exercise 28.2 here, and the identification of $\mathbb{R}^2$ and $\mathbb{C}$.

[14] Because $|e^{ikx}| = 1$.

**Theorem 28.16.** *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*[15]*. Then (28.20) defines a smooth function $\widehat{f} : \mathbb{R} \to \mathbb{C}$ satisfying the estimate*

$$|\widehat{f}(k)| \leq \frac{1}{\sqrt{2\pi}\, k^2} \int_{-\infty}^{\infty} |f''(x)|\, dx = \frac{|f''|_1}{\sqrt{2\pi}\, k^2}$$

*for every real $k \neq 0$, so*

$$|\widehat{f}\,|_1 = \int_{-\infty}^{\infty} |\widehat{f}(k)|\, dk < \infty.$$

*Moreover,*

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \widehat{f}(k) e^{ikx}\, dk \quad \text{and} \quad \int_{-\infty}^{\infty} |f(x)|^2\, dx = \int_{-\infty}^{\infty} |\widehat{f}(k)|^2\, dk.$$

**Remark 28.17.** *Let $C_0 \cap L^1$ be the space of continuous functions $f : \mathbb{R} \to \mathbb{C}$ with $f(x) \to 0$ as $|x| \to \infty$ and $\int_{-\infty}^{\infty} |f| < \infty$. The definition of $\widehat{f}$ in (28.20) and the derivative formula's*

$$\widehat{f}(k) = \frac{\widehat{f'}(k)}{ik} = \frac{\widehat{f''}(k)}{(ik)^2}$$

*in (28.23), with the estimates*

$$|\widehat{f}(k)| \leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |f(x)|\, dx, \quad |\widehat{f}(k)| \leq \frac{1}{\sqrt{2\pi}|k|} \int_{-\infty}^{\infty} |f'(x)|\, dx,$$

$$|\widehat{f}(k)| \leq \frac{1}{\sqrt{2\pi} k^2} \int_{-\infty}^{\infty} |f''(x)|\, dx,$$

*are valid if $f, f', f'' \in C_0 \cap L^1$. These statements about $\widehat{f}$ follow via integration by parts and the obvious estimates as before.*

**Exercise 28.18.** Prove the statements in the Remark 28.17. What can you conclude if also $f''' \in C_0 \cap L^1$?

**Remark 28.19.** *If $f$ is in $C_c^\infty$, the class of smooth compactly supported $\mathbb{C}$-valued functions*[16] *on $\mathbb{R}$, it not only follows from Theorem 27.3 that $\widehat{f}$ is*

---

[15] So $f, f, f'' \in C_c$.

[16] The test functions used in the theory of *distributions*.

*in $C^\infty$, but also that every derivative of $\widehat{f}$ is itself a Fourier transform. Along the same lines as above we then have that*

$$\forall_{n\in\mathbb{N}_0}\,\forall_{p\in\mathbb{N}_0}\,\exists_{C>0}\,\forall_{k\in\mathbb{R}}\,:\,|k|^p\,|(\widehat{f}\,)^{(n)}(k)|\leq C. \tag{28.25}$$

*Details will be given in Section 28.5 below, also for the stronger statement in Theorem 28.20 about the so-called Schwarz class of functions in $C^\infty$ defined by (28.25). In the less than optimal result*

$$f\in C_c^\infty \implies \widehat{f}\in\mathcal{S}$$

*the conclusion $\widehat{f}\in\mathcal{S}$ cannot be replaced by $\widehat{f}\in C_c^\infty$.*

**Theorem 28.20.** *The stronger implication*

$$f\in\mathcal{S} \implies \widehat{f}\in\mathcal{S}$$

*holds for all $f:\mathbb{R}\to\mathbb{C}$.*

## 28.5 Differentiation under integrals over the real line

In Section 7.9 we considered integrals over $\mathbb{R}$. It is necessary for the proof of Theorem 28.25 to have a version of the statements in Section 27.2 for integrals

$$J(t)=\int_{-\infty}^{\infty} f(t,x)\,dx,\quad t\in I, \tag{28.26}$$

since

$$2\pi\,\widehat{f}(t)=\int_{-\infty}^{\infty} f(x)e^{-itx}\,dx=\int_{-\infty}^{\infty} f(x)\cos(tx)\,dx-i\int_{-\infty}^{\infty} f(x)\sin(tx)\,dx.$$

We discuss the real valued version of the statement we need first. It applies to the two integrals on the right hand side just above if $f$ is a continuous real valued function for which

$$\int_{\infty}^{\infty}|\underbrace{xf(x)}_{f_1(x)}|\,dx$$

exists[17] as a real number via Definition 7.30. The statement then is that the derivatives exist as integrals, namely

$$\frac{d}{dt}\int_{-\infty}^{\infty} f(x)\cos(tx)\,dx=-\int_{-\infty}^{\infty} xf(x)\sin(tx)\,dx,$$

[17] This is often sloppily formulated as $\int_{-\infty}^{\infty}|xf(x)|\,dx<\infty$.

$$\frac{d}{dt}\int_{-\infty}^{\infty} f(x)\sin(tx)\,dx = \int_{-\infty}^{\infty} xf(x)\cos(tx)\,dx,$$

and are continuous in $t$. By the definition of the complex integrals in the Fourier transforms this is equivalent

$$\widehat{f}' = -i\widehat{f_1}, \quad f_1(x) = xf(x), \tag{28.27}$$

a formula which mirrors

$$\widehat{f'}(k) = ik\widehat{f}(k).$$

**Theorem 28.21.** *Let the improper Riemann integrals*

$$J(t) = \int_{-\infty}^{\infty} f(t,x)\,dx$$

*be well defined for t in some interval I. Assume that $g(t,x) = f_t(t,x)$ is the continuous partial derivative of $f$ with respect to $t$ and that $|g(t,x)| \le M(x)$ for all $\in \mathbb{R}$ and all $t$ in some interval $[a,b] \subset I$. If*

$$\int_{-\infty}^{\infty} M$$

*exists then*

$$j(t) = \int_{-\infty}^{\infty} g(t,x)\,dx$$

*defines a function $j \in C([a,b])$, and $J$ is a primitive function of $j$ on $[a,b]$.*

**Proof**. Via Theorem 7.87 this $j(t)$ is not only well defined as

$$j(t) = \lim_{R\to\infty} \int_{-R}^{R} g(t,x)\,dx$$

for all $t \in [a,b]$, but also the uniform $\varepsilon$-estimate

$$|j(t) - \int_{-R}^{R} g(t,x)\,dx| \le \bar{M} - \int_{-R}^{R} M(x)\,dx < \varepsilon \tag{28.28}$$

holds for $R$ sufficient large[18]. Since

$$\int_s^t \int_{-R}^{R} g = \int_{-R}^{R}\int_s^t g = \int_{-R}^{R} f(t,x)\,dx - \int_{-R}^{R} f(s,x)\,dx \tag{28.29}$$

[18]Check this!

for all $s, t \in [a, b]$, we can write the difference $j(t) - j(s)$ as

$$\underbrace{j(t) - \int_{-R}^{R} g(t,x)\,dx}_{<\varepsilon} + \int_{-R}^{R} g(t,x)\,dx - \int_{-R}^{R} g(s,x)\,dx + \underbrace{\int_{-R}^{R} g(s,x)\,dx - j(s)}_{<\varepsilon}.$$

With $R$ already chosen via the uniform $\varepsilon$-estimate (28.28) the continuity of $g$ allows for a by now standard proof that the difference is less than $3\varepsilon$ by choosing $\delta > 0$ accordingly to have the middle term smaller than $\varepsilon$ for $|t - s| < \delta$. Finish the proof by doing Exercise 28.22. $\square$

**Exercise 28.22.** Use Theorem 10.10 to complete the proof. Hint: use (28.28), (28.29) and

$$J(t) = \lim_{R\to\infty} \int_{-R}^{R} f(t,x)\,dx$$

to show that

$$\int_s^t j = J(t) - J(s).$$

**Exercise 28.23.** Use Exercise 8.17 and modified versions of (28.8) to formulate and prove a version of Theorem 28.21 for complex valued functions. Use it to verify (28.27) direclty.

## 28.6 The Fourier transform as a bijection

The pairing[19]

$$\widehat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x)e^{-i\xi x}\,dx \quad \text{and} \quad f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \widehat{f}(\xi)e^{i\xi x}\,d\xi \quad (28.30)$$

discovered with (28.23) 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

$$\widehat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x)e^{-i\xi x}\,dx \tag{28.31}$$

[19] I now prefer $\xi$ as name for the Fourier variable.

of $\hat{f}$. For $\widehat{f}$ to be well defined it certainly suffices that $f$ is in $C \cap L^1$, i.e.

$$f : \mathbb{R} \to \mathbb{C} \text{ is continuous and } |f|_1 = \int_{\mathbb{R}} |f| < \infty. \tag{28.32}$$

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 $\infty$-norm.

**Theorem 28.24.** *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 (28.32).*

**Proposition 28.25.** *If $f \in C \cap L^1$ then $\widehat{f} \in C_0$. If we write*

$$f(x) \quad \stackrel{\widehat{}}{\to} \quad \widehat{f}(\xi)$$

*to say that $\widehat{f}$ is the Fourier transform of $f$, then for every $y, \eta \in \mathbb{R}$ and $a > 0$ it holds that*

$$e^{i\eta x} f(a(x-y)) \quad \stackrel{\widehat{}}{\to} \quad \frac{1}{a}\, e^{-iy\xi} \widehat{f}\left(\frac{\xi - \eta}{a}\right), \tag{28.33}$$

*and*

$$e^{-\frac{1}{2}x^2} \quad \stackrel{\widehat{}}{\to} \quad e^{-\frac{1}{2}\xi^2}.$$

**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

$$\sqrt{2\pi}\,|\widehat{f_n}(\xi) - \widehat{f}(\xi)| = \left|\int_{-\infty}^{\infty} f_n(x)e^{-i\xi x}\,dx - \int_{-\infty}^{\infty} f(x)e^{-i\xi x}\,dx\right| =$$

$$\left|\int_{-\infty}^{\infty} (f_n(x) - f(x))e^{-i\xi x}\,dx\right| \leq \int_{-\infty}^{\infty} |f_n(x) - f(x))| = |f_n - f|_1 \to 0,$$

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

$$\widehat{f_n}(\xi) = \frac{\widehat{f_n'}(\xi)}{i\xi}, \quad \text{whence} \quad |\widehat{f_n}(\xi)| \leq \frac{|f_n'|_1}{|\xi|} \quad \text{and} \quad \widehat{f_n} \in C_0.$$

Since $C_0$ is closed in $C_b$ it follows that $\widehat{f} \in C_0$. The statement in (28.33) is easily checked. The exercises below deal with the statement about the Fourier transform of the Gaussian. $\square$

**Exercise 28.26.** Suppose that $f$ has a continuous derivative $f'$, and that $f(x) \to 0$ as $|x| \to \infty$. Explain again that

$$i\xi\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \lim_{R\to\infty} \int_{-R}^{R} f'(x)e^{-i\xi x}\,dx.$$

Thus the Fourier transform of $f'(x)$ is given by $i\xi\hat{f}(\xi)$ if $\int_{-\infty}^{\infty} |f'| < \infty$.

**Exercise 28.27.** Show again that for $f : \mathbb{R} \to \mathbb{C}$ with $\int_{-\infty}^{\infty} |f| < \infty$ the formula

$$\frac{d}{d\xi}\int_{-\infty}^{\infty} f(x)e^{-i\xi x}\,dx = -i\int_{-\infty}^{\infty} xf(x)e^{-i\xi x}\,dx$$

is justified if in addition it holds that $\int_{-\infty}^{\infty} |xf(x)|\,dx < \infty$. Hint: use

$$|\frac{\partial}{\partial\xi}f(x)e^{-i\xi x}| \leq |xf(x)|.$$

For such $f$ the Fourier transform is thus differentiable and $\widehat{f}'$ is the Fourier transform of the function $f_1$ defined bu

$$x \xrightarrow{f_1} -ixf(x).$$

**Exercise 28.28.** The function $u_a$ is defined by $u_a(x) = e^{-ax^2}$. Use reasoning as above to establish that the function $v$ defined by $v(a,\xi) = \widehat{u_a}(\xi)$ voor $a > 0$ and $\xi \in \mathbb{R}$ has a partial derivative with respect to $a$ which is given by the Fourier transform of $x \to -x^2 u_a(x)$. Then show that $v$ is a (smooth) solution of the partial differential equation

$$\frac{\partial v}{\partial a} = \frac{\partial^2 v}{\partial \xi^2},$$

a PDE that is known as the linear heat equation in space dimension 1.

**Exercise 28.29.** Show directly from the definition of $\widehat{u_a}$ that $v$ has the selfsimilar form

$$v(a,\xi) = \widehat{u_a}(\xi) = \frac{1}{\sqrt{a}}\widehat{u_1}(\frac{\xi}{\sqrt{a}}).$$

**Exercise 28.30.** Show that $F = \widehat{u_1}$ satisfies the ordinary differential equation

$$\text{(ODE)} \qquad F''(\eta) + \frac{1}{2}\eta F'(\eta) + \frac{1}{2}F(\eta) = 0,$$

in which $\eta = \frac{\xi}{\sqrt{a}}$, a so-called similarity variable .

**Exercise 28.31.** Show that

$$F(\eta) = F(0)e^{-\frac{1}{4}\eta^2}.$$

Hint: $F(0)$ and $F'(0)$ determine the solution of (ODE) uniquely. Why is $F'(0) = 0$?

**Exercise 28.32.** Determine $F(0)$ and thereby $\widehat{u_a}$ for every $a > 0$. Verify the last statement in Proposition 28.25.

**Proposition 28.33.** *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 order[20] 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 $\xi$, and of $f$ and $g$. □

**Theorem 28.34.** *Let*

$$X = \{f \in C_0 \cap L^1 : \widehat{f} \in C_0 \cap L^1\}. \tag{28.34}$$

*Then $C_c^2 \subset X$, so $X$ is nonempty. The Fourier transform is a bijection between $X$ and itself in the sense that*

$$g(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-i\xi x}\, dx \iff f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(\xi) e^{ix\xi}\, d\xi \tag{28.35}$$

*for all $f, g \in X$. In particular the equalities in (28.35) and (28.30) hold pointwise for every $x$ and every $\xi$ when $f$ is in $X$. Moreover,*

$$\int_{-\infty}^{\infty} |\widehat{f}(\xi)|^2\, d\xi = \int_{-\infty}^{\infty} |f(x)|^2\, dx \le |f|_{max} |f|_1,$$

*so the bijection uniquely extends to the completion of $C_c^2$ with respect to the 2-norm, regardless of how that completion is defined: abstractly or in some larger space constructed by other means.*

[20]Why is this allowed? You know it for integrals of continuous functions over rectangles.

**Proof**. We observe that $X$ contains $C_c^2$ because of Theorem 28.16 in Section 28.4.This statement is independent of the rest of the theorem, which we prove next. So let $g$ be defined by the left hand side of (28.35). Note that $\hat{f} \in C_0$ because of Proposition 28.25. In this context the extra assumption $\widehat{f} \in L^1$ is equivalent to the existence of

$$\int_{-\infty}^{\infty} |\widehat{f}|$$

as a number in $\mathbb{R}$. This is needed when we apply Proposition 28.33 with the function $g(\xi)$ and $\hat{g}(x)$ that appear there[21] replaced by the functions on both sides of this application:

$$e^{i\eta\xi} e^{-\frac{1}{2}a^2\xi^2} \quad \overset{\frown}{\to} \quad \frac{1}{a}\, e^{-\frac{1}{2a^2}(x-\eta)^2}.$$

The left hand side of the equality Proposition 28.33 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 in Proposition 28.33 becomes

$$\int_{-\infty}^{\infty} f(x) \frac{1}{a}\, e^{-\frac{1}{2a^2}(x-\eta)^2}\, dx = \int_{-\infty}^{\infty} \frac{1}{a}\, e^{-\frac{1}{2a^2}x^2} f(\eta - x)\, dx.$$

**Exercise 28.35.** Put $a^2 = 2t$ to recognise, up to a usual factor, the solution formula for the partial differential equation $u_t = u_{xx}$ with initial data $u(0,x) = f(x)$ by convolution with the (heat) kernel

$$E_t(x) = \frac{1}{2\sqrt{\pi t}} e^{-\frac{x^2}{4t}}.$$

This is very much like Exercise 28.13 for (28.13). Prove for this good kernel that $E_t * f \to f$ uniformly[22] as $t \downarrow 0$ for every $f \in C_0$. Get the prefactor right to conclude that the right hand side of (28.35) holds.

This finishes the proof of $\Longrightarrow$ in (28.35). The proof of $\Longleftarrow$ in (28.35) is of course similar. $\square$

[21]Not to be confused with the $g$ in (28.35).

[22]For the pointwise convergence $f \in C_b \cap L^1$ more than suffices.

**Remark 28.36.** *If we denote the space of measurable complex valued Lebesgue measurable integrable functions by* $L^1$*, then*

$$f \in L^1 \implies \widehat{f} \in C_0.$$

*There is no point in considering possible versions of Theorem 28.34 with* $C_0$ *replaced by* $C_b$ *or even* $C$ *in (28.34). By itself, the assumption* $f \in C \cap L^1$ *is sufficient to conclude*

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{ix\xi}\, d\xi. \tag{28.36}$$

**Remark 28.37.** *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*[23] *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*

$$\int_{-\infty}^{\infty} f(x)\, \overline{g(x)}\, dx = \int_{-\infty}^{\infty} \widehat{f}(\xi)\, \overline{\widehat{g}(\xi)}\, d\xi, \tag{28.37}$$

*which you should compare to Theorem 28.33. If* $f \in L^1 \cap L^2$ *we can copy (28.20), replacing the integral by a Lebesgue integral.*

**Theorem 28.38.** *The space* $X$ *contains the Schwarz class* $\mathcal{S}$*. Thus (28.30) 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 (28.35) for* $f, g$ *in* $X$*.*

**Proof**. Theorem 28.20 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 (28.35) holds for $f, g \in S$. □

[23] In Chapter 31 we indicate how for all practical purposes this concept may be avoided.

## 28.7 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 $\Xi$, 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 $\Xi$ but the shifts do. They change $X$ and $\Xi$ 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 $\Xi$ with zero expectation. Integrating the integral for the squared 2-norm of $f$ by parts with the 1-trick[24] 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 $f_1, \phi_1$ are defined by

$$f_1(x) = xf(x) \quad \text{and} \quad \phi_1(x) = x\phi(x).$$

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

$$\int_{-\infty}^{\infty} |f_1(x)|^2\,dx = \int_{-\infty}^{\infty} |x|^2\,|f(x)|^2\,dx,$$

$$\int_{-\infty}^{\infty} |\phi_1(x)|^2\,dx = \int_{-\infty}^{\infty} |x|^2\,|\phi(x)|^2\,dx,$$

[24]Recall $\int_1^x \ln s\,ds = \int_1^x 1\,\ln s\,ds = [s\ln s]_1^x - \int_1^x s\,\frac{1}{s}\,ds = x\ln x + x - 1$.

this establishes the estimate

$$4\,EX^2\,E\,\Xi^2 \geq 1$$

for the product of the variations of $X$ and $\Xi$. For the standard deviations the conclusion is that

$$2\sigma(X)\,\sigma(\Xi) \geq 1, \tag{28.38}$$

which corresponds to the Heisenberg Uncertainty Principle.

## 28.8 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

$$-u''(x) + u(x) = f(x) \tag{28.39}$$

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 you will find out in Exercise 28.42 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 $\pi$-dependent prefactor. And likewise for problems in which $x$ is taken modulo $2\pi$, 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[25] $u'' = f - u$ is also continuous, so we know *a priori* that the Fourier coefficients $\hat{u}(n)$ must have a quadratic decay. Indeed,

$$\hat{u''}(n) = n^2\hat{u}(n)$$

[25] This line of reasoning only works for ordinary differential equations unfortunately.

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 $2\pi$-periodic integrable functions is defined by

$$(f*g)(x) = \int_{-\pi}^{\pi} f(x-y)g(y)\,dy = \int_{-\pi}^{\pi} f(y)g(x-y)\,dy \tag{28.40}$$

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

$$\int_{-\pi}^{\pi} (f*g)(x)\,e^{-inx}\,dx = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x-y)g(y)\,dy\,e^{-inx}\,dx$$

$$= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x-y)g(y)\,e^{-inx}\,dy\,dx$$

$$= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x-y)g(y)\,e^{-in(x-y)}\,e^{-iny}\,dx\,dy$$

$$= \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x)g(y)\,e^{-inx}\,e^{-iny}\,dx\,dy = \underbrace{\int_{-\pi}^{\pi} f(x)\,e^{-inx}\,dx}_{2\pi\hat{f}(n)}\,\underbrace{\int_{-\pi}^{\pi} g(y)\,e^{-iny}\,dy}_{2\pi\hat{g}(n)}.$$

Upto an annoying factor $2\pi$ the Fourier coefficients of $f*g$ are the products of the Fourier coefficients of $f$ and of $g$. This allows to conclude that the statement in this theorem holds.

**Theorem 28.39.** *Let $f, g \in C_{2\pi}$. Then*

$$(f*g)(x) = 2\pi \sum_{n=-\infty}^{\infty} \hat{f}(n)\hat{g}(n)e^{inx} \quad \text{for all} \quad f, g \in C_{2\pi}, \tag{28.41}$$

*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*

$$|\hat{f}\hat{g}|_1 = \sum_{n=-\infty}^{\infty} |\hat{f}(n)\hat{g}(n)| \le \sqrt{\sum_{n=-\infty}^{\infty} |\hat{f}(n)|^2}\sqrt{\sum_{n=-\infty}^{\infty} |\hat{g}(n)|^2} = |\hat{f}|_2\,|\hat{g}|_2.$$

Summing up, the Fourier coefficients of $f * g$ were obtained 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

$$u = G * f, \quad G(x) = \frac{1}{2\pi} \sum_{n \in \mathbb{Z}} \frac{1}{1+n^2} \, e^{inx} = \frac{1}{\pi} \left( \frac{1}{2} + \sum_{n=1}^{\infty} \frac{\cos nx}{1+n^2} \right). \tag{28.42}$$

We note that this avoids the use of solutions with $f$ replaced by the Dirac $\delta$-function. In principal we can avoid 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 28.9 and illustrate the advantage of the use of the $\delta$-function with an example. Usually way before this theory is well understood, solutions of equations with $\delta$ 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 $2\pi$-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 $2\pi$), because $u'' = u - \delta$, and the "integral" of $\delta(x)$ over any interval $(-\varepsilon, \varepsilon)$ is equal to 1. Combined with symmetry and $2\pi$-periodicity this then implies that[26] the solution of (28.39) with $f(x)$ replaced by $\delta(x)$ is given by

$$u(x) = G(x) = \frac{\cosh(x - \pi)}{2 \sinh \pi} \quad \text{for} \quad 0 \leq x \leq 2\pi,$$

and you easily check that indeed[27]

$$u(x) = G(x) = \sum_{n \in \mathbb{Z}} \frac{e^{inx}}{n^2 + 1} \quad \text{for all} \quad x \in \mathbb{R}.$$

Apparently we have that in the class of $2\pi$-periodic functions the solution operator for (28.39) maps $f = \delta$ to $G$.

Next we consider the Fourier coefficients of $fg$. This is more difficult. Again

$$f(x) \sim \sum_{n=-\infty}^{\infty} \hat{f}(n) e^{inx} \quad \text{and} \quad g(x) \sim \sum_{n=-\infty}^{\infty} \hat{g}(n) e^{inx}$$

[26] You may like to compare $-G'(x + \pi)$ to the sawtooth in $Z(x)$ in Section 28.1.

[27] Draw the $2\pi$-periodic graph and compute $2\pi \hat{G}(n)$ as $\int_0^{2\pi} \cosh(x - \pi) \, e^{-inx} \, dx$.

have a clear meaning if

$$|\hat{f}|_{1} = \sum_{n=-\infty}^{\infty} |\hat{f}(n)| < \infty \quad \text{and} \quad |\hat{g}|_{1} = \sum_{n=-\infty}^{\infty} |\hat{g}(n)| < \infty, \tag{28.43}$$

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 $f(x)g(x) =$

$$\sum_{k=-\infty}^{\infty} \hat{f}(k)e^{ikx} \sum_{m=-\infty}^{\infty} \hat{g}(m)e^{imx} = \sum_{n=-\infty}^{\infty} \underbrace{\sum_{k+m=n} \hat{f}(k)\hat{g}(m)}\, e^{inx}, \tag{28.44}$$

so if (28.43) holds it must be that the underbraced factor is the $n$-th Fourier coefficient of $fg$. We rewrite this factor as

$$\sum_{k+m=n} \hat{f}(k)\hat{g}(m) = \sum_{k=-\infty}^{\infty} \hat{f}(k)\hat{g}(n-k), \tag{28.45}$$

a *discrete convolution*. For its partial sums we have that

$$\begin{aligned}
(2\pi)^2 \sum_{k=-N}^{N} \hat{f}(k)\hat{g}(n-k) &= \sum_{k=-N}^{N} \int_{-\pi}^{\pi} f(x)e^{-ikx}\,dx \int_{-\pi}^{\pi} g(y)e^{-i(n-k)y}\,dy \\
&= \sum_{k=-N}^{N} \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x)e^{-ik(x-y)}\,dx\, g(y)\, e^{-iny}\,dy \\
&= \int_{-\pi}^{\pi} \underbrace{\int_{-\pi}^{\pi} f(x+y) \sum_{k=-N}^{N} e^{-ikx}\,dx}_{\text{in view of (28.10) this is } 2\pi\, S_N f(y)}\; g(y)\, e^{-iny}\,dy,
\end{aligned} \tag{28.46}$$

so the conclusion should be that

$$\begin{aligned}
\sum_{k=-N}^{N} \hat{f}(k)\hat{g}(n-k) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} S_N f(y) g(y)\, e^{-iny}\,dy \\
&\to \frac{1}{2\pi} \int_{-\pi}^{\pi} f(y)g(y)\, e^{-iny}\,dy
\end{aligned} \tag{28.47}$$

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 28.40.** *Let $f, g \in C_{2\pi}$. Then the complex Fourier coefficients of $fg$ are given by*

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} f(y)g(y)\, e^{-iny}\, dy = \sum_{k=-\infty}^{\infty} \hat{f}(k)\hat{g}(n-k). \tag{28.48}$$

Now let $f, g \in C_b \cap L^1$. Then for the Fourier transform of the convolution[28]

$$f * g(x) = \int_{-\infty}^{\infty} f(x-y)g(y)\, dy = \int_{-\infty}^{\infty} f(y)g(x-y)\, dy \tag{28.49}$$

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

$$F(x) = Kf(x) = \int_{-\infty}^{\infty} K(x,y)f(y)\, dy, \tag{28.50}$$

in which $K(x,y)$ is continuous with

$$\int_{-\infty}^{\infty} |K(x,y)|\, dx \leq C \quad \text{and} \quad \int_{-\infty}^{\infty} |K(x,y)|\, dy \leq C$$

[28]See Exercise 7.90.

for all $x, y \in \mathbb{R}$ and some fixed $C > 0$. For $f \in C_b \cap L^1$ we have

$$|F(x)| \leq \int_{-\infty}^{\infty} |K(x,y)|\,|f(y)|\,dy \leq \int_{-\infty}^{\infty} |K(x,y)|\,dy\,|f|_{\infty} \leq C\,|f|_{\infty},$$

and also

$$|F|_{1} = \int_{-\infty}^{\infty} \Big| \int_{-\infty}^{\infty} K(x,y)f(y)\,dy\Big|\,dx \leq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |K(x,y)f(y)|\,dy\,dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |K(x,y)f(y)|\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |K(x,y)|\,dx\,|f(y)|\,dy$$

$$\leq C \int_{-\infty}^{\infty} |f(y)|\,dy = C|f|_{1}.$$

You should check that in fact $F \in C_b \cap L^1$. These estimates can be applied to (28.49) with $K(x,y) = f(x-y)$ or $K(x,y) = g(x-y)$, all this proves the following theorem about $f * g$.

**Theorem 28.41.** *Let $f, g \in C \cap L^1$. Then the convolution $f * g$, defined by*

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-y)g(y)\,dy = \int_{-\infty}^{\infty} f(y)g(x-y)\,dy, \tag{28.51}$$

*is in $C \cap L^1$ as well, and its Fourier transform is given by*

$$\widehat{f * g}(\xi) = \sqrt{2\pi}\,\widehat{f}(\xi)\widehat{g}(\xi). \tag{28.52}$$

*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*

$$|\widehat{f * g}|_{1} \leq \sqrt{2\pi}|\widehat{f}\widehat{g}|_{2} = \sqrt{2\pi}|\widehat{f}g|_{2} \leq \sqrt{2\pi}|\widehat{f}|_{1}\,|g|_{\infty},$$

*and finally Remark 28.36 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 (28.34) we have

$$2\pi f(x)g(x) = \int_{-\infty}^{\infty} \widehat{f}(k)e^{ikx}\,dk \int_{-\infty}^{\infty} \widehat{g}(m)e^{imx}\,dm$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \widehat{f}(k)\widehat{g}(m)\,e^{i(k+m)x}\,dm\,dk$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \widehat{f}(k)\widehat{g}(m-k)\, e^{i(k+m-k)x}\, dm\, dk$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \widehat{f}(k)\widehat{g}(m-k)\, dk\, e^{imx}\, dm$$

$$= \int_{-\infty}^{\infty} (\widehat{f} * \widehat{g})(m)\, e^{imx}\, dm,$$

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

$$\widehat{fg}(m) = \frac{1}{\sqrt{2\pi}}(\widehat{f} * \widehat{g})(m). \tag{28.53}$$

So let $f, g \in C \cap L^1$.

A direct calculation in the spirit of what followed after (28.45) gives that

$$2\pi \int_{-R}^{R} \widehat{f}(k)\widehat{g}(m-k)\, dk = \int_{-R}^{R} \int_{-\infty}^{\infty} f(x)e^{-ikx}\, dx \int_{-\infty}^{\infty} g(y)e^{-i(m-k)y}\, dy\, dk$$

$$= \int_{-R}^{R} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)e^{-ikx} g(y)e^{-i(m-k)y}\, dx\, dy\, dk$$

$$= \int_{-R}^{R} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x+y)e^{-ikx} g(y)e^{-imy}\, dx\, dy\, dk$$

$$= \int_{-\infty}^{\infty} \int_{-R}^{R} \int_{-\infty}^{\infty} f(x+y)e^{-ikx}\, dx\, dk\, g(y)e^{-imy}\, dy,$$

$$= \int_{-\infty}^{\infty} \underbrace{\int_{-\infty}^{\infty} f(x+y) \int_{-R}^{R} e^{-ikx}\, dk\, dx}\; g(y)e^{-imy}\, dy,$$

which you should compare to (28.46), 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 (28.53) holds for $f,g\in C\cap L^1$.

**Exercise 28.42.** Consider again (28.39), but now in the class of integrable functions. Follow the same reasoning as for $2\pi$-periodic functions, relating to Theorem 28.41 instead of Theorem 28.39. You should arrive at

$$u(x)=G(x)=\exp(-\frac{|x|}{2})$$

as the solution with $f=\delta$. What would $\widehat{\delta}$ and $\widehat{G}$ be?

## 28.9 Remark on Fourier transforms of distributions

This section could build on an earlier section not included here yet about the distributional definition of generalised functions such as the $\delta$-function[29]. With Proposition 28.33 we showed that $f,\phi\in C\cap L^1$ implies $\widehat{f},\widehat{\phi}\in C_0$ and

$$\langle \widehat{f},\phi\rangle=\int_{-\infty}^{\infty}\widehat{f}(\xi)\phi(\xi)\,d\xi=\int_{-\infty}^{\infty}f(x)\widehat{\phi}(x)\,dx=\langle f,\widehat{\phi}\rangle, \tag{28.54}$$

in which we use the notation

$$\langle f,g\rangle=\int_{-\infty}^{\infty}fg. \tag{28.55}$$

So certainly (28.54) holds for all $\phi\in\mathcal{S}$ if $f\in C\cap L^1$. We now take (28.54) as the defining property of $\widehat{f}:\mathcal{S}\to\mathbb{C}$ if $f:\mathcal{S}\to\mathbb{C}$ is a linear functional

$$\phi\to\langle f,\phi\rangle$$

[29]And solutions of equations like $u-u''=\delta$.

defined for $\phi \in \mathcal{S}$. Likewise the definition of the weak derivative $f'$ copies

$$\langle f', \phi \rangle = \int_{-\infty}^{\infty} f'(x)\phi(x)\, dx = -\int_{-\infty}^{\infty} f(x)\phi'(x)\, dx = -\langle f, \phi' \rangle \tag{28.56}$$

for e.g. $f, \phi \in C^1 \cap C_0$ to define the linear functional $f' : C_c^\infty \to \mathbb{C}$ by[30]

$$\langle f', \phi \rangle = -\langle f, \phi' \rangle, \tag{28.57}$$

which is well defined for every linear functional $f$ on $C_c^\infty$. It may happen that $f'$ is in fact a function of course.

**Exercise 28.43.** Examine what the above definition of $f'$ gives if $f$ is given by the function defined by $f(x) = |x|$ for all $x \in \mathbb{R}$. Characterise $f''$ as well before your read on.

An example of such a linear functional is $\delta_s$ defined by

$$\langle \delta_s, \phi \rangle = \phi(s).$$

We note that $\delta_s$ 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 a dual[31] space $X^*$ of some space $X$ for $\phi$, via its action on the space for $\phi$ 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

$$\int f(s)\delta_s\, ds = f, \quad \text{informally written in turn as} \quad \int f(s)\delta(x-s)\, ds = f(x).$$

Solving equations like (28.39) with right hand side $\delta_s$ we obtain a Green's function $G_s$ and can then actually prove that the convolution of $G_s$ and $f$ solves (28.39) with right hand side $f$.

[30] For (28.54) the assumption on $f$ is stronger, it needs to be a linear functional on $\mathcal{S}$.

[31] A space of all Lipschitz continuous linear functions from some normed space $X$ to $\mathbb{R}$.

## 28.10 Playing with Fourier series solutions of PDE's

Recall from (28.3) that we write

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{inx} = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx),$$

with

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx \, dx, \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx \, dx,$$

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} \, dx$$

for $2\pi$-periodic integrable functions. For $2\pi$-periodic continuous functions it is in general not true that

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx} = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx)$$

for every $x$, as we only have square summability of the coefficients, which (therefore certainly) go to zero as $n \to \infty$.

**Exercise 28.44.** Let $f$ be such a $2\pi$-periodic continuous function. Show that

$$u(t,x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \underbrace{\exp(-n^2 t)}_{e^{-n^2 t}}$$

defines a smooth solution $u(t,x)$ of $u_t = u_{xx}$ defined for all $x \in \mathbb{R}$ and for all $t > 0$. Use the boundedness of the Fourier coefficients.

**Exercise 28.45.** Let again $f$ be a $2\pi$-periodic continuous function. Show that polar coordinates $x = r\cos\theta, y = r\sin\theta$ and

$$u(x,y) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos n\theta + b_n \sin n\theta) r^n$$

define a smooth solution of $u_{xx} + u_{yy} = 0$ defined for all $x, y \in \mathbb{R}$ with $x^2 + y^2 < 1$. Use the formula's for the coefficients $a_n$ and $b_n$ to show that

$$u(x,y) = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_r(\theta - \phi) f(\phi) \, d\phi,$$

in which the $2\pi$-periodic symmetric good Poisson kernel $P_r$ is defined by

$$P_r(\phi) = 1 + 2\sum_{n=1}^{\infty} r^n \cos(n\phi) = \frac{1-r^2}{1+r^2-2r\cos\phi}$$

and ranges between

$$\frac{1+r}{1-r} \quad \text{and} \quad \frac{1-r}{1+r} \quad \text{in} \quad \phi = 0 \quad \text{and} \quad \phi = \pi.$$

**Exercise 28.46.** Use

$$\int_{-\pi}^{\pi} P_r = 2\pi$$

and the estimate

$$\underbrace{P_r(\psi)}_{\to 0 \text{ as } r\uparrow 1} \geq P_r(\phi) \geq \frac{1-r}{1+r} > 0 \quad \text{for} \quad 0 < \psi \leq \phi \leq \pi$$

to show for $v$ defined by

$$v(r,\phi) = 1 + 2\sum_{n=1}^{\infty} r^n \cos(n\phi)$$

that $v(r,\phi) \to f(\phi_0)$ if $r \to 1$ and $\phi \to \phi_0$. Hint: show that

$$|v(r,\phi) - f(\phi_0)| = \big|\frac{1}{2\pi}\int_{-\pi}^{\pi} P_r(\theta-\phi)(f(\phi) - f(\phi_0))\,d\phi\big| < \varepsilon$$

if $|\phi - \phi_0| < \delta$ and $0 < 1 - r < \delta$ with $\delta > 0$ depending on $\varepsilon > 0$. This is a stronger statement than in Exercise 28.13 for convolution with the good kernel $F_N$.

**Exercise 28.47.** Formulate and prove that stronger statement in Exercise 28.13.

**Remark 28.48.** *Let again $f$ be a $2\pi$-periodic continuous function. Is it true that*

$$u(t,x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos nx + b_n \sin nx)e^{-n^2 t} = \frac{1}{2\pi}\int_{-\pi}^{\pi} H_t(x-y)f(y)\,dy$$

*for $x \in \mathbb{R}$ and $t > 0$, with a heat kernel $H_t$ that has properties similar to that of $P_r$? Clearly $H_t$ should be given by*

$$H_t(x) = 1 + 2\sum_{n=1}^{\infty} e^{-n^2 t}\cos(nx),$$

*but there is no complex geometric series calculation that gives an expression for the sum of the series.*

**Exercise 28.49.** Show that

$$\begin{aligned} 1 + 2\sum_{n=1}^{N} r^n \cos(n\phi) &= P_r(\phi) + 2r^{N+1}\frac{r\cos N\phi - \cos(N+1)\phi}{1 + r^2 - 2r\cos\phi} \\ &= \frac{1 - r^2 + 2r^{N+1}(r\cos N\phi - \cos(N+1)\phi)}{1 + r^2 - 2r\cos\phi} \end{aligned}$$

and examine sign and size of these partial sums for $r$ close to 1. Check that for $r = 1$ this sum reduces to the Dirichlet kernel (28.11).

**Remark 28.50.** *In view of Exercise 28.49 we cannot expect to see the good properties of $H_t$ in Exercise 28.48 from its partial sums. But recall that in Exercise 28.35 we encountered*[32]

$$E_t(x) = \frac{1}{2\sqrt{\pi t}} e^{-\frac{x^2}{4t}} \quad \text{and} \quad u(t,x) = \int_{\mathbb{R}} E_t(x-y) f(y)\,dy$$

*for $u_t = u_{xx}$ evaluates for $2\pi$-periodic continuous $f$ as*

$$\begin{aligned} u(t,x) = \sum_{n\in\mathbb{Z}} \int_{(2n-1)\pi}^{(2n+1)\pi} E_t(x-y) f(y)\,dy &= \sum_{n\in\mathbb{Z}} \int_{-\pi}^{\pi} E_t(x-y-2n\pi) f(y)\,dy \\ &= \int_{-\pi}^{\pi} \sum_{n\in\mathbb{Z}} E_t(x-2n\pi-y) f(y)\,dy, \end{aligned}$$

*which strongly suggests that*

$$\frac{1}{2\pi} H_t(x) = \sum_{n\in\mathbb{Z}} E_t(x - 2n\pi)$$

*whence*

$$H_t(x) = 1 + 2\sum_{n=1}^{\infty} e^{-n^2 t}\cos(nx) = \sqrt{\frac{\pi}{t}} \sum_{n\in\mathbb{Z}} e^{-\frac{(x-2n\pi)^2}{4t}}$$

*would follow. How would you prove this? Plot some partial sums.*

[32] Show that $E_t * f \to f$ uniformly if $f$ is uniformly continuous.

## 28.11 Fast Fourier Transform

These are rough notes for a session of JB's numerics course I took over a few years ago. For AUC students as I remember. My first encounter with the subject. Consider an $L$-periodic function $f = f(t)$. Write

$$f(t) = \sum_{k=-\infty}^{\infty} c_k e^{\frac{2\pi kit}{L}}, \quad c_k = \frac{1}{L}\int_0^L f(t)e^{-\frac{2\pi kit}{L}}dt$$

If we only know $f(t)$ in the points $t = 0, \frac{L}{N}, \frac{2L}{N}, \frac{3L}{N}, \ldots$ then the obvious approximation for $c_k$ is

$$\tilde{c}_k = \frac{1}{L}\sum_{j=0}^{N-1} \frac{L}{N} f(\frac{jL}{N})e^{-\frac{2\pi ki\frac{jL}{N}}{L}} = \frac{1}{N}\sum_{j=0}^{N-1} f(\frac{jL}{N})e^{-\frac{2\pi jki}{N}}$$

This involves precisely $N$ samples of $f$.

An approximation of $f(t)$ with $c_k$ replaced by $\tilde{c}_k$ does not make sense since the $\tilde{c}_k$ are $N$-periodic in $k$. Therefore

$$f(t) \not\approx \sum_{k=-\infty}^{\infty} \tilde{c}_k e^{\frac{2\pi kit}{L}}$$

since the latter is not defined. However, the natural finite sum with $c_k$ replaced by $\tilde{c}_k$ does a perfect job in $t = \frac{mL}{N}$ :

$$\sum_{k=0}^{N-1} \tilde{c}_k e^{\frac{2\pi kim\frac{L}{N}}{L}} = \sum_{k=0}^{N-1} \frac{1}{N}\sum_{j=0}^{N-1} f(\frac{jL}{N})e^{-\frac{2\pi jki}{N}}e^{\frac{2\pi kmi}{N}} = f(\frac{mL}{N})$$

This follows by changing the order of summation and the fact that

$$\sum_{k=0}^{N-1} e^{-\frac{2\pi jki}{N}}e^{\frac{2\pi kmi}{N}} = N\delta_{jm}$$

NB. Summing over a sequence $k$ symmetric around $k = 0$ would perhaps look more natural in terms of the usual convergence results for

$$\sum_{k=-n}^{n} c_k e^{\frac{2\pi kit}{L}} \to f(t),$$

as $n \to \infty$. But for the sample points there is no difference.

For MatLab reasons we number the values of $f$ in $t = 0, \frac{L}{N}, \frac{2L}{N}, \frac{3L}{N}, \dots$ as $f_1, f_2, \dots, f_N$ and write $F_k = N\tilde{c}_{k-1}$. The relation between the $N$-vectors $F$ and $f$ is then given by

$$F = \Omega f, \quad \Omega_{jk} = \omega^{(j-1)(k-1)} \quad (j, k = 1, \dots, N), \quad \omega = e^{-\frac{2\pi i}{N}}$$

To go from $f$ to $F$ thus takes $N^2$ multiplications and additions, but this can be improved!

First note that if $N = 2n$ is even, we can set

$$\tilde{\omega} = \omega^2, \quad \tilde{\Omega}_{jk} = \tilde{\omega}^{(j-1)(k-1)} \quad (j, k = 1, \dots, n)$$

and decompose the vector $f$ in $f_o = (f_1, f_3, \dots, f_{N-1})$ and $f_e = (f_2, f_4, \dots, f_N)$. Reshuffling the rows of $\Omega$ accordingly we find for $F^+ = (F_1, F_2, \dots, F_n)$ and $F^- = (F_{n+1}, F_{n+2}, \dots, F_{2n})$ that

$$F^+ = \tilde{\Omega} f_o + D\tilde{\Omega} f_e, \quad F^- = \tilde{\Omega} f_o - D\tilde{\Omega} f_e,$$

in which $D$ is a diagonal matrix with entries $\omega^{j-1}$ with $j = 1, \dots, n$. We used $\omega^n = -1$ here.

If $N = 2^p$ then we can use this reduction to recursively compute $F$ in $p$ steps, each of which contains approximately $2^p$ operations. This is because the lenght of the vectors involved in each step times the number of vectors involved in each step remains the same. This way the order of the number of computations required for $F$ is about $N \log_2 N$.

# 29 The Fredholm alternative

This chapter is about solutions $x$ of $Ax = y$ when $A : X \to X$ is linear and continuous, $X$ a real Banach space, $y \in X$ given, $X$ not finite-dimensional. The Fredholm alternative[1] extends what you may know[2] for square matrices $A$ to more general $A$ of the form $A = I - K$, $K : X \to X$ having a suitable compactness property, $I : X \to X$ the identity map. Perhaps the first simple fact to observe about such operators

$$A = I - K : X \to X$$

in Section 29.1 below is that their *null spaces are finite-dimensional*. This requires a simple lemma with a simple consequence.

**Exercise 29.1.** (Riesz' Lemma) Let $X$ be a real normed space, $L \subsetneq X$ a closed subspace. Show there exists $x_0 \in X$ with $|x_0| = 1$ and[3]

$$d_0 = d(x_0, L) = \inf_{x \in L} |x - x_0| > \frac{1}{2}.$$

Hint: choose $y_0 \in X$, $y_1 \in L$, $d(y_0, L) = 1$, $|y_0 - y_1| > \frac{1}{2}$; try $x_0 = \mu(y_0 - y_1)$.

**Exercise 29.2.** Let $X$ be a real normed space which is not finite dimensional. Use the Riesz' Lemma to exhibit a sequence $x_n$ with $|x_n| = 1$ and $|x_n - x_m| > \frac{1}{2}$. if $m \neq n$. Thus only[4] finite-dimensional normed spaces have the[5] property that every bounded sequence has a convergent subsequence.

## 29.1 Injectivity implies surjectivity

For $I - K$ that is. While reading further observe that we don't use knowledge about finite- versus not finite-dimensional spaces[6] to prove[7] Theorem 29.3, the main result of this section.

[1] Compare the statement in Theorem 29.30 to the similar statement for square matrices.
[2] $Ax = y$ only solvable for $y$ perpendicular to the kernel of the transpose of $A$.
[3] In your proof the number $\frac{1}{2}$ may be replaced by any $\varepsilon \in (0, 1)$.
[4] See Exercise 5.31.
[5] The Heine-Borel property holds in finite dimension only.
[6] Exercise 5.28 and further classify these in terms of compactness.
[7] For the converse of Theorem 29.3 we do: Exercise 29.4.

**Theorem 29.3.** *Let $X$ be a real Banach space, and assume that a linear map $K : X \to X$ has the* (compactness) property that for every bounded sequence $x_n \in X$ the sequence $K(x_n)$ has a convergent subsequence. *Then the implication*

$$N(I-K) = \{0\} \implies R(I-K) = X$$

*holds true. In other words, $I - K$ is surjective if it is injective*[8]*, i.e. if its kernel $N(I-K)$ is trivial. Moreover, not only $I-K$ but also its inverse, which then exists, are continuous if $K$ is such an, as we say, compact linear map.*

**Exercise 29.4.** Much simpler, let $X$ be a real normed space and $K : X \to X$ such a compact linear map. Prove that $N(I-K)$ is finite-dimensional.

Hint: every bounded sequence in $N(I-K)$ has a convergent subsequence.

**Exercise 29.5.** (continued) Suppose that $X_0$ is a closed subspace with[9]

$$X_0 \cap N(I-K) = \{0\}.$$

Prove that there exists a constant $\gamma > 0$ such

$$\gamma|x| \leq |x - K(x)| \quad \text{for all} \quad x \in X_0.$$

Hint: otherwise there exists $x_n \in X_0$ with $|x_n| = 1$ and $x_n - K(x_n) \to 0$.

**Exercise 29.6.** (*continued*) Show that the range of $I-K$ restricted to $X_0$, i.e.

$$R_0 = \{x - K(x) : x \in X_0\},$$

is closed if $X$ is complete.

Hint: show $I-K$ is a bijection between $X_0$ and $R_0$ continuous in both directions.

**Proof of Theorem 29.3** . If not then $X_1 = R(I-K)$ is a closed subspace of $X_1$ with $X_1 \neq X$, and

$$X = X_0 \xrightarrow{I-K} X_1 \xrightarrow{I-K} X_2 \xrightarrow{I-K} X_3 \xrightarrow{I-K} X_4 \xrightarrow{I-K} X_5 \xrightarrow{I-K} \cdots \tag{29.1}$$

[8] Theorem 29.13: no compact linear $K : X \to X$ has $I-K$ surjective but not injective.

[9] Special case: $N(I-K) = \{0\}$, $X_0 = X$ complete.

is a chain of continuous (in both directions) bijections with $X_{n-1} \supsetneq X_n$ for all $n \in \mathbb{N}$. Exercise 29.1 then provides us with a sequence $x_n \in X$ with $x_n \in X_{n-1}$, $d(x_n, X_n) > \frac{1}{2}$ and $|x_n| = 1$. This implies that

$$|K(x_n) - K(x_m)| = |x_n + \underbrace{(I-K)(x_m) - (I-K)(x_n) - x_m}_{\text{in } X_n}| > \frac{1}{2}$$

for $m > n$, contradicting the compactness of $K$. □

## 29.2 The Hahn-Banach property

To deal with $I - K$ not injective we need a tool to get $X_0$ as in Exercise 29.5 such that $X = X_0 + N(I - K)$, in which case we write

$$X = X_0 \oplus N(I - K). \tag{29.2}$$

We formulate this tool using the dual spaces introduced in Theorem 5.39.

**Definition 29.7.** *A real normed space $X$ has the Hahn-Banach property if for every subspace $L$ of $X$ and every $f \in L^*$ there exists $F \in X^*$ with $|F| = |f|$ and $F(x) = f(x)$ for all $x \in X$.*

**Remark 29.8.** *From here on we will always assume that every normed space $X$ under consideration has the Hahn-Banach property.*

**Remark 29.9.** *For what it's worth: the Hahn-Banach property holds in all real normed spaces by virtue of Zorn's Lemma.*

**Remark 29.10.** *If $X$ is a real normed space which contains a sequence such that every point of $X$ is a limit point of that sequence*[10] *then $X$ has the Hahn-Banach property by a direct iterative construction starting from $L$ and $f$.*

**Exercise 29.11.** The Hahn-Banach property reformulated: show that if $L$ is a closed subspace of a normed space $X$ and $x_0$ in $X$ is not in $L$, there exists

$$F \in X^* \quad \text{with} \quad F(x) = 0 \quad \text{for all} \quad x \in L, \tag{29.3}$$

$$F(x_0) = d_0 = \inf_{x \in L} |x - x_0| > 0 \quad \text{and} \quad |F| = 1.$$

Hint: apply Definition 29.7 to $f$ with $f(x) = 0$ for all $x \in L$ and $f(x_0) = d_0$.

[10] $X$ is then called separable.

**Exercise 29.12.** Complementing $N(I-K)$ to get (29.2). Assume that $N(I-K)$ has dimension $d \in \mathbb{N}$. Thus, as in Exercise 5.28, there are $e_1, \ldots, e_d$ in $N(I-K)$ such that every $x \in N(I-K)$ is uniquely written as

$$x = \xi_1 e_1 + \cdots + \xi_d e_d, \quad \xi_1, \ldots, \xi_d \in \mathbb{R}.$$

Define $f_1, \ldots, f_d \in N(I-K)^*$ by

$$f_i e_j = \delta_{ij} = \begin{matrix} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{matrix}$$

to obtain extensions $F_1, \ldots, F_d \in X^*$ via the Hahn-Banach property, and let

$$X_0 = N(F_1) \cap \cdots \cap N(F_d).$$

Show that every $x \in X$ is uniquely written as $x = x_0 + x_1$ with $x_0 \in X_0$ and $x_0 \in N(I-K)$ to conclude that

$$X = X_0 \oplus N(I-K),$$

a decomposition of $X$ as the direct sum of the closed subspaces $X_0$ and $N(I-K)$.

## 29.3 Surjectivity implies injectivity

For $I-K$ that is. It's here that we first need that the identity $I$ itself is not compact, unless $X$ is finite-dimensional.

**Theorem 29.13.** *Let $X$ be a Banach space, and $K : X \to X$ linear and compact. Then the implication*

$$R(I-K) = X \implies N(I-K) = \{0\}$$

*and thereby, with Theorem 29.3, the equivalence*

$$R(I-K) = X \iff N(I-K) = \{0\}$$

*holds true: $I-K$ is injective*[11] *if and only if $I-K$ is surjective.*

**Proof of Theorem 29.13.** Assume that $I-K$ is surjective and that in

$$\begin{array}{rl} X = N(I-K) \oplus & X_0 \\ & \downarrow I-K \\ & X \end{array}$$

[11] We need the dual space $X^*$ and the adjoint of $A = I-K$ to continue if it is not.

the finite-dimensional kernel $N(I-K)$ is non-trivial. Then

$$X_0 \xrightarrow{I-K} X$$

is a bijection, continuous in both directions by Exercise 29.5. We then obtain a never ending chain[12]

$$\cdots \xrightarrow{I-K} N_4 \xrightarrow{I-K} N_3 \xrightarrow{I-K} N_2 \xrightarrow{I-K} N_1 = N(I-K)$$

of bijections and a chain of surjective maps

$$\cdots \xrightarrow{I-K} N_1 \oplus N_2 \oplus N_3 \xrightarrow{I-K} N_1 \oplus N_2 \xrightarrow{I-K} N_1 \tag{29.4}$$

with increasing dimensions

$$1 \leq d = \dim(N_1) < 2d = \dim(N_1 \oplus N_2) < 3d = \dim(N_1 \oplus N_2 \oplus N_3) < \cdots,$$

allowing the same reasoning as in the proof of Theorem 29.3 for (29.1), with

$$x_n \in N_1 \oplus \cdots \oplus N_n, \quad |x_n| = 1, \quad d(x_n, N_1 \oplus \cdots \oplus N_{n-1}) > \frac{1}{2}$$

to obtain a contradiction. $\square$

## 29.4 Annihilators

**Exercise 29.14.** Annihilators. Let $X$ be a real normed space and $L \subset X$. Show that for every $L \subset X$ the *annihilator*[13]

$$L^0 = \{f \in X^* : \forall_{x \in L}\ f(x) = 0\}$$

is a closed subspace of $X^*$. Remember the definition of $L^0$ also as

$$f \in L^0 \iff f(x) = 0 \quad \text{for all} \quad x \in L.$$

**Exercise 29.15.** (*continued*) Let $X$ be a real normed space and let $M \subset X^*$. Show that

$${}^0M = \{x \in X : \forall_{f \in M}\ f(x) = 0\}$$

is a closed subspace of $X$, and that by definition

$${}^0M = X \iff M = \{0\}.$$

[12] If the theorem were false that is.

[13] So $F$ in (29.3) is in $L^0$.

**Theorem 29.16.** *Let $X$ be a real normed space, and let $L \subset X$ be a subspace. Then the closure of $L$ is given by*[14]

$$\bar{L} = {}^0(L^0). \tag{29.5}$$

**Exercise 29.17.** By definition $L \subset {}^0(L^0)$ and therefore $\bar{L} \subset {}^0(L^0)$ because ${}^0(L^0)$ is closed. Prove Theorem 29.16 by assuming there exists $x_0$ in ${}^0(L^0)$ but not in $\bar{L}$.

Hint: use the Hahn-Banach property via Exercise 29.11 to get a contradiction.

**Exercise 29.18.** More[15] about annihilators. By definition $\{0\}^0 = X^*$ and $X^0 = \{0\}$. Use the Hahn-Banach property to show that

$$L = \{0\} \iff L^0 = X^*,$$

and also that

$$M = X^* \implies {}^0M = \{0\}.$$

**Remark 29.19.** *We note that $L^0 = \{0\}$ is equivalent to*

$$\forall_{f \in X^*}: \quad f \neq 0 \implies \exists x \in L \quad f(x) \neq 0,$$

*and that ${}^0M = \{0\}$ is equivalent to*

$$\forall_{x \in X}: \quad x \neq 0 \implies \exists f \in M \quad f(x) \neq 0.$$

*These minimality statements for $L^0$ and ${}^0M$ may be interpreted as separation statements for $L$ with respect to $X^*$ and $M$ with respect to $X$, but do not imply that $L$ or $M$ is maximal.*

**Remark 29.20.** *Each of two minimality statements for $L$ and $M$, namely $L = \{0\}$ and $M = \{0\}$ trivially implies by definition the maximality of $L^0$ or ${}^0M$.*

**Remark 29.21.** *Each of the four maximality statements for $L, {}^0M, M, L^0$ implies minimality of the corresponding $L^0, M, {}^0M, L$. When using maximality statements in $X^*$ the Hahn-Banach property is used in the proof. When using maximality statements in $X$ only the definition is used.*

[14] You can now jump to Definition 29.22 and skip what's next about annihilators.

[15] Not needed in relation to $I - K$.

## 29.5 Adjoints and finite-rank perturbations

**Definition 29.22.** *Let $X$ and $Y$ be real normed spaces, and let $A : X \to Y$ be a linear map*[16]*. We define the* adjoint operator *$A^*$ by*

$$A^*g = g \circ A, \quad i.e. \quad (A^*g)(x) = g(A(x)) \quad \text{for all} \quad x \in X, g \in Y^*.$$

**Exercise 29.23.** Let $X$ and $Y$ be real normed spaces and assume that $A : X \to Y$ is linear and continuous. Explain why the kernel

$$N(A) = \{x \in X : A(x) = 0\}$$

of $A$ is a closed linear subspace of $X$.

**Exercise 29.24.** (*continued*). Denoting the range $A$ in $Y$ by

$$R(A) = \{Ax : x \in X\},$$

explain why the kernel of $A^*$ is given by

$$N(A^*) = \{g \in Y^* : A^*(g) = 0 \in X^*\} = R(A)^0.$$

**Exercise 29.25.** (*continued*) Use the Hahn-Banach property in $Y$ to show that the closure of the range $R(A)$ is the subspace

$$\overline{R(A)} = {}^0N(A^*). \tag{29.6}$$

**Remark 29.26.** Solvability condition for $x - K(x) = y$. *From (29.6) with $A = I - K$, $K : X \to X$ compact linear, $X$ a Banach space, we find*

$$R(I - K) = {}^0N(I^* - K^*),$$

*so the equation*

$$x - K(x) = y$$

*for $x$ is solvable if and only if $y \in {}^0N(I^* - K^*)$. In what follows we avoid invoking the compactness*[17] *of $K^* : X^* \to X^*$, when we consider the different possibilities for $N(I^* - K^*)$ in case $N(I - K)$ is nontrivial*[18].

[16] Special case: $X = Y$, $A = I - K$, $K$ compact.

[17] Not so easy, an easier statement we will also not use is that $I^*$ is the identity on $X^*$.

[18] Otherwise $R(I - K) = X$ and thereby also $N(I^* - K^*)$ is trivial.

**Exercise 29.27.** Modification of $I - K$ in Exercise 29.12. Assume $g_1, \dots, g_{d+1}$ exist in $N(I^* - K^*)$ and $x_1, \dots, x_{d+1}$ in $X$ such that $g_i(x_j) = \delta_{ij}$. In particular $x_i \notin {}^0N(I^* - K^*) = R(I - K)$. Modify the operator $A = I - K$ on $N(A)$ by sending each $e_i$, $i = 1, \dots, n$, to $x_i$ in $X$, rather than to $0$ in $X$, but don't modify $A$ on $X_0$. This gives what we call a *finite rank perturbation of $A$. This modification* $\tilde{A}$ of $A$ is as in Theorem 29.3, with a modification $\tilde{K}$, which in turn is a finite rank perturbation of the original $K$. Verify that $\tilde{K}$ is still compact, $N(I - K) = \{0\}$, but $x_{d+1} \notin {}^0N(I^* - \tilde{K}^*) = R(I - \tilde{K})$, in contradiction with Theorem 29.3.

**Exercise 29.28.** (*continued*) Conclude that the kernel of $N(I^* - K^*)$ is finite-dimensional, and that its dimension is at most the dimension of $N(I - K)$.

**Exercise 29.29.** Getting the dimension of $N(I^* - K^*)$ right. Suppose the dimension of the by now finite-dimensional kernel of $N(I^* - K^*)$ is smaller than the dimension of $N(I - K)$. If $N(I^* - K^*)$ is nontrivial we have again $g_1, \dots, g_k$ in $N(I^* - K^*)$ for some $1 \leq k < d$ with $x_1, \dots, x_k \in X$ such that $g_i(x_j) = \delta_{ij}$. Verify that

$$X = R(I - K) \oplus [x_1] \oplus \cdots \oplus [x_k],$$

and modify $A$ by sending each $e_i$ to $x_i$, to obtain a modification $\tilde{A} = I - \tilde{K}$ which is surjective, but has a nontrivial kernel spanned by the remaining $e_i$. We now have a contradiction because Theorem 29.13 said there is in fact no such compact linear $K : X \to X$ with $I - K$ surjective but not injective. This completes the proof of the Fredholm alternative for $I - K$.

**Theorem 29.30.** *(Fredholm Alternative) Let $X$ be a Banach space and let $K : X \to X$ be linear and compact. Then both $N(I - K)$ and $N(I^* - K^*)$ are finite-dimensional with the same dimension $d$. If the kernels are trivial then $A = I - K$ has a continuous linear inverse. If the kernels are nontrivial then*

$$A(x) = x - K(x) = y$$

*is solvable if and only if $y$ is in ${}^0N(A^*) = {}^0N(I^* - K^*)$.*

## 29.6 Fredholm operators

If we define the operator $B$ to be the inverse of

$$A = I - K : X_0 \to R(A)$$

on $R(A) = R(I - K)$, and just $0 \in X$ on $L$ then

$$N(A) \oplus R(B) = X = R(A) \oplus N(B),$$

with $N(A)$ and $N(B)$ of the same finite dimension, $R(A)$ and $R(B)$ closed.

This framework generalises to Banach spaces $X, Y$, operators $A \in L(X, Y)$, $B \in L(Y, X)$, kernels $N(A)$, $N(B)$ of finite dimension, closed ranges $R(A)$, $R(B)$, with similar decompositions of $X$ and $Y$. If in the scheme

$$\begin{aligned} X = \underbrace{N(A)}_{\substack{\downarrow \\ 0 \in Y}} \oplus \underbrace{R(B)}_{A\downarrow} \qquad & \\ \overbrace{R(A)}^{\uparrow B} \oplus \overbrace{N(B)}^{\substack{0 \in X \\ \uparrow}} &= Y \end{aligned}$$

the operators $A$ and $B$ are each others inverses on the closed ranges, that is, modulo the finite-dimensional kernels, then $(A, B)$ is called a Fredholm pair of Fredholm operators $A$ and $B$.

## 29.7 More on adjoints for later perhaps

**Exercise 29.31.** (*Continuity of the adjoint*) Let $X$ and $Y$ be real normed spaces, and let $A \in L(X, Y)$. Prove that $A^* \in L(Y^*, X^*)$ and $|A^*| \leq |A|$.

**Exercise 29.32.** (*continued*) Show that $|A^*| = |A|$ if the space $Y$ has the Hahn-Banach property.

Hint: choose $x_\varepsilon \in X$ with $|x_\varepsilon| = 1$ such that $|A(x_\varepsilon)|$ is $\varepsilon$-close to $|A|$, and then use Definition 29.7 with $L = \{0\} \subset Y$ and $x_0 = A(x_\varepsilon) \notin L$.

# 30 Some real Hilbert space theory

This chapter is in part a set of more or less *do it yourself* notes. But first we recall the real Banach space $C([a,b])$. This first function space one encounters is *not* a Hilbert space, because the inner product defined by

$$x \cdot y = \int_a^b x(t)y(t)\,dt$$

for $x, y \in C([a,b]$ does not render $C([a,b])$ complete. In Section 30.1 the spectral Theorem 30.10 for compact symmetric linear operators is formulated to also apply to such spaces. It generalises the statements from linear algebra for symmetric real $n$ by $n$ matrices about eigenvalues and eigenvectors. Of course the treatment also applies to the case that the inner product space $V$ is just $\mathbb{R}^2$. The only difference then is that the trick below with (30.9) and the generalised Cauchy-Schwarz inequality is not needed because the unit circle is compact. In fact you may then like to use polar coordinates to verify the statements from a calculus perspective. The example $V = \mathbb{R}^2$ is also natural for considering projections on closed convex sets in Section 30.2. This leads to the Riesz Representation Theorem that says that in Hilbert spaces the continuous linear functions are all of the form $x \to a \cdot x$, just as in $V = \mathbb{R}^2$.

A real Hilbert space $H$ is a real vector space with an inner product

$$(x, y) \in H \times H \to (x, y)_H = x \cdot y$$

in which Cauchy sequences are convergent. In terms of the inner product this completeness property says that if a sequence $x_n$ in $H$ has the property that

$$(x_n - x_m) \cdot (x_n - x_m) \to 0$$

as $m, n \to \infty$, then there must exist a (unique) $\overline{x} \in H$ 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*[1] on $H$. Subscripts $H$ will be dropped, unless they are needed to avoid confusion.

[1]See Chapter 5.

**Exercise 30.1.** Derive and prove the Cauchy-Schwarz[2] 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.

$$|x + y|^2 + |x - y|^2 = 2\,|x|^2 + 2\,|y|^2.$$

**Exercise 30.2.** Let $V$ be a complex vector space with an complex inner product, so $w \cdot z = \overline{z \cdot w}$ and $(\lambda z) \cdot w = \lambda\,(z \cdot w) = z \cdot \overline{\lambda} w$ for all $\lambda \in \mathbb{C}$ and all $z, w \in V$, we have, assuming that $|z|^2 = z \cdot z = 1$, that

$$\begin{aligned} 0 \leq (\lambda z + w) \cdot (\lambda z + w) &= \lambda\overline{\lambda} + \lambda z \cdot w + w \cdot \lambda z + w \cdot w \\ &= \lambda\,(\overline{\lambda} + z \cdot w) + \overline{\lambda}\,w \cdot z + w \cdot w \\ &= (\lambda + w \cdot z)\,(\overline{\lambda} + z \cdot w) - w \cdot z\, z \cdot w + w \cdot w \\ & |\lambda + w \cdot z|^2 - |z \cdot w|^2 + |w|^2 \end{aligned}$$

Prove for all $z, w \in V$ that

$$|z \cdot w| \leq |z|\,|w|,$$

with equality only if $w = \lambda z$ for some $\lambda \in \mathbb{C}$ or $z = \mu w$ for some $\mu \in \mathbb{C}$. Also show that

$$|z + w|^2 + |z - w|^2 = 2\,|z|^2 + 2\,|w|^2.$$

## 30.1 Compact symmetric linear operators

Let $V$ be a real vector space with an inner product denoted by $x \cdot y$ for $x, y \in V$, and let $S : V \to V$ be a continuous linear map which is symmetric in the sense that

$$Sx \cdot y = x \cdot Sy$$

for all $x, y \in V$. We say that $S$ is nonnegative if in addition

$$Sx \cdot x \geq 0$$

[2] See Exercise 30.3 below if you forgot the proof given in your linear algebra course.

for all $x \in V$. If in addition $Sx \cdot x = 0$ only occurs for $x = 0$ then $S$ is called positive. In this section we do not assume that Cauchy sequences in $V$ are convergent. Continuity for linear maps is equivalent to the operator norm[3]

$$|S|_{op} = \sup_{0 \neq x \in V} \frac{|Sx|}{|x|} = \sup_{0 \neq x \in V} \sqrt{\frac{Sx \cdot Sx}{x \cdot x}} = \sup_{x \cdot x = 1} \sqrt{Sx \cdot Sx} \tag{30.7}$$

being finite. From here on we shall call $S$ a symmetric bounded[4] linear operator.

**Exercise 30.3.** Derive the Cauchy-Schwarz inequality for $x, y \in V$ by inspection of the minimum of the nonnegative function

$$\lambda \xrightarrow{q} (\lambda y - x) \cdot (\lambda y - x),$$

and show that the same reasoning leads to a generalised Cauchy-Schwarz inequality, namely

$$|Sx \cdot y| \leq \sqrt{Sx \cdot x}\,\sqrt{Sy \cdot y}$$

for all nonnegative symmetric bounded linear operators $S : V \to V$ and all $x, y \in V$. Don't forget the possibility that the function $q$ is a linear.

**Exercise 30.4.** Use the method in Exercise 30.2 to generalise to complex inner product spaces $V$ and $S : V \to V$ continuous linear, with $S$ symmetric[5] in the sense that

$$Sx \cdot y = x \cdot Sy = \overline{Sy \cdot x}$$

for all $x, y \in V$, and nonnegative in the sense that

$$Sx \cdot x \geq 0$$

for all $x \in V$.

**Exercise 30.5.** Let $S : V \to V$ be a symmetric continuous linear operator. Use the Cauchy-Schwarz inequality and the definition of the operator norm to show that

$$|Sx \cdot x| \leq |S|_{op}\, x \cdot x, \quad \text{whence} \quad M = \sup_{0 \neq x \in V} \frac{|Sx \cdot x|}{x \cdot x} = \sup_{x \cdot x = 1} |Sx \cdot x| \leq |S|_{op}. \tag{30.8}$$

[3]This terminology was introduced for matrices in Section 18.2, see also Remark 11.15.
[4]It is is bounded on bounded sets.
[5]The more common term is self-adjoint

Then write

$$4\,Sx \cdot y = S(x+y)\cdot(x+y) - S(x-y)\cdot(x-y)$$

and estimate the right hand side using first the triangle inequality, then twice (30.8), and finally the parallellogram law. For all $x, y \in V$ with $|x| = |y| = 1$ this should give that $|Sx \cdot y| \leq M$. Explain why thus $|Sx \cdot y| \leq M|x|\,|y|$ for all $x, y \in V$ and choose $y = Sx$ to conclude that $|S|_{op} = M$.

**Exercise 30.6.** Generalise Exercise 30.5 to complex inner product spaces $V$ and $S : V \to V$ continuous linear and symmetric.

The map

$$x \to Q(x) = Sx \cdot x$$

defined by $S$ is called a quadratic form. Exercise 30.5 states the remarkable fact that the suprema of $x \to |Q(x)|$ and $x \to |Sx|$ on the unit ball coincide.

Ignoring the trivial case that $M = 0$ we next observe[6] that the generalised Cauchy-Schwarz inequality in Exercise 30.3 also holds with $S$ replaced by $M - S = MI - S$, $I$ being the identity map. Thus it holds that

$$|(M-S)x \cdot w| \leq \sqrt{(M-S)x \cdot x}\,\sqrt{(M-S)w \cdot w},$$

whence, varying $w$ over the unit ball, we have

$$|(M-S)x| \leq \sqrt{(M-S)x \cdot x}\,\sqrt{|M-S|_{op}}. \tag{30.9}$$

Now take a sequence $x_n \in V$ with $|x_n| = 1$ and $Sx_n \cdot x_n \to \pm M$. In case $Sx_n \cdot x_n \to M$ it follows that $(M-S)x_n \cdot x_n = M - Sx_n \cdot x_n \to 0$, and thus

$$Mx_n - Sx_n \to 0$$

by (30.9). If the sequence $x_n$ can be chosen to have $Sx_n$ converging to a limit $y \in V$, it follows that also $Mx_n \to y$ and that $M = |y| > 0$. But then $w = \frac{y}{M}$ is a unit eigenvector of $S$ with eigenvalue $M$. In case $Sx_n \cdot x_n \to -M$ we replace $S$ by $-S$ and apply the same reasoning. We have thereby proved the following theorem.

**Theorem 30.7.** *Let $V$ be a real inner product space and $S : V \to V$ linear, symmetric, $Sx \neq 0$ for at least one $x \in V$. If for every bounded sequence $x_n$ in $V$ it holds that $Sx_n$ has a convergent subsequence, then*

$$M_1 = \max_{0 \neq x \in V} \frac{|Sx \cdot x|}{x \cdot x} > 0 \tag{30.10}$$

[6] Following Evans' treatment in one of the appendices to his PDE book.

*exists, and $M_1$ or $-M_1$ (possibly both) is an eigenvalue $\lambda_1$ of $S$. The corresponding eigenvectors are precisely the maximizers*[7] *of the quotient under consideration.*

**Exercise 30.8.** Generalise Theorem 30.7 to the case of complex inner product spaces $V$ and $S : V \to V$ linear and symmetric with the same compactness property.

**Remark 30.9.** *Let $V$ and $W$ be normed spaces. Then a linear operator $S : V \to W$ is called compact if for every bounded sequence $x_n$ in $V$ it holds that $Sx_n$ has a convergent subsequence in $W$. You should have no difficulty proving that such operators are bounded.*

Given such an eigenvector $v_1$ with $|v_1| = 1$ it easily follows that $S$ maps

$$V_1 = \{x \in V : x \cdot v_1 = 0\}$$

to itself. Unless $V_1$ is[8] the null space of $S$ it then follows that

$$M_2 = \max_{x \cdot v_1 = 0 \neq x \in V} \frac{|Sx \cdot x|}{x \cdot x} \neq 0 \tag{30.11}$$

is also the absolute value of an eigenvalue $\lambda_2$ of $S$, with eigenvector $v_2$ with $|v_2| = 1$.

Repeating the argument with

$$V_2 = \{x \in V : x \cdot v_1 = x \cdot v_2 = 0\}$$

we obtain a sequence of eigenvalues

$$|\lambda_1| \geq |\lambda_2| \geq \cdots > 0,$$

which either terminates[9], or has the property that $\lambda_n \to 0$ as $n \to \infty$. The latter statement is a consequence of the compactness assumption: the corresponding mutually perpendicular unit eigenvectors

$$v_1, v_2, \ldots,$$

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. We have proved:

[7] Genericly only multiples of one eigenvector.

[8] Here we include the possibility that $V_1 = \{0\}$.

[9] If the range of $V$ is spanned by $v_1, \ldots, v_N$ for some $N \in \mathbb{N}$.

**Theorem 30.10.** *Let $V$ be a real or complex inner product vector space. If the null space of a compact symmetric linear operator $S : V \to V$ is trivial then $V$ has an orthonormal (Hilbert) basis consisting of eigenvectors $v_1, v_2, \ldots$ of $S$, obtained as maximizers of (30.10), (30.11), .....; this statement applies also to $S$ restricted to*

$$\{x \in V : Sx = 0\}^\perp$$

*if the null space of $S$ is not trivial.*

## 30.2 Projections on closed convex sets

This section is do it yourself. Take $H = \mathbb{R}^2$ first and draw pictures to see what should be true for general closed convex sets. The special case that $K$ is a line through the origin should be familiar. We conclude with a general statement about parabola's.

**Exercise 30.11.** Let $H$ be a Hilbert space, $K \subset H$ a non-empty closed convex[10] subset, and $a \in H$. Show there exists a unique $p \in K$ that realises the distance

$$|p - a| = \inf_{x \in K} |x - a| = d(a, K)$$

from $a$ to $K$ as a minimum, 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. Consider first the case that $a \notin K$ and verify the not so interesting case that $a \in K$.

**Exercise 30.12.** 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 $a, b \in H$. In other words, it is Lipschitz continuous with Lipschitz constant 1. Hint: take $p = P_K(b)$ in the characterisation in Exercise 30.11 for $P_K(a)$ and $p = P_K(a)$ for $P_K(b)$.

**Exercise 30.13.** Use Exercise 30.2 to generalise to complex Hilbert spaces.

**Exercise 30.14.** Let $H$ be a Hilbert space, $L \subset H$ a closed linear subspace. Prove that $P_L : H \to L$ linear, and that

$$M = N(P_L) = \{x \in H : P_L(x) = 0\} = L^\perp = \{x \in H : x \cdot y = 0 \; \forall y \in L\},$$

[10] If $a, b \in K$ then $[a, b] = \{ta + (1 - t)b : t \in [0, 1]\} \subset K$.

the *null space* of $P_L$, is a closed linear subspace with $M \cap L = \{0\}$. Show that $M + L = H$ and conclude that every $x \in H$ is uniquely written as $x = p + q$ with $p \in L$ and $q \in M$. Notation: $L \oplus M = H$.

**Exercise 30.15.** 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 $K$. Use Exercise 30.11 to prove this statement.

## 30.3 Riesz representation of linear Lipschitz functions

On $\mathbb{R}^2$ all linear functions are continuous. It is a remarkable fact that this statement is[11] false for every infinite-dimensional normed space. For Hilbert spaces we have the Riesz Representation Theorem below, which is proved and put in some pespective in the exercises that follow. Lipschitz continuity is the natural concept here.

**Theorem 30.16.** *Let $H$ be a Hilbert space. The continuous linear functions on $H$ are precisely the functions $f : H \to \mathbb{R}$ of the form $f(x) = a \cdot x$ with $a \in H$. Such an $a$ is called the Riesz representation of $f$, notation $a = R_H(f)$.*

**Exercise 30.17.** Let $X$ be a normed[12] 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 for $f$. 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 $X = X$.

[11] In fact it is false if and only if the normed space is infinite-dimensional.

[12] See Exercise 5.26 for a definition.

The result in Exercise 30.17 is only of interest if there are such Lipschitz Lipschitz continuous functions on $X$. The dual space $X^*$ is by definition the space of all Lipschitz continuous *linear* functions from $X$ to $\mathbb{R}$. For $f : X \to \mathbb{R}$ linear and $x \in X$ the notation

$$\langle f, x \rangle = f(x)$$

is common. The brackets denote what is called the duality between $X^*$ and $X$ and are not to be confused with inner product brackets, although Theorem 30.16 does tempt us to do so.

In case of $X = H$ a Hilbert space every $a \in H$ defines a linear $\phi_a$ in $Lip_0(H)$ by

$$\phi_a(x) = a \cdot x,$$

with smallest Lipschitz constant $|a|$. Thus $a \to \phi_a$ defines map

$$\Phi := H \to Lip_0(H).$$

and the range of $\Phi$ is contained in $H^*$, the (normed) space of all Lipschitz continuous *linear* functions $f : H \to \mathbb{R}$.

**Exercise 30.18.** 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 all $x, x_1, x_2 \in H$, and that $[\Phi(x)]_{Lip} = |x|$. Thus $\Phi$ is linear and norm preserving.

Is $\Phi$ surjective, i.e. is every $f \in H^*$ of the form $\phi_a$? Towards a positive answer we consider[13] its null space

$$N_f = \{x \in H : f(x) = 0\}.$$

**Exercise 30.19.** Show that $N_f \subset H$ is a closed linear subspace.

**Exercise 30.20.** By 30.14 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 $f(x) = f(e)e \cdot x$.

[13] We write $N_f$ instead of $N(f)$, to distinguish between $f$ and $P_L$.

**Exercise 30.21.** Riesz Representation Theorem restated: explain why Exercise 30.19 says that $\Phi : H \to H^*$ a linear isometry. The inverse of $\Phi$ is called the Riesz representation of $H^*$. We denote the inverse of $\Phi$ by $R_H$, and its domain is $H^* \subsetneq Lip_0(H)$.

**Exercise 30.22.** Show that there are many nonlinear functions in $Lip_0(H)$. Hint: use 30.11.

## 30.4 Bilinear forms and the Lax-Milgram theorem

This section modifies the approach in Evans' PDE book. We use $u$ and $v$ to denote elements in a Hilbert space $H$, greek letters for elements of the dual space $H^*$, and establish a generalisation of the Riesz Representation Theorem 30.16 and its restatement in Exercise 30.21.

**Theorem 30.23.** *Let $H$ be a Hilbert space and $B : H \times H \to \mathbb{R}$ be a bounded coercive bilinear form. This means that*

*(a) for every $u \in H$ fixed the map $v \to B(u,v)$ is linear;*
*(b) for every $v \in H$ fixed the map $u \to B(u,v)$ 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*

$$H^* \ni \phi \xrightarrow{S} u \in H$$

*with $|S| \leq \frac{1}{\beta}$, which is the inverse of the continuous linear map*

$$H \ni u \xrightarrow{A} \phi \in H^*$$

*defined by*

$$\langle \phi, v \rangle = \langle Au, v \rangle = B(u,v) \quad \text{for all} \quad v \in H, \tag{30.12}$$

*which has $|A| \leq \alpha$.*

**Proof.** Observe that (30.12) 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

$$Au : H \to \mathrm{I\!R}$$

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

$$\beta|u| \le |Au| \le \alpha|u| \tag{30.13}$$

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 $\Phi$ be as in the Riesz Representation Theorem[14] and let

$$L = \Phi^{-1}(R(A) \subset H.$$

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 $v \ne 0$. 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 30.23. $\square$

[14]Exercise 30.21.

## 30.5 Hilbert spaces in disguise

In Section 15.6 we mentioned that Theorem 15.10, the Morse lemma, is not restricted to the case that $X$ is a Hilbert space in disguise[15]. But if we start the reasoning in Section 30.4 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 30.24.** *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 the map $v \xrightarrow{\phi} B(u,v)$ is linear;*
*(b) for every $v \in X$ fixed the map $u \xrightarrow{\psi} B(u,v)$ 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 30.25.** *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 30.26 below.*

**Theorem 30.26.** *Every bounded bilinear form[16] on a normed space $X$ is of the form*

$$(u,v) \to B(u,v) = \langle Au, v \rangle \in \mathbb{R} \tag{30.14}$$

[15] A complete metric vector space which allows an equivalent inner product norm.
[16] See Section 15.2 for the second derivative as an example of the symmetric case.

*with* $A \in L(X, X^*)$, *and*[17]

$$\sup_{u,v\in X\setminus\{0\}} \frac{|B(u,v)|}{|u|\,|v|} = |A|. \tag{30.15}$$

*If* $X$ *is complete and* $B$ *is coercive then* $X$ *is a Hilbert space in disguise, and* $A$ *is a bijection*[18] *between* $X$ *and* $X^*$ *with*

$$\beta\,|u| \le |Au| \le \alpha|u|$$

*for all* $u \in X$, $0 < \beta \le \alpha$, *as in Definition 30.24.*

**Proof.** We use (a) again to define $A$ by $Au = \phi$, so (30.14) 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

$$|Au| \le \alpha\,|u|, \tag{30.16}$$

and (b) implies that $A : X \to X^*$ is linear. We conclude that $A \in L(X, X^*)$ and $|A| \le \alpha$.

**Exercise 30.27.** Prove (30.15) 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 (30.13) holds and

$$X \xrightarrow{A} R(A) = \{Au : u \in X\}$$

[17] The norms have subscripts that we omit in this section.

[18] 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 $\phi$.

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

$$^0R(A) = \{v \in X : \forall_{\phi \in R(A)}\ \phi(v) = 0\} = \{v \in X : \forall_{u \in X}\ B(u,v) = 0\}.$$

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[19] 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 30.26. □

## 30.6 The standard Hilbert space

We review the standard example of a real Hilbert space.

**Exercise 30.28.** Show that

$$l^{(2)} = \{x = (x_1, x_2, x_3, \dots) : x_n \text{ is a sequence in } \mathbb{R},\ \sum_{n=1}^{\infty} x_n^2 < \infty\}$$

is a Hilbert space with respect to the inner product defined by

$$x \cdot y = \sum_{n=1}^{\infty} x_n y_n.$$

We wrote

$$x = \sum_{n=1}^{\infty} x_n e_n, \quad e_1 = (1,0,0,\dots), \quad e_2 = (0,1,0,\dots), \dots,$$

but we often prefer a notation with column vectors instead.

Every infinite dimensional separable[20] 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

[19] Holds for reflexive spaces, spaces $X$ for which $(X^*)^* = \{f \to \langle f, x\rangle : x \in X\}$.

[20] See Section 5.6, this means that $H$ contains a sequence $a_n$ as in what follows.

element in $H$ is a limit point of this sequence. We apply the Gramm-Schmidt procedure. Let

$$e_1 = \frac{1}{|a_1|} a_1$$

if $a_1 \neq 0$, otherwise throw $a_1$ away, renumber the sequence. Repeat until you have $a_1 \neq 0$ and $e_1$ as above. 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

$$H = \{x = \sum_{n=1}^{\infty} x_n e_n : , x_n \text{ a sequence in } \mathbb{R}, \ \sum_{n=1}^{\infty} x_n^2 < \infty\}.$$

**Remark 30.29.** *We thus showed that 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 Exercise 30.30 below.*

We may view $l^{(2)}$ as $H = l^{(2)} = L^2(\mathbb{N})$ with the counting measure on $\mathbb{N}$. Then elements $u$ in $H$ 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 put these in a column vector

$$u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{pmatrix}.$$

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

$$u \cdot v = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \end{pmatrix} = u_1v_1 + u_2v_2 + u_3v_3 + \cdots = \sum_{k=1}^{\infty} u_k v_k = (u, v)_H.$$

This inner product is the integral of the product function $uv$ with respect to the counting measure on $\mathbb{N}$. In general $uv$ is not[21] in $l^{(2)} = L^2(\mathbb{N})$.

**Exercise 30.30.** 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.

## 30.7 Other inner products

The examples below derive from the observation that the counting measure is not the only measure on $\mathbb{N}$: every sequence of positive numbers

$$0 < \lambda_1 \leq \lambda_2 \leq \lambda_2 \leq \cdots \tag{30.17}$$

definines a measure on $\mathbb{N}$ by assigning measure $\lambda_n$ 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 some (maximal) subspace $V$ of our standard space $H = l^{(2)}$. This subspace is not closed in $H$ if $\lambda_n \to \infty$.

**Exercise 30.31.** Why not? Assume that (30.17) holds. Show that $V$ with $((\cdot, \cdot))$ is a Hilbert space, and that $V = H$ if and only if $\lambda_n$ is a bounded sequence.

So $V \subset H$, and the norm on $V$ is given by

$$|u|_V = ||u|| = \sqrt{\sum_{n=1}^{\infty} \lambda_n u_n^2},$$

[21]Many function spaces are not algebra's and this is one of them.

whence

$$||u||^2 \geq \lambda_1 |u|^2$$

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 30.32.** Check that $\overline{i(V)} = \overline{V} = H$.

## 30.8 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

$$\phi(v) = (f, v) = ((u, v)) \quad \forall v \in V.$$

Thus the linear continuous functions

$$V \ni v \xrightarrow{f\in H} (f, v)_H \in \mathbb{R}$$

and

$$V \ni v \xrightarrow{u\in V} (u, v)_V \in \mathbb{R}$$

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 30.33.** Assume $0 < \lambda_1 \leq \lambda_2, \ldots$ is unbounded. Why is not every continuous linear $\psi : V \to \mathbb{R}$ of the form $\phi(v) = (f, v)$ with $f \in H$?

If this nondecreasing sequence $\lambda_n > 0$ is unbounded it is easier 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 $R_H$ as defined in Exercise 30.21, then

$$V \subsetneq H = H^* \subsetneq V^*,$$

which then conflicts with an identification of $V^*$ and $V$ via $R_V$.

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 30.34.** Prove that $i^* : H^* \to V^*$ is linear and continuous. Hint: consider the norm of $i^*(\phi) = \phi \circ i$.

## 30.9 A more general abstract perspective

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 above is much the same. We do not need[22] to assume that $V \subset H$. It is instructive to see how the injectivity of $i : V \to H$ and the density of its range being dense come into play.

**Exercise 30.35.** 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

$$(u, v)_V = (f, i(v))_H$$

[22]In applications to elliptic boundary values problems we do have $V \subset H$.

for all $v \in H$ is given by

$$S = R_V \circ i^* \circ (R_H)^{-1},$$

and has norm $|S| = |i^*|$.

**Remark 30.36.** *We think of $S$ as a solution operator. If the inner product on $V$ is replaced by a nonsymmetric coercive bilinear form, the Lax-Milgram theorem replaces the Riesz Representation theorem. In Section 30.4 we discussed why this approach still requires a Hilbert space setting. In Section 30.9 we consider the operator $S$ in Exercise 30.35 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 30.42 below and how they are used in Exercise 30.44. They all relate to the solution operator, but one of them does not need the solution operator.*

**Exercise 30.37.** (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

$$\langle i^*(\phi), v\rangle = \langle \phi, i(v)\rangle \tag{30.18}$$

for every $v \in V$ and every $\phi \in H^*$. In case we identify $H$ and $H^*$ this reads

$$\langle i^*(\phi), v\rangle = (\phi, i(v))_H. \tag{30.19}$$

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 $\phi_n \in H^*$ 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 30.38.** 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 (30.18) 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 $v \in V$.

**Exercise 30.39.** Assume $H$ and $V$ Hilbert spaces, $i : V \to H$ linear and continuous. Let $S : H \to V$ be given via Exercise 30.35 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 $f \in H$.

**Exercise 30.40.** 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 30.41.** 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 30.42.** 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 18.3, see also (18.12), 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 30.35 and 30.37.

**Exercise 30.43.** (continued) If the first supremum is a maximum then its maximizer $\phi$ is an eigenvector with eigenvalue $\lambda = |S_0|_{L(H,H)}$. You should give a direct proof of this, but see Remark 18.7. Same statement for $S_1$ and the second supremum of course.

**Exercise 30.44.** Any eigenvector $\phi$ of $S_0$ makes for an eigenvector $\psi = S\phi$ of $S_1$ with the same eigenvalue, unless $S\phi = 0$. Likewise, any eigenvector $\psi$ 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 30.42 for the norm of $S_0$ is a maximum, then so is the supremum for the norm of $S_1$ and vice versa.

**Remark 30.45.** *Each linear, injective, continuous*[23] *compact*

$$i : V \to H \quad \text{with} \quad \overline{i(V)} = H$$

*defines via Exercises 30.35, 30.40 and 30.41 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*[24].

[23] Follows from compactness of $i$.

[24] See the Brézis book on functional analysis.

# 31 Lebesgue spaces

The definitions of Lebesgue spaces such as $L^p(\Omega)$ usually come at the end[1] of a course on measure theory and Lebesgue integration[2]. For $p \geq 1$ and $\Omega \subset \mathbb{R}^N$ open an alternative approach[3] to define $L^p(\Omega)$ is provided in Section 31.6, via *equivalence classes of Cauchy sequences* of compactly supported continuous functions, *mollified as functions* in Section 31.8.

In both approaches $L^p(\Omega)$ is the $p$-norm closure of $C_c(\Omega)$, the space of continuous functions $f : \Omega \to \mathbb{R}$ with compact support in $\Omega$, and it will be convenient to consider $C_c(\Omega)$ as being contained in $C_c(\mathbb{R}^N)$. The Cauchy sequence approach only requires the use of Riemann integrals, and properties of the $p$-norm, defined by

$$|f|_p^p = \int_{\mathbb{R}^N} |f|^p$$

for $f \in C_c(\mathbb{R}^N)$. After recalling these properties in Section 31.1 you can therefore jump to Remark 31.17 in Section 31.3 about the *mollified functions* $f^\varepsilon = \eta_\varepsilon * f$ used in Chapter 32 for the theory of Sobolev spaces.

The *mollifiers* $\eta_\varepsilon$ are introduced in Exercise 31.16 as radially symmetric smooth nonnegative convolution kernels of the form

$$\eta_\varepsilon(x) = \frac{1}{\varepsilon^N}\eta(\frac{x}{\varepsilon}),$$

in which $\eta$ is chosen with compact support in the open unit ball $B$, and

$$\int_B \eta = 1.$$

The *uniform convergence* of $f^\varepsilon$ to $f$ in Exercise 31.16 is essential for our purposes in Chapter 32.

We combine the good behaviour of $f^\varepsilon$ for $f \in C_c(\mathbb{R}^N)$ with Remark 31.24 about integrals of equivalence classes of Cauchy sequences. This remark takes you from the Cauchy sequences in Section 31.6 straight to Theorem 31.32: the equivalence class $F$ of a Cauchy sequence $f_n$ in $C_c(\mathbb{R}^N)$ mollifies to a (smooth) *function* $F_\varepsilon$ in[4] $C_0(\mathbb{R}^N)$ with finite $p$-norm. In particular

$$|F_\varepsilon|_p^p = \lim_{n\to\infty} \int_{\mathbb{R}^N} |f_n^\varepsilon|^p = \int_{\mathbb{R}^N} |F_\varepsilon|^p < \infty.$$

[1] If time permits....

[2] See e.g. Folland's Real Analysis book, Chapters 2 and 6.

[3] Similar tot the construction of $\mathbb{R}$ out of $\mathbb{Q}$.

[4] Recall that $C_0(\mathbb{R}^N)$ is the space of continuous functions $f : \mathbb{R}^N \to \mathbb{R}$ with

$$\forall_{\delta>0}\, \exists_{R>0}\, \forall_{x\in\mathbb{R}^N} : |x| \geq R \implies |f(x)| < \delta.$$

## 31.1 Hölder's inequality

We recall from Section 17.3 that

$$|\sum_{i=1}^{n} a_i b_i| \leq \sum_{i=1}^{n} |a_i b_i| \leq |a|_p \, |b|_q \quad \text{for} \quad p, q > 1 \quad \text{with} \quad \frac{1}{p} + \frac{1}{q} = 1. \quad (31.1)$$

This is Hölder's inequality for finite sums of real numbers. Memorise that

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

Via any kosher definition of the integral it also holds that

$$|\int_\Omega fg| \leq \int_\Omega |fg| \leq |f|_p \, |g|_q, \quad (31.2)$$

if the norms are finite, e.g. if $f, g \in C_c(\Omega) \subset C_c(\mathrm{I\!R}^{\mathrm{N}})$.

**Exercise 31.1.** The triangle inequality for the 1-norm is equivalent to

$$\int_\Omega |f+g| \leq \int_\Omega |f| + \int_\Omega |g|.$$

Use this inequality to derive the triangle inequality for the $p$-norm from Hölder's inequality applied to

$$\int_\Omega |f+g|^{p-1}|f| \quad \text{and} \quad \int_\Omega |f+g|^{p-1}|g|.$$

**Exercise 31.2.** No estimate of the type

$$|u|_p \leq C_{pqN}|u|_q$$

with $p > q \geq 1$ can hold for all $u \in C_c(\mathrm{I\!R}^{\mathrm{N}})$. Why? Hint: scale the spatial variable.

**Remark 31.3.** *The limit case $p = \infty$: the $\infty$-norm $|f|_\infty$ of a measurable function $f$ is the smallest number $M$ such that*

$$|\{x \in \Omega : \, |f(x)| \leq M\}| = 0.$$

*If such $M \geq 0$ exists we say that $f \in L^\infty(\Omega)$, and then (31.2) holds with $p = \infty$ and $q = 1$ if $g \in L^1(\Omega)$. For $f \in C_0(\mathrm{I\!R}^{\mathrm{N}})$ the $\infty$-norm of $f$ is equal to the maximum norm*

$$|f|_{max} = \max_{x \in \mathrm{I\!R}^{\mathrm{N}}} |f(x)|$$

*of $f$.*

**Exercise 31.4.** Show that

$$|g|_p \leq |g|_1^{\frac{1}{p}} \, |g|_\infty^{1-\frac{1}{p}}.$$

Hint: use (31.2) with $q = \infty$, $p = 1$.

**Exercise 31.5.** Apply (31.2) to $f^a$ and $f^b$ to obtain

$$|f|_{a+b}^{a+b} \leq |f|_{ap}^{a} |f|_{bq}^{b}.$$

Then solve the equations $1 \leq ap = r < a + b = s < bq = t$ and $(p-1)(q-1) = 1$ to prove the (interpolation) inequality

$$|f|_s \leq |f|_r^{\frac{r}{s}\frac{t-s}{t-r}} \, |f|_t^{\frac{t}{s}\frac{s-r}{t-r}}.$$

This inequality leads to

$$L^r(\Omega) \cap L^t(\Omega) \subset L^s(\Omega) \quad \text{for} \quad r < s < t.$$

Check that the limit case $r = 1$, $t = \infty$ is consistent with Exercise 31.4.

## 31.2 Lebesgue spaces as Banach spaces

In the treatment below I need that you know what measurable sets and measurable functions are, and what it means for a measurable function to be integrable.

**Definition 31.6.** *Let $\Omega \subset \mathbb{R}^N$ be open and $p \geq 1$ a real number. A (Lebesgue) measurable function $u : \Omega \to \mathbb{R}$ is said to be in $L^p_{loc}(\Omega)$ if the (Lebesgue) integral*

$$\int_B |f|^p$$

*is finite for every open ball $B$ with $\bar{B} \subset \Omega$, and in $L^p(\Omega)$ if*

$$|f|_p^p = \int_\Omega |f|^p < \infty. \tag{31.3}$$

*This defines the (semi)norm $|f|_p$, which is called the $p$-norm of $f$ in $L^p(\Omega)$.*

**Exercise 31.7.** Explain why the spaces $L^p_{loc}(\Omega)$ are nested:

$$L^p_{loc}(\Omega) \subset L^q_{loc}(\Omega) \subset L^1_{loc}(\Omega) \quad \text{if} \quad p \geq q \geq 1.$$

Hint: use (31.2) with $g \equiv 1$ to show that the spaces $L^p(\Omega)$ are nested if $\Omega$ is bounded.

We need some technicalities to deal with $|f|_p = 0$ only implying that

$$\{x \in \Omega : \, f(x) \neq 0\}$$

is a set of zero measure[5]. To turn $L^p(\Omega)$ into a Banach space with its norm defined by (31.3), we consider equivalence classes in $L^p(\Omega)$ for the equivalence relation

$$f \sim g \iff |\{x \in \Omega : \, f(x) \neq g(x)\}| = 0 \iff \int_\Omega |f-g|^p = 0. \tag{31.4}$$

For $f \in L^p(\Omega)$ the $p$-norm of $[f]$ is well defined by

$$|[f]|_p^p = \int_\Omega |f|^p,$$

and makes[6]

$$\{[f] : \, f \in L^p(\Omega)\}$$

a Banach space, and the (equivalence classes of) compactly supported continuous functions form a dense subspace of this Banach space[7]. The density is formulated as

$$\forall_{f \in L^p(\Omega)} \forall_{\varepsilon > 0} \exists_{g \in C_c(\Omega)} \quad |f-g|_p < \varepsilon. \tag{31.5}$$

Equivalently, for every $L^p(\Omega)$ there exists a sequence $f_n \in C_c(\Omega)$ such that $f_n \to f$ in $L^p(\Omega)$. There are many such sequences, and every such sequence is a Cauchy sequence[8] with respect to the $p$-norm.

**Remark 31.8.** *Every $f \in L^p(\Omega)$ extends to $f \in L^p(\mathbb{R}^N)$ by setting $f(x) = 0$ for $x \notin \Omega$. No such general[9] statement holds for $f \in L^p_{loc}(\Omega)$ and $L^p_{loc}(\mathbb{R}^N)$.*

**Remark 31.9.** *We note that $|A| = 0$ if and only if for every $\varepsilon > 0$ there[10] exist a sequence of open balls $B_n = B(x_n, r_n)$ indexed by $n \in \mathbb{N}$ such that*

$$A \subset \cup_{n\in\mathbb{N}} B_n \quad \text{and} \quad \sum_{n\in\mathbb{N}} |B_n| \leq \varepsilon.$$

*Recall that*

$$|B_n| = \omega_N r_n^N.$$

[5] Here $|A|$ denotes the Lebesgue measure of a Lebesgue measurable subset $A \subset \mathbb{R}^N$.
[6] This is a theorem that requires the full machinery of the Lebesgue integral.
[7] The brackets around $f$ are dropped from the notation in everyday practice.
[8] Definition 31.20 introduces the notion of equivalence.
[9] Example: $p = N = 1$, $f(x) = \frac{1}{x}$, $\Omega = \mathbb{R}_+$.
[10] This is in fact the statement that the Lebesgue *outer* measure of $A$ is at most $\varepsilon > 0$.

*This zero measure concept was already introduced in Section 8.4. It does not 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 31.10.** 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.

## 31.3 Statement of Lebesgue's Differentiation Theorem

In Section 31.2 we discussed the standard approach with equivalence classes of measurable functions to define $L^p(\Omega)$. For what it's worth we now identify a unique and in some sense best function $\tilde{f}$ in every equivalence class $[f]$ of measurable functions equal to $f$ almost everywhere in $\Omega$. Theorem 31.12 may very well be formulated and proved *before* the machinery of Lebesgue integration is introduced. The proof only uses properties that any kosher extension of the Riemann integral should have.

Since much of all this is for a PDE course we recall a theorem from Chapter 1 of

`https://www.few.vu.nl/~jhulshof/NOTES/ellpar.pdf`

about harmonic functions, i.e. solutions of the partial differential equation

$$\Delta f = 0.$$

It says that a continuous function $f : \Omega \to \mathbb{R}$ defined on an open set $\Omega \subset \mathbb{R}^{\mathrm{N}}$ is harmonic if and only if for every closed ball $\bar{B}(x, r) \subset \Omega$ it holds that

$$f(x) = \underbrace{\frac{1}{|B(x,r)|} \int_{B(x,r)} f}_{A_f(x,r)}\,. \tag{31.6}$$

Part of the proof of this characterising *mean value property* uses the much weaker statement

$$A_f(x_0, r) = \frac{1}{|B(x_0,r)|} \int_{B(x_0,r)} f \to f(x_0) \quad \text{as} \quad r \to 0 \tag{31.7}$$

for the local averages $A_f(x_0, r)$. This statement holds for every locally integrable $f : \Omega \to \mathbb{R}$ that is continuous in a given point $x_0 \in \Omega$, and it is needed for the proof of Theorem 31.12. It is in turn a special case of statements about convolutions, see Exercise 31.16 and Remark 31.17, which are essential for our purposes in Chapter 32.

**Definition 31.11.** *The good set of a function $f \in L^1_{loc}(\Omega)$ is defined by*

$$G_f = \{x \in \Omega : \lim_{r \downarrow 0} A_f(x,r) = f(x)\}. \tag{31.8}$$

*For every $x \in \Omega$ the existence and value of the limit in (31.8) rely only on the equivalence class*[11] *$[f]$ to which $f$ belongs. Thus the set*

$$\mathcal{N}_f = \mathcal{N}_{[f]} = \{x \in \Omega : \lim_{r \downarrow 0} A_f(x,r) \, \text{does not exist}\},$$

*may be called the bad set of the equivalence class $[f]$.*

**Theorem 31.12.** *(The Lebesgue Differentiation or Good Set Theorem) For every $f \in L^1_{loc}(\Omega)$ the good set $G_f$ has a complement in $\Omega$ with zero measure. This complement contains the set $\mathcal{N}_f$, which thereby also has zero measure. Thus there is a unique $\tilde{f}$ in the equivalence class $[f]$ for which*

$$\lim_{r \downarrow 0} A_{\tilde{f}}(x,r) = \lim_{r \downarrow 0} A_f(x,r) = \tilde{f}(x)$$

*for all $x \notin \mathcal{N}_{\tilde{f}} = \mathcal{N}_f$, and $\tilde{f}(x) = 0$ for all $x \in \mathcal{N}_{\tilde{f}}$. Therefore $\Omega$ is the disjoint union of $G_{\tilde{f}}$ and $\mathcal{N}_{\tilde{f}}$.*

**Theorem 31.13.** *Let $f \in L^1_{loc}(\Omega)$. Then for almost all $x$ in $\Omega$ it holds that*

$$⨍_{B(x,r)} |f - f(x)| \to 0 \quad \text{as} \quad r \to 0.$$

*For such $x$ this is a stronger statement than the limit statement in Theorem 31.12.*

**Exercise 31.14.** Apply Theorem 31.12 to the function $s \to |f(s) - q|$ for every $q \in \mathbb{Q}$ to prove Theorem 31.13. Hint: with the integration variable in

$$|f(s) - f(x)| \leq |f(s) - q| + |q - f(x)|$$

being $s$, it follows that

$$⨍_{B(x,r)} |f - f(x)| \leq \underbrace{⨍_{B(x,r)} |f - q|}_{\to |f(x) - q| \text{ if } x \in G_{|f-q|}} + |q - f(x)|.$$

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.

[11] The first $\iff$ in (31.4) defines the equivalence classes.

**Exercise 31.15.** Generalise the statements in Theorems 31.12,31.14 to $L^1_{loc}(\Omega)$.

**Exercise 31.16.** Prove the statements in Theorem 31.12 for $f \in C_c(\mathrm{I\!R}^N)$ by showing that $G_f = \mathrm{I\!R}^N$, i.e. the limit exists for every $x \in \mathrm{I\!R}^N$ and is what it should be. Hint: this is very much like Exercise 28.35 if you specify a convolution kernel $K_r$ such that

$$A_{x,r}f = (K_r * f)(x).$$

Show that the convergence is in fact uniform, and likewise for the smooth compactly supported functions $f^\varepsilon$ defined by

$$f^\varepsilon(x) = (\eta_\varepsilon * f)(x) = \int_{\mathrm{I\!R}^N} \eta_\varepsilon(x-y)f(y)\,dy, \tag{31.9}$$

with families of kernels given by

$$\eta_\varepsilon(x) = \frac{1}{\varepsilon^N}\eta(\frac{x}{\varepsilon}),$$

in which[12]

$$\eta \in C_c^\infty(B) \subset C_c^\infty(\mathrm{I\!R}^N), \quad \eta(x) = \eta(|x|) \geq 0, \quad \int_{\mathrm{I\!R}^N} \eta = 1.$$

**Remark 31.17.** *Let $\eta_\varepsilon$ be convolution kernels such as in Exercise 31.16. Hölder's inequality with*

$$\frac{1}{p} + \frac{1}{q} = 1$$

*applied to (31.9), i.e. to*

$$f^\varepsilon(x) = (\eta_\varepsilon * f)(x) = \int_{\mathrm{I\!R}^N} \eta_\varepsilon(x-y)f(y)\,dy,$$

*gives*

$$|f^\varepsilon(x)| = |\int_{\mathrm{I\!R}^N} \eta_\varepsilon(x-y)f(y)\,dy| \leq \underbrace{|\eta_\varepsilon|_q}_{\varepsilon^{-\frac{N}{p}}|\eta|_q} |f|_p,$$

*but also*

$$|f^\varepsilon(x)| \leq \int_{|y|\leq\varepsilon} |f(x+y)|\,\eta_\varepsilon(y)^{\frac{1}{p}+\frac{1}{q}}\,dy$$

[12] We write $B$ for the open unit ball.

$$\leq \left(\int_{|y|\leq\varepsilon} |f(x+y)|^p \,\eta_\varepsilon(y)\,dy\right)^{\frac{1}{p}} \underbrace{\left(\int_{|y|\leq\varepsilon} \eta_\varepsilon(y)\,dy\right)^{\frac{1}{q}}}_{=1},$$

*whence*

$$\int_{\mathbb{R}^N} |f^\varepsilon(x)|^p\,dx \leq \int_{x\in\mathbb{R}^N}\int_{y\in\mathbb{R}^N} |f(x+y)|^p\,\eta_\varepsilon(y)\,dy\,dx =$$

$$\int_{y\in\mathbb{R}^N} \underbrace{\int_{x\in\mathbb{R}^N} |f(x+y)|^p\,dx}_{=\int_{\mathbb{R}^N}|f(x)|^p\,dx}\;\eta_\varepsilon(y)\,dy = \int_{\mathbb{R}^N} |f(x)|^p\,dx.$$

*Summing up we have*

$$|f^\varepsilon(x)| \leq \underbrace{\varepsilon^{-\frac{N}{p}}|\eta|_q}_{|\eta_\varepsilon|_q}\,|f|_p \quad \text{and} \quad |f^\varepsilon|_p \leq \underbrace{|\eta_\varepsilon|_1}_{1}\,|f|_p \tag{31.10}$$

*for $f \in C_c(\mathbb{R}^N)$. These estimates prepare to have in Chapter 32 that $f^\varepsilon \to f$ in p-norm as $\varepsilon \to 0$, not only for $f \in C_c(\mathbb{R}^N)$, but for all $f \in L^p(\mathbb{R}^N)$.*

## 31.4 Proof of Lebesgue's Differentiation Theorem

This section is not essential for our purposes in Chapter 32, but the proof is of independent interest. It invokes the Hardy-Littlewood function, defined by

$$H_f(x) = \sup_{r>0} A_{|f|}(x,r) = \sup_{r>0} \frac{1}{\omega_N r^N}\int_{B(x,r)} |f| \in [0,\infty], \tag{31.11}$$

the supremum of all average of $|f|$ on balls centered in $x$. Here we assume that $f \in L^1(\mathbb{R}^N)$ for simplicity.

We shall use in the proof of Theorem 31.12 that for every every $B(x,r)$ the integral

$$\int_{B(x,r)} f,$$

which is independent of the choice of $f$ in $[f]$, varies continuously with $x$ and $r > 0$. This continuity is combined with monotonicity properties such as

$$B(x,r) \subset B(y,s) \implies |\int_{B(x,r)} f| \leq \int_{B(x,r)} |f| \leq \int_{B(y,s)} |f|,$$

and the finite additivity of the integral, e.g.

$$\int_{B_1\cup\cdots\cup B_n} f = \int_{B_1} f + \cdots + \int_{B_n} f$$

for balls $B_1, \ldots, B_n$ with $B_i \cap B_j = \emptyset$ if $i \neq j$.

Note that

$$|A_f(x,r)| \leq A_{|f|}(x,r) \leq H_f(x) = H_{|f|}(x) \leq \infty. \tag{31.12}$$

In view of the decay estimate

$$0 \leq A_{|f|}(x,r) \leq \frac{1}{\omega_N r^N} \int_{\mathrm{I\!R}^\mathrm{N}} |f|,$$

and the continuity of $A_{|f|}(x,r)$, the supremum $H_f(x)$ in (31.11) is finite unless $A_{|f|}(x,r) \to \infty$ as $r \to 0$.

We next examine the averages $A_f(x,r)$ via $H_{f-g}(x)$ with $g \in C_c(\mathrm{I\!R}^\mathrm{N})$ chosen to have

$$\int_{\mathrm{I\!R}^\mathrm{N}} |f-g| > 0 \tag{31.13}$$

small. Writing

$$A_f(x,r) - f(x) = \underbrace{A_f(x,r) - A_g(x,r)}_{A_{f-g}(x,r)} + A_g(x,r) - g(x) + g(x) - f(x),$$

we use (31.12) with $f-g$. In the resulting inequality, which reads

$$|A_f(x,r) - f(x)| \leq H_{f-g}(x) + \underbrace{|A_g(x,r) - g(x)|}_{\to 0 \text{ as } r\to 0} + |g(x) - f(x)|,$$

the dependence on $r$ in the right hand side disappears as $r \to 0$, thanks to Exercise 31.16. Thus

$$\limsup_{r\to 0} |A_f(x,r) - f(x)| \leq H_{f-g}(x) + |g(x) - f(x)|. \tag{31.14}$$

If the left hand side of (31.14) is not small then the first or the third term is not small. Or both. We therefore consider the sets[13]

$$O^\varepsilon_{f-g} = \{x \in \mathrm{I\!R}^\mathrm{N} : H_{f-g}(x) > \varepsilon\};$$

$$S^\varepsilon_{f-g} = \{x \in \mathrm{I\!R}^\mathrm{N} : |f(x) - g(x)| > \varepsilon\},$$

[13] $O$ is for open.

and let $\mathcal{N}_f^\varepsilon$ be the set of all points $x \in \mathbb{R}^N$ for which the statement

$$\limsup_{r\to 0} |A_f(x,r) - f(x)| \le 2\varepsilon$$

fails. Then it must be that

$$\mathcal{N}_f^\varepsilon \subset O_{f-g}^\varepsilon \cup S_{f-g}^\varepsilon, \tag{31.15}$$

and these sets are nested:

$$0 < \eta < \varepsilon \implies O_{f-g}^\varepsilon \subset O_{f-g}^\eta, \quad S_{f-g}^\varepsilon \subset S_{f-g}^\eta \quad \text{and} \quad \mathcal{N}_f^\varepsilon \subset \mathcal{N}_f^\eta.$$

Via[14]

$$\int_{\mathbb{R}^N} |f-g| \ge \int_{S_{f-g}^\varepsilon} |f-g| \ge \varepsilon |S_{f-g}^\varepsilon|$$

we conclude from (31.15) that

$$|\mathcal{N}^\varepsilon| \le |O_{f-g}^\varepsilon| + \frac{1}{\varepsilon}\int_{\mathbb{R}^N} |f-g|. \tag{31.16}$$

Now suppose that

$$|O_{f-g}^\varepsilon| \le \frac{C_N}{\varepsilon}\int_{\mathbb{R}^N} |f-g| \tag{31.17}$$

for some universal $N$-dependent constant $C_N$. We can then choose $g$ to make the integral (31.13) as small we like to establish for every $\varepsilon > 0$ that

$$|\mathcal{N}^\varepsilon| = 0.$$

This will complete the proof because $G_f$ is the complement of the union $\mathcal{N}_f$ of the sets

$$\mathcal{N}_f^1 \subset \mathcal{N}_f^{\frac{1}{2}} \subset \mathcal{N}_f^{\frac{1}{3}} \subset \mathcal{N}_f^{\frac{1}{4}} \subset \mathcal{N}_f^{\frac{1}{5}} \subset \mathcal{N}_f^{\frac{1}{6}} \subset \dots,$$

and thereby, see Exercise 31.10, the complement of a set of measure zero.

It thus remains to estimate $H_{f-g}(x)$ and establish (31.17), but this argument will not depend on the choice of $g$. So we take $g \equiv 0$ and note that the set

$$O_f^\varepsilon = \{x \in \mathbb{R}^N : H_f(x) > \varepsilon\}$$

is open because

$$x \in O_f^\varepsilon \iff \exists r > 0 : \underbrace{\int_{B(x,r)} |f|}_{\text{continuous in } r,x} > \varepsilon |B(x,r)|. \tag{31.18}$$

[14]This does require to have the Lebesgue measure of $S_{f-g}^\varepsilon$ well-defined.

The measure of this set $O_f^\varepsilon$ is the supremum of all the measures of compact subsets $K$ of $O_f^\varepsilon$, and every such $K$ is covered[15] by only finite many balls as in (31.18), say

$$K \subset B_1 \cup \cdots \cup B_m.$$

If these balls were always disjoint it would follow that

$$\varepsilon|K| \leq \varepsilon(|B_1+\cdots+|B_m|) < \int_{B_1} |f|+\cdots+\int_{B_m} |f| = \int_{B_1\cup\cdots\cup B_m} |f| \leq \int_{\mathbb{R}^N} |f|,$$

and we would get (31.17) with $C_N = 1$, but of course we cannot expect this to be the case. It does however hold that

$$\varepsilon|K| \leq 3^N \int_{\mathbb{R}^N} |f|, \tag{31.19}$$

which gives (31.17) with $C_N = 3^N$.

To prove (31.19) we choose the 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 in the old collection, let $B_{j_2}$ be the largest of the remaining balls, 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[16] $B_{j_1},\ldots,B_{j_k}$ are disjoint, most likely don't cover $K$, but we do have

$$K \subset B_1 \cup \cdots \cup B_m \subset 3B_{j_1} \cup \cdots \cup 3B_{j_k}.$$

Therefore

$$\begin{aligned}\int_{\mathbb{R}^N} |f| \geq \int_{B_1\cup\cdots\cup B_m} |f| \geq \int_{B_{j_1}\cup\cdots\cup B_{j_k}} |f| = \int_{B_{j_1}} |f| + \cdots + \int_{B_{j_k}} |f| \\ > \varepsilon(|B_{j_1}| + \cdots + |B_{j_k}|) = 3^{-N}\varepsilon(|3B_{j_1}| + \cdots + |3B_{j_k}|) \\ \geq 3^{-N}\varepsilon|3B_{j_1} \cup \cdots \cup 3B_{j_k}| \geq 3^{-N}\varepsilon|B_1 \cup \cdots \cup B_m| \geq 3^{-N}\varepsilon|K|.\end{aligned}$$

So indeed (31.19) holds for all compact $K \in O_f^\varepsilon$ and (31.17) with $g \equiv 0$ follows. This completes the proof that the complement of the good set (31.8) has zero measure. □

**Exercise 31.18.** Generalise the proof to $f \in L^1_{loc}(\Omega)$.

[15]Both compactness and measurability rely on definitions with coverings.

[16]The statement that follows is called Vitali's covering lemma.

## 31.5 Another proof of the Hardy-Littlewood estimate

This section is not essential for our purposes in Chapter 32. For $f \in L^1(\Omega)$, $\Omega \subset \mathrm{I\!R}^{\mathrm{N}}$ open[17], we reformulate the estimate in (31.17) with $g \equiv 0$ as a separate theorem. A modified proof uses countable coverings of $\Omega$ with open balls. This may be of some independent interest.

**Theorem 31.19.** *For $f \in L^1(\Omega)$ and $\varepsilon > 0$ let*

$$H_f(x) = \sup_{\substack{r>0 \\ B(x,r)\subset\Omega}} \fint_{B(x,r)} |f| \quad \textit{and} \quad O_f^\varepsilon = \{x \in \Omega : \, H_f(x) > \varepsilon\}.$$

*Then there exists an $f$-dependent family of balls $B_k$ indexed by a subset $K$ of $\mathrm{I\!N}$ such that*

$$O_f^\varepsilon \subset \cup_{k\in K} B_k \quad \textit{with} \quad \sum_{k\in K} |B_k| \leq \frac{6^N}{\varepsilon} |f|_{_1}.$$

*The number 6 can be replaced by any real number larger than 3. Thus the Lebesgue outer measure of $O_f^\varepsilon$ is at most*

$$\frac{3^N}{\varepsilon} |f|_{_1}.$$

**Proof**. We use the continuity of $A_{|f|}(x,r)$ to prove the statement in Theorem 31.19 about the set $O_f^\varepsilon$ of points $x$ which allow an average

$$\fint_{B(x,r)} |f| > \varepsilon$$

with $B(x,r) \subset \Omega$. Indeed, since

$$x \in O_f^\varepsilon \iff \exists r > 0 : \underbrace{\int_{B(x,r)} |f|}_{\text{continuous in } r,x} > \varepsilon |B(x,r)|, \, B(x,r) \subset \Omega,$$

the set $O_f^\varepsilon$ is open. Moreover, $O_f^\varepsilon$ is contained in the union of all such balls $B(x,r)$ and in every $B(x,r)$ there is an open ball $B$ with rational center and rational radius such that

$$\int_B |f| > \varepsilon |B| \quad \text{and} \quad x \in B.$$

[17]Before we took $\Omega = \mathrm{I\!R}^{\mathrm{N}}$, now we work with open balls $B(x,r) \subset \Omega$.

We conclude that there is a *countable* family of open balls $B_n \subset \Omega$ such that

$$O_f^\varepsilon \subset \cup_{n\in\mathbb{N}} B_n \quad \text{with} \quad \int_{B_n} |f| > \varepsilon |B_n|.$$

We will show that a subcollection of enlarged balls will do the job.

To see how let $r_n > 0$ be the corresponding sequence of radii and denote the distances between the centers of the balls $B_m$ and $B_n$ by $d_{mn} \geq 0$. Since

$$\varepsilon |B_n| < \int_\Omega |f|,$$

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, which we denote by $6\tilde{B}_1$, 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[18]

$$r_{n_2} > \frac{R_2}{2},$$

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[19] open balls $\tilde{B}_k$ indexed by $k$, and for every finite sum indexed by a finite subset $K$ of $\mathbb{N}$ we have

$$\varepsilon \sum_{k\in K} |B_k| < \sum_{k\in K} \int_{B_k} |f| = \int_{\cup_{k\in K} B_k} |f| \leq \int_\Omega |f|.$$

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

$$O_f^\varepsilon \subset \cup_{k\in\mathbb{N}} 6\tilde{B}_k,$$

which we view as

$$n = \infty \quad \text{in} \quad O_f^\varepsilon \subset \cup_{k=1}^n 6\tilde{B}_k \tag{31.20}$$

for the case that the process does stop at some $k = n \in \mathbb{N}$.

[18]The $2 > 1$ in the denominator leads to $3 \cdot 2 = 6 > 3$, any other $p > 1$ will also do.

[19]Nontouching because $d_{kl} > 2R_k \geq R_k + R_l$ for $l > k \geq 1$.

For every $m \in \mathbb{N}$ with $m \leq n$ we then have that

$$\varepsilon \sum_{k=1}^{m} |6\tilde{B}_k| = 6^N \varepsilon \sum_{k=1}^{m} |\tilde{B}_k| < 6^N \sum_{k=1}^{m} \int_{\tilde{B}_k} |f| = 6^N \int_{\cup_{k=1}^{m} \tilde{B}_k} |f| \leq 6^N \int_{\Omega} |f|,$$

so we conclude that

$$O_f^\varepsilon \subset \cup_{k=1}^{n} 6\tilde{B}_k \quad \text{with} \quad \sum_{k=1}^{n} |6\tilde{B}_k| \leq \frac{6^N}{\varepsilon} \int_{\Omega} |f| \quad \text{and} \quad n \in \mathbb{N} \cup \{\infty\}. \tag{31.21}$$

It remains to rename these balls $6\tilde{B}_k$ as $B_k$. Note that the number 6 appears as $2 \cdot 3$. Choosing $p > 1$ instead of 2 the estimate is improved in that 6 is replaced by $3p$. This completes the proof of Theorem 31.19. □

## 31.6 Lebesgue spaces via Cauchy sequences

If you are unfamiliar with Lebesgue integration then this section will help you to avoid it. We define $L^p(\Omega)$ as consisting of equivalence classes of Cauchy sequences of compactly supported continuous functions $f_n$, and explain how to integrate them. You may have arrived here skipping Theorem 31.12 and all that followed except Exercise 31.16 and Remark 31.17. After this section you can jump to Section 31.8 which adapts Exercise 31.16 to our later purposes in Chapter 32.

**Definition 31.20.** *Let $\Omega \subset \mathbb{R}^N$ be a nonempty open set. Two sequences $f_n$ and $g_n$ in $C_c(\Omega)$ are equivalent in the p-norm if*[20]

$$|f_n - g_n|_p \to 0$$

*as $n \to \infty$. We write*

$$F = [f_n] \tag{31.22}$$

*for the equivalence class of a sequence $C_c(\Omega)$ with*

$$|f_n - f_m|_p \to 0$$

*as $m, n \to \infty$. Such sequences are called p-Cauchy sequences, and we let $L^p(\Omega)$ be the Banach space of all such equivalence classes $F$ equipped with the norm*

$$|F|_p = \lim_{n\to\infty} |f_n|_p .$$

*This is nothing but the abstract completion procedure applied to the vector space $C_c(\Omega)$ and the p-norm. In view of Exercise 31.16 it is no restriction to assume that every $f_n$ in (31.22) is smooth.*

[20] If this holds for sequences $f_n$, $g_n$ in $C_c(\mathbb{R}^N)$ we call them equivalent on $\Omega$.

**Remark 31.21.** *We extend*[21] *functions in* $C_c(\Omega)$ *to functions in* $C_c(\mathbb{R}^N)$, *and thereby automatically have that* $L^p(\Omega) \subset L^p(\Omega') \subset L^p(\mathbb{R}^N)$ *for every* $\Omega'$ *open with* $\Omega \subset \Omega' \subset \mathbb{R}^N$. *To restrict elements of* $L^p(\Omega')$ *to* $L^p(\Omega)$ *we simply choose for every p-Cauchy sequence* $g_n$ *in* $C_c(\Omega')$ *a p-Cauchy sequence* $f_n$ *in* $C_c(\Omega)$ *with* $\int_\Omega |f_n - g_n|^p \to 0$ *as* $n \to \infty$.

We first restrict the attention to $p = 1$. Cauchy sequences with respect to the 1-norm are characterised by the property that

$$\int_\Omega |f_n - f_m| \to 0$$

as $m, n \to \infty$. If $f_n \in C_c(\Omega)$ is a Cauchy sequence and $f_n \sim g_n$ for some other sequence $g_n \in C_c(\Omega)$, then also $g_n$ is a Cauchy sequence. In particular $F = [f_n]$ is the zero element if and only if $|f_n|_1 = \int_\Omega |f_n| \to 0$ as $n \to \infty$.

Now consider two such equivalent Cauchy sequences in $C_c(\Omega) \subset C_c(\mathbb{R}^N)$, and let $A \subset \mathbb{R}^N$ be any bounded set over which we can integrate continuous functions by means of the Riemann integral. the sequences

$$\int_A f_n \quad \text{and} \quad \int_A g_n$$

are (equivalent) Cauchy sequences in $\mathbb{R}$ with the same limit. Thus

$$\int_A F = \lim_{n\to\infty} \int_A f_n, \tag{31.23}$$

is the natural definition of the integral[22] of the equivalence class (31.22) over $A$. The functional

$$F \to \int_A F \tag{31.24}$$

is linear, and

$$\int_A |F| = \lim_{n\to\infty} \int_A |f_n| \geq |\lim_{n\to\infty} \int_A f_n| = |\int_A F|,$$

in which $|F| = [|f_n|]$ is the equivalence class[23] of the Cauchy sequence $|f_n|$.

**Remark 31.22.** *Likewise*

$$\int_\Omega F\phi = \lim_{n\to\infty} \int_\Omega f_n \phi$$

[21]By setting $f_n(x) = 0$ for all $x \notin \Omega$.

[22]If $A \supset \Omega$ then $\int_A F$ defines $\int_\Omega F$.

[23]See Exercise 31.23 for some details.

*defines the integral of the product of the equivalence class $F$ and the function $\phi \in C_c(\mathrm{I\!R}^{\mathrm{N}})$ as the limit of the Cauchy sequence $\int_\Omega f_n\phi$. For $\phi \in C_c(\Omega)$ this only requires equivalence classes for the equivalence relation*

$$f_n \sim g_n \iff \lim_{n\to\infty}\int_B |f_n - g_n| = 0 \quad \textit{for every ball} \quad B \quad \textit{with} \quad \bar{B} \subset \Omega$$

*of sequences $f_n$ for which*

$$\lim_{m,n\to\infty}\int_B |f_n - f_m| = 0 \quad \textit{for every ball} \quad B \quad \textit{with} \quad \bar{B} \subset \Omega.$$

*We can think of such equivalence classes as constituting the space $L^1_{loc}(\Omega)$.*

**Exercise 31.23.** Let $F = [f_n]$ be as in Definition 31.20 with $p = 1$. Write

$$f_n(x) = f_n^+(x) - f_n^-(x),$$

with $f_n^+$ and $f_n^-$ nonnegative. Let $F^+ = [f_n^+]$, $F^- = [f_n^-]$, and $|F| = [|f_n|]$. Referring to (31.23) show that

$$\int_A F = \int_A F^+ - \int_A F^- \quad \text{and} \quad \int_A |F| = \int_A F^+ + \int_A F^-.$$

Hint: prove and use the linearity of (31.24).

**Remark 31.24.** *Now that we can integrate equivalence classes of Cauchy sequences you can jump to Section 31.8 as you wish, where we will need a variant of the following proposition. The proof below is instructive.*

**Proposition 31.25.** *Let $\Omega$ be open, and let $F$ be an equivalence class of a Cauchy sequence $f_n$ in $C_c(\Omega)$ with respect to the 1-norm. Then the function*[24]

$$(x, r) \to \int_{B(x,r)} F \tag{31.25}$$

*is continuous, uniformly*[25] *in $x$ and $r$.*

**Proof**. We estimate

$$|\int_{B(x,r)} F - \int_{B(y,s)} F| \leq$$

[24]Compare with (31.18).

[25]A restriction to bounded sets is not necessary because $f_N \in C_c(\mathrm{I\!R}^{\mathrm{N}})$ .

$$|\int_{B(x,r)} F - \int_{B(x,r)} f_n| + |\int_{B(x,r)} f_n - \int_{B(y,s)} f_n| + |\int_{B(y,s)} f_n - \int_{B(y,s)} F|$$

$$\leq \underbrace{\int_{B(x,r)} |F - f_n|}_{\leq \varepsilon} + |\int_{B(x,r)} f_n - \int_{B(y,s)} f_n| + \underbrace{|\int_{B(y,s)} |f_n - F|}_{\leq \varepsilon} \tag{31.26}$$

for $n \geq N$, if $N$ corresponds to $\varepsilon > 0$ via the definition of $f_n$ being a Cauchy sequence[26]. The definition of the integral of the class $F$ in (31.23), as the limit of the Cauchy sequence

$$\int_{B(x,r)} f_n,$$

has that same $N$ implying the $\varepsilon$-bounds for all balls $B(x,r)$ and $B(y,s)$ in (31.26) simultaneously if $n \geq N$.

We then fix $n = N$ and ask for the second term in (31.26) that

$$|\int_{B(x,r)} f_N - \int_{B(y,s)} f_N| \leq \varepsilon.$$

Since $f_N \in C_c(\Omega) \subset C_c(\mathbb{R}^N)$, this can be done uniformly in terms of the smallness of $|r - s|$ and $|x - y|$. $\square$

## 31.7 From Cauchy sequences to functions

For what it's worth we ask again the question: does the equivalence class define a function? So let $F$ be an equivalence class of a Cauchy sequence $f_n$ as in Definition 31.20. Then we can define

$$A_F(x,r) = ⨍_{B(x,r)} F = \frac{1}{|B(x,r)|} \int_{B(x,r)} F \tag{31.27}$$

for all $x \in \mathbb{R}^N$ and $r > 0$. If you wish[27] we can use (31.27) to turn $F$ into a function from $\mathbb{R}^N$ to $\mathbb{R}$ by means of a variant of Theorem 31.12.

**Theorem 31.26.** *Let $F$ be an equivalence class of Cauchy sequences in $C_c(\Omega)$ with respect to the 1-norm, and let $\mathcal{N}_F$ be the set of points $x$ for which*

$$\lim_{r \to 0} A_F(x,r) \tag{31.28}$$

*does not exist. Then $\mathcal{N}_F$ is a zero measure set.*

[26] With respect to the 1-norm.
[27] If not jump to Section 31.8.

**Proof**. We examine $\mathcal{N}_F$ using[28]

$$H_F(x) = \sup_{r>0} A_{|F|}(x,r) = \sup_{r>0} \frac{1}{|B(x,r)|} \int_{B(x,r)} |F| \in [0,\infty], \tag{31.29}$$

and reason as in Section 31.4. Replacing the estimates that lead to (31.14) by

$$|A_F(x,r) - A_F(x,s)| \leq$$
$$|A_F(x,r) - A_{f_m}(x,r)| + |A_{f_m}(x,r) - A_{f_m}(x,s)| + |A_{f_m}(x,s) - A_F(x,s)|$$
$$\leq A_{|F-f_m|}(x,r) + |A_{f_m}(x,r) - A_{f_m}(x,s)| + A_{|f_m-F|}(x,s)$$

for $0 < s < r$, it follows that

$$|A_F(x,r) - A_F(x,s)| \leq 2H_{F-f_m}(x) + \underbrace{|A_{f_m}(x,r) - A_{f_m}(x,s)|}_{\to\ 0 \text{ as } r,s \to 0.}. \tag{31.30}$$

The first term on the right hand side of (31.30) is twice the upper bound

$$H_{F-f_m}(x) = H_{|F-f_m|}(x)$$

for

$$A_{|F-f_m|}(x,r) \quad \text{and} \quad A_{|f_m-F|}(x,s).$$

Here $F - f_m$ with $m$ fixed denotes the equivalence class of the Cauchy sequence[29] $f_n - f_m$, and $|F - f_m|$ the equivalence class of the Cauchy sequence $|f_n - f_m|$. It follows that[30]

$$\limsup_{r,s\to 0} |A_F(x,r) - A_F(x,s)| \leq 2H_{F-f_m}(x), \tag{31.31}$$

Now let

$$\mathcal{N}_F^\varepsilon = \{x \in \mathrm{IR}^{\mathrm{N}} : \limsup_{r,s\to 0} |A_F(x,r) - A_F(x,s)| > 2\varepsilon\}$$

Then (31.31) gives[31]

$$\mathcal{N}_F^\varepsilon \subset O_{F-f_m}^\varepsilon = \{x \in \mathrm{IR}^{\mathrm{N}} : H_{F-f_m}(x) > \varepsilon\},$$

and similar to (31.18) the continuity established with Proposition 31.25 implies that $O_{F-f_m}^\varepsilon$ is open because

$$x \in O_{F-f_m}^\varepsilon \iff \exists r > 0 : \underbrace{\int_{B(x,r)} |F - f_m|}_{\text{continuous in } r \text{ and } x} > \varepsilon |B(x,r)|. \tag{31.32}$$

[28]Mainly for notational convenience we don't restrict to balls $B(x,r) \subset \Omega$.
[29]Indexed by $n$.
[30]This looks cleaner than (31.14) in fact.
[31]Compare to (31.15), in which we could have used subscripts $g$ on the right.

**Exercise 31.27.** Modify the proof of Theorem 31.19 to show that $\mathcal{N}_F^{\varepsilon}$ is a set of zero measure for every $\varepsilon > 0$. Hint: use (31.21).

Thus the limit in (31.28) exists outside the union $\mathcal{N}_F$ of the sets

$$\mathcal{N}_F^1 \subset \mathcal{N}_F^{\frac{1}{2}} \subset \mathcal{N}_F^{\frac{1}{3}} \subset \mathcal{N}_F^{\frac{1}{4}} \subset \mathcal{N}_F^{\frac{1}{5}} \subset \mathcal{N}_F^{\frac{1}{6}} \subset \dots,$$

a set of measure zero as before. □

**Definition 31.28.** *Let $F$ be the equivalence class in Theorem 31.26. We define the* function $F$ *by*

$$F(x) = \lim_{r \to 0} A_F(x,r) \quad \textit{for} \quad x \notin \mathcal{N}_F \quad \textit{and} \quad F(x) = 0 \quad \textit{for} \quad x \in \mathcal{N}_F,$$

*just like we did in Theorem 31.12 for equivalence classes of functions. We call $\mathcal{G}_F = (\mathcal{N}_F)^c$ the good set of $F$.*

**Exercise 31.29.** Referring to (31.31) and Exercise 31.23 show that[32]

$$\limsup_{r,s \to 0} |A_{F^+}(x,r) - A_{F^+}(x,s)| \leq 2H_{F-f_m}(x),$$

likewise for $F^- = [f_n^-]$, and also

$$\limsup_{r,s \to 0} |A_{|F|}(x,r) - A_{|F|}(x,s)| \leq 2H_{F-f_m}(x).$$

**Exercise 31.30.** Define $F^+(x)$, $F^-(x)$ and $|F|(x)$ as in Theorem 31.26. For which $x$ can you conclude that $|F|(x) = |F(x)| = F^+(x) + F^-(x)$ and $F(x) = F^+(x) - F^-(x)$?

In relation to Theorem 31.13, and its proof in Exercise 31.14, we observe that for fixed $q \in \mathbb{Q}$ we can also modify the proof of (31.31) to conclude that

$$\limsup_{r,s \to 0} |A_{|F-q|}(x,r) - A_{|F-q|}(x,s)| \leq 2H_{F-f_m}(x). \tag{31.33}$$

Now that we have our function $F$ and its good set $\mathcal{G}_F$ from Section 31.7, we can ask: given $x \in \mathcal{G}_F$, how does the value $F(x)$ provided by Definition 31.28 relate to $f_n(x)$? We observe for such $x$ that

$$|f_n(x) - F(x)| \leq$$

[32] The right hand sides are the same as in (31.31).

$$\underbrace{|f_n(x) - A_{f_n}(x,r)|}_{\to 0 \quad \text{as } r \to 0} + \underbrace{|A_{f_n}(x,r) - A_F(x,r)|}_{\leq H_{f_n - F}(x)} + \underbrace{|A_F(x,r) - F(x)|}_{\to 0 \quad \text{as } r \to 0},$$

whence

$$|f_n(x) - F(x)| \leq H_{f_n - F}(x), \tag{31.34}$$

and again the set

$$\{x \in \mathrm{IR}^{\mathrm{N}} : H_{f_n - F}(x) > \varepsilon\}$$

comes into play via

$$\{x \in \mathcal{G}_F : |f_n(x) - F(x)| > \varepsilon\}.$$

We can deal with it just as in Exercise 31.27. It is covered by an at most countable union of open balls with a bound

$$\frac{6^N}{\varepsilon} \int_\Omega |f_n - F|$$

for the sum of the measures of the covering balls. This allows to prove statements about the existence of convergent subsequences only[33].

**Exercise 31.31.** Let $\delta > 0$. For $k, m \in \mathbb{N}$ choose $n = N_{km}$ for which

$$m 6^N \int_\Omega |f_n - F| < \frac{\delta}{2^k}$$

to prove the existence of a subsequence of $f_n$ which converges uniformly on the complement of a set $\mathcal{N}_F^\delta$ with $|\mathcal{N}_F^\delta| < \delta$. Then choose $\delta = \frac{1}{j}$ to construct a subsequence that converges in every $x \in \Omega$ outside a set $\mathcal{N}_F$ of measure zero.

## 31.8 Mollifying functions and equivalence classes

Refering to

$$A_{x,r} f = (K_r * f)(x)$$

in Exercise 31.16, we recall that thanks to (31.23) we also had

$$A_{x,r} F = (K_r * F)(x) = \lim_{n \to \infty} (K_r * f_n)(x) \tag{31.35}$$

at our disposal for equivalence classes $F$ of Cauchy sequences $f_n \in C_c(\Omega)$. We now use the $p$-norm to mollify equivalence classes of the Cauchy sequences introduced in Section 31.6.

[33]Convergent subsequences is the best we can hope for in general.

**Theorem 31.32.** *Let $f_n \in C_c(\mathbb{R}^N)$ have the Cauchy property with respect to the p-norm, i.e.*

$$\int_{\mathbb{R}^N} |f_n - f_m|^p \to 0 \quad as \quad m, n \to \infty,$$

*and let $F$ be its equivalence class. Then*

$$F_\varepsilon(x) = (\eta_\varepsilon * F)(x) = \lim_{n\to\infty} \underbrace{(\eta_\varepsilon * f_n)}_{f_n^\varepsilon}(x) \tag{31.36}$$

*defines*[34] *a function $F_\varepsilon \in C_0(\mathbb{R}^N)$ for every $\varepsilon > 0$. The convergence is uniform and*

$$\int_{\mathbb{R}^N} |f_n^\varepsilon - F_\varepsilon|^p \to 0 \quad as \quad n \to \infty,$$

*and*

$$\int_{\mathbb{R}^N} |f_n^\varepsilon|^p \to \int_{\mathbb{R}^N} |F_\varepsilon|^p \quad as \quad n \to \infty.$$

*The integrals containing $F_\varepsilon$ are finite improper Riemann integrals of nonnegative continuous functions. We say that $F_\varepsilon \in C_0(\mathbb{R}^N)$ is p-integrable. In fact $F_\varepsilon$ is smooth, see Theorem 31.37, and all its derivatives are also p-integrable and in $C_0(\mathbb{R}^N)$.*

**Proof**. We note that $f_n^\varepsilon$ is the convolution of $C_c(\mathbb{R}^N)$ and $\eta_\varepsilon \in C_c^\infty(\mathbb{R}^N)$. Thereby $f_n^\varepsilon$ is smooth and its support is contained in a closed $\varepsilon$-neighbourhood of the support of $f_n$. We use the estimates in (31.10) applied to $f_n$ and $f_n - f_m$. The uniform bounds

$$|f_n^\varepsilon(x)| = |\int_{\mathbb{R}^N} \eta_\varepsilon(x-y) f_n(y)\,dy| \leq |\eta_\varepsilon|_q \, |f_n|_p,$$

$$|f_m^\varepsilon(x) - f_n^\varepsilon(x)| = |\int_{\mathbb{R}^N} \eta_\varepsilon(x-y)(f_m(y) - f_n(y))\,dy| \leq |\eta_\varepsilon|_q \, |f_m - f_n|_p,$$

make that the *function* $F_\varepsilon$ is well defined by (31.36) as the uniform limit of a bounded sequence of functions in $C_c(\mathbb{R}^N)$, and thereby in $C_0(\mathbb{R}^N)$. But we also have that

$$||f_m^\varepsilon|_p - |f_n^\varepsilon|_p| \leq |f_m^\varepsilon - f_n^\varepsilon|_p \leq |f_m - f_n|_p.$$

Thereby $f_n^\varepsilon \in C_c(\mathbb{R}^N)$ is a $p$-Cauchy sequence in $L^p(\mathbb{R}^N)$, and $|f_n^\varepsilon|_p$ is a Cauchy sequence in $\mathbb{R}$. Thus

$$L = \lim_{n\to\infty} \int_{\mathbb{R}^N} |f_n^\varepsilon|^p$$

[34]Superscript in $f_n^\varepsilon$ as in Evans, subscript in $F_\varepsilon$ to distinguish from $F^\varepsilon$ in Exercise 31.34.

exists. Suppose that the possibly infinite improper Riemann integral

$$\int_{\mathbb{R}^N} |F_\varepsilon|^p = \lim_{R\to\infty} \int_{B(0,R)} |F_\varepsilon|^p$$

is larger than $L$. Using the two limit definitions it follows that there exist $\delta > 0$ and $R > 0$ such that for all $n$ sufficiently large

$$\int_{B(0,R)} |F_\varepsilon|^p \geq \int_{\mathbb{R}^N} |f_n^\varepsilon|^p + \delta \geq \int_{B(0,R)} |f_n^\varepsilon|^p + \delta.$$

But this is impossible since

$$\int_{B(0,R)} |f_n^\varepsilon|^p \to \int_{B(0,R)} |F_\varepsilon|^p$$

by the uniform convergence of $f_n^\varepsilon$ to $F_\varepsilon$. Thus the improper Riemann integral of $|F_\varepsilon|^p$ over $\mathbb{R}^N$ is finite and

$$\int_{\mathbb{R}^N} |F_\varepsilon|^p \leq \lim_{n\to\infty} \int_{\mathbb{R}^N} |f_n^\varepsilon|^p = L.$$

Can this inequality be strict? In that case there exists $\delta > 0$ such that for all $n$ sufficiently large there exists $R_n > 0$ such that

$$\left(\int_{\mathbb{R}^N} |f_n^\varepsilon|^p\right)^{\frac{1}{p}} > \left(\int_{B(0,R_n)} |f_n^\varepsilon|^p\right)^{\frac{1}{p}} = \left(\int_{\mathbb{R}^N} |F_\varepsilon|^p\right)^{\frac{1}{p}} + \delta.$$

Pick such an $n$ and call it $m$. Since $f_n^\varepsilon$ converges uniformly on $B(0,R_m)$ we have

$$\left(\int_{B(0,R_m)} |f_n^\varepsilon|^p\right)^{\frac{1}{p}} < \left(\int_{B(0,R_m)} |F_\varepsilon|^p\right)^{\frac{1}{p}} + \frac{\delta}{2} \leq \left(\int_{\mathbb{R}^N} |F_\varepsilon|^p\right)^{\frac{1}{p}} + \frac{\delta}{2}$$

$$= \left(\int_{B(0,R_m)} |f_m^\varepsilon|^p\right)^{\frac{1}{p}} - \frac{\delta}{2}.$$

for all $n > m$ sufficiently large. Then

$$\left(\int_{B(0,R_m)} |f_n^\varepsilon|^p\right)^{\frac{1}{p}} < \left(\int_{B(0,R_m)} |f_m^\varepsilon|^p\right)^{\frac{1}{p}} - \frac{\delta}{2},$$

and thereby

$$|f_n^\varepsilon - f_m^\varepsilon|_p \geq \left(\int_{B(0,R_m)} |f_n^\varepsilon - f_m^\varepsilon|^p\right)^{\frac{1}{p}} \geq$$

$$\Big(\int_{B(0,R_m)} |f_m^\varepsilon|^p\Big)^{\frac{1}{p}} - \Big(\int_{B(0,R_m)} |f_n^\varepsilon|^p\Big)^{\frac{1}{p}} > \frac{\delta}{2}.$$

Since $m$ can be chosen as large as we like this prohibits $f_n^\varepsilon$ to be a Cauchy sequence in $L^p(\mathbb{R}^N)$, a contradiction.

This proves

$$\lim_{n\to\infty} \int_{\mathbb{R}^N} |f_n^\varepsilon|^p = \int_{\mathbb{R}^N} |F_\varepsilon|^p, \tag{31.37}$$

and in the following exercise you will complete[35] the proof. □

**Exercise 31.33.** Use (31.37) to show that

$$\lim_{n\to\infty} \int_{\mathbb{R}^N} |f_n^\varepsilon - F_\varepsilon|^p = 0.$$

**Exercise 31.34.** Denote the *equivalence class* $[f_n^\varepsilon]$ by $F^\varepsilon$. Show that

$$\int_{\mathbb{R}^N} F^\varepsilon \phi = \int_{\mathbb{R}^N} F_\varepsilon \phi$$

for every $\phi \in C_c(\mathbb{R}^N)$. Hint: use the uniform convergence of $f_n^\varepsilon$.

**Remark 31.35.** *In Exercise 31.34 we have a p-equivalence class $F^\varepsilon$ of the p-Cauchy sequence $f_n^\varepsilon$ in $C_c(\mathbb{R}^N)$, and a p-integrable function $F_\varepsilon$ in $C_0(\mathbb{R}^N)$. It is easy to see that $F_\varepsilon$ can be represented by a p-Cauchy sequence in $C_c(\mathbb{R}^N)$ as well, for instance by*

$$F_n^\varepsilon(x) = F_\varepsilon(x)\eta(\frac{x}{n}).$$

*As a consequence the sequence $z_n^\varepsilon = F_n^\varepsilon - f_n^\varepsilon$ is also a p-Cauchy sequence in $C_c(\mathbb{R}^N)$. Its equivalence class $Z^\varepsilon$ then has the property that $\int_{\mathbb{R}^N} Z^\varepsilon \phi = 0$ for every $\phi \in C_c(\mathbb{R}^N)$ by the very statement of the exercise. Of course we shall want to conclude that $Z^\varepsilon$ is therefore equal to the equivalence class of the sequence $0, 0, \ldots,$ and Proposition 32.7 will indeed take care of this wish.*

**Exercise 31.36.** Discuss why $F_\varepsilon \in C_c(\mathbb{R}^N)$ if $\Omega$ is bounded. What can you say about its support?

[35] Except for the statements about the derivatives in Theorem 31.37.

**Theorem 31.37.** *Let the function $F_\varepsilon \in C_0(\mathbb{R}^N)$ be defined as in Theorem 31.32 as the limit of the $\varepsilon$-mollifications $f_n^\varepsilon$ of the $p$-Cauchy sequence $f_n$. Then $F_\varepsilon$ is smooth and all its derivatives are $p$-integrable as well. In fact $D^\alpha f_n^\varepsilon \to D^\alpha F_\varepsilon$ uniformly and in $p$-norm as $n \to \infty$ for every multi-index $\alpha$. If $\Omega$ is bounded then $F_\varepsilon \subset C_c^\infty(\mathbb{R}^N)$ and its support is contained in a closed $\varepsilon$-neighbourhood of $\Omega$.*

**Proof**. We define

$$F_\varepsilon^\alpha(x) = (D^\alpha \eta_\varepsilon * F)(x) = \lim_{n\to\infty} \underbrace{(D^\alpha \eta_\varepsilon * f_n)}_{D^\alpha f_n^\varepsilon}(x), \tag{31.38}$$

in which $D^\alpha$ is any partial derivative with multi-index

$$\alpha = (\alpha_1, \ldots, \alpha_N) \quad \text{of order} \quad |\alpha| = |\alpha_1| + |\alpha_N|.$$

We saw in the proof of Theorem 31.32 that for $\varepsilon > 0$ fixed the bounded continuous function $F_\varepsilon = \eta_\varepsilon * F$ is the uniform $n \to \infty$ limit of $f_n^\varepsilon \in C_c^\infty(\mathbb{R}^N)$. Likewise the bounded continuous function $F_\varepsilon^\alpha = D^\alpha \eta_\varepsilon * F$ is the uniform $n \to \infty$ limit of $D^\alpha f_n^\varepsilon$, because

$$\begin{aligned} |D^\alpha f_m^\varepsilon(x) - D^\alpha f_n^\varepsilon(x)| &= |\int_{\mathbb{R}^N} D^\alpha \eta_\varepsilon(x-y)(f_m(y) - f_n(y))\, dy| \\ &\le |D^\alpha \eta_\varepsilon|_q \, |f_m - f_n|_p . \end{aligned}$$

Thus the limit $n \to \infty$ and the derivative $D^\alpha$ commute when acting on $f_n^\varepsilon$. Recalling both (31.36) and (31.38) it follows that $D^\alpha F_\varepsilon$ exists and is given by

$$D^\alpha \eta_\varepsilon * F = F_\varepsilon^\alpha = D^\alpha F_\varepsilon = D^\alpha(\eta_\varepsilon * F). \tag{31.39}$$

The $p$-integrability also follows as in the proof of Theorem 31.32, thanks to the estimate

$$|D^\alpha f_m^\varepsilon - D^\alpha f_n^\varepsilon|_p \le |D^\alpha \eta_\varepsilon|_1 \, |f_m - f_n|_p .$$

In particular

$$\int_{\mathbb{R}^N} |D^\alpha f_n^\varepsilon|^p \to \int_{\mathbb{R}^N} |D^\alpha F_\varepsilon|^p \quad \text{as} \quad n \to \infty.$$

$\square$

**Remark 31.38.** *With the standard definition of Lebesgue integrals the kernels $\eta_\varepsilon$ introduced in Exercise 31.16 are used to mollify locally integrable Lebesgue measurable functions $f$. For $f$ defined on the whole of $\mathbb{R}^N$ we write*[36]

$$f^\varepsilon(x) = (\eta_\varepsilon * f)(x) = \underbrace{\int_{\mathbb{R}^N} \eta_\varepsilon(x-y)f(y)\,dy}_{\text{smooth in } x} = \int_{\mathbb{R}^N} \eta_\varepsilon(y)f(x-y)\,dy, \tag{31.40}$$

*whence*

$$f^\varepsilon(x) = \int_{|y|\le r} \eta_\varepsilon(y)f(x+y)\,dy = \int_{|y-x|\le r} \eta_\varepsilon(y-x)f(y)\,dy.$$

*Such mollified functions $f^\varepsilon$ are globally well defined if $f \in L^1_{loc}(\mathbb{R}^N)$, which contains every $L^p(\Omega)$ with $\Omega$ open and bounded. Repeated differentiation of the first integral in (31.40) with respect to $x$ is then allowed*[37] *and gives*

$$D^\alpha f^\varepsilon(x) = \int_{\mathbb{R}^N} D^\alpha \eta_\varepsilon(x-y)f(y)\,dy \tag{31.41}$$

*for every partial differential operator with multi-index $\alpha$. If $\Omega$ is bounded and $f \in L^p(\Omega)$ then $f^\varepsilon$ is compactly supported and its support is contained in a closed $\varepsilon$-neighbourhood of $\Omega$.*

Think of $D^\alpha$ first as a (partial) derivative with respect to $x$, but then also as an operator acting on smooth functions $u$ to produce new functions $D^\alpha u$. In particular $D^\alpha \eta_\varepsilon$ is a globally defined (smooth) function (with compact support). It appears in (31.41), evaluated in $x-y$ as

$$(D^\alpha \eta_\varepsilon)(x-y) = \underbrace{D^\alpha}_{\text{acts on } x}\ \underbrace{\eta_\varepsilon(x-y)}_{\text{varies with } x}.$$

## 31.9 Towards Sobolev spaces

In case $f_n \in C_c^{|\alpha|}(\Omega)$ the first term in (31.39), that is the first term in

$$D^\alpha \eta_\varepsilon * F = F^\alpha_\varepsilon = D^\alpha F_\varepsilon = D^\alpha(\eta_\varepsilon * F),$$

is the uniform $n \to \infty$ limit of the sequence

$$D^\alpha \eta_\varepsilon * f_n = \eta_\varepsilon * D^\alpha f_n, \tag{31.42}$$

[36] We use a superscript $\varepsilon$ for $f^\varepsilon$ to be consistent with the notation in [Evans].

[37] Theorem 2.27 in [Folland] requires a majorant for which we can take a multiple of $f$.

integrating the defining Riemann integrals by parts. This leads us to consider sequences $f_n \in C_c^k(\Omega)$ for which $D^\alpha f_n$ is a $p$-Cauchy sequence for every $\alpha$ with $|\alpha| \leq k$. Equivalence classes of such sequences define new spaces that we encounter in Chapter 32.

Referring to the density statement in (31.5) we shall consider functions $u$, $u_n$ and $v$ rather than $f$, $f_n$ and $g$, and often think of $v$ as taken from a sequence $u_n$ that converges to $u$ in the $p$-seminorm. Statements about the equivalence class $[u]$ of a function $u$ in $L^p(\Omega)$ can be derived from statements about any $p$-Cauchy sequence $u_n \in C_c(\Omega)$ *which converges to an element*[38] *$u$ of that equivalence class* $[u]$ in the sense that $|u_n - u|_p \to 0$. Likewise statements about the equivalence class $U$ of a $p$-Cauchy sequence $u_n \in C_c(\Omega)$ are derived from statements about any Cauchy sequence *in that class* $U$. To prove something about $[u]$ or $U$ we will pick a function $u_m$ sufficiently far in the $p$-Cauchy sequence $u_n \in C_c(\Omega)$, and it saves us an index to call it $v$. We can choose to consider such $v$ as $v \in [v]$, or as defining the equivalence class $V$ of the sequence $v, v, v, \ldots$.

In both approaches it is needed to also use the mollified functions $v^\varepsilon = u_m^\varepsilon$, sometimes with $\varepsilon > 0$ sufficiently small to have $v^\varepsilon$ in $C_c^\infty(\Omega)$. When considering the entire sequence $u_n^\varepsilon$ with $\varepsilon > 0$ fixed, we have Theorem 31.32 at our disposal which represents the equivalence class $U^\varepsilon$ of the mollified Cauchy sequence $u_n^\varepsilon$ by a function $U_\varepsilon \in C_c^\infty(\mathrm{I\!R}^\mathrm{N})$, supported in an $\varepsilon$-neighbourhood of $\Omega$. This function coincides[39] with the function $u^\varepsilon$ defined by (31.40) for $u \in L^p(\Omega) \subset L^1_{loc}(\mathrm{I\!R}^\mathrm{N})$.

**Remark 31.39.** *It is tempting to mystify notation writing $u$ both for $U$ and for $[u]$, likewise $v$ for $V$ and $[v]$, and all epsilons as superscripts, see Exercise 31.34. Thereby we largely forget about the differences between the two approaches in Chapter 31. But do remember on occasion that $u_m - u$ is either the equivalence class $u_m - U$ of the (Cauchy) sequence $u_m - u_n$ indexed by $n$, or it is chosen in the equivalence class $[u_m - u]$ of the function $u_m - u$.*

[38]To every such element in fact.

[39]Do we ever use this?

# 32 Sobolev spaces with subscript zero

You may try to read this chapter before reading Chapter 31 and flip back when necessary. Let $\Omega$ be an open set in $\mathbb{R}^N$. Recalling that for $1 \le p < \infty$ the $p$-norm of a function $v \in C_c(\Omega)$ is defined by

$$|v|_p^p = \int_{\mathbb{R}^N} |v|^p, \tag{32.1}$$

the Sobolev $W^{1,p}$-norm of a function $v \in C_c^1(\Omega)$ is defined by

$$|v|_{1,p}^p = |v|_p^p + |v_{x_1}|_p^p + \cdots + |v_{x_N}|_p^p. \tag{32.2}$$

No matter which approach to $L^p(\Omega)$ we prefer, the Sobolev space $W_0^{1,p}(\Omega)$ will be the closure[1] of $C_c^1(\Omega)$ with respect to the $W^{1,p}$-norm[2], just like $L^p(\Omega)$ is the closure of $C_c(\Omega)$ with respect to the $p$-norm. The closure is either taken in a not yet defined larger space $W^{1,p}(\Omega)$, or in the abstract sense with equivalence classes, just as described in Section 31.6 for $L^p(\Omega)$ with equivalence classes[3] $U = [u_n]$ of $p$-Cauchy sequences[4] $u_n$ in $C_c(\Omega)$, and definitions like

$$\int_\Omega U\phi = \int_\Omega \phi U = \lim_{n\to\infty} \int_{\mathbb{R}^N} \phi\, u_n, \quad |U|_p = \lim_{n\to\infty} |u_n|_p,$$

and so on[5]. Let us start keeping the latter perspective for $L^p(\Omega)$. Of course these definitions are theorems if we replace $U$ by the limit $u \in L^p(\Omega)$ of the Cauchy sequence with $L^p(\Omega)$ as in Definition 31.6.

**Exercise 32.1.** We know, one way or another, that $L^p(\Omega)$ is the closure of $C_c(\Omega)$ with respect to the $p$-norm. Explain why it is then also the closure of $C_c^1(\Omega)$ with respect to the $p$-norm. Hint: for instance by considering mollifications of functions in $C_c(\Omega)$ as introduced in Exercice 31.16.

## 32.1 Cauchy sequences in the Sobolev norm

Section 31.8 ended with the consideration of sequences $u_n \in C_c^k(\Omega)$ for which $D^\alpha u_n$ is a Cauchy sequence with respect to the $p$-norm for every $\alpha$ with

[1] We consider $C_c^1(\Omega)$ as a subspace of $C_c^1(\mathbb{R}^N)$.

[2] The notation is consistent with [Evans], except for the domains being called $\Omega$.

[3] Denoted by capitals until we decide otherwise and just write $u$ for the class $[u_n]$.

[4] Which have $||u_m|_p - |u_n|_p| \le |u_m - u_n|_p \to 0$ as $m, n \to \infty$, see Definition 31.20.

[5] As in Remark 31.22, with $\phi \in C_c(\mathbb{R}^N)$.

$|\alpha| \leq k$. Restricting to $k = 1$ we denote the first order partial derivatives of a sequence $u_n \in C_c^1(\Omega)$ by $D_1, \ldots, D_N$ acting on $u_n$. From the notations

$$D_i u_n = \frac{\partial u_n}{\partial x_i} = (u_n)_{x_i}$$

we pick the first.

Now observe that a sequence $u_n$ in $C_c^1(\Omega) \subset C_c^1(\mathrm{I\!R}^\mathrm{N})$ is a Cauchy sequence with respect to the norm defined in (32.2) if and only if the sequences $u_n$, $D_1 u_n$, ..., $D_N u_n$ are $p$-Cauchy sequences. This is a much stronger statement than the statement that $u_n$ is a $p$-Cauchy sequence: for any $W^{1,p}$-Cauchy sequence $u_n \in C_c^1(\mathrm{I\!R}^\mathrm{N})$ we have the $p$-equivalence classes of not only the $p$-Cauchy sequence $u_n$ but also of the $p$-Cauchy sequences $D_1 u_n, \ldots, D_N u_n$. For now we denote these classes by capitals $U, W_1, \ldots, W_N$. Note that if we take a sequence independent of $n$, say $v, v, v \ldots$, with $v \in C_c^1(\mathrm{I\!R}^\mathrm{N})$, then we can identify $v$ with its equivalence class, and likewise for the sequences $D^i v, D^i v, D^i v \ldots$ of course.

The equivalence class of such a $W^{1,p}$-Cauchy sequence $u_n$ is given by all sequences $v_n \in C_c^1(\Omega) \subset C_c^1(\mathrm{I\!R}^\mathrm{N})$ with

$$|u_n - v_n|_{1,p} \to 0 \quad \text{as} \quad n \to \infty. \tag{32.3}$$

This class is strictly contained in the $p$-equivalence class $U$, and it comes with $p$-equivalence classes $W_1, \ldots, W_N$, as just explained.

Now recall that Theorem 27.7 stated that

$$\int_\Omega D_i v = \int_{\partial\Omega} \nu_i v, \tag{32.4}$$

for $\Omega$ bounded open in $\mathrm{I\!R}^\mathrm{N}$, $\nu$ the outer normal vector on $\partial\Omega \in C^1$, and $v : \bar{\Omega} \to \mathrm{I\!R}$ continuously differentiable. Applied to $v = \phi u_n$, and $\Omega$ replaced by some large open ball $B_n$ containing the support of a function $u_n \in C_c^1(\mathrm{I\!R}^\mathrm{N})$ under consideration here, we actually don't need Theorem 27.7 to conclude that

$$\int_{\mathrm{I\!R}^\mathrm{N}} \phi\, D_i u_n = \int_{B_n} \phi\, D_i u_n = -\int_{B_n} u_n D_i \phi = -\int_{\mathrm{I\!R}^\mathrm{N}} u_n D_i \phi,$$

because $v = \phi u_n$ vanishes outside $B_n$ and therefore

$$\int_{B_n} \phi\, D_i u_n + u_n D_i \phi = \int_{B_n} D_i(\phi u_n) = 0$$

for every $i$ and every such $\phi$. Thus

$$\forall_{\phi \in C_c^1(\mathrm{I\!R}^\mathrm{N})} \forall_{n \in \mathrm{I\!N}} \quad \underbrace{\int_{\mathrm{I\!R}^\mathrm{N}} \phi\, D_i u_n}_{\to \int \phi W_i} = \underbrace{-\int_{\mathrm{I\!R}^\mathrm{N}} u_n\, D_i \phi}_{\to \int U\, D_i \phi}.$$

By the integral definition in Remark 31.22 this implies as indicated that

$$\forall_{\phi \in C_c^1(\mathrm{I\!R}^{\mathrm{N}})} \quad \int_{\mathrm{I\!R}^{\mathrm{N}}} \phi\, W_i = -\int_{\mathrm{I\!R}^{\mathrm{N}}} U\, D_i \phi. \tag{32.5}$$

By itself this is statement about $p$-equivalence classes *only*, that happens to hold for $U, W_1, \ldots, W_n$ obtained from the $W^{1,p}$-sequence $u_n$ in $C_c^1(\Omega)$ we started out with.

## 32.2 Weak derivatives of equivalence classes

The statement in (32.5) leads us to consider the $p$-equivalence class $W_i$ as the derivative $D^i U$ of the $p$-equivalence class $U$ in some generalised sense, to be made precise with a suitable class of so called test functions $\phi$ for which the identity in (32.5) should hold. For good reasons the test functions $\phi$ are now restricted to $\phi \in C_c^1(\Omega)$, because otherwise the definition fails for functions as in Theorem 27.7, in view of the possible presence of a boundary integral term avoided in (32.5).

After the above discussion the following definition is natural.

**Definition 32.2.** *Let $U$ and $W_i$ be equivalence classes as in Definition 31.20 of $L^p(\Omega)$. We say that $W_i = D_i U$ is the generalised or weak derivative of $U$ with respect to the $i^{th}$ variable if*

$$\forall_{\phi \in C_c^1(\Omega)} \quad \int_\Omega \phi\, W_i = -\int_\Omega U\, D_i \phi. \tag{32.6}$$

*By agreement $W^{1,p}(\Omega)$ without subscript 0 consists of all $U \in L^p(\Omega)$ for which such $W_1, \ldots, W_N \in L^p(\Omega)$ exist. Does it thereby contain $W_0^{1,p}(\Omega)$?*

**Remark 32.3.** *So $W^{1,p}(\Omega)$ is defined above as a subset of*

$$L^p(\Omega) = \{U = [u_n] : \ u_n \in C_c(\Omega) \text{ is a } p\text{-Cauchy sequence}\}$$

*by*

$$W^{1,p}(\Omega) = \{U \in L^p(\Omega) : \ D_1 U, \ldots, D_N U \text{ exist in } L^p(\Omega)\}.$$

*Since*

$$W_0^{1,p}(\Omega) = \{\text{equivalence classes of } W^{1,p}\text{-Cauchy sequences } u_n \in C_c^1(\Omega)\},$$

*the obvious linear injective map from $W_0^{1,p}(\Omega)$ to $W^{1,p}(\Omega)$ is defined by*

$$\underbrace{[u_n]}_{\text{class in } W_0^{1,p}(\Omega)} \quad \to \quad \underbrace{[u_n]}_{\text{class in } L^p(\Omega)} \tag{32.7}$$

*for every $W^{1,p}$-equivalence class of a every $W^{1,p}$-Cauchy sequence $u_n \in C_c^1(\Omega)$.*

**Remark 32.4.** *Changing to lower cases for $U$ and $W_i$ the formulation in (32.6) is consistent with the defining statement in Theorem 32.46 below*[6]*, where we use the standard definition of $L^p(\Omega)$ in Definition 31.6.*

**Remark 32.5.** *To have uniqueness of the weak derivatives $W_i = D_i U$ we need*[7]*, and indeed prove below, that $\int_\Omega W_i \phi = 0$ for all $\phi \in C^1_c(\Omega)$ implies that $W_i$ is the equivalence class of the sequence $0, 0, 0, \dots$ in $C_c(\Omega)$. The obvious definition of the $W^{1,p}$-norm of $U \in W^{1,p}(\Omega)$ is then via*

$$|U|^p_{1,p} = |U|^p_p + |D_1 U|^p_p + \cdots + |D_N U|^p_p.$$

**Remark 32.6.** *Consider $W^{1,p}_0(\Omega)$ as the abstract closure of $C^1_c(\Omega)$ in the $W^{1,p}$-norm. Remark 32.3 then implies that this abstract Banach space is linearly and isometrically embedded by*

$$\underbrace{[u_n]}_{\text{class in } W^{1,p}_0(\Omega)} \to \underbrace{[u_n]}_{\text{class in } L^p(\Omega)}$$

*in the normed space $W^{1,p}(\Omega)$. This latter space is itself also*[8] *a Banach space and contains $C^1_c(\Omega)$, as well as the closure of $C^1_c(\Omega)$. Thus we can also consider $W^{1,p}_0(\Omega)$ as the closure of $C^1_c(\Omega)$ in $W^{1,p}(\Omega)$.*

**Proposition 32.7.** *Let $\Omega \subset \mathbb{R}^N$ be open and let $U = [u_n]$ be the equivalence class of a $p$-Cauchy sequence $u_n$ in $C_c(\Omega)$. If $\int_\Omega U\phi = 0$ for all $\phi \in C^1_c(\Omega)$ then $U = 0$, i.e. $\int_\Omega |u_n|^p \to 0$ as $n \to \infty$.*

**Proof**. We consider $p = 1$ first and argue by contradiction. Suppose that

$$|U|_1 = \lim_{n\to\infty} |u_n|_1 = \lim_{n\to\infty} \int_\Omega |u_n| > 0.$$

We can assume without loss of generality that $|u_n|_1 = 1$ for all $n \in \mathbb{N}$, dividing every $u_n$ by its 1-norm, skipping $n$ for which $|u_n|_1 = 0$. It is easy to see[9] that this does not affect the Cauchy property of the sequence.

As a consequence any function $\phi \in C^1_c(\Omega)$ with $|\phi|_{\max} \le 1$ allows the estimate

$$|\int_\Omega u_n \phi - \int_\Omega u_m \phi| = |\int_\Omega (u_n - u_m)\,\phi| \le |u_n - u_m|_1 < \frac{1}{4}$$

[6] Where we use subscripts $x_i$ to denote $D_i$.
[7] We stumbled on this issue in Section 31.8, see Exercise 31.35.
[8] Exercise 32.9.
[9] Formulate and do the exercise.

for all $m, n$ sufficiently large[10]. It is then no restriction to assume that this estimates holds for all $m, n$. Thus it suffices to exhibit such a $\phi$ for which

$$\int_\Omega u_1\phi > \frac{1}{3}$$

and thereby conclude that

$$\int_\Omega u_n\phi = \underbrace{\int_\Omega u_1\phi}_{>\frac{1}{3}} + \underbrace{\int_\Omega u_n\phi - \int_\Omega u_1\phi}_{>-\frac{1}{4}} > \frac{1}{12}$$

for all $n$. But then

$$\int_\Omega U\phi \geq \frac{1}{12},$$

in contradiction to the asumption that such integrals vanish. A nice but not very sharp[11] exercise about Riemann integrals of compactly supported continuous functions completes the proof for $p = 1$, and also for $p > 1$. □

**Exercise 32.8.** Let $f \in C_c(\mathbb{R}^N)$ have $\int_{\mathbb{R}^N} |f| = 1$. Show there exists a smooth function $\phi$ with compact support contained in the support of $f$ such that

$$|\phi|_{\max} \leq 1 \quad \text{and} \quad \int_\Omega f\phi > \frac{1}{3}.$$

Likewise, if $\int_{\mathbb{R}^N} |f|^p = 1$ for $p > 1$ show there exists such a $\phi$ with

$$|\phi|_q \leq 1 \quad \text{and} \quad \int_\Omega f\phi > \frac{1}{3},$$

and prove Proposition 32.7 for $p > 1$.

**Exercise 32.9.** Remark 32.5 makes $W^{1,p}(\Omega)$ a normed space, thanks to Proposition 32.7. Prove that $W^{1,p}(\Omega)$ is complete. Hint: consider a Cauchy sequence and use that $L^p(\Omega)$ is by definition complete as the abstract closure of $C_c(\Omega)$ in the $p$-norm.

In Definition 32.2 we can replace $\phi \in C_c^1(\Omega)$ by $\phi\zeta$ with $\zeta \in C^1(\mathbb{R}^N)$ to get

$$\int_\Omega \zeta\phi\, D_iU = \int_\Omega W_i\, \zeta\phi = -\int_\Omega U(\zeta D_i\phi + \phi D_i\zeta),$$

[10] For $p > 1$ you would use $|\int_\Omega (u_n - u_m)\,\phi| \leq |u_n - u_m|_q$ with $\frac{1}{p} + \frac{1}{q} = 1$.
[11] We arbitrarily chose $0 < \frac{1}{4} < \frac{1}{3} < 1$.

whence

$$\forall_{\phi\in C_c^1(\Omega)} \quad \int_\Omega (W_i\zeta + UD_i\zeta)\,\phi = -\int_\Omega U\zeta\, D_i\phi. \tag{32.8}$$

This proves the best Leibniz rule we can expect in the $L^p(\Omega)$ context.

**Theorem 32.10.** *Let $\zeta \in C^1(\mathrm{I\!R}^{\mathrm{N}})$ and let $U \in L^p(\Omega)$ have a weak derivative $D_iU \in L^p(\Omega)$, then $\zeta U$ has a weak derivative given by*

$$D_i(U\zeta) = (D_iU)\zeta + U(D_i\zeta)$$

*in $L^p(\Omega)$.*

**Remark 32.11.** *The defining statement in (32.6) of the weak derivative $W_i = D_iU$ also makes sense for $U$ and $W_i$ in $L^1_{loc}(\Omega)$ as in Remark 31.22. The uniqueness of $W_i$ follows along similar lines and Theorem 32.10 holds with $L^p(\Omega)$ replaced by $L^1_{loc}(\Omega)$.*

## 32.3 Mollifiers and density tricks, compactness!

This section could be part of Chapter 31 but is included here to have the present chapter be more self-contained. We first explain how in the Lebesgue approach from Section 31.2 to $L^p(\Omega)$ mollification and density arguments combine to have convergence of mollifiers in $p$-norm. Then we will see that the equivalence class approach to $L^p(\Omega)$ only requires Theorem 32.13.

Recall from Exercise 31.16 and (31.40) that we use mollifiers

$$\eta_\varepsilon(x) = \frac{1}{\varepsilon^{\mathrm{N}}}\eta(\frac{x}{\varepsilon})$$

with

$$\eta(x) = \eta(|x|) \geq 0, \quad \eta \in C_c^\infty(B), \quad \int_{\mathrm{I\!R}^{\mathrm{N}}} \eta = 1,$$

$B$ denoting the open unit ball. As discussed in Section 31.9,

$$u^\varepsilon(x) = (\eta_\varepsilon * u)(x) = \underbrace{\int_{\mathrm{I\!R}^{\mathrm{N}}} \eta_\varepsilon(x-y)u(y)\,dy}_{\text{implies smoothness}} = \underbrace{\int_{|y|\leq\varepsilon} \eta_\varepsilon(y)u(x\pm y)\,dy}_{\text{good for } \varepsilon\to 0} \tag{32.9}$$

defines a function $u^\varepsilon \in C_c^\infty(\mathrm{I\!R}^{\mathrm{N}})$ for all $u$ in the Lebesgue space $L^1_{loc}(\mathrm{I\!R}^{\mathrm{N}})$, but below we prefer to make statements about convergence in $L^p(\mathrm{I\!R}^{\mathrm{N}})$ only, namely

$$u \in L^p(\mathrm{I\!R}^{\mathrm{N}}) \implies |u^\varepsilon - u|_p \to 0 \quad \text{as} \quad \varepsilon \to 0. \tag{32.10}$$

which applies to any $u \in L^p(\Omega)$ extended to the whole of $\mathbb{R}^{\mathrm{N}}$. For bounded $\Omega$ and $u_n$ a bounded sequence in $W_0^{1,p}(\Omega)$ it is then no big deal to obtain a subsequence that converges in $p$-norm.

For $f \in C_c(\mathbb{R}^{\mathrm{N}})$ we noted the Hölder estimate

$$|f^\varepsilon(x)| \leq \int_{|y|\leq\varepsilon} |f(x+y)|\,\eta_\varepsilon(y)^{1=\frac{1}{p}+\frac{1}{q}}\,dy$$

$$\leq \left(\int_{|y|\leq\varepsilon} |f(x+y)|^p\,\eta_\varepsilon(y)\,dy\right)^{\frac{1}{p}} \underbrace{\left(\int_{|y|\leq\varepsilon} \eta_\varepsilon(y)\,dy\right)^{\frac{1}{q}}}_{=1},$$

whence

$$\int_{\mathbb{R}^{\mathrm{N}}} |f^\varepsilon(x)|^p\,dx \leq \int_{x\in\mathbb{R}^{\mathrm{N}}} \int_{y\in\mathbb{R}^{\mathrm{N}}} |f(x+y)|^p\,\eta_\varepsilon(y)\,dy\,dx =$$

$$\int_{y\in\mathbb{R}^{\mathrm{N}}} \underbrace{\int_{x\in\mathbb{R}^{\mathrm{N}}} |f(x+y)|^p\,dx}_{=\int_{\mathbb{R}^{\mathrm{N}}}|f(x)|^p\,dx}\,\eta_\varepsilon(y)\,dy = \int_{\mathbb{R}^{\mathrm{N}}} |f(x)|^p\,dx.$$

This special case[12] of Theorem 32.12 is included in Theorem 32.13 below.

**Theorem 32.12.** *Let $u \in L^p(\mathbb{R}^{\mathrm{N}})$. Then*

$$\int_{\mathbb{R}^{\mathrm{N}}} |u^\varepsilon(x)|^p\,dx \leq \int_{\mathbb{R}^{\mathrm{N}}} |u(x)|^p\,dx, \tag{32.11}$$

*i.e. the p-norm of $u^\varepsilon$ is at most equal to the p-norm of $u$ itself.*

As a consequence we have[13]

$$|u^\varepsilon - v^\varepsilon|_p \leq |u-v|_p \tag{32.12}$$

for $u \in L^p(\mathbb{R}^{\mathrm{N}})$ and $v \in C_c(\mathbb{R}^{\mathrm{N}})$, and likewise that

$$|u_m^\varepsilon - u_n^\varepsilon|_p \leq |u_m - u_n|_p \tag{32.13}$$

for sequences $u_n \in C_c(\mathbb{R}^{\mathrm{N}})$. The latter will suffice for the equivalence class approach in Sections 31.6 and 31.8. The estimates are uniform in the mollifier parameter $\varepsilon$, and are complemented by a convergence estimate for $v \in C_c^1(\mathbb{R}^{\mathrm{N}})$, applicable also to $u_m$ and $u_n$ in (32.13).

[12] The Lebesgue setting requires Fubini's theorem, or a limit argument.

[13] You already noticed we refrain from the use of $\varepsilon$ in smallness estimates.

**Theorem 32.13.** *Let $v \in C_c^1(\mathbb{R}^N)$ and $1 \leq p < \infty$. Then in addition to the estimate*

$$|v^\varepsilon|_p \leq |v|_p,$$

*that holds for all $v \in C_c(\mathbb{R}^N)$ from the reasoning above, we have that*

$$|v^\varepsilon - v|_p \leq \varepsilon \, |\nabla v|_p, \tag{32.14}$$

*as a consequence of the intermediate result*

$$|v^\varepsilon(x) - v(x)| \leq \left( \int_{\mathbb{R}^N} |v(x + \varepsilon y) - v(x)|^p \, \eta(y) dy \right)^{\frac{1}{p}}. \tag{32.15}$$

The right hand side in (32.14) is the $p$-norm of the Euclidean length of $\nabla v$.

**Remark 32.14.** *For $v \in C_c(\mathbb{R}^N)$ we already know from Exercise 31.16 that $v^\varepsilon \in C_c^\infty(\mathbb{R}^N)$ satisfies*

$$|v^\varepsilon|_{max} \leq |v|_{max} \quad \text{and} \quad |v^\varepsilon - v|_{max} \to 0$$

*as $\varepsilon \to 0$.*

Before we prove Theorem 32.13 we note that it is via the above two theorems and the splitting

$$|u - u^\varepsilon|_p \leq |u - v|_p + \underbrace{|v - v^\varepsilon|_p}_{\leq \varepsilon |\nabla v|_p} + \underbrace{|v^\varepsilon - u^\varepsilon|_p}_{\leq |u-v|_p} \tag{32.16}$$

with $v \in C_c^1(\Omega)$ that the following theorem follows.

**Theorem 32.15.** *Let $u \in L^p(\mathbb{R}^N)$, then*

$$|u^\varepsilon|_p \leq |u|_p \quad \text{and} \quad |u^\varepsilon - u|_p \to 0$$

*as $\varepsilon \to 0$. The same statement holds for $p$-equivalence classes $U = [u_n]$ if $U^\varepsilon$ is defined as the $p$-equivalence $U^\varepsilon = [u_n^\varepsilon]$, in which $u_n \in C_c^1(\Omega)$ is a $p$-Cauchy sequence.*

**Exercise 32.16.** Prove the first statement in Theorem 32.15 using Theorems 32.12 and 32.13. Hint: use (32.11) with $u - v$, the density of $C_c^1(\Omega)$ in $L^p(\mathbb{R}^N)$ and (32.14).

**Exercise 32.17.** Let $u_n \in C^1_c(\Omega)$ be a $p$-Cauchy sequence. Use the splitting

$$|u_n - u_n^\varepsilon|_p \leq |u_n - u_m|_p + \underbrace{|u_m - u_m^\varepsilon|_p}_{\leq \varepsilon |\nabla u_m|_p} + \underbrace{|u_m^\varepsilon - u_n^\varepsilon|_p}_{\leq |u_n - u_m|_p}$$

to show for the $p$-equivalence classes $U = [u_n]$ and $U^\varepsilon = [u_n^\varepsilon]$ that

$$|U^\varepsilon|_p \leq |U|_p \quad \text{and} \quad |U - U^\varepsilon|_p \to 0$$

as $\varepsilon \to 0$. This is the second statement in Theorem 32.15 .

**Remark 32.18.** *Theorems 31.32 and 31.37 imply that the p-equivalence class $U^\varepsilon$ is represented by a smooth p-integrable function $U_\varepsilon$ with all its derivatives in $C_0(\mathbb{R}^N)$. It follows from Remark 31.35 and Proposition 32.7 that*

$$U_\varepsilon \to U$$

*in p-norm. In case $\Omega$ is bounded the functions $U_\varepsilon$ are in $C^\infty_c(\mathbb{R}^N)$, just like[14] the function $u^\varepsilon$ defined by (32.9), with the supports contained in closed $\varepsilon$-neighbourhoods of $\Omega$.*

**Proof of Theorem 32.13.** We now prove the $\varepsilon$-estimate (32.14), crucially used for the middle terms in the 3-splittings above. Recall this is about[15] functions $v$ in $C^1_c(\mathbb{R}^N)$. We write

$$v^\varepsilon(x) - v(x) = \int_{\mathbb{R}^N} \eta_\varepsilon(y) v(x+y)\, dy - v(x) = \int_{\mathbb{R}^N} \eta_\varepsilon(y)\left(v(x+y) - v(x)\right)\, dy$$

$$= \int_{\mathbb{R}^N} \eta(y)\left(v(x+\varepsilon y) - v(x)\right)\, dy = \int_{\mathbb{R}^N} \eta(y)^{\frac{1}{q}} \eta(y)^{\frac{1}{p}}\left(v(x+\varepsilon y) - v(x)\right)\, dy.$$

Hölder's inequality gives

$$|v^\varepsilon(x) - v(x)| \leq \underbrace{\left(\int_{\mathbb{R}^N} \eta(y)\, dy\right)^{\frac{1}{q}}}_{=1} \left(\int_{\mathbb{R}^N} \eta(y)|v(x+\varepsilon y) - v(x)|^p\, dy\right)^{\frac{1}{p}}.$$

whence (32.15) follows.

We next use the mean value theorem in integral form, see (12.1), and the Hölder estimate

$$|\int_0^1 f|^p \leq |\int_0^1 |f|^p \qquad (1 \leq p < \infty).$$

[14]See Remark 31.38.

[15]We don't need to know here in which sense this holds for (which?) other functions.

The $x$-integral in the right hand side of (32.15) is estimated for $|y| \leq \varepsilon$ as

$$\int_{\mathrm{IR^N}} |v(x+\varepsilon y) - v(x)|^p \, dx = \int_{\mathrm{IR^N}} |[v(x+t\varepsilon y)]_{t=0}^{t=1}|^p \, dx$$

$$= \int_{\mathrm{IR^N}} | \int_0^1 \nabla v(x+\varepsilon t y) \cdot \varepsilon y \, dt|^p \, dx$$

$$\leq \varepsilon^p \int_{\mathrm{IR^N}} \left( \int_0^1 |\nabla v(x+\varepsilon t y)| \, dt \right)^p dx \leq \varepsilon^p \int_{\mathrm{IR^N}} \int_0^1 |\nabla v(x+\varepsilon t y)|^p \, dt \, dx$$

whence (32.15) gives

$$\int_{\mathrm{IR^N}} |v^\varepsilon(x) - v(x)|^p \, dx \leq \int_{\mathrm{IR^N}} \eta(y) \int_{\mathrm{IR^N}} |v(x+\varepsilon t y) - v(x)|^p \, dx \, dy \leq$$

$$\leq \int_{\mathrm{IR^N}} \eta(y) \, \varepsilon^p \int_{\mathrm{IR^N}} \int_0^1 |\nabla v(x+\varepsilon t y)|^p \, dt \, dx \, dy,$$

and (32.14) follows changing integration order from $dt \, dx \, dy$ to $dx \, dt \, dy$. $\square$

We next establish the compactness of the embedding $W_0^{1,p}(\Omega) \to L^p(\Omega)$ for bounded open sets $\Omega$.

**Theorem 32.19.** *Let $p \geq 1$ and let $\Omega \subset \mathrm{IR^N}$ be bounded and open. Then the embedding*

$$W_0^{1,p}(\Omega) \to L^p(\Omega)$$

*is compact: every bounded sequence $u_n$ in $W_0^{1,p}(\Omega)$ has a subsequence which, considered as a sequence in $L^p(\Omega)$, is convergent. Here $W_0^{1,p}(\Omega) \subset L^p(\Omega)$ is defined one way or another*[16] *as a Banach space in which $C_c^1(\Omega)$ is dense with respect to the $W^{1,p}$-norm.*

**Proof**. We use the splitting

$$|u_n - u_m|_p \leq \underbrace{|u_n - v_n|_p}_{\to 0} + \underbrace{|v_n - v_n^\varepsilon|_p}_{\leq \varepsilon \, |\nabla v_n|_p} + |v_n^\varepsilon - v_m^\varepsilon|_p + \underbrace{|v_m^\varepsilon - v_m|_p}_{\leq \varepsilon \, |\nabla v_m|_p} + \underbrace{|v_m - u_m|_p}_{\to 0},$$

and deal with with the first and fifth term by density of $C_c^1(\Omega)$ in $W_0^{1,p}(\Omega)$ taking $m, n$ large. To be precise, choose $v_n \in C_c^1(\Omega)$ with $|u_n - v_n|_{1,p} \to 0$ as $n \to \infty$. Then $\nabla v_n$ is bounded in $p$-norm, whence the second and fourth term are controled by (32.14) and as small as we like taking $\varepsilon > 0$ small independent of $m, n$. For the middle term we invoke the Ascoli-Arzelà

[16]The choice between the two approaches to introduce $L^p(\Omega)$ is not relevant here.

Theorem[17], using $\varepsilon$-dependent bounds on $v_m^\varepsilon(x), v_n^\varepsilon(x)$ and $\nabla v_m^\varepsilon(x), \nabla v_n^\varepsilon(x)$ uniform in $x$ and $m, n$, to get Cauchy subsequences of $v_n^\varepsilon$ in the maximum norm, for every $\varepsilon > 0$ fixed. These subsequences are also $p$-Cauchy sequences. Exercise 32.20 then completes the proof. $\square$

**Exercise 32.20.** Use a diagonal argument to show that $u_n$ itself has a $p$-Cauchy subsequence and explain why this completes the proof.

## 32.4 Estimates and embeddings for $W_0^{1,p}(\Omega)$

Estimates derived for functions in $C_c^1(\Omega)$ carry over to functions in $W_0^{1,p}(\Omega)$. We discuss two important such estimates, namely the Hölder continuity estimate (32.19) for $p > N$, and the Gagliardo-Nirenberg-Sobolev estimate

$$|u|_q \leq C_{p,\mathrm{N}} |\nabla u|_p \quad \text{for} \quad \frac{1}{q} = \frac{1}{p} - \frac{1}{\mathrm{N}} \quad \text{if} \quad 1 \leq p < \mathrm{N}. \tag{32.17}$$

The proof of (32.17) relies on repeated application of the one-dimensional estimate[18]

$$|u|_{\max} \leq \frac{1}{2} |u'|_1 \quad \text{for} \quad u \in C_c^1(\mathbb{R}), \tag{32.18}$$

first in the special case that $p = 1$, and for $N \geq 3$ by a clever repeated[19] use of the generalised Hölder's inequality in which the exponents are all taken equal. The general case in (32.17) then follows from putting $u^\gamma$ 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 $\gamma$ to get the exponents right. The constant $C_{p,\mathrm{N}}$ blows up as $p \to \mathrm{N}$ (from below). For the computational details and additional exercises see Section 32.5.

The second (Morrey) estimate is usually stated as

$$|u(x_1) - u(x_2)| \leq C_{p,\mathrm{N}} |\nabla u|_p |x_1 - x_2|^\alpha \quad \text{for} \quad \alpha = 1 - \frac{\mathrm{N}}{p} \quad \text{if} \quad p > \mathrm{N},$$

but the $p$-norm of $\nabla u$ 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

$$|u(x_1) - u(x_2)| \leq C_{p,N} |\nabla u|_{L^p(W_{x_1 x_2})} |x_1 - x_2|^{1-\frac{N}{p}}, \tag{32.19}$$

[17]Theorem 8.13 with $[0,1]$ replaced by a large closed box.

[18]The case $p = N = 1, q = \infty$ in (32.17), does not generalise to $p = N > 1, q = \infty$.

[19]For $N = 2$ a single application of Cauchy-Schwartz.

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

$$\int_{C_R} |u - u(0)| \leq \frac{R^{\mathrm{N}}}{\mathrm{N}} \int_{C_R} \frac{|u_r|}{r^{\mathrm{N}-1}} \tag{32.20}$$

for cones described in polar coordinates as

$$C_R = \{x = r\omega : 0 \leq r \leq R,\, \omega \in A\},$$

with $A$ a nice subset of the unit sphere, and $u_r$ denoting the radial derivative. The $r$-part of the integral on the left in (32.20) is in some sense the counter part of (32.18), and the integral on the right hand side is estimated using a follow your nose estimate invoking Hölder's inequality. The Morrey estimate (32.19) is then proved estimating

$$|u(x_1) - u(x_2)| \leq |u(x_1) - u(x)| + |u(x) - u(x_2)|,$$

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. For the details and additional exercises see Section 32.6. Again the constant $C_{p,\mathrm{N}}$ blows up as $p \to \mathrm{N}$ (from above).

## 32.5 Proof of the GNS-estimates

In this section we use subscripts for partial derivatives. With Remark 32.22 we establish a statement that implies (32.17). Let's do the arguments for increasing values of the dimension. It is easy to see that

$$|f(x)| \leq \frac{1}{2} \int_{-\infty}^{\infty} |f'(x)|\, dx$$

for $f \in C_c^1(\mathbb{R})$. Applied to $x \to u(x,y)$ and $y \to u(x,y)$ we have for $u \in C_c^1(\mathbb{R}^2)$ that

$$|u(x,y)|^2 \leq \underbrace{\frac{1}{2} \int_{-\infty}^{\infty} |u_x(x,y)|\, dx}_{\text{depends on } y \text{ only}} \; \underbrace{\frac{1}{2} \int_{-\infty}^{\infty} |u_y(x,y)|\, dy}_{\text{depends on } x \text{ only}},$$

and thus it follows for $u \in C_c^1(\mathbb{R}^2)$ that

$$\iint_{\mathbf{R}^2} |u|^2 \leq \frac{1}{4} \iint_{\mathbf{R}^2} |u_x| \iint_{\mathbf{R}^2} |u_y|,$$

whence

$$|u|_2 \le \frac{1}{2}|u_x|_1^{\frac{1}{2}}|u_y|_1^{\frac{1}{2}}.$$

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 $p_1 = p_2 = \frac{1}{2}$ to the successive $x, y, z$-integrations, 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, the subscripts indicate which variable has been integrated away.

**Exercise 32.21.** Prove that

$$|u|_{\frac{3}{2}} \le \frac{1}{2}|u_x|_1^{\frac{1}{3}}|u_y|_1^{\frac{1}{3}}|u_z|_1^{\frac{1}{3}}$$

generalises to

$$|u|_{\frac{N}{N-1}} \le \frac{1}{2}|u_{x_1}|_1^{\frac{1}{N}}\cdots|u_{x_N}|_1^{\frac{1}{N}} = \frac{1}{2}\Pi_{i=1,\ldots,N}|u_{x_i}|_1^{\frac{1}{N}} \le \frac{1}{2}|\nabla u|_1$$

for $u \in C_c^1(\mathrm{I\!R}^{\mathrm{N}})$ via $N$ integrations and Hölder's generalised inequality with $N-1$ equal exponents $\frac{1}{N-1}$ applied in every step. Use that 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[20]

$$\frac{1}{p} + \frac{1}{p'} = 1$$

[20]We use $p'$ instead of $q$ which is already in use.

that

$$|u|^{\gamma}_{\frac{\gamma N}{N-1}} = |u^{\gamma}|_{\frac{N}{N-1}} \le \frac{1}{2}\,\Pi_{i=1,\dots,N}|\gamma u^{\gamma-1}u_{x_i}|_1^{\frac{1}{N}} \le \frac{\gamma}{2}\,\Pi_{i=1,\dots,N}|u^{\gamma-1}|_{p'}^{\frac{1}{N}}|u_{x_i}|_p^{\frac{1}{N}}$$

$$= \frac{\gamma}{2}|u|^{\gamma-1}_{\frac{(\gamma-1)p}{p-1}}\,\Pi_{i=1,\dots,N}|u_{x_i}|_p^{\frac{1}{N}}$$

in which $\gamma$ can be chosen to have equal subscripts of $|u|$ in the first and last expression in this chain.

**Exercise 32.22.** 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 $\gamma$? Dividing by $|u|_q^{\gamma-1}$ on both sides you get

$$C_{Np}|u_{x_1}|_p^{\frac{1}{N}}\cdots|u_{x_N}|_p^{\frac{1}{N}} \le C_{Np}|\nabla u|_p$$

with an explicit constant $C_{Np}$. Give this value. Check again that $1 \le p < N$ is the assumption to make here.

**Remark 32.23.** *We have proved for $u \in C_c^1(\Omega)$ that*

$$|u|_q \le C_{Np}|u_{x_1}|_p^{\frac{1}{N}}\cdots|u_{x_N}|_p^{\frac{1}{N}} \le C_{Np}|\nabla u|_p$$

*This estimate holds in fact for all $u \in W_0^{1,p}(\Omega)$ if $1 \le p < N$ and*

$$\frac{1}{q} = \frac{1}{p} - \frac{1}{N}.$$

*In Exercise 32.27 we play a bit towards similar statements about $W_0^{2,p}(\Omega)$.*

**Exercise 32.24.** Let $1 \le p < \mathrm{N}$ and let $\Omega$ be bounded. Use (32.17) to prove that

$$W_0^{1,p}(\Omega) \subset L^q(\Omega) \quad \text{if} \quad \frac{1}{q} \ge \frac{1}{p} - \frac{1}{\mathrm{N}},$$

and that $|\nabla u|_p$ defines an equivalent norm on $W_0^{1,p}(\Omega)$. In particular

$$|u|_p \le C_{pq\Omega}|\nabla u|_p \quad \text{for all} \quad u \in W_0^{1,p}(\Omega),$$

with $C_{pq\Omega}$ a constant depending on $p$, $q$ and $\Omega$ only[21]. Make $C_{pq\Omega}$ as explicit as possible in terms of $p$, $N$ and the measure of $\Omega$.

**Exercise 32.25.** A special case in Exercise 32.24 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 $W_0^{1,p}(\Omega)$, which was defined as the closure of $C_c^1(\Omega)$ in $W^{1,p}(\Omega)$ 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 $N \le p < \infty$.

**Exercise 32.26.** Let $1 \le p < \mathrm{N}$ and $\Omega \subset \mathrm{I\!R}^\mathrm{N}$ open and bounded. Prove that the embedding

$$W_0^{1,p}(\Omega) \to L^q(\Omega)$$

is compact if

$$\frac{1}{q} > \frac{1}{p} - \frac{1}{\mathrm{N}}.$$

Hint: use Theorem 32.19 and interpolation inequalities[22] with the $p$-norms.

**Exercise 32.27.** Show for $N > 2$ that[23]

$$|u|_{\frac{N}{N-2}} \le \frac{1}{4} \max_{i \ne j} |u_{x_i x_j}|_1$$

for $u \in C_c^2(\mathrm{I\!R}^\mathrm{N})$. Only the mixed derivatives are needed[24].

---

[21] And thereby also on the dimension $N$.
[22] Exercise 31.5.
[23] There are similar estimates for $u \in C_c^3(\mathrm{I\!R}^\mathrm{N})$, $u \in C_c^4(\mathrm{I\!R}^\mathrm{N})$, . . . .
[24] Nice project: versions similar to Exercise 32.23 for $W_0^{2,p}$ with only mixed derivatives.

## 32.6 Proof of the Morrey estimates

We first take $N = 3$ and $u \in C^1(\mathbb{R}^3)$. Write $u(x, y, z) = v(r, \theta, \phi)$, with spherical coordinates

$$x = r \sin\theta \cos\phi,\ y = r \sin\theta \sin\phi,\ z = r \cos\theta,$$

and let $C_{R,\psi}$ be the set in $\mathbb{R}^3$ defined by $0 \le r \le R$, $0 \le \theta \le \psi$ and $\phi$ free. If $0 < \psi < \frac{\pi}{2}$ and $R > 0$, then[25] $\psi$ is called[26] the opening angle of the closed cone $C_{R,\psi}$, and $R$ is called the radius of the cone[27].

Assuming that $u(0, 0, 0) = v(0, \theta, \phi) = 0$ we use

$$|v(r, \theta, \phi)| \le \int_0^r |v_r(\rho, \theta, \phi)|\, d\rho$$

to estimate

$$\int_{C_{R,\psi}} |u| = \int_0^\psi \int_0^{2\pi} \int_0^R |v(r, \theta, \phi)|\, r^2 \sin\theta\, dr\, d\phi\, d\theta$$

$$\le \int_0^\psi \int_0^{2\pi} \int_0^R \int_0^r |v_r(\rho, \theta, \phi)|\, d\rho\, r^2 \sin\theta\, dr\, d\phi\, d\theta$$

(interchanging the order of the integrations with respect to $r$ and $\rho$, throwing a away one negative term and replacing $\rho$ by $r$ again)

$$\le \frac{R^3}{3} \int_0^\psi \int_0^{2\pi} \int_0^R \frac{|v_r|}{r^2}\, r^2 \sin\theta\, dr\, d\phi\, d\theta = \frac{R^3}{3} \int_{C_{R,\psi}} \frac{|u_r|}{r^2},$$

in which $u_r = xu_x + yu_y + zu_z = v_r$. Thus

$$\int_{C_{R,\psi}} |u| \le \frac{R^3}{3} \int_{C_{R,\psi}} \frac{|u_r|}{r^2} \tag{32.21}$$

**Exercise 32.28.** Use generalised polar coordinates

$$x_1 = r\omega_1 = r\cos\theta_1 = rc_1,\ x_2 = r\omega_2 = r\sin\theta_1\cos\theta_2 = rs_1c_2,\ x_3 = r\omega_3 = rs_1s_2c_3,$$

$$\ldots, x_{N-2} = r\omega_{N-1} = rs_1 \cdots s_{N-2}c_{N-1},\ x_N = r\omega_N = rs_1 \cdots s_{N-2}s_{N-1}$$

[25] $C_{R,\psi}$ a not cone for $\psi > \frac{\pi}{2}$, it's a half ball with radius $R$ if $\psi = \frac{\pi}{2}$ and a ball if $\psi = \pi$.
[26] And not $2\psi$.
[27] Evans only integrates over balls.

to generalise and improve[28] (32.21) as

$$\int_{C_{R,\psi}} |u| + \frac{1}{N}\int_{C_{R,\psi}} r|u_r| \le \frac{R^N}{N}\int_{C_{R,\psi}} \frac{|u_r|}{r^{N-1}} \tag{32.22}$$

for $C_{R,\psi}$ in $\mathbb{R}^{\mathrm{N}}$ defined by $0 \le r \le R$, $0 \le \theta_1 \le \psi$ and $\theta_2, \ldots, \theta_{N-1}$ free.

**Exercise 32.29.** (continued) Let $\omega \in \mathbb{R}^{\mathrm{N}}$ with $|\omega| = 1$. Explain why estimate (32.22) holds with $C_{R,\psi}$ replaced by the closed cone

$$C_{R,\omega,\psi} = \{x \in \mathbb{R}^{\mathrm{N}} : |x| \le R,\, x \cdot \omega \ge |x| \cos\psi\}, \tag{32.23}$$

a cone with direction[29] $\omega$ and opening angle $\psi \in (0, \frac{\pi}{2})$.

**Exercise 32.30.** Not so important: explain why the measure of the cone $C_{1,\omega,\psi}$ defined by (32.23) is given by

$$\int_{C_{1,\omega,\psi}} 1 = \frac{2\pi}{N}\int_0^\psi \sin^{N-2}\theta_1\, d\theta_1 \int_0^\pi \sin^{N-1}\theta_2\, d\theta_2 \cdots \int_0^\pi \sin\theta_{N-2}\, d\theta_{N-2}. \tag{32.24}$$

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 $\omega_{N-1}$ is the measure of the unit ball in $\mathbb{R}^{N-1}$. Correct my mistakes if this formula is wrong[30].

**Exercise 32.31.** Hölder's inequality[31] gives[32]

$$\int_{C_{R,\omega,\psi}} \frac{|u_r|}{r^{N-1}} \le \left(\int_{C_{R,\omega,\psi}} \left(\frac{1}{r^{N-1}}\right)^{p'}\right)^{\frac{1}{p'}} \underbrace{\left(\int_{C_{R,\omega,\psi}} |u_r|^p\right)^{\frac{1}{p}}}_{|u_r|_{L^p(C_{R,\omega,\psi})}}$$

---

[28] Don't throw that negative term away now but bring it to the other side.

[29] Note $\omega = e_3$ in the 3-dimensional example, $\omega = e_1$ in the $N$-dimensional example.

[30] Is there a quicker way? Does the integral simplify if $\psi = \frac{\pi}{3}$? Not so important.

[31] Which for integrals follows from the inequality in Section 17.3.

[32] With $\frac{1}{p} + \frac{1}{p'} = 1$.

for the right hand side of (32.22). Show that

$$\int_{C_{R,\omega,\psi}} \frac{|u_r|}{r^{N-1}} \leq \left( C_{N\psi} \, \frac{p-1}{p-N} \right)^{1-\frac{1}{p}} |u_r|_{L^p(C_{R,\omega,\psi})} \, R^{1-\frac{N}{p}} \tag{32.25}$$

if $p > N$. Explain why the estimate holds for all $u \in C^1(C_{R,\omega,\psi})$. Why does the estimate fail for $p \leq N$?

Combining (32.22) and (32.25) we have

$$\frac{N}{R^N} \int_{C_{R,\omega,\psi}} |u| \leq \left( C_{N\psi} \, \frac{p-1}{p-N} \right)^{1-\frac{1}{p}} |u_r|_{L^p(C_{R,\omega,\psi})} \, R^{1-\frac{N}{p}}, \tag{32.26}$$

in which $\psi$ can have any value in $[0, \pi]$.

**Exercise 32.32.** For $R > 0$, $x_1, x_2, \omega_1, \omega_2 \in \mathrm{I\!R}^{\mathrm{N}}$ with $|\omega_1| = |\omega_2| = 1$ and angles $\psi_1, \psi_2$, consider $C_1 = x_1 + C_{R,\omega_1,\psi_1}$ and $C_2 = x_2 + C_{R,\omega_2,\psi_2}$. Use

$$|C_1 \cap C_2| \, |u(x_1) - u(x_2)| \leq \int_{C_1} |u(x_1) - u(x)| dx + \int_{C_2} |u(x) - u(x_2)| dx$$

and (32.26) to show that

$$|C_1 \cap C_2| \, |u(x_1) - u(x_2)| \leq \frac{R^N}{N} \left( C_{N\psi_1}^{1-\frac{1}{p}} + C_{N\psi_2}^{1-\frac{1}{p}} \right) \left( \frac{p-1}{p-N} \right)^{1-\frac{1}{p}} |\nabla u|_{L^p(C_1 \cup C_2)} \, R^{1-\frac{N}{p}}.$$

**Exercise 32.33.** In Exercise 32.32 take[33]

$$R = |x_1 - x_2|, \; \omega_1 = \frac{x_2 - x_1}{R} = -\omega_2, \; \psi = \frac{\pi}{3},$$

and prove that

$$|u(x_1) - u(x_2)| \leq C(N,p) |\nabla u|_{L^p(B_{|x_1-x_2|}(x_1) \cap (B_{|x_1-x_2|}(x_2))} \, |x_1 - x_2|^{1-\frac{N}{p}}, \tag{32.27}$$

in which $C(N,p)$ is a constant that can be made explicit if you insist.

[33] Sketch the balls with centers $x_1$ and $x_2$ and radius $|x_1 - x_2|$ to see what's going on.

**Exercise 32.34.** If so, show that

$$C(N,p) = c_N \frac{\left(\int_0^{\frac{\pi}{3}} \sin^{N-2}\theta \, d\theta\right)^{1-\frac{1}{p}}}{\omega_{N-1}{}^{\frac{1}{p}}} \tag{32.28}$$

with $c_N$ to be specified. 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 $2A_N R^N$ 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 $N > 3$.

**Exercise 32.35.** Let $p > \mathrm{N}$. Prove that $W_0^{1,p}(\Omega) \subset C^\alpha(\Omega)$ for $\alpha = 1 - \frac{\mathrm{N}}{p}$, in which

$$C^\alpha(\Omega) = \{u \in C(\Omega) : [u]_\alpha < \infty\} \quad \text{where} \quad [u]_\alpha = \sup_{x_1 \neq x_2} \frac{|u(x_1) - u(x_2)|}{|x_1 - x_2|^\alpha},$$

the supremum taken over the whole of $\Omega$.

**Exercise 32.36.** Let $\Omega$ be bounded and $\alpha \in (0,1]$. Then[34] every $u \in C^\alpha(\Omega)$ extends to a continuous function on $\bar{\Omega}$, and $C^\alpha(\Omega)$ is a Banach space with norm defined by[35] $|u|_\alpha = |u|_{\max} + [u]_\alpha$.

**Exercise 32.37.** Let $p > \mathrm{N}$, $\Omega$ bounded, $\alpha = 1 - \frac{\mathrm{N}}{p}$. Use the Ascoli-Arzelà Theorem to prove that every bounded sequence $u_n$ in $W_0^{1,p}(\Omega)$ has a subsequence that, considered as a sequence in $C(\bar{\Omega})$, converges uniformly to a limit $u$, which is also in $C_0^\alpha(\Omega)$, the subspace consisting of functions $u \in C^\alpha(\Omega)$ which vanish on $\partial\Omega$. Verify that this subspace has the seminorm $[\cdot]_\alpha$ as an equivalent norm.

[34]This is to convince you that maybe it is better to rename $C^\alpha(\Omega)$ and write $C^\alpha(\bar{\Omega})$.

[35]If you don't use Greek letters for Lebesgue norms this will not confuse.

**Remark 32.38.** *In view of Exercise 32.37 the embedding*

$$W_0^{1,p}(\Omega) \to C(\bar{\Omega})$$

*is compact for* $p > \mathrm{N}$ *if* $\Omega$ *is bounded. It then also holds that*

$$W_0^{1,p}(\Omega) \to L^p(\Omega) \quad \text{is compact if} \quad \Omega \quad \text{is bounded,} \tag{32.29}$$

*but this was already shown for all* $p \geq 1$ *in Theorem 32.19 via different arguments*[36].

Now recall the definitions of $W^{1,p}(\Omega)$ and $W_0^{1,p}(\Omega)$ for $\Omega$ in $\mathrm{I\!R}^{\mathrm{N}}$ bounded, open and connected,

$$u \in W^{1,p}(\Omega) \iff u, u_{x_1}, \dots, u_{x_N} \in L^p(\Omega),$$

and, for $1 \leq p < \infty$, the space $W_0^{1,p}(\Omega)$ being the closure of $C_c^1(\Omega)$ in the Banach space $W^{1,p}(\Omega)$.

**Exercise 32.39.** Let $u \in W_0^{1,p}(\Omega)$, $\Omega$ in $\mathrm{I\!R}^{\mathrm{N}}$ bounded, open and connected, $N < p < \infty$ and let $\alpha = 1 - \frac{N}{p}$. Take a sequence $u_n \in C_c^\infty(\Omega)$ with $u_n \to u$ in $W^{1,p}(\Omega)$. Prove that $u_n$ is a Cauchy sequence in $C^\alpha(\bar{\Omega})$, and that its limit $\bar{u}$ in $C^\alpha(\bar{\Omega})$ has the property that $|u - \bar{u}|_p = 0$. Prove that the map $u \to \bar{u}$ is linear and continuous from $W_0^{1,p}(\Omega)$ to $C^\alpha(\bar{\Omega})$.

**Exercise 32.40.** (continued) A rough estimate for the seminorm

$$[u]_\alpha = \sup_{\substack{x,y\in\Omega \\ x\neq y}} \frac{|u(x) - u(y)|}{|x-y|^\alpha}$$

with $\alpha = 1 - \frac{N}{p}$: show that

$$[u]_{1-\frac{N}{p}} \leq C(p,N)|\nabla u|_{L^p(\Omega)},$$

and also show that

$$|u|_\infty \leq \tilde{C}(p,N,\Omega)|\,|\nabla u|_{L^p(\Omega)}$$

for some constant $\tilde{C}(p,N,\Omega)$ you can make as precise as you want. Give just one. Hint: first for $u \in C_c^1(\Omega)$, reason as in Exercise 32.39 to get the estimate for all $u \in W_0^{1,p}(\Omega)$ if $p > N$.

[36] Including the Ascoli-Arzelà Theorem.

**Exercise 32.41.** Show that $C^\alpha(\bar\Omega)$ is a Banach space.

**Exercise 32.42.** Show that $W_0^{1,p}(\Omega)$ is compactly embedded in $C_0(\Omega)$. Hint: use the Ascoli-Arzelà theorem via Exercise 32.40.

**Exercise 32.43.** Let $0<\beta<\alpha<1$. Show that

$$[u]_\beta \le [u]_\alpha + C_{\alpha\beta}|u|_\infty,$$

in which $C_{\alpha\beta}$ is a constant depending on $\alpha$ and $\beta$ only. Hint: it is easy to estimate $[u]_\alpha$ by a product of powers of $[u]_\beta$ and $|u|_\infty$. Use Young's inequality

$$ab \le \frac{\varepsilon^p a^p}{p} + \frac{b^q}{q\varepsilon^q} \quad \text{for} \quad \varepsilon>0,\ a,b\ge 0,\ p,q\ge 1 \quad \text{with} \quad \frac{1}{p}+\frac{1}{q}=1$$

to conclude.

**Exercise 32.44.** Use Exercises 32.42 and 32.43 to conclude that $W_0^{1,p}(\Omega)$ is compactly embedded in $C^\beta(\bar\Omega)$ is $0<\beta<1-\frac{N}{p}$.

**Exercise 32.45.** Show that $W_0^{1,p}(\Omega)$ is embedded in $h^\alpha(\bar\Omega)$, the closed subspace[37] of $C^\alpha(\bar\Omega)$ for which

$$\sup_{\substack{x,y\in\Omega\\ 0<|x-y|\le\varepsilon}} \frac{|u(x)-u(y)|}{|x-y|^\alpha} \to 0 \quad \text{as} \quad \varepsilon\to 0,$$

if $\alpha = 1-\frac{N}{p}$ and $p>N$.

## 32.7 Weak derivatives of Lebesgue functions

The approach in Section 31.2 and the concept of weak derivative lead to the following theorem.

[37]These are the so-called little Hölder spaces, unlike $C^\alpha(\bar\Omega)$ they are separable.

**Theorem 32.46.** *Suppose that*

$$\int_\Omega v\phi = -\int_\Omega u\phi_{x_i} \tag{32.30}$$

*for every $\phi \in C_c^1(\Omega)$, for some given $u$ and $v$ in $L^1_{loc}(\Omega)$, $\Omega \subset \mathrm{I\!R}^{\mathrm{N}}$. Then $v$ is unique in $L^1_{loc}(\Omega)$ for $u \in L^1_{loc}(\Omega)$. We say that $v$ is the weak derivative of $u$ with respect to its $i^{th}$ variable, notation $v = D_i u = u_{x_i}$.*

**Proof**. The proof uses the full Lebesgue machinery that we happily avoided in Section 32.2. Suppose that some other $v$, say $\tilde{v} \in L^1_{loc}(\Omega)$ also satisfies this condition, then the difference $w = v - \tilde{v}$ is also $L^1_{loc}(\Omega)$ and satisfies

$$\int_\Omega w\phi = 0$$

for every $\phi \in C_c^1(\Omega)$. Take an open ball $B \subset \Omega$ and redefine $w(x) = 0$ for $x \notin B$. The mollifier $w^\varepsilon$ is then identically equal to zero on $\mathrm{I\!R}^{\mathrm{N}}$ for all $\varepsilon > 0$. By Remark 31.38 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 31.12 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 $B \subset \Omega$ open. Thus $v = \tilde{v}$ in $\Omega$ outside a set of measure zero. $\square$

**Definition 32.47.** *For $1 \leq p < \infty$ and $\Omega \subset \mathrm{I\!R}^{\mathrm{N}}$ the space $W^{1,p}(\Omega)$ is defined as the space of functions*[38] *$u \in L^p(\Omega)$ for which the weak derivatives $u_{x_1}, \dots, u_{x_N}$ exist and are in $L^p(\Omega)$.*

**Exercise 32.48.** Prove that $W^{1,p}(\Omega)$ is a Banach space with the norm defined by

$$|u|^p_{1,p} = |u|^p_p + |u_{x_1}|^p_p + \cdots + |u_{x_{\mathrm{N}}}|^p_p. \tag{32.31}$$

Hint: you need to use that $L^p(\Omega)$ is a Banach space when you consider a Cauchy sequence $u_n$ in $W^{1,p}(\Omega)$.

**Remark 32.49.** *Once again $W_0^{1,p}(\Omega)$ may be defined as the closure of $C_c^1(\Omega)$ in $W^{1,p}(\Omega)$. Since $W^{1,p}(\Omega)$ is Banach space, the space $W_0^{1,p}(\Omega)$ is again complete as a closed subspace of $W^{1,p}(\Omega)$.*

[38]With the equivalence relation that identifies functions differing on zero measure sets.

**Remark 32.50.** *Every $u \in W_0^{1,p}(\Omega)$ extends to a $u \in W_0^{1,p}(\mathbb{R}^N)$ by setting $u$ equal to zero outside $\Omega$. The similar statement $u \in C_c^1(\Omega)$ and $C_c^1(\mathbb{R}^N)$ already implied the same statement in the equivalence class context. And we have already proved the following theorem in Exercise 32.20.*

**Remark 32.51.** *The definitions of $W^{2,p}(\Omega)$ and of $W_0^{2,p}(\Omega)$ (the closure of $C_c^k(\Omega)$ with $k \geq 2$) should be obvious, starting from the norm defined by*

$$|u|_{2,p}^p = |u|_p^p + \sum_{1\leq i\leq N} |u_{x_i}|_p^p + \sum_{1\leq i\leq j\leq N} |u_{x_i x_j}|_p^p.$$

**Exercise 32.52.** Fill in the details of Remark 32.51 and generalise to $W^{k,p}(\Omega)$ and $W_0^{k,p}(\Omega)$ with $k \geq 2$. Again you may prefer the approach in which functions in $L^p(\Omega)$ are equivalence classes of $p$-Cauchy sequences in $C_c(\Omega)$, or $C_c^\infty(\Omega)$ for that matter.

## 32.8 Sobolev spaces for $\Omega = \mathbb{R}^N$

We finish this chapter by showing that the subscript zero has no meaning when $\Omega = \mathbb{R}^N$.

**Theorem 32.53.** *Let $u \in W^{1,p}(\mathbb{R}^N)$, $1 \leq p < \infty$. Then*

$$u^\varepsilon \to u \quad \text{in} \quad W^{1,p}(\mathbb{R}^N) \quad \textit{as} \quad \varepsilon \to 0.$$

**Proof**. In the Lebesgue approach let $u \in W^{1,p}(\mathbb{R}^N)$ and $u_{x_i} = D_i u \in L^p(\mathbb{R}^N)$. Then[39]

$$(u_{x_i})^\varepsilon(x) = (D_i u)^\varepsilon(x) = \int_{\mathbb{R}^N} \eta_\varepsilon(x-y)(D_i u)(y)\,dy \tag{32.32}$$

$$= \int_{\mathbb{R}^N} (D_i \eta_\varepsilon)(x-y)u(y)\,dy = (\int_{\mathbb{R}^N} \eta_\varepsilon(x-y)u(y)\,dy)_{x_i} = (D_i u^\varepsilon)(x),$$

so

$$D_i(u^\varepsilon) = (u^\varepsilon)_{x_i} = (u_{x_i})^\varepsilon = (D_i u)^\varepsilon. \tag{32.33}$$

Since Theorems 32.12 and 32.13 apply to both $u$ and $D_i u$ the limit statement in the theorem follows.

Alternatively we consider $u$ and $w_i = D_i u$ as $p$-equivalence classes $U$ and $W_i = D_i U$, and use the functions $F_\varepsilon$ as in Exercise 31.34 of Section 31.8,

[39]Note that $D_i$ acts on $u$ to give $D_i u$ which we can evaluate in $x$, $x-y$, $y$ and so on.

with $F$ replaced by $U$ and $V_i = D_i U$. Note carefully that the mollification of the equivalence class $W_i = D_i U$ rewrites as

$$(D_i U)_\varepsilon(x) = (\eta_\varepsilon * D_i U)(x) = (D_i \eta_\varepsilon * U)(x) = (D_i U_\varepsilon)(x),$$

by first Definition 32.2 which puts $D_i$ on $\eta_\varepsilon$, and then the same reasoning as in the proof of Theorem 31.37. As a consequence $U_\varepsilon$ and $D_i U_\varepsilon = (D_i U)_\varepsilon$ are smooth $p$-integrable functions in $C_0(\mathrm{I\!R}^{\mathrm{N}})$ that converge in $p$-norm to $U$ and $D_i U$, i.e. $U_\varepsilon$ converges to $U$ in $W^{1,p}$-norm. □

**Theorem 32.54.**

$$W_0^{1,p}(\mathrm{I\!R}^{\mathrm{N}}) = W^{1,p}(\mathrm{I\!R}^{\mathrm{N}}).$$

**Exercise 32.55.** Prove this theorem. Hint: if $u$ is $W^{1,p}(\mathrm{I\!R}^{\mathrm{N}})$ then by Theorem 32.10 so is the function $u_n$ defined by $u_n(x) = u(x)\eta(\frac{x}{n})$, $\eta$ as in the definition of $\eta_\varepsilon$. Apply Theorem 32.53 to $u_n$.

# 33 Sobolev spaces without subscript zero

Recall that if $u, v \in L^p(\Omega)$ satisfy

$$\int_\Omega v\phi = -\int_\Omega u\phi_{x_i}$$

for every $\phi \in C_c^1(\Omega)$, then $v$ is the (unique) weak derivative of $u$ with respect to its $i^{th}$ variable, notation $v = D_i u = u_{x_i}$. If so, then for every continuously differentiable $\zeta : \mathbb{R}^N \to \mathbb{R}^N$ the product $\zeta u$ is also in $L^p(\Omega)$ and has a (unique) weak derivative with respect to its $i^{th}$ variable, given by the (weak version of the) Leibniz rule[1]

$$D_i\zeta u = uD_i\zeta + \zeta D_i u. \tag{33.1}$$

As in Remark 33.3 we define

$$W^{1,p}(\Omega) = \{u \in L^p(\Omega) : \, D_1 u, \dots, D_N u \text{ exist in } L^p(\Omega\},$$

and note that the implication

$$\begin{matrix} u \in W^{1,p}(\Omega) \\ \zeta \in C_c^1(\mathbb{R}^N) \end{matrix} \implies \zeta u \in W^{1,p}(\Omega)$$

is at our disposal, with $\zeta u \in W^{1,p}(\mathbb{R}^N)$ if $\operatorname{supp}\zeta \subset \Omega$.

From here on we no longer resist the temptation of Remark 31.39. It is up to you whether you think of $u$ as an equivalence class of Lebesgue measurable functions[2], or as an equivalence class of $p$-Cauchy sequences[3] $u_n$ in $C_c(\Omega)$. Both approaches allow us to use that every $u \in L^p(\Omega) \subset L^p(\mathbb{R}^N)$ can be approximated in $p$-norm with functions $v \in C_c(\Omega) \subset C_c(\mathbb{R}^N)$, if we like taken from a suitably chosen $p$-Cauchy sequence $u_n$ in $C_c(\Omega) \subset C_c(\mathbb{R}^N)$. Writing $u^\varepsilon$ for the mollification of $u$ we then know that the limit

$$u^\varepsilon = \lim_{n\to\infty} \underbrace{(\eta_\varepsilon * u_n)}_{u_n^\varepsilon} \tag{33.2}$$

exists as a smooth function for every $\varepsilon > 0$ fixed as in (31.36). The convergence is in $p$-norm and in maximum norm, also for all derivatives. Thus $D^\alpha u^\varepsilon$ is $p$-integrable and in $C_0(\mathbb{R}^N)$ for every multi-index $\alpha$. If $\Omega$ is bounded we can conclude that $u^\varepsilon \in C_c^\infty(\mathbb{R}^N)$, with its support contained in an $\varepsilon$-neighbourhood of $\Omega$.

[1]See Theorem 32.10.
[2]See Definition 31.6.
[3]See Section 31.6.

We know that

$$|u^\varepsilon|_p \le |u|_p \quad \text{and} \quad |u^\varepsilon - u|_p \to 0$$

as $\varepsilon \to 0$, thanks to[4] density and the estimate

$$|v^\varepsilon - v|_p \le \varepsilon\, |\nabla v|_p$$

for $v \in C^1_c(\mathrm{I\!R}^{\mathrm{N}})$. The latter estimate was not yet used for $u \in W^{1,p}(\mathrm{I\!R}^{\mathrm{N}})$, but the previous chapter did end with the statement that

$$u^\varepsilon \to u \quad \text{in} \quad W^{1,p}(\mathrm{I\!R}^{\mathrm{N}}) \quad \text{as} \quad \varepsilon \to 0 \tag{33.3}$$

for every $u \in W^{1,p}(\mathrm{I\!R}^{\mathrm{N}})$, thanks to[5]

$$D_i(u^\varepsilon) = (D_i u)^\varepsilon.$$

Thereby

$$W^{1,p}_0(\mathrm{I\!R}^{\mathrm{N}}) = W^{1,p}(\mathrm{I\!R}^{\mathrm{N}}), \tag{33.4}$$

and thus

$$|u^\varepsilon - u|_p \le \varepsilon\, |\nabla u|_p$$

also holds for all $u \in W^{1,p}(\mathrm{I\!R}^{\mathrm{N}})$.

However it is only in the special case that $\Omega = \mathrm{I\!R}^{\mathrm{N}}$ that $W^{1,p}$ is the closure of $C^1_c$ in the $W^{1,p}$-norm[6]. Whereas statements about $W^{1,p}(\mathrm{I\!R}^{\mathrm{N}})$ can be proved via functions in $C^1_c(\mathrm{I\!R}^{\mathrm{N}})$ and the methods of calculus, things are much more complicated for $W^{1,p}(\Omega)$ with $\Omega$ strictly contained in $\mathrm{I\!R}^{\mathrm{N}}$. Although given $u \in W^{1,p}(\Omega)$ we can extend $u$ and its derivatives $w_1 = D_1 u, \ldots, w_N = D_N u$ to $\mathrm{I\!R}^{\mathrm{N}}$ by setting[7] $u(x) = w_1(x) = \cdots = w_N(x) = 0$ for $x \notin \Omega$, we certainly cannot expect to have $w_i = D_i u$ across $\partial\Omega$. We still have that

$$u^\varepsilon \to u \quad \text{and} \quad w_i^\varepsilon \to w_i \quad \text{in} \quad L^p(\mathrm{I\!R}^{\mathrm{N}}) \quad \text{as} \quad \varepsilon \to 0,$$

but convergence in $W^{1,p}(\Omega)$ fails because we can only have that

$$w_i^\varepsilon = D_i u^\varepsilon \quad \text{in} \quad L^p(\Omega_\varepsilon), \quad \Omega_\varepsilon = \{x \in \Omega : B(x,\varepsilon) \subset \Omega\}. \tag{33.5}$$

The following observation will help to find an alternative.

[4]See (32.16) in Section 32.3, where we used $v$; $u_n$ was used in Exercise 32.17.

[5]Using superscripts only, see Remark 31.39 and Exercise 31.34.

[6]In other words, $C^1_c(\mathrm{I\!R}^{\mathrm{N}})$ is dense in $W^{1,p}(\mathrm{I\!R}^{\mathrm{N}})$.

[7]Reasoning in the Lebesgue setting for convenience.

**Remark 33.1.** *Let $\Omega \subset \mathrm{I\!R}^{\mathrm{N}}$ be open and bounded, with boundary $\partial\Omega \in C^1$. Any statement that we want make here about functions $u$ in $W^{1,p}(\Omega)$ can be localised using partitions of unity, splitting $u$ via $\zeta_0, \zeta_1, \ldots, \zeta_n$ in $C_c^1(\mathrm{I\!R}^{\mathrm{N}}) \subset C_c^\infty(\mathrm{I\!R}^{\mathrm{N}})$, with $\zeta_0 + \zeta_1 + \cdots + \zeta_n \equiv 1$ on $\bar{\Omega}$, as*

$$u = u_0 + u_1 + \cdots + u_n = \zeta_0 u + \zeta_1 u + \cdots + \zeta_n u,$$

*in which we take $\zeta_0 \in C_c^1(\Omega)$ and $\zeta_1, \ldots, \zeta_n \in C_c^1(\mathrm{I\!R}^{\mathrm{N}})$. This is just like in Sections 27.4 and 27.5 for the proof of Green's Theorem 27.7. The weak Leibniz rule (33.1) applies*[8] *to each $\zeta_i u$.*

## 33.1 Density via shifts, localisation and mollification

In this section we prove that $C^1(\bar{\Omega})$ is dense in $W^{1,p}(\Omega)$ if $\Omega$ bounded and $\partial\Omega$ is not too bad[9] , e.g. if $\partial\Omega \in C^1$. The proof involves a translation trick that we record here separately. It will be applied to the individual terms $\zeta_i u$ in Remark 33.1 to establish the density of $C^1(\bar{\Omega})$ in $W^{1,p}(\Omega)$. This corresponds to Theorem 3 in Evans' Section 5.3.3, which we improve by making the formulation functional with the construction of a continuous linear $\varepsilon$-dependent map $u \to u_\varepsilon$ from $W^{1,p}(\Omega)$ to $C_c^\infty(\mathrm{I\!R}^{\mathrm{N}})$ with the desired properties.

**Theorem 33.2.** *Let $e$ be a unit vector $e$ and $h > 0$ as in (33.6). For $u$ in $L^p(\mathrm{I\!R}^{\mathrm{N}})$ define $u_{he} \in L^p(\mathrm{I\!R}^{\mathrm{N}})$ by*

$$u_{he}(x) = u(x + he). \tag{33.6}$$

*Then the mollified translates satisfy*

$$u_{he}^\varepsilon \to u \quad \text{in} \quad L^p(\mathrm{I\!R}^{\mathrm{N}}) \quad \textit{as} \quad h \to 0 \quad \textit{and} \quad \varepsilon \to 0.$$

**Proof**. Since clearly

$$(u_{he})^\varepsilon = (u^\varepsilon)_{he}$$

we can write $u_{he}^\varepsilon$ for the mollified and shifted $u$. It follows that

$$|u_{he}^\varepsilon - u_{he}|_p = |u^\varepsilon - u|_p \to 0 \quad \text{in} \quad L^p(\mathrm{I\!R}^{\mathrm{N}}) \quad \text{as} \quad \varepsilon \to 0$$

by Theorem 32.15. But then

$$|u_{he}^\varepsilon - u|_p \le \underbrace{|u_{he}^\varepsilon - u_{he}|_p}_{=|u^\varepsilon - u|_p} + |u_{he} - u|_p = \underbrace{|u^\varepsilon - u|_p}_{\to 0} + \underbrace{|u_{he} - u|_p}_{\to 0}, \tag{33.7}$$

[8] Evans' stronger assumption $\zeta \in C_c^\infty(\Omega)$ for Leibniz' rule leads to cumbersome details.
[9] See Chapter 27.3 for $C^1$-boundaries. To do later perhaps: non smooth boundaries!

the latter because

$$|u_{he} - u|_p \le \underbrace{|u_{he} - v_{he}|_p}_{=|v-u|_p} + |v_{he} - v|_p + |v - u|_p \tag{33.8}$$

for every $v \in C_c(\mathrm{I\!R}^{\mathrm{N}})$, similar to (32.16). This completes the proof. $\square$

**Exercise 33.3.** Use (33.8) with $v \in C_c(\mathrm{I\!R}^{\mathrm{N}})$ and $|v - u|_p$ as small as desired to prove that $u_{he} \to u$ in $L^p(\mathrm{I\!R}^{\mathrm{N}})$. Hint: use the uniform continuity of each such $v$.

**Exercise 33.4.** Verify that the proof works for both definitions[10] of $L^p(\Omega)$.

**Theorem 33.5.** *Let $\Omega$ be open bounded with $\partial\Omega \in C^1$. Then there exists a partition of unity as in Remark 33.1, unit vectors $e_1, \ldots, e_n$, and a number $\lambda > 0$ such that for every $u$ in $W^{1,p}(\Omega)$ the function*

$$u_\varepsilon = (\zeta_0 u)^\varepsilon + \sum_{i=1}^{n} (\zeta_i u)^\varepsilon_{\lambda\varepsilon e_i},$$

*is in $C_c^\infty(\mathrm{I\!R}^{\mathrm{N}})$, and converges to $u$ in $W^{1,p}(\Omega)$.*

**Remark 33.6.** *In fact the result is valid under the assumption that $\partial\Omega$ is compact and uniformly Lipschitz continuous*[11]*. This is not so important for our purposes here as we will need $\partial\Omega$ to be $C^1$ for other reasons later, but we do note the number $\lambda$ is $\sqrt{1+L^2}$, in which $L$ is the largest of the $n$ Lipschitz constants that occur in the proof. Moreover, if $u \in W^{k,p}(\Omega)$ with $k \in \mathrm{I\!N}$ and $1 \le p < \infty$ then $u_\varepsilon \to u$ in $W^{k,p}(\Omega)$.*

**Proof of Theorem 33.5**. We localise $u$ by multiplying it with a function $\zeta \in C_c^1(\mathrm{I\!R}^{\mathrm{N}})$ and then consider shifts $\tilde{u}_{he}$ of $\tilde{u} = \zeta u$ defined by the choice of a fixed unit vector $e$ to be chosen in relation to the local description of $\Omega$ and its boundary in and near the support of $\zeta$. Note that

$$\tilde{u} \in W^{1,p}(\tilde{\Omega}), \quad \tilde{\Omega} = \Omega \cup (\operatorname{supp} \zeta)^c, \quad \tilde{\Omega}^c = \Omega^c \cap \operatorname{supp} \zeta$$

and that

$$\tilde{u}_{he} \in W^{1,p}(\tilde{\Omega}_{he})$$

[10] See Sections 31.2 and 31.6.

[11] Give a definition of what this should mean.

is defined by

$$\tilde{u}_{he}(x) = \tilde{u}(x+eh) \quad \text{for} \quad x \in \tilde{\Omega}_{he} = \{x \in \mathbb{R}^{\mathrm{N}} : x + he \in \tilde{\Omega}\}.$$

We extend $\tilde{u}$ and $\tilde{w}_i = D_i\tilde{u}$, defined in $L^p(\tilde{\Omega})$, to $L^p(\mathbb{R}^{\mathrm{N}})$ by setting

$$\tilde{u}(x) = \tilde{w}_i(x) = 0 \quad \text{for} \quad x \in \tilde{\Omega}^c = \Omega^c \cap \operatorname{supp}\zeta \quad \text{and} \quad i = 1, \dots, N$$

as before, and know from (33.2) that $u^\varepsilon, w_i^\varepsilon \in C_c^\infty(\mathbb{R}^{\mathrm{N}})$ have the property that

$$u_{he}^\varepsilon \to u \quad \text{and} \quad w_{ih}^\varepsilon \to w_i \quad \text{in} \quad L^p(\mathbb{R}^{\mathrm{N}}) \quad \text{as} \quad h \to 0 \quad \text{and} \quad \varepsilon \to 0.$$

To conclude that

$$\tilde{u}_{he}^\varepsilon \to \tilde{u} \quad \text{in} \quad W^{1,p}(\Omega)$$

it suffices to have

$$\tilde{w}_{ih}^\varepsilon = D_i\tilde{u}_{he}^\varepsilon \quad \text{in} \quad L^p(\Omega), \tag{33.9}$$

and in view of (33.5) this holds if

$$\Omega \subset \{x \in \tilde{\Omega}_{he} : B(x,\varepsilon) \subset \tilde{\Omega}_{he}\}. \tag{33.10}$$

The above reasoning will now be applied to a partition of unity as in Section 27.4 with each $\zeta_1, \dots, \zeta_n$ taking care of some part of the boundary of $\Omega$, and $\zeta_0 \in C_c^\infty(\Omega)$. Thus $\zeta_0 u$ is in $W^{1,p}(\mathbb{R}^{\mathrm{N}})$ and taken care of by Theorem 32.53:

**Exercise 33.7.** Use Theorem 32.53 to show that $(\zeta_0 u)^\varepsilon \to \zeta_0 u$ in $W^{1,p}(\Omega)$ as $\varepsilon \to 0$.

Without loss of generality we continue the proof reasoning in the 2-dimensional setting, with[12]

$$\operatorname{supp}\,\zeta = [\tilde{a}, \tilde{b}] = [\tilde{a}_1, \tilde{b}_1] \times [\tilde{a}_2, \tilde{b}_2],$$

$$a_1 < \tilde{a}_1 < \tilde{b}_1 < b_1, \quad a_2 < \tilde{a}_2 < \tilde{b}_2 < b_2,$$

$$\Omega \cap (a,b) = \{(x,y) : a_1 < x < b_1,\ f(x) < y < b_2\},$$

in which

$$f : [a_1, b_1] \to (\tilde{a}_2, \tilde{b}_2)$$

[12] Make a sketch.

is Lipschitz continuous with Lipschitz constant $L$. Dropping the (second unit) vector $e$ from the notation we write

$$\tilde{u}_h(x, y) = \tilde{u}(x, y + h)$$

and note that

$$\tilde{\Omega}_h \supset \{(x, y) : a_1 < x < b_1, \, y > f(x) - h\}.$$

Now let $\lambda = \sqrt{1 + L^2}$ and $h = \lambda\varepsilon$. Then every point in $[\tilde{a}, \tilde{b}] \cap \Omega$ 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 $(a, b) \cap \Omega$, provided $\varepsilon$ is smaller than the distance from $[\tilde{a}, \tilde{b}]$ to the boundary of $(a, b)$. This implies (33.10) holds with $h = \lambda\varepsilon$, whence

$$\tilde{u}^{\varepsilon}_{\lambda\varepsilon} \to \tilde{u} \quad \text{in} \quad W^{1,p}(\Omega) \tag{33.11}$$

as $\varepsilon \to 0$.

**Exercise 33.8.** To convince yourself of the statement preceding (33.11) 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 $\varepsilon$ to that line, and the point $Q_\varepsilon$ 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 $\varepsilon$.

The above argument applies to every $\zeta_1, \ldots, \zeta_n$. Combined with Exercise 33.7 this concludes the proof Theorem 33.5. $\square$

## 33.2 Statements for $W^{1,p}(\Omega)$ via the extension operator

To extend results for $W_0^{1,p}(\Omega)$ to $W^{1,p}(\Omega)$ we use a well behaved extension operator that maps $W^{1,p}(\Omega)$ into $W_0^{1,p}(\tilde{\Omega})$ with $\tilde{\Omega}$ slightly larger than $\Omega$. The extension operator is first defined for $u \in C^1(\bar{\Omega})$ and requires the boundary $\partial\Omega$ to be bounded and $C^1$ (locally the graph of a $C^1$-function). This uses the same partitions of unity used in the proof of Theorem 33.5 to establish that $W^{1,p}(\Omega)$ is the closure of $C^1(\bar{\Omega})$ if $\partial\Omega$ is bounded and $C^1$. The extension map

$$u \in W^{1,p}(\Omega) \xrightarrow{E} \tilde{u} \in W_0^{1,p}(\tilde{\Omega})$$

comes out linear and bounded. Results such as the compactess in Theorem 32.19 and the embeddings[13] in Section 32.4 then carry over to $W^{1,p}(\Omega)$.

Here I don't follow Evans' approach in which the extensions are first defined locally for $u$ and then glued together using a partition of unity. It is much simpler to first cut up $u$ in smaller pieces $\zeta_i u$ as explained in Remark 33.1 and use the globally defined extensions $\tilde{u}_i$ of $\zeta_i u$ rather than locally defined extensions of $u$. With local $C^1$-extensions $\tilde{u}_i$ defined by linear maps

$$u \xrightarrow{E_i} \tilde{u}_i \tag{33.12}$$

with

$$|\tilde{u}_i|_{1,p} \leq C_i\, |u|_{1,p}, \tag{33.13}$$

we simply define

$$u \in C^1_c(\bar{\Omega}) \xrightarrow{E} C^1_c(\tilde{\Omega}) \quad \text{by} \quad u \to \zeta_0 u + \tilde{u}_1 + \cdots + \tilde{u}_n.$$

In view of the Leibniz rule[14]

$$(\zeta u)_{x_j} = \zeta u_{x_j} + \zeta_{x_j} u$$

it follows that

$$|Eu|_{1,p} \leq C\, |u|_{1,p},$$

with $C$ some horrible constant depending on $\tilde{\Omega}$ and $\Omega$ via the norms of $\zeta_i$ in $C^1$. This cuts a long story short. The map

$$u \in C^1(\bar{\Omega}) \xrightarrow{E} \tilde{u} \in C^1_c(\tilde{\Omega})$$

satisfies the desired $W^{1,p}$-estimates, and thereby extends as a map from $W^{1,p}(\Omega)$ to $W^{1,p}_0(\tilde{\Omega})$ using the density of $C^1(\bar{\Omega})$ in $W^{1,p}(\Omega)$ established with Theorem 33.5.

Note that we take the supports of $\zeta_1, \ldots, \zeta_n$ as compact subsets of open cylinders rather than balls. After a permutation of coordinates each of these cylinders is described as $C_i = B_i \times I_i$, with $B_i$ an open ball, $I_i$ a bounded interval, and chosen such that

$$\Omega \cap C_i = \{x = (x_1, \ldots, x_{N-1}, x_N) \in C : \, x_N > \gamma_i(x_1, \ldots, x_{N-1})\}.$$

Here $\gamma_i : \bar{B}_i \to I_i$ is $C^1$, and the supports of the $\zeta_i$ are then contained in smaller cylinders $\tilde{C}_i = \tilde{B}_i \times \tilde{I}_i \subset\subset C_i$. The extensions $\tilde{u}_i$ of $\zeta_i u$ with compact support in $C_i$ are defined via transformations and higher order reflections as in Appendix C.1 and in Section 5.4 in of Evans.

[13]With different constants of course.

[14]I'm using subscripts $x_j$ for partial derivatives here, so $\zeta_{x_j} = D_j \zeta$.

## 33.3 The trace operator and its kernel

The other important operator is the bounded linear trace operator

$$W^{1,p}(\Omega) \xrightarrow{T} L^p(\partial\Omega)$$

in Section 5.4 of Evans, which extends

$$u \in C^1(\bar{\Omega}) \to u|_{\partial\Omega} \in C^1(\partial\Omega).$$

Evans defines it locally, first under the assumption that $\partial\Omega$ is flat and $u \in C^1(\bar{\Omega})$. The same splitting as in Theorem 33.5 can be used to first define $T(\zeta_i u)$ instead, for $u \in C^1(\bar{\Omega})$, so

$$u \in C^1(\bar{\Omega}) \to \zeta_i u = u_i \in C^1(\bar{\Omega} \cap \bar{C}_i) \to u_i|_{\partial\Omega} \in C^1(\partial\Omega).$$

The local coordinate transformation flattening $\partial\Omega \cap \bar{C}_i$ 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 $x_{\mathrm{N}}$ 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 $\nabla u_i$, 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 $u \in C_c^2(\mathbb{R}^2)$:

$$\int_{-\infty}^\infty |u(x,y)|^p \, dx \leq 2^p \left( \int_{-\infty}^\infty |u(x,0)|^p \, dx + y^{p-1} \int_0^y \int_{-\infty}^\infty |u_y|^p \right). \quad (33.14)$$

**Exercise 33.9.** Prove (33.14) and explain why it holds for $u \in W^{1,p}(\mathbb{R}_+^2)$ with compact support in $\mathbb{R} \times [0, \infty)$.

**Exercise 33.10.** 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 33.14 below.

**Exercise 33.11.** Let $\Omega$ 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 $\zeta u \in W^{1,p}(\Omega)$ if $u \in W^{1,p}(\Omega)$.

**Exercise 33.12.** Introduce new coordinates $\xi, \eta$ by

$$x = x_0 + a\xi + b\eta, \quad y = y_0 + c\xi + d\eta,$$

with $ad \neq bc$, define $V$ by $(\xi,\eta) \in V \iff (x,y) \in \Omega$, and write $v(\xi,\eta) = u(x,y)$. Show that

$$u \in W^{1,p}(\Omega) \iff v \in W^{1,p}(V)$$

and that this correspondence defines a linear homeomorphism between the two Sobolev spaces.

**Exercise 33.13.** Assume $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$ is $C^1$, injective on $\bar{\Omega}$, with invertible Jacobian matrix in every $(x,y) \in \bar{\Omega}$. Then $u \to u \circ \Phi^{-1} = v$ defines a bijective map from $C^1(\bar{\Omega}) \to C^1(\bar{V})$ where $V = \Phi(\Omega)$. Show that

$$\frac{1}{C}|v|_{W^{1,p}(V)} \leq |u|_{W^{1,p}(\Omega)} \leq C|v|_{W^{1,p}(V)}$$

for some $C > 1$.

**Exercise 33.14.** Explain why this map uniquely extends to a bijection from $W^{1,p}(\Omega)$ to $W^{1,p}(V)$ if $\partial\Omega \in C^1$.

**Exercise 33.15.** The intersection of $W^{1,p}(\Omega)$ and $C(\bar{\Omega})$ 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 $C^1(\bar{\Omega})$ with respect to this norm.

# 34 Elliptic boundary value problems

This chapter corresponds to Chapter 6 in Evans' PDE book, but we first consider the partial differential equation

$$Lu = -(a^{ij}u_{x_i})_{x_j} + cu = f \quad \text{with boundary condition} \quad u = 0 \tag{34.1}$$

for $u = u(x)$, $x \in \Omega$, $\Omega$ a bounded domain[1] in $\mathbb{R}^N$, $\partial\Omega$ 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.

Clearly the existence of classical solutions, i.e. solutions $u$ with $u \in C^2(\Omega)$ and $u \in C(\bar{\Omega})$, requires smoothness conditions on the coefficients $a^{ij} = a^{ij}(x)$ and $c = c(x)$, and on the right hand side $f$. The *uniform ellipticity condition* (4) on the coefficients $a^{ij} = a^{ij}(x)$ is that there exists $\theta > 0$ such that

$$a^{ij}(x)\xi_i\xi_j \geq \theta|\xi|^2$$

for all $x \in \Omega$ and for all $\xi \in \mathbb{R}^N$. It is used to verify the coercivity condition in the Lax-Milgram Theorem 30.23 for bilinear forms such as the form $B$ introduced in (34.3) below. We stress that in the special case of (34.1) we shall only need the Riesz Representation Theorem, see Section 30.3. The advantage of the boundary condition $u = 0$ is that we can work in the space $H_0^1(\Omega) = W_0^{1,2}(\Omega)$, which is defined as the closure of $C_c^1(\Omega)$ in $H^1(\Omega)$.

Evans starts with the more general equation

$$Lu = -(a^{ij}u_{x_i})_{x_j} + b^i u_{x_i} + cu = f,$$

for which the Lax-Milgram Theorem is required, see Section 30.4. This theorem reduces to the Riesz Representation Theorem in the case that the bilinear form in Theorem 30.23 is symmetric. In Sections 30.7, 30.8 we presented a simple example, starting from the standard Hilbert space in Section 30.6, of the general framework that we then recognise when solving (34.1) in the space $H_0^1(\Omega)$. In this framework the Spectral Theorem 30.10 applies to the solution operator

$$f \xrightarrow{S} u$$

with respect to two different inner products, as explained in Section 30.9. This corresponds to Theorems 1 and 2 in Evans' Section 6.5.

[1] Domains are denoted by $U$ in Evans.

## 34.1 Weak solutions

In the weak solution approach we multiply the partial differential equation in (34.1) by a $v \in C^1(\bar{\Omega})$, integrate over $\Omega$ and use integration by parts to rewrite the terms with $a^{ij}$ as

$$-\int_\Omega \underbrace{(a^{ij}u_{x_i})_{x_j}}_{w_{x_j}} v = -\int_{\partial\Omega} \underbrace{\nu_j a^{ij} u_{x_i}}_{\nu_j w} v + \int_\Omega \underbrace{a^{ij}u_{x_i}}_{w} v_{x_j}, \tag{34.2}$$

in which $\nu_j$ is the $j^{th}$ component of the outward normal on $\partial\Omega$. This requires $\partial\Omega$ to be piecewise $C^1$, the coefficients $a^{ij} \in C^2(\bar{\Omega})$, $c \in C(\bar{\Omega})$, the right hand side $f \in C(\bar{\Omega})$, and $u \in C^2(\bar{\Omega})$. The boundary integral disappears if $v = 0$ on $\partial\Omega$, leading to the identity[2]

$$\underbrace{\int_\Omega a^{ij}u_{x_i}v_{x_j} + \int_\Omega cuv}_{B[u,v]} = \underbrace{\int_\Omega fv}_{(f,v)} \tag{34.3}$$

for all $v \in C^1(\bar{\Omega})$ with $v = 0$ on $\partial\Omega$. The assumptions on $a^{ij}, c, f$ can be weakened, since this identity certainly make sense for $u \in C^1(\bar{\Omega})$, and so does the boundary condition $u = 0$. In fact, the standard weak solution approach requires solutions which have their first order derivatives in $L^2(\Omega)$. Thus the natural (Hilbert) space for $u, v$ to live is $H_0^1(\Omega)$, and $u \in H_0^1(\Omega)$ is called a weak solution of (34.1) if (34.3) holds for all $v \in H_0^1(\Omega)$.

We will also say that $u \in H^1(\Omega)$ is a weak solution of[3]

$$Lu = -(a^{ij}u_{x_i})_{x_j} + cu = f \quad \text{with boundary condition} \quad \nu_j a^{ij} u_{x_i} = 0$$

if (34.3) holds for all $v \in H^1(\Omega)$, but for now we restrict the attention to the reformulation of (34.1), with $u$ and $v$ both in $H_0^1(\Omega)$. Note that (34.3) requires $a^{ij}, c \in L^\infty(\Omega)$ only, and $f \in L^2(\Omega)$ more than suffices for the right hand side of (34.3) to makes sense because $H_0^1(\Omega)$ is contained in $L^2(\Omega)$.

Recall that $H_0^1(\Omega)$ is the closure of $C_c^1(\Omega)$ in $H^1(\Omega) = W^{1,2}(\Omega)$. The right hand side of (34.3) is equal to the inner product of $f$ and $v$ in $L^2(\Omega)$ and it defines a linear functional $F$ via

$$v \in H_0^1(\Omega) \xrightarrow{F} \int_\Omega fv = (f,v)_{L^2(\Omega)} = \underbrace{F(v) = \langle F, v\rangle}_{\text{two notations for } F}, \tag{34.4}$$

the latter notation being the one used in the Lax-Milgram Theorem in Section 6.2.1. In (34.4) we chose not to write $f$ for $F$ in $\langle F, v\rangle$, as $f$ acts on $v$ via the

[2]The underbraces indicate the notation in (8) of Evans' Section 6.1.
[3]Section 5.8.1 is important for this problem.

$L^2$-inner product, and *not* via the inner product in $H^1(\Omega)$, which is defined by the left hand side of (34.3) with $a^{ij} = \delta^{ij}$ and $c = 1$, i.e.

$$(u,v)_{H^1(\Omega)} = \underbrace{\int_\Omega u_{x_i} v_{x_i}}_{\text{first order terms}} + \int_\Omega uv = \int_\Omega \nabla u \cdot \nabla v + \int_\Omega uv. \tag{34.5}$$

Note that the $H^1(\Omega)$ inner product is the bilinear form corresponding to the partial differential equation $\Delta u + u = f$. This is the easiest example of (34.3). The next easiest example is when $c(x) \geq 0$ for all $x \in \Omega$, and the left hand side of (34.3) still defines an inner product, also if $c(x) = 0$ for all $x \in \Omega$.

## 34.2 The Lax-Milgram Theorem

It is only for equations of the form

$$Lu = -(a^{ij} u_{x_i})_{x_j} + b^i u_{x_i} + cu = f$$

that we need the Lax-Milgram Theorem of Section 30.4. Its statement and proof will be recalled below. The symmetric case was also discussed in Section 30.8. The $f$ in Theorem 1 in Evans' Section 6.2.1 is really the $F$ in (34.4) if the theorem is applied to the bilinear form in (34.3) and the Hilbert space $H_0^1(\Omega)$. This $F$ was defined via the inner product of the larger Hilbert space $L^2(\Omega)$. Have a look at Section 30.8 for a first example of this double dealing with the Riesz Theorem. Our Sobolev space $H_0^1(\Omega)$ here corresponds to $V$ there, and $L^2(\Omega)$ to $H$. Note that Evans considers a more general right hand side in (10), which is related to a characterisation of the dual of $H_0^1(\Omega)$ in his Section 5.9.1.

The discussion of the Lax-Milgram Theorem below uses the notation in Evans[4]. If a bilinear form $B : H \times H \to \mathbb{R}$ is bounded, i.e. if

$$\forall u, v \in H \quad |B[u,v]| \leq \alpha |u|\,|v|,$$

then for each $u \in H$ the map

$$v \in H \xrightarrow{Au} B[Au, v] = \underbrace{(Au)(v) = \langle Au, v\rangle}_{\text{two notations for Au}} \tag{34.6}$$

is linear and bounded, since

$$|\langle Au, v\rangle| = |B[u,v]| \leq \alpha\, |u|\, |v|$$

[4] And this $H$ corresponds to $X$ in Theorem 30.26 of Section 30.5.

implies that

$$|Au| \le \alpha\,|u|$$

for all $u \in H$. Thus $A : H \to H^*$ is linear and bounded. Recall that the dual space

$$H^* = \{f : H \to \mathrm{I\!R} : \ f \text{ is linear and bounded}\}$$

is normed by

$$|f| = \sup_{0 \neq v \in H} \frac{|\langle f, v\rangle|}{|v|},$$

and 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 $H = H_0^1(\Omega)$ this is not the inner product in the right hand side to (34.3).

If the bilinear form is also coercive on $H$, i.e. if

$$\forall u \in H \quad B[u,u] \ge \beta\,|u|^2,$$

then

$$\beta\,|u|^2 \le B[u,u] = \langle Au, u\rangle \le |Au|\,|u| \quad \text{whence} \quad |Au| \ge \beta\,|u|$$

for all $u \in H$ and it follows that $A$ is a bijection between $H$ and $A(H)$, a subspace of $H^*$, and that this bijection is bounded in both directions. Thus $A(H)$ is complete, and thereby a closed subspace of $H^*$ which[5] coincides with $H^*$.

The (linear) solution operator[6]

$$F \in H^* \xrightarrow{S} u \in H \quad \text{ is defined by } \quad B[u,v] = \langle F, v\rangle \quad \text{for all} \quad v \in H, \tag{34.7}$$

and has the property that

$$|u|_H = |SF|_H \le \frac{1}{\beta}\,|F|_{H^*}$$

In the application to boundary value problems any right hand side of (34.1) that defines an $F$ in the dual of the Sobolev space used is allowed. In the case of $H_0^1(\Omega)$ this dual space is denoted by $H^{-1}(\Omega)$ 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.

[5] Via the Riesz Representation or the Hahn-Banach Theorem and the reflexivity of $H$.
[6] See also Section 30.9.

## 34.3 Checking the 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

$$\nu_j a^{ij} u_{x_i} = 0 \quad \text{on} \quad \partial\Omega, \tag{34.8}$$

which is a special case of the Robin boundary condition

$$\nu_j a^{ij} u_{x_i} + bu = 0 \quad \text{on} \quad \partial\Omega, \tag{34.9}$$

in which $b = b(x)$ is assumed to be bounded and integrable. 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 $\Omega$ with heat exchange at the boundary[7]. In (34.2) this condition gives the additional term

$$\int_{\partial\Omega} buv$$

which should now be included in the left hand side of (34.3). The natural Sobolev space to pose

$$\int_\Omega a^{ij} u_{x_i} v_{x_i} + \int_\Omega cuv + \int_{\partial\Omega} buv = \int_\Omega fv \tag{34.10}$$

in is $H^1(\Omega)$.

In the case of the Neumann condition (34.8) this extra term is not there and the only difference with the Dirichlet problem is the choice of the Sobolev 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 a natural boundary condition, it has to be forced on the solution by the choice of the smaller Sobolev space $H^1_0(\Omega)$.

## 34.4 Checking 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 (34.5). In the case of $H^1_0(\Omega)$ the Poincaré inequality

$$\int_\Omega u^2 \le C_\Omega \int_\Omega |\nabla u|^2 \tag{34.11}$$

[7]This physical context forces the exchange coefficient to be positive.

helps. In particular the bilinear form

$$B[u,v] = \int_\Omega \nabla u \cdot \nabla v$$

used for solving

$$-\Delta u = f \quad \text{with boundary condition} \quad u = 0$$

is coercive on $H^1_0(\Omega)$ considered as a subspace of $H^1(\Omega)$ with the norm derived from (34.5).

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

$$\tilde{H}^1(\Omega) = \{u \in H^1(\Omega) : \int_\Omega u = 0\}$$

as the Sobolev space to be used in the weak formulation. For functions in this space it holds that

$$|u|_2 \leq C_\Omega |\nabla u|_2,$$

see Evans' Section 5.8.1.

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 (34.10). Coercivity requires some positivity.

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 to have $u \in H^2_0(\Omega)$, the closure of $C^2_c(\Omega)$ in the $W^{2,2}$-norm. Can you think of natural boundary conditions that lead to a formulation in $H^2(\Omega)$, or a mix of natural and unnatural boundary conditions that require $H^2(\Omega) \cap H^1_0(\Omega)$ as the space to be used? Note that for the coercivity of the bilinear form you need the regularity theory in Section 6.3.

## 34.5 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 $Lu$. Section 6.2.3 uses the adjoint operator and the Fredholm alternative towards Theorem 4. The Fredholm alternative is applied to the solution operator $S_\mu$ for the bilinear form

$$B_\mu[u,v] = B[u,v] + \mu \int_\Omega uv$$

with one choice $\gamma$ of $\mu$, chosen sufficiently large to make $B_\gamma$ coercive, and thereby $B_\mu$ for all $\mu \geq \gamma$. This latter trick is independent of the Fredholm alternative approach that follows after Theorem 3 in Evans' Section 6.2, and requires the energy estimates in Section 6.2.2. Young's inequality is repeatedly used here, and will also be used in Section 35 below.

## 34.6 The symmetric case

Evans Section 6.5. See again Section 30.6. 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

$$(Sf, g)_{L^2(\Omega)} = (f, Sg)_{L^2(\Omega)} \quad \text{and} \quad B(Su, v) = B(u, Sv)$$

as you should verify, and it is compact from $L^2(\Omega)$ to $L^2(\Omega)$ as well as from the Sobolev space to itself. The eigenvalue formula's 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.

## 34.7 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`

# 35 Regularity

This chapter corresponds to Evans' Section 6.3. Read his motivation with (3) for (2). The upshot is that we want to use the pure second order derivatives of a weak solution $u$ as $v$ in the weak formulation of

$$Lu = -(a^{ij}u_{x_i})_{x_j} + b^i u_{x_i} + cu = f, \tag{35.1}$$

e.g. with boundary condition $u = 0$. This is not allowed, but we can get around this obstruction if we take finite second order differences instead, as Evans does in (11) of his Section 6.3. The circumvention is explained in Section 35.6, which you can actually read now if you like[1], but I explain first how Evans' approach can be turned around in such a way that his (11) is only used *without* the $\zeta^2$ factor between $D_k^{-h}$ and $D_k^h$.

The first goal is to prove that a weak solution $u \in H_0^1(\Omega)$ is in $H^2(\Omega)$ under minimal assumptions on $L$ and $\partial\Omega$, provided $f$ is in $L^2(\Omega)$, as a first step towards higher regularity and smoothness of solutions. In the proof we need a correspondence between bounds on integrals of squared first order difference quotients uniformly in the step size $h$ and bounds on the 2-norm of weak derivatives. This will require the use of weak limits of

$$D_k^h u(x) = \frac{1}{h}(u(x + he_k) - u(x))$$

along sequences $h_n \to 0$, given a uniform $L^2$-bound on $D_k^h u$. Note that Evans' exposition in Section 5.8.2 relies on his Appendix D4, but we only need the Hilbert space case[2] of the weak compactness statement, which is much easier to prove using Remark 30.29.

## 35.1 An a priori energy estimate

Recall that $u \in H^1(\Omega)$ is called a weak solution of $Lu = f$ if

$$\underbrace{\int_\Omega a^{ij}u_{x_i}v_{x_j} + \int_\Omega b^i u_{x_i} v + \int_\Omega cuv}_{B[u,v]} = \underbrace{\int_\Omega fv}_{(f,v)} \tag{35.2}$$

holds for all $v \in H_0^1(\Omega)$. We assume that $a^{ij} = a^{ji}$, $b^i$ and $c$ are continuous on $\bar{\Omega}$, and that the uniform ellipticity condition

$$a^{ij}\xi_i\xi_j \geq \theta|\xi|^2 \tag{35.3}$$

[1]Have a look Theorem 35.3 before your read on.

[2]Since this is about 2-norms.

holds. In case $u \in H_0^1(\Omega)$ we can take $v = u$ and thus have

$$\theta \int_\Omega |\nabla u|^2 \leq \int_\Omega a^{ij} u_{x_i} u_{x_j} = \int_\Omega (f - cu - b^i u_{x_i}) u. \tag{35.4}$$

It then follows that the so-called *energy estimate*[3]

$$\int_\Omega |\nabla u|^2 \leq \int_\Omega f^2 + C \int_\Omega u^2 \tag{35.5}$$

holds, for some constant $C$ that can be specified upon demand in terms of $\theta$ and the bounds on $b^i$ and $c$. We emphasize that (35.5) is an explicit *a priori bound: it holds for all* solutions $u \in H_0^1(\Omega)$.

**Exercise 35.1.** Use Young's inequality for

$$ab = \varepsilon a \, \frac{b}{\varepsilon}$$

with $p = q = 2$ and a suitable choice of $\varepsilon > 0$ to prove (35.5).

We next want to establish that the second order (weak) derivatives of a weak solution $u \in H_0^1(\Omega)$ exist in $L^2(\Omega)$, with *a priori bounds similar* to (35.5). To do so we assume the first order derivatives $a^{ij}_{x_k}$ exist and are continuous on $\bar{\Omega}$, and that $\partial\Omega \in C^2$.

## 35.2 Localise it

We write

$$u = \zeta_0 u + \cdots + \zeta_n u = \sum_{m=0}^{n} u_m, \tag{35.6}$$

with $\zeta_0, \ldots, \zeta_n$ smooth, compactly supported, nonnegative, the support of $\zeta_0$ contained in $\Omega$, just as in Remark 33.1. Evans' method can be modified by observing that each of the terms $\hat{u} = u_m = \zeta_m u$ in (35.6) is itself a weak solution of an equation $\hat{L}u = \hat{f}$ with the same coefficients $a^{ij}$. This is just like (9) and (10) in Evans' Section 6.3, and follows replacing $v$ in (35.2) by $\zeta_m v$.

Indeed, dropping the subscript we have that

$$\int_\Omega f\zeta v = B[u, \zeta v] = \int_\Omega a^{ij} u_{x_i} (\zeta v)_{x_j} + \int_\Omega b^i u_{x_i} \zeta v + \int_\Omega cu\zeta v$$

[3]See Evans' Section 6.2.2.

$$= \underbrace{\int_\Omega a^{ij} u_{x_i} \zeta v_{x_j} + \int_\Omega a^{ij} u_{x_i} \zeta_{x_j} v}_{\text{Leibniz rule for } (\zeta v)_{x_j}} + \int_\Omega b^i u_{x_i} \zeta v + \int_\Omega c u \zeta v.$$

Taking the three $v$-terms to the other side we use $\zeta u_{x_i} = (\zeta u)_{x_i} - u\zeta_{x_i}$ to get

$$\int_\Omega (f\zeta - c\zeta u - b^i u_{x_i}\zeta - a^{ij} u_{x_i} \zeta_{x_j}) v = \int_\Omega a^{ij} u_{x_i} \zeta v_{x_j}$$
$$= \int_\Omega a^{ij} (\zeta u)_{x_i} v_{x_j} \underbrace{- \int_\Omega a^{ij} u \zeta_{x_i} v_{x_j}}_{\text{Leibniz again}}.$$

By the definition of weak derivative the last term equals

$$\int_\Omega (a^{ij} u \zeta_{x_i})_{x_j} v = \int_\Omega a^{ij}_{x_j} u \zeta_{x_i} v + \int_\Omega a^{ij} u_{x_j} \zeta_{x_i} v + \int_\Omega a^{ij} u \zeta_{x_i x_j} v,$$

where we have use the Leibniz rule[4] for the product $a^{ij} u \zeta_{x_i}$. Taking also the two new terms with $v$ to the other side we arrive at

$$\int_\Omega a^{ij} \hat{u}_{x_i} v_{x_j} = \int_\Omega a^{ij} (\zeta u)_{x_i} v_{x_j} = \int_\Omega \hat{f} v \tag{35.7}$$

with

$$\hat{f} = \zeta(f - cu - b^i u_{x_i}) - \zeta_{x_i}(2a^{ij} u_{x_j} + a^{ij}_{x_j} u) - \zeta_{x_i x_j} a^{ij} u. \tag{35.8}$$

In other words $\hat{u} = \zeta u$ satisfies

$$-(a^{ij} \hat{u}_{x_i})_{x_j} = \hat{f} \tag{35.9}$$

in the usual weak sense, and clearly[5]

$$|\hat{f}|_{_2} \le C |u|_{_{1,2}} \tag{35.10}$$

with the constant $C$ depending only on uniform bounds for

$$a^{ij}, a^{ij}_{x_k}, b^i, c, \zeta, \zeta_{x_i}, \zeta_{x_i x_j}.$$

In the case that $\zeta \in C^2_c(\Omega)$ we see that both $u$ and $f$ have compact support in $\Omega$. The function $\hat{u}$ is then in fact a weak solution of (35.9) in $H^1(\mathrm{I\!R}^{\mathrm{N}}) = H^1_0(\mathrm{I\!R}^{\mathrm{N}})$. Compare this to Evans' Section 6.7 Exercise 4[6], which then applies. The upshot is that there is really no need for the use of a cut-off function in (11) of Evans' Section 6.3.1. We test the equation for $\hat{u} = \zeta u$ with second order difference quotients

$$D_k^{-h} D_k^h \hat{u}$$

only, rather then testing the equation for $u$ with (11).

---

[4]Note carefully that $a^{ij}\zeta_{x_i}\phi$ is an allowable test function if $\phi \in C^1_c(\Omega)$.

[5]Recall (32.48).

[6]Exercise 7 in the second edition.

## 35.3 Flatten it

The trick with $\hat{u} = \zeta u$ and $\hat{f}$ in (35.10) also works if $\zeta$ is one of the functions as in Remark 33.1. Or as in the proof of Theorem 27.7 for that matter, compactly supported in a window $W$ in which the boundary of $\Omega$ is given by the graph of a $C^1$-function $\gamma$. From that proof we now borrow the flattening trick that Evans explains[7] in Appendix C1. We examine the action of these flattening transformations[8] on test functions $v$ with support in $W$, and on the solution $\hat{u} = \zeta u$. Define new coordinates

$$x_i = \tilde{x}_i \quad \text{for} \quad i = 1, \dots, N-1 \quad \text{and} \quad \tilde{x}_N = x_N - \gamma(\underbrace{x_1, \dots, x_{N-1}}_{\substack{\uparrow \\ x' = (x_1, \dots, x_{N-1}) = \tilde{x}'}}),$$

let $\tilde{u}, \tilde{v}$ be defined by

$$\hat{u}(x) = \tilde{u}(\tilde{x}), \quad v(x) = \tilde{v}(\tilde{x}), \tag{35.11}$$

and assume that in the support of $\zeta$ we have that $x \in \Omega$ if and only if $\tilde{x}_N > 0$. Then for $v \in C_c^1(W)$ the chain rule implies that

$$D_N v = D_N \tilde{v} \quad \text{and} \quad D_i v = D_i \tilde{v} - D_N \tilde{v} D_i \gamma$$

for $i = 1, \dots, N-1$. Since

$$D_i \gamma = \gamma_{x_i} = \gamma_{\tilde{x}_i},$$

we may write

$$v_{x_N} = \tilde{v}_{\tilde{x}_N} \quad \text{and} \quad v_{x_i} = \tilde{v}_{\tilde{x}_i} - \gamma_{\tilde{x}_i} \tilde{v}_{\tilde{x}_N},$$

or simply

$$v_{x_i} = \tilde{v}_{\tilde{x}_i} - \gamma_{\tilde{x}_i} \tilde{v}_{\tilde{x}_N} \tag{35.12}$$

for all $i = 1, \dots, N$, since $D_N \gamma = 0$. By density this also holds for $v$ in $H_0^1(\Omega)$ with support in $W$, and likewise

$$\hat{u}_{x_i} = \tilde{u}_{\tilde{x}_i} - \gamma_{\tilde{x}_i} \tilde{u}_{\tilde{x}_N}. \tag{35.13}$$

In fact a test function argument[9] shows that we also have (35.13) starting with $u \in H^1(\Omega)$.

As a consequence the bilinear form[10] in (35.7) is transformed via

$$a^{ij} \hat{u}_{x_i} v_{x_j} = \tilde{a}^{ij} (\tilde{u}_{\tilde{x}_i} - \gamma_{\tilde{x}_i} \tilde{u}_{\tilde{x}_N})(\tilde{v}_{\tilde{x}_j} - \gamma_{\tilde{x}_j} \tilde{v}_{\tilde{x}_N}),$$

[7] Take balls rather than cylinders as neighbourhoods.

[8] Postponing the difference quotients $D_k^h$ once more.

[9] In which $dx = d\tilde{x}$, we will need this later, requires $\gamma \in C^1$ only.

[10] If you like with also $v$ localised.

and so are the terms in (35.8) via $\hat{f}(x) = \tilde{f}(x)$.

The bounded first order derivatives of $\gamma$ do not qualitatively change the bound in (35.10), only the constant changes. The coercivity estimate (35.3) changes likewise. The quadratic form in (35.3) transforms as

$$Q(\xi) = a^{ij}\xi_i\xi_j = \tilde{a}^{ij}(\underbrace{\tilde{\xi}_i - \varepsilon_i\tilde{\xi}_N}_{\xi_i})(\underbrace{\tilde{\xi}_j - \varepsilon_j\tilde{\xi}_N}_{\xi_j}) = \hat{a}^{ij}\tilde{\xi}_i\tilde{\xi}_j = \tilde{Q}(\tilde{\xi}), \tag{35.14}$$

in which $a^{ij} = a^{ij}(x) = \tilde{a}^{ij}(\tilde{x}) = \tilde{a}^{ij}$, $\varepsilon_i = \gamma_{x_i}$, and the coefficients[11] $\hat{a}^{ij}(\tilde{x})$ is defined by the third equality[12]. Since $\varepsilon_N = 0$ and the other $\varepsilon_i$ are uniformly bounded, we have that

$$|\xi| \geq C|\tilde{\xi}| \tag{35.15}$$

for some constant $C > 0$, whence

$$\tilde{Q}(\tilde{\xi}) \geq \tilde{\theta}C^2|\tilde{\xi}|^2.$$

Thus the operator

$$\tilde{L} = -\frac{\partial}{\partial\tilde{x}_j}\hat{a}^{ij}(\tilde{x})\frac{\partial}{\partial\tilde{x}_i}$$

is uniformly elliptic, and $\tilde{u}$ is a solution of

$$\tilde{L}u = \tilde{f}$$

in $H_0^1(\mathbb{R}^N_+)$ with compact support, with $\tilde{f}$ defined by (35.8) and

$$\tilde{f}(\tilde{x}) = \hat{f}(x). \tag{35.16}$$

## 35.4 Flattening as a Sobolev map

We consider the flattening transformation again. Recall (35.13) as

$$u(x) = \tilde{u}(\tilde{x}), \quad u_{x_i} = w_i = \tilde{u}_{\tilde{x}_i} - \gamma_i\tilde{u}_{\tilde{x}_N}, \quad \gamma_i = \gamma_{x_i}, \quad \gamma_N = 0, \tag{35.17}$$

in which

$$x_i = \tilde{x}_i \quad \text{for} \quad i = 1, \ldots, N-1 \quad \text{and} \quad \tilde{x}_N = x_N - \gamma(x_1, \ldots, x_{N-1})$$

is a transformation that is clearly invertible. In Section 35.8 we need the following theorem[13] with $k = 2$.

---

[11] We use hats as superscripts because the tildes were already in use.

[12] Here we see that we need $\partial\Omega \in C^2$ to have $\hat{a}^{ij} \in C^1$.

[13] We can replace $H^k$ by $W^{1,k}$ of course.

**Theorem 35.2.** *Assume that*

$$\gamma \in C^k([a_1, b_1] \times \cdots \times [a_{N-1}, b_{N-1}]),$$

*and let*

$$U = \{x = (x', x_N) : x' \in (a_1, b_1) \times \cdots \times (a_{N-1}, b_{N-1}),\ x_N > \gamma(x')\};$$

$$\tilde{U} = \{\tilde{x} = (x', \tilde{x}_N) : x' \in (a_1, b_1) \times \cdots \times (a_{N-1}, b_{N-1}),\ \tilde{x}_N > \gamma(x')\}.$$

*Then (35.17) defines a linear bijection between $H^k(U)$ and $H^k(\tilde{U})$ which is continuous in both directions.*

**Proof**. Let $k = 1$. Every $\tilde{u}$ in $H^1(\tilde{U})$ *defines* functions $u, w_1, \ldots, w_N$ in $L^2(U)$ via (35.17). We have to show that $u \in H^1(U)$ with $w_i = u_{x_i}$ in $L^2(U)$. By the chain rule we know that that every $\tilde{v} \in C_c^1(\tilde{U})$ defines $v \in C_c^1(U)$ via $v(x) = \tilde{v}(\tilde{x})$ and vice versa, and

$$v_{x_i} = \tilde{v}_{\tilde{x}_i} - \gamma_i \tilde{v}_{\tilde{x}_N}.$$

Since $d\tilde{x} = dx$ it follows that

$$\int_U w_i v + \int_U u v_{x_i} = \int_{\tilde{U}} (\tilde{u}_{\tilde{x}_i} - \gamma_i \tilde{u}_{\tilde{x}_N})\tilde{v} + \int_{\tilde{U}} \tilde{u}(\tilde{v}_{\tilde{x}_i} - \gamma_i \tilde{v}_{\tilde{x}_N}). \tag{35.18}$$

and with $i = N$ this reads

$$\int_U w_N v + \int_U u v_{x_N} = \int_{\tilde{U}} \tilde{u}_{\tilde{x}_N} \tilde{v} + \int_{\tilde{U}} \tilde{u} \tilde{v}_{\tilde{x}_N} = 0,$$

because $\tilde{u} \in H^1(\tilde{U})$. This proves $w_N = u_{x_N}$ in $L^2(U)$. Thereby the terms with $\gamma_i$ in (35.18) cancel, because $\gamma_i \tilde{v} \in C_c^1(\tilde{U})$ and $(\gamma_i \tilde{v})_{\tilde{x}_N} = \gamma_i \tilde{v}_{\tilde{x}_N}$. But then we're left with

$$\int_U w_i v + \int_U u v_{x_i} = \int_{\tilde{U}} \tilde{u}_{\tilde{x}_i} \tilde{v} + \int_{\tilde{U}} \tilde{u} \tilde{v}_{\tilde{x}_i} = 0,$$

again because $\tilde{u} \in H^1(\tilde{U})$. This proves $w_i = u_{x_i}$ in $L^2(U)$ for all the other $i$. So (35.17) maps $\tilde{u} \in H^1(\tilde{U})$ to $u \in H^1(U)$.

Likewise we have that (35.17) maps $u \in H^1(U)$ to $\tilde{u} \in H^1(\tilde{U})$. Thus (35.17) defines a (linear) bijection between $H^1(U)$ and $H^1(\tilde{U})$, which is continuous in both directions, because as in (35.15) we have

$$|\nabla \tilde{u}| \leq C|\nabla u| \quad \text{and} \quad |\nabla u| \leq C|\nabla \tilde{u}|.$$

Note that

$$u_{x_i} = \tilde{u}_{\tilde{x}_i} - \gamma_i \tilde{u}_{\tilde{x}_N}$$

in (35.17) inverts as

$$\tilde{u}_{\tilde{x}_i} = u_{x_i} + \gamma_i u_{x_N}. \tag{35.19}$$

This completes the proof for $k = 1$. The proof for $k > 1$ is similar but not needed, as the statement in the theorem can be applied to the terms in (35.19) separately, to give

$$\tilde{u}_{\tilde{x}_i\tilde{x}_j} = u_{x_i x_j} + \gamma_j u_{x_N x_j} + \gamma_{ij} u_{x_N} + \gamma_i u_{x_N x_j} + \gamma_i \gamma_j u_{x_N x_N} \tag{35.20}$$

for $k = 2$ and so on. □

## 35.5 Garbage distribution

For later purposes we spell out what (35.16) does to (35.8), in which we distribute the garbage as

$$\begin{aligned} \hat{f} &= \zeta(f - cu - b^i u_{x_i}) - \zeta_{x_i}(2a^{ij} u_{x_j} + a^{ij}_{x_j} u) - \zeta_{x_i x_j} a^{ij} u \\ &= \zeta f - \underbrace{(\zeta c + \zeta_i a^{ij}_j + \zeta_{ij} a^{ij})}_{\Gamma} u - \underbrace{(\zeta b^i + 2\zeta_j a^{ij})}_{\beta^i} \underbrace{w_i}_{u_{x_i}}, \end{aligned}$$

with subscripts for derivatives of known functions. The first order derivatives of the unknown solution are denoted by $w_i$. Recall that the function $\zeta$ is compactly supported and smooth, and so are its derivatives. Thus the coefficients $\Gamma$ and $\beta_i$ are compactly supported, and as smooth as the smoothness of $a^{ij}$ and $b^i$ allow.

Applying (35.16) we use the notation as in (35.11) for $v$ for $f$ and $u$, but since $\tilde{f}$ and $\tilde{u}$ are already used in (35.16) and (35.11), we set

$$f(x) = \tilde{f}_0(\tilde{x}), \quad u(x) = w_0(x) = \tilde{w}_0(\tilde{x}) = U(\tilde{x}),$$

with $x$ restricted to the window in which $\zeta$ is supported.

For the first order derivatives of $u$ we *do not use this transformation rule*. To be consistent with (35.13) we *define the functions* $W_i = \tilde{w}_i$ as functions of $\tilde{x}$ by

$$u_{x_i} = w_i = \tilde{w}_i - \gamma_i \tilde{w}_N, \tag{35.21}$$

in which $\gamma_i$ is the partial derivative of $\gamma$ with respect to $x_i$. Recall that $\gamma$ and $\gamma_i$ do not change under the flattening transformation.

For the other (given) functions and coefficients we use again the same transformation rule as in (35.11) for $v$ and write

$$\zeta(x) = \tilde{\zeta}(\tilde{x}), \quad \beta^i(x) = \tilde{\beta}^i(\tilde{x}), \quad \Gamma(x) = \tilde{\Gamma}(\tilde{x}) = C(\tilde{x}),$$

and find that $\hat{f}$ as a function of $x$ transforms as

$$\tilde{f} = \underbrace{\tilde{\zeta}\tilde{f}_0}_{F} - CU - \underbrace{\tilde{\beta}^i(\tilde{w}_i - \gamma_i \tilde{w}_N)}_{B^i W_i} \tag{35.22}$$

as a function of $\tilde{x}$, i.e.

$$\tilde{f}(\tilde{x}) = F(\tilde{x}) - C(\tilde{x})U(\tilde{x}) - B^i(\tilde{x})W_i(\tilde{x}),$$

with $W_i = \tilde{w}_i$ defined by (35.21). The first term $F$ in (35.22) is the localised and then transformed original right hand side $f$ in (35.1). If $f$ is in $L^2(\Omega)$ then $F$ is in $L^2(\mathbb{R}^N_+)$. This only requires the function $\gamma$ used in the flattening transformation to be continuous[14].

The second[15] term $CU$ is the product of a compactly supported known function $C$ and $U$, the transformed but not localised unknown solution $u = w_0$. As for the third term, note that $B$ is related to $\beta$, but not simply as $B = \tilde{\beta}$. Still, every term $B^i W_i$ is the product of a compactly supported known function $B^i$, and the unknown function $W_i$ defined by (35.21) in terms of $w_1, \ldots, w_N$. Since we already know that $u = w_0$ is in $H^1(\Omega)$, the unknown functions $w_0, w_1, \ldots, w_N$ are all in $L^2(\Omega)$. Therefore $CU$ and $BW$ are in $L^2(\mathbb{R}^N_+)$ if $C$ and $B$ are bounded, which is certainly the case here because $c$ and $b^i$ are bounded by assumption.

## 35.6 Difference quotients of weak derivatives

We are now ready to use second order difference quotients $D_k^{-h} D_k^h \tilde{u}$ and observe that for $k < N$ these are in $H^1_0(\mathbb{R}^N_+)$, because we started out with $u \in H^1_0(\Omega)$. In what follows we drop the hats and tildes and note that the arguments are valid for any compactly supported[16] weak solution of

$$-(a^{ij} u_{x_i})_{x_j} = f \tag{35.23}$$

in $H^1_0(\mathbb{R}^N_+)$ if $f$ is in $L^2(\mathbb{R}^N_+)$. We have

$$\int_{\mathbb{R}^N_+} a^{ij} D_i u D_j v = \int_{\mathbb{R}^N_+} a^{ij} u_{x_i} v_{x_j} = \int_{\mathbb{R}^N_+} f v \quad \text{for all} \quad v \in H^1_0(\mathbb{R}^N_+), \tag{35.24}$$

and insert

$$v = D_k^{-h} D_k^h u \in H^1_0(\mathbb{R}^N_+) \quad \text{with} \quad k < N. \tag{35.25}$$

[14]For what's to come: $F$ is in $H^m(\mathbb{R}^N_+)$ if $f$ is in $H^m(\Omega)$ and $\gamma$ is in $C^m$.
[15]We chose to write $C$ for $\tilde{\Gamma}$.
[16]This is for convenience only, and suffices for our needs.

Note that for any $v$ in $H_0^1(\mathrm{I\!R}^N_+)$ we have that

$$D_k^h v(x) = \frac{1}{h}(v(x+he_k) - v(x))$$

defines $D_k^h v \in H_0^1(\mathrm{I\!R}^N_+)$ if $k < N$. It is easy to see that $D_i$ and $D_k^h$ commute, and that

$$D_i D_k^h v = D_k^h D_i v \in L^2(\mathrm{I\!R}^N_+) \tag{35.26}$$

for all $k < N$ and all $i$, including $i = N$. Moreover

$$D_k^h(uv)(x) = \frac{1}{h}(u(x+he_k)v(x+he_k) - u(x)v(x)) =$$

$$u(x+he_k)\frac{1}{h}(v(x+he_k) - v(x)) + \frac{1}{h}(u(x+he_k) - u(x))v(x) =$$

$$u(x+he_k)D_k^h v(x) + v(x)D_k^h u(x)$$

gives[17]

$$D_k^h(uv) = u_{he_k}D_k^h v + vD_k^h u, \tag{35.27}$$

and all terms in this Leibniz rule are certainly in $L^1(\mathrm{I\!R}^N_+)$ if $u$ and $v$ are in $H_0^1(\mathrm{I\!R}^N_+)$. We thus have

$$\int_{\mathrm{I\!R}^N_+} D_k^h(uv) = \int_{\mathrm{I\!R}^N_+} u_{he_k}D_k^h v + \int_{\mathrm{I\!R}^N_+} vD_k^h u.$$

The left hand side vanishes if $u$ or $v$ has compact support, in which case it follows that

$$\int_{\mathrm{I\!R}^N_+} vD_k^h u = -\int_{\mathrm{I\!R}^N_+} uD_k^{-h} v, \tag{35.28}$$

because

$$\int_{\mathrm{I\!R}^N_+} u(x+he_k)\frac{1}{h}(v(x+he_k) - v(x))\,dx = \int_{\mathrm{I\!R}^N_+} u(x)\underbrace{\frac{1}{h}(v(x) - v(x-he_k))}_{D_k^{-h}v(x)}\,dx.$$

We insert (35.25), which by the commutation rule (35.26) is equal to

$$D_j v = D_j D_k^{-h} D_k^h u = D_k^{-h} D_k^h D_j u,$$

in (35.24) and use the discrete integration by parts rule (35.28) for the second order terms to obtain

$$\int_{\mathrm{I\!R}^N_+} D_k^h(a^{ij}D_i u)D_k^h D_j u + \int_{\mathrm{I\!R}^N_+} fD_k^{-h}D_k^h u = 0.$$

[17] See (33.6).

The discrete Leibniz rule (35.27) applied to the first factor in the first integral then gives

$$\int_{\mathbb{R}^N_+} \underbrace{a^{ij}_{he_k} D^h_k D_i u \, D^h_k D_j u}_{\geq \theta \, |D^h_k \nabla u|^2} + \int_{\mathbb{R}^N_+} D^h_k a^{ij} \, D_i u \, D^h_k D_j u + \int_{\mathbb{R}^N_+} f D^{-h}_k D^h_k u = 0,$$

in which the first term is good but the middle term is bad[18] for our purposes. The third term will turn out to be quite innocent, just as the integral of $fu$ in (35.4). We'll take it from here with tricks that deserve a separate section.

## 35.7 Young estimates for the good and the bad

Taking the bad term on the left to the right we have

$$\theta \int_{\mathbb{R}^N_+} |D^h_k \nabla u|^2 + \int_{\mathbb{R}^N_+} f \, D^{-h}_k D^h_k u \leq - \int_{\mathbb{R}^N_+} \underbrace{D^h_k a^{ij} \, D_i u}_{(D^h_k a \, \nabla u)^j} \, \underbrace{D^h_k D_j u}_{(D^h_k \nabla u)_j}$$

$$\leq \frac{\theta}{2} \int_{\mathbb{R}^N_+} |D^h_k \nabla u|^2 + \frac{1}{2\theta} \int_{\mathbb{R}^N_+} \underbrace{|D^h_k a|_2^2}_{\substack{\text{squared Frobenius} \\ \text{matrix norm}}} \quad |\nabla u|^2,$$

in which we used Young's inequality with $p = q = 2$. i.e. with two squares, and $\varepsilon = \theta$.

Working towards (35.29) below, we picked the factor to appear squared as the bad coefficient of the $\theta$-term to be just half of the coefficient of the $\theta$-term on the left. This square is the bad term not under control yet. We also simplified the other square in the $\frac{1}{\theta}$-term using (16.10) with the Frobenius norm

$$|D^h_k a|_2^2 = \sum_{i,j=1}^{N} (D^h_k a^{ij})^2.$$

It follows that

$$\frac{\theta}{2} \int_{\mathbb{R}^N_+} |D^h_k \nabla u|^2 + \int_{\mathbb{R}^N_+} f D^{-h}_k D^h_k u \leq \frac{M_k}{2\theta} \int_{\mathbb{R}^N_+} |\nabla u|^2, \tag{35.29}$$

for some constant $M_k$ that depends[19] only on the bounds on the partial derivatives of the coefficients $a^{ij}$ with respect to $x_k$.

[18] But not too bad.

[19] To be precise, we can take

$$M_k = \sum_{i,j=1}^{N} |a^{ij}_{x_k}|_{\max}^2 .$$

The left hand side of (35.29) now contains the term we want to bound,

$$\int_{\mathbb{R}^N_+} |D_k^h \nabla u|^2 = \int_{\mathbb{R}^N_+} |\nabla D_k^h u|^2,$$

with a new prefactor $\frac{\theta}{2}$. We estimate the other term in the left hand side of (35.29) using[20]

$$\int_{\mathbb{R}^N_+} (D_k^{-h} w)^2 \leq \int_{\mathbb{R}^N_+} (D_k w)^2 \leq \int_{\mathbb{R}^N_+} |\nabla w|^2 \tag{35.30}$$

for $w = D_k^h u \in H_0^1(\mathbb{R}^N_+)$ via again Young's inequality as

$$\begin{aligned} \Big|\int_{\mathbb{R}^N_+} f D_k^{-h} D_k^h u\Big| &\leq \frac{\theta}{4} \int_{\mathbb{R}^N_+} (D_k^{-h} \underbrace{D_k^h u}_{w})^2 + \frac{1}{\theta} \int_{\mathbb{R}^N_+} f^2 \\ &\leq \frac{\theta}{4} \int_{\mathbb{R}^N_+} |\nabla \underbrace{D_k^h u}_{w}|^2 + \frac{1}{\theta} \int_{\mathbb{R}^N_+} f^2. \end{aligned}$$

Again we picked the factor to appear squared as the bad coefficient of the $\theta$-term to be just half of the coefficient of the $\theta$-term on the left that we want to estimate. It thus follows that

$$\frac{\theta}{4} \int_{\mathbb{R}^N_+} |D_k^h \nabla u|^2 \leq \frac{M_k}{2\theta} \int_{\mathbb{R}^N_+} |\nabla u|^2 + \frac{1}{\theta} \int_{\mathbb{R}^N_+} f^2. \tag{35.31}$$

By now the first term on the right hand side of (35.29) is under control, because

$$\theta \int_{\mathbb{R}^N_+} |\nabla u|^2 \leq \int_{\mathbb{R}^N_+} a^{ij} u_{x_i} u_{x_j} = \int_{\mathbb{R}^N_+} f u \leq \frac{\varepsilon}{2} \int_{\mathbb{R}^N_+} f^2 + \frac{1}{2\varepsilon} \int_{\mathbb{R}^N_+} u^2.$$

This time the choice of $\varepsilon > 0$ in Young's inequality is not so important anymore. The estimate is even easier than, but similar to the energy estimate in (35.5), which is of course already in the bag in case we deal with a solution obtained from a solution in $H_0^1(\Omega)$ via localisation and flattening.

Combined with (35.31) it follows that with some constant $C_k$ that can be made explicit in terms of $\theta$ and $M_k$ only, the a priori estimate

$$\int_{\mathbb{R}^N_+} |D_k^h \nabla u|^2 \leq C_k \int_{\mathbb{R}^N_+} (u^2 + f^2) \tag{35.32}$$

[20] To do: this holds via density and straightforward calculations for $w \in C_c^1(\mathbb{R}^N_+)$.

holds for all $k = 1, \dots, N-1$. A weak limit argument for $h \to 0$ then shows[21] that the weak $x_k$-derivatives of all $u_{x_i}$ exist in $L^2(\mathrm{I\!R}^N_+)$.

The only second order derivative not yet in $L^2(\mathrm{I\!R}^N_+)$ is the pure second order derivative $u_{x_N x_N}$, but note that (35.24) can now be written as

$$\int_{\mathrm{I\!R}^N_+} a^{NN} u_{x_N} v_{x_N} = \int_{\mathrm{I\!R}^N_+} v\big(f + \sum_{(i,i)\neq(N,N)} (a^{ij} u_{x_i x_j} + a^{ij}_{x_j} u_{x_i})\big) \tag{35.33}$$

for all $v \in H^1_0(\mathrm{I\!R}^N_+)$. Every term in the big factor on the right exists[22] in $L^2(\mathrm{I\!R}^N_+)$, so $a^{NN} u_{x_N}$ has a weak derivative with respect to $x_N$ in $L^2(\mathrm{I\!R}^N_+)$, and thereby[23] so does $u_{x_N}$. All terms in the partial differential equation (35.23) now exist in $L^2(\mathrm{I\!R}^N_+)$, and the equation also holds in $L^2(\mathrm{I\!R}^N_+)$! Writing $|\nabla\nabla u|^2$ for the sum of all the squared partial derivatives we have proved the following theorem for compactly[24] supported $u \in H^1_0(\mathrm{I\!R}^N_+)$.

**Theorem 35.3.** *A weak solution $u \in H^1_0(\mathrm{I\!R}^N_+)$ of*

$$-(a^{ij} u_{x_i})_{x_j} = f$$

*is in $H^2(\mathrm{I\!R}^N_+)$ provided $f \in L^2(\mathrm{I\!R}^N_+)$, the coefficients $a^{ij}$ and their first order derivatives are continuous and bounded, and for some $\theta > 0$ it holds that*

$$a^{ij}(x)\xi_i\xi_j \geq \theta|\xi|^2$$

*for all $x \in \mathrm{I\!R}^N_+$ and all $\xi \in \mathrm{I\!R}^N$. Moreover, the equation*

$$\underbrace{a^{ij} u_{x_i x_j} + a^{ij}_{x_j} u_{x_i}}_{(a^{ij} u_{x_i})_{x_j}} + f = 0 \tag{35.34}$$

*is satisfied (with each term) in $L^2(\mathrm{I\!R}^N_+)$, and there exists a constant $C > 0$ depending only on the ellipticity constant $\theta > 0$ and the bounds on $a^{ij}$ and $a^{ij}_{x_k}$ such that*

$$\int_{\mathrm{I\!R}^N_+} |\nabla\nabla u|^2 + \int_{\mathrm{I\!R}^N_+} |\nabla u|^2 \leq C \int_{\mathrm{I\!R}^N_+} (u^2 + f^2)$$

*for every such solution $u$. Vice versa, if $u \in H^2(\mathrm{I\!R}^N_+) \cap H^1_0(\mathrm{I\!R}^N_+)$, then (35.34) defines $f \in L^2(\mathrm{I\!R}^N_+)$.*

[21] To do: also not very difficult, and again only needed on half spaces.
[22] By Definition 32.6 and the Leibniz rule in Theorem 32.10.
[23] Again by Theorem 32.10.
[24] Remember this restriction is only for convenience, but suffices for what follows.

It remains to prove the statement

$$\int_{\mathbb{R}^N_+} (D_k^{-h} w)^2 \leq \int_{\mathbb{R}^N_+} (D_k w)^2 \leq \int_{\mathbb{R}^N_+} |\nabla w|^2$$

in (35.30) for all $w \in H_0^1(\mathbb{R}^N_+)$, which we needed to get to (35.31), and its counterpart to get from (35.32) to the same estimate for $\nabla u_{x_k}$. The first one follows from a calculation for $w \in C_c^1(\mathbb{R}^N_+)$ that goes along lines we have seen before. We follow Step 1 of Evans' proof in Section 5.8.2.a and use for $k < N$ and $h \in \mathbb{R}$ that

$$w(x + he_k) - w(x) = h \int_0^1 D_k w(x + the_k)\, dt,$$

whence

$$\int_{\mathbb{R}^N_+} |D_k^h w|^2 \leq \int_{\mathbb{R}^N_+} \int_0^1 |D_k w(x + the_k)|^2 \, dt\, dx = \int_{\mathbb{R}^N_+} |D_k w|^2,$$

an estimate which then also holds for all $w \in H_0^1(\mathbb{R}^N_+)$ by the density.

For the counterpart we use the discrete integration by parts formula

$$\int_{\mathbb{R}^N_+} w\, D_k^h \phi = - \int_{\mathbb{R}^N_+} \phi\, D_k^{-h} w \tag{35.35}$$

for $\phi \in C_c^1(\mathbb{R}^N_+)$ and $w \in L^2(\mathbb{R}^N_+)$ with

$$\int_{\mathbb{R}^N_+} |D_k^h w|^2 \leq C$$

for all $h \in \mathbb{R}$. Letting $h \to 0$ the left hand side of (35.35) converges to

$$\int_{\mathbb{R}^N_+} w\, \phi_{x_k}$$

if $h \to 0$. Since the functions $D_k^{-h} w$ are bounded in $L^2(\mathbb{R}^N_+)$ by $C$, there is a sequence $h_n \to 0$ and a function $v_k$ in $L^2(\mathbb{R}^N_+)$ with

$$\int_{\mathbb{R}^N_+} |v_k|^2 \leq C, \tag{35.36}$$

such that the right hand side converges to

$$- \int_{\mathbb{R}^N_+} \phi\, v_k$$

for every $\phi \in L^2(\mathrm{I\!R}^{\mathrm{N}}_+)$. It follows that

$$\int_{\mathrm{I\!R}^{\mathrm{N}}_+} w\, \phi_{x_k} = -\int_{\mathrm{I\!R}^{\mathrm{N}}_+} \phi\, v_k,$$

so $v_k$ is the weak $x_k$-derivative of $w$, and (35.36) holds, as desired in relation to (35.32) for the conclusion in Theorem 35.3. Note that we have used

**Theorem 35.4.** *Let $H$ be a Hilbert space, $C > 0$, and $x_n$ a sequence in $H$ with $|x_n| \le C$ for all $n$. Then there exists $x \in H$ with $|x| \le C$ and a strictly increasing sequence $n_k$ in $\mathrm{I\!N}$ such that*

$$x_{n_k} \cdot y \to x \cdot y$$

*for all $y \in H$. We say that $x_{n_k}$ converges weakly $x$ in $H$.*

**Exercise 35.5.** Prove this theorem for the standard Hilbert space in Exercise 30.28 and use Remark 30.29 to conclude that the above reasoning for the right hand side of (35.35) is valid.

**Remark 35.6.** *If $u \in H^1_0(\mathrm{I\!R}^{\mathrm{N}}_+)$ and $D_k u \in L^2(\mathrm{I\!R}^{\mathrm{N}}_+)$ then $D^h_k u$ is bounded in $H^1_0(\mathrm{I\!R}^{\mathrm{N}}_+)$ by the argument above for (35.30). By Theorem 35.4 $D^h_k u$ converges weakly in $H^1_0(\mathrm{I\!R}^{\mathrm{N}}_+)$ along some sequence $h_n \to 0$, and it follows that $D_k u \in H^1_0(\mathrm{I\!R}^{\mathrm{N}}_+)$.*

## 35.8 Regularity for zero boundary data

Everything we did in the previous two sections applies to solutions obtained from localizing and flattening solutions $u \in H^1_0(\Omega)$ as we did in Section 35.2 and Section 35.3. Thus we now obtain the same result for solutions of

$$Lu = -(a^{ij} u_{x_i})_{x_j} + b^i u_{x_i} + cu = f$$

by translating Theorem 35.3 for $\tilde{u}$ defined from the terms in (35.6) by (35.11) back to each $\zeta u$. Here we only need that $\tilde{u} \in H^2(\mathrm{I\!R}^{\mathrm{N}}_+)$ implies that $\zeta u \in H^2(\Omega)$, which is the inverse of the statement we announced in Section 35.5 as a footnote. For general coordinate transformations this statement is not obvious[25], but for the flattening transformation it follows directly from the definition of weak derivatives[26]. The first result for solutions $u \in H^1_0(\Omega)$ as in Section 35.1 is now in the pocket.

[25] See also Exercise 33.14 and the exercises above it.

[26] See Section 35.4.

**Theorem 35.7.** *If the coefficients $a^{ij}$ are in $C^1(\bar{\Omega})$, $b^i$ and $c$ are in $C(\bar{\Omega})$, $\partial\Omega$ is in $C^2$, and $f$ is in $L^2(\Omega)$, then indeed every weak solution $u \in H_0^1(\Omega)$ of $Lu = f$ is in $H^2(\Omega)$ and*

$$\int_\Omega |\nabla\nabla u|^2 + \int_\Omega |\nabla u|^2 \le C \int_\Omega (u^2 + f^2).$$

*The constant $C$ only depends on $\theta$, $\Omega$ and the bounds on $a^{ij}, a^{ij}_{x_k}, b^i, c$.*

## 35.9 Higher order regularity

To see what we need to get $u$ in $H^3$ we assume that $u$ is as in Theorem 35.3, $\phi \in C_c^\infty(\mathrm{IR}^{\mathrm{N}}_+)$, and take $v = \phi_{x_k}$ with $k < N$ in (35.24) to obtain

$$\int_{\mathrm{IR}^{\mathrm{N}}_+} a^{ij} u_{x_i} \phi_{x_k x_j} = \int_{\mathrm{IR}^{\mathrm{N}}_+} f \phi_{x_k}.$$

Using $\phi_{x_k x_j} = \phi_{x_j x_k}$, Definition 32.6 and the Leibniz rule, this rewrites as

$$\int_{\mathrm{IR}^{\mathrm{N}}_+} f_{x_k} \phi = \int_{\mathrm{IR}^{\mathrm{N}}_+} (a^{ij} u_{x_i})_{x_k} \phi_{x_j} = \int_{\mathrm{IR}^{\mathrm{N}}_+} a^{ij} u_{x_i x_k} \phi_{x_j} + \int_{\mathrm{IR}^{\mathrm{N}}_+} a^{ij}_{x_k} u_{x_i} \phi_{x_j}.$$

Then another application of Definition 32.6 and the Leibniz rule give

$$\int_{\mathrm{IR}^{\mathrm{N}}_+} a^{ij} u_{x_k x_i} \phi_{x_j} = \int_{\mathrm{IR}^{\mathrm{N}}_+} (f_{x_k} - (a^{ij}_{x_k} u_{x_i})_{x_j}) \phi$$

$$= \int_{\mathrm{IR}^{\mathrm{N}}_+} (f_{x_k} - a^{ij}_{x_k} u_{x_i x_j} - a^{ij}_{x_k x_j} u_{x_i}) \phi,$$

provided also the second order derivatives of $a^{ij}$ are continuous, and $f_{x_k}$ is in $L^2(\mathrm{IR}^{\mathrm{N}}_+)$ . It follows that $u_{x_k}$ is a weak solution of (35.23) with $f$ replaced by

$$f_k = f_{x_k} - a^{ij}_{x_k} u_{x_i x_j} - a^{ij}_{x_k x_j} u_{x_i}, \tag{35.37}$$

and we have another another theorem for free[27]. Applying Theorem 35.3 to all $u_{x_k}$ with $k < N$ it follows that all third order derivatives are in $L^2(\mathrm{IR}^{\mathrm{N}}_+)$ with similar estimates.

**Theorem 35.8.** *A weak solution $u \in H_0^1(\mathrm{IR}^{\mathrm{N}}_+)$ of*

$$-(a^{ij} u_{x_i})_{x_j} = f$$

[27] Remark 35.6 says that $u_{x_k}$ is in $H_0^1(\mathrm{IR}^{\mathrm{N}}_+)$.

*is in* $H^3(\mathbb{R}^N_+)$ *provided* $f \in H^1(\mathbb{R}^N_+)$, *the coefficients* $a^{ij}$ *and their first and second order derivatives are continuous and bounded, and for some* $\theta > 0$ *it holds that*
$$a^{ij}(x)\xi_i\xi_j \geq \theta|\xi|^2$$
*for all* $x \in \mathbb{R}^N_+$ *and all* $\xi \in \mathbb{R}^N$. *Moreover, there exists a constant* $C > 0$ *depending only on the ellipticity constant* $\theta > 0$ *and the bounds on* $a^{ij}$, $a^{ij}_{x_k}$, $a^{ij}_{x_k x_l}$ *such that*

$$\int_{\mathbb{R}^N_+} |\nabla\nabla\nabla u|^2 + \int_{\mathbb{R}^N_+} |\nabla\nabla u|^2 + \int_{\mathbb{R}^N_+} |\nabla u|^2 \leq C \int_{\mathbb{R}^N_+} (u^2 + f^2 + |\nabla f|^2)$$

*for every such solution* $u$.

Thus we can give the wheel another turn if for our solution $\tilde{u} \in H^1_0(\mathbb{R}^N_+)$ at hand it holds that it solves the equation in Theorem 35.3 with $\tilde{f} \in H^1(\mathbb{R}^N_+)$. In view of (35.37) this amounts to having

$$\tilde{f}_k = \tilde{f}_{\tilde{x}_k} - \tilde{a}^{ij}_{\tilde{x}_k}\tilde{u}_{\tilde{x}_i\tilde{x}_j} - \tilde{a}^{ij}_{\tilde{x}_k\tilde{x}_j}\tilde{u}_{\tilde{x}_i} \in L^2(\mathbb{R}^N_+),$$

of which we already know that the second term is in $L^2(\mathbb{R}^N_+)$. So is the third if the second order derivatives of $\tilde{a}^{ij}$ are continuous and bounded, which is certainly the case if $a^{ij} \in C^2(\bar{\Omega})$ and $\partial\Omega \in C^2$. What do we need to have the weak partial derivatives $\tilde{f}_{\tilde{x}_k}$ in $L^2(\mathbb{R}^N_+)$? We use Theorem 35.2 to answer the question by verifying what we need to have $\tilde{f} \in H^1(\mathbb{R}^N_+)$.

Recall that $\tilde{f}$ in (35.22) was defined from

$$\hat{f} = \zeta f - \underbrace{(\zeta c + \zeta_i a^{ij}_j + \zeta_{ij}a^{ij})u}_{g_0} - \underbrace{(\zeta b^i + 2\zeta_j a^{ij})u_{x_i}}_{g_i}$$

via $\tilde{f}(\tilde{x}) = \hat{f}(x)$, whence[28]

$$\tilde{f} = \tilde{\zeta}\tilde{f}_0 - \tilde{g}_0 - \tilde{g}_1 - \cdots - \tilde{g}_N.$$

The first term is in $H^1(\mathbb{R}^N_+)$ if we started with $f$ in $H^1(\Omega)$.

If we assume that $c \in C^1(\bar{\Omega})$ then the second term is in $H^1_0(\mathbb{R}^N_+)$ because $u$ is in $H^1_0(\Omega)$, and its multiplied by the $C^1$-function $\zeta c + \zeta_i a^{ij}_j + \zeta_{ij}a^{ij}$, in which the terms with $a^{ij}$ are already $C^1$ by assumption. This gives a localized

$$g_0 = (\zeta c + \zeta_i a^{ij}_j + \zeta_{ij}a^{ij})\, u \in H^1_0(\Omega),$$

[28]We don't use (35.22) now.

which has the same support as $\zeta$, and maps to[29] a localized $\tilde{g}_0 \in H^1_0(\mathrm{IR}^{\mathrm{N}}_+)$.

As for the third term, Theorem 35.7 says that each $u_{x_i}$ is in $H^1(U)$. If we assume that also all $b^i$ are in $C^1(\bar{\Omega})$, then multiplying the functions $u_{x_i}$ by $C^1$-functions $\zeta b^i + 2\zeta_j a^{ij}$ gives localised functions

$$g_i = (\zeta b^i + 2\zeta_j a^{ij})\, u_{x_i} \in H^1(\Omega).$$

Again these have the same support as $\zeta$, and map to localised functions $\tilde{g} \in H^1(\mathrm{IR}^{\mathrm{N}}_+)$ by Theorem 35.2 with $k = 1$.

**Theorem 35.9.** *If the coefficients $a^{ij}$ are in $C^2(\bar{\Omega})$, $b^i$ and $c$ are in $C^1(\bar{\Omega})$, $\partial\Omega$ is in $C^3$, and $f$ is in $H^1(\Omega)$, then every weak solution $u \in H^1_0(\Omega)$ of $Lu = f$ is in $H^3(\Omega)$, and*

$$\int_\Omega |\nabla\nabla\nabla u|^2 + |\nabla\nabla u|^2 + \int_\Omega |\nabla u|^2 \leq C \int_\Omega (u^2 + |\nabla f|^2 + f^2).$$

*The constant $C$ only depends on $\theta$, $\Omega$ and all the bounds on the coefficients and the derivatives used.*

**Proof**. The assumptions make that $\tilde{f}$ is in $H^1(\Omega)$, as explained above. Therefore Theorem 35.8 applies to $\tilde{u}$. It only remains to apply Theorem 35.2 with $k = 3$ to conclude that $\zeta u$ is in $H^3(\Omega)$. $\square$

And then the wheel keeps on turning. The GNS- and Morrey-estimates finish the job, with eigenfunctions of $L$ as a special case of interest. Again all arguments are easiliest done for $\tilde{u} \in H^1_0(\mathrm{IR}^{\mathrm{N}}_+)$ with bounded support.

[29] We only need $\tilde{g}_0 \in H^1(\mathrm{IR}^{\mathrm{N}}_+)$.

# 36 Exercises about weak solutions

We write $H^1(0,1)$ for the Sobolev space of $u \in L^2(0,1)$ with $u' \in L^2(0,1)$, and

$$||u||^2_{H^1} = \int_0^1 (u^2 + u'^2)$$

defines the $H^1$-norm on $H^1(0,1)$. Recall that $u' \in L^2(0,1)$ means that the weak derivative of $u'$ of $u$ satisfies

$$\int_0^1 u'\phi + \int_0^1 u\phi' = 0$$

for all $\phi \in C^1_c(0,1)$, and that the closure of $C^1_c(0,1)$ in the $H^1$-norm is denoted by $H^1_0(0,1)$.

If you like the space $L^2(0,1)$ may be defined as the abstract closure of $C^1_c(0,1)$ or $C_c(0,1)$ with respect to the 2-norm. For every $f \in L^2(0,1)$ the mollified functions $f^\varepsilon$ are defined by convolution with the usual mollifiers as smooth compactly supported functions by

$$f^\varepsilon(x) = \int_{\mathbb{R}} \eta_\varepsilon(x-y) f(y)\, dy,$$

in which the integrals may be obtained via (with respect to the 2-norm) Cauchy sequences $f_n \in C^\infty_c(0,1) \subset C^\infty_c(\mathbb{R})$. In particular Theorem 32.13 applies.

**Exercise 36.1.** Let $u_n \in C^1_c(0,1)$ be a Cauchy sequence with respect to the $H^1$-norm.

a) Prove that $u_n$ is uniformly convergent, and explain why its limit $u$ in $H^1_0(0,1)$ is also in $C([0,1])$, with $u(0) = u(1) = 0$.

b) Use Young's inequality with $p = q = 2$ and $(u_n^2)' = 2u_n u_n'$ to prove that

$$\int_0^1 u_n^2 \leq \int_0^1 u_n'^2 \quad \text{and thereby} \quad \int_0^1 u^2 \leq \int_0^1 u'^2,$$

the Poincaré inequality for $u$ in $H^1_0(0,1)$. Hint: integrate from $0$ to $x \leq \frac{1}{2}$ for control on $(0, \frac{1}{2})$ and then use symmetry for control on $(\frac{1}{2}, 1)$.

**Exercise 36.2.** Prove that the closure of $C^1([0,1])$ in $H^1(0,1)$ is $H^1(0,1)$.

**Exercise 36.3.** Let $\tilde{H}^1(0,1)$ be the set of all $u \in H^1(0,1)$ with $\int_0^1 u = 0$. Prove that $\tilde{H}^1(0,1)$ is a closed subspace of $H^1(0,1)$, and that the Poincaré inequality

$$\int_0^1 u^2 \leq 4 \int_0^1 u'^2,$$

holds for all $u \in \tilde{H}^1(0,1)$. Hint: prove that $u \in C([0,1])$ and use the intermediate value theorem.

**Exercise 36.4.** Consider the partial differential operator $L$ defined by

$$Lu = -(au')' + bu' + cu$$

for $u \in C^2([0,1])$, with coefficients $a, b, c \in C([0,1])$, and assume that $a(x) \geq \theta > 0$ and $c(x) \geq \mu > 0$ for all $x \in [0,1]$. Recall that $u \in H^1(0,1)$ is called a weak solution of $Lu = f$ if $B(u,v) = \int_0^1 fv$ for all $v \in H_0^1(0,1)$.

a) Show there exists $\delta > 0$ such that the bilinear form defined by

$$B(u,v) = \int_0^1 (au'v' + bu'v + cuv)$$

has the property that

$$\exists_{\beta>0} \forall_{u \in H^1(0,1)} \quad B(u,u) \geq \beta ||u||^2_{H^1(0,1)}$$

if $b(x)^2 < 4\mu\theta$ for all $x \in [0,1]$. Hint: use Young's inequality with $p = q = 2$ in the form

$$ab \leq \varepsilon a^2 + \frac{b^2}{4\varepsilon}.$$

b) If so show $Lu = f$ has a unique weak solution in $H_0^1(0,1)$ for all $f \in L^2(0,1)$.

c) Show that $au \in H^1(0,1)$ if $a \in C^1([0,1])$ and $u \in H^1(0,1)$.

d) If so show that a weak solution $u \in H^1(0,1)$ is in

$$H^2(0,1) \cap H_0^1(0,1) = \{u \in H_0^1(0,1) : u' \in H^1(0,1)\}$$

if $f \in L^2(0,1)$. Hint: so you have to show that $u'' \in L^2(0,1)$.

e) Explain why $L$ is then a linear homeomorphism between $H^2(0,1) \cap H_0^1(0,1)$ and $L^2(0,1)$.

f) Generalise the results to the case that $b \equiv c \equiv 0$.

**Exercise 36.5.** Let $L$ be as in Exercise 36.4 with the same assumptions on the coefficients $a, b, c$. The unique weak solution in $H_0^1(0,1)$ is called the weak solution of $Lu = f$ with homogeneous Dirichlet boundary condition $u = 0$ on the boundary of the open interval $(0,1)$.

a) Explain why a weak solution $u \in H^1(0,1)$ of $Lu = f$ with homogeneous Neumann boundary condition $u_x = 0$ on the boundary is defined by $B(u,v) = \int_0^1 fv$ for all $v \in H^1(0,1)$. Under the assumptions on $a, b, c$ prove that there exists a unique weak solution (of the homogeneous Neumann boundary problem) $u$ in $H^1(0,1)$ for all $f \in L^2(0,1)$.

b) Assume that $a \in C^1([0,1])$. Prove that every weak solution $u \in H^1(0,1)$ of $Lu = f$ is in $H^2(0,1)$ if $f \in L^2(0,1)$.

c) Now consider the case that $b \equiv 0$ and $c(x) = \frac{1}{n}$. Denote the solution of the homogeneous Neumann boundary problem for $L$ by $u_n$. Derive a condition on $f$ necessary for the convergence of $u_n$ in $H^1(0,1)$. Hint: take a convenient function $v \in H^1(0,1)$ in the definition.

d) Let $f \in L^2(0,1)$. Prove that there exists a unique

$$u \in \tilde{H}^1(0,1) = \{u \in H^1(0,1) : \int_0^1 u = 0\}$$

such that

$$\int_0^1 au'v' = \int_0^1 fv$$

for every $v \in \tilde{H}^1(0,1)$.

e) Under which condition on $f$ does it hold that this $u$ is a weak solution of the differential equation $-(au')' = f$ on $(0,1)$?

f) Consider again the case that $b \equiv 0$, and consider the solution set $U_\lambda$ of all nonzero weak solutions $u \in H^1(0,1)$ of $Lu = \lambda u$ with homogeneous Neumann boundary condition $u_x = 0$. Explain why this set is empty for all $\lambda \le 0$ and give the Raleigh formula for the smallest $\lambda > 0$ for which $U_\lambda$ is not empty.

g) Derive a Raleigh formula for the smallest $\lambda$ for which $-(au')' = \lambda u$ has a nonzero weak solution with homogeneous Neumann boundary condition. Explain why this $\lambda$ is $0$ and describe $E_0$.

h) Use the Raleigh formula with $u$ restricted to the closed subspace of functions with $\int_0^1 uw =$ for every $w \in E_0$ to characterise the next smallest $\lambda$ for which $-(au')' = \lambda u$ has a nonzero weak solution with homogeneous Neumann boundary condition.

i) Take $a \equiv 1$ and compare to Exercise 36.3.

**Exercise 36.6.** Before we consider a weak solution approach for equations like

$$(x(1-x)u'(x))' + f(x) = 0$$

on $(0,1)$, we compare solving this equation to solving $u'' + f = 0$. Referring to the special case $\lambda = 1$ in (28.39) of Section 28.8 we also consider the *resolvent equations*

$$\lambda u(x) - u''(x) = f(x) \quad \text{and} \quad \lambda u(x) - (x(1-x)u'(x))' = f(x)$$

with $\lambda > 0$. Recall you solved $-u'' = f$ with homogeneous Dirichlet boundary conditions $u(0) = u(1) = 0$ as

$$u(x) = \int_0^1 g(x,s) f(s)\, ds,$$

with

$$g(x,s) = \begin{cases} (1-s)x & \text{for } x < s \\ s(1-x) & \text{for } x > s \end{cases}$$

by integrating $-u'' = f$ twice.

a) Read on after (28.42) and explain why $g(x,s)$ is the unique solution of

$$-u''(x) = \delta_s(x) = \delta(x-s)$$

with $u(0) = u(1) = 0$ if $0 < s < 1$.

b) For $\lambda > 0$ the function

$$u_\lambda(x) = \frac{\sinh \sqrt{\lambda}\, x}{\sqrt{\lambda}} = x + \frac{\lambda}{3!}x^3 + \frac{\lambda^2}{5!}x^5 + \frac{\lambda^4}{7!}x^7 + \cdots$$

is the unique solution of $\lambda u(x) - u''(x) = 0$ with $u(0) = 0$ and $u'(0) = 1$. Determine $a_0(s,\lambda), b_0(s,\lambda) > 0$ such that, for every $s \in (0,1)$,

$$g_\lambda(x,s) = \begin{cases} a_0(s,\lambda) u_\lambda(x) & \text{for } x < s \\ b_0(s,\lambda) u_\lambda(1-x) & \text{for } x > s \end{cases}$$

defines the (unique) solution of $\lambda u - u'' = \delta_s$ with homogeneous Dirichlet boundary conditions $u(0) = u(1) = 0$. You should get

$$g_\lambda(s,x) = g_\lambda(x,s) = \frac{\sinh \sqrt{\lambda}\,(1-s) \sinh \sqrt{\lambda}\, x}{\sqrt{\lambda} \sinh \sqrt{\lambda}}$$

$$= \frac{(1 - s + \frac{\lambda(1-s)^3}{3!} + \frac{\lambda^2(1-s)^5}{5!} + \cdots)(x + \frac{\lambda x^3}{3!} + \frac{\lambda^2 x^5}{5!} + \cdots)}{1 + \frac{\lambda}{3!} + \frac{\lambda^2}{5!} + \cdots}$$

for $0 \le x \le s \le 1$. Notice how well behaved this is for $\lambda \to 0$.

c) Still for $\lambda > 0$: show that

$$\tilde{u}_\lambda(s,x) = \tilde{u}_\lambda(x,s) = \frac{\cosh\sqrt{\lambda}\,(1-s)\,\cosh\sqrt{\lambda}\,x}{\sqrt{\lambda}\,\sinh\sqrt{\lambda}} \quad (0 \le x \le s \le 1)$$

defines the (unique) solution of $\lambda u - u'' = \delta_s$ with homogeneous Neumann boundary conditions. Notice how ill behaved this is for $\lambda \to 0$.

d) Show for $\lambda > 0$ that nontrivial solutions of $\lambda u(x) - (x(1-x)u'(x))' = 0$ with $u(x)$ bounded as $x \to 0$ are power series with radius of convergence $1$ and $u(0) \neq 0$. In particular there is a unique power series solution

$$u(x) = U_\lambda(x) = 1 + \alpha_1(\lambda)x + \alpha_2(\lambda)x^2 + \cdots = \sum_{n=0}^{\infty} \alpha_n x^n$$

with $u(0) = 1$ and recurrence relation

$$\alpha_n(\lambda) = \frac{\lambda + n(n-1)}{n^2}\,\alpha_{n-1}(\lambda)$$

for $n \in \mathbb{N}$, and every solution is of the form

$$u(x) = A(1 + BV_\lambda(x))U_\lambda(x),$$

with $V_\lambda(x)$ a primitive of

$$\frac{1}{x(1-x)U_\lambda(x)}$$

that behaves like $\ln x$ as $x \downarrow 0$.

e) Use $U_\lambda$ to obtain

$$G_\lambda(x,s) = \begin{cases} A_0(s,\lambda)U_\lambda(x) & \text{for } x < s \\ B_0(s,\lambda)U_\lambda(1-x) & \text{for } x > s \end{cases},$$

as a solution of $\lambda u(x) - (x(1-x)u'(x))' = \delta(x-s)$ for $s \in (0,1)$, by choosing $A_0, B_0$ to make $G_\lambda(x,s)$ continuous in $x = s$ with a suitable jump condition for its $x$-derivative. You should get $A_0, B_0$ as the solution of

$$U_\lambda(s)A = U_\lambda(1-s)B \quad U_\lambda'(s)A + U_\lambda'(1-s)B = \frac{1}{s(1-s)}.$$

This suggest that for $\lambda > 0$ the solution of $\lambda u(x) - (x(1-x)u'(x))' = f(x)$ does not involve boundary conditions. What goes wrong for $\lambda = 0$?

f) Write and solve the recurrence relation as

$$\ln\alpha_n - \ln\alpha_{n+1} = \ln(1 - \frac{1}{n} + \frac{\lambda}{n^2}) = \ln(1 - \frac{1}{n}) + \ln(1 + \frac{\lambda}{n(n-1)})$$

$$= \ln\lambda + \underbrace{\sum_{k=2}^{n} \ln(1-\frac{1}{k})}_{-L_n} + \underbrace{\sum_{k=2}^{n} \ln(1+\frac{\lambda}{k(k-1)})}_{F_n(\lambda)=F(\lambda)-f_n(\lambda)},$$

with

$$F(\lambda) = \sum_{k=2}^{\infty} \ln(1+\frac{\lambda}{k(k-1)}), \quad f_n(\lambda) \sim \lambda \int_n^{\infty} \frac{dx}{x(x+1)} \sim \frac{\lambda}{n},$$

$$L_n = -\sum_{k=2}^{n} \ln(1-\frac{1}{k}) = \sum_{k=2}^{n} \frac{1}{k} + \underbrace{\sum_{m=2}^{\infty} \frac{1}{m} \sum_{k=2}^{n} \frac{1}{k^m}}_{<1}$$

$$= \sum_{k=2}^{n} \frac{1}{k} + \underbrace{\sum_{m=2}^{\infty} \frac{1}{m} \sum_{k=2}^{\infty} \frac{1}{k^m}}_{<1} - \sum_{m=2}^{\infty} \frac{1}{m} \underbrace{\sum_{k=n+1}^{\infty} \frac{1}{k^m}}_{<\frac{n^{1-m}}{m-1}}$$

$$= \ln n + \gamma - 1 + \sum_{m=2}^{\infty} \frac{1}{m} \sum_{k=2}^{\infty} \frac{1}{k^m} + O(\frac{1}{n})$$

to conclude that for $n \geq 2$ the coefficients are given by

$$\alpha_n = \lambda \exp(\sum_{k=2}^{n} \ln(1-\frac{1}{k}) \exp(\sum_{k=2}^{n} \ln(1+\frac{\lambda}{k(k-1)}) \sim \frac{C_\lambda}{n}$$

as $n \to \infty$, with[30]

$$C_\lambda = \lambda \exp(1 - \gamma - \sum_{m=2}^{\infty} \frac{1}{m} \sum_{k=2}^{\infty} \frac{1}{k^m} + F(\lambda)).$$

g) Verify that

$$U_\lambda(x) - C_\lambda \ln \frac{1}{1-x}$$

is a multiple of $U_\lambda(1-x)$.

h) For which complex $\lambda$ are the above calculations valid?

i) Solve the equation $C_\lambda = 0$.

j) Multiply $(x(1-x)u'(x))' = \lambda u(x)$ by $v$, interchange $u$, $v$ and subtract, to find that $x(1-x)(u'(x)v(x) - u(x)v'(x))$ is constant. Which constant do you get for $u(x) = U_\lambda(x)$ and $v(x) = U_\lambda(1-x)$? Evaluate $G_\lambda(x, s)$.

[30] If I have my expansions right.

**Exercise 36.7.** Consider the differential equation

$$Lu = -(au')' + bu' + cu = f$$

on some open interval $I \subset \mathbb{R}$, with $a, b, c$ sufficiently smooth and $a(x) > 0$ for all $x \in I$. Let $U$ be a solution with $f \equiv 0$ for which $U(x)$ has some well defined behaviour at the left endpoint of $I$ that makes $U$ unique up to a multiplicative constant, and $V$ be a solution for which $V(x)$ has some well defined behaviour at the right endpoint of $I$ that makes $V$ unique up to a multiplicative constant.

a) Verify for $s \in I$ that

$$G(x,s) = \frac{1}{a(s)(U'(s)V(s)) - (U(s)V'(s))} \begin{cases} V(s)U(x) & \text{for } x < s \\ U(s)V(x) & \text{for } x > s \end{cases}$$

is the unique solution of $Lu = \delta_s$ with these behaviours at both endpoints, provided the Wronskian $W(s) = U'(s)V(s)) - (U(s)V'(s)$ is nonzero.

b) Verify that $aW' = bW$ to conclude that $W$ is either identically zero or nowhere zero on $I$ to distinguish between either the existence of a solution operator or the existence of nontrivial solutions of the homogeneous equation, with the behaviours under consideration.

c) Exercise 36.6 concerned $I = (0,1)$, $b \equiv 0$, $c \equiv \lambda$, $V(x) = U(1-x)$, with various very special choices of $a$, and various choices of boundary behaviours. See if you want to rewrite $C_\lambda$ and $G_\lambda$.

d) Choose a new independent variable $z$ such that

$$x(1-x)\frac{d}{dx} = \frac{d}{dz}$$

and re-examine the equation in $z$ on $\mathbb{R}$.

**Exercise 36.8.** Consider the differential equation

$$Lu = -(xu')' = f$$

on the interval $(0,1]$.

a) Solve the equation with boundary condition $u(1) = 0$ with $f(x) \equiv 1$, $n \in \mathbb{N}$ and show that it has precisely one solution $u_n$ which is bounded on $(0,1]$. Which solution?

b) Same question but now with a general $f \in C([0,1])$. Prove that there exists a unique bounded solution $u$. Hint: use $F(x) = \int_0^x f$ and show in particular that

$$u(0) = \int_0^1 \frac{F(x)}{x}\, dx.$$

c) The weak solution approach for $Lu = f$ leads to the bilinear form

$$B(u,v) = \int_0^1 xu'(x)v'(x)\, dx$$

Use the result in (b) to show that

$$B(u,v) = \int_0^1 fv$$

for all $v \in C^1([0,1])$, provided $v$ satisfies a certain condition. Which condition?

d) The weak formulation involves the standard $L^2$-inner product which defines the 2-norm, and what we here call the $B$-norm defined by

$$||u||_B^2 = \int_0^1 xu'(x)^2\, dx = B(u,u).$$

Why does the latter define a norm on

$$V = \{u \in C^1([0,1]) :\, u(1) = 0\}?$$

e) Show that

$$|u|_2 \leq ||u||_B$$

for all $u \in V$. Hint: use the Cauchy-Schwarz inequality for

$$u(x) = -\int_x^1 u'(s)\, ds = -\int_x^1 \frac{1}{\sqrt{s}}\sqrt{s}u'(s)\, ds \quad \text{and} \quad \int_0^1 \ln = -1.$$

f) We can thus take the closure of $V$ with respect to the $B$-norm in $L^2(0,1)$. Denote this closure by $H$. Explain why for all $f \in L^2(0,1)$ there exists a unique $u \in H$ such that

$$\forall_{v\in H} \quad B(u,v) = \int_0^1 fv,$$

and why for $f \in C([0,1])$ this solution is given by (b).

g) Continue from (b) to show that

$$|u'(x)| \leq \frac{|f|_2}{\sqrt{x}},$$

and thereby

$$|u(x) - u(y)| \leq 2|f|_2|\sqrt{x} - \sqrt{y}|$$

for all $x, y \in (0,1]$.

h) Prove that the solution operator $f \to u$ is compact from $C([0,1])$ to itself with respect to the 2-norm. Hint: use (g) to show that $f_n$ bounded in 2-norm implies that the corresponding solutions $u_n$ are uniformly equicontinuous.

i) Prove that the same solution operator is symmetric with respect to the $L^2$-inner product and that there exists a unique sequence of positive eigenvalues $\mu_n$.

j) Show that the power series

$$\sum_{n=0}^{\infty} \frac{(-x)^n}{n!^2}$$

has a countable number of positive zero's. Hint: use (i).

**Exercise 36.9.** Consider the bilinear form

$$B_\lambda(u,v) = \int_0^1 \big(a(x)u'(x)v'(x) + \lambda u(x)v(x)\big)\, dx$$

in relation to the equations studied in Exercise 36.6.

**Exercise 36.10.** We use the closure $H_0^2(0,1)$ of $C_c^4(0,1)$ in the space

$$H^2(0,1) = \{u \in H^1(0,1) : u' \in H^1(0,1)\} \quad \text{with} \quad ||u||_{H^2}^2 = \int_0^1 (u^2 + u'^2 + u''^2)$$

(which defines the $H^2$-norm of $u$), to define weak solutions of $u'''' = f$ with boundary conditions $u = u' = 0$ in $x = 0$ and $x = 1$.

a) Assume that $f \in C([0,1])$. Integrate the equation four times to show that there exists a unique classical solution of this boundary value problem, which we call (BVP).

b) Explain why every classical solution of $u'''' = f$ satisfies

$$\int_0^1 u''v'' = \int_0^1 fv \quad \text{for all} \quad v \in H_0^2(0,1).$$

c) For $f \in L^2(0,1)$ we say that $u \in H_0^2(0,1)$ is a weak solution of (BVP) if (b) holds. Prove that (BVP) has a unique solution for every $f \in L^2(0,1)$.

d) Denote the solution $u$ of (BVP) by $S(f)$. Prove that $\int_0^1 S(f)g = \int_0^1 fS(g)$ for all $f, g \in L^2(0,1)$. We say that $S$ is symmetric with respect to the 2-norm.

e) The solution map $S$ maps $H_0^2(0,1)$ to itself. Explain why $S$ is not symmetric with respect to the $H^2$-norm. Which equivalent inner product norm does make it symmetric?

f) Explain why $S:\ H_0^2(0,1) \to H_0^2(0,1)$ is compact.

g) Give a Rayleigh type formula for the largest eigenvalue of $S$.

h) Give a boundary value problem for a fourth order differential equation for which the solution operator $S:\ H_0^2(0,1) \to H_0^2(0,1)$ is symmetric with respect to the $H^2$-norm.

i) Which boundary conditions should you impose for $u'''' = f$ to get

$$\int_0^1 u''v'' = \int_0^1 fv \quad \text{for all} \quad v \in H^2(0,1)$$

as definition of a weak solution $u \in H^2(0,1)$?

**Exercise 36.11.** Consider the weak formulation in the last item in Exercise 36.10. Under which conditions on $f$ does it allow solutions, and which additional conditions are needed to make the solution unique?

**Exercise 36.12.** Play with (Raleigh quotients for) the bilinear form

$$B(u,v) = \int_0^1 (a(x)u'(x)v'(x) + c(x)u(x)v(x))\,dx,$$

in which $a, c \in C([0,1])$ and $a > 0$ on $[0,1]$.

**Exercise 36.13.** Play with the bilinear form

$$B(u,v) = \int_0^1 (A(x)u''(x)v''(x) + a(x)u'(x)v'(x) + c(x)u(x)v(x))\,dx,$$

in which $A, a, c \in C([0,1])$ and $A > 0$ on $[0,1]$.

**Exercise 36.14.** In Chapter 35 we considered weak solutions $u \in H^1(\Omega)$ of (35.1) using (35.2) with $v \in H_0^1(\Omega)$. As in Exercise 36.5 we say that $u \in H^1(\Omega)$ is weak solution of the homogeneous Neumann problem for

$$Lu = -(a^{ij}u_{x_i})_{x_j} + b^i u_{x_i} + cu = f$$

if (35.2) holds for all $v \in H^1(\Omega)$. What are the (smoothness and) boundary conditions that we have to impose on ($a, b, c, f$, $\partial\Omega$ and) a classical solution $u$ for $u$ to be such a weak solution?

**Exercise 36.15.** The homogeneous Neumann problem

$$Lu = -(a^{ij}u_{x_i})_{x_j} + b^i u_{x_i} + cu = f$$

leads to $B(u, v)$ with $u$ and $v$ in the large space $H^1(\Omega)$. It also comes with the question as to what is needed to have the weak solution $u \in H^2(\Omega)$. See if you can modify the approach in Chapter 35. Can your prove that in the end this $u$ also satisfies the boundary condition you imposed in Exercise 36.14?

**Exercise 36.16.** In Section 35.5 we considered a localised form of

$$Lu = -(a^{ij}u_{x_i})_{x_j} + b^i u_{x_i} + cu = f,$$

derived in Section 35.2 using $\zeta v$ in the weak formulation with the binear form $B(u, v)$. Take $N = 1$ and derive conditions that allow a choice of $\zeta$ that make for $\zeta u$ solving a problem with a symmetric bilinear form. Discuss why in general this does not fly for $N > 1$.

**Exercise 36.17.** Let $\Omega \subset \mathbb{R}^N$ be a bounded domain. Reason as in the above exercise to show that there exists a constant $C_\Omega > 0$ such that

$$\int_\Omega u^2 \leq C_\Omega \int_\Omega |\nabla u|^2$$

for all $u \in H^1(\Omega)$ with $\int_\Omega u = 0$, the Poincaré inequality in

$$\tilde{H}^1(\Omega) = \{u \in H^1(\Omega) : \int_\Omega u = 0\}.$$

**Exercise 36.18.** Exercise 36.17 is an alternative for Section 5.8.1 when $p = 2$. Here's the same result reasoning from contradiction. Let $\Omega \subset \mathbb{R}^{\mathrm{N}}$ be a bounded domain with sufficiently smooth boundary to allow an extension operator as in Theorem 1 of Section 5.4 in Evans. Then the embedding $H^1(\Omega)$ in $L^2(\Omega)$ is compact, and thereby also the embedding of $\tilde{H}^1(\Omega)$ in $L^2(\Omega)$. Prove that there cannot be a sequence $u_n \in \tilde{H}^1(\Omega)$ with $\int_\Omega u_n^2 = 1$ and $\int_\Omega |\nabla u_n|^2 \to 0$. Conclude there exists a constant $C_\Omega$ such that the Poincaré inequality

$$\int_\Omega u^2 \leq C_\Omega \int_\Omega |\nabla u|^2$$

holds for all $u \in \tilde{H}^1(\Omega)$.

# 37 Bio-related stuff

All this relates to what Bio-Bob and I did and are doing with Frank, Bas and the SysBio group.

## 37.1 Cellular chemical networks

We consider cellular networks of reacting metabolites with enzymes catalyzing the chemical reactions[1]. We are interested in the question as to if and how the cell is capable of tuning its enzyme concentrations to maximise certain output flows when the metabolic concentrations are in steady state. The steady state depends both on the enzyme concentrations and the concentrations of external metabolites that take part in the cellular network under consideration.

The simplest of such networks are linear chains

$$X_0 \overset{e_1}{\longleftrightarrow} X_1 \overset{e_2}{\longleftrightarrow} X_2 \overset{e_3}{\longleftrightarrow} \dots \overset{e_N}{\longleftrightarrow} X_N \overset{e_{N+1}}{\longleftrightarrow} X_{N+1}, \tag{37.1}$$

in which $X_0$ and $X_{N+1}$ are external and $X_1, \dots, X_N$ internal metabolites, and the $e_j$, $j \in J = \{1, \dots, N+1\}$, stand for the enzyme concentrations. The dynamics of this linear chain are specified by

$$\dot{x}_i = v_i - v_{i+1} \tag{37.2}$$

for $i = 1, \dots, N$, in which $x_i$ are the concentrations of the (internal) metabolites $X_i$, $i \in I = \{1, \dots, N\}$, and $v_i$ and $v_{i+1}$ are the reaction rates of the reaction that produces product $X_i$ and the reaction that consumes substrate $X_i$. The output flow is $v_{N+1}$ and in steady state it is equal to the input flow:

$$v_{N+1} = \dots = v_1.$$

The reactions are typically modelled by

$$v_j = v_j(e_j, \boldsymbol{x}) = e_j f_j(\boldsymbol{x}), \tag{37.3}$$

in which the metabolic concentrations are grouped in a vector $\boldsymbol{x}$ which includes also the external concentrations. Thus

$$\boldsymbol{x} = (\boldsymbol{x}_E, \boldsymbol{x}_I),$$

[1]This is about joint work with the Systems Biology group at the VU.

with $E$ the index set of external concentrations and $I$ the index set of internal ones. The external concentrations appearing in the in- and output reaction rates are often considered to be prescribed and constant, implying that

$$\dot{x}_i = 0 \tag{37.4}$$

for $i \in E$, with $E = \{0, N+1\}$ in case of (37.1), when the steady state equations for $x_1, \ldots, x_N$ have the external concentrations $x_0$, $x_{N+1}$ and the enzyme concentrations as parameters. Only the ratio[2]

$$e_1 : e_2 : \cdots : e_{N+1}$$

is of direct interest here. Doubling the enzyme concentrations doubles the reaction rates, which has no effect on the steady state, but it does double the steady state flow through the network. Clearly there must be some restriction in the model to have the maximization problem make sense.

For more general networks models like

$$\dot{x}_i = \sum_{j \in J} N_{ij} v_j(e_j, \boldsymbol{x}) \quad (i \in I) \tag{37.5}$$

are used, in which $N$ is a stoichiometry matrix. In the case of the linear chain (37.1) this matrix has entries

$$N_{ii} = 1, \quad N_{ij} = -1 \quad \text{for} \quad j = i+1, \quad N_{ij} = 0$$

for $j < i$ and $j > i+1$. The reaction functions $f_j$ may depend only on the substrates and products of the corresponding reaction, so

$$f_j(\boldsymbol{x}) = f_j(x_{j-1}, x_j) \tag{37.6}$$

in case of (37.1), or on other metabolic concentrations, for instance if metabolites can form complexes with enzymes catalyzing reactions in which they do not appear as substrate or product. These functions $f_j(\boldsymbol{x})$ are often referred to as the saturation levels of the corresponding enzyme. In Michaelis-Menten kinetics they are derived from mass action kinetics involving different time scales.

## 37.2 Michaelis-Menten kinetics

Enzyme reaction rates are derived from mass action kinetics for reaction blocks which in the simplest case are of the form

$$S + E \rightleftarrows ES \rightleftarrows EP \rightleftarrows P + E, \tag{37.7}$$

[2]Think of projective coordinates.

in which $S$ and $P$ are substrate and product of a reversible reaction

$$S \overset{e}{\longleftrightarrow} P$$

catalyzed by enzyme $E$, and $ES$ and $EP$ are complexes of the enzyme with the substrate $S$ and the product $P$. In a linear chain every link

$$X_i \overset{e_i}{\longleftrightarrow} X_{i+1}$$

is of this form.

Denoting concentrations by

$$s = [S],\ p = [P],\ c_0 = [E] = [C_0],\ c_1 = [ES] = [C_1],\ c_2 = [EP] = [C_2],$$

mass action kinetics for each of the six arrows in (37.7) gives a coupled system of differential equations

$$\frac{ds}{dt} = -k_1 s c_0 + k_2 c_1; \quad \frac{dc_0}{dt} = -k_1 s c_0 + k_2 c_1 + k_5 c_2 - k_6 p c_0;$$

$$\frac{dc_1}{dt} = k_1 s c_0 - k_2 c_1 - k_3 c_1 + k_4 c_2;$$

$$\frac{dc_2}{dt} = k_3 c_1 - k_4 c_2 - k_5 c_2 + k_6 p c_0; \quad \frac{dp}{dt} = k_5 c_2 - k_6 p c_0,$$

in which we have used $k_{1,3,5}$ to denote the reaction constants of the forward and $k_{2,4,6}$ for the backward reactions in (37.7). The constant $k_1$ corresponds to the rate of $S$ and $E$ binding, the constant $k_2$ to the rate of the complex $ES$ unbinding, and likewise for $k_6$ and $k_5$. The constant $k_3$ and $k_4$ correspond to the rate of the complex $ES$ turning into the complex $EP$ and vice versa[3].

The right hand sides of these equations are linear in $c_0, c_1, c_2$, so we can write

$$\begin{pmatrix} \dot{s} \\ \dot{c}_0 \\ \dot{c}_1 \\ \dot{c}_2 \\ \dot{p} \end{pmatrix} = \begin{pmatrix} -k_1 s & k_2 & 0 \\ -k_1 s - k_6 p & k_2 & k_5 \\ k_1 s & -k_2 - k_3 & k_4 \\ k_6 p & k_3 & -k_4 - k_5 \\ -k_6 p & 0 & k_5 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix},$$

in which we recognise that the total enzyme concentration

$$c_0 + c_1 + c_2 = e_{tot} = \varepsilon$$

[3]Ignore smaller molecular groups consumed or produced in $ES \rightleftarrows EP$?

is constant[4], a positive constant denoted by $\varepsilon$ and assumed to be small in what follows.

The system is of the form

$$\begin{aligned}\dot{\boldsymbol{x}} &= A(\boldsymbol{x})\boldsymbol{c} \\ \dot{\boldsymbol{c}} &= B(\boldsymbol{x})\boldsymbol{c}\end{aligned} \quad \text{for} \quad \boldsymbol{x} = \begin{pmatrix} s \\ p \end{pmatrix} \quad \text{and} \quad \boldsymbol{c} = \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix},$$

with $A(\boldsymbol{x})$ and $B(\boldsymbol{x})$ matrices depending on $\boldsymbol{x}$. Assuming the enzyme concentrations to be small compared to the concentrations of $S$ and $P$ we scale the free and bound enzyme concentrations with $\varepsilon$ and set

$$\boldsymbol{c} = \varepsilon\boldsymbol{\gamma}.$$

This shows that a splitting of time-scales appears when $\varepsilon$ is small because

$$\dot{\boldsymbol{x}} = \varepsilon A(\boldsymbol{x})\boldsymbol{\gamma}; \quad \dot{\boldsymbol{\gamma}} = B(\boldsymbol{x})\boldsymbol{\gamma}.$$

Introducing a new time variable $\tau = \varepsilon t$, we write

$$\dot{\boldsymbol{x}} = \frac{d\boldsymbol{x}}{dt} = \varepsilon\frac{d\boldsymbol{x}}{d\tau} = \varepsilon\boldsymbol{x}', \quad \dot{\boldsymbol{\gamma}} = \frac{d\boldsymbol{\gamma}}{dt} = \varepsilon\frac{d\boldsymbol{\gamma}}{d\tau} = \varepsilon\boldsymbol{\gamma}'$$

and conclude that

$$\boldsymbol{x}' = A(\boldsymbol{x})\boldsymbol{\gamma}, \, \varepsilon\boldsymbol{\gamma}' = B(\boldsymbol{x})\boldsymbol{\gamma}.$$

For $\varepsilon = 0$ this reduces to

$$\boldsymbol{x}' = A(\boldsymbol{x})\boldsymbol{\gamma} \quad \text{with} \quad B(\boldsymbol{x})\boldsymbol{\gamma} = 0 \quad \text{and} \quad \gamma_0 + \gamma_1 + \gamma_2 = 1,$$

and leads[5] to

$$\dot{p} = \frac{\varepsilon(k_1k_3k_5s - k_2k_4k_6p)}{k_2k_4 + k_2k_5 + k_3k_5 + k_1(k_3 + k_4 + k_5)s + (k_2 + k_3 + k_4)k_6p} \tag{37.8}$$

as the modelling equation for

$$S \overset{e}{\longleftrightarrow} P$$

**Exercise 37.1.** Explain why (37.8) would be plausible. What do you get for $\dot{s}$ via the same reasoning? Hint: the rank of the matrix $B(x)$ is 2 and its kernel is given by

$$k_2k_4 + k_2k_5 + k_3k_5 : k_1s(k_4 + k_5) + k_4k_6p : k_1sk_3 + (k_2 + k_3)k_6p$$

in projective coordinates, which sum up to

$$k_2k_4 + k_2k_5 + k_3k_5 + k_1s(k_3 + k_4 + k_5) + (k_2 + k_3 + k_4)k_6p.$$

[4]We also have that $\dot{s} + \dot{c}_1 + \dot{c}_2 + \dot{p} = 0$.

[5]You can wonder if there's a theorem to support this reduction.

**Exercise 37.2.** Explain[6] why

$$\frac{k_2k_4 + k_2k_5 + k_3k_5}{k_1 s(k_3 + k_4 + k_5) + (k_2 + k_3 + k_4)k_6 p}$$

is the ratio between free enzyme and bound enzyme.

The resulting differential equation (37.8) is of the form

$$\dot{p} = \frac{\varepsilon(k_{135}s - k_{246}p)}{k_{2345} + k_{1345}s + k_{2346}p}, \tag{37.9}$$

but often[7] written as

$$\dot{p} = \frac{\frac{V^+}{K_s}(s - \frac{p}{K_{eq}})}{1 + \frac{s}{K_s} + \frac{p}{K_p}},$$

in which $K_{eq}$ is the value at which $S$ and $P$ are in thermodynamic equilibrium and therefore called the equilibrium constant. Note that this equilibrium is a constant, in the sense that it does not depend on the amount of enzyme invested in the reaction. Enzymes lower the threshold (chemical potential) necessary to drive the reaction but the equilibrium (the tipping point when there is enough substrate relative to product to get a net positive reaction rate) remains always the same.

**Exercise 37.3.** Verify that

$$K_s = \frac{k_{2345}}{k_{1345}} = \frac{k_2k_4 + k_2k_5 + k_3k_5}{k_1(k_3 + k_4 + k_5)}, \quad K_p = \frac{k_{2345}}{k_{2346}} = \frac{k_2k_4 + k_2k_5 + k_3k_5}{(k_2 + k_3 + k_4)k_6},$$

$$K_{eq} = \frac{k_{135}}{k_{246}} = \frac{k_1k_3k_5}{k_2k_4k_6}, \quad V^+ = \frac{k_{135}}{k_{1345}} = \frac{\varepsilon k_3k_5}{k_3 + k_4 + k_5}.$$

**Exercise 37.4.** Examine (37.9), $K_{eq}$, $K_s$ and $K_p$ when $k_1, k_2 \to \infty$ with

$$\frac{k_1}{k_2} = \kappa_s$$

fixed. Same question for $k_5, k_6 \to \infty$ with

$$\frac{k_6}{k_5} = \kappa_p$$

fixed.

[6] Does this relate to the term saturation level?

[7] See Appendix 1 of Teusink's FEBS 2000 paper for examples from yeast glycolysis.

### 37.2.1 Directed graphs and trees

We put the derivation of (37.8) in a graph theoretic framework which should help you to derive similar ODE's for more complicated[8] reactions. We change the notation in the differential system and write it as

$$\begin{pmatrix} \dot{s} \\ \dot{c}_0 \\ \dot{c}_1 \\ \dot{c}_2 \\ \dot{p} \end{pmatrix} = \begin{pmatrix} -k_{10}s & k_{01} & 0 \\ -k_{10}s - k_{20}p & k_{01} & k_{02} \\ k_{10}s & -k_{01} - k_{21} & k_{12} \\ k_{20}p & k_{21} & -k_{02} - k_{12} \\ -k_{20}p & 0 & k_{02} \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix}, \tag{37.10}$$

to exhibit the graph structure in which $C_0, C_1, C_2$ are the nodes. The constants $k_{ij}$ systematically correspond to the reactions

$$C_i \leftarrow C_j$$

and the graph has precisely one link between every pair of nodes. Every link

$$C_i - C_j$$

comes with constants $k_{ij}$ and $k_{ji}$. The additional structure in this particular example is that the constant $k_{10}$ multiplies $s$ and the constant $k_{20}$ multiplies $p$.

If we drop $s$ and $p$ from the notation we get the square matrix

$$K_2 = \begin{pmatrix} -k_{10} - k_{20} & k_{01} & k_{02} \\ k_{10} & -k_{01} - k_{21} & k_{12} \\ k_{20} & k_{21} & -k_{02} - k_{12} \end{pmatrix} \tag{37.11}$$

which labels every arrow in the directed graph[9]

$$C_0 \rightleftarrows C_1 \rightleftarrows C_2 \rightleftarrows C_0 \tag{37.12}$$

accordingly.

**Exercise 37.5.** Determine the null space of $K_2$. Show that it is spanned by a vector in which every entry is the sum of three terms, each of which is the product of two different $k_{ij}$. The first[10] entry contains the term $k_{01}k_{02}$ which we can view as corresponding to the subgraph

$$C_1 \rightarrow C_0 \leftarrow C_2$$

of (37.12). List the graphs corresponding to all nine terms.

[8] It was already complicated....

[9] Drawn with two copies of $C_0$ for linear convenience.

[10] We number the 3 entries $0, 1, 2$.

**Exercise 37.6.** Re-examine (37.7) and observe it corresponds to the full matrix $K_2$. In projective form the coordinates of the null vector of $K_2$ are

$$k_{01}k_{12} + k_{02}k_{21} + k_{01}k_{02} : k_{10}k_{02} + k_{12}k_{20} + k_{10}k_{12} : k_{20}k_{01} + k_{21}k_{10} + k_{20}k_{21}$$

1. Multiply $k_{10}$ by $s$ and $k_{20}$ by $p$ to get a solution $(c_0, c_1, c_2)$ of $\dot{c}_0 = \dot{c}_1 = \dot{c}_2 = 0$ for $s$ and $p$ fixed and divide by $c_0+c_1+c_2$ to get a solution with $c_0+c_1+c_2 = 1$.
2. Multiply this solution by $\varepsilon$ and substitute it in
$$\dot{s} = -k_{01}sc_0 + k_{01}c_1; \quad \dot{p} = k_{02}c_2 - k_{20}pc_0,$$
to get an equation for $\dot{s}$ and $\dot{p}$.
3. The numerator in the resulting expression for $\dot{p}$ is the product of $\varepsilon$ and the difference of two terms, each of which corresponds to a directed subgraph with three arrows. Which subgraphs?

**Remark 37.7.** *What you found is (37.8) with the constants*

$$k_1, k_2, k_3, k_4, k_5, k_6$$

*replaced by*

$$k_{10}, k_{01}, k_{21}, k_{12}, k_{02}, k_{20},$$

*and every product of k's can be seen as corresponding to a subgraph.*

**Exercise 37.8.** From (37.11) it is clear what the matrix $K_3$ should be. Use a computer algebra package to find the null vector $K_3$ in a form similar to what you found in Exercise 37.8. The first entry contains the term $k_{01}k_{02}k_{23}$ which we can view as corresponding to the subgraph

$$C_1 \to C_0 \leftarrow C_2 \leftarrow C_3$$

of the complete directed graph with nodes $C_0, C_1, C_2, C_3$. This first entry is the sum of 16 such terms. Draw the 16 corresponding directed graphs[11].

**Exercise 37.9.** The directed graphs corresponding to $k_{01}k_{02}$ in Exercise 37.5 and $k_{01}k_{02}k_{23}$ in Exercise 37.8 may be seen as trees rooted in $C_0$. Verify that the terms in the first entry of the null vector of $K_3$ exhibit all trees rooted in $C_0$. In which entries do the trees rooted in $C_1$ appear? How do these differ from the trees rooted in $C_0$? Same question for $C_2$ and $C_3$.

[11] More complicated reactions have more complexes but typically sparser matrices.

**Exercise 37.10.** State a theorem that generalises what you just found to $n = 4, 5, \ldots$ and see if you can prove[12] it.

**Remark 37.11.** *The rooted trees become undirected trees if we undirect the arrows of their directed graphs. The special role of the root then disappears. Each entry of the null vector then generates all trees with nodes $C_0, \ldots, C_n$. Computer algebra packages will allow you to guess a formula for the total number of such trees from brute force calculation of the kernel of $K_n$ for $n = 3, 4, 5$. Prüfer codes provide a proof*[13].

### 37.2.2 More complicated reactions

Consider

$$A + E \rightleftarrows EA$$

$$B + E \rightleftarrows EB$$

$$B + EA \rightleftarrows EAB \rightleftarrows A + EB$$

$$EAB \rightleftarrows P + E$$

as a model for

$$A + B \overset{e}{\longleftrightarrow} P \tag{37.13}$$

and introduce

$$a = [A],\ b = [B],\ p = [P],$$

$$c_0 = [E] = [C_0],\ c_1 = [EA] = [C_1],\ c_2 = [EB] = [C_2],\ c_3 = [EAB] = [C_3]$$

as the concentrations. In this model[14]

$$EAB \rightleftarrows P + E$$

is a lumped simplification of

$$EAB \rightleftarrows EP \rightleftarrows P + E.$$

This model has a graph in which every pair of nodes $C_i$ and $C_j$ is linked, except for the $C_1$ and $C_2$. Its matrix $K_3$ has $k_{12} = k_{21} = 0$.

[12] Using the graph interpretation rather than linear algebra.

[13] Look elsewhere for this proof.

[14] The analogous model for $S \overset{e}{\longleftrightarrow} P$ is examined in Exercise 37.21.

**Exercise 37.12.** We write the null vector projectively as

$$k_{01}k_{02}k_{03} + \cdots : \cdots : \cdots : k_{30}k_{31}k_{32} + \cdots .$$

The constant $k_{10}$ has to be multiplied by $a$, $k_{20}$ by $b$, and $k_{30}$ by $p$, but first you should determine all terms in the null vector from the graph structure. Do so, and then write the equation for $\dot{p}$ like you did in Exercise 37.6, using (the reactions in) the link

$$C_3 \rightleftarrows C_0$$

and the null vector of $K_3$ with $k_{12} = k_{21} = 0$ and $a, b, p$ included in the appropriate coefficients. The denominator will be big and also appears in the derivatives $\dot{a}$ and $\dot{b}$. Which links do you need for $\dot{a}$? Note the symmetry in $a$ and $b$. How do $\dot{a}$ and $\dot{p}$ relate? And $\dot{a}$ and $\dot{b}$?

**Exercise 37.13.** We say that (37.13) is in steady state if $\dot{a} = \dot{b} = \dot{p} = 0$. This reduces to one single equation for $a, b, p$. Write this equation and examine its solution set.

**Exercise 37.14.** (continued). The oriented closed loop $0 \to 1 \to 3 \to 2 \to 0$ comes with $4$ reaction coefficients and so does $0 \to 2 \to 3 \to 1 \to 0$. Is there an a priori constraint on these reaction coefficients? Two hints: what properties should the solution set in Exercise 37.13) have? What about the reaction constants in the closed loops $0 \to 1 \to 3 \to 0$ and $0 \to 3 \to 1 \to 0$? Think of Escher's eternal climbing or descending staircase.

**Exercise 37.15.** (continued). The constraint in Exercise 37.14 causes the numerators to factorise. Relate the terms in the factors to subgraphs.

**Exercise 37.16.** Modify the above model for (37.13) with an extra reaction step

$$EAB \rightleftarrows EP \rightleftarrows P + E$$

and do the same analysis as in the previous exercises. Hint: you now have $5$ nodes.

**Exercise 37.17.** Model the reaction

$$A + B \overset{e}{\longleftrightarrow} P + Q$$

in a way which is similar to Exercise 37.16 and symmetric in substrates and products. How many nodes? There are now two loops with constraints. Examine the results in relation to the graphs.

**Exercise 37.18.** Refering to Exercise 37.4, see how the above complicated formulas simplify if you take similar limits for all reactions involving $E$.

**Exercise 37.19.** See how the complicated formulas for (37.13) simplify if you take out one of the two reaction paths from $EAB$ to $E$.

**Exercise 37.20.** Examine the extremely complicated equation for

$$A + B \overset{e}{\longleftrightarrow} P + Q$$

in Exercise 37.17. How does it simplify under modifications as in the previous two exercises?

**Exercise 37.21.** Back to $S \overset{e}{\longleftrightarrow} P$. Compare (37.7) to the simplified model

$$S + E \rightleftarrows ES \rightleftarrows P + E, \tag{37.14}$$

which has the graph

$$C_0 = C_1$$

in which there are two links between the nodes $C_0$ and $C_1$, one for $S + E \rightleftarrows ES$ and one for $ES \rightleftarrows P + E$, with constants

$$k_{01}^s, k_{10}^s, k_{01}^p, k_{10}^p.$$

Verify that the method with rooted trees leads to

$$\dot{p} = \frac{k_{01}^p k_{10}^s s - k_{10}^p k_{01}^s p}{k_{01}^s + k_{01}^p + k_{10}^s s + k_{10}^p p},$$

and relate this to taking $k_{21}$ and $k_{12}$ in the system (37.10) for (37.7) large but of the same order.

**Exercise 37.22.** Derive the simplest model for $A + B \overset{e}{\longleftrightarrow} P + Q$ which still has the two loops in its graph.

## 37.3 Linear chains

We go back to (37.1),

$$X_0 \overset{e_1}{\longleftrightarrow} X_1 \overset{e_2}{\longleftrightarrow} X_2 \overset{e_3}{\longleftrightarrow} \dots \overset{e_N}{\longleftrightarrow} X_N \overset{e_{N+1}}{\longleftrightarrow} X_{N+1},$$

in which every link is a reaction like (37.7) modelled with (37.6) derived as in Section 37.2 and therefore a quotient of the form (37.9). This leads to some projects which we formulate more open than the modelling of the individual reactions above.

### 37.3.1 Steady states

Given the enzyme concentrations $e_i$ and the exterior concentrations $x_0$ and $x_{N+1}$, the system of ODE's derived starting from (37.2) is bound to have a steady state. Show that there is a unique steady state. Hint: use monotonicity properties of the functions $f_i$. Distinguish between

$$v_{N+1} = \cdots = v_1 > 0 \quad \text{and} \quad v_{N+1} = \cdots = v_1 < 0,$$

and explain why this distinction does not depend on the enzyme concentrations but only on $x_0$ and $x_{N+1}$. How? Play with interchanging the roles of solutions $x_1, \dots, x_N$ and parameters $e_1, \dots, e_{N+1}$. Which values of the concentrations allow a choice of enzyme concentrations that make them steady states?

### 37.3.2 Global behaviour of nonsteady state solutions

Consider the ODE system for given fixed enzyme concentrations $e_i$ and exterior concentrations $x_0$ and $x_{N+1}$, and initial concentrations at time $t = 0$ for $x_1, \dots, x_N$. Try to show that this system of ODE's is globally stable. Hints: take $N = 2$ to get started and some ideas; use the monotonicity properties of the reaction functions $f_i(x_{i-1}, x_i)$ to derive monotonicity properties for the maximum and the minimum of $x_i(t)$ over $i = 1, \dots, N$ and order-preserving properties between solutions with different initial data; maybe derive ODE's for $v_i$.

### 37.3.3 Optimisation problems

Largely independent of the previous two sections. Given exterior concentrations $x_0$ and $x_{N+1}$ such that steady states have

$$v_{N+1} = \cdots = v_1 = J > 0,$$

examine the problem of finding steady states which maximize $J$ varying the positive enzyme concentrations under the constraint that

$$e_0 + \cdots e_{N+1} = e_T > 0.$$

Hint: how does this problem depend on $e_T$? Reformulate the problem as a problem in which the enzyme concentrations appear only after solving the problem.

### 37.3.4 Self-steering networks

Suppose the cell is able to tune its enzyme concentration based on the internal metabolic concentrations. How would it choose its enzyme concentrations to steer towards an optimal steady state as in Section 37.3.3?

### 37.3.5 Inhibition

The reactions in (37.1) may be more complicated than (37.7) if other metabolites in the chain also bind to the enzyme that catalyzes the reaction

$$X_{i-1} \rightleftarrows X_i$$

for some of the single substrate single product reaction in the chain. Explore how the reaction rates derived for (37.7) have to be modified, and how this affects what we did above.

## 37.4 General networks

The linear chains come with an ODE system (37.5) which is special for another reason: the kernel of $N$ is one-dimensional. Up to a multiple, only one vector of reaction rates is allowed in steady state, and all entries of the vector are nonzero. The chain may be part of a larger network in which all the other reactions have zero reaction rate, due to the corresponding enzyme concentrations being zero. The set (or vector) of nonzero reaction rates is called a flux mode. The linear chain itself only has one flux mode.

### 37.4.1 Networks with one flux mode and one output

Cook up (small) nonlinear networks with one flux mode and one output in which more complicated reactions as in Section 37.2.2 occur and see what you can do concerning the issues in the subsections of Section 37.3 if the output flow to is to be maximised.

### 37.4.2 Networks with more flux modes and one output

What are the questions to be asked and how would you proceed for the optimization problem?

## 37.5 Stable polynomials

You may have seen that we often need to solve equations of the form

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

with real coefficients $a_0, a_1 \ldots, a_{n-1}, a_n = 1$, and that we want to know if the solutions all have negative real part.

Write

$$f(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0 = p(z^2) + zq(z^2)$$

and introduce

$$P(z) = p(-z^2), \quad Q(z) = zq(-z^2).$$

For example, if

$$f(z) = 1 + z + z^2 + z^3 + z^4 + z^5,$$

then

$$P(z) = 1 - z^2 + z^4, \quad Q(z) = z - z^3 + z^5.$$

Assume that $f(z)$ factorizes as

$$f(z) = (z - z_1)(z - z_2)\cdots(z - z_n)$$

(this is always true), in which all $z_j$ have negative real part. In the course I showed that this implies that

$$\text{Im}(Q(z)\overline{P(z)}) > 0 \quad \text{if} \quad \text{Im}z > 0.$$

For nonreal $z$ this implies that $P(z) \neq 0 \neq Q(z)$ so that we can write (without worrying about dividing by zero!)

$$\frac{Q(z)}{P(z)} = \frac{Q(z)\overline{P(z)}}{|P(z)|^2} \quad \text{and} \quad \frac{P(z)}{Q(z)} = \frac{P(z)\overline{Q(z)}}{|Q(z)|^2},$$

and conclude that for Im$z > 0$ the first quotient has positive imaginairy part and the second negative imaginairy part.

We also see that $P(z)$ and $Q(z)$ have only real zero's. Except for $z = 0$ for $Q(z)$, these appear in pairs $z = \pm\zeta$ with $\zeta > 0$, and we cannot have $P(\zeta) = Q(\zeta) = 0$, because in such a case $z = i\zeta$ is a zero of both $p(z^2)$ and $q(z^2)$, and therefore of $f(z)$. This would contradict the assumption that all zero's of $f(z)$ have negative real part.

Moreover, we can write for any such $\zeta$, e.g. with $P(\zeta) = 0$, using long division (in dutch: staartdeling) that

$$P(z) = (z - \zeta)R(z),$$

with $R(z)$ a polynomial with real coefficients. We say that $z = \zeta$ is a simple zero of $P(z)$ if $R(\zeta) \neq 0$. If $\zeta$ is not a simple zero you can divide out more factors $z - \zeta$ and obtain that

$$P(z) = (z - \zeta)^k R_k(z)$$

with $k > 1$ and $R_k(\zeta) \neq 0$, so that in the end

$$\frac{P(z)}{Q(z)} = (z - \zeta)^k \frac{R_k(z)}{Q(z)} = (z - \zeta)^k S(z),$$

with $S(\zeta) \neq 0$. You should be able to see that this cannot be true if the imaginary part of this expression is negative close to $\zeta$ in the upper half plane.

Thus

$$P(z) = (z - \zeta)R(z) \quad \text{with} \quad R(\zeta) = P'(\zeta) \neq 0$$

and the quotient is of the form

$$\frac{P(z)}{Q(z)} = (z - \zeta)S(z) \quad \text{with} \quad S(\zeta) = \frac{P'(\zeta)}{Q(\zeta)} \neq 0.$$

The sign of the last quotient tells you the sign of the imaginary part of this expression for $z$ a little above $\zeta$. Conclude that

$$P(\zeta) = 0 \quad \Longrightarrow \quad \frac{P'(\zeta)}{Q(\zeta)} < 0,$$

and, by the same reasoning,

$$Q(\zeta) = 0 \quad \Longrightarrow \quad \frac{Q'(\zeta)}{P(\zeta)} > 0.$$

It follows that the zeros of $P(z)$ and $Q(z)$ are interlaced!

A remarkable consequence is also that all coefficients of $f(z)$ have the same sign and are positive:

$$a_0 > 0,\ a_1 > 0,\ \ldots, a_n > 0,\ a_{n-1} = 1 > 0.$$

To see this consider for instance $p(-z^2)$, replace $z^2$ by $w$ and write

$$p(w) = a_0 - a_2 w + a_4 w^2 - \ldots.$$

All zero's of $p(w)$ are real, positive and simple. Convince yourself that polynomials with this property have alternating coefficients, just like the series for the $\cos x$ and $\sin x$ that I wrote on the board.

The final conclusion is that

- all the coefficients in $f(z)$ have the same sign;
- the zero's of $P(z)$ and $Q(z)$ are all real, simple, and interlaced.

It is another nice exercise to show that if these the two properties hold, all zero's of $f(z)$ have negative real part.

## 37.6 Hurwitz, rough stuff

We give an analytic proof, the origin of which we do not know. We found it in the book of Hurwitz, chapter 3, §9, see also the paper of Velleman. The proof uses the following fact from the theory of Laurent series.

**Theorem 37.23.** *[Cauchy's estimate] Let $q(z) = \sum_{n=-\infty}^{\infty} a_n z^n$ be a Laurent series, converging for $r_1 < |z| < r_2$. Let $\rho$ be a real number between $r_1$ and $r_2$. Let $M := \max_{|z|=\rho} |q(z)|$. Then for all $n \in \mathbb{Z}$ we have $|a_n|\rho^n \leq M$.*

*Proof.* We first proof this for $n = 0$. Let $\varepsilon > 0$. The Laurent series is uniformally convergent on the compact circle $\{|z| = \rho\}$, we can find an $n \in_N$ with $\left|q(z) - \sum_{k=-n}^{n} a_k z^k\right| < \varepsilon$ for all $z \in \mathbb{C}$ with $|z| = \rho$. Write $f(z) := a_0 + \sum_{k=-n}^{n}{}' a_k z^k$, where the prime means that we do not take the summand for $k = 0$. It follows that $|f(z)| < M + \varepsilon$ for all $z$ with $|z| = \rho$. Let $\xi$ be a complex number with $|\xi| = 1$ and $\xi^k \neq 1$ for all $k \in_N$. Example: $\xi = (2-i)/(2+i)$. If $\xi^k = 1$ for some $k$, a simple calculation shows that $(2i)^k = (2-i)(A+Bi)$ for some $A, B \in \mathbb{Z}$. Then $4^k = 5(A^2 + B^2)$, which is impossible. Put $z_i := \xi^i \rho$. Then by using the geometric series:

$$\frac{f(z_0) + \ldots + f(z_{s-1})}{s} = a_0 + \frac{1}{s}\sum_{k=-n}^{n}{}' a_k \rho^k \frac{\xi^{ks} - 1}{\xi^k - 1}$$

Notice that $|\xi^{ks}-1| \leq 2$ and with $\lambda := \sum_{k=-n}^{n}{}' \left|\frac{a_k\rho^k}{\xi^k-1}\right|$ we have

$$\left|\sum_{k=-n}^{n}{}' \frac{a_k\rho^k}{\xi^k-1}\right| \leq 2\lambda,$$

and that $\lambda$ does not depend on $s$. Hence

$$|a_0| \leq \frac{|f(z_0)| + \ldots + |f(z_{s-1})|}{s} + \frac{2\lambda}{s}.$$

This holds for all $s$, hence $|a_0| \leq M + \varepsilon$. This holds for all $\varepsilon > 0$, hence $|a_0| \leq M$.

If $n$ is arbitrary, apply the result just proved to the Laurent series $z^{-n}q(z)$, whose maximum for $|z| = \rho$ is equal to $\rho^{-n}M$. □ □

**Theorem 37.24.** *Suppose $f(z) = \sum_{n=1}^{\infty} a_n z^n$ with $a_n \in \mathbb{C}$ is a power series with convergence radius $R > 0$. Then there exist an $a \in \mathbb{C}$, $|a| = R$ which is a singularity for $f$.*

*Proof.* Suppose the Theorem is wrong. This means, that for all $a$ with $|a| = R$ there exists a local function $f(z, a)$ around $a$ extending $f(z)$. By the monodromy Theorem, we get an analytic function $f(z)$ which is well defined on $|z| < R + \sigma$ for some $\sigma > 0$.[1] Let $M$ be the maximum of $|f(z)|$ on $|z| \leq R + \sigma$. Let $p \in \mathbb{C}$ with $|p| < R$. We develop $f(z)$ around $p$:

$$f(z) = \sum_{k=0}^{\infty} f^{(k)}(p)\frac{(z-p)^k}{k!}. + \ldots$$

By Lemma 37.23, it follows $|f^{(k)}(p)/k!|\sigma^k \leq M$, that ist

$$|f^{(k)}(p)| \leq \frac{M \cdot k!}{\sigma^k}.$$

However, also

$$f^{(k)}(p) = \sum_{n=k}^{\infty} n(n-1) \cdot \ldots \cdot (n-k+1)a_n p^{n-k}.$$

We again can apply Lemma 37.23, supposing $p$ is on a circle of radius $\alpha < R$ with center 0 to conclude:

$$n(n-1) \cdot \ldots \cdot (n-k+1) \cdot |a_n| \cdot \alpha^{n-k} \leq \frac{M \cdot k!}{\sigma^k}.$$

[1] Here we do not need the monodromy theorem, as the extension is given by $1/p(z)$.

This holds for all $\alpha < R$, so

$$\binom{n}{k}|a_n|R^{n-k}\sigma^k \leq M$$

This holds for $k = 0, \ldots, n$. Adding the results and by the binomial formula: $|a_n|(R+\sigma)^n \leq (n+1)M.$, it follows that for all $z \in \mathbb{C}$:

$$|a_n z^n| \leq M \cdot (n+1)\left(\frac{|z|}{R+\sigma}\right)^n$$

It follows that the convergence radius of $\sum_{n=0}^{\infty} a_n z^n$ is at least $R + \sigma > R$, which is a contradiction. $\square$ $\square$

From this we give a proof of the Fundamental Theorem of Algebra.

*Proof.* Suppose the polynomial $p(z) = z^n + a_{n-1} + \ldots + a_0$ does not have a zero. Then the power series

$$\frac{1}{p(z)} = b_0 + b_1 z + b_2 z^2 + \cdots$$

is absolute convergent for all $z \in \mathbb{C}$. We will show that there exist $c, r > 0$ such that there exist infinitely many $k$ with $|b_k|r^k > c > 0$, so that the above series is not convergent at all, contradiction. To prove this claim, consider

$$1 = (a_0 + a_1 z + \ldots + a_{n-1}z^{n-1} + z^n) \cdot (b_0 + b_1 z + b_2 z^2 + \cdots)$$

given for all $k \geq 0$:

$$a_0 b_{k+n} + a_1 b_{k+n-1} + \ldots + a_{n-1}b_{k+1} + b_k = 0.$$

Notice that $b_0 = 1/p(0) \neq 0$. Choose $c$ with $0 < c < |b_0|$ and choose $r > 0$ so small that

$$|a_0|r^n + |a_1|r^{n-1} + \ldots + |a_{n-1}r \leq 1.$$

(we can do this by continuity of the left hand side.) Obviously $|b_0| > cr^0$. Given $k$ with $|b_k| > cr^k$. We will show that there exist an $1 \leq i \leq n$ with $|b_{k_i}| > cr^{k+i}$, which would prove the claim. Suppose not. Then

$$\begin{array}{rcl} |b_k| & = & |a_0 b_{k+n} + a_1 b_{k+n-1} + \ldots + a_{n-1}b_{k+1}| \\ & \leq & |a_0|cr^{k+n} + |a_1|cr^{k+n-1} + \ldots + |a_{n-1}|r \\ & \leq cr^k & \end{array}$$

contradiction. $\square$

# 38 The Navier-Stokes equations

Read about the Navier-Stokes equations in Dutch on a very introductionary and informal level in `http://www.math.vu.nl/~jhulshof/handoutNS.pdf`. Next we consider these 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.

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 $(\mathbf{N})$ 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 $(\mathbf{N})$? 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 $(\mathbf{N})$ 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 $\Omega$ 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 $\Omega$ to derive that

$$\frac{1}{2}\frac{d}{dt}\int_\Omega |w|^2 + \int_\Omega |Dw|^2 = -\int_\Omega w\cdot(w\cdot\nabla)v \le \int_\Omega |Dv|\,|w|^2.$$

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 $Ax\cdot x \le |A|\,|x|^2$ for $2\times 2$ 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

$$\frac{d}{dt}\int_\Omega |w|^2 + 2\int_\Omega |Dw|^2 \le 2\Big(\int_\Omega |Dv|^2\Big)^{\frac{1}{2}}\Big(\int_\Omega |w|^4\Big)^{\frac{1}{2}}.$$

(d) Insert the inequality

$$\int_\Omega |w|^4 \le \int_\Omega |w|^2 \int_\Omega |Dw|^2$$

in (c) to derive that

$$\frac{d}{dt}\int_\Omega |w|^2 \le \frac{1}{2}\int_\Omega |Dv|^2 \int_\Omega |w|^2.$$

Hint: in the right hand side you get a product which contains the factor $a = \int_\Omega |Dw|^2$. Use the inequality $2ab \le a^2 + b^2$ and observe that $a^2$ als appears on the left hand side.

(e) Derive from (d) and (a) with $u$ replaced by $v$ that

$$\int_\Omega |w(t)|^2 \le \int_\Omega |w_0|^2 e^{\frac{1}{4}\int_\Omega |v_0|^2}.$$

(f) Prove the inequality used in (d) for compactly supported smooth vectorfields on $\mathbb{R}^2$ by first showing that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x_1,x_2)^4\,dx_1dx_2 \le$$

$$\Big(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u^2\Big)\Big(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u_{x_1}^2\Big)^{\frac{1}{2}}\Big(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u_{x_2}^2\Big)^{\frac{1}{2}}$$

for compactly supported smooth functions $u$. In short

$$|u|_4^4 \le |u|_2^2 \, |u_{x_1}|_2 \, |u_{x_2}|_2.$$

Hint: write $u(x_1, x_2)^4 = u(x_1, x_2)^2 u(x_1, x_2)^2$ and show first that

$$u(x_1, x_2)^2 \le \int_{-\infty}^{\infty} u(\xi, x_2) u_{x_1}(\xi, x_2) d\xi$$

and likewise

$$u(x_1, x_2)^2 \le \int_{-\infty}^{\infty} u(x_1, \eta) u_{x_2}(x_1, \eta) d\eta.$$

# 39 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

$$I(u) = \int_a^b L(t, u(t), \dot{u}(t))\, dt \tag{39.1}$$

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:

$$\frac{d}{dt}\frac{\partial L}{\partial p^i} - \frac{\partial L}{\partial u^i} = 0 \quad (i = 1, \ldots, n) \tag{39.2}$$

We also saw the Jacobi equations, obtained from (1.3.6) and the linearised Lagrangian

$$\phi = \frac{\partial^2 L}{\partial p^i \partial p^j}\pi^i\pi^j + 2\frac{\partial^2 L}{\partial u^i \partial p^j}\pi^i\eta^j + \frac{\partial^2 L}{\partial u^i \partial u^j}\eta^i\eta^j \tag{39.3}$$

The Euler-Lagrange equations of (39.3) are the Jacobi equations

$$\frac{d}{dt}\frac{\partial \phi}{\partial \pi^i} - \frac{\partial \phi}{\partial \eta^i} = 0 \quad (i = 1, \ldots, n) \tag{39.4}$$

These Jacobi equations are the linearised Euler-Lagrange equations. Verify this!

For Lagrangians independent of $t$ we noticed a conservation law. When you multiply (39.2) by $p^i(t) = \dot{u}^i(t)$ you get

$$0 = p^i(t)\frac{d}{dt}\frac{\partial L}{\partial p^i} - \dot{u}^i(t)\frac{\partial L}{\partial u^i} = \frac{d}{dt}\left(p^i\frac{\partial L}{\partial p^i}\right) \underbrace{-\dot{p}^i(t)\frac{\partial L}{\partial p^i} - \dot{u}^i(t)\frac{\partial L}{\partial u^i}}_{-\frac{dL}{dt}}$$

$$= \frac{d}{dt}\left(p^i\frac{\partial L}{\partial p^i} - L\right)$$

## 39.1 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

$$V = V^i x_i = V^i(u)x_i(u) = V^1(u)x_1(u) + \cdots + V^n(u)x_n(u), \tag{39.5}$$

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

## 39.2 Covariant differentiation

If we differentiate a vector field $V$ as given by (39.5) 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 (39.5) with respect to $u^j$ can be written as

$$\frac{\partial V}{\partial u^j} = \frac{\partial V^i}{\partial u^j} x_i + V^i x_{ij}, \quad x_{ij} = \frac{\partial x_i}{\partial u^j} = \frac{\partial^2 x}{\partial u^j \partial u^i} = \frac{\partial^2 x}{\partial u^i \partial u^j} = x_{ji} \tag{39.1}$$

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 $\Gamma_{ijk}$ is obtained from $\Gamma^l_{ij}$ using $g_{lk}$. Introducing $g^{kl} = g^{lk}$ by

$$g_{lk} g^{km} = \delta^m_l,$$

we also obtain $\Gamma^m_{ij}$ from $\Gamma_{ijk}$:

$$g^{mk} \Gamma_{ijk} = \Gamma^l_{ij} g_{lk} g^{km} = \Gamma^l_{ij} \delta^m_l = \Gamma^m_{ij}$$

The relation between both $\Gamma$-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 $\Gamma_{ijk}$. 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 $\Gamma$'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

$$\Gamma_{ijk} = \frac{1}{2}\left(g_{jk,i} + g_{ik,j} - g_{ij,k}\right), \quad \Gamma^m_{ij} = \frac{1}{2}g^{mk}\left(g_{jm,i} + g_{im,j} - g_{ij,m}\right) \tag{39.2}$$

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 (39.1) 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

$$D_{u^j}V := (\frac{\partial V}{\partial u^j})_T = \left(\frac{\partial V^l}{\partial u^j} + V^i\Gamma^l_{ij}\right)x_l, \quad V = V^i x_i \tag{39.3}$$

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

$$V^i \quad \text{and} \quad (D_{u^j}V)^l = \frac{\partial V^l}{\partial u^j} + V^i\Gamma^l_{ij}$$

## 39.3 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

$$\dot{\gamma}(t) = X(\gamma(t)), \quad \gamma(0) = x_0 \in M, \tag{39.1}$$

where $X$ is a tangent vector field defined near $x_0$. The differential equation in (39.1) is called the *flow equation* for $X$. Using coordinates $u$, with $u = u_0$ corresponding to $x_0$, the expressions in (39.1) 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

$$\dot u^i = X^i(u), \quad u(0) = u_0. \tag{39.2}$$

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) $\phi : M \to \mathrm{I\!R}$, given in local coordinates as

$$\phi = \phi(u^1, \ldots, u^n),$$

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 $\phi$ 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 $\phi$ is extended to a neighbourhood of $M$ in $\mathrm{I\!R}^d$. 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 $\phi$ 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 $\nabla_X$ 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.

## 39.4 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 (39.1), 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 (39.2) around $u(t)$,

$$\dot{\xi}^i = \underbrace{(\frac{\partial X^i}{\partial u^j})}_{\text{in } (u(t)} \xi^j(t), \quad \xi^j(0) = Y^i(u_0) \tag{39.1}$$

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

$$[X, Y] = \mathcal{L}_X Y \tag{39.2}$$

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

$$[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0 \tag{39.3}$$

holds.

## 39.5 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 (39.3) 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

$$\nabla_X V = X^j \left( \frac{\partial V^l}{\partial u^j} + V^i \Gamma^l_{ij} \right) \frac{\partial}{\partial u^i} \tag{39.1}$$

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 $M$.

## 39.6 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

$$\bar{\nabla}_X (N \cdot Y) = \bar{\nabla}_X N \cdot Y + N \cdot \bar{\nabla}_X Y, \tag{39.1}$$

in which $\bar{\nabla}$ is the (standard covariant) derivative in $\mathbb{R}^d$. On $M$ the left hand side of (39.1) 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

$$\bar{\nabla}_X N \cdot Y = -N \cdot \mathrm{I\!I}(X, Y) \quad \text{on } M. \tag{39.2}$$

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

$$h(X, Y) = N \cdot \mathrm{I\!I}(X, Y) = -\bar{\nabla}_X N \cdot Y = h_{ij} X^i Y^j \tag{39.3}$$

## 39.7 Curvature

The equality

$$\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z = \nabla_{[X,Y]} Z + R(X, Y) Z \tag{39.1}$$

[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

$$(R(X,Y)Z)^\alpha = R^\alpha_{ijk}Z^iX^jY^k. \tag{39.2}$$

So $Z$ comes first[7] and then $X$ and $Y$. Using (39.1) and writing

$$\Gamma^\alpha_{ij,k}=\frac{\partial\Gamma_{ij}}{\partial u^k}$$

you should verify that[8]

$$R^\alpha_{ijk}=\Gamma^\beta_{ik}\Gamma^\alpha_{\beta j}-\Gamma^\beta_{ij}\Gamma^\alpha_{\beta k}+\Gamma^\alpha_{ik,j}-\Gamma^\alpha_{ij,k} \tag{39.3}$$

and the zero $ijk$ and $jk$ cyclic sums

$$R^\alpha_{ijk}+R^\alpha_{kij}+R^\alpha_{jki}=0=R^\alpha_{ijk}+R^\alpha_{ikj} \tag{39.4}$$

If $W$ is another tangent field then[9]

$$Rm(X,Y,Z,W)=R(X,Y)Z\cdot W=R^\alpha_{ijk}Z^iX^jY^kg_{\alpha l}W^l=R_{lijk}W^lZ^iX^jY^k, \tag{39.5}$$

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 $2 \times 2$ determinants you can get from the matrix $h_{ij}$. Note that similarly

$$(W \cdot X)(Y \cdot Z) - (W \cdot Y)(X \cdot Z) = \underbrace{(g_{ik}g_{lj} - g_{ij}g_{lk})}_{G_{lijk}} W^l Z^i X^j Y^k, \tag{39.6}$$

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

$$Rm(X, Y, Z, W) = \mathrm{I\!I}(X, W)\mathrm{I\!I}(Y, Z) - \mathrm{I\!I}(X, Z)\mathrm{I\!I}(Y, W), \tag{39.7}$$

which in the codimension 1 case (39.3) reduces to

$$Rm(X, Y, Z, W) = h(W, X)h(Y, Z) - h(W, Y)h(X, Z),$$

so

$$Rm(X, Y, Z, W) = \underbrace{(h_{ik}h_{lj} - h_{ij}h_{lk})}_{R_{lijk}} W^l Z^i X^j Y^k, \tag{39.8}$$

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 (39.7) and (39.6). In $\mathbb{R}^3$ this is the scalar ratio between (39.8) and (39.6).

## 39.8 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 $\gamma$'s in the book. My $\gamma(t)$ is what is $c(t)$ in the book. The energy is defined by

$$E = \frac{1}{2}\int_a^b |\dot\gamma(t)|^2\, dt = \frac{1}{2}\int_a^b \dot\gamma(t)\cdot\dot\gamma(t)\, dt = \frac{1}{2}\int_a^b x_i\dot u^i \cdot x_j\dot u^j\, dt$$

$$= \frac{1}{2}\int_a^b \dot u^i\dot u^j g_{ij}(u)\, dt = \int_a^b L(u(t),\dot u(t))\, dt,$$

in which

$$L = L(u,p) = \frac{1}{2}p^i p^j g_{ij}(u). \tag{39.9}$$

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 $\gamma$ 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

$$\ddot u^i + \Gamma^i_{\alpha\beta}\dot u^\alpha \dot u^\beta = 0 \tag{39.10}$$

and are called the geodesic equations. Indeed,

$$\Gamma^i_{\alpha\beta} = \frac{1}{2}g^{ik}\left(g_{\alpha k,\beta} + g_{\beta k,\alpha} - g_{\alpha\beta,k}\right),$$

the symbols computed in (39.2). 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

$$g_{ij}(u) = g(|u|)\delta_{ij} \tag{39.11}$$

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 (39.11) 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 (39.11) in $\mathbb{R}^2$?

[10] implying $0 = g'(0) = g'''(0) = g'''''(0) = \cdots$

## 39.9 The Jacobi equations

Consider the Lagrangian (39.9).

- Show that the Jacobi equations (39.4) for (39.9) are

$$\ddot{\eta}^i + 2\Gamma^i_{jk}\dot{u}^j\dot{\eta}^k + \Gamma^i_{jk,l}\dot{u}^j\dot{u}^k\eta^l = 0 \tag{39.12}$$

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 (39.10) may be written as

$$D_t\dot{\gamma} = 0, \quad \dot{\gamma} = \dot{u}^i x_i$$

- Derive that (39.12) may be written as

$$(D_t^2\eta)^i + \dot{u}^\alpha R^i_{\alpha\beta k}\eta^\beta\dot{u}^k = 0, \quad \text{i.e.} \quad D_t^2\eta + R(\eta,\dot{\gamma})\dot{\gamma} = 0$$

# 40 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

$$u \to u - f'(u)^{-1} f(u) \tag{40.1}$$

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

$$u \to u + f'(0)^{-1}(v - f(u)) \tag{40.2}$$

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 (40.2) 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.

## 40.1 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

$$|f'(u)| \leq M_1 \quad \text{and} \quad |f''(u)| \leq M_2 \quad \forall u \in B. \tag{40.3}$$

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,

$$f(u_n) = \underbrace{f(u_{n-1}) + f'(u_{n-1})(u_n - u_{n-1})}_{\text{linear aproximation}} + Q_f(u_{n-1}, u_n), \tag{40.4}$$

in which

$$|Q_f(u_{n-1}, u_n)| \leq \frac{M_2}{2}|u_n - u_{n-1}|^2, \tag{40.5}$$

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 (40.3) which says that the linear map $f'(u) : X \to X$ satisfies

$$|f'(u)v| \leq M_1|v| \quad \forall u \in B \quad \forall u \in X, \tag{40.6}$$

but a similar bound

$$|L(u)| \leq C \tag{40.7}$$

for maps $L(u)$, that act as right inverses of $f'(u)$ in the sense that

$$f'(u_{n-1})L(u_{n-1})f(u_{n-1}) = f(u_{n-1}), \tag{40.8}$$

is essential. Writing

$$p_n = |u_n - u_{n-1}| \quad \text{and} \quad q_n = |f(u_n)| \tag{40.9}$$

the Newton scheme

$$u_n = u_{n-1} - L(u_{n-1})f(u_{n-1}) \quad (n \in \mathbb{N}), \tag{40.10}$$

starting with $u_0 = 0$, then defines $u_n \in B$ as long as

$$p_1 + p_2 + \cdots + p_n < 1, \tag{40.11}$$

and the inequalities

$$p_n \leq Cq_{n-1} \quad \text{and} \quad q_n \leq \frac{1}{2}M_2 p_n^2 \tag{40.12}$$

are immediate from (40.4,40.5,40.10). Note that (40.10) kills the linear approximation in (40.4). The inequalities in (40.12) are complemented by

$$q_0 = |f(0)| \quad \text{and} \quad p_1 \leq Cq_0 = C|f(0)|. \tag{40.13}$$

## 40.2 The optimal result

Clearly (40.12) and (40.13) combine as

$$p_n \leq \mu p_n^2 \quad \text{with} \quad \mu = \frac{1}{2}MC \quad \text{and} \quad p_1 \leq C|f(0)|, \tag{40.14}$$

and the condition to be stated is which $\bar{P} = \bar{P}(\mu)$ guarantees that the implication

$$C|f(0)| < \bar{P} \implies \sum_{n=1}^{\infty} p_n < 1 \tag{40.15}$$

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 (40.10) with $u_0 = 0$. Note that with $C|f(0)| \leq \bar{P}$ 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 (40.14) and (40.15) by inequalities. This leads to

$$p_n = \mu p_{n-1}^2 \ \text{ for } \ n \in \mathbb{N}; \quad p_1 = \bar{P}\,; \quad \sum_{n=1}^{\infty} p_n = 1. \tag{40.16}$$

Via $\xi_n = \mu p_n$ and $\xi_n = \xi_{n-1}^2$ this is easily seen to be equivalent to

$$\mu = G(\mu\bar{P}) \quad \text{with} \quad G(\xi) = \xi + \xi^2 + \xi^4 + \xi^8 + \xi^{16} + \cdots \tag{40.17}$$

but this does not yield a simple formula for $\bar{P} = \bar{P}(\mu)$.

## 40.3 A suboptimal result

A rough estimate

$$G(\xi) < \xi + \xi^2 + \xi^3 + \xi^4 + \xi^5 + \cdots = \frac{\xi}{1-\xi} \tag{40.18}$$

leads to a simple but suboptimal formula:

$$\bar{P} = \frac{1}{1+\mu} \quad \text{or} \quad \mu = \frac{1}{\bar{P}} - 1. \tag{40.19}$$

## 40.4 Alternative proof of convergence

The alternative approach to (40.12) and (40.13) in [KP] is to derive an estimate of the form

$$p_n \le e^{-\gamma\lambda^n} \tag{40.20}$$

via induction starting from

$$p_1 \le C|f(0)| < \bar{P} = e^{-\gamma\lambda}, \tag{40.21}$$

with choices of $\gamma$ and $\lambda$ that guarantee both

$$\sum_{n=1}^{\infty} e^{-\gamma\lambda^n} \le 1 \tag{40.22}$$

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

$$\ln\mu \le \gamma\lambda^{n-1}(2-\lambda) \quad \forall n \ge 1.$$

## 40.5 The optimal alternative result

For a given $\mu$ this is equivalent to

$$\ln\mu \le \gamma\lambda(2-\lambda) \quad \text{and} \quad \lambda \le 2 \tag{40.23}$$

if we make the obvious restriction that $\gamma$ and $\lambda$ be positive. Conditions (40.21) and (40.23) suggest $\alpha = \gamma\lambda$ and $\lambda$ as the more relevant parameter so we have to pick $\alpha > 0$ and $1 < \lambda \le 2$ with

$$\ln\mu \le \alpha(2-\lambda), \quad \sum_{n=0}^{\infty} e^{-\alpha\lambda^n} \le 1 \quad \text{and} \quad \bar{P} = e^{-\alpha} \quad \text{maximal.} \tag{40.24}$$

For $\mu > 1$ the inequalities define a set in the first quadrant of the $\lambda,\alpha$-plane bounded by the two curves given by

$$\ln\mu = \alpha(2-\lambda) \;\text{ and }\; \sum_{n=0}^{\infty} e^{-\alpha\lambda^n} = 1, \tag{40.25}$$

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

$$1 = \sum_{n=0}^{\infty} \bar{P}^{(2+\frac{\ln\mu}{\ln\bar{P}})^n} \quad \text{and} \quad \mu \geq 1 \tag{40.26}$$

hits the curve defined by (40.17) in $\mu = 1$ and lies below (40.17) of course, but above (40.19) 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 (40.24) is given by

$$\sum_{n=0}^{\infty} \bar{P}^{2^n}.$$

## 40.6 A suboptimal alternative result

A more explicit formula is again obtained via a rough estimate

$$\sum_{n=1}^{\infty} e^{-\gamma\lambda^n} \leq \underbrace{\sum_{n=1}^{\infty} e^{-\gamma(1+n(\lambda-1))}}_{\text{a geometric series}} = \frac{e^{-\gamma\lambda}}{1-e^{-\gamma(\lambda-1)}} = \frac{e^{-\alpha}}{1-e^{\gamma}e^{-\alpha}} \tag{40.27}$$

and replacing (40.24) 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

$$1 \leq \mu = \frac{1}{\bar{P}^{2+\frac{\ln\bar{P}}{\ln(\frac{1}{\bar{P}}-1)}}} < \frac{1}{\bar{P}} - 1 \tag{40.28}$$

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 $\bar{P} \to 0$.

## 40.7 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.

## 40.8 A much better suboptimal alternative result

Actually the first rough estimate above works better with $\alpha$ dan with $\gamma$, as I only noticed May 21. Directly in terms of $\gamma$ and $\lambda$ we have

$$\sum_{n=1}^{\infty} e^{-\gamma\lambda^n} = \sum_{n=1}^{\infty} e^{-\gamma\lambda\lambda^{n-1}} \leq \sum_{n=1}^{\infty} e^{-\gamma\lambda(1+(n-1)(\lambda-1))} = \frac{e^{-\gamma\lambda}}{1-e^{-\gamma\lambda(\lambda-1)}} \leq 1 \tag{40.29}$$

if

$$2-\lambda \leq \frac{\ln(e^{\gamma\lambda}-1)}{\gamma\lambda} = \frac{\ln(e^{\alpha}-1)}{\alpha},$$

so that we arrive at

$$\ln\mu \leq \alpha(2-\lambda), \quad \alpha(2-\lambda) \leq \ln(e^{\alpha}-1) \quad \text{and} \quad \bar{P} = e^{-\alpha} \quad \text{maximal.} \tag{40.30}$$

This is the optimal estimate using the Bernoulli type inequality

$$\lambda^n \geq 1 + n(\lambda-1). \tag{40.31}$$

With equality in the final inequality in (40.29) we arrive at

$$\ln\mu \leq \ln(e^{\alpha}-1) = \ln(\frac{1}{\bar{P}}-1),$$

which for $\mu > 1$ coincides with (40.19) and we can forget about the annoying (40.28) above. Note that factoring out another $\lambda$ in the exponent in (40.29) 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).

# 41 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

$$N = \frac{\bar{k} - k}{k - m} \tag{41.1}$$

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

$$|S(t)u|_k \leq K_{kl} t^{k-l} |u|_l \quad \text{and} \quad |(I - S(t))u|_l \leq \frac{K^{kl}}{t^{k-l}} |u|_k \tag{41.2}$$

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 (41.2) says that

$$|S(t)|_{kl} \leq K_{kl} t^{k-l} \quad \text{and} \quad |(I - S(t))|_{lk} \leq \frac{K^{kl}}{t^{k-l}}. \tag{41.3}$$

Besides (41.3) 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 $|u|_k \leq 1$.

## 41.1 The modified scheme

The idea of Nash was to modify Newton's scheme into

$$u_n = u_{n-1} - S(t_{n-1})L(u_{n-1})f(u_{n-1}), \tag{41.4}$$

with a suitable choice of $t_n \to \infty$ as $n \to \infty$. In (41.4) 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 (40.4) is modified when combined with (41.4). We have

$$f(u_n) = \underbrace{f(u_{n-1}) + f'(u_{n-1})(u_n - u_{n-1})}_{\text{vanishes with (40.10)}} + Q_f(u_{n-1}, u_n)$$

$$= \underbrace{f'(u_{n-1})(I - S_{n-1})L(u_{n-1})f(u_{n-1})}_{\text{because of (41.4)}} + Q_f(u_{n-1}, u_n),$$

so that, with

$$p_n = |u_n - u_{n-1}|_k \quad \text{and} \quad q_n = |f(u_n)|_l,$$

the estimate

$$q_n \leq \underbrace{M_1^{lk}|I - S_{n-1}|_{km}|L(u_{n-1})f(u_{n-1})|_m}_{\text{new error like term}} + \frac{1}{2}M_2^{lk}p_n^2 \tag{41.5}$$

holds.

## 41.2 The new error term

The third factor in the error like term in (41.5) 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 (40.10) and that should guarantee that quadratic term in (41.5) will still allow us to establish a conclusion like

(40.15). 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

$$|L(u)f(u)|_{\bar{m}} \leq A_{\bar{m}\bar{k}}(1 + |u|_{\bar{k}}) \tag{41.6}$$

with

$$\bar{k} - \bar{m} = k - m,$$

which requires an additional estimate for

$$r_n = 1 + |u_n|_{\bar{k}} \tag{41.7}$$

to be used in combination with

$$q_n \leq M_1^{lk} \underbrace{|I - S_{n-1}|_{k\bar{m}}}_{\text{controled by (41.3)}} r_{n-1} + \frac{1}{2} M_2^{lk} p_n^2 \tag{41.8}$$

and the estimate for $p_n$. Via (41.4) the latter now reads

$$p_n \leq |S_{n-1}|_{km} C_{ml} q_{n-1} \tag{41.9}$$

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 (41.4). 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

$$1 \leq r_n \leq 1 + A_{\bar{m}\bar{k}} \sum_{j=1}^{n} |S_{j-1}|_{\bar{k}\bar{m}} r_{j-1}. \tag{41.10}$$

The "error" terms accumulate but can be kept under control as we shall see below.

The system of inequalities (41.9,41.8,41.10) and initial inequalities for $q_0$, $r_0 = 1$ and $r_1$ allows again estimates of the form (40.20), provided $\bar{k} - k = \bar{m} - m$ is sufficiently large in terms of (41.1). The idea is to get the first term in (41.8) 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 (41.9) can be chosen not to large so as still to have (40.20) 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

$$t_{n-1} = e^{\beta\lambda^{n-1}} \tag{41.11}$$

with $\beta > 0$ to be chosen in terms of $\gamma$. Note that this gives $\lambda^n$ in the exponents of the exponential bounds for $S_n$ and $I - S_n$.

Here we choose to keep $\lambda$ as a parameter in a range as large as possible, like we did in the analysis of (40.10). 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

$$p_n \leq e^{-\gamma\lambda^n} \quad \text{and} \quad r_n \leq e^{\delta\lambda^n} \tag{41.12}$$

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.

## 41.3 The system of inequalities

With (41.11) we have the system of inequalities

$$p_n \leq K_{km} e^{(k-m)\beta\lambda^{n-1}} C_{ml} q_{n-1}; \tag{41.13}$$

$$q_n \leq M_1^{lk} K^{k\bar{m}} e^{(k-\bar{m})\beta\lambda^{n-1}} A_{\bar{m}\bar{k}} \underbrace{r_{n-1}}_{\leq e^{\delta\lambda^{n-1}}} + \frac{1}{2} M_2^{lk} \underbrace{p_n^2}_{\leq e^{-2\gamma\lambda^n}}; \tag{41.14}$$

$$1 \leq r_n \leq 1 + \underbrace{A_{\bar{m}\bar{k}} K_{\bar{k}\bar{m}}}_{\mu_3} \sum_{j=1}^{n} e^{(\bar{k}-\bar{m})\beta\lambda^{j-1}} \underbrace{r_{j-1}}_{\leq e^{\delta\lambda^{j-1}}}, \tag{41.15}$$

and we aim for a proof of (41.12) via induction, using the underbraced estimates in the three inequalities above as induction hypothesis. In (41.14) the estimate of the first term is controled by the estimate of the second term if

$$e^{(k-\bar{m})\beta\lambda^{n-1}} e^{\delta\lambda^{n-1}} \leq e^{-2\gamma\lambda^n},$$

requiring

$$(\bar{m} - k)\beta \geq \delta + 2\gamma\lambda, \tag{41.16}$$

which says that in the $\lambda, \beta$-plane we must be above a line that comes down as $\bar{m}$ is increased.

Combining the first two inequalities we arrive at

$$p_n \le e^{(k-m)\beta\lambda^{n-1}}(\mu_1 e^{(k-\bar{m})\beta\lambda^{n-2}} r_{n-2} + \mu_2 p_{n-1}^2) \quad r_n \le 1 + \mu_3 \sum_{j=0}^{n-1} e^{(\bar{k}-\bar{m})\beta\lambda^j} r_j, \tag{41.17}$$

the constants $\mu_{123}$ given by

$$\mu_1 = \underbrace{K_{km}C_{ml}}_{C} M_1^{lk} \underbrace{K^{k\bar{m}}A_{\bar{m}\bar{k}}}_{A}, \quad \mu_2 = \frac{1}{2}\underbrace{K_{km}C_{ml}}_{C} M_2^{lk}, \quad \mu_3 = \underbrace{K_{\bar{k}\bar{m}}A_{\bar{m}\bar{k}}}_{\bar{A}}. \tag{41.18}$$

## 41.4 Estimating the increments

Under the assumption that (41.16) holds, the induction hypotheses for $p_{n-1}$ and $r_{n-2}$ produce the desired inequality for $p_n$ from (41.17) 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

$$\ln(\mu_1 + \mu_2) \le \lambda(\gamma(2-\lambda) - (k-m)\beta) \quad \text{with} \quad (k-m)\beta \le \gamma(2-\lambda), \tag{41.19}$$

a sharp upper bound for $\beta$ 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

$$\lambda\gamma(2-\lambda) \le \ln(e^{\gamma\lambda} - 1). \tag{41.20}$$

## 41.5 Estimating the error terms

For the inductive construction of the upper bound for $r_n$ we set

$$b = (\bar{k} - \bar{m})\beta = (k-m)\beta > 0 \tag{41.21}$$

and conclude from the inequality in (41.17) that (shifting the index)

$$r_n \leq 1 + \mu_3 \sum_{j=0}^{n-1} e^{b\lambda^j} r_j \leq 1 + \mu_3 \sum_{j=0}^{n-1} e^{b\lambda^j} e^{\delta\lambda^j}$$

in view of the induction assumption for (all) smaller $n$. Thus we need the inequality

$$1 + \mu_3 \sum_{j=0}^{n-1} e^{(b+\delta)\lambda^j} \leq e^{\delta\lambda^n} \tag{41.22}$$

for all $n \geq 2$. Recall that we start with $r_0 = 1 \leq e^\delta$ and

$$1 \leq r_1 \leq e^{\delta\lambda} \quad \text{(and also } p_1 \leq e^{\gamma\lambda} \text{ of course)} \tag{41.23}$$

via a smallness assumptions on $q_0$ still to be discussed.

Dividing by the right hand side, (41.22) is equivalent to

$$e^{-\delta\lambda^n} + \mu_3 (e^{(b+\delta-\delta\lambda)\lambda^{n-1}} + e^{-\delta\lambda^n} \sum_{j=0}^{n-2} e^{(b+\delta)\lambda^j}) \leq 1 \tag{41.24}$$

in which we have separated the probably dominant term with $j = n-1$ from the sum. Neglecting the sum in (41.24) a sufficient (and in any case necessary) condition for the induction step to work for all $n \geq 2$ would be that

$$\ln \mu_3 + (b + \delta - \delta\lambda)\lambda^{n-1} \leq 0 \quad \text{with} \quad b \leq \delta(\lambda - 1), \tag{41.25}$$

so that in particular we now need to impose two inequalities on $b$, namely

$$b < \delta(\lambda - 1) \quad \text{and} \quad b < \gamma(2 - \lambda), \tag{41.26}$$

the latter being the (strict) inequality from (41.19).

These two bounds severely restrict the bound in (41.16), which in terms of $b$ becomes

$$\frac{\bar{m} - k}{k - m} b \geq \delta + 2\gamma\lambda, \tag{41.27}$$

and this does not really depend on how we turn the necessary condition (41.25) into a sufficient condition, which we do next, rewriting it as

$$e^{-\delta\lambda^n} + \mu_3 (e^{(b+\delta-\delta\lambda)\lambda^{n-1}} + \sum_{j=0}^{n-2} e^{(b+\delta-\delta\lambda^{n-j})\lambda^j}) \leq 1.$$

In view of (41.26) and using Bernoulli's inequality (40.31) 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

$$e^{-\delta\lambda^2}+\mu_3 e^{(b+\delta-\delta\lambda)\lambda}(1+\frac{e^{-(b+\delta)(\lambda-1)}}{1-e^{(\lambda-1)(b+\delta-\delta\lambda^2)}})\leq 1 \tag{41.28}$$

Note that the first term on the right hand side of (40.31) 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 $\mu_3$. We note that in [KP] the proof is without the 1 in (40.31) 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

$$e^{-\delta\lambda^2}+3\mu_3 e^{(b+\delta-\delta\lambda)\lambda}\leq 1 \quad \text{suffices if} \quad \ln 2\leq\delta\lambda(\lambda-1)^2, \tag{41.29}$$

[1] And it is not a proof of the theorem actually stated.

and the first inequality in (41.29) certainly holds if it holds with the first exponential replaced by the larger second exponential. Thus we arrive at

$$\ln(1+3\mu_3) \leq \lambda(\delta(\lambda-1)-b) \quad \text{and} \quad \ln 2 \leq \delta\lambda(\lambda-1)^2 \tag{41.30}$$

as the final condition needed.

## 41.6 Sufficient conditions for a convergence result

Summing up, with the condition on $q_0$ still to be imposed we arrive at

$$\lambda\gamma(2-\lambda) \leq \ln(e^{\gamma\lambda}-1), \tag{41.31}$$

$$\ln 2 \leq \delta\lambda(\lambda-1)^2, \tag{41.32}$$

$$(\bar{m}-k)\beta \geq \delta + 2\gamma\lambda, \tag{41.33}$$

$$(k-m)\beta < \gamma(2-\lambda) \quad \text{and} \quad (\bar{k}-\bar{m})\beta < \delta(\lambda-1) \tag{41.34}$$

as conditions on the parameters that we still have to choose.

The first inequality, (41.31), 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

$$\sum_{n=1}^{\infty} e^{-\gamma\lambda^n} \leq 1.$$

The second, (41.32), was a technical condition to bound the sum of the geometric series in (41.28) by 2. The third, (41.33), allows to bound the error term in estimate (41.14) 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 (41.12) for $p_n$ and $r_n$. Thus, given $\mu_1, \mu_2, \mu_3$, we need to choose $1 < \lambda < 2$ and $\gamma, \beta, \delta > 0$ such that

$$\lambda(\gamma(2-\lambda)-(m-k)\beta) \geq \ln(\mu_1+\mu_2); \tag{41.35}$$

$$\lambda(\delta(\lambda-1)-(\bar{m}-\bar{k})\beta) \geq \ln(1+3\mu_3). \tag{41.36}$$

After a simultaneous rescaling of $\gamma, \beta, \delta$, this is always possible once the first 5 conditions are satisfied. The inequalities in (41.34) 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 (41.18) 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 $f : C^k \to C^l$.

## 41.7 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

$$C q_0 \leq e^{-(k-m)\beta}e^{-\gamma\lambda} \quad \text{and} \quad \bar{C} q_0 \leq e^{-(\bar{k}-m)\beta}(e^{\delta\lambda} - 1) \tag{41.37}$$

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 41.6 to make the induction steps work in the proof of the desired estimates (41.12) for $p_n$ and $r_n$.

## 41.8 The optimal choice of parameters

At this point we compare (41.35) and (41.36) to (40.23). The strict inequalities in (41.34) 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

$$\frac{\bar{m} - k}{k - m} = \frac{\bar{m} - m}{k - m} - 1 = \frac{\bar{m} - m}{\bar{k} - \bar{m}} - 1 = \frac{\bar{k} - k}{k - m} - 1 = N - 1 \tag{41.38}$$

has to be sufficiently large for the set of allowable $b$, as defined by (41.21), to be nonempty. Note that in Nash' strategy to get around the ill-posedness of Newton's method, (41.1) 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 (41.34) equal to one another, so as to maximize the allowable upper bound for $\beta$. Thus we choose $1 < \lambda < 2$ such that

$$\gamma(2 - \lambda) = \delta(\lambda - 1) \quad \text{whence} \quad \lambda = \frac{2\gamma + \delta}{\gamma + \delta} \tag{41.39}$$

and (41.33,41.34) become

$$\frac{4\gamma^2 + 3\gamma\delta + \delta^2}{\gamma + \delta} \le (N - 1)b < (N - 1)\frac{\gamma\delta}{\gamma + \delta} \tag{41.40}$$

for

$$b = (k - m)\beta = (\bar{k} - \bar{m})\beta$$

in terms of $\gamma, \delta, N$, subject to (41.31,41.32) which reduce to

$$e^{\frac{2\gamma+\delta}{\gamma+\delta}\frac{\delta}{\gamma+\delta}} + 1 \le e^{\gamma\frac{2\gamma+\delta}{\gamma+\delta}} \quad \text{and} \quad \ln 2 \le \delta\,\frac{2\gamma + \delta}{\gamma + \delta}\left(\frac{\gamma}{\gamma + \delta}\right)^2 . \tag{41.41}$$

In particular (41.40) requires

$$N > \frac{4\gamma}{\delta} + 4 + \frac{\delta}{\gamma} \ge 8, \tag{41.42}$$

the minimum 8 being realised by

$$\delta = 2\gamma. \tag{41.43}$$

The further choice of parameters depends on the constants which are as indicated in (41.18), at the end of Section 41.6 and in (41.37):

$$C = K_{km} C_{ml};\ \bar{C} = K_{\bar{k}m} C_{ml};\ A = K^{k\bar{m}} A_{\bar{m}\bar{k}};\ \bar{A} = K_{\bar{k}\bar{m}} A_{\bar{m}\bar{k}}; \tag{41.44}$$

$$\mu_1 + \mu_2 = C(M_1 A + \frac{1}{2} M_2); \quad \mu_3 = \bar{A}. \tag{41.45}$$

We collect these constants in one single constant $\Theta$ as

$$\Theta = \frac{3}{4} \max(\ln C + \ln(M_1 A + \frac{1}{2} M_2), \ln(1 + 3\bar{A})) \tag{41.46}$$

and, depending on $N$, the remaining parameters $\gamma, b$ have to be chosen to control these constants via

$$\Theta \leq \frac{2\gamma}{3} - b \tag{41.47}$$

and

$$\frac{14\gamma}{3} \leq (N-1)b < \frac{2\gamma}{3}(N-1), \quad e^{\frac{8}{9}} + 1 \leq e^{\frac{4\gamma}{3}}, \quad \ln 2 \leq \frac{8\gamma}{27}, \tag{41.48}$$

which is (41.40,41.41) with $\delta = 2\gamma$. The last inequality now implies the one preceding it.

For the initial condition $q_0$ we arrive via (41.39) 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

$$Cq_0 \leq e^{-\frac{4\gamma}{3}} e^{-b} \quad \text{and} \quad \bar{C} q_0 \leq (e^{\frac{8\gamma}{3}} - 1) e^{-(N+1)b}. \tag{41.49}$$

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 $\rho$ 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 $\rho$ we arrive at

$$b \geq \max\left(\frac{63}{4} \frac{\ln 2}{N-1}, \frac{7\Theta}{N-8}\right) \quad \text{and} \quad q_0 \leq \min\left(\frac{1}{C} e^{-\frac{2N+5}{7N} b}, \frac{80}{\bar{C}} e^{-(N-1)b}\right) \tag{41.50}$$

as sufficient conditions, to be used as: given $\Theta$ 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 $|u|_k < 1$.

## 41.9 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 (41.47,41.48). The conditions on $\gamma$ rewrite as

$$\gamma \geq \max\left(\frac{3}{4} \ln\left(\frac{1}{\varepsilon} + e^{\frac{8}{9}}\right), 2^{\frac{27}{8}}\right) \sim \varepsilon^{-\frac{3}{4}}$$

as $\varepsilon \to 0$. This forces a larger choice of $b$ and thereby via (41.49) a smaller (exponentially small in terms of $\varepsilon$ in fact) bound on $q_0$ for the Nash scheme to converge within the ball of $k$-radius $\varepsilon$, as was to be expected of course. The fact that the limit $u$ is a solution of $f(u) = 0$ is immediate from (41.14).

Note that for the standard Newton method the constructed solution of $f(u) = 0$ will have $|u| < \varepsilon$ if we take equalities in (40.30) 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 $q_0 \leq C\bar{P}_\varepsilon$.

# 42 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 39. Requires a deeper discussion of the smoothing operators used in the proof of Theroem 32.19 and the Fourier transfrom.

# 43 Hartman-Grobman stelling

Some material prepared for this very enjoyable event (see also Section 14):

`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 $\phi$ 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 $\eta$ wellicht nog te kiezen, dan zien we dat de onbekende functie $\phi$ 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 $\phi$ differentieerbaar is volgt (als $\lambda \neq 0$) dat

$$\phi'(x) = \phi'(\lambda x)$$

voor alle $x \in \mathbb{R}$, en daarmee zijn heel veel $\phi'$ waarden gelijk aan elkaar, tenzij $|\lambda| = 1$. Als $\phi'$ 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 $\phi$ wel maken is echter zeer de vraag. In het iteratieproces helpt de $\lambda$ 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

$$\phi_1(x) = x + \frac{x^2}{\lambda},$$

$$\phi_2(x) = x + \left(\frac{1}{\lambda} + 1\right) x^2 + \frac{2x^3}{\lambda} + \frac{x^4}{\lambda^2},$$

$$\phi_3(x) = x + \left(\frac{1}{\lambda} + 1 + \lambda\right) x^2 + \left(\frac{2}{\lambda} + 2 + 2\,\lambda\right) x^3 + \left(\frac{1}{\lambda^2} + \frac{1}{\lambda} + 6 + \lambda\right) x^4$$
$$+ \left(\frac{6}{\lambda} + 4\right) x^5 + \left(\frac{2}{\lambda^2} + \frac{6}{\lambda}\right) x^6 + \frac{4x^7}{\lambda^2} + \frac{x^8}{\lambda^3},$$

$$\phi_4(x) = x + \left(\frac{1}{\lambda} + 1 + \lambda + \lambda^2\right) x^2 + \left(\frac{2}{\lambda} + 2 + 4\,\lambda + 2\,\lambda^2 + 2\,\lambda^3\right) x^3$$
$$+ \left(\frac{1}{\lambda^2} + \frac{1}{\lambda} + 7 + 7\,\lambda + 6\,\lambda^3 + \lambda^4 + 7\,\lambda^2\right) x^4 + \cdots + \frac{x^{16}}{\lambda^4},$$

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

$$\lambda\phi(x, y) = \phi(\lambda x + a(x, y), \mu y + b(x, y))$$
$$\mu\psi(x, y) = \psi(\lambda x + a(x, y), \mu y + b(x, y))$$

kunnen oplossen naar de functies $\phi, \psi$ 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 $\phi$, in de tweede de functie $\psi$. Die voor $\phi$ 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 $\xi, \eta$. 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

$$(|\lambda| + \varepsilon)^\delta < |\lambda|.$$

# 44 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 $\gamma_\zeta$ 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 $\gamma_\zeta$ 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 $\zeta$-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"

$\phi_\zeta$. If $M_\zeta$ is the maximum of $\operatorname{Re}\Phi(z;\zeta)$ along $\gamma_\zeta$, realised in some $z = m_\zeta$, one may also factor out $e^{M_\zeta}$ and write

$$\operatorname{Ai}(\zeta) = \frac{e^{M_\zeta + i\phi_\zeta}}{2\pi} \underbrace{\int_{\gamma_\zeta} e^{-f_0(z;\zeta)} dz}_{\to\,?\ \text{as}\ |\zeta|\to\infty},$$

in which $f_0(z;\zeta) \geq 0$ along $\gamma_\zeta$. Typically $f_0(z;\zeta)$ has a unique global minimum zero along $\gamma_\zeta$ and $f_0(z;\zeta) \to +\infty$ as $|z| \to \infty$ along $\gamma_\psi$. Note though that the integrand is likely to be ill-behaved as $\rho = |\zeta| \to \infty$, also because the contour $\gamma_\zeta$ 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 $\gamma_\zeta$.

Thus, returning to the definition of $\operatorname{Ai}(\zeta)$ one writes $\operatorname{Ai}(\rho e^{i\psi}) =$

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \underbrace{e^{i(\rho e^{i\psi} x + \frac{x^3}{3})}}_{\text{scale}\, x = \rho^{\frac{1}{2}} u} dx = \frac{\rho^{\frac{1}{2}}}{2\pi} \int_{-\infty}^{\infty} e^{i\rho^{\frac{3}{2}}(e^{i\psi} u + \frac{u^3}{3})} du = \frac{\rho^{\frac{1}{2}}}{2\pi} \int_{-\infty}^{\infty} e^{\rho^{\frac{3}{2}}\Psi(u)} du,$$

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 $\rho$.

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

$$\frac{df}{dt} = f_u\dot{u} + f_v\dot{v} = f_u^2 + f_v^2 > 0\,, \quad \frac{dg}{dt} = g_u\dot{u} + g_v\dot{v} = g_u f_u + g_v f_v = 0.$$

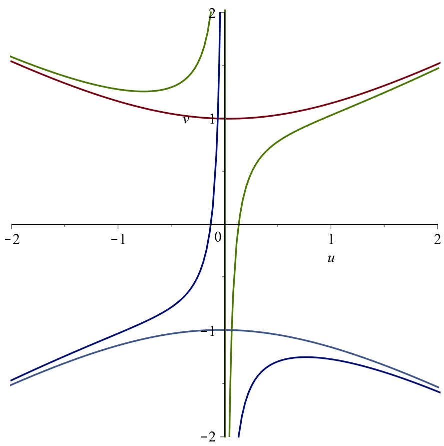



Figure 6: orbits for $\psi = \frac{\pi}{24}$

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 $\psi$ 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 $\psi$ is taken larger. Note that although $\mathrm{Ai}(\rho e^{i\psi})$ is $2\pi$-periodic in $\psi$, the parametrisation of the saddle point $m_\psi$ is only $4\pi$-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 $m_\psi$. Expansion gives

$$\varphi_\psi(-s+p) = c + \frac{-1+c}{s}p + \frac{1}{3}\frac{2c-1}{c+1}p^2 + \ldots.$$

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 $\psi$ there are changes as $\psi$ crosses $\frac{\pi}{3}$ and $\frac{2\pi}{3}$. For all $0 \leq \psi < \frac{2\pi}{3}$ 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 $M_\psi$, 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.

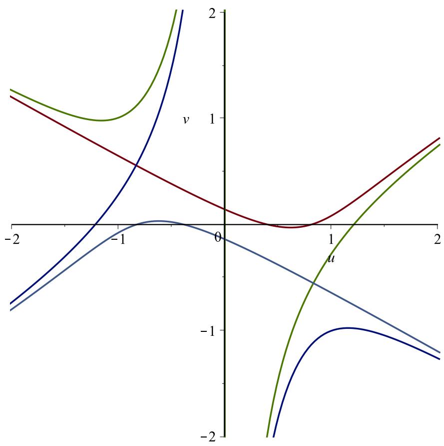



Figure 7: orbits for $\psi = \frac{15\pi}{24}$

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 $\psi$ is increased to $\psi = \pi$ both $M_\pi$ 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 $\infty$ 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

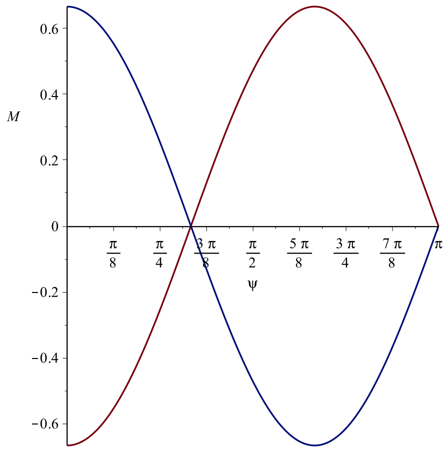



Figure 8: $M_\psi$ and $M_{\psi+2\pi}$

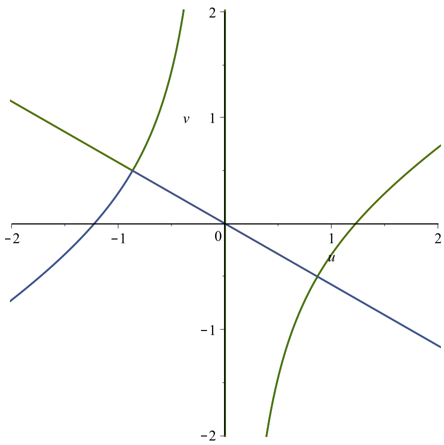



Figure 9: orbits for $\psi = \frac{16\pi}{24} = \frac{2\pi}{3}$

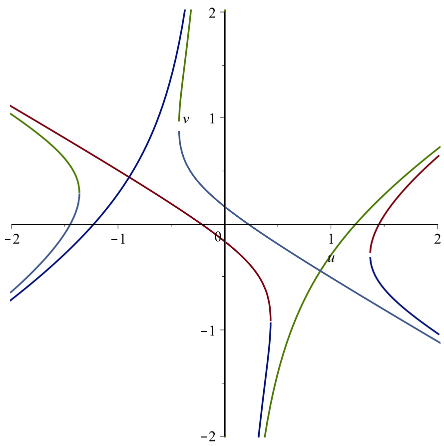



Figure 10: orbits for $\psi = \frac{17\pi}{24}$

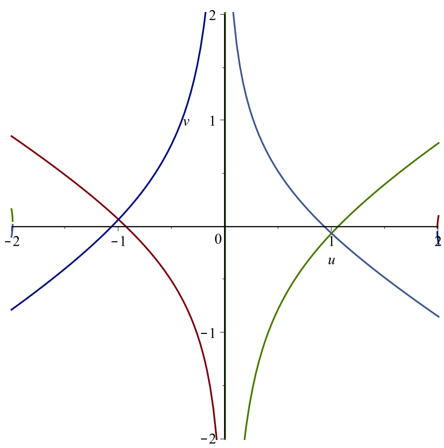



Figure 11: orbits for $\psi = \frac{23\pi}{24}$

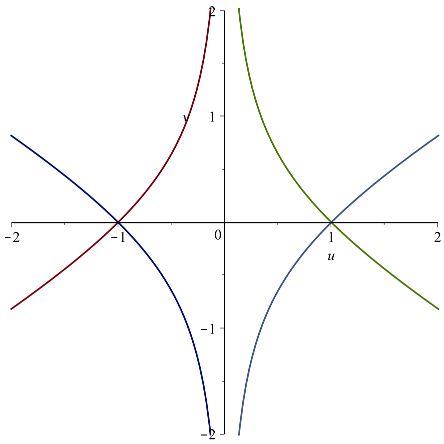



Figure 12: orbits (stable manifolds: blue curves) for $\psi = \frac{24\pi}{24} = \pi$

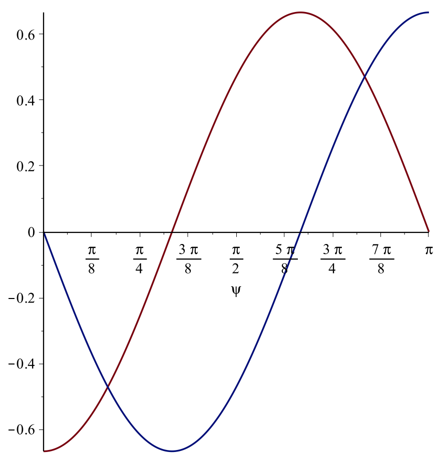



Figure 13: Amplitude $M_\psi$ (red) and phase $\phi_\psi$, changes at $\psi = 0, \frac{\pi}{3}, \frac{2\pi}{3}, \pi$.

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 $\gamma$ 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 $\rho$ 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 $\phi$ by solving

$$D = \frac{dD}{dx} = 0.$$

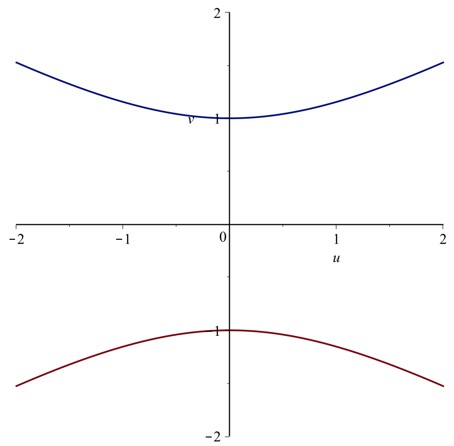



Figure 14: $u, v$-plane for $\psi = 0$, the vertical axis also consist of orbits

The phase $\phi$ 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 $t, c, \xi, a$, the $y$-coordinate $y = y_c$. The phase $\phi_c$ 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 $y, \phi, M$ in the other saddle point replace $x_c$ by $-x_c$. Note that $x_c = x_c(t)$ and likewise for $y_c, \phi_c, M_c$ (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 $\Phi$ 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 $G_c$, $\varphi_c$ 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

$$F_c(u, \varphi_c(u;t);t) = -\frac{2}{9}\frac{tRu^2\left(3\,x_c + 2\,u\right)^2}{\left(x_c + u\right)^2} + \frac{2}{3}\frac{ak_0u^2\left(3\,x_c + 2\,u\right)}{\left(x_c + u\right)^2}$$

$$-\frac{2}{3}\frac{{k_0}^2a^2\left(Rx_ct - ak_0\right)u^2\left(3\,x_c + u\right)}{t^2{x_c}^3\left(x_c + u\right)^3}$$

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 $\delta$-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

$$u(t,x) = \frac{1}{(3t)^{\frac{1}{3}}} \mathrm{Ai}(\frac{x}{(3t)^{\frac{1}{3}}})$$

   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 $\xi \to -\infty$.

# 45 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:

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

$$(x,y) \xrightarrow{d} d(x,y) = |x-y|, \tag{45.1}$$

een afbeelding[1] $d$ van $X \times X$ naar $[0,\infty)$ met de eigenschappen dat voor alle $x,y,z \in X$ geldt dat

$$(i) \quad d(x,y) = 0 \iff x = y; \quad (ii) \quad d(x,y) = d(y,x);$$

$$(iii) \quad d(x,y) \leq d(x,z) + d(z,x). \tag{45.2}$$

Iedere niet-lege deelverzameling $A$ van $X$ is zo een metrische ruimte, waarbij we de algebraische vectorruimte operaties nu vergeten.

**Definition 45.1.** *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 (45.2) hierboven gelden voor alle $x,y,z \in X$.*

**Exercise 45.2.** De $\varepsilon, N$-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 45.3.** 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.

[1]De $d$ van distance, $a$ van afstand doen we maar niet.

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 $a \in X$.

**Definition 45.4.** *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 45.5.** 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 45.6.** *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 45.7.** Bewijs dat rijkompakte metrische ruimten volledig zijn.

**Exercise 45.8.** 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 45.9.** 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.

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 45.10.** 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 45.9 en Opgave 45.10 volgt dat de stelling over maxima en minima van continue functies op gesloten begrensde deelverzamelingen van $\mathbb{R}^{\mathrm{N}}$.

**Theorem 45.11.** *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 45.12.** 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 45.13.** 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 45.14.** 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 45.12 leidt dat tot deze karakterisatie van equivalente normen op $X$.

**Exercise 45.15.** 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 45.16.** 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 45.17.** *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 45.18.** Wel een kluifje: bewijs Stelling 45.17. 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 45.5.

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 45.19.** 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 45.15 en de karakterisatie daaronder en boven van open verzamelingen met behulp bollen, gesloten of open, zoals $B_\varepsilon^p(\xi) = \{x \in \mathbb{R}^2 : |x - x_0|_p < \varepsilon\}$ met $\xi \in \mathbb{R}^2$ en $\varepsilon > 0$.

**Exercise 45.20.** De bollen $B^1$ en $B^\infty$ 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 45.21.** 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 45.20, 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 45.22.** 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 45.13. 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 45.13. 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 45.23.** 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 45.24.** 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 45.25.** 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 45.26.** 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 $x \to \max(|x_1|, |x_2|)$.

# 46 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.

## 46.1 Punten en vectoren in het platte vlak

**Exercise 46.1.** 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 46.2.** 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 46.3.** 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

$$a_1 x_1 + a_2 x_2 = c$$

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].

## 46.2 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 46.1. 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 46.4.** Wat is de kortste afstand tussen $\underline{a} = (1, 1)$ en de lijn met vergelijking $3x_1 + x_2 = 1$?

**Exercise 46.5.** 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 46.6.** Reflecteer[11] op wat het begrip loodrecht met het begrip afstand te maken heeft.

**Exercise 46.7.** 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 46.8.** Een deelverzameling $K \subset \mathbb{R}^2$ heet convex als met elk tweetal punten $\underline{a}$ en $\underline{b}$ in $K$ ook het lijnstuk

$$\{t\underline{a} + (1-t)\underline{b} : \, 0 \le t \le 1\}$$

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.

## 46.3 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 46.9.** Bewijs dat $|\vec{a} \cdot \vec{b}| \leq |\vec{a}||\vec{b}|$, met andere woorden, dat

$$(a_1b_1 + a_2b_2)^2 \leq (a_1^2 + a_2^2)(b_1^2 + b_2^2).$$

Hint: breng alles naar de rechterkant, doe de algebra en herken het kwadraat. Doe vervolgens ook

$$(a_1b_1 + a_2b_2 + a_3b_3)^2 \leq (a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2),$$

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 46.10.** 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 46.11.** Leid met Opgave 46.9 en Opgave 46.10 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 46.12.** 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 46.13.** 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?

## 46.4 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 46.14.** 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 46.14 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 46.15.** $A \subset \mathbb{R}^2$ 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 46.16.** 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 46.17.** 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 46.18.** 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 46.16 nu een gelijkheid voor alle $\underline{x} \in l$? Gebruik dit en reken $P_l(\underline{b})$ gewoon uit voor gegeven $\underline{b}$.

**Exercise 46.19.** 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.

## 46.5 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 46.20.** 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]

$$B(\vec{x}, \vec{y}) = \sum_{i,j=1}^{2} a^{ij} x_i y_j,$$

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 46.21.** 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 46.22.** 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 46.23.** Bewijs direct, dus zonder cosinussen en sinussen, dat voor elke symmetrische (hier nog twee bij twee) matrix $A$ geldt dat het maximum $\mu$ 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 $\mu$ of bij $-\mu$ (of bij allebei in bijzonder gevallen).

**Exercise 46.24.** De eigenvector in Opgave 46.23 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 46.19 toe[22] op $l = l_1$ en noem $m = l_2$. Laat zien dat $A$ deze $l_2$ op zichzelf afbeeldt.

## 46.6 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 $\lambda_1$ en $\lambda_2$,

$$A\vec{v}_1 = \lambda_1\vec{v}_1 \quad \text{and} \quad A\vec{v}_2 = \lambda_2\vec{v}_2.$$

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 46.16. Niet benadrukt nog is wat de essentie was van het bewijs dat je in Opgave 46.23 hebt gegeven. Waar het resultaat in Opgave 46.16 via Opgave 46.14 en een convergente minimaliserende rij tot stand kwam, is in Opgave 46.23 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 46.19. 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?

## 46.7 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 $2\pi$ worden gerekend. En $2\pi$ 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 46.25.** 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 46.26.** 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 $z = x + iy$.

# 47 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 46 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 $2\pi$-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 $2\pi$-periodieke functies $a$ en $x$, (goed) gedefinieerd voor continue functies als gewone Riemann integraal[4].

**Exercise 47.1.** Voor $n = 1, 2, 3, \ldots$ zijn de $2\pi$-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 47.2.** Er is nog een functie die loodrecht staat op al deze cosinussen en sinussen. Welke functie?

## 47.1 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.

## 47.2 Symmetrische matrices

Net als in de twee-dimensionale context heeft iedere symmetrische $n \times n$ matrix

$$A = (a^{ij})_{i,j=1,\dots,n}$$

(een basis van) eigenvectoren $\vec{v}_1, \dots, \vec{v}_n \in \mathbb{R}^n$ 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

$$\lambda_1, \dots, \lambda_n,$$

die gemaakt worden door Opgave 46.23 herhaald toe te passen. Dit geeft in ieder stap zowel een nieuwe $\lambda_k$ 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 $k = 1, \dots, n$.

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 $x_1, x_2, x_3, \dots$ 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.

## 47.3 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 47.3.** 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 47.4.** 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 46.14. 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 46.9.

**Exercise 47.5.** 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 47.4 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 47.6.** 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 47.6. 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 $\phi_y$ 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 $\phi_y$, 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 47.7.** 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 47.8.** 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 $\phi_y$ 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 $\Phi$ is bevat in $H^*$, de (genormeerde) ruimte van Lipschitz continue *lineaire* functies $f : H \to \mathbb{R}$, en $\Phi$ is zelf weer lineair[19]:

**Exercise 47.9.** 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 $\Phi$ surjectief is: is elke $f \in H^*$ van de vorm $\phi_y$? 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 47.10.** Bewijs dat $N_f \subset H$ een gesloten lineaire deelruimte is.

In het bijzonder bestaat nu dankzij Opgave 47.5 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 47.11.** 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 $\phi_e$: $f(x) = f(e)e \cdot x$.

**Exercise 47.12.** Leg uit waarom met het resultaat in Opgave 47.10 de afbeelding $\Phi : H \to H^*$ een lineaire isometrie is.

Lineaire isometriën zijn de mooiste continue afbeeldingen die er bestaan. De inverse van $\Phi$ 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 $\Phi$ 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 47.13.** Gebruik Opgave 47.4 om aan te tonen dat er plenty niet-lineaire functies in $Lip_0(H)$ zijn.

## 47.4 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 LaTe$\chi$ 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 47.14.** 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 47.15.** 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 $2\pi$, 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 $a_1, a_2, a_2, \dots$ 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 $\dots$, enzovoorts. Dit produceert een rij $e_1, e_2, e_3, \dots$ van vectoren waarvoor

$$(e_i, e_j) = \delta^{ij},$$

en deze vectoren spannen een lineaire deelruimte op in $H$.

**Exercise 47.16.** 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 $l^{(2)}$ geidentificeerd *kan* worden.

[28]Indien gewenst.

[29]Nou ja, als er geen dubbelen in de rij voorkomen dan...

# 48 Fourier series: inner product approach

This is from a slow set of notes for real Fourier series only. The series in (28.1) 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 48.1.** Let $f$ be an integrable[1] $2\pi$-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 48.2.** Derive the equality in (28.12) by writing

$$\sin \frac{x}{2} + \cdots + \sin \frac{(N+1)x}{2}$$

[1] 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 $2\pi$. 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 $\delta$ this upper bound is small when $N$ is large. Note that $F_{_N}(x)$ is even and $2\pi$-periodic. Make plots of $F_{_N}$ for some values of $N$.

**Exercise 48.3.** Let $f$ be $2\pi$-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 48.4.** Let $f$ be $2\pi$-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 48.5.** 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 48.3 shows that in the space of continuous functions $2\pi$-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 48.6.** 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 48.7.** Use Exercise 48.6 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 (28.4) 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 48.8.** Let $f : (0, \pi) \to \mathbb{R}$ be given by $f(x) = 1$ and choose a $2\pi$-periodic even extension $f : \mathbb{R} \to \mathbb{R}$. Determine all Fourier coefficients $a_n$ and $b_n$.

**Exercise 48.9.** Let $f : (0, \pi) \to \mathbb{R}$ be given by $f(x) = 1$ and choose a $2\pi$-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 48.10.** Let $f : (0, \pi) \to \mathbb{R}$ be given by $f(x) = \sin x$. Choose an even $2\pi$-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 48.11.** Let $f : (0, \pi) \to \mathbb{R}$ be given by $f(x) = \cos x$ and choose an odd $2\pi$-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 48.9.
3. Now take the odd $2\pi$-periodic extension of $f(x) = 1 - \cos x$ (the difference of the function in Exercise 48.9 and the function here). Investigate numerically the behaviour of $S_N f$ near $x = 0$ for large $N$.

**Exercise 48.12.** Let $f : (0, \pi) \to \mathbb{R}$ be given by $f(x) = \pi - x$ and choose an odd $2\pi$-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 48.13.** Let $f : (0, \pi) \to \mathbb{R}$ be given by $f(x) = x(\pi - x)$ and choose an odd $2\pi$-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 $2\pi$-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

$$f \cdot g = \int_{-\pi}^{\pi} f(x) g(x)\, dx \tag{48.1}$$

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

$$|f|_2 = \sqrt{f \cdot f}, \tag{48.2}$$

the length of $f$ considered as a vector. Pythagoras could have told us told us that

$$f \cdot g = 0 \quad \Rightarrow \quad |f + g|_2^2 = |f|_2^2 + |g|_2^2. \tag{48.3}$$

Below we write

$$S_N g(x) = \frac{c_0}{2} + \sum_{k=1}^{N} (c_k \cos kx + d_k \sin kx). \tag{48.4}$$

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 48.14.** 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 $0 \leq \int_{-\pi}^{\pi} (f(x) - g(x))^2\, dx = \dots$.
2. Prove the Cauchy-Schwartz inequality. Hint: apply 1 to $f(x)/|f|_2$ and $g(x)/|g|_2$.
3. Prove that
$$|f + g|_2 \leq |f|_2 + |g|_2. \tag{48.5}$$

**Exercise 48.15.** 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 48.16.** 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 48.17.** Define $R_{\scriptscriptstyle N} f = f - S_{\scriptscriptstyle N} f$ and, with

$$\sigma_{\scriptscriptstyle N} f = \frac{1}{N+1}(S_0 f + S_1 f + \cdots S_{\scriptscriptstyle N} f),$$

let $\rho_{\scriptscriptstyle N} f = f - \sigma_{\scriptscriptstyle N} f$.

1. Show that $R_{\scriptscriptstyle N} f \cdot S_{\scriptscriptstyle N} f = 0$.

2. Show that $R_{\scriptscriptstyle N} f \cdot \sigma_{\scriptscriptstyle N} f = 0$.

3. Show that
$$|S_{\scriptscriptstyle N} f|_2^2 + |R_{\scriptscriptstyle N} f|_2^2 = |f|_2^2,$$
whence $|S_{\scriptscriptstyle N} f|_2 \leq |f|_2$ and (Bessel's inequality)
$$\frac{1}{2}a_0^2 + \sum_{k=1}^{N}(a_k^2 + b_k^2) \leq \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)^2\,dx. \tag{48.6}$$

4. Show that
$$|R_{\scriptscriptstyle N} f|_2^2 + |\sigma_{\scriptscriptstyle N} f - S_{\scriptscriptstyle N} f|_2^2 = |\rho_{\scriptscriptstyle N} f|_2^2.$$

5. In Exercise 48.3 we showed for $f$ continuous and $2\pi$-periodic that $\sigma_{\scriptscriptstyle N} f \to f$ uniformly on $\mathbb{R}$ as $N \to \infty$. Prove that then also $|R_{\scriptscriptstyle N} f|_2 \to 0$, so that (Parceval equality)
$$\frac{1}{2}a_0^2 + \sum_{k=1}^{\infty}(a_k^2 + b_k^2) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)^2\,dx. \tag{48.7}$$
Hint: use part 4.

6. Show that
$$\begin{aligned} f \cdot g &= (S_{\scriptscriptstyle N} f + R_{\scriptscriptstyle N} f) \cdot (S_{\scriptscriptstyle N} g + R_{\scriptscriptstyle N} g) \\ &= S_{\scriptscriptstyle N} f \cdot S_{\scriptscriptstyle N} g + R_{\scriptscriptstyle N} f \cdot R_{\scriptscriptstyle N} g. \end{aligned}$$

7. For $f$ and $g$ continuous and $2\pi$-periodic show that
$$\frac{1}{2}a_0 c_0 + \sum_{k=1}^{\infty}(a_k c_k + b_k d_k) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)g(x)\,dx = \frac{1}{\pi}\, f \cdot g. \tag{48.8}$$
Hint: use part 6, Exercise 48.16 and apply the Cauchy-Schwartz inequality to $R_{\scriptscriptstyle N} f \cdot R_{\scriptscriptstyle N} g$.

**Exercise 48.18.** We want to show that Parceval's equality (48.7) holds for $2\pi$-periodic peiceswise continuous functions. Let $f : \mathbb{R} \to \mathbb{R}$ be such a function. Show that there exists a sequence of $2\pi$-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 48.19.** Prove (48.7) 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 (48.5) 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.

A direct construction[2] 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 $\mathbb{R}$.

[2]Choices to be made in relation to Section 31.6.

# 49 Fourierreeksen

Het ligt voor de hand om de abstracte constructie van $H$ uit $C(\mathrm{I\!R}_{2\pi})$ te zien als gebeurende in het platte vlak $\mathrm{I\!R}^2$, waarbij de grafiek $G$ van een functie $f : \mathrm{I\!R}_{2\pi} \to \mathrm{I\!R}$ 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 $C(\mathrm{I\!R}_{2\pi})$ moeten hetzelfde element uit de te maken $H$ zijn als geldt dat

$$|f_n - g_n| \to 0$$

voor $n \to \infty$. Opgave 49.10 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 49.1.** Maak Opgave 49.10 nog een keer maar anders. Teken de grafieken van een rij functies $\tilde{Z}_n \in C(\mathrm{I\!R}_{2\pi})$ 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 $n \in \mathrm{I\!N}$ geldt dat $\tilde{Z}_n(0) = 1$.

**Exercise 49.2.** Maak Opgave 49.1 maar nu met $\tilde{Z}_n(0) = 0$.

**Exercise 49.3.** Maak Opgave 49.2 maar nu met $\tilde{Z}_n(0) = 2$.

**Exercise 49.4.** Maak Opgave 49.2 maar nu met $\tilde{Z}_n(0) = n$.

**Exercise 49.5.** Laat in Opgaven 49.1,49.2,49.3,49.4 hierboven zien dat $|Z_n - \tilde{Z}_n| \to 0$ als $n \to \infty$, waarbij $Z_n$ is als in Opgave 49.1.

[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 49.6.** 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 49.6, 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 49.7.** 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

$$\sum_{n=1}^{\infty} a_n c_n + \sum_{n=1}^{\infty} b_n s_n \tag{49.1}$$

kunnen nemen, met coefficiënten als in Opgave 49.7, 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 49.8.** 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 $C(\mathbb{R}_{2\pi})$.

## 49.1 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 49.10. 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 $2\pi$ 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 $\pi$ 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 $\mathit{Ka}$ uitbreiden met $\mathit{Ka}(2\pi n) = \frac{\pi}{2}$ in de gehele veelvouden $2\pi n$ van $2\pi$ ($n \in \mathbb{Z}$). Al deze functies zijn instructief als voorbeeld bij de vraag of ze te schrijven zijn als een oneindige som van de vorm (49.1). Met name de zaagtand is een bron van leerzaam vermaak zoals we zullen zien.

**Exercise 49.9.** 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

$$f = a_0 \frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} a_n \frac{c_n}{\sqrt{\pi}} + \sum_{n=1}^{\infty} b_n \frac{s_n}{\sqrt{\pi}} \tag{49.2}$$

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 (49.2) 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 (49.2) maken.

## 49.2 Functies op de cirkel

Als we afspreken dat twee getallen in $\mathbb{R}$ eigenlijk hetzelfde zijn als ze een geheel veelvoud van $2\pi$ verschillen dan maken $\mathbb{R}$ en $2\pi$ 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 $\pi$ is hier een positief reëel getal, waarvoor we op enig moment de $\pi$ 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 $2\pi$-periodiek zijn kunnen we ook opvatten als functies $f : \mathbb{R}_{2\pi} \to \mathbb{R}$, en omgekeerd. De verzameling van continue $2\pi$-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 49.10.** 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 $C_0(\mathbb{R}_{2\pi})$, 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 $C_0(\mathbb{R}_{2\pi})$ 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 $C_0(\mathbb{R}_{2\pi})$ 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?

## 49.3 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 30.31 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 $2\pi$-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 $C_0$ is alleen voor nu even.

[9] Waarom niet?

**Exercise 49.11.** 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 49.12.** Zijn er functies $f$ in $L^2(-\pi,\pi)$ waarvoor aan $f(0)$ geen betekenis[10] kan worden gegeven?

**Exercise 49.13.** Er is maar één $2\pi$-periodieke oneven[11] functie $\mathit{Ka}$ die voldoet aan $\mathit{Ka}(x) = 1$ voor $0 < x < \pi$. Schets de grafiek van $\mathit{Ka}$ en maak een rij $2\pi$-periodieke oneven continue functies $\mathit{Ka}_1, \mathit{Ka}_2 \dots$ waarvoor geldt dat $|\mathit{Ka}_n - \mathit{Ka}|_2 \to 0$ als $n \to \infty$. Hint: schets eerst de grafieken van $\mathit{Ka}_n$.

**Exercise 49.14.** Bewijs dat een oneven $2\pi$-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 49.7. 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] $\mathit{Ka}(-x) = -\mathit{Ka}(x)$ voor alle $x \in \mathbb{R}$.

**Exercise 49.15.** Verifieer dat zo'n $F$ een begrensde ($2\pi$-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 49.16.** 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 $2\pi$-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 46.22 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 $x$.

## 49.4 Blipfuncties

Het formulevoorschrift

$$x \xrightarrow{\mathit{blip}} \begin{cases} \exp(-\frac{1}{x}) & \text{for } x > 0 \\ 0 & \text{for } x \le 0, \end{cases}$$

definieert de functie $\mathit{blip} : \mathbb{R} \to [0,1)$ die met goed recht zowel oogverblindend mooi als gruwelijk lelijk genoemd[15] mag worden.

**Exercise 49.17.** Schets de grafiek van *blip* en onderzoek het gedrag van $\mathit{blip}'(x)$ als $x \downarrow 0$. En van $\mathit{blip}''(x)$. En van alle afgeleiden van *blip*. Concludeer dat *alle* afgeleiden van *blip* als continue functies van $\mathbb{R}$ naar $\mathbb{R}$ bestaan!

**Exercise 49.18.** Je kunt *blip* ook schalen. Definieer $\mathit{blip}_n$ door

$$\mathit{blip}_n(x) = \mathit{blip}(nx) = \exp(-\frac{1}{nx})$$

en bepaal $\lim_{n\to\infty} \mathit{blip}_n(x)$ voor elke $x \in \mathbb{R}$. De limietfunctie heet de Heaviside[16] functie, hier genoteerd als $\mathit{He}(x)$. 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 $\mathit{He}(0)$ als limietwaarde van $\mathit{blip}_n(0)$ is $0$, maar net zo vaak wordt $\mathit{He}(0) = \frac{1}{2}$ of $\mathit{He}(0) = 1$ genomen. Of zelfs $\mathit{He}(0) = [0, 1]$.

**Exercise 49.19.** De functies *blip* en *He* zitten niet in $L^2(\mathbb{R})$. Waarom niet? Maar $\mathit{He} - \mathit{blip}$ wel. Waarom? En $\mathit{He} - \mathit{blip}_n$ ook. De affiene ruimte

$$\mathit{He} + L^2(\mathbb{R}) = \{f = \mathit{He} + g : g \in L^2(\mathbb{R})\}$$

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 $d(\mathit{blip}_n, \mathit{He}) \to 0$ als $n \to \infty$.

**Exercise 49.20.** Definieer de functies $\mathit{blok}_n$ door

$$\mathit{blok}_n i(x) = \mathit{blip}_n(x)\mathit{blip}_n(\pi - x)$$

en laat zien dat

$$\lim_{n\to\infty} \mathit{blok}_n(x) = \chi_{(0,\pi)}(x)$$

voor alle $x \in \mathbb{R}$. Waarom geldt dat $\mathit{blok}_n \to \chi_{(0,\pi)}$ in 2-norm?

**Exercise 49.21.** Dezelfde vragen als in Opgave 49.20 maar nu voor $\mathit{blok}_n$ gedefinieerd door

$$\mathit{blok}_n i(x) = \mathit{blip}_n(x - \frac{1}{n})\mathit{blip}_n(\pi - \frac{1}{n} - x),$$

Opgave 49.21 laat zien dat $\chi_{(0,\pi)}$, opgevat als

**Exercise 49.22.** 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.

## 49.5 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 $x = (x_1, \ldots, x_n) \in \mathbb{R}^N$.

**Exercise 49.23.** 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 $\mathbb{R}^N$. Hint: leg eerst uit waarom het *geen* beperking is om aan te nemen dat $|x|_p = |y|_q = 1$.

**Exercise 49.24.** Bewijs dat $x \to |x|_p$ een norm is op $\mathbb{R}^N$.

**Exercise 49.25.** Bewijs dat $|x|_p \to |x|_\infty$ als $p \to \infty$.

**Exercise 49.26.** 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 $C[-\pi, \pi]$.

# 50 Welke fundamenten?

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. Als het onderstaande goed leesbaar is dan kun je rustig op weg met wat er verder komt in dit boek. Zo niet, dan zou het *groene* boekje met Ronald Meester[1] je wat op weg kunnen helpen. 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

$$k + \sum_{n=1}^{\infty} \frac{d_n}{10^n} \tag{50.1}$$

is. In (50.1) 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[2] 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[3] 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 $\Sigma$ leidt tot controverses, alsmede het gebruik van het symbool $\infty$ 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”.

[1] vuuniversitypress.com/15-voor-auteurs/overige-content/108-wiskunde-in-je-vingers

[2] NB, 0 is niet positief, $\mathbb{N} = \{n \in \mathbb{Z} : n > 0\}$, $\mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$.

[3] Ontbind teller en noemer in priemfactoren en streep gemeenschappelijke factoren weg.

Het gebruik van het symbool $\infty$ is wellicht te vermijden door

$$\sum_{n\in\mathbb{N}} \quad \text{in plaats van} \quad \sum_{n=1}^{\infty}$$

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*[4]. In het derde hoofdstuk komen we hier nog op terug.

Tenslotte merken we op dat de schrijfwijze in (50.1) niet altijd uniek is omdat getallen van de vorm

$$k + \sum_{n=1}^{m} \frac{d_n}{10^n} \tag{50.2}$$

nu eenmaal twee representaties hebben, bijvoorbeeld

$$1 = \frac{9}{10} + \frac{9}{100} + \frac{9}{1000} + \frac{9}{10000} + \frac{9}{100000} + \cdots = \sum_{n=1}^{\infty} \frac{9}{10^n}, \tag{50.3}$$

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?

## 50.1 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[5] subsectie kan overgeslagen worden bij eerste lezing[6], maar een opmerking van Jan

[4]Rekenen met $\infty$ doen wij hier niet.
[5]We gaan niet recht op een doel af nu.
[6]En ook bij tweede lezing.

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 (50.1)? Een rekenvraag die we nu rekenend en redenerend willen beantwoorden zonder over verzamelingen van getallen te spreken.

Terzijde, helemaal consequent is de schrijfwijze in (50.1) niet. De $k$ is duidelijk anders dan de rest van de termen in deze aftelbare som[7]. Als we ons beperken tot de positieve reële getallen, die we als verzameling[8] gezien aanduiden met $\mathbb{R}^+$, dan is het eleganter om het "Romeinse" perspectief van eenheden, tientallen, tienden, honderdtallen, honderdsten, 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

$$8765{,}4321 \tag{50.4}$$

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 (50.4) hoeven verder geen misverstanden te bestaan. Links van de komma tellen de decimalen van rechts af de 1-tallen, 10-tallen, $10 \times 10$-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

$$8765{,}432143214321 \quad \text{of} \quad 8765{,}43219999999999999999999999,$$

[7]Som van een eindeloze rij, zoals Marjolein Kool dat zo mooi noemt.

[8]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 (50.4) waarbij er voor de komma alleen maar 0-en staan, en achter de komma alleen maar 1-en, zonder dat het rechts afbreekt[9].

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

$$0{,}0101010101010101010101010101\ldots \tag{50.5}$$

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 (50.5) 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,

[9]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, 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[10] 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[11] terecht als een minder gelukkige en te mijden conventie gezien. Afbrekende kommagetallen kunnen wellicht 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 (50.5) achtennegentig keer bij zichzelf op dan zien we dat 99 keer $\frac{1}{99}$ niet alleen gelijk is aan 1 maar ook aan

$$0{,}\underline{9} = 0{,}9999999999999999999999999\ldots,$$

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.

[10] In het PO heeft de onhandige methode veelal de voorkeur gekregen.

[11] Waar de laatste decimaal meestal met meetnauwkeurigheid te maken heeft.

### 50.1.1 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[12] kan worden, zoals bijvoorbeeld in

$$\begin{array}{rl} 0{,}9999999 & \\ 0{,}9999999 & \\ \hline & + \\ 1{,}8000000 & \\ 0{,}1800000 & \\ 0{,}0180000 & \\ 0{,}0018000 & \\ 0{,}0001800 & \\ 0{,}0000180 & \\ 0{,}0000018 & \\ \hline & + \\ 1{,}9999998 & \end{array}$$

Immers, met doorlopende negens gaat dit onhandig *cijferen* precies hetzelfde en duurt nauwelijks langer dan hierboven:

$$\begin{array}{rl} 0{,}99999\ldots & \\ 0{,}99999\ldots & \\ \hline & + \\ 1{,}80000\ldots & \\ 0{,}18000\ldots & \\ 0{,}01800\ldots & \\ 0{,}00180\ldots & \\ 0{,}000180.. & \\ 0{,}000018.. & \\ \ldots\ldots\ldots\ldots & \\ \hline & + \\ 1{,}99999\ldots & \end{array} \tag{50.6}$$

Gelukkig: $1 + 1 = 2!$[13] Weliswaar schendt de realistisch tussenstap hier wel de regel dat we rechts geen doorlopende 0-en mogen hebben, de uitkomst

[12] Onhandig cijferen wordt ook wel kolomrekenen genoemd.
[13] Lees: één en één is twee uitroepteken.

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[14] 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 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

$$\begin{array}{ll}
0{,}99999\ldots & \\
0{,}99999\ldots & \\
0{,}99999\ldots & \\
\hline & +\\
2{,}70000\ldots & \\
0{,}27000\ldots & \\
0{,}02700\ldots & \\
0{,}00270\ldots & \\
0{,}000270.. & \\
0{,}000027.. & \\
\ldots\ldots\ldots & \\
\hline & +\\
2{,}99999\ldots &
\end{array} \tag{50.7}$$

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

[14] Wel doen!

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[15] of we er eerst 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

$$\begin{array}{r} 99 \\ 88 \\ 77 \\ \text{—} + \\ 264 \end{array}$$

nu ook (juist wel handig) van rechts af per kolom uitgecijferd[16] 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.

[15] Lees: $a + b + c = (a + b) + c = c + (a + b)$ met alle gepermuteerde variaties.

[16] Ook kolomcijferen, maar wordt meestal mechanisch rekenen genoemd.

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

$$\begin{array}{ll}
0{,}909090909090909090... & \\
0{,}222222222222222222... & \\
\hline
 & - \\
0{,}686868686868686868... &
\end{array}$$

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 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:

$$\begin{array}{ll}
0{,}12345\ldots & \\
0{,}99999\ldots & \\
\hline
 & + \\
1{,}00000\ldots & \\
0{,}11000\ldots & \\
0{,}01200\ldots & \\
0{,}00130\ldots & \\
0{,}000140.. & \\
0{,}000015.. & \\
................. & \\
\hline
 & + \\
1{,}12345\ldots &
\end{array} \tag{50.8}$$

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 50.1.4.

### 50.1.2 Vermenigvuldigen?

Kunnen we ook vermenigvuldigen? Dit gemene[17] sommetje bijvoorbeeld?

$$\begin{array}{r} 0,999999.....\\ 0,999999.....\\ \hline \text{??????} \end{array} \times$$

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 (50.6) en (50.7) de uitwerking van de som symmetrisch in de bijdragen van de aparte termen was.

$$\begin{array}{l} 0,999999.....\\ 0,999999.....\\ \hline 0,810000.....\\ 0,081000.....\\ 0,081000.....\\ 0,008100.....\\ 0,008100.....\\ 0,008100.....\\ 0,000810.....\\ 0,000810.....\\ 0,000810.....\\ 0,000810.....\\ 0,0000810...\\ 0,0000810...\\ 0,0000810...\\ 0,0000810...\\ 0,0000810...\\ ......................\\ \hline \text{????????} \end{array} \times \tag{50.9}$$

[17]Denk nog niet meteen aan $0{,}\underline{6} \times 0{,}\underline{6}$.

Dat ziet er een stuk ingewikkelder uit dan (50.6). 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[18] 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:

$$00, 01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 12, 14, 15, 16, 18, 20, 24,$$

$$25, 27, 28, 30, 32, 35, 36, 40, 42, 45, 48, 49, 54, 56, 63, 64, 72, 81.$$

Met alle blokjes gelijk aan 81 gaat het in (50.9) om de som van de getallen in het schema dat begint met

$$\begin{array}{cccccc}
 & & \frac{81}{10^2} & & & \\
 & & \frac{81}{10^3} \quad \frac{81}{10^3} & & & \\
 & & \frac{81}{10^4} \quad \frac{81}{10^4} \quad \frac{81}{10^4} & & & \\
 & & \frac{81}{10^5} \quad \frac{81}{10^5} \quad \frac{81}{10^5} \quad \frac{81}{10^5} & & & \\
 & & \frac{81}{10^6} \quad \frac{81}{10^6} \quad \frac{81}{10^6} \quad \frac{81}{10^6} \quad \frac{81}{10^6} & & & \\
 & & \frac{81}{10^7} \quad \frac{81}{10^7} \quad \frac{81}{10^7} \quad \frac{81}{10^7} \quad \frac{81}{10^7} \quad \frac{81}{10^7} & & &
\end{array} \tag{50.10}$$

en naar beneden breder en breder doorloopt. Let wel, de volgorde waarin we cijferend optellen in (50.9) komt overeen met per regel optellen in (50.10) en leidt in de somnotatie tot

$$81 \times \sum_{n=1}^{\infty} \frac{n}{10^{n+1}} \tag{50.11}$$

[18]Het betreft $1 \times 1 = 1$, maar dat terzijde.

als maximale uitkomst (vast wel gelijk[19] aan 1) van een produkt van twee *nul komma (minstens) nog wat getallen.*

En met drie zulke getallen gaat het om maximaal

$$
\begin{array}{cccccccccc}
 & & & & \frac{729}{10^3} & & & & & \\
 & & & \frac{729}{10^4} & \frac{729}{10^4} & \frac{729}{10^4} & & & & \\
 & & \frac{729}{10^5} & \frac{729}{10^5} & \frac{729}{10^5} & \frac{729}{10^5} & \frac{729}{10^5} & \frac{729}{10^5} & & \\
\frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6}
\end{array}
\tag{50.12}
$$

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[20] 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 (50.9) 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 | ... |
| zien we | 1 | 3 | 5 | 7 | 9 | 11 | 13 | 15 | ... |

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 | ... |
| maximaal | $1 \times 9$ | $3 \times 9$ | $5 \times 9$ | $7 \times 9$ | 9 | $11 \times 9$ | $13 \times 9$ | $15 \times 9$ | ... |

voor de kolomsommen, hetgeen leidt tot de vraag of

$$9 \times (\frac{1}{10} + \frac{3}{10^2} + \frac{5}{10^3} + \frac{7}{10^3} + \frac{9}{10^4} + \cdots) = 9 \times \sum_{n=1}^{\infty} \frac{2n-1}{10^n}$$

[19]Equivalent met

$$\sum_{n=1}^{\infty} \frac{n}{10^n} = \frac{10}{81}, \quad \text{wat zou} \quad \sum_{n=1}^{\infty} \frac{n^2}{10^n} \quad \text{zijn? Zie verder.}$$

[20]Het betreft immers $1 \times 1 \times 1 = 1$.

een decimaal ontwikkelbaar getal definieert waar *elke* eindige som van termen in (50.9) niet boven kan komen, een vraag vergelijkbaar met de minder ruw afgeleide vraag over (50.11). Maar het moge duidelijk zijn dat we opnieuw afdwalen van de basisschoolstof waar het hier toch om zou moeten gaan[21].

Het *cijferen* geeft wellicht meer begrip. Onhandig kolomcijferend zien we in (50.9) 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

$$\begin{array}{cccccccccccccccccccc}
8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 & 9 & 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 & 9 \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 8 & 9 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 7 & 8 \\
 & & & & & & & & & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}$$

en alles loopt niet alleen naar rechts maar ook naar beneden door.

Opnieuw optellen per kolom geeft

$$\begin{array}{cccccccccccccccccccc}
9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 8 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 8 & 8 \\
 & & & & & & & & 1 & & & & & & & & & & 1 & 1
\end{array}$$

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[22]. 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 (50.12), als de breedte maar niet te snel toeneemt. Dat idee verkennen we in de volgende subsectie, waarin we opnieuw afdwalen van het cijferen.

### 50.1.3 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

$$\sum_{n=1}^{\infty} \frac{a_n}{10^n} \tag{50.13}$$

[21]Voor een analysecursus zijn dit vragen om te onthouden!

[22]Lees: $a \times b \times c = (a \times b) \times c = c \times (a \times b)$, weer met alle gepermuteerde variaties.

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 (50.13) te praten.

Een noodzakelijke voorwaarde is dat de eindige sommen

$$A_1 = \frac{a_1}{10}, \quad A_2 = \frac{a_1}{10} + \frac{a_2}{10^2}, \quad A_3 = \frac{a_1}{10} + \frac{a_2}{10^2} + \frac{a_3}{10^3}, \quad \dots \tag{50.14}$$

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 $0{,}d_1\underline{9}$ 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 $0{,}d_1\underline{9}$, $0{,}d_1d_2\underline{9}$, $0{,}d_1d_2d_3\underline{9}$, $\dots$, en voor de bijbehorende

$$D_1 = \frac{d_1}{10}, \quad D_2 = \frac{d_1}{10} + \frac{d_2}{10^2}, \quad D_3 = \frac{d_1}{10} + \frac{d_2}{10^2} + \frac{d_3}{10^3}, \quad \dots,$$

geldt dat

$$D_1 + \frac{1}{10}, \quad D_2 + \frac{1}{10^2}, \quad D_3 + \frac{1}{10^3}, \quad \dots$$

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 $n = 1, 2, 3, \dots$, 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

$$D_k < A_n < D_k + \frac{1}{10^k} \tag{50.15}$$

voor alle $n$ vanaf $n = n_k$, en dat zou moeten betekenen dat

$$A_n \to D = 0{,}d_1d_2d_3d_4 \dots, \tag{50.16}$$

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 $\alpha_1$. 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 $\alpha_2$, 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[23] we nu uit (50.15) 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 $\alpha_n$ 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 (50.13) definieert dus ondubbelzinnig een *nul komma minstens nog wat getal*, mits we weten dat alle eindige sommen in (50.14) kleiner zijn dan $0{,}\underline{9}$. Maar wie voor (50.10) en (50.12) meteen ziet dat dat inderdaad zo is mag het zeggen. We zijn er dus nog niet uit wat betreft produkten van positieve kommagetallen.

### 50.1.4 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

[23] Nog even nagaan dit dus.

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 $0{,}9 \times 0{,}\underline{9}$:

$$\begin{array}{l} 0{,}999999..... \\ 0{,}9 \\ \overline{\phantom{0{,}999999}} \times \\ 0{,}810000..... \\ 0{,}081000..... \\ 0{,}008100..... \\ 0{,}000810..... \\ 0{,}0000810... \\ ...................... \\ \overline{\phantom{0{,}999999}} \times \\ 0{,}899999..... \end{array}$$

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.

### 50.1.5 Produkten van kommagetallen

Als het bovenstaande eenmaal in in één keer lukt als

$$\begin{array}{l} 0{,}999999..... \\ 0{,}9 \\ \overline{\phantom{0{,}999999}} \times \\ 0{,}899999..... \end{array}$$

dan kan daarna

$$
\begin{array}{l}
0{,}999999.....\\
0{,}999999.....\\
\hline \times \\
0{,}899999.....\\
0{,}089999.....\\
0{,}008999.....\\
0{,}000899.....\\
0{,}000089.....\\
.....................
\end{array}
\tag{50.17}
$$

ook, en vervolgens kunnen we dan van boven af de kommagetallen term voor term optellen met wat we kolomcijferend geleerd hebben in sommetjes als (50.8).

De eerste stap is

$$
\begin{array}{l}
0{,}899999.....\\
0{,}089999.....\\
\hline + \\
0{,}899999.....\\
0{,}009999.....\\
0{,}080000.....\\
\hline + \\
0{,}909999.....\\
0{,}080000.....\\
\hline + \\
0{,}989999.....
\end{array}
\tag{50.18}
$$

In (50.18) 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 (50.17), waarbij in dit voorbeeld de 8 eentje opgeschoven is naar rechts.

Zo gaat dat verder. Nu we met (50.17) zijn gevorderd tot

$$
\begin{array}{l}
0,999999.....\\
0,999999.....\\
\rule{3cm}{0.4pt}\ \times\\
0,989999.....\\
0,008999.....\\
0,000899.....\\
0,000089.....\\
.....................
\end{array}
\tag{50.19}
$$

zien we dat het patroon zich herhaalt in

$$
\begin{array}{l}
0,989999.....\\
0,008999.....\\
\rule{3cm}{0.4pt}\ +\\
0,998999.....
\end{array}
$$

met als resultaat de som van de eerste drie kommagetallen in (50.17). De 8 is weer eentje opgeschoven en dat gaat zo door. In de volgende stap zien we

$$
\begin{array}{l}
0,999999.....\\
0,999999.....\\
\rule{3cm}{0.4pt}\ \times\\
0,998999.....\\
0,000899.....\\
0,000089.....\\
.....................
\end{array}
\tag{50.20}
$$

met nu boven de drie nullen na de komma in (50.20) 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.

## 50.2 Kleinste bovengrenzen

Net als de aftelbare som in (50.1) met $k \geq 0$ is (50.3) een mooi voorbeeld van

$$\sum_{n=0}^{\infty} a_n \tag{50.21}$$

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 (50.21), notatie

$$S = \sum_{n=0}^{\infty} a_n.$$

In het geval van (50.1) 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 (50.21) een eindig getal kan komen is zo vanuit (50.1) vanzelfsprekend, ook al dacht ene Zeno daar destijds anders over. Mooie voorbeelden waarbij er uit de som geen eindig getal komt zijn

$$S = \sum_{n=0}^{\infty} 1 \quad \text{met} \quad S_N = N, \quad \text{en} \quad S = \sum_{n=0}^{\infty} \frac{1}{n}. \tag{50.22}$$

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[24]. 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 (50.22), waarover we opmerken dat

$$1 + \underbrace{\frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{>\frac{1}{2}}}_{>1} + \underbrace{\underbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}_{>\frac{1}{2}} + \underbrace{\frac{1}{9} + \frac{1}{10} + \frac{1}{11} + \frac{1}{12} + \frac{1}{13} + \frac{1}{14} + \frac{1}{15} + \frac{1}{16}}_{>\frac{1}{2}}}_{>1},$$

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[25], 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[26] 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 50.3 te spreken.

[24] Overtuigd?

[25] Ik meen $\ln 2$.

[26] Kies positieve termen om boven $S$ te komen, dan negatieve om onder $S$, dan ...

## 50.3 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 (50.1) een voorbeeld geweest van (50.21) zonder de a priori informatie dat $a_n \geq 0$ maar *wel* met de eigenschap dat

$$\sum_{n=0}^{\infty} |a_n| < \infty, \tag{50.23}$$

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

$$S = \sum_{n=0}^{\infty} a_n, \tag{50.24}$$

en hernummeren van de som verandert niets aan die uitkomst. Reeksen van de vorm (50.21) waarvoor (50.23) 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

$$|S| = |\sum_{n=0}^{\infty} a_n| \leq \sum_{n=0}^{\infty} |a_n|. \tag{50.25}$$

Wat betreft het bewijs van (50.24) gegeven (50.23), de invariantie onder hernummeren en de aftelbare 3-hoeksongelijkheid (50.25): 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 [27] de "voor alle $p > 0$" nog door een

[27] 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 $\varepsilon$-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 (50.24) 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.

## 50.4 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[28], vanuit wat men de leer van verzamelingen noemt.

Deze verzamelingenleer is iets waarover Paul Halmos in zijn mooie boekje Naive Set Theory [29] schreef: alle wiskundigen vinden dat je er wat van

[28] Een vriendje van Einstein heeft helaas laten zien dat dat nooit bevredigend zal lukken.
[29] Vertaald ooit als Prisma pocket verkrijgbaar.

gezien 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[30] 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[31], 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[32].

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[33], dat er op neerkomt dat[34] als je voor $N = 1$ de uitspraak controleert, en je vervolgens laat zien dat de *implicatie*

$$(P_N) \implies (P_{N+1}) \tag{50.26}$$

geldt voor alle $N$ waarvoor je hem nodig hebt, namelijk om via herhaald toepassen van de inductiestap (50.26) 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 (50.26) 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.

[30]In LaTe$\chi$: $\emptyset$. Op het schoolbord liever $\phi$.
[31]Tellend is $\mathbb{N}$ met nul een beetje flauwe kul, op $0^0$ komen we nog terug.
[32]Gaat er dus eigenlijk om hoe je al die getallen zonder stippels tussen accolades vangt.
[33]Naamgeving volledig intimiderend, vriendelijker is: dominoprincipe.
[34]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

$$(P_1) \wedge (P_2) \wedge \cdots \wedge (P_N) \implies (P_{N+1}) \tag{50.27}$$

is[35].

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[36].

Du moment dat er over het bestaan van[37] $\mathbb{N}$ geen twijfel meer is, worden in de verzamelingsleer, $\mathbb{Z}$, $\mathbb{Q}$ en uiteindelijk $\mathbb{R}$ *wiskundig netjes* geconstrueerd. De constructie van $\mathbb{R}$ is in 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[38].

Het is instructief om de constructies van $\mathbb{Z}$ en $\mathbb{Q}$ uit $\mathbb{N}$ met elkaar te vergelijken. 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

$$q = \frac{t}{d}$$

[35] De $\wedge$ staat voor "en", dat is logisch. Denken aan dominosteentjes is nu lastiger.

[36] https://www.youtube.com/watch?v=DOzqUyW7jog

[37] Eventueel via Peano's axioma's.

[38] 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

$$\frac{t_1}{d_1} = \frac{t_2}{d_2} \quad \text{als} \quad t_1 d_2 = t_2 d_1.$$

## 50.5 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[39],

en daar moet je mee omgaan. Leuk is dat (ii) ons dan later[40] 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[41] 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[42] fraai, omdat ze afstand nemen

[39]Waarom is dat een paradox?
[40]Maar niet hier.
[41]Lees: in maatheoretische zin.
[42]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[43].

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[44] 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

$$[r_n] = \{q_n \in \mathbb{Q} : \, q_n \sim r_n\} \tag{50.28}$$

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[45] op de verzameling van alle reële rijen. Een voorbeeld is

$$n! \sim (\frac{n}{e})^n \sqrt{2\pi n},$$

uitvoerig besproken in [HM].

[43] In je hoofd of op de getallenlijn.

[44] Ook dat woord heeft een *wiskundige* definitie natuurlijk.

[45] Wel een goede vraag hierboven is: wat is de beste representant?

## 50.6 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[46] 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[47] 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[48].

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 hier 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

$$\text{de lijn door } (x, y) = (a, F(a) \text{ en } (x, y) = (b, F(b))$$

[46]Van het Latijnse fungor (deponens: een passieve vorm met actieve betekenis).

[47]Vaak slordig: de grafiek $y = f(x)$ in het $x, y$-vlak.

[48]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

afgeleide van $F'(x)$ van $F(x)$ 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[49] 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*[50] 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 in 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, niet alleen voor wie dat stuk heeft overgeslagen. 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[51] voor $n!$ als $n \to \infty$ een mooi voorbeeld[52] 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

[49] Denk ook aan valsspelen met meetwaarden.

[50] Geen sport zonder *techniek*.

[51] De voorbeeldformule met $\sim$ een paar pagina's terug, uit te spreken als "twiddles".

[52] En buitengewoon relevant voor probleemstellingen in de natuurkunde.

$\mathbb{C}$ naar $\mathbb{C}$ en afbeeldingen en functies met meerdere variabelen. Lineaire 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[53] ligt hier voor de hand.

Gewone differentiaalvergelijkingen, al aan de orde geweest in de context van machtreeksen, motiveren het opnemen van een hoofdstuk waarin we ook de calculus voor functies op en naar Banachruimten introduceren, 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. Dat doen we in één moeite door in combinatie de multiplicatorenmethode van Lagrange[54] 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[55]. De meer abstracte formulering van de methode van Lagrange is opgenomen for amuse-

[53]Door mijn medestudenten destijds ook wel Bananachalgebra's genoemd.

[54]De eerste stelling die ik ooit zelf aan anderen uitlegde, maar nu heel anders.

[55]Zie www.encyclo.nl/begrip/Submersie èn www.encyclo.nl/begrip/Immersie.

ment. Ook wat pittiger is de behandeling van tweede orde afgeleiden die we pas in abstracte setting in meer detail doen. Het hoofdresultaat is het Lemma van Morse, 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.

Natuurlijk behandelen we ook het gebruikelijke jargon met metrieken, omgevingen en open en gesloten verzamelingen samenvatten voor wie zich minder gelukkig voelt met informele uitdrukkingen als in de buurt van, zoals ook Adams die gebruikt in zijn nu vrijwel overal gebruikte calculus boek, dat door de theoretisch hoofdstukken in dit boek stevig wordt onderbouwd.