"Fundamentals of analysis"

and other stuff in various stages of preparation

Chapters 1-11: a first course in Analysis?
Chapter 6: linear continuous functions in general settings
Chapter 11: Morse lemma and quadratic functions
Chapters 12-13: from multivariate differential calculus to varieties.
Chapters 14: applications is systems biology.
Chapters 15-16: integral calculus and various applications (in Dutch)
Chapter 17: complex polygon integrals & functional calculus (in Dutch)
Chapters 18-20: Gauss, Green and Stokes, first analysis, then the algebra
Chapters 21-28: PDE/VM related stuff (bachelor and master courses)
Chapters 29-31: Nash Implicit Function Theorem
And other parts in Dutch written for different audiences

JH&FF, Oegstgeest, Amsterdam (2018)

ISBN

NUR

Nederlandse samenvatting Analyse 1

De eerste elf hoofdstukken (zeg maar Deel 1) betreffen alles dat met $y = f(x)$ te maken heeft. We beginnen met de vraag wat er gebeurt als je f als input-output machientje gebruikt. Dat begint bij het kleitablet hierboven en benaderingen van de wortel uit twee ongeveer 3700 jaar geleden, en de getallenrij van Heron tweeduizend jaar later: de facto een implicatie van de methode van Newton uit wat dan nu wel onze moderne tijd mag heten. De van school bekende begrippen als limieten, continuïteit en differentieerbaarheid komen hier op een natuurlijke manier langs.

Een andere startvraag, oppervlaktebepaling van gebieden in het platte vlak beschreven met formules als $y = f(x)$, leidt ons terug naar de tijd van Archimedes en de formule voor de inhoud van aan piramide met de voorfactor $\frac{1}{3}$, die mooie breuk waarmee in de referentiekaders rekenen stb-2010-265 niet meer wordt gerekend.

De gangbare niet puur meetkundige afleiding van deze formule leidt tot basale vragen over de getallen waarmee we onze wiskunde bedrijven, zoals ook de wortel uit twee dat al deed. De moderne integraalrekening is vervolgens op de antwoorden gebaseerd.

Het verband tussen differentiaalrekening en integraalrekening wordt eerst geïllustreerd met het voorbeeld van polynomen en machtreeksen, en wel zo algebraïsch mogelijk zonder limieten. Daarna komt pas de analytische aanpak die het verband in algemenere context precies maakt.

Het oplossen van vergelijkingen met twee variabelen x en y naar y met x als parameter leidt tot het begrip impliciete functie en een tweede herbezinning op het kleitablet, de rij van Heron en de methode van Newton. De impliciete functiestelling is het resultaat.

Alle definities en stellingen worden geformuleerd voor x en y gewone reële variabelen, maar zo verwoord dat overschrijven naar de situatie dat x in X en y in Y zit, wat X en Y dan later ook mogen zijn, zoveel mogelijk een copy-paste oefening is. Daarbij wordt $|x - y|$ vervangen door $d(x, y)$, spreek uit de afstand tussen x en y.

Het is een neiging van wiskundigen om getallen in verzamelingen te stoppen. In $y = f(x)$ zitten x en y doorgaans in $\mathbb{R}$, de verzameling van de reële getallen, waar dikke boeken over volgeschreven zijn. De f in $y = f(x)$ is een ander object. We zullen meer aandacht besteden aan verzamelingen waarvan de elementen f-jes zijn, en deze verzamelingen kunnen weer als een X of een Y een rol spelen in wat via copy-paste inmiddels als bouwwerk staat.

Opmerking 22.14 is aanleiding de mogelijkheid te verkennen om de Lesbesgue ruimten in te voeren zonder maatheorie. Zie Sectie 22.3 waarin de van de maatheorie en topologie bekende overdekkingen wel worden gebruikt, maar zonder het begrip kompaktheid, en zonder maten van nietnulverzamelingen.

Een intermezzo is het platte vlak als voorbeeld van een Hilbertruimte.

Een eerste toepassing van de abstracties is de toepassing van de inmiddels bewezen fundamentele stelling, geïntroduceerd in de context van het kleitablet, over oplossingen van vergelijkingen van de vorm $x = f(x)$, maar met x in X nu, op het oplossen van differentiaalvergelijkingen.

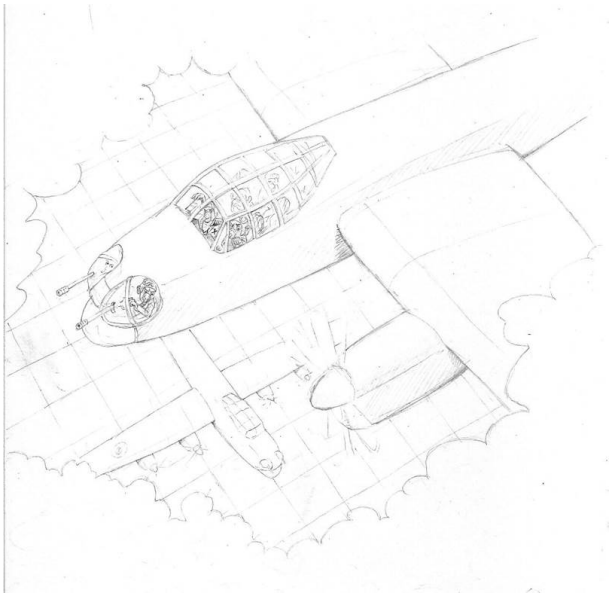
We besluiten het eerste deel van dit boek met Hoofdstuk 11, waarin een vraag wordt behandeld die in het voorbeeld dat x en y allebei in $\mathbb{R}$ zitten voor f eindelijk nooit zo expliciet gesteld wordt. Dat is namelijk de vraag of en onder welke voorwaarden een functie f met $f(0) = f'(0) = 0$ een kwadratische functie in vermomming is. Het antwoord op die vraag brengt ons op miraculeuze wijze terug tot het worteltrekken waar het hele verhaal met het kleitablet mee begonnen is.

Het tweede deel van het boek begint met Hoofdstuk 12. Uitpakken van wat we al weten na het eerste deel en interpreteren wat het is in andere concrete situaties.

Over en aan de rest van dit pdf pak kom ik nog te schrijven.

JH, 2018.

Contents



'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

This manuscript includes notes[1] for the analysis course for first year students of the Bachelor Mathematics at the VU. Topics covered in this course 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, to introduce some of the notation along the way, as well as a few basic principles. I assume that you have some familiarity with highschool calculus: limits, continuity, differentiability and integration in the context of real valued functions $f(x)$ of a real variable x. In particular you have probably seen the integration formula

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

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

Perhaps you have also seen the Newton scheme

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

for solving the equation $f(x) = 0$ numerically. The example $f(x) = x^2 - 2$ takes us way back to Babylonian times, and the origins of differential calculus, from the modern viewpoint if you like. It concerns $\sqrt{2}$, a geometric number which appears as the length of the diagonal in the unit square. This square root of 2 cannot be written as the quotient of two positive integers. The proof of this statement is not very hard but we will not include it here.

Theorem 1.1. *Let p and q be positive integers. Then $p^2 \neq 2q^2$.*

[1] Until Chapter 11.

1.1 One third of what?

Another geometric number is $\frac{1}{3}$, which appears as the volume V of a pyramid with unit square base and unit height. To see how and why we divide this pyramid into 10 horizontal layers of height $\frac{1}{10}$ and write n for 10. The maximal width of each layer varies from 1 at the bottom to $\frac{1}{10} = \frac{1}{n}$ at the top[2]. From top to bottom these maximal widths are

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

so 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 are

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

so V is larger than

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

Combining the two bounds we have

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

in which we don't really have to exhaust ourselves to take n different from 10 as large as we want.

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

$$0 < W - V < \bar{S}_n - \underline{S}_n = \frac{1}{n}.$$

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

[2] Picture it!

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.

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. Here z ranges from $z = 0$ at the bottom to $z = 1$ at the top of the pyramid. We recognise $V = \frac{1}{3}$ as the coefficient of n^3 in (C_n).

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: 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 also (C_1) via

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

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

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

This trick to prove (C_n) for all positive integers n is also called proof by induction. Think of the n^{th} statement (C_n) as being written on the n^{th} domino. Put all dominos in a never ending cue. Kick the first domino $(n = 1)$ over and watch. The statements still to be checked are the dominos still standing.

You may have noted that

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

which belongs to a family of integrals

$$I_1 = \int_0^1 x\,dx = \frac{1}{2}, \quad I_2 = \int_0^1 x^2\,dx = \frac{1}{3}, \quad I_3 = \int_0^1 x^3\,dx = \frac{1}{4}, \ldots,$$

expressions that you must have seen before for the area I_p of the region

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

in the xy-plane.

Archimedean type expressions for sums of powers can be used to show directly that this sequence continues like suggested, but the sum formulas for exponents p larger than 3 become a bit more cumbersome. The inequalities[3]

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

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

$$\underbrace{\frac{1}{n^{p+1}} \sum_{k=0}^{n-1} k^p}_{lower\ \ sum} < \frac{1}{p+1} < \underbrace{\frac{1}{n^{p+1}} \sum_{k=0}^{n} k^p}_{upper\ \ sum}$$

for lower and upper approximations of I_p. Since the lower and upper sums differ by $\frac{1}{n}$, Archimedes tells us again that

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

for every positive integer p.

[3] Proved in Section 5.1.

1.2 A comment

The positive integers form the set of natural numbers

$$\mathbb{N} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, \ldots\},$$

numbers you have probably learnt at the age of 3. The elegant reasoning above involves calculations and inequalities with fractions of such positive integers. These form the set

$$\mathbb{Q}_+ = \{\, \frac{p}{q} : \, p, q \in \mathbb{N} \,\}$$

of positive rational numbers. We agree that the number $\frac{p}{q}$ in $\mathbb{Q}_+$ does not change if we skip common factors in p and q, and we also agree that $\frac{p}{1} = p$.

It is a curious fact that such illuminating calculations have been widely outbanned in Dutch educational law, in which nondecimal fractions such as $\frac{1}{3}$ simply do not appear in examples of calculations that children should master in elementary school. Of course $\frac{1}{3}$ still exist as a the name one third for certain quantities like amounts of liquid in 33% bottles, but in calculations $\frac{1}{3}$ is often replaced by 0.33, or even 0.34 if for some arbitrary lack of no reason the fraction $\frac{2}{3}$ happens to be taken equal to 0.66. The number $\frac{1}{3}$ as Archimedes knew it has been forced out.

So far for the rational numbers and the origins of integral calculus. The other perhaps older geometric question we put forward above takes us out of $\mathbb{Q}_+$ and relates to differential calculus.

1.3 The square root of two

The first recorded attempt[4] 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 our decimal notation one of these number sequences is

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

and stands for[5]

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

[4] That I know of.

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

which is a hexagesimal approximation of

$$\frac{577}{408} = 1.414\underline{2156862745098039} \approx \sqrt{2} = 1.4142135\ldots,$$

and therebye a remarkably good approximation of the square root of 2.

This approximation is believed to have resulted from calculations employing the approximation

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

The right hand side is an expression in x which we now relate to the line defined by

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

in the Cartesian plane, while

$$y = \sqrt{1+x}$$

defines half of the parabola given by

$$1 + x = y^2.$$

Now refresh your highschool knowledge of tangent lines and the differential calculus: if F is the function defined by

$$F(x) = \sqrt{1+x},$$

then $\frac{1}{2}$ is the value of $F'(0)$, the derivative of F in $x = 0$.

Replacing x by $y - 1$ the above approximation is equivalent to

$$g(y) = \sqrt{y} \approx 1 + \frac{1}{2}(y-1) = \frac{1}{2} + \frac{y}{2},$$

which relates to $g'(1) = \frac{1}{2}$. Again the right hand side is "linear": y only appears with a multiplicative constant, which happens to be equal to the additive constant $\frac{1}{2}$,

Now let $r > 0$ be a possibly not so very good approximation of $\sqrt{2}$. Then the above approximation with $y = r^2$ gives

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

possibly a better approximation of $\sqrt{2}$. Starting with $r = 1$ for instance the new approximation of $\sqrt{2}$ is $\frac{3}{2}$, which is not that bad really.

Note that the coefficient $\frac{1}{2r}$ is the value of the derivative $g'(y)$ in $y = r^2$. You may recognise Newton's method if you consider the function f defined by

$$f(x) = x^2 - 2,$$

and the method is also known as Heron's method. Redoing the approximation with $r = \frac{3}{2}$ gives $\frac{17}{12}$, much better, and $r = \frac{17}{12}$ in turn gives

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

the Babylonian approximation on YBC7289.

Try to find this expansion with long division and the hexigesimal table of

$$408 = 360 + 48 = 06\,48$$

if you like[6]. Or use long division and the table of 17 to find the expansion of

$$\frac{12}{17}$$

in hexigesimal form, which you can add to

$$\frac{17}{24} = \frac{85}{120} = \frac{42}{60} + \frac{50}{60^2} = 00{,}42\,50$$

to conclude.

1.4 The Archimedean Principle

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

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

the set of all integers and the set of all rationals, the former being a subset of the latter, indicated by the use of the inclusion symbol: $\mathbb{Z} \subset \mathbb{Q}$. We think of $\mathbb{Z}$ as a bi-infinite sequence of marked points on a number line with no endpoints, the other numbers of $\mathbb{Q}$ lying 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 $0 < q < 1$.

[6] You need tables which run to 60 times the divisor.

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

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

Beginning with $\mathbb{N}$ these are all 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 such an enumeration of $\mathbb{R}$. Then $\mathbb{R}$ is completely covered by the intervals[7]

$$(x_1 - \frac{1}{4}, x_1 + \frac{1}{4}),\ (x_2 - \frac{1}{8}, x_1 + \frac{1}{8}),\ (x_3 - \frac{1}{16}, x_3 + \frac{1}{16}),\ (x_4 - \frac{1}{32}, x_4 + \frac{1}{32}),$$

$$(x_4 - \frac{1}{64}, x_4 + \frac{1}{64}),\ (x_4 - \frac{1}{128}, x_4 + \frac{1}{128}),\ (x_4 - \frac{1}{256}, x_4 + \frac{1}{256}),$$

et cetera. 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. This is an absurdity that we are not willing to accept, since the total length of the real number line should be larger than any positive number. Have we proved the following theorem?

Theorem 1.2. *The set* $\mathbb{R}$ *of real numbers is not enumerable. In other words,* $\mathbb{R}$ *is not a sequence.*

A more direct proof of Theorem 1.2 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

$$x_n = \sum_{j=1}^{\infty} \frac{d_{nj}}{10^j},$$

and put the digits[8] d_{nj} in a block

$$d_{11} \quad d_{12} \quad d_{13} \quad d_{14} \quad d_{15} \quad d_{16} \quad d_{17} \quad d_{18} \quad \ldots$$

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

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

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

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

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

cannot appear as any x_n, a contradiction.

To make such 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

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

whence

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

This is just like

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

which relates to binary representations of the real numbers.

The equalities in the above expressions relate to the Archimedean principle. For instance, the nonnegative difference between 1 and the last sum with infinitely many terms has to be smaller than every power of $\frac{1}{2}$. This difference is then smaller then every $\frac{1}{n}$ and thus equal to zero according to Archimedes. We honour Archimedes by stating his principle as a theorem in which we use the modern symbols $\forall$ and $\exists$.

Theorem 1.3. *The Archimedean Principle:*

$$\forall_{\varepsilon\in\mathbb{R}_+}\,\exists_{n\in\mathbb{N}}:\ \frac{1}{n}<\varepsilon.$$

The statement in this theorem is often written as

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

and pronounced as

for every real number $\varepsilon > 0$ there exists a positive integer n with $\frac{1}{n} < \varepsilon$.

One of our tasks in this course will be to prove and understand this principle.

1.5 The geometric series

See

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

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

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

and substituting $10 = n$ we "discover" that

$$\frac{1}{n}+\frac{1}{n^2}+\frac{1}{n^3}+\frac{1}{n^4}+\frac{1}{n^5}+\cdots=\frac{1}{n-1}. \tag{1.1}$$

It's easy to convince yourself why (1.1) 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. Then have a look at

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

before you read on.

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

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

is called a geometric series. Finite sums[9]

$$\sum_{n=0}^{N} x^n = 1 + x + x^2 + \cdots + x^N = \frac{1 - x^{N+1}}{1 - x} \tag{1.2}$$

lead to the conclusion that

[9] Check the formula!

Theorem 1.4. *For $x \in \mathbb{R}$ it holds that*

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \quad \text{if} \quad |x| < 1. \tag{1.3}$$

If you have been born with n fingers ($n > 1$) you are likely to discover (1.1) as a fact of every day arithmetic life, long before you eat pizza's or hear of Theorem 1.4.

The mathematical proof of Theorem 1.4 is based on algebra that justifies (1.2), and a limit argument that involves the norm or absolute value of x, denoted by $|x|$, and reasoning of the type

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

We think of the norm $|x|$ as the distance of x from 0.

In everyday algebra the number 0 is the neutral element for addition, and the number 1 is the neutral element for multiplication. The algebra in (1.2) also involves $x \in \mathbb{R}$, and in the right hand side we divide by $1 - x$. We can rewrite this right hand side as

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

the multiplicative inverse of $1 - x$ acting on $1 - x^{N+1}$.

Exercise 1.5. Use (1.2) to show for $n \in \mathbb{N}$ that

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

Normed algebra with addition, multiplication and norm estimates is not limited to $\mathbb{R}$, $+$, $\times$ and $|\ |$. It should be no surprise that conclusions from (1.2) can be drawn in a far more general setting. For instance, we may replace x by a square matrix A with real entries, and 1 by the identity matrix I of the same size as A. If the matrix A has the property that, for some suitable matrix norm, $|A| < 1$ implies that

$$A^n \to 0 \quad \text{as} \quad n \to \infty,$$

then

$$\sum_{n=0}^{\infty} A^n = (I - A)^{-1} \quad \text{provided} \quad |A| < 1, \tag{1.4}$$

a formula for the inverse matrix of $I - A$.

A possible suitable matrix norm is the square root of the sum off all the squared entries of A,

$$|A|_2 = \sqrt{A_{11}^2 + A_{12}^2 + A_{21}^2 + A_{22}^2} \quad \text{if} \quad A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

It has the property that

$$|AB|_2 \leq |A|_2 \, |B|_2 \quad \text{and} \quad |A + B|_2 \leq |A|_2 + |B|_2$$

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

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

Let's examine Heron's method in more detail. It is defined by the iterative scheme

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

starting from $x_0 = 1$, and produces numbers

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

and so on. Writing

$$\tilde{x} = f(x) = \frac{x}{2} + \frac{1}{x} \tag{2.1}$$

for $x > 0$ we think of (2.1) as an input-output relation defined by a formula $f(x)$ or function f, with input some freely chosen x, and output some other $\tilde{x}$, defined by (2.1). Every x_n is obtained as $\tilde{x}$ from every previous $x = x_{n-1}$, starting from a positive value x_0, e.g. $x_0 = 1$.

We note that

$$\tilde{x}^2 - 2 = (\frac{x}{2} + \frac{1}{x})^2 - 2 = (\frac{x}{2} - \frac{1}{x})^2,$$

and $\tilde{x}$ differs from x by

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

If $x_0 > 0$ has $x_0^2 \neq 2$ it easily follows that

$$x_n^2 > 2 \quad \text{and} \quad 0 < x_{n+1} < x_n \quad \text{for all} \quad n \in \mathbb{N}. \tag{2.2}$$

We view the numbers x_n as a sequence indexed by n. This sequence was designed by Heron to solve the equation

$$x^2 = 2, \tag{2.3}$$

most likely only for $x > 0$. The index set for n may be $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ or[1] just $\mathbb{N}$. We shall want to be able to conclude that

$$x_n \to \sqrt{2} \tag{2.4}$$

as n gets larger and larger running upwards through the set $\mathbb{N}$.

[1] The notation $\mathbb{N}_k = \{n \in \mathbb{N} : n \geq k\}$ for $k \in \mathbb{N}_0$ may be convenient.

The reasoning should be that

$$x_n = f(x_{n-1}) \to \bar{x} = f(\bar{x}) \quad \text{as} \quad n \to \infty \tag{2.5}$$

for some limiting value $\bar{x}$ which solves

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

a purposefully perverted[2] equivalent version of the equation $x^2 = 2$. We therefore have an urgent need for a meaningful statement of

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

and well as the statement that then also

$$f(x_n) \to f(\bar{x}).$$

2.1 Heron's sequence is bounded and decreasing

The idea behind Heron's sequence is that $\tilde{x}$ as defined by (2.1) may be better than x as far as approximately solving $x^2 - 2 = 0$ is concerned. We note that

$$\tilde{x}^2 - 2 = (\frac{x}{2} + \frac{1}{x})^2 - 2 = (\frac{x}{2} - \frac{1}{x})^2,$$

and that $\tilde{x}$ differs from x by

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

For $x_0 > 0$ with $x_0^2 \neq 2$ it then easily follows that

$$x_n^2 > 2 \quad \text{and} \quad 0 < x_{n+1} < x_n \quad \text{for all} \quad n \in \mathbb{N}. \tag{2.6}$$

Restricting the index set to $\mathbb{N}$, Heron's sequence has the property

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

a decreasing sequence of rational numbers bounded from below by a rather arbitrary bound $\frac{4}{3}$.

Exercise 2.1. Prove that $\frac{4}{3}$ is indeed a lower bound for the sequence, but that there are larger rational lower bounds.

The largest lower bound for this rational sequence is the irrational number $\sqrt{2}$ we are looking for. Why does this largest lower bound exist, and why is its square equal to 2?

[2] Check this out!

2.2 Bounded monotone sequences have limits!

Heron's sequence, indexed by $n \in \mathbb{N}$ is strictly decreasing and bounded. Sequences of numbers[3] x_n with either

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

will be called monotone sequences. There are two types of monotone sequences: nondecreasing and nonincreasing. If such a sequence is bounded we think of it as approximating a number, possibly a rational number. The sequence

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

for instance is bound to approximate the rational number 1. But most nondecreasing bounded sequences will define a number which is not rational.

Exercise 2.2. Show that there exist 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 with n digits with the property that $x_n^2 < 2$.

The idea is to add to $\mathbb{Q}$ all the lowest upper bounds of bounded nondecreasing sequences which do not approximate a rational number, and then automatically also the largest lower bounds of bounded nonincreasing sequences which do not approximate a rational number either. The resulting[4] set $\mathbb{R}$ has the property that it contains $\mathbb{Q}$, and is just like $\mathbb{Q}$ as far as the algebraic operations and ordering are concerned.

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

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

which is the unique limit of that sequence in terms of a definition that will follow shortly, and likewise for every nonincreasing bounded sequence and its largest lower bound (infimum)

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

Let's make these notions more precise.

[3] For the moment rational numbers.

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

Definition 2.3. *Let x_n be a sequence of numbers in $\mathbb{R}$ indexed by $n \in \mathbb{N}$. Then the sequence is called*

nondecreasing if

$$\forall_{n\in\mathbb{N}} : x_n \leq x_{n+1},$$

i.e. $x_n \leq x_{n+1}$ for every natural number n;

strictly increasing if

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

nonincreasing if

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

strictly decreasing if

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

bounded from above if

$$\exists_{M\in\mathbb{R}} \forall_{n\in\mathbb{N}} : x_n \leq M,$$

i.e. there exists a real number M such that for all natural numbers n

$$x_n \leq M;$$

bounded from above if[5]

$$\exists_{m\in\mathbb{R}} \forall_{n\in\mathbb{N}} : x_n \geq m;$$

bounded if it is both bounded from above and bounded from below.

The number M, if it exists, is called an upper bound, the number m, if it exists, is called a lower 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 with $M < S$. 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 with $m > L$.

Heron's sequence is a strictly decreasing bounded sequence, bounded from above by its first number $M = x_1 = 1$, bounded from below by $m = \frac{4}{3}$.

Theorem 2.4. *Every nonincreasing bounded sequence in $\mathbb{R}$ has an infimum in $\mathbb{R}$. Equivalently: every nondecreasing bounded sequence in $\mathbb{R}$ has a supremum in $\mathbb{R}$.*

We will not prove this theorem. It follows from every proper construction of $\mathbb{R}$, for instance via decimal expansions as in Exercise 2.2. Applied to Heron's sequence it gives us L, the largest lower bound of this decreasing sequence.

[5] m is a real number here.

2.3 The limit definition: epsilons

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

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

We thus conclude that

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N} \,:\, \underbrace{|x_n - L|}_{d(x_n,L)} < \varepsilon, \tag{2.7}$$

a statement to be pronounced as: for all (real) $\varepsilon > 0$ there exists a natural number N such that for all natural numbers n with $n \geq N$ it holds that the distance from x_n to L is smaller than ε.

Definition 2.5. *A sequence of real numbers x_n indexed by $n \in \mathbb{N}$ is called convergent if there exists $L \in \mathbb{R}$ such that*

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

If so we say that $x_n \to L$ as $n \to \infty$, or equivalently

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

L being called the limit of the sequence.

The distance from x_n to L, notation $d(x_n, L)$, is the absolute value of the difference. So

$$d(x_n, L) = |x_n - L| < \varepsilon$$

holds for all $n \geq N$, N depending on $\varepsilon > 0$. The Greek letters always stand for real numbers and the letters in the middle of the alphabet for integers, unless explicitly stated otherwise.

The statement in (2.7) makes sense for every real L and every real sequence[6], not just for monotone sequences. A sequence is called convergent if such an L exists and we don't have to call it L. Thus convergence of the sequence x_n means that

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

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

Remark 2.6. *The negation of (2.8) reads*

$$\forall_{\bar{x}\in\mathbb{R}}\,\exists_{\varepsilon>0}\,\forall_{N\in\mathbb{N}}\,\exists_{n\geq N}\;:\;|x_n-\bar{x}|\geq\varepsilon, \tag{2.9}$$

obtained from (2.8) by negating the statement following the semi-colon, and making every $\exists$ *a* $\forall$ *and vice versa. Sequences for which (2.9) holds are called divergent.*

2.4 Basic theorems about convergent sequences

In Section 2.3 we taylered the definition of convergence to make Section 2.2 imply that the following theorem holds.

Theorem 2.7. *In* $\mathbb{R}$ *bounded monotone sequences are convergent.*

This theorem in turn implies that the limit in the following theorem exists as the largest lower bound of the sequence

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

but it does not tell us that the limit is 0.

Theorem 2.8. *(The Archimedean Principle)*

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

By Theorem 2.7 the limit exists as the largest lower bound of the sequence $\frac{1}{n}$. So why is the largest lower bound equal to 0? It is clear that 0 is a lower bound. Could there be a larger lower bound? If so this would imply that there is a positive lower bound[7] m for the sequence, i.e.

$$\frac{1}{n}\geq m\quad\text{and thus}\quad n\leq\frac{1}{m}=M$$

voor all $n\in\mathbb{N}$, which looks absurd: how could the sequence

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

be bounded?

Actually in $\mathbb{R}$ it cannot, because a bounded set $\mathbb{N}$ would have a lowest upper bound $S\in\mathbb{R}$. But then $S-\frac{1}{2}$ is not an upper bound whence there

[7] m is a real number here.

exists $n \in \mathbb{N}$ with $n > S - \frac{1}{2}$. Thus[8] the number $n+1 \in \mathbb{N}$ satisfies $n+1 > S + \frac{1}{2} > S$ contradicting S being an upper bound for $\mathbb{N}$. This completes the proof of Theorem 2.8, which in particular says that

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

often stated as a separate statement that bears the name of Archimedes, see Theorem 1.3, which we now proved.

Exercise 2.9. Prove that

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

as $n \to \infty$.

Exercise 2.10. Note that (2.7) is the first occurrence of an absolute[9] value in a definition. We recall that $|x| = x$ for $x \geq 0$ and $|x| = -x$ for $x < 0$. Prove

1. *The triangle inequality*. For all $a, b \in \mathbb{R}$ it holds that

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

2. *The reverse triangle inequality*. For all $a, b \in \mathbb{R}$ it holds that

$$||a| - |b|| \leq |a+b| \quad \text{and thus also} \quad ||a| - |b|| \leq |a-b|,$$

the latter inequality being a nice statement about the map[10] $|\cdot| : \mathbb{R} \to \mathbb{R}$.

3. *The repeated triangle inequality*. Let $N \in \mathbb{N}$. Then

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

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

Exercise 2.11. Let $a, b, c, d \in \mathbb{R}$. Prove that

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

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

[9] $|x|$ will also be called the norm of x.

[10] The first time we use this expression, a function is also called a map.

Exercise 2.12. Prove that a convergent sequence is bounded, i.e. contained in $[-R, R]$ for some $R > 0$. Hint: use the definition with one[11] particular choice of ε and the triangle inequality.

Exercise 2.13. Prove that a convergent sequence can only have one limit. Hint: suppose there are two limits, say L_1 and L_2, and take $\varepsilon = \frac{1}{2}|L_1 - L_2| > 0$ to derive a contradiction using again the triangle inequality.

Exercise 2.14. Suppose that the sequence x_n is convergent with limit L. Prove that $x_n^2 \to L^2$ as $n \to \infty$, that is to say, the sequence x_n^2 is convergent with limit L^2.

Exercise 2.15. Prove that the limit $\bar{x}$ of a convergent sequence x_n with all $x_n \geq a$ for some number a also satisfies $\bar{x} \geq a$. And likewise with $\geq$ replaced by $\leq$.

Theorem 2.16. *If x_n is a convergent sequence indexed by $n \in \mathbb{N}$, with limit $\bar{x}$, then so is $|x_n|$, with limit $|\bar{x}|$.*

Exercise 2.17. Prove Theorem 2.16.

Theorem 2.18. *If x_n and y_n are convergent sequences indexed by $n \in \mathbb{N}$ with limits $\bar{x}$ and $\bar{y}$, then so are $x_n + y_n$, $x_n - y_n$ and $x_n y_n$, respectively with limits $\bar{x} + \bar{y}, \bar{x} - \bar{y}, \bar{x}\bar{y}$.*

To prove the statement about the product we have to estimate $x_n y_n - \bar{x}\bar{y}$ whereas we only know that $x_n - \bar{x}$ and $y_n - \bar{y}$ are under control. The universal trick is to substract and add a suitable term that brings $x_n - \bar{x}$ and $y_n - \bar{y}$ into play[12]. If we choose this term to be $\bar{x}y_n$ we get

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

whence

$$|x_n y_n - \bar{x}\bar{y}| \leq |x_n y_n - \bar{x}y_n| + |\bar{x}y_n - \bar{x}\bar{y}| = |x_n - \bar{x}|\,|y_n| + |\bar{x}|\,|y_n - \bar{y}|.$$

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

[12] Did you do Exercise 2.11?

We now use what we know, namely that every $\tilde{\varepsilon} > 0$ inserted for ε in (2.7) with $L = \bar{x}$ gives the existence of some $N_1 \in \mathbb{N}$ such that $|x_n - \bar{x}| < \tilde{\varepsilon}$ for all $n \geq N_1$. Likewise, there exists $N_2 \in \mathbb{N}$ such that $|y_n - \bar{y}| < \tilde{\varepsilon}$ for all $n \geq N_2$.

We also use Exercise 2.12 to conclude that $|y_n| \leq R$ for some $R > 0$ and if needed choose R larger to ensure that $|x| \leq R$. It then follows that

$$|x_n y_n - \bar{x}\bar{y}| < 2R\tilde{\varepsilon}$$

if $n \geq N = \max(N_1, N_2)$. This N depends on $\tilde{\varepsilon}$. To complete the proof let $\varepsilon > 0$, define $\tilde{\varepsilon}$ by $2R\tilde{\varepsilon} = \varepsilon$ and let N be as just described. Then

$$|x_n y_n - \bar{x}\bar{y}| < \varepsilon \quad \text{for all} \quad n \geq N.$$

Since $\varepsilon > 0$ was arbitrary this concludes the proof.

Exercise 2.19. Prove the statements in Theorem 2.18 for $x_n + y_n$ and $x_n - y_n$.

Theorem 2.18 does not deal with quotients. Suppose x_n is a convergent sequence with limit $\bar{x} \neq 0$, meaning (2.8) holds via this $\bar{x}$. We want to state and prove that

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

Observe that

$$|x_n - \bar{x}| < \varepsilon \iff x_n \in (\bar{x} - \varepsilon, \bar{x} + \varepsilon) \tag{2.10}$$

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

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

for $n \in \mathbb{N}$ as in (2.7). In both cases it follows that

$$|x_n| > \frac{1}{2}|\bar{x}| \quad \text{whence} \quad |\frac{1}{x_n}| < \frac{2}{|\bar{x}|} \tag{2.11}$$

and then also

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

Theorem 2.20. *Let x_n be a convergent sequence with limit $\bar{x} \neq 0$. Then there exists $N \in \mathbb{N}$ such the sequence $\frac{1}{x_n}$ indexed by $n \in \mathbb{N}$ with $n \geq N$ is a well defined convergent sequence with limit $\frac{1}{\bar{x}}$.*

Exercise 2.21. Prove Theorem 2.20.

Exercise 2.22. The intermediate step (2.11) relied on (2.10) with two distinct cases. Use the reverse triangle inequality to avoid this use of intervals in the derivation of (2.11) .

Exercise 2.23. Show that the statement

$$\forall_{\varepsilon>0} \ldots\ldots\ldots\ldots\ldots < \varepsilon$$

in Definition 2.5 is equivalent to the at first glance weaker statement with $\leq \varepsilon$, which is sometimes easier to obtain.

2.5 Suprema and infima of nonempty sets

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

$$\{x_n : n \in \mathbb{N}\}$$

of $\mathbb{R}$.

Definition 2.24. *A nonempty subset of 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.*

Of course we want to show that a nonempty subset A of $\mathbb{R}$ which is bounded from above has a lowest upper bound. This is a bit of a project. Suppose that A is such a set. Take an $a_0 \in A$ and consider

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

If m_0 is 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 has an infimum that we will call S.

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

It may or may not happen that $S = \bar{a} \in A$, but in both cases the conclusion is stated by:

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 denoted by $I = \inf A \in \mathbb{R}$.*

Remark 2.27. *If A is not bounded from above we say that $\sup A = \infty$. If A is not bounded from below we say that $\inf A = -\infty$.*

2.6 What about Heron's limit?

Heron's method to generate the sequence x_n may be written as

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

whence

$$2x_{n-1}(x_n - x_{n-1}) = 2 - x_{n-1}^2.$$

Exercise 2.28. Shifting the index we also have

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

Prove that $x_n^2 \to 2$ as $n \to \infty$ if the sequence x_n is convergent. Hint: first show that $x_{n+1} - x_n \to 0$.

In view of Exercise 2.14 it follows for the largest upper bound L of Heron's sequence that $L^2 = 2$. By construction $L > 0$ because $L \geq \frac{4}{3}$. Note that we did not really use that $x_0 = 1$ for the starting value . For every $x_0 > 0$ with $x_0^2 \neq 2$ the conclusions are the same: the sequence x_n indexed by $n \in \mathbb{N}$ is strictly decreasing with limit L.

Exercise 2.29. Prove that L is the only positive real number which squares to 2. This then justifies the conclusion that $L = \sqrt{2}$. Tie this up with Exercise 2.2 which produced a bounded nondecreasing sequence which therefore has a supremum S. Prove that S too is equal to 2: $S = L = \sqrt{2}$.

3 Banach contraction fixed point theorem

Let's look at another approach to arrive at the conclusion that limits of sequences defined via (2.1) exist, an approach in which we do not use the monotonicity of the sequence, but look at the size of the increments or steps

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

These steps reproduce x_n via

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

from x_0. Clearly the existence of

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

corresponds to the existence of

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

a limit that we will denote by

$$\sum_{n=1}^{\infty} \xi_n = \sum_{k=1}^{\infty} \xi_k = \lim_{n\to\infty} S_n = \lim_{n\to\infty} \sum_{k=1}^{n} \xi_k. \tag{3.2}$$

3.1 Estimates for increments

Returning to x we note that

$$\frac{x}{2} + \frac{1}{x} > \frac{x}{2} + \frac{1}{2x} > 1 \quad \text{for} \quad x > 0,$$

so every sequence defined by

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

has $x_n > 1$ for all $n \in \mathbb{N}$ if $x_0 > 0$. We then have

$$\xi_2 = x_2 - x_1 = \frac{x_1}{2} + \frac{1}{x_1} - \frac{x_0}{2} - \frac{1}{x_0} = \frac{x_1 - x_0}{2} + \frac{1}{x_1} - \frac{1}{x_0}$$

$$= \left(\frac{1}{2} - \frac{1}{x_0 x_1}\right)(x_1 - x_0) = \left(\frac{1}{2} - \frac{1}{x_0 x_1}\right)\xi_1,$$

and likewise

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

for every $n \in \mathbb{N}$.

Starting from $n = 2$ the prefactor on the right is between $-\frac{1}{2}$ and $\frac{1}{2}$ because $x_n > 1$. For $n = 1$ the same conclusion holds because

$$x_0 x_1 = \frac{x_0^2}{2} + 1 > 1.$$

Thus

$$|x_{n+1} - x_n| = |\xi_{n+1}| < \frac{1}{2}|\xi_n| = |x_n - x_{n-1}| \quad \text{for all} \quad n \in \mathbb{N}.$$

The first step may be large, depending on x_0, but every next step is under control by the first step:

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

and so on:

$$|\xi_n| < \frac{1}{2^n}|\xi_1|$$

The steps get smaller and smaller exponentially fast.

If we denote the defining map[1] for Heron's sequence by f, i.e.

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

then[2]

$$|x_{n+1} - x_n| = |f(x_n) - f(x_{n-1})| \leq \frac{1}{2}|x_n - x_{n-1}|. \tag{3.3}$$

for all $n \in \mathbb{N}$ if $x_0 > 0$. In fact f has the property that

$$|f(x) - f(y)| \leq \frac{1}{2}|x - y| \tag{3.4}$$

for all x and y with $x \geq 1$ and $y \geq 1$. We say that f is 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:

[1] We shall prefer to use the word map for functions which are not $\mathbb{R}$-valued.

[2] We slightly weaken the statement replacing $<$ by $\leq$.

Definition 3.1. *Let $A \subset \mathbb{R}$. A function*[3] *$f : A \to \mathbb{R}$ is called Lipschitz continuous with Lipschitz constant $L > 0$ if for all $x, y \in A$ it holds that*

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

If $L < 1$ then f is called contractive with contraction factor L. If $L = 1$ then f is called nonexpanding.

Exercise 3.2. Show that the map $x \to |x|$ is nonexpanding.

The contraction property implies that there can be atmost one solution of $x = f(x)$ in A if f is such a contractive map from A to $\mathbb{R}$, even if the range

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

is not a subset of A.

Exercise 3.3. A warming up exercise for what's to come: let X be a subset of $\mathbb{R}$ and suppose that f is a contractive map from X to X with contraction factor $\frac{1}{2}$. Suppose that the sequence x_n defined by $x_n = f(x_{n-1})$ for $n \in \mathbb{N}$ and some given $x_0 \in X$ converges to a limit $\bar{x}$ in X. Prove that $\bar{x}$ is the only solution of $f(x) = x$ in X. What can you conclude about sequences starting from other values in X?

3.2 Properties of Heron's sequence due to contraction

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

$$x_m - x_n = S_m - S_n = \xi_{n+1} + \cdots + \xi_m$$

and thus via (3.3) estimated by

$$|x_m - x_n| = |S_m - S_n| \leq |\xi_{n+1}| + \cdots + |\xi_m| \tag{3.5}$$

$$\leq \frac{|\xi_1|}{2^{n+1}} + \cdots + \frac{|\xi_1|}{2^m} \leq \frac{|\xi_1|}{2^N}$$

if $m > n \geq N$, a bound we can make as small as we like by choosing N large. To be precise, given any given real number $\varepsilon > 0$, we can choose a positive integer N such that

$$\frac{|\xi_1|}{2^N} < \varepsilon,$$

[3] We prefer to write $f : A \to \mathbb{R}$ and not $f : A \mapsto \mathbb{R}$.

by the Archimedean principle again. If not there would be an $\varepsilon > 0$ such that

$$N \leq 2^N \leq \frac{|\xi_1|}{\varepsilon}$$

for all $N \in \mathbb{N}$, contradicting the Archimedean principle, which says that the set $\mathbb{N}$ is not bounded from above by any real number, or equivalently that

$$1, 2, 3, 4, 5, 6, 7, \ldots$$

is not a bounded sequence.

It follows that

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

with N depending on $\varepsilon > 0$ of course.

3.3 Cauchy sequences, monotone subsequences

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

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \forall_{m,n\geq N} \,:\, \underbrace{|x_n - x_m|}_{d(x_n,x_m)} < \varepsilon, \tag{3.6}$$

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 ε.

Definition 3.4. *A sequence of real numbers x_n indexed by $n \in \mathbb{N}$ is called Cauchy, or a Cauchy sequence, if (3.6) holds.*

We already knew that Heron's sequence is convergent. Compare this definition to the definition of a convergent sequence in Section 2.3. Unlike Definition 2.5 the new definition does not involve any number L that candidates for being the limit of the sequence. Can it be used as an alternative definition of convergence?

Exercise 3.5. Prove that a convergent sequence is Cauchy. Hint: $2\tilde{\varepsilon} = \varepsilon$.

Theorem 3.6. *The collection of Cauchy sequences in the set of real numbers is precisely the collection of convergent sequences in the real numbers.*

Proof of Theorem 3.6: this theorem is an immediate consequence of Exercise 3.5 and the rather pleasant fact that every sequence of real numbers has a monotone subsequence, which we state as separate theorem.

Theorem 3.7. *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*

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

such that the subsequence x_{n_k}, which is indexed by the same k, is monotone. The statement also holds for a sequence of rational numbers.

Exercise 3.8. Prove Theorem 3.7. 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 nonincreasing 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.

Once you know Theorem 3.7 you observe that Cauchy sequences are bounded, and thus so is the monotone subsequence provided by Theorem 3.7, which then has a limit in view of Theorem 2.7. This limit turns out to be the limit of the whole sequence as well, in view of the following exercise which completes the proof of Theorem 3.6.

Exercise 3.9. Suppose that x_n is a Cauchy sequence of real numbers which has a subsequence which is convergent. Prove that the sequence itself is convergent and has the same limit as the convergent subsequence.

3.4 Bolzano-Weierstrass: convergent subsequences

We note that Theorems 2.7 and 3.7 also immediately imply the following theorem which will be essential for proving theorems about continuous[4] functions. Simply observe that Theorem 3.7 implies that every bounded sequence has a monotone (also bounded) subsequence, and that Theorem 2.7 says this subsequence is convergent. That is:

Theorem 3.10. *(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.*

[4] You probably already know this term, but see Definition 3.34 if not.

The standard proof of Theorem 3.10 involves a diagonal subsequence argument reminiscent of the proof of Theorem 1.2. You can skip it for now. The diagonal argument will also be used and explained in the proof of Theorem 4.44.

In case you don't skip: 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

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

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

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

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

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

and the diagonal subsequence has

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

for every k.

Exercise 3.11. Finish this proof of Theorem 3.10. Hint: $a_k \leq x_{n_{kk}} \leq b_k$ and the sequences a_k, b_k are monotone.

Remark 3.12. *Limits of convergent subsequences of a sequence are called limit points of the sequence.*

Exercise 3.13. Prove that $\bar{x}$ is a limit point of the sequence x_n if and only if

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

Thus Theorem 3.10 states for bounded sequences x_n of real numbers that

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

a statement that looks very much like the statement (2.8) for convergence, so be careful.

Exercise 3.14. Prove that a bounded sequence of real numbers is convergent if and only if it has only one limit point.

Exercise 3.15. 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 3.16. Determine all limit points of the sequences defined by $x_n=(-1)^n$, $x_n=(-1)^n+\frac{1}{n}$, $x_n=(-1)^n+(-1)^{2n}$.

Definition 3.17. *A subset A of $\mathbb{R}$ is called closed in $\mathbb{R}$ if every convergent sequence $x_n\in A$ has its limit in A as well.*

Definition 3.18. *A subset $\mathcal{O}$ of $\mathbb{R}$ is called open if for every point $x_0\in\mathcal{O}$ there exists $\delta_0>0$ such that $x\in\mathcal{O}$ for all $x\in\mathbb{R}$ with $|x-x_0|<\delta_0$.*

Exercise 3.19. Prove that $A\subset\mathbb{R}$ is closed if and only if $A^c=\{x\in\mathbb{R}:\,x\notin A\}$ is open.

Exercise 3.20. Let $a,b\in\mathbb{R}$ with $a<b$. Prove $[a,b]$ is closed in $\mathbb{R}$.

Exercise 3.21. Let A and B be closed subsets of $\mathbb{R}$. Prove that $A\cup B$ and $A\cap B$ are closed.

Exercise 3.22. Let I be any index set and let $A_i\subset\mathbb{R}$ be closed for every $i\in I$. Prove that the intersection

$$\cap_{i\in I}A_i=\{x\in\mathbb{R}:\,\forall_{i\in I}:\,x\in A_i\}$$

is closed.

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

We have seen in Section 3.2 that if f is a contractive[5] map from a subset X of $\mathbb{R}$ to itself then every sequence defined by $x_n = f(x_{n-1})$ starting from any $x_0 \in X$ is a Cauchy sequence, and thus convergent to some limit $\bar{x} \in \mathbb{R}$ in view of Theorem 3.6. Now assume that X is closed. Then $\bar{x} \in X$ and by Exercise 3.3 it is the unique solution of the equation $f(x) = x$ in X.

We have now proved a special case of the Banach contraction theorem, namely for closed sets $X \subset \mathbb{R}$ and contractive maps f from X to X with contraction factor $\frac{1}{2}$. In fact the contraction factor can be any number θ with $0 < \theta < 1$, the condition being that

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

for all $x, y \in X$, i.e. the distance between two outputs of f is atmost θ times the distance of the inputs.

Exercise 3.23. Prove that $\theta^n \to 0$ as $n \to \infty$ if $\theta \in (0, 1)$. Hint: you have many ways now to do this but it has to be done. This exercise generalises Exercise 2.9. Note that the sequence θ^n is decreasing. Also, it is defined by $x_0 = 1$ and $x_n = \theta x_{n-1}$ for $n \in \mathbb{N}$.

Theorem 3.24. *(Banach contraction theorem for closed subsets of $\mathbb{R}$) Let X be a closed subset of $\mathbb{R}$ and let f be a map from X to X which is contractive with contraction factor $\theta \in (0, 1)$, meaning that (3.7) holds for all $x, y \in X$. Then f has a unique fixed point $\bar{x} \in X$. For every $x_0 \in X$ this $\bar{x}$ is the limit of the sequence x_n defined by $x_n = f(x_{n-1})$ for all $n \in \mathbb{N}$.*

Exercise 3.25. Prove this theorem. Hint: you have seen the proof for $\theta = \frac{1}{2}$ that started from (3.3) and via (3.5) led to the sequence of iterates being a Cauchy sequence. With $\frac{1}{2}$ replaced by θ you should arrive at

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

for $m > n \geq N$ by taking N large.

[5] For the moment with contraction factor $\frac{1}{2}$.

3.6 Generalisation to metric spaces

We write

$$d(x, y) = |x - y| \tag{3.8}$$

for the distance between x and y in X. We call d, which assigns to every pair of elements of X a number in $\mathbb{R}$, a metric on X because

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

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

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

and we say that X is a metric space with metric d. Now forget about the elements of $\mathbb{R}$ outside of X and about X being a subset of $\mathbb{R}$, and forget about any algebraic operations in X.

Definition 3.26. *A set X is called a metric space if for every $x, y \in X$ a real number $d(x, y)$ is defined such that the above three properties hold.*

In particular $X = \mathbb{R}$ is an example of a metric space, and so is every subset of a metric space.

Exercise 3.27. Think about other examples. Subsets of $\mathbb{R}^2$ with the Pythagorean distance. 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.

Copying earlier definitions replacing absolute values $|x - y|$ of differences by distances $d(x, y)$ we have

Definition 3.28. *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[6] *if every Cauchy sequence in X is convergent*[7]*.*

[6] We don't speak of closed here since we forgot about everything outside of X.

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

Theorem 3.29. *(Banach contraction theorem for complete metric spaces) Let X be a complete metric space and let Φ be a map from X to X which is contractive with contraction factor $\theta \in (0,1)$, meaning that*

$$d(\Phi(x), \Phi(y)) \leq \theta\, d(x,y)$$

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

Theorem 3.30. *$\mathbb{R}$ is complete with $d(x,y) = |x-y|$. And so is every closed subset of $\mathbb{R}$.*

Exercise 3.31. Prove Theorem 3.30.

Exercise 3.32. Prove Theorem 3.29. Hint: replace (3.3) by

$$d(x_{n+1}, x_n) = d(\Phi(x_{n+1}), \Phi(x_n)) \leq \theta\, d(x_n, x_{n-1}) \tag{3.9}$$

and follow the reasoning from there. You don't need the algebra that defines the difference $x_n - x_m$ of two elements of the sequence!

3.7 Examples of complete metric spaces

An important general example is formulated in this theorem.

Theorem 3.33. *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, i.e. every sequence in A which is convergent in X has its limit in A.*

Important concrete examples are (closed subsets of) $\mathbb{R}$, $\mathbb{R}^2$ with the Euclidean distance[8], and, as we shall see in Chapter 4, (closed subsets of)

$$C([a,b]) = \{f : [a,b] \to \mathbb{R} : \ f \text{ is continuous in every point of } [a,b]\}.$$

The space $C([a,b])$ will be used for the construction of solutions of differential equations via a transformation to integral equations. Integral equations will be discussed in Section 6.6 and solved via Theorem 3.29 with

[8] More about norms and metrics on $\mathbb{R}^2$ in Section 6.3.

$X = C([a,b])$. We note that $C([a,b])$ will also be a natural function space on which to consider the (linear) map

$$f \to \int_a^b f(x)\,dx.$$

The above definition of $C([a,b])$, defined for $a < b$, requires the concept of continuity of a function in a point, a concept you are probably already familiar[9] with. In the metric context this is the natural definition.

Definition 3.34. *Let X be a subset of $\mathbb{R}$ or any other metric space, and let $f : X \to \mathbb{R}$ be a real valued function. Then f is called continuous in $\xi \in X$ if the implication*

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

holds for every sequence x_n in X. If this holds for every ξ in X then f is called continuous.

Definition 3.35. *Let X, Y be metric spaces and $\Phi : X \to Y$ a map. Then Φ is called continuous in $\xi \in X$ if the implication*

$$x_n \to \xi \implies \Phi(x_n) \to \Phi(\xi)$$

holds for every sequence x_n in X. If this holds for every ξ in X then Φ is called continuous.

With $X = [a,b]$ and $Y = \mathbb{R}$ the set $C([a,b])$ is now properly defined. In the next chapter we will show that

$$d(f,g) = \max_{a \le x \le b} |f(x) - g(x)| \tag{3.10}$$

exists for every $f, g \in C([a,b])$, as a consequence of Theorem 4.29.

Exercise 3.36. Take $a = 0, b = 1$, $f(x) = x^2$, $g(x) = x(1-x)$. Compute $d(f,g)$. Sketch the graphs of $y = f(x)$ and $y = g(x)$ in the xy-plane and explain what $d(f,g)$ is.

Exercise 3.37. Assume that (3.10) exists for all $f, g \in C([a,b])$. Prove that it defines a metric[10] on $C([a,b])$.

[9] Lipschitz continuity is a much stronger global property.

[10] In Theorem 4.31 the completeness of $C([a,b])$ will be established.

We will take this over the top: another example of a complete metric space is

$$C([a,b],Y) = \{f : [a,b] \to Y : \ f \text{ continuous in every point of } [a,b]\},$$

the set of all continuous maps f from $[a,b]$ to a given complete metric space Y, equipped with the metric defined by

$$d(f,g) = \max_{a \leq x \leq b} d_{_Y}(f(x), g(x)),$$

in which $d_{_Y}$ is the metric on Y. The case that Y is itself a space like $C([a,b])$ will be of special interest when we consider the dependence of solutions of differential equations on parameters, culminating in Exercise 10.4.

4 Continuous functions on metric spaces

In the last section of the previous chapter we introduced the concept of continuity of a function or map in a given point via sequences. For real valued functions of a real variable x ranging over some interval I we restate the definition.

Definition 4.1. *Let $I \subset \mathbb{R}$ be an interval, let $f : I \to \mathbb{R}$ be a function and let $\xi \in I$. Then f is called continuous in ξ if $f(x_n) \to f(\xi)$ for every sequence x_n in I with $x_n \to \xi$. If f is continuous in every $\xi \in I$ then $f : I \to \mathbb{R}$ is called continuous.*

4.1 Limits and continuity via epsilons and deltas

At some point though it will be more convenient to use the ε, δ-formulation of continuity of f in ξ:

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{x\in I}\ : \underbrace{|x-\xi|}_{d(x,\xi)} < \delta \implies \underbrace{|f(x)-f(\xi)|}_{d(f(x),f(\xi))} < \varepsilon. \tag{4.1}$$

Exercise 4.2. 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 (4.1) says.

Exercise 4.3. Let $\xi = 2$, $f(x) = 2x + 1$ and $\varepsilon = \frac{1}{n}$ with $n \in \mathbb{N}$. Which $\delta > 0$ will validify (4.1)? Same question for $f(x) = x^2$ and $f(x) = \frac{1}{x}$.

Theorem 4.4. *Let X, Y be metric spaces, let $\Phi : X \to Y$ and let $\xi \in X$. Then Φ is continuous in ξ if and only if*

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{x\in X}\ : d_X(x,\xi) < \delta \implies d_Y(\Phi(x),\Phi(\xi)) < \varepsilon, \tag{4.2}$$

in which d_X is the metric on X and d_Y is the metric on Y.

Exercise 4.5. Prove Theorem 4.4 in the case that $X = Y = \mathbb{R}$ and $\Phi = f$.

Exercise 4.6. Prove Theorem 4.4.

Exercise 4.7. Let $I \subset \mathbb{R}$ be an interval. By Theorem 4.2 with $X = I$ and $Y = \mathbb{R}$ the function $f : I \to \mathbb{R}$ is continuous if

$$\forall_{\xi \in I} \, \forall_{\varepsilon > 0} \, \exists_{\delta > 0} \, \forall_{x \in I} \, : |x - \xi| < \delta \implies |f(x) - f(\xi)| < \varepsilon,$$

so $\delta > 0$ depends on $\xi \in I$ and $\varepsilon > 0$. We sometimes write $\delta = \delta_{\xi\varepsilon}$. Determine the smallest possible $\delta_{\xi\varepsilon}$ for $f : I \to \mathbb{R}$ defined by $f(x) = x^2$ if $I = [0, 1]$ and if $I = \mathbb{R}$. In which of the two cases is het possible to choose $\delta > 0$ depending[1] on $\varepsilon > 0$ only? Same question for $I = (0, 1)$ and $f(x) = \frac{1}{x}$.

A slight modification of (4.2) for $\Phi : X \to Y$ and $L \in Y$ is the statement that $\Phi(x) \to L$ as $x \to \xi$:

$$\forall_{\varepsilon > 0} \, \exists_{\delta > 0} \, \forall_{x \in X} \, : 0 < d_X(x, \xi) < \delta \implies d_Y(\Phi(x), L) < \varepsilon, \tag{4.3}$$

a definition which can be modified to maps with a domain D_Φ which is not the whole of X.

Definition 4.8. *Let X and Y be metric spaces and $\Phi(x) \in Y$ be defined for every $x \in D_\Phi \subset X$. For $L \in Y$ and $\xi \in X$ we write*

$$L = \lim_{D_\Phi \ni x \to \xi} \Phi(x)$$

or just

$$L = \lim_{x \to \xi} \Phi(x)$$

if

$$\forall_{\varepsilon > 0} \, \exists_{\delta > 0} \, \forall_{x \in D_\Phi} \, : 0 < d_X(x, \xi) < \delta \implies d_Y(\Phi(x), L) < \varepsilon. \tag{4.4}$$

Informally we say $\Phi(x) \to L$ as $D_\Phi \ni x \to \xi$ or just $\Phi(x) \to L$ as $x \to \xi$.

Exercise 4.9. Reformulate Definition 4.4 as a statement for sequences in D_Φ.

Continuity of Φ in ξ means

$$\lim_{D_\Phi \ni x \to \xi} = \Phi(\xi). \tag{4.5}$$

This is three statements in one: the existence of the limit, the statement that $\xi \in D_\Phi$, and the limit being equal to $\Phi(\xi)$. Note that the statement in (4.3) may be completely meaningless if $\xi \notin D_f$.

[1] More on this issue in Section 4.5, we need $\delta > 0$ independent of ξ in Section 5.4.

Exercise 4.10. Let $A = D_\Phi \subset X$ be the domain of the Y-valued function Φ. Show that the limit L in Definition 4.8 is unique if and only if[2]

$$\forall_{\delta>0}\, \exists_{x\in A}\ : 0 < d(x,\xi) < \delta. \tag{4.6}$$

If (4.6) fails then the statement in the definition holds for every $L \in Y$.

Remark 4.11. *Every statement of the form*

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \ldots\ldots\ldots\ldots : \ldots\ldots < \delta \implies \ldots\ldots < \varepsilon$$

is equivalent to the at first glance stronger statement with $\leq \delta$*, and the at first glance weaker statement with* $\leq \varepsilon$*. The first is sometimes less cumbersome to continue from, and the latter less cumbersome to establish.*

4.2 Closed subsets, interior points and all that

Exercise 4.6 takes us to terminology used before for $X = \mathbb{R}$. See Definition 3.17 and further. The two basic definitions we will be needing are stated first

Definition 4.12. *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. The name is best explained rephrasing the statement as: taking limits of sequences contained in A you cannot get out of A.*

Theorem 4.13. *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 in a_n with $a_n \in A$. Then $\bar{A}$ is the smallest closed subset of X which contains A, and it is called the closure of A.*

Exercise 4.14. Prove Theorem 4.13. Hint: first show that $\bar{A}$ is closed, then show that there is no closed subset $\tilde{A}$ with $A \subset \tilde{A} \subset \bar{A}$ and $\tilde{A} \neq \bar{A}$.

Definition 4.15. *Let A be a subset of a metric space X. A point $x_0 \in A$ is called an interior point of A if there exists $\delta_0 > 0$ such that $x \in A$ for all $x \in X$ with $d(x, x_0) < \delta_0$, i.e.*

$$\exists_{\delta_0>0}\, \forall_{x\in X}\, :\, d(x,x_0) < \delta_0 \implies x \in A.$$

If so we say that x_0 is in the interior of A, notation $x_0 \in \mathrm{int}(A)$. [3]

[2] Replacing $< \delta$ by $\leq \delta$ in (4.6) gives an equivalent statement, like in Remark 4.16.

[3] The use of the term interior should be intuitively clear.

Exercise 4.16. The statement

$$\exists_{\delta_0>0} \ldots\ldots\ldots\ldots\ldots < \delta_0 \implies \ldots\ldots$$

in Definition 4.15 is equivalent to the at first glance stronger statement with $\leq \delta_0$.

The definitions and theorems below are strictly speaking not needed for our purposes here. For the moment we leave the proofs to the reader.

Definition 4.17. *Let X be a metric space and $A \subset X$. Then $\xi \in X$ is called an accumulation point of A if (4.6) holds. The name is explained by the following theorem.*

Theorem 4.18. *Let X be a metric space and $A \subset X$. Then $\xi \in X$ is an accumulation point of A if and only if there exists a sequence $x_n \in A$ such that $d(x_n, \xi)$ is a strictly decreasing sequence of positive numbers converging to zero.*

Definition 4.19. *A subset $\mathcal{O}$ of X is called open*[4] *if every point $x_0 \in \mathcal{O}$ is an interior point of $\mathcal{O}$. For $\delta > 0$ the set*

$$B_\delta(x_0) = \{x \in X : d(x, x_0) < \delta\}$$

is called an open ball with center x_0 and radius δ.

Theorem 4.20. *Let X be a metric space, $x_0 \in X$ and $\delta > 0$. Then the open ball $B_\delta(x_0)$ is open. A subset A of X is closed if and only if the complement*

$$A^c = X \backslash A = \{x \in X : x \notin A\}$$

of A in X is open.

Theorem 4.21. *Unions of open subsets of a metric space X are open. The intersection of two open subsets of X is also open.*

Theorem 4.22. *Let X, Y be metric spaces and $\Phi : X \to Y$. Then Φ is continuous in every point of X if and only if the inverse image*

$$\Phi^{-1}(\mathcal{O}) = \{x \in X : \Phi(x) \in \mathcal{O}\}$$

of $\mathcal{O}$ under Φ is open in X for every open subset $\mathcal{O}$ of Y.

[4] The use of the term open is intuitively not very clear, it is not the negation of closed!

Exercise 4.23. Refering to Definition 4.19 with $X = \mathbb{R}^2$: 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[5]

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

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

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

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

4.3 A global monotone inverse function theorem

Theorem 4.24. *Let I be an open interval in $\mathbb{R}$ and $f : I \to \mathbb{R}$ a continuous function. For $a, b \in I$ with $a < b$ let*

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

be the image of $[a,b]$ under f. Then

$$f(a) < f(b) \implies [f(a), f(b)] \subset f([a,b]),$$

and

$$f(a) > f(b) \implies [f(b), f(a)] \subset f([a,b]).$$

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 (4.1). 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, contradicting ξ being an upper bound.

Can it be that $f(\xi) > c$? Choose $\varepsilon > 0$ with $\varepsilon < f(\xi) - c$ and apply (4.1). 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, making $\xi - \delta$ an upper bound, contradicting ξ being the lowest upper bound. This completes the proof for the case that $f(a) < f(b)$. The other case is of course similar.

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

Theorem 4.25. *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

$$I = g(J) = \{g(y) : y \in J\}.$$

For the proof observe that 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.

If $c, d \in J$ with $c < d$ then $c = f(a)$ and $d = f(b)$, and $[c, d] \subset J$ by Theorem 4.24. 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 $\delta_0 > 0$, whence $[f(x_0 - \delta_0), f(x_0 + \delta_0)] \subset J$ so y_0 is an interior point because $f(x_0 - \delta_0) < f(x_0) < f(x_0 + \delta_0)$. We conclude that J is an open interval.

It remains to prove the continuity of g, so let $y_0 = f(x_0)$ and $\varepsilon > 0$. It is no limitation to choose $\varepsilon < \delta_0$, δ_0 as just above. Then

$$(f(x_0 - \varepsilon), f(x_0 + \varepsilon)) \subset [f(x_0 - \delta_0), f(x_0 + \delta_0)] \subset J$$

and we can choose $\delta > 0$ such that

$$f(x_0-\delta_0) < \underbrace{f(x_0 - \varepsilon)}_{\substack{\downarrow g \\ x_0-\varepsilon}} < y_0-\delta < \underbrace{f(x_0) = y_0}_{\substack{\downarrow g \\ x_0=g(y_0)}} < y_0+\delta < \underbrace{f(x_0 + \varepsilon)}_{\substack{\downarrow g \\ x_0+\varepsilon}} < f(x_0+\delta_0),$$

whence

$$g((y_0 - \delta, y_0 + \delta)) \subset (g(y_0) - \varepsilon, g(y_0) + \varepsilon).$$

This completes the proof.

Exercise 4.26. 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 4.25? Specify J and compute $g(y)$. What is J if $I = (0, \infty)$?

Exercise 4.27. Formulate Theorem 4.25 for strictly decreasing functions.

4.4 Maxima and minima and the maximum norm

One of the highlights of analysis in $\mathbb{R}^n$ is that a real valued function defined and continuous on a closed bounded subset A of $\mathbb{R}^n$ has a global maximum and global minimum on A. Here's the version for $[a,b] \subset \mathbb{R}$.

Definition 4.28. *Let X be a set and $f : [a,b] \to \mathbb{R}$ a real valued function. If $\bar{x} \in X$ has the property that $f(x) \leq f(\bar{x})$ for every $x \in X$, then $M = f(\bar{x})$ is called a global maximum of f and $\bar{x}$ is called a maximizer of f. Likewise, If $\underline{x} \in X$ has the property that $f(x) \geq f(\underline{x})$ for every $x \in X$, then $m = f(\underline{x})$ is called a global minimum of f and $\underline{x}$ is called a minimizer of f.*

Theorem 4.29. *Let $a, b \in \mathbb{R}$ with $a < b$. If $f : [a,b] \to \mathbb{R}$ is continuous in every point of $[a,b]$ then f has a global maximum and a global minimum. As a consequence the metric (3.10) is well defined[6], and in particular*

$$|f| = \max_{a \leq x \leq b} |f(x)|$$

is called the maximum norm of f.

For the proof we use Definition 4.1. Let

$$R_f = \{f(x) : x \in [a,b]\} \subset \mathbb{R}$$

be the range of f. This set may be bounded from above. If it is then by Theorem 2.26 it has a smallest lower bound which we will call M, in which case every $M - \frac{1}{n}$ with $n \in \mathbb{N}$ is not an upper bound and therefore $x_n \in [a,b]$ exists with $M - \frac{1}{n} < f(x_n) \leq M$. It follows that $f(x_n) \to M$.

If R_f is not bounded then no $n \in \mathbb{N}$ is an upper bound and thus there exists $x_n \in [a,b]$ with $f(x_n) > n$. In both cases the sequence x_n has a convergent subsequence x_{n_k} with limit $\bar{x} \in [a,b]$ in view of Exercise 3.20. Since f is continuous in $\bar{x}$ it follows that $f(x_{n_k}) \to f(\bar{x})$. Therefore $f(x_{n_k})$ is a bounded sequence. This excludes the second case whence we're in the first case and conclude that $f(x_{n_k}) \to M$. But then $M = f(\bar{x})$ because the limit of the sequence $f(x_{n_k})$ is unique. We conclude that the supremum of R_f is attained by f: $M = f(\bar{x})$ is the global maximum of f and $\bar{x}$ is the maximizer. The argument for the global minimum is similar.

[6] Because $x \to |f(x) - g(x)|$ is continuous if f and g are.

Theorem 4.30. *With algebra in $C([a,b])$ defined by*

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

for all $x \in [a,b]$ and $f, g \in C([a,b])$ we have

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

for all $f, g \in C([a,b])$.

There is not much to prove. Theorems 2.16,2.18 and Definition 4.1 imply that with $f, g \in C([a,b])$ also $|f|, |g|, f+g, fg, |f+g|, |fg| \in C([a,b])$. Let $\bar{x}, \bar{y}, \bar{z}, \bar{w} \in [a,b]$ be maximizers for $|f|, |g|, |f+g|, |fg|$. Then

$$|f+g| = |f(\bar{z}) + g(\bar{z})| \le |f(\bar{z})| + |g(\bar{z})| \le |f(\bar{x})| + |g(\bar{y})|$$

and

$$|fg| = |f(\bar{w})|\,|g(\bar{w})| \le |f(\bar{x})|\,|g(\bar{y})|$$

Theorem 4.31. *With the metric well defined by (3.10), i.e.*

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

the space $C([a,b])$ is complete.

The proof is easy except for one issue which is discussed in Theorem 4.32 below. The easy part of the reasoning is as follows. If f_n is a Cauchy sequence in $C([a,b])$ then for each fixed $x_0 \in [a,b]$

$$|f_n(x_0) - f_m(x_0)| \le \max_{a \le x \le b} |f_n(x) - f_m(x)| = d(f_n, f_m) < \varepsilon, \tag{4.7}$$

provided $m, n \ge N$ with N depending on $\varepsilon > 0$. Thus $f_n(x_0)$ is a Cauchy sequence in $\mathbb{R}$ and thereby convergent. The limit is denoted by $f(x_0)$. Since $x_0 \in [a,b]$ was arbitrary, this defines a function $f : [a,b] \to \mathbb{R}$.

Also we can take the limit of the left hand side of (4.7) as $m \to \infty$. This gives

$$|f_n(x_0) - f(x_0)| \le \varepsilon, \tag{4.8}$$

for $n \ge N$. Note that N depends on $\varepsilon > 0$ but not on x_0.

If we would have $f \in C([a,b])$ we can take the maximum over $x_0 \in [a,b]$ and conclude that

$$d(f_n, f) = \max_{a \le x \le b} |f_n(x) - f(x)| \le \varepsilon$$

for all $n \ge N$, still the same N which only depends on ε. This would prove[7] the theorem. All that is needed to conclude is the following theorem.

[7] See Exercise 2.23.

Theorem 4.32. *Let f_n be sequence of functions from $[a,b]$ to $\mathbb{R}$ and f be another function from $[a,b]$ to $\mathbb{R}$. If*

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

and all f_n are continuous in $x_0\in[a,b]$, then so is f.

Definition 4.33. *Bearing in mind Remark 2.23 the sequence f_n is called uniformly convergent on $[a,b]$ with limit f if (4.9) holds.*

The theorem says that the limit of a uniformly convergent sequence of functions f_n inherits the continuity properties of f_n. For a proof we need to estimate $f(x)-f(x_0)$ and we split this in three terms that allow us to use what we know. Writing

$$f(x)-f(x_0)=f(x)-f_n(x)+f_n(x)-f_n(x_0)+f_n(x_0)-f(x_0)$$

we have

$$|f(x)-f(x_0)|\leq\underbrace{|f(x)-f_n(x)|}_{\leq\tilde{\varepsilon}}+|f_n(x)-f_n(x_0)|+\underbrace{|f_n(x_0)-f(x_0)|}_{\leq\tilde{\varepsilon}}, \tag{4.10}$$

with the inequalities holding for all $n\geq N$ depending on $\tilde{\varepsilon}$ via the assumption in the theorem. Now pick just one such n, for instance $n=N$, and use the continuity of f_N in x_0. It follows that for some $\delta>0$ it holds that

$$|f_N(x)-f_N(x_0)|\leq\tilde{\varepsilon}$$

if $x\in[a,b]$ and $|x-x_0|\leq\delta$. Combining with (4.10) we see that

$$|f(x)-f(x_0)|\leq 3\tilde{\varepsilon}.$$

So given $\varepsilon>0$ we take $\tilde{\varepsilon}=\frac{1}{3}\varepsilon$ and choose the corresponding N. This completes the proof of Theorem 4.32 and thereby the proof that $C([a,b])$, equipped with

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

is a complete metric space.

Exercise 4.34. Show that there are bounded sequences in $C([a,b])$ which do not have any convergent subsequence. Hint: $[a,b]=[0,1]$, $f_n(x)=x^n$.

4.5 Uniform epsilon statements and continuity

Definition 4.33 is the first example of a $\forall_{\varepsilon>0}$-statement in which what follows contains an additional parameter or variable, x in this case, that does not affect the quantifiers that precede it, nor the statement that follows: the sequence $f_n(x)$ converges to $f(x)$ as $n \to \infty$ with a choice of $N \in \mathbb{N}$ depending on $\varepsilon > 0$ which is independent of x. This is why (4.9), i.e.

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

says that $f_n(x)$ converges uniformly to $f(x)$. The equivalent[8] statement of uniform convergence is

$$\forall_{\varepsilon>0}\, \exists_{N\in\mathbb{N}}\, \underbrace{\forall_{x\in[a,b]}}_{\text{uniform}}\, \forall_{n\geq N} \,:\, |f_n(x) - f(x)| \leq \varepsilon. \tag{4.11}$$

Uniform convergence is stronger than pointwise convergence, which[9] only says that

$$\forall_{\varepsilon>0}\, \underbrace{\forall_{x\in[a,b]}}_{\text{pointwise}}\, \exists_{N\in\mathbb{N}}\, \forall_{n\geq N} \,:\, |f_n(x) - f(x)| \leq \varepsilon, \tag{4.12}$$

or equivalently[10] that

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

and allows N to depend on both $\varepsilon > 0$ and $x \in [a, b]$, thereby weakening the statement made by (4.9) which has N depending on $\varepsilon > 0$ only.

Remark 4.35. *The uniform statement (4.11) and the non-uniform pointwise statement (4.12) differ by only one $\forall$-$\exists$ swop.*

Another example of a nonuniform versus a uniform ε-statement is continuity in every $\xi \in I$, see (4.1). The non-uniform statement is

$$\underbrace{\forall_{\xi\in I}}_{\text{every } \xi}\, \forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{x\in I} \,:\, |x - \xi| < \delta \implies |f(x) - f(\xi)| < \varepsilon,$$

with δ depending on both ε and ξ, and $\forall_{\xi\in I}$ is at the front. Equivalently:

$$\forall_{\varepsilon>0}\, \underbrace{\forall_{\xi\in I}}_{\text{every } \xi}\, \exists_{\delta>0}\, \forall_{x\in I} \,:\, |x - \xi| < \delta \implies |f(x) - f(\xi)| < \varepsilon,$$

[8] There is no difference between $\forall_{n\geq N}\, \forall_{x\in[a,b]}$ and $\forall_{x\in[a,b]}\, \forall_{n\geq N}$.

[9] Interchanging $\forall_{x\in[a,b]}$ and $\exists_{N\in\mathbb{N}}$.

[10] There is no difference between $\forall_{\varepsilon>0}\, \forall_{x\in[a,b]}$ and $\forall_{x\in[a,b]}\, \forall_{\varepsilon>0}$.

The uniform statement (one $\forall$-$\exists$ swop) is

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\ \underbrace{\forall_{\xi\in I}}_{\text{uniform}}\ \forall_{x\in I}\ :|x-\xi|<\delta\implies|f(x)-f(\xi)|<\varepsilon, \tag{4.13}$$

δ depending only on ε. Equivalently:

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\,\forall_{x\in I}\ \underbrace{\forall_{\xi\in I}}_{\text{uniform}}\ :|x-\xi|<\delta\implies|f(x)-f(\xi)|<\varepsilon,$$

in which the $\forall_{\xi\in I}$ is at the end of the quantor sequence that precedes the implication.

Exercise 4.36. Let $I=\mathbb{R}$ 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 geometrically what the implication in (4.13) says.

We prefer to pull $\forall_{x\in I}$ and $\forall_{\xi\in I}$ together in the uniform statement of continuity, and we write y for ξ in the definition we memorize.

Definition 4.37. *Let $f:I\to\mathbb{R}$. Then f is called uniformly continuous on I if*

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\,\forall_{x,y\in I}\ :|x-y|<\delta\implies|f(x)-f(y)|<\varepsilon.$$

Exercise 4.38. Let $f(x)=2x+1$. Prove that f is uniformly continuous on $\mathbb{R}$.

Exercise 4.39. Let $f(x)=x^2$ and $I=(0,1)$. Prove that f is uniformly continuous on I. Is f uniformly continuous on $\mathbb{R}$?

Exercise 4.40. Let$f(x)=\frac{1}{x}$ and $I=(1,\infty)$. Prove that f is uniformly continuous on I. Is f uniformly continuous on $(0,1)$?

A function may be uniformly continuous because it satisfies an explicit estimate such as (3.7) with $\theta>0$, in which case δ can be chosen via $\theta\delta=\varepsilon$. An example with $\theta=1$ is the absolute value function, see Exercise 5.36. A pointwise continuous function f may be automatically uniformly continuous without any further assumptions on f if the domain has a nice property that we introduce in the proof of the next theorem.

Theorem 4.41. *Let $f : [a, b] \to \mathbb{R}$ be continuous in every $\xi \in [a, b]$. Then $f : [a, b] \to \mathbb{R}$ is uniformly continuous on $[a, b]$.*

The proof argues from the negation of the statement in Definition 4.37. This negation reads

$$\exists_{\varepsilon>0}\, \forall_{\delta>0}\, \exists_{x,y\in I} \,: |x - y| < \delta \quad \text{and} \quad |f(x) - f(y)| \geq \varepsilon,$$

obtained by 3 $\exists$-$\forall$ replacements and the negation of the final statement, which here is an implication. This negation has to be used to derive a contradiction. The $\exists_{\varepsilon>0}$ leaves no choice, but the subsequent $\forall_\delta$ is only used in full strength by taking infinitely many smaller and smaller values of δ, for instance $\delta = \frac{1}{n}$, with $n \in \mathbb{N}$. Then the negation provides us with $x_n, y_n \in [a, b]$ for which it holds that

$$|x_n - y_n| < \frac{1}{n} \quad \text{and} \quad |f(x_n) - f(y_n)| \geq \varepsilon.$$

Both x_n and y_n are bounded sequences. As in the proof of Theorem 4.29 it is Theorem 3.10 that provides us with a converging subsequence x_{n_k} having limit $\bar{x} \in [a, b]$ and the continuity of f in $\bar{x}$ implies that $f(x_{n_k}) \to f(\bar{x})$. Since

$$|x_{n_k} - y_{n_k}| < \frac{1}{n_k} \leq \frac{1}{k},$$

we also have $y_{n_k} \to \bar{x}$ and $f(y_{n_k}) \to f(\bar{x})$, whence

$$f(x_{n_k}) - f(y_{n_k}) \to 0, \quad \text{contradicting} \quad |f(x_{n_k}) - f(y_{n_k})| \geq \varepsilon \quad \text{for all } k \in \mathbb{N}.$$

This completes the proof of Theorem 4.41.

Theorem 4.41 will be essential in Chapter 5 for proving that

$$\int_a^b f(x)\, dx$$

is defined for every $f \in C([a, b])$, and other properties of the map

$$f \in C([a, b]) \to \int_a^b f(x)\, dx \in \mathbb{R} = \int_a^b$$

as well as the map

$$f \in C([a, b]) \to F \in C([a, b]) \quad \text{defined by} \quad F(x) = \int_a^x,$$

which will be contractive if $b - a < 1$.

4.6 Uniform convergence and equicontinuity

If f_n is a sequence of functions defined on a subset X of $\mathbb{R}$ we can[11] speak of continuity of f_n in a point $x_0 \in X$ which is uniform in n, and continuity of f_n which is uniform in both x and n.

Definition 4.42. *Let $f_n : X \to \mathbb{R}$ be a sequence of functions. Then f_n is called uniformly equicontinuous on X if*

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\,\forall_{x,y\in I}\,\forall_{n\in\mathbb{N}} : \underbrace{|x-y|}_{d(x,y)} < \delta \implies |f_n(x) - f_n(y)| < \varepsilon.$$

Remark 4.43. *This definition carries over to sequences of functions on any metric space X. The theorem below carries over to sequentially compact X, but is formulated for $X = [a, b] \subset \mathbb{R}$. The policeman's lemma is needed for the general case.*

Theorem 4.44. *(Arzéla-Ascoli) Let $f_n : [a, b] \to \mathbb{R}$ be a bounded sequence of uniformly equicontinuous functions. Then f_n has a convergent subsequence in $C([a, b])$ with limit $f \in C([a, b])$.*

The proof uses Theorem 3.10, which provides us with pointwise convergent subsequences. For simplicity assume that $a = 0$ and $b = 1$. The sequence $f_n(0)$ is bounded in $\mathbb{R}$ and therefore contains a convergent subsequence. Denote the limit by $f(0)$. By Theorem 3.10 this subsequence of f_n contains a further subsequence which converges in $x = 1$ as well. Denote the limit by $f(1)$. Along another further subsequence $f_n(\frac{1}{2})$ converges. The limit defines $f(\frac{1}{2})$. And so on.

We conclude that we can define the values of a desired limit function f in $0, 1, \frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8}, \ldots$, and if the indices[12] of all these subsequences are given by

$$\begin{array}{llllllll}
n_{11} & n_{12} & n_{13} & n_{14} & n_{15} & n_{16} & \ldots & \text{for convergence in } 0 \text{ only,} \\
n_{21} & n_{22} & n_{23} & n_{24} & n_{25} & n_{26} & \ldots & \text{for convergence also in } 1, \\
n_{31} & n_{32} & n_{33} & n_{34} & n_{35} & n_{36} & \ldots & \text{for convergence also in } \dfrac{1}{2}, \\
n_{41} & n_{42} & n_{43} & n_{44} & n_{45} & n_{46} & \ldots & \text{for convergence also in } \dfrac{1}{4}, \\
n_{51} & n_{52} & n_{53} & n_{54} & n_{55} & n_{56} & \ldots & \text{for convergence also in } \dfrac{3}{4},
\end{array}$$

[11] We won't consider pointwise equicontinuiy.

[12] Have a look at the proof Theorem 1.2.

$$\begin{array}{llllllll} n_{61} & n_{62} & n_{63} & n_{64} & n_{65} & n_{66} & \dots & \text{for convergence also in } \frac{1}{8}, \\ n_{71} & n_{72} & n_{73} & n_{74} & n_{75} & n_{76} & \dots & \text{for convergence also in } \frac{3}{8}, \\ n_{81} & n_{82} & n_{83} & n_{84} & n_{85} & n_{86} & \dots & \text{for convergence also in } \frac{5}{8}, \\ n_{91} & n_{92} & n_{93} & n_{94} & n_{95} & n_{96} & \dots & \text{for convergence also in } \frac{7}{8}, \end{array}$$

and so on, then each of these sequences is a subsequence of the previous sequence, and has the diagonal subsequence f_{kk} as a further subsequence. It follows that

$$f_{kk}(a) \to f(a)$$

for every

$$a \in A = \{\frac{j}{2^m} : j, m \in \mathbb{N}_0,\, j \leq 2^m\}.$$

Exercise 4.45. Prove that $f : A \to \mathbb{R}$ has the property that

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{a,b\in A} : |a-b| < \delta \implies |f(a)-f(b)| \leq \varepsilon.$$

Hint: Let δ correspond to ε as in Definition 4.42 and estimate

$$|f(a)-f(b)| \leq \underbrace{|f(a)-f_{kk}(a)|}_{\to 0 \quad \text{as } k\to\infty} + \underbrace{|f_{kk}(a)-f_{kk}(b)|}_{<\varepsilon \quad \text{if } |a-b|<\delta} + \underbrace{|f_{kk}(b)+f(b)|}_{\to 0 \quad \text{as } k\to\infty}.$$

Exercise 4.46. Prove that $f : A \to \mathbb{R}$ as in Exercise 4.45 extends uniquely to $f : [0,1] \to \mathbb{R}$ satisfying

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{x,y\in[0,1]} : |x-y| < \delta \implies |f(x)-f(y)| \leq \varepsilon.$$

Hint: have a look at Theorem 4.53 and also prove Theorem 4.55.

Exercise 4.47. Completion of the proof of Theorem 3.10: let $f : [0,1] \to \mathbb{R}$ be as in Exercise 4.46. Prove that $f_{kk} \to f$ in $C([0,1])$. Hint: establish

$$|f_{kk}(x)-f(x)| \leq \underbrace{|f_{kk}(x)-f_{kk}(a)|}_{<\varepsilon \quad \text{if } |x-a|<\delta} + |f_{kk}(a)-f(a)| + \underbrace{|f(a)-f(x)|}_{\leq\varepsilon \quad \text{if } |x-a|<\delta} < 2\varepsilon$$

for all $x \in [0,1]$ by a δ-dependent choice of a finite subset $A_\delta \subset A$ such that each $x \in [0,1]$ has $|x-a| < \delta$ for some x-dependent $a \in A_\delta$, and the convergence of $f_{kk}(a)$ to $f(a)$ for all $a \in A_\delta$.

4.7 Over the top

The arguments in Sections 4.4 and 4.5 about continuous functions generalise to continuous functions $f : X \to \mathbb{R}$ on so-called (sequentially) compact metric spaces X, of which $[a, b]$ is the first example. We recall the definition of continuity via sequences.

Definition 4.48. *Let X be a metric space, $f : X \to \mathbb{R}$ a function and $\xi \in X$. Then f is called continuous in ξ if $f(x_n) \to f(\xi)$ for every sequence x_n in X with $x_n \to \xi$. The equivalent ε-δ-statement is*

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{x\in X} \,:\, d(x, y) < \delta \implies |f(x) - f(y)| < \varepsilon.$$

Definition 4.49. *Let X be a metric space and $f : X \to \mathbb{R}$ a function. Then f is called uniformly continuous on X if*

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{x,y\in X} \,:\, d(x, y) < \delta \implies |f(x) - f(y)| < \varepsilon.$$

Definition 4.50. *A metric space X is compact if every sequence in X has a convergent subsequence.*

A compact metric space is complete, but there are many[13] noncompact complete metric spaces.

Theorem 4.51. *If X is a compact metric space and $f : X \to \mathbb{R}$ is continuous then f is uniformly continuous on X and f has[14] a global maximum and minimum on X.*

Theorem 4.52. *If X is a compact metric space and Y a complete metric space, and $\Phi : X \to Y$ is continuous then Φ is uniformly continuous on X. The metric defined by*

$$d(\Phi, \Psi) = \max_{x\in X} d_{_Y}(\Phi(x), \Psi(x))$$

makes $C(X, Y)$, the space of all continuous maps from X to Y a complete metric space.

Theorem 4.53. *Let X be a metric space and $f : X \to \mathbb{R}$ a uniformly continuous function. Suppose that $x_n \in X$ is a Cauchy sequence in X. Then $f(x_n)$ is a Cauchy sequence in $\mathbb{R}$ and thereby convergent. Any other Cauchy sequence y_n in X with $d(x_n, y_n) \to 0$ has $f(y_n)$ convergent with the same limit as $f(x_n)$.*

[13] Examples: $\mathbb{R}$, $\{f \in C([a, b]) : |f|_\infty \leq 1\}$.
[14] A striking application: Section 12.1.

The proof is easy, we have to make sure that $|f(x_m) - f(x_n)| < \varepsilon$ if $\varepsilon > 0$ is given. The uniform continuity says there exists $\delta > 0$ such that this is the case if $d(x_m, x_n) < \delta$. But the sequence being Cauchy allows to take this δ as ε in the definition of Cauchy sequence and conclude that there exists $N \in \mathbb{N}$ such that $d(x_m, x_n) < \delta$ for all $m, n \geq N$, and thereby also $|f(x_m) - f(x_n)| < \varepsilon$.

Exercise 4.54. Prove the last part of Theorem 4.53 about the sequence $f(y_n)$.

Theorem 4.55. *Let X be a complete metric space, $A \subset X$ and $f : A \to \mathbb{R}$ uniformly continuous. Then f extends uniquely to a uniformly continuous function from $\bar{A}$ to $\mathbb{R}$.*

Exercise 4.56. Formulate and prove Theorem 4.44 for X a sequentially compact metric space, Y a complete Banach space, $\Phi_n : X \to Y$ a bounded sequence of uniformly equicontinuous maps.

5 Integration

Let $a, b \in \mathbb{R}$ and $f : [a, b] \to \mathbb{R}$ 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

$$A_- = \{(x, y) \in \mathbb{R}^2 : f(x) < y < 0,\, a < x < b\}$$

as subsets of

$$\mathbb{R}^2 = \{(x, y) \in \mathbb{R}^2 :\}.$$

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

$$\int_a^b f(x)\, dx = |A_+| - |A_-|.$$

Here we will not bother to define the area of general subsets of the plane, but opt for a definition of the integral only, a definition that will not make you to uncomfortable in relation to what your intuition says the area of the sets $|A_+|$ and $|A_-|$ should be.

Starting point is the consensus that the area of the square in the xy-plane bounded by the lines $x = 0$, $x = 1$, $y = 0$, $y = 1$ is equal to 1, and that line segments and graphs $y = f(x)$ of nice functions $f : [a, b] \to \mathbb{R}$ have zero area. Thus the areas $|A_+|$ and $|A_-|$ do not change if every $<$ in the definition of A_+ and A_- is replaced by $\leq$.

5.1 Integrals of monomials

We first examine the case that $A_- = \emptyset$ and consider the example

$$f(x) = x^2.$$

We should come with a definition of

$$I_2 = \int_0^1 x^2\, dx$$

that coincides with what we think is the area of

$$A_2 = \{(x, y) \in \mathbb{R}^2 : 0 \leq y \leq x^2 \leq 1\}$$

in the Cartesian xy-plane.

Now evaluate $y = x^2$ with

$$0 = \frac{0}{10}, \frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \frac{4}{10}, \frac{5}{10}, \frac{6}{10}, \frac{7}{10}, \frac{8}{10}, \frac{9}{10}, \frac{10}{10} = 1.$$

These particular x-values give you points in the unit square S, the square consisting of all points (x, y) with $0 \leq x \leq 1$ and $0 \leq y \leq 1$. We agreed that the area of S is equal to 1. Look at the region A_2 in S bounded by $y = 0$, $x = 1$ and $y = x^2$. Can you convince yourself that the area $|A_2|$ of A_2 is less than

$$\frac{1}{1000} + \frac{4}{1000} + \frac{9}{1000} + \frac{16}{1000} + \frac{25}{1000} + \frac{36}{1000} + \frac{49}{1000} + \frac{64}{1000} + \frac{81}{1000} + \frac{100}{1000},$$

but more than

$$\frac{1}{1000} + \frac{4}{1000} + \frac{9}{1000} + \frac{16}{1000} + \frac{25}{1000} + \frac{36}{1000} + \frac{49}{1000} + \frac{64}{1000} + \frac{81}{1000}?$$

To do so you have to relate every term in the sums above to the area of a rectangle[1] with 4 corner points, one of which is the point on the curve. There are two obvious ways to do so, providing you with the above (upper and lower) sums[2] for the area of A_2. If this worked out, you can also convince yourself that

$$|A_2| < \frac{1}{n^3} \sum_{k=1}^{n} k^2 < \frac{1}{3} + \frac{1}{n}$$

for every natural number n. As discussed in Chapter 1 we know[3] that

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

It follows that

$$|A_2| < \frac{1}{n^3} \sum_{k=1}^{n} k^2 = \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2} < \frac{1}{3} + \frac{1}{n}$$

Likewise we have

$$|A_2| > \frac{1}{n^3} \sum_{k=1}^{n-1} k^2 = \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2} - \frac{1}{n} = \frac{1}{3} - \frac{1}{2n} + \frac{1}{6n^2} > \frac{1}{3} - \frac{1}{n}.$$

[1] Parallel to the axes.

[2] These are called Riemann sums.

[3] We proved it using the domino principle.

Combining it follows that

$$\forall_{n\in\mathbb{N}}: \frac{1}{3}-\frac{1}{n}<|A_2|<\frac{1}{3}+\frac{1}{n}, \tag{5.1}$$

which should leave no other possibility then

$$|A_2|=\frac{1}{3}.$$

Exercise 5.1. Use Theorem 1.3 to show that (5.1) implies that $|A_2|=\frac{1}{3}$. Hint: read Section 1.1 again.

Thus it is really the Archimedean principle which implies that the area[4] of the set

$$A_2=\{(x,y)\in\mathbb{R}^2: 0\le y\le x^2,\, x\le 1\}$$

can only be $\frac{1}{3}$, the same number[5] $\frac{1}{3}$ as in the formula for the volume of a pyramid.

Now consider

$$I_p=\int_0^1 x^p\,dx,$$

in which p can be any positive integer. The mathematical definition of this integral in Section 5.2 below will imply that I_p is equal to what we understand to be to the area of the region

$$A_p=\{(x,y)\in\mathbb{R}^2: 0\le y\le x^p,\, x\le 1\}$$

in the plane. With $p=1$ this integral is the area of the triangle A_1, and geometrically is is clear that this area is half of the area of A_0 and therefore equal to $\frac{1}{2}$.

All other values of p require calculus to compute I_p. Here's a nice approach based on the identities

$$a^{p+1}-b^{p+1}=(a-b)(a^p+\cdots+b^p)\quad(p\in\mathbb{N}), \tag{5.2}$$

which generalise

$$a^2-b^2=(a-b)(a+b).$$

Actually these identities will also be our starting point for differential calculus, with the functions f_p defined by $f_p(x)=x^p$ as first examples in Section 7.4.

[4] Once properly defined in a course on measure theory.

[5] As a number abolished in Dutch educational law.

Exercise 5.2. Fill in the dots in (5.2) for $p = 3, 4, 5$, and complete the expression below:

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

Exercise 5.3. Assume $a > b > 0$. Show from Exercise 5.2 that

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

and reflect on the limit case $a = b$ in relation to the derivative of x^{p+1} with respect to x at $x = a$.

Exercise 5.4. Put $a = k + 1$ and $b = k$ in Exercise 5.3 and take the sum over $k = 0, 1, \dots, n - 1$ to show that[6]

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

for $p, n \in \mathbb{N}$.

Exercise 5.5. Use Exercise 5.4 to show that

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

even before we have defined areas or integrals.

5.2 Integrals of monotone functions via finite sums

For $a, b \in \mathbb{R}$ with $a < b$ and $f : [a, b] \to \mathbb{R}$, with $f(x) \geq 0$ for all $a \leq x \leq b$, the definition of

$$\int_a^b f \; = \int_a^b f(x) \, dx \tag{5.3}$$

gives a meaning to the area $|A|$ of the set

$$A = \{(x, y) \in \mathbb{R}^2 : 0 \leq y \leq f(x), \, a \leq x \leq b\}. \tag{5.4}$$

We first assume that $f : [a, b] \to \mathbb{R}$ is (both nonnegative and) nondecreasing.

[6] Frits Beukers showed me this neat trick.

Definition 5.6. *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 $x_1, x_2 \in [a, b]$.

For such f we consider left endpoint sums[7]

$$L_N = \sum_{k=1}^{N} f(x_{k-1})(x_k - x_{k-1}) \tag{5.5}$$

$$= f(x_0)(x_1 - x_0) + \cdots + f(x_{N-1})(x_N - x_{N-1}),$$

in which

$$a = x_0 \leq x_1 \leq \cdots \leq x_N = b \quad \text{with} \quad N \in \mathbb{N}. \tag{5.6}$$

Such a choice of $x_1, \ldots, x_{N-1}$ is called a *partition* P of $[a, b]$, and each term in (5.5) is the area of a rectangle[8]

$$(x_{k-1}, x_k) \times (0, f(x_{k-1})) = \{(x, y) \in \mathbb{R}^2 : 0 < y < f(x_{k-1}),\ x_{k-1} < x < x_k\}$$

contained in A. These rectangles are mutually disjoint. Thus any sensible definition of $|A|$ should come with the conclusion that

$$L_N \leq |A|,$$

because we used the values of f in every left endpoint of the intervals $[x_{k-1}, x_k]$ defined by the partition. We therefore say that L_N is a lower sum for $|A|$ and write

$$\underline{S}_N = L_N.$$

In the same fashion the rectangles[9]

$$[x_{k-1}, x_k] \times [0, f(x_k)] = \{(x, y) \in \mathbb{R}^2 : 0 \leq y \leq f(x_k),\ x_{k-1} \leq x \leq x_k\}$$

with k running from 1 to N cover A completely, so the right endpoint sums

$$R_N = \sum_{k=1}^{N} f(x_k)(x_k - x_{k-1}) \tag{5.7}$$

[7] Make a sketch in which you see what these sums are.
[8] Possibly empty, if $x_{k-1} = x_k$ of $f(x_{k-1}) = 0$.
[9] Possibly reducing to line segments or points with zero area.

should have

$$R_N \geq |A|$$

and we say that

$$\bar{S}_N = R_N$$

is an upper sum for $|A|$.

We conclude that the number $|A|$ we are looking for must have

$$\underline{S}_N = L_N \leq |A| \leq R_N = \bar{S}_N \tag{5.8}$$

for all possible choices of the partition P. Since for equidistant partitions, i.e. partitions with

$$x_k - x_{k-1} = \frac{b-a}{N},$$

we have[10]

$$\bar{S}_N - \underline{S}_N = R_N - L_N = \sum_{k=1}^{N} (f(x_k) - f(x_{k-1})) \frac{b-a}{N} = (f(b) - f(a)) \frac{b-a}{N} \tag{5.9}$$

with $N \in \mathbb{N}$ arbitrary, there is at most one such number.

Exercise 5.7. In general the sequences $\underline{S}_N = L_N$ and $\bar{S}_N = R_N$ may not be monotone. Show that restricting to $N = 2^n$ we have

$$L_1 \leq L_2 \leq L_4 \leq L_8 \leq \cdots \leq R_8 \leq R_4 \leq R_2 \leq R_1.$$

and that[11]

$$\sup_{n \in \mathbb{N}} \underline{S}_{2^n} = \sup_{n \in \mathbb{N}} L_{2^n} = \inf_{n \in \mathbb{N}} R_{2^n} = \inf_{n \in \mathbb{N}} \bar{S}_{2^n}.$$

Exercise 5.8. Verify that for nonincreasing nonnegative functions the story is the same, except for reversed roles of the left and right endpoints: $L_N = \bar{S}_N$, $R_N = \underline{S}_N$, with minor changes in (5.9).

Since we never used the nonnegativity of f in the arguments we can now define the integral for every monotone function $f : [a, b] \to \mathbb{R}$:

[10] This finite sum is called a telescoping sum.

[11] We'll come back to this issue.

Definition 5.9. *For $f : [a,b] \to \mathbb{R}$ monotone*[12] *the integral of f from a to b is by definition*

$$\int_{[a,b]} f = \int_a^b f = \int_a^b f(x)\,dx = \sup_{n\in\mathbb{N}} \underline{S}_{2^n} = \inf_{n\in\mathbb{N}} \bar{S}_{2^n},$$

defined via (5.5) and (5.7) as in Exercises 5.7 and 5.8.

In the notation x is a *dummy* variable, which may be replaced by any other letter[13].

We used a very special class of partitions to identify the real number that we want assign to the integral of a monotone function from $[a,b]$ to $\mathbb{R}$. What about other partitions than equidistant ones?

Exercise 5.10. Let again $f : [a,b] \to \mathbb{R}$ be nondecreasing, let P be a partition as in (5.6) and let another partition Q be given by

$$a = y_0 \le y_1 \le \cdots \le y_M = b \quad \text{with} \quad M \in \mathbb{N}.$$

Take the upper sum[14]

$$\bar{S} = \sum_{k=1}^{N} f(x_k)(x_k - x_{k-1})$$

as in (5.7) and the lower sum

$$\underline{S} = \sum_{l=1}^{M} f(y_{l-1})(y_l - y_{l-1}).$$

Show that $\underline{S} \le \bar{S}$. Hint: try this by yourself first, and then have a look at (5.18).

Theorem 5.11. *Let $f : [a,b] \to \mathbb{R}$ be nonincreasing*[15]*. Then there is a unique $I \in \mathbb{R}$ such that*

$$\underline{S} \le I \le \bar{S}$$

for every lower sum $\underline{S} = L_N$ as in (5.5) and every upper sum as in (5.7). This real number I is by definition the integral of $f(x)$ from $x = a$ to $x = b$, notation

$$I = \int_a^b f(x)\,dx.$$

[12] Meaning that f is either nondecreasing or nonincreasing on $[a,b]$.
[13] Preferably not a, b, d or f.
[14] No index in the notation here.
[15] The statement for nondecreasing functions is similar.

The proof relies on (5.9), complemented by Exercise 5.10. Let A be the nonempty set consisting of all $\underline{S}$ and B the nonempty set consisting of all $\bar{S}$. Then $A \leq B$ in the sense that $\alpha \leq \beta$ for all $\alpha \in A$ and all $\beta \in B$, whence both $\sup A$ and $\inf B$ exist.

Can it be that

$$\sup A > \inf B?$$

If so, then $\inf B$ is not an upper bound for A because $\sup A$ is the lowest upper bound for A, so there exists $\alpha \in A$ with

$$\alpha > \inf B.$$

Since $\inf B$ is the largest lower bound for B, α is not a lower bound for B. So there exists

$$\beta \in B$$

with $\beta < \alpha$, contradicting $A \leq B$.

Can it be that

$$\sup A < \inf B?$$

Then $\varepsilon = \inf B - \sup A > 0$, whence

$$\beta - \alpha \geq \inf B - \sup A = \varepsilon > 0$$

for all $\alpha \in A$ and all $\beta \in B$. But (5.9) provides us with $\alpha \in A$ and $\beta \in B$ such that also

$$0 < \varepsilon \leq \beta - \alpha \leq (f(b) - f(a))\frac{b-a}{N}.$$

It would follow that

$$0 < \frac{\varepsilon}{(b-a)(f(b)-f(a))} \leq \frac{1}{N}$$

for all $N \in \mathbb{N}$, a contradiction with Archimedes' principle, which completes the proof.

5.3 Scaling arguments and the natural logarithm

Exercise 5.12. 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 $p + 1$. Show this by scaling the units on the axes. In integral notation this says that

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

Exercise 5.13. 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

$$\int_a^b x^p \, dx = [\frac{x^p}{p+1}]_a^b = \frac{b^{p+1}}{p+1} - \frac{a^{p+1}}{p+1}.$$

Explain why.

Definition 5.14. *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 5.15. Let $0 < a < b$, $c > 0$ and $f : [a, b] \to \mathbb{R}$ monotone. Show directly from the definitions that

$$\int_a^b f(x) \, dx = \frac{1}{c} \int_{ca}^{cb} f(\frac{x}{c}) \, dx,$$

and use this for $x > 1$ and $y > 1$ to write the integral for $\ln y$ as an integral from x to xy. Conclude that

$$\ln xy = \ln x + \ln y,$$

and prove this identity for all $x, y \in \mathbb{R}^+$. Hint: show separately that

$$\ln x + \ln \frac{1}{x} = 0$$

for all $x > 0$.

Exercise 5.16. It follows that $\ln$ is a strictly increasing function on $\mathbb{R}^+$. Prove and use

$$\ln n \geq \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n} = \sum_{k=2}^{n} \frac{1}{k}$$

to show that $\ln x \to \infty$ as $x \to \infty$. What can you conclude for $x \to 0$?

5.4 Integrals of uniformly continuous functions

So monotone functions are integrable via lower and upper sums which use the values of f in the endpoints of the intervals defined by (5.6). Every (other) choice of function values $f(\xi_k)$ with

$$x_{k-1} \le \xi_k \le x_k \quad (k = 1, \dots, N)$$

also produces a sum

$$S = \sum_{k=1}^{N} f(\xi_k)(x_k - x_{k-1}). \tag{5.10}$$

If $f : [a, b] \to \mathbb{R}$ fails to be monotone, none of these may be a lower of upper sum, but they all relate to what $\int_a^b f(x)\,dx$ should be. We now reason starting from the assumption that f is bounded[16] on $[a, b]$, i.e. the range

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

of f is a bounded set. Given (5.6) there are then[17] $m_k, M_k \in \mathbb{R}$ such that

$$m_k \le f(x) \le M_k \quad \text{for all} \quad x \in [x_{k-1}, x_k], \tag{5.11}$$

for $k - 1, \dots, N$. Then

$$\underline{S} = \sum_{k=1}^{N} m_k(x_k - x_{k-1}) \le \sum_{k=1}^{N} f(\xi_k)(x_k - x_{k-1}) \le \sum_{k=1}^{N} M_k(x_k - x_{k-1}) = \bar{S}, \tag{5.12}$$

so varying the choice of (5.6) the $\underline{S}$ and $\bar{S}$ are lower and upper sums for the integral $\int_a^b f(x)\,dx$ we wish to define, which will force the other sums S to follow if refining the partition P the lower and upper sums get closer and closer to the unique real number I we hope for.

Of course the optimal choice for m_k and M_k, see Theorem 4.29, is

$$m_k = \inf_{[x_{k-1}, x_k]} f = \min\{f(x) : x \in [x_{k-1}, x_k]\}; \tag{5.13}$$

$$M_k = \max_{[x_{k-1}, x_k]} f = \sup\{f(x) : x \in [x_{k-1}, x_k]\}. \tag{5.14}$$

[16] A monotone function from $[a, b]$ to $\mathbb{R}$ is trivially bounded.

[17] If f is not bounded on $[a, b]$ this statement is wrong for at least one k.

Denoting the set of all lower sums $\underline{S}$ by A, and the set of all upper sums $\bar{S}$ by B, we have[18] as in Exercise 5.10 that, $A \leq B$, both sets nonempty, and hence $\sup A \leq \inf B$.

We then need an estimate of the form

$$\bar{S} - \underline{S} = \sum_{k=1}^{N} (M_k - m_k)(x_k - x_{k-1}) \leq \varepsilon(b-a) \tag{5.15}$$

to guarantee that

$$\sup A = \inf B,$$

leading to

$$\int_a^b f(x)\,dx = I = \sup A = \inf B$$

as the natural definition of the integral, just like before for monotone functions: there is only one number I such that $\alpha \leq I \leq \beta$ for all $\alpha \in A$ and all $\beta \in B$.

We will be done if for all $\varepsilon > 0$ a partition P can be chosen such that

$$M_k - m_k < \varepsilon \tag{5.16}$$

for $k = 1, \ldots, N$, which happens to be the case if f is uniformly continuous. We recall the definition of uniform continuity as

Definition 5.17. *Let $a, b \in \mathbb{R}$ with $a < b$ and $f : [a,b] \to \mathbb{R}$. Then f is called uniformly continuous if for every $\varepsilon > 0$ a $\delta > 0$ can be found such that for all $x, y \in [a,b]$ it holds that*

$$|x - y| < \delta \implies |f(x) - f(y)| < \varepsilon.$$

Exercise 5.18. Let $a, b \in \mathbb{R}$ with $a < b$ and $f : [a,b] \to \mathbb{R}$ uniformly continuous. Show that the estimate (5.16) and thereby (5.15) is established for any partition with

$$\max_{k=1,\ldots,N} |x_k - x_{k-1}| < \delta,$$

provided we take

$$m_k = \min_{[x_{k-1}, x_k]} f \quad \text{and} \quad M_k = \max_{[x_{k-1}, x_k]} f.$$

Verify that we have proved[19] the following theorem once we know for sure that the set A of all lower sums and the set B of all upper sums are ordered by $A \leq B$.

[18] Details to follow in Section 5.5.

[19] Have a look at Section 6.5 for a different approach towards $\int_a^b f(x)\,dx$ for $f \in C([a,b])$.

Theorem 5.19. *Let $a, b \in \mathbb{R}$ with $a < b$ and $f : [a, b] \to \mathbb{R}$ uniformly continuous. Then there exists a unique number $I \in \mathbb{R}$ such that*

$$\underline{S} = \sum_{k=1}^{N} m_k(x_k - x_{k-1}) \leq I \leq \sum_{k=1}^{N} M_k(x_k - x_{k-1}) = \bar{S}, \tag{5.17}$$

for all partitions (5.6) and numbers m_k, M_k such that (5.11) holds for all $k = 1, \ldots, N$. The number I is called the integral of f over $[a, b]$, and is usually denoted by

$$I = \int_a^b f(x)\, dx.$$

5.5 Integrability: general definition and technicalities

Let us begin this section by showing that for the set of lower sums A and the set of upper sums B it indeed holds that $A \leq B$, no matter which bounded function $f : [a, b] \to \mathbb{R}$ we consider. For P as in (5.6) and Q as in Exercise 5.10, the common refinement

$$a = z_0 \leq z_1 \leq \cdots \leq z_K = b, \tag{5.18}$$

is obtained by putting $x_1 \leq \cdots \leq x_{N-1}$ and $y_1 \leq \cdots \leq y_{M-1}$ in joint order. So $K - 1 = M - 1 + N - 1$ and every z_i is either an x_k or a y_l. Via the optimal choices

$$m_l = \inf_{[y_{l-1}, y_l]} f, \quad \tilde{m}_i = \inf_{[z_{i-1}, z_i]} f \leq \sup_{[z_{i-1}, z_i]} f = \tilde{M}_i, \quad M_k = \sup_{[x_{k-1}, x_k]} f$$

we have

$$\sum_{l=1}^{M} m_l(y_l - y_{l-1}) \leq \sum_{i=1}^{K} \tilde{m}_i(z_i - z_{i-1}) \leq \sum_{i=1}^{K} \tilde{M}_l(z_i - z_{i-1}) \leq \sum_{k=1}^{N} M_k(x_k - x_{k-1})$$

for the lower sum obtained obtained from Q using the infima[20] of f on every interval and the upper sum obtained from P using the suprema[21] of f.

It follows for every lower sum $\underline{S}$ and every upper sum $\bar{S}$ that $\underline{S} \leq \bar{S}$. Copy/paste of the statement in Theorem 5.19 provides us with the natural definition for Riemann integrability of bounded functions on $[a, b]$.

[20] The optimal choice from below.
[21] The optimal choice from above.

Definition 5.20. *Let $a, b \in \mathbb{R}$ with $a < b$ and $f : [a, b] \to \mathbb{R}$ a bounded function. Suppose there exists a unique number $I \in \mathbb{R}$ such that*

$$\underline{S} = \sum_{k=1}^{N} m_k(x_k - x_{k-1}) \leq I \leq \sum_{k=1}^{N} M_k(x_k - x_{k-1}) = \bar{S},$$

for all partitions (5.6) and numbers m_k, M_k such that (5.11) holds for all $k = 1, \ldots, N$. Then the number I is called the integral of f over $[a, b]$, which we write as

$$I = \int_a^b f(x)\, dx.$$

Exercise 5.21. Explain why a bounded $f : [a, b] \to \mathbb{R}$ is integrable if and only if for every $\varepsilon > 0$ a partition P as in (5.6) can be found such that for a lower sum $\underline{S}$ and an upper sum $\bar{S}$ for that same partition it holds that $\bar{S} - \underline{S} < \varepsilon$.

Exercise 5.22. With

$$M_k = \sup\{f(x) : x_{k-1} < x < x_k\} \quad \text{and} \quad m_k = \inf\{f(x) : x_{k-1} < x < x_k\}$$

in Definition 5.20 we get the same class of integrable functions with the same values for the integral.

Exercise 5.23. Prove that the function f defined by $f(x) = 1$ for all $x \in \mathbb{Q}$ and $f(x) = 0$ for all $x \notin \mathbb{Q}$ is not niet integrable on $[0, 1]$.

Exercise 5.24. A bounded integrable function $f : [a, b] \to \mathbb{R}$ can be modified in the points of a partition

$$a = x_0 < x_1 < \cdots < x_N = b$$

by introducing the function $g : [a, b] \to \mathbb{R}$ defined by $g(x_k) = c_k$ for $k = 0, \ldots, N$, and $g(x) = f(x)$ for all $x \in [a, b]$ not belonging to P. Prove that $g : [a, b] \to \mathbb{R}$ is integrable and that $\int_a^b f(x)\, dx = \int_a^b g(x)\, dx$, no matter what the numbers of $c_k \in \mathbb{R}$ actually are.

Exercise 5.25. Every bounded function $f : [a, b] \to \mathbb{R}$ for which a partition

$$a = x_0 < x_1 < \cdots < x_N = b$$

exists such that f is monotone or uniformly continuous on every open interval (x_{k-1}, x_k) is integrable. Prove this statement and show that

$$\int_a^b f(x)\,dx = \sum_{k=1}^{N} \int_{a_k}^{b_k} f(x)\,dx$$

Definition 5.26. *If $a > b$ then*

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

if $f : [b, a] \to \mathbb{R}$ is bounded and integrable[22].

Exercise 5.27. 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

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

5.6 Two limit theorems

We formulate two theorems of the type

$$\text{if } f_n \to f \text{ then } \int_a^b f_n \to \int_a^b f. \tag{5.19}$$

In the next section we formulate such theorems as continuity results for the map $f \to \int_a^b f(x)\,dx$.

The first theorem below is usually not stated in the context of Riemann integrals, as it is a special case of the monotone convergence theorem in Lebesgue's theory integrals for a much larger class of functions, the so-called measurable functions.

[22] Consistent with the intuition that dx in $\int_a^b f(x)\,dx$ is negative if $a < b$.

Theorem 5.28. *Let $f_n : [a,b] \to \mathbb{R}$ be a sequence on nondecreasing functions indexed by $n \in \mathbb{N}$ and suppose that for every $x \in [a,b]$ this sequence is also nondecreasing in n. Assume that $f_n(b)$ is bounded. Then*

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

exists for every $x \in [a,b]$, the function f thus defined is also nondecreasing, and

$$\int_a^b f_n(x)\,dx$$

is a nondecreasing sequence which converges to

$$\int_a^b f(x)\,dx$$

as $n \to \infty$. Similar statements hold for $f_n : [a,b] \to \mathbb{R}$ nonincreasing in both n and x, nonincreasing in n and nondecreasing in x, and nonincreasing in x and nondecreasing in n.

For the proof consider, dropping the subscripts N, a lower sum L and an upper sum R for the limit function f as in (5.5) and (5.7), with the partition chosen such that $R - L < \varepsilon$, i.e.

$$R - \varepsilon < L.$$

This is possible in view of Exercise 5.7. Denote the lower sum for $\int_a^b f_n$ with the same partition by L_n. Then we have

$$L_n \le \int_a^b f_n \le \int_a^b f \le R,$$

while L_n is a nondecreasing sequence with $L_n \to L$ in view of $f_n(x_k) \to f(x_k)$ as $n \to \infty$ for every $k = 1, \ldots, N$. In particular it follows that for some $N \in \mathbb{N}$ it holds that

$$R - \varepsilon < L_n \le \int_a^b f_n \le \int_a^b f \le R$$

if $n \ge N$ whence

$$0 \le \int_a^b f - \int_a^b f_n < \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary this completes the proof.

The proofs for the other 3 cases are similar. Remember this result as the statement that $f_n(x)$ monotone in both n and x implies that the (5.19) holds if the limit function exists, which is the case if $f_n(a)$ and $f_n(b)$ converge, with one of the two implying the other.

Theorem 5.29. *Let $f_n : [a,b] \to \mathbb{R}$ be a sequence of bounded integrable functions indexed by $n \in \mathbb{N}$ and suppose that f_n is uniformly convergent on $[a,b]$. Then the limit*

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

defines a function f which is also bounded and integrable, and

$$\int_a^b f_n(x)\,dx \to \int_a^b f(x)\,dx$$

as $n \to \infty$.

For a proof let $\varepsilon > 0$ and take $N \in \mathbb{N}$ such that for all $n \geq N$ and all $x \in [a,b]$ it holds that

$$|f_n(x) - f(x)| \leq \varepsilon,$$

and take lower and upper sums $\underline{S}_n$ and $\bar{S}_n$ for $\int_a^b f_n$ such that

$$\bar{S}_n - \underline{S}_n < \frac{\varepsilon}{b-a}.$$

Then $\underline{\sigma}_n = \underline{S}_n - \varepsilon$ and $\bar{\sigma}_n = \bar{S}_n + \varepsilon$ are lower and upper sums for $\int_a^b f$, an integral that we need to show to exist, as well as for $\int_a^b f_n$, an integral which we assumed to exists. Since

$$\bar{\sigma}_n - \underline{\sigma}_n = \bar{S}_n + \varepsilon - \underline{S}_n + \varepsilon < 3\varepsilon$$

in which $\varepsilon > 0$ was arbitrary, it follows from Exercise 5.21 that $\int_a^b f$ exists, and

$$\underline{\sigma}_n \leq \int_a^b f \leq \bar{\sigma}.$$

But for $\int_a^b f_n$ we also have

$$\underline{\sigma}_n \leq \int_a^b f_n \leq \bar{\sigma},$$

so the difference has

$$\left|\int_a^b f_n - \int_a^b f\right| \leq \bar{\sigma}_n - \underline{\sigma}_n < 3\varepsilon \quad \text{for all } n \geq N.$$

Since $\varepsilon > 0$ was arbitrary this completes the proof.

5.7 Integrals as linear functionals

Functions defined on the same domain can be added and multiplied. For instance, if $f : [a, b] \to \mathbb{R}$ and $g : [a, b] \to \mathbb{R}$ are functions from $[a, b]$ to $\mathbb{R}$ then so are $f + g : [a, b] \to \mathbb{R}$ and $fg : [a, b] \to \mathbb{R}$ defined by

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

for all $x \in [a, b]$. We say that the functions from $[a, b]$ to $\mathbb{R}$ form an algebra that includes the constant functions and in particular the neutral elements for addition and multiplication. A sub-algebra of these functions is

$$\mathrm{RI}([a, b]) = \{f : [a, b] \to \mathbb{R} : f \text{ is bounded and integrable}\},$$

a statement that amounts to the following theorem.

Theorem 5.30.

$$f, g \in \mathrm{RI}([a, b]) \implies f + g \in \mathrm{RI}([a, b]) \quad \textit{and} \quad fg \in \mathrm{RI}([a, b]).$$

In particular $\lambda f \in \mathrm{RI}([a, b])$ for every $\lambda \in \mathbb{R}$ if $f \in \mathrm{RI}([a, b])$. We conclude that $\mathrm{RI}([a, b])$ is also a vector space over $\mathbb{R}$. There's no expression for the integral of the product in terms of the product of the integrals, but we do have the linearity of the map

$$f \to \int_a^b f(x)\,dx$$

from $\mathrm{RI}([a, b])$ to $\mathbb{R}$, which is the statement in the following theorem.

Theorem 5.31. *For $f, g \in \mathrm{RI}([a, b])$ and $\lambda \in \mathbb{R}$ it holds that*

$$\int_a^b (f(x) + g(x))\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx;$$

$$\int_a^b \lambda f(x)\,dx = \lambda \int_a^b f(x)\,dx.$$

In other words:

$$\phi(f) = \int_a^b f(x)\,dx = I, \tag{5.20}$$

I as in Definition 5.20, defines a linear map

$$\phi : \mathrm{RI}([a, b]) \to \mathbb{R}.$$

Exercise 5.32. Not so easy: prove the statement about fg in the first theorem. Hint:

$$\sup_I fg - \inf_I fg = \sup_{x,y\in I} |f(x)g(x) - f(y)g(y)|$$

for subintervals $I \subset [a,b]$; use

$$f(x)g(x) - f(y)g(y) = (f(x) - f(y))g(x) + f(y)(g(x) - g(y)),$$

estimate in terms of (bounds on) $|f|$ and $|g|$, $\sup_I f - \inf_I f$, $\sup_I g - \inf_I g$; take a partition refining two partitions chosen for f and g and conclude.

Exercise 5.33. Doable: use the inequality[23]

$$|\sum_{i=1}^{n} a_i b_i| \leq \left(\sum_{i=1}^{n} |a_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |b_i|^q\right)^{\frac{1}{q}},$$

which holds for $p, q > 1$ with

$$\frac{1}{p} + \frac{1}{q},$$

to show that

$$|\int_a^b f(x)g(x)\,dx| \leq \left(\int_a^b |f(x)|^p\,dx\right)^{\frac{1}{p}} \left(\int_a^b |g(x)|^q\,dx\right)^{\frac{1}{q}}$$

for such p and q and $f, g \in \mathrm{RI}([a,b])$. Hint: use the conclusion of Exercise 5.32 and combine the inequality for the sums with the definition of integrability via finite sums.

Exercise 5.34. Easier: prove both statements about $f+g$ in these theorems. Hint: you still need to take a partition refining two partitions chosen for f and g.

Exercise 5.35. Much easier: prove both statements about λf in these theorems.

Exercise 5.36. If $f \in \mathrm{RI}([a,b])$ then also $x \to |f(x)|$ is in $\mathrm{RI}([a,b])$, and

$$|\phi(f)| = |\int_a^b f(x)\,dx| \leq \int_a^b |f(x)|\,dx = |f|_1 \leq (b-a)|f|_\infty. \tag{5.21}$$

[23] See (22.2).

Give a proof directly from the definitions and exhibit a function $f : [a, b] \to \mathbb{R}$ not in $\mathcal{RI}([a, b])$ for which $x \to |f(x)|$ is.

An important example of a linear map between function spaces is the map

$$A : C([a, b]) \to C([a, b]) \quad \text{defined by} \quad (A(f))(x) = \int_a^x f(s)\, ds. \tag{5.22}$$

Since $F = A(f)$ satisfies

$$|F(x) - F(y)| = |\int_x^y f| \leq |f|_\infty \, |x - y|$$

for all $x, y \in [a, b]$ this map is well defined[24]. Similar to (5.21) for (5.20) it satisfies the estimate

$$|A(f)|_\infty \leq |f|_1 \leq (b - a)|f|_\infty. \tag{5.23}$$

[24] More on this primitive function F is Chapter 9.

6 Normed spaces and continuous linear maps

In (5.21) we introduced the notation for what is usually called the 1-norm of f, although we shall see shortly that it is only a norm on the smaller vector space $C([a,b])$. This norm belongs to a family of norms[1]

$$|f|_p = \left(\int_a^b |f(x)|^p \, dx \right)^{\frac{1}{p}}, \tag{6.1}$$

in which p varies from $p = 1$ to the limit case $p = \infty$ which we introduce next.

Recall that (3.10) defined a metric

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

on the set $C([a,b])$, which is contained in $\mathrm{RI}([a,b])$ in view of Theorems 4.41 and 5.19. It is not a metric on $\mathrm{RI}([a,b])$ because the maximum may not exist, but

$$d(f,g) = \sup_{a \le x \le b} |f(x) - g(x)| \tag{6.2}$$

generalises the definition, and defines a metric on the space $B([a,b])$ of all bounded functions $f : [a,b] \to \mathbb{R}$. We have

$$C([a,b]) \subset \mathrm{RI}([a,b]) \subset B([a,b]). \tag{6.3}$$

Exercise 6.1. Explain why all three spaces in (6.3) are complete.

Definition 6.2. *A vector space X over $\mathbb{R}$ is a normed space if every x has a norm $|x| \in \mathbb{R}$ such that*

$$|x| > 0 \quad \text{if} \quad x \neq 0 \quad \text{and} \quad |\lambda x| = |\lambda| \, |x|$$

for all $x \in X$ and all $\lambda \in \mathbb{R}$, and

$$|x + y| \le |x| + |y|$$

for all $x, y \in X$. If X is also an algebra[2] then X is called a normed algebra if in addition

$$|xy| \le |x| \, |y|$$

[1] The triangle inequality relies on Hölder's inequality, see Section 12.9.

[2] Vector space with multiplication and usual rules: e.g. the function spaces in (6.3).

for all $x, y \in X$. If every Cauchy sequence in X is convergent[3] *then X is called complete. A complete normed vector space over $\mathbb{R}$ is called a real Banach space. In case X is also a normed algebra it is called a (real) Banach algebra.*

All three spaces in (6.3) are normed algebras with the supnorm defined defined by

$$|f|_\infty = d(f, 0) = \sup_{a \le x \le b} |f(x)|.$$

Exercise 6.3. Prove that $|f + g|_\infty \le |f|_\infty + |g|_\infty$ and $|fg|_\infty \le |f|_\infty |g|_\infty$ for all $f, g \in B([a, b])$. Since $B([a, b])$ is complete, it is a Banach algebra, and so are its closed sub-algebras $C([a, b])$ and $\mathrm{RI}([a, b])$.

Note that

$$d(f, g) = |f - g|_\infty$$

is the metric in (6.2), which reduces to (3.10) on $C([a, b])$.

6.1 Lipschitz continuous linear maps

Refering to Theorem 5.31 and (5.20), the map

$$f \xrightarrow{\phi} \int_a^b f(x)\, dx = \phi(f)$$

from $\mathrm{RI}([a, b])$ to $\mathbb{R}$ is not only linear (Theorem 5.31), but also has the property that

$$|\phi(f)| \le (b - a)\, |f|_\infty \tag{6.4}$$

for all $f \in \mathrm{RI}([a, b])$, simply because by contruction the integral $I = \int_a^b f(x)\, dx$ has

$$(b - a) \inf_{[a,b]} f \le I \le (b - a) \sup_{[a,b]} f.$$

From (6.4) and the linearity of ϕ we have

$$|\phi(f) - \phi(g)| = |\phi(f - g)| \le (b - a)\, |f - g|_\infty = (b - a)\, d(f, g)$$

for all $f, g \in \mathrm{RI}([a, b])$, which says that ϕ is a Lipschitz continuous map from $X = \mathrm{RI}([a, b])$ to $\mathbb{R}$ with Lipschitz constant $L = b - a$.

[3] With by definition its limit contained in X.

Exercise 6.4. Let X be a normed space and $\phi : X \to \mathbb{R}$ a linear map. Assume that ϕ is continuous in a point $x_0 \in X$, formulated as[4]

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{x\in X} \,:\, |x - x_0|_X \leq \delta \implies |\phi(x) - \phi(x_0)| \leq \varepsilon. \tag{6.5}$$

Pick one $\varepsilon > 0$ and take the corresponding $\delta > 0$. Prove that ϕ is Lipschitz continuous with Lipschitz constant

$$L = \frac{\varepsilon}{\delta}.$$

In particular[5]

$$|\phi(x)| \leq L|x|_X$$

for all $x \in X$. The smallest L for which this estimate holds is called the norm of ϕ.

6.2 Dual and other spaces of continuous linear maps

Theorem 6.5. *Let X be a normed space. Then X^* is by definition the set of all Lipschitz continuous linear maps from X to $\mathbb{R}$. With the norm defined by*

$$|\phi|_{X^*} = \sup_{0\neq x\in X} \frac{|\phi(x)|}{|x|_X}$$

for $\phi \in X^*$, and linear algebra by

$$(\phi + \psi)(x) = \phi(x) + \psi(x), \quad (\lambda\phi)(x) = \lambda\phi(x) \quad (\phi, \psi \in X^*, \lambda \in \mathbb{R}, x \in X),$$

this so-called dual space X^* is a Banach space.

Exercise 6.6. Prove Theorem 6.5.

Remark 6.7. *Elements of X^* are in fact $\mathbb{R}$-valued functions on X. They are usually called functionals, to distinguish them from the functions they act on when X is a function space. Function spaces inspired the notation*

$$\phi(f) = \langle \phi, f \rangle$$

for $\phi \in X^*$ and $f \in X$, a notation which is also commonly used outside the function space context.

[4] We now choose to use the characterisation with $\leq \delta$ and $\leq \varepsilon$.

[5] This is also called boundedness of ϕ, not on X but on $B = \{x \in X : |x| \leq 1\}$.

Theorem 6.8. *Let X be a normed space with the Hahn-Banach property*[6] *that for every $x \in X$ there exists $\psi \in X^*$ with $|\psi|_{X^*} = 1$ and*

$$\psi(x) = \langle \psi, x \rangle = |x|_X .$$

Then X is isometrically isomorphic with a subspace of the dual space $X^{**} = (X^*)^*$ of its dual space X^* via the map

$$x \xrightarrow{hat} \hat{x}$$

defined by

$$\hat{x}(\phi) = \langle \hat{x}, \phi \rangle = \langle \phi, x \rangle = \phi(x)$$

for all $\phi \in X^*$. If X is not complete, then its closure $\bar{X}$ in X^{**} is.

It's important to see how easy the proof of Theorem 6.8 is. Let $x \in X$. The linearity of $\hat{x} : X^* \to \mathbb{R}$ is immediate from the definition and the fact that X^* is a linear space:

$$\hat{x}(\lambda\phi + \mu\psi) = (\lambda\phi + \mu\psi)(x) = \lambda\phi(x) + \mu\psi(x) = \lambda\hat{x}(\phi) + \mu\hat{x}(\psi),$$

for all $\lambda, \mu \in \mathbb{R}$ and all $\phi, \psi \in X^*$. Since

$$|\hat{x}(\phi)| = |\phi(x)| \le |\phi|_{X^*} \, |x|_X$$

it follows that $\hat{x} \in (X^*)^* = X^{**}$ and

$$|\hat{x}|_{X^{**}} \le |x|_X .$$

Taking $\psi \in X^*$ as in the assumption of the theorem, it follows that

$$|\hat{x}|_{X^{**}} = \sup_{0 \ne \phi \in X^*} \frac{|\hat{x}(\phi)|}{|\phi|_{X^*}} \ge |\hat{x}(\psi)| = \psi(x) = |x|_X .$$

Thus $|\hat{x}|_{X^{**}} = |x|_X$ so

$$X \ni x \xrightarrow{hat} \hat{x} \in X^{**}$$

is norm preserving. The linearity of this map follows from

$$\widehat{\lambda x + \mu y}(\phi) = \phi(\lambda x + \mu y) = \lambda\phi(x) + \mu\psi(x) = \lambda\hat{x}(\phi) + \mu\hat{y}(\phi) = (\lambda\hat{x} + \mu\hat{y})(\phi)$$

for all $\lambda, \mu \in \mathbb{R}$, all $x, y \in X$, and all $\phi \in X^*$, whence

$$\widehat{\lambda x + \mu y} = \lambda\hat{x} + \mu\hat{y}$$

for all $\lambda, \mu \in \mathbb{R}$ and all $x, y \in X$. Theorem 4.13 says that the closure of

$$\hat{X} = \{\hat{x} : x \in X\} \quad \text{in} \quad X^{**} \tag{6.6}$$

is closed.

[6] Every separable normed space has this property.

Definition 6.9. *A normed space X is called reflexive if $X^{**} = \hat{X}$.*

Exercise 6.10. Theorem 6.5 about

$$X \xrightarrow{\phi} \mathbb{R}$$

generalises to Theorem 6.11 below about

$$X \xrightarrow{A} Y.$$

Prove this theorem. It's formulated below.

Theorem 6.11. *Let X, Y be a normed spaces. Let $L(X,Y)$ be the space of continuous linear maps A from X to Y. Then $L(X,Y)$ is a normed vector space with norm defined by*

$$|A|_{L(X,Y)} = \sup_{0 \neq x \in X} \frac{|A(x)|_Y}{|x|_X},$$

and linear algebra defined by

$$(A+B)(x) = A(x) + B(x), \quad (\lambda A)(x) = \lambda\, A(x)$$

for $A, B \in L(X,Y), \lambda \in \mathbb{R}, x \in X$. We often write $Ax = A(x)$. If Y is a Banach space then so is $L(X,Y)$. And $L(X, L(X,Y))$.

Exercise 6.12. Take $X = \mathbb{R}^m$ and $Y = \mathbb{R}^n$ in Theorem 6.11. Prove that every linear map[7] A from $\mathbb{R}^m$ to $\mathbb{R}^n$ is continuous, and thereby automatically in $L(X,Y)$.

Exercise 6.13. Take $X = Y = \mathbb{R}$ in Theorem 6.11. Then[8]

every linear map $\phi : \mathbb{R} \to \mathbb{R}$ is of the form $\phi_A(x) = Ax$

with $A \in \mathbb{R}$, and thereby automatically in $\mathbb{R}^*$. The map $A \to \phi = \phi_A$ is a linear isometric bijection between $\mathbb{R}$ and its dual $\mathbb{R}^*$. In particular

$$|\phi_A|_{\mathbb{R}^*} = |A|$$

for all $A \in \mathbb{R}$.

[7] If you like you can think of A as a matrix.

[8] This is a special case of Theorem 6.31.

Exercise 6.14. For $f \in \mathrm{RI}([a,b])$ let

$$|f|_1 = \int_a^b |f(x)\,dx.$$

Explain why this does not define a norm on $\mathrm{RI}([a,b])$, but that on $C([a,b])$ it does. Show that $C([a,b])$ is not complete with this norm.

The estimate in Exercise 6.14 says that

$$|\phi(f)| \leq |f|_1 \tag{6.7}$$

for all $f \in C([a,b]) \subset \mathrm{RI}([a,b])$ so ϕ is also a Lipschitz continuous map from $C([a,b])$ to $\mathbb{R}$ if we use the 1-norm on $C([a,b])$. Note that it is harder for a linear functional on $C([a,b])$ to be continuous if we use this 1-norm:

Exercise 6.15. Show that the point evaluations

$$f \to f(x_0) \tag{6.8}$$

with $x_0 \in [a,b]$ fixed are linear but not continuous if we use the 1-norm on $C([a,b])$. They are continuous with respect to the maximum norm.

Whenever we speak of $C([a,b])$ we tacitly mean the Banach space $C([a,b])$ with the maximum norm[9]

$$|f|_\infty = \max_{x\in[a,b]} |f(x)|,$$

unless explicit stated otherwise. Next to $\mathbb{R}$ it is one of our favourite Banach spaces[10]. Its dual space includes both the linear functionals[11] defined by the point evaluations is (6.8), as well as the map ϕ from $X = C([a,b])$ to $\mathbb{R}$, defined by (5.20). The "primitive" map A from $X = C([a,b])$ to itself defined by (5.22) is linear and Lipschitz continuous with Lipschitz constant

$$|A|_{L(X,X)} = (b-a).$$

[9] Which is the limit as $p \to \infty$ of $|f|_p$.

[10] Banach algebras in fact.

[11] A functional is a function whose domain is a function space.

6.3 The plane as product space: equivalent norms

The Banach space $C([a,b])$ contains the closed linear subspace $\mathcal{A}f([a,b])$ of functions $s \to As + B$. These functions[12] are often called linear, a bit unfortunate in view of the terminology in e.g. Exercise 6.13. The other commonly used adjective for these functions is affine.

Clearly the map

$$C([a,b]) \ni f \xrightarrow{A} (f(a), f(b)) = (x_1, x_2) \in \mathbb{R}^2.$$

is linear.

Exercise 6.16. Prove directly that

$$x = (x_1, x_2) \to |x|_1 = |x_1| + |x_2| \quad \text{and} \quad x = (x_1, x_2) \to |x|_\infty = \max(|x_1|, |x_2|)$$

define norms on $\mathbb{R}^2$.

Exercise 6.17. Prove that

$$\mathcal{A}f([a,b]) \ni f \xrightarrow{A} (f(a), f(b)) \in \mathbb{R}^2$$

is a (linear) bijection, and that $|Af|_\infty = |f|_\infty$ for all $f \in \mathcal{A}f([a,b])$.

The standard norm on $\mathbb{R}^2$ is given by

$$|x|_2 = \sqrt{x_1^2 + x_2^2},$$

the Euclidean length of the 2-vector

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

It belongs to a family of norms

$$x = (x_1, x_2) \to |x|_p = (|x_1|^p + |x_2|^p)^{\frac{1}{p}},$$

which are indeed norms because of the following theorem which we state without proof.

[12] Note that $\mathcal{A}f([a,b])$ is not an algebra.

Theorem 6.18. $|x+y|_p \leq |x|_p + |y|_p$.

Exercise 6.19. Prove that $|x|_p \to |x|_\infty$ as $p \to \infty$.

Exercise 6.20. Prove that

$$|x|_\infty \leq |x|_2 \leq |x|_1;$$

$$|x|_1 \leq \sqrt{2}\,|x|_2; \quad |x|_2 \leq \sqrt{2}\,|x|_\infty.$$

Exercise 6.21. Refering to Definition 4.19: if two norms

$$x \to |x|_1 \quad \text{en} \quad x \to |x|_2$$

define the same collection $\mathcal{O}$ of open sets in a vector space X then there are constants C_1 en C_2 such that for all $x \in X$ it holds that

$$|x|_1 \leq C_2|x|_2 \quad \text{en} \quad |x|_2 \leq C_1|x|_1. \tag{6.9}$$

Prove this statement. The other way around is also true but easier to prove. Do so.

Definition 6.22. *Let X be a vector space over $\mathbb{R}$ and suppose that (6.9) holds for the norms $x \to |x|_1$ and $x \to |x|_2$. Then these norms are called equivalent.*

Exercise 6.23. Let X and Y be normed spaces, each with two equivalent norms. For a map $\Phi : X \to Y$ to be Lipschitz continuous it does not matter which of these norms is used in the definition of Lipschitz continuity. In case $\Phi = A$ is linear as in Theorem 6.11, there are four ways to define the norm of A, and all these four norms are equivalent.

Theorem 6.24. *Let X and Y be normed space, each with two equivalent norms. For a map $\Phi : X \to Y$ to be continuous in a point $x_0 \in X$ it does not matter which of these norms is used in the definition of continuity.*

Theorem 6.25. *Let X and Y be normed spaces and $p \geq 1$. Then*

$$(x, y) \to (|x|_X^p + |y|_Y^p)^{\frac{1}{p}}$$

defines a family of equivalent norms on $X \times Y$ with

$$(|x|_X^p + |y|_Y^p)^{\frac{1}{p}} \to \max(|x|_X, |y|_Y)$$

as $p \to \infty$. Each of these norms defines the same class of open sets in $X \times Y$. Example: $X = Y = \mathbb{R}, X \times Y = \mathbb{R}^2$.

Exercise 6.26. Let X and Y be normed spaces. Prove that every bounded linear function on $X \times Y$ is of the form

$$(x, y) \to \phi(x) + \psi(y)$$

with $\phi \in X^*, \psi \in Y^*$ uniquely determined. Determine the norm of this functional if the norm on $X \times Y$ is chosen to be $(x, y) \to |x| + |y|$. Same question if the norm is chosen to be $(x, y) \to \max(|x|, |y|)$. And two other questions you can think of.

6.4 The plane as a Hilbert space: Riesz representation

Assume the norm on a Banach space X comes from an inner product. We then write H for X,

$$(u, v) \in H \times H \to (u, v)_H = u \cdot v$$

for the inner product, and say that H is a Hilbert space. The norm satisfies the Cauchy-Schwarz inequality

$$|u \cdot v| \leq |u|\,|v|$$

for all $u, v \in H$. A trivial example is $H = \mathbb{R}$. An easy illuminating example is $H = \mathbb{R}^2$, which we often think of as a space of vectors $\vec{a} = \binom{a_1}{a_2}$ rather then a space of points $a = (a_1, a_2)$.

Exercise 6.27. Back to school and vectors in the plane. Draw two vectors $\vec{a}$ en $\vec{b}$, shift $\vec{b}$ to put them in head-tail configuration, draw $\vec{a} + \vec{b}$ and the usual triangle. Also draw the triangle obtained by shifting $\vec{a}$ instead. Work out $(\vec{a} + \vec{b}) \cdot (\vec{a} + \vec{b})$ and $(\vec{a} - \vec{b}) \cdot (\vec{a} - \vec{b})$, take the sum, derive what is called the parallellogram law and relate this law to your sketch.

Exercise 6.28. Let H be a Hilbert space, $K \subset H$ a closed convex subset, and $a \in H$. Prove the existence of a unique $p \in K$ which minimizes the distance $d(a,x)$ as x varies over K. Hint: use the parallellogram law to show that a minimizing sequence is Cauchy. Also show that $(p-a)\cdot(x-p) \geq 0$ for all $x \in K$.

Exercise 6.29. (continued) Show that the map $P_K : H \to K$ defined by $P_K(a) = p$ has the property that $|P_K(a) - P_K(b)| \leq |a-b|$ for all $a, b \in H$. Hint: play with the inequalities for a and $p = P_k(a)$, b and $q = P_K(b)$, and $x \in K$.

Exercise 6.30. Let H be a Hilbert space, $L \subset H$ a closed linear subspace. Prove that $P_L : H \to L$ is linear and that

$$M = N(P_L) = \{x \in H : P_L(x) = 0\} = L^\perp = \{x \in H : x \cdot y = 0 \ \forall\, y \in L\}$$

is also a closed subspace with $M \cap L = \{0\}$, and that every $x \in H$ is uniquely decomposed as $x = p + q$ with $p \in L$ and $q \in M$.

Theorem 6.31. *For $u \in H$ define $\phi_u : H \to \mathbb{R}$ by*

$$v \to u \cdot v = \phi_u(v) = \langle \phi_u, v \rangle.$$

Then[13]

$$H \ni u \xrightarrow{\Phi} \phi_u \in H^*$$

is a linear isometric bijection[14] *between H and H^*. This bijection allows*[15] *to identify H and H^*, and implies that H^* is also a Hilbert space, with inner product defined by*

$$\phi_u \cdot \phi_v = \frac{1}{2}\left(|\phi_u + \phi_v|^2 - |\phi_u|^2 - |\phi_v|^2\right) = \frac{1}{2}\left(|u+v|^2 - |u|^2 - |v|^2\right) = u \cdot v.$$

We shall denote the inverse of Φ by R_H and call it the Riesz representation map. If $\phi \in H^*$ then $u = R_H(\phi)$ is called its Riesz representation in H.

[13] Exercise 6.13 discussed the case that $H = \mathbb{R}$.

[14] The Riesz representation theorem is a special case of the Lax-Milgram Theorem 11.14.

[15] But it may be convenient not do so.

We sketch the proof. The Cauchy-Schwarz inequality gives

$$|\phi_u(v)| = |u \cdot v| \leq |u||v|,$$

so $\phi_u \in H^*$, and $v = u$ gives $|\phi_H| = |u|$. This defines the linear isometric map Φ, which we claim is surjective. To see why let $\phi \in H^*$ and

$$N_\phi = \{v \in H : \phi(x) = 0\}.$$

Exercise 6.32. Prove that $N_f \subset H$ is a closed linear subspace and that P_N is a scalar multiple of ϕ_e if $e \in N_f^\perp$ with $|e| = 1$.

Exercise 6.33. Explain why every every Hilbert space has the Hahn-Banach property mentioned in Theorem 6.8.

6.5 Integrals of continuous X-valued functions

In view of (5.12,5.17) we may just as well use the sums

$$S = \sum_{k=1}^{N} f(\xi_k)(x_k - x_{k-1}) \quad \text{with} \quad \xi_k \in [x_{k-1}, x_k] \tag{6.10}$$

for the definition of the integral of a uniformly continuous function

$$f : [a, b] \to \mathbb{R}.$$

The point to make here is that the ordering of $\mathbb{R}$ was not really needed to define

$$\int_a^b f(x)\, dx$$

for continuous $\mathbb{R}$-valued functions of $x \in [a, b]$ and that an alternative approach generalises for free to X-valued functions of $x \in [a, b]$, if X is a complete[16] normed space.

Suppose we take two such sums, with the same partition, but with different intermediate points, say $\xi_k, \tilde{\xi}_k \in [x_{k-1}, x_k]$, and consider S and

$$\tilde{S} = \sum_{k=1}^{N} f(\tilde{\xi}_k)(x_k - x_{k-1}).$$

[16] We will have to take limits of Cauchy sequences.

Then clearly

$$S - \tilde{S} = \sum_{k=1}^{N} (f(\xi_k) - f(\tilde{\xi}_k))(x_k - x_{k-1}).$$

Now assume that $x_k - x_{k-1} < \delta$ for all $k = 1, \ldots, N$, with $\delta > 0$ chosen for $\varepsilon > 0$ as in the ε-statement for uniform continuity of $f : [a, b] \to \mathbb{R}$. Then also $\xi_k - \tilde{\xi}_{k-1} < \delta$ whence

$$|S - \tilde{S}| \leq \sum_{k=1}^{N} |f(\xi_k) - f(\tilde{\xi}_k)| \, (x_k - x_{k-1}) < \sum_{k=1}^{N} \varepsilon \, (x_k - x_{k-1}) = \varepsilon(b - a).$$

Exercise 6.34. Take two such sums, one with partition points x_k and intermediate points $\xi_k \in [x_{k-1}, x_k]$, the other as in Exercise 5.10 with partition points y_l, and intermediate points $\eta_l \in [y_{l-1}, y_l]$, and consider

$$S = \sum_{k=1}^{N} f(\xi_k)(x_k - x_{k-1}) \quad \text{and} \quad T = \sum_{l=1}^{M} f(\eta_l)(y_l - y_{l-1}).$$

Use the common refinement as in Section 5.5 to show that

$$|S - T| < \varepsilon(b - a)$$

if $x_k - x_{k-1} < \delta$ for all $k = 1, \ldots, N$ and $y_l - y_{l-1} < \delta$ for all $l = 1, \ldots, M$, with $\delta > 0$ as in the ε-statement for uniform continuity.

Exercise 6.35. Take equidistant partitions and the left endpoint sums L_{2^n} as in Exercise 5.7. Use Exercise 6.34 to show that these form a Cauchy sequence and thereby a convergent sequence with a limit $I \in \mathbb{R}$. Then prove the following theorem for the case that $X = \mathbb{R}$. The proof for general Banach spaces X is identical.

Theorem 6.36. *Let X be a real Banach space and let $f : [a, b] \to X$ be continuous*[17]*. Then there is a unique $I \in X$ such that every*

$$S = \sum_{k=1}^{N} f(\xi_k)(x_k - x_{k-1}) \quad \text{with} \quad \xi_k \in [x_{k-1}, x_k]$$

as in (6.10) has $|S - I| < \varepsilon(b - a)$, provided $x_k - x_{k-1} < \delta$ for all $k = 1, \ldots, N$, with $\delta > 0$ chosen for $\varepsilon > 0$ as in the definition of uniform continuity of

[17] By Theorem 4.52 it is also uniformly continous.

$f : [a,b] \to X$. *This I is called the integral of f from a to b, notation as before*

$$I = \int_a^b f(x)\,dx.$$

Exercise 6.37. Note that $x \to |f(x)|$ is continuous from $[a,b]$ to $\mathbb{R}$ with the same uniform modulus of continuity as f, meaning that the ε-dependent δ-choice in the uniform continuity statement of f also does the job for $x \to |f(x)|$. Explain why it follows directly from the simultaneous construction of the integrals via (6.10) that

$$|\int_a^b f(x)\,dx| \leq \int_a^b |f(x)|\,dx. \tag{6.11}$$

6.6 Integral equations

Let X be a Banach space and $T > 0$. Consider the space

$$U = C([0,T], X)$$

of all continuous functions from the interval $[0,T]$ to X. We shall denote its elements by u, and consider every u as an X-valued function of t. If you think of t as corresponding to a horizontal t-axis, and $x \in X$ as corresponding to a vertical x-axis if $X = \mathbb{R}$, then $x = u(t)$ defines the graph of u.

Now let[18] $F : X \to X$ be Lipschitz continuous with Lipschitz constant L, meaning that

$$|F(x_1) - F(x_2)| \leq L\,|x_1 - x_2|$$

for all $x_1, x_2 \in X$, and consider the X-valued function v defined by

$$v(t) = \int_0^t F(u(s))\,ds \quad (t \in [0,T]). \tag{6.12}$$

Exercise 6.38. For $u \in U$ let v be defined by (6.12). Show that $v \in U$. Thus (6.12) defines a map from U to itself.

Now recall that the norm in U is given by

$$|u|_U = \max_{0 \leq t \leq T} |u(t)|_X,$$

the maximum value of the norm of $u(t)$ in X. We shall drop the subscripts.

[18] A word on notation: before we had $X \xrightarrow{\Phi} X$, here we prefer $X \xrightarrow{F} X$ and $U \xrightarrow{\Phi} U$.

Exercise 6.39. Let $\Phi : U \to U$ be the map defined in Exercise 6.12, i.e. $\Phi(u) = v$. Prove that Φ is Lipschitz continuous with Lipschitz constant LT.

Exercise 6.40. Let $\xi \in X$, X a Banach space, $F : X \to X$ Lipschitz continuous with Lipschitz constant L, and T a real number with

$$0 < T < \frac{1}{L}.$$

Prove there exists a unique continuous function $u : [0, T] \to X$ such that

$$u(t) = \xi + \int_0^t F(u(s))\, ds$$

for all $t \in [0, T]$. Hint: use Theorem 3.29.

Exercise 6.41. There is no reason to restrict to $t \geq 0$ in (6.12), nor to start in $t = 0$ with the integral. Prove the following theorem step by step by step, starting with $C([-T, T], X)$.

Theorem 6.42. *Let X be a Banach space, $F : X \to X$ a Lipschitz continuous function, $\xi \in X$. Then there exists a unique continuous function $u : \mathbb{R} \to X$ such that*

$$u(t) = \xi + \int_0^t F(u(s))\, ds \tag{6.13}$$

for all $t \in \mathbb{R}$.

Remark 6.43. *Note that $u(0) = \xi$ is u is a solution of (6.13). In the context of (differential or integral) equations for functions $u(t)$ of t one often writes $u(0) = u_0$.*

Note that the values $u(t)$ for $t \in [-T, T]$ with T smaller than the reciprocal of the Lipschitz constant of F follow from an application of the Banach contraction theorem in $C([-T, T]; X)$, or by studying the scheme

$$u_n(t) = \xi + \int_0^t F(u_{n-1}(s))\, ds \quad (n \in \mathbb{N}) \quad \text{starting from} \quad u_0(t) \equiv \xi, \tag{6.14}$$

i.e. starting from the function u_0 which is defined by $u_0(t) = \xi$ for all t.

Exercise 6.44. Let $X = \mathbb{R}$, $\xi = 1$ and $F(x) = x$. Evaluate u_1, u_2, $u_3, \ldots$ using (6.14), and characterise the limit $u(t)$ as a power series in t, i.e. as

$$u(t) = \sum_{k=0}^{\infty} \alpha_k t^k.$$

Determine the coefficients α_k.

Exercise 6.45. Write the function in Theorem 6.42 as $u(t;\xi)$ indicating both the dependence on t and the dependence on ξ and introduce the t-dependent maps

$$S(t) : X \to X$$

by $S(t)\xi = u(t;\xi)$. Show that $S(t+s) = S(t) \circ S(s)$ for all $s, t \in \mathbb{R}$.

If we write $u(t;\xi) = u_\xi(t)$ we can think of

$$\xi \to u_\xi$$

as a map from a ξ-interval to $C([-T,T])$ which describes the solution set of

$$\Psi(\xi, u) = u - \Xi_\xi - \Phi(u) = 0,$$

in which $\Xi_\xi(t) = \xi$ for all $t \in [-T,T]$ and Φ is the map introduced in Exercise 6.39. We say that $\xi \to u_\xi$ is an implicit function for the equation $\Psi(\xi, u) = 0$. Implicit functions are discussed in Chapter 10 and we will come back to this example in Exercise 10.4.

7 Power series

The message for this chapter is that calculus for polynomials is based on calculus for monomials, that is calculus for

$$f_n(x) = x^n \quad (n \in \mathbb{N}), \tag{7.1}$$

and that the step from polynomials to power series like

$$p(x) = 1 + 2x + 3x^2 + \cdots \tag{7.2}$$

is just a matter of observing that every such power series comes with a critical radius R: for $x \in \mathbb{R}$ with $|x| < R$ the calculus is the same as that for polynomials, if $|x| > R$ the expressions have no meaning.

7.1 Polynomials and power series

The second degree polynomial

$$p_2(x) = 1 + 2x + 3x^2$$

is defined for all $x \in \mathbb{R}$. You will no doubt have seen that 1 is the derivative of x, $2x$ the derivative of x^2, $3x^2$ the derivative of x^3, and so on. In particular p_2 is the derivative of P_3, where P_3 is defined by the third degree polynomial

$$P_3(x) = x + x^2 + x^3,$$

a third degree polynomial. We have

$$x + x^2 + x^3 = P_3(x) \quad \text{and} \quad P_3'(x) = p_2(x) = 1 + 2x + 3x^2$$

for all x in $\mathbb{R}$, which for a start is the natural domain of both P_3 and p_2.

Whatever meaning[1] you may assign to these statements, the upshot of the step from polynomials to power series will be that $p(x)$ in (7.2) is the derivative of

$$P(x) = x + x^2 + x^3 + \cdots, \tag{7.3}$$

with usually a quickly made observation about the validity of both expressions depending on the size $|x|$. In (7.2) and (7.3) the individual terms become larger and larger if $|x| > 1$, while they rapidly become small if $|x| < 1$.

Thus in this example R, the radius of convergence, happens to be 1 and

$$P'(x) = p(x) \quad \text{holds if } |x| < 1, \tag{7.4}$$

[1] Chapter 8 will be full of meaning.

while it is meaningless if $|x| > 1$. If you know your geometric series, see Section 1.5, then you easily verify that

$$P(x) = \frac{x}{1-x},$$

and if you know from highschool that

$$P'(x) = \frac{1}{(1-x)^2},$$

it follows via (7.4) that

$$p(x) = 1 + 2x + 3x^2 + \cdots = \frac{1}{(1-x)^2} \quad \text{holds if } |x| < 1,$$

a truly remarkable formula.

Another example is

$$P(x) = 1 + x + 2x^2 + 6x^3 + 24x^4 + 120x^5 + \cdots,$$

which has no meaning al all[2], unless $x = 0$, meaning that $R = 0$. On the other hand the power series

$$P(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!} = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

has a gloryful meaning[3] for all $x \in \mathbb{R}$. For this power series we say that $R = \infty$.

7.2 Unconditional convergence of good series

Look again at (3.1),(3.2) and (3.5), and suppose we start with a sequence ξ_n of real numbers indexed by $n \in \mathbb{N}$. We introduce the (partial) sums[4]

$$S_n = \sum_{k=1}^{n} \xi_k \quad \text{and} \quad M_n = \sum_{k=1}^{n} |\xi_k|.$$

Then

$$|S_n| \leq \sum_{k=1}^{n} |\xi_k| = M_n,$$

[2] Its finite sums do serve as approximations of mysteriously beautiful functions.

[3] It happens to have itself as derivative.

[4] With dummy index k.

and M_n is a nondecreasing sequence. If this sequence is bounded, then

$$\bar{M} = \lim_{n\to\infty} M_n = \sup_{n\in\mathbb{N}} M_n \in \mathbb{R}$$

exists. In that case we define the sum of the absolute values to be

$$\sum_{k=1}^{\infty} |\xi_k| = \lim_{n\to\infty} \sum_{k=1}^{n} |\xi_k| = \bar{M}. \tag{7.5}$$

Equation 7.5 thus gives a meaning to the sum of an infinite number of non-negative terms enumerated by $k \in \mathbb{N}$.

For $m, n \in \mathbb{N}$ with $m < n$ we then have that

$$|S_n - S_m| = |\sum_{k=m+1}^{n} \xi_k| \leq \sum_{k=m+1}^{n} |\xi_k| = M_n - M_m,$$

so S_n is a Cauchy sequence and thereby convergent because the sequence M_n is convergent and thereby a Cauchy sequence. We conclude that

$$S = \lim_{n\to\infty} S_n = \lim_{n\to\infty} \sum_{k=1}^{n} \xi_k \tag{7.6}$$

exists, a limit which is called the sum of the series[5]

$$\sum_{n=1}^{\infty} \xi_n.$$

Exercise 7.1. Prove that $|S| \leq \bar{M}$.

Informally we write (7.5) as

$$\sum_{n=1}^{\infty} |\xi_n| < \infty$$

and say that the series

$$\sum_{n=1}^{\infty} \xi_n \tag{7.7}$$

is absolutely convergent.

[5] And denoted by the same expression.

It may of course happen that the sequence M_n is not bounded. Then (7.5) has no meaning but (7.6) may still hold, in which case we still define the sum to be

$$\sum_{n=1}^{\infty} \xi_n = S$$

and say that the series (7.7) is convergent with sum S, but not absolutely convergent. Summing up we have the following theorem.

Theorem 7.2. *Let ξ_n be a sequence of real numbers indexed by $n \in \mathbb{N}$. If*

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

is absolutely convergent then it is also convergent and

$$|\sum_{k=1}^{\infty} \xi_n| \leq \sum_{k=1}^{\infty} |\xi_n|.$$

Series for which the sequence M_n is unbounded are no good.

Exercise 7.3. Think about

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

We want to compute and manipulate with sums like

$$S = \sum_{k=0}^{\infty} \xi_k = \xi_0 + \xi_1 + \xi_2 + \cdots \tag{7.8}$$

we do with finite sums. For instance,

$$\xi_0+\xi_1+\xi_2 = \xi_0+\xi_2+\xi_1 = \xi_1+\xi_0+\xi_2 = \xi_1+\xi_2+\xi_0 = \xi_2+\xi_0+\xi_1 = \xi_2+\xi_1+\xi_0$$

is 6 ways to write the same sum

$$\sum_{k=0}^{3} \xi_k,$$

which we define via

$$\xi_0 + \xi_1 + \xi_2 = (\xi_0 + \xi_1) + \xi_2.$$

Every permutation $\phi : \{0,1,2\} \to \{0,1,2\}$ gives the same sum

$$(\xi_{\phi(0)} + \xi_{\phi(1)}) + \xi_{\phi(2)} = (\xi_0 + \xi_1) + \xi_2,$$

which is why we write the sum of 3 terms as any of the 6 sums we like best.

For sums as in (7.8) we would similarly like to have that

$$S = \sum_{k=0}^{\infty} \xi_{\phi(n)} = \sum_{k=0}^{\infty} \xi_k \tag{7.9}$$

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

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

that

$$S_n^\phi \to S \quad \text{and} \quad \bar{M}_n^\phi \to \bar{M} \tag{7.10}$$

as $n \to \infty$.

The basic assumption to make is again (7.5). What we know then is that

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

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

$$\bar{M} - \varepsilon < \sum_{k=0}^{n_\varepsilon} |\xi_k| \le \bar{M}, \tag{7.11}$$

for otherwise $\bar{M}$ is not the lowest upperbound, and then also

$$\bar{M} - \varepsilon < \sum_{k=0}^{n} |\xi_k| \le \bar{M},$$

for all $n \ge n_\varepsilon$, for otherwise $\bar{M}$ is not an *upper bound*. This is just the proof that

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

redone. Subtracting the partial sum in (7.11) from (7.11) we obtain in particular that

$$\sum_{k=n_\varepsilon+1}^{\infty} |\xi_k| - \varepsilon < 0 \le \sum_{k=n_\varepsilon+1}^{\infty} |\xi_k|,$$

whence

$$\sum_{k=n_\varepsilon+1}^{\infty} |\xi_k| < \varepsilon. \tag{7.12}$$

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

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

via ϕ with $m \in \mathbb{N}_0$, then

$$\{0, 1, \ldots, n_\varepsilon\} \subset \{\phi(0), \phi(1), \ldots, \phi(m_\varepsilon)\}$$

for some $m_\varepsilon \in \mathbb{N}_{n_\varepsilon}$ Therefore

$$\bar{M} - \varepsilon < M_{n_\varepsilon} \leq \bar{M}_{m_\varepsilon}^\phi \leq \bar{M}_m^\phi \leq \bar{M}.$$

if $m \geq m_\varepsilon$. This proves that $\bar{M}_m^\phi \leq \bar{M}$ as $m \to \infty$, and we also have, for the other partial sums, that

$$|S_m^\phi - S_{n_\varepsilon}| \leq \sum_{k=n_\varepsilon+1}^{\infty} |\xi_k| < \varepsilon,$$

because $S_m^\phi - S_{n_\varepsilon}$ is a finite sum of terms ξ_k with $k > n_\varepsilon$ if $m \geq m_\varepsilon$.

Exercise 7.4. Show that S_n^ϕ converges to the same sum $S \in \mathbb{R}$.

Theorem 7.5. *Let $\xi_k \in \mathbb{R}$ be indexed $k \in \mathbb{N}_0$. If the partial sums*

$$M_n = \sum_{k=0}^{n} |\xi_k|$$

are bounded, then

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

are convergent sequences for every bijection $\phi : \mathbb{N}_0 \to \mathbb{N}_0$, with limits S and $\bar{M}$ independent of the bijection ϕ, and $|S| \leq \bar{M}$. We write

$$S = \sum_{k=0}^{\infty} \xi_{\phi(n)} \quad \textit{and} \quad |\sum_{k=0}^{\infty} \xi_{\phi(n)}| \leq \sum_{k=0}^{\infty} |\xi_{\phi(n)}| = \bar{M}.$$

The same statement holds for every bijection $\phi : \mathbb{N}_0 \to A$ if A is some other enumerable set, for numbers $\xi_\alpha \in \mathbb{R}$ with $\alpha \in A$. We can thus write

$$S = \sum_{\alpha \in A} \xi_\alpha, \quad \bar{M} = \sum_{\alpha \in A} |\xi_\alpha|, \quad \textit{with again} \quad |S| \leq \bar{M}, \tag{7.13}$$

if for some bijection $\phi : \mathbb{N} \to A$ the partial sums

$$\sum_{k=1}^{n} |\xi_{\phi(k)}|$$

are a bounded sequence indexed by $n \in \mathbb{N}$.

7.3 Integral calculus for power series

Consider

$$p(x) = \sum_{n=1}^{\infty} \alpha_n x^n. \tag{7.14}$$

In Exercise 5.13 we saw that

$$\int_a^b x^n \, dx = [\frac{x^n}{n+1}]_a^b = \frac{b^{n+1}}{n+1} - \frac{a^{n+1}}{n+1} \tag{7.15}$$

for $0 \leq a < b$. Via Theorem 5.11, Exercise 5.25, and Definition 5.26 this restriction on a and b disappears:

Exercise 7.6. Verify that (7.15) holds for all $n \in \mathbb{N}$ and any $a, b \in \mathbb{R}$.

Theorem 5.31 then implies for

$$p_k(x) = \sum_{n=1}^{k} \alpha_n x^n = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_k x^k, \tag{7.16}$$

the partial polynomial sums of (7.14), that

$$\int_a^b p_k(x) \, dx = P_{k+1}(b) - P_{k+1}(a), \tag{7.17}$$

with P_{k+1} defined by

$$P_{k+1}(x) = \alpha_0 x + \frac{\alpha_1}{2} x^2 + \frac{\alpha_2}{3} x^3 + \cdots + \frac{\alpha_k}{k+1} x^{k+1}. \tag{7.18}$$

Of course 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

$$\sum_{n=1}^{\infty} |\alpha_n| r^n < \infty. \tag{7.19}$$

The estimate in Theorem 7.2 with $\xi_n = \alpha_n x^n$ implies, for all $x \in \mathbb{R}$ with $|x| \leq r$, that

$$|p_k(x) - p(x)| = |\sum_{n=k+1}^{\infty} \alpha_n x^n| \leq \sum_{n=k+1}^{\infty} |\alpha_n x^n| \leq \sum_{n=k+1}^{\infty} |\alpha_n| r^n,$$

provided $[a, b] \subset [-r, r]$. It follows that $p_k \to p$ in $C([a, b])$ and thus by Theorem 5.29 that

$$\int_a^b p_k(x)\, dx \to \int_a^b p(x)\, dx \tag{7.20}$$

as $k \to \infty$. Combining (7.18) and (7.20) we arrive at the statements in the following theorem for a, b, x contained in $[-r, r]$.

Theorem 7.7. *If α_n is a sequence of real coefficients indexed by $n \in \mathbb{N}_0$, then there exists a maximal $R \in [0, \infty]$ such*

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots \tag{7.21}$$

exists for all $x \in \mathbb{R}$ with $|x| < R$. For those values

$$P(x) = \alpha_0 x + \frac{\alpha_1}{2} x^2 + \frac{\alpha_2}{3} x^3 + \cdots = \sum_{n=0}^{\infty} \frac{\alpha_n}{n+1} x^{n+1} = \sum_{n=1}^{\infty} \frac{\alpha_{n-1}}{n} x^n \tag{7.22}$$

also exists. Moreover

$$\int_a^b p(x)\, dx = P(b) - P(a)$$

whenever $|a| < R$ and $|b| < R$.

Exercise 7.8. Finish the proof of Theorem 5.29. Hint: consider the set of values $r > 0$ for which (7.19) holds. It is either empty, the whole of $\mathbb{R}_+$, or an interval of the form $(0, R)$ or $(0, R]$ with $R \in \mathbb{R}_+$.

Exercise 7.9. A bit of a project. Avoid the epsilons and prove Theorem 5.29 without using results on the integrability of continuous functions. Hint: use Exercise 5.27 to reduce to integrals in which x does not change sign, Theorems 5.30, 5.31 to split and further reduce to integrals of monotone functions (Section 5.2) and Theorem 5.28 to conclude. How does R appear?

Exercise 7.10. 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 $|x| < R$.

7.4 Differential calculus for powerseries

You will be familiar with

$$f'(a) = \lim_{h\to 0} \frac{f(a+h) - f(a)}{h} = \lim_{x\to a} \frac{f(x) - f(a)}{x - a} \tag{7.23}$$

as the usual definition of the derivative $f'(a)$ of a given real valued function f of a real variable x in a given point $x = a$ on the real line, and the notation

$$f'(x) = \frac{df}{dx} = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}. \tag{7.24}$$

If the limit of the difference quotient in (7.24) exists it is called the *differential quotient* of f in x, formally denoted as a fraction[6] with denominator df and numerator dx. The difference quotient itself is often denoted as

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

with $\Delta x = h \neq 0$.

The domain D_f of f is the subset of all $x \in \mathbb{R}$ for which $f(x)$ is defined, one way or another. The difference quotient in (7.23) is only defined for $a, x \in D_f$ with $x \neq a$. For the simplest examples to consider first, e.g. monomials such as

$$f_{42}(x) = x^{42},$$

the domain is the whole of $\mathbb{R}$ and the difference quotient is naturally defined for $x = a$ as well. For instance,

$$\frac{x^{42} - a^{42}}{x - a} = x^{41} \underbrace{+ \cdots +}_{\text{Exercise 5.2!}} a^{41}$$

[6] This also suggests to write $df = f'(x)dx$.

is equal to $42a^{41}$ for $x = a$. Who would need a limit concept[7] here? Whatever the value of a, clearly

$$f'_{42}(a) = 42a^{41}$$

so

$$f_{42}(x) = x^{42} \quad \text{has} \quad f'_{42}(x) = 42x^{41}.$$

But that's algebra. Algebra is easy. Analysis is hard. In analysis we need the concept of limits.

For (7.23) we need Definition 4.8 with $X = Y = \mathbb{R}$ and the difference quotient

$$\frac{f(x) - f(a)}{x - a}$$

rather than $f(x)$. Unfortunately this quotient is not defined in the case of general X and $f : X \to \mathbb{R}$, which is a motivation to want to avoid (7.23), as we prefer to do from here on, also when $X = \mathbb{R}$.

7.5 Powerseries: the fundamental theorem

In what follows we will avoid limits of difference quotients and think of differentiation as a method to best approximate a given (nonlinear) $f(x)$ by

$$f(a) + A(x - a) = Ax + B,$$

best in terms of the properties of the remainder term

$$R_a(x) = f(x) - Ax - B$$

near $x = a$. We often write

$$f(x) = f(a) + A(x - a) + R_a(x)$$

and identify $f'(a)$ as the unique value of A for which the remainder term $R_a(x)$ has a smallness property that fails for other choices of A.

For simple choices of $f(x)$, e.g. $x^2, x^3, \ldots$, elementary algebra show us what to do, with a nice explicit expression for A that you have seen before via (7.23), and a resulting elegant formula for the remainder term. A first theorem that we shall then prove using this approach is a differentiation theorem that complements Theorem 7.21 for power series

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots$$

with real coefficients $\alpha_0, \alpha_1, \alpha_2, \ldots$.

[7] Jaap Murre: "U dacht natuurlijk dat u alleen kunt differentiëren in de analyse".

Theorem 7.11. *Every power series expression of the form*

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n$$

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 $p'(x)$ is also absolutely convergent for all $x \in \mathbb{R}$ with $|x| < R$, and the convergence of both $p(x)$ and $p'(x)$ is uniform on every $\{x \in \mathbb{R} : |x| < r\}$ 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 sums exists.

7.6 Linear approximations of monomials

To illustrate the approach with linear approximations consider a difference quotient as above for the function f_7 defined by $f_7(x) = x^7$. A little algebra[8] tells 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 may rewrite as

$$x^7 = a^7 + (x^6 + ax^5 + a^2x^4 + a^3x^3 + a^4x^2 + a^5x + a^6)(x - a) = \qquad (7.25)$$

$$\underbrace{a^7 + 7a^6(x-a)}_{\text{linear approximation } Ax+B} + \underbrace{(x^5 + 2ax^4 + 3a^2x^3 + 4a^3x^2 + 5a^4x + 6a^5)(x-a)^2}_{\text{remainder term}}.$$

The first 2 terms can be seen as the best approximation of the form

$$Ax + B = a^7 + 7a^6(x - a)$$

[8] Long division for instance.

to $f_7(x) = x^7$ when x is close to a, in a meaning to be properly chosen and explained. The particular choice of A and B followed from putting $x = a$ in the 7 terms of the typographically large prefactor in (7.25), but not in $(x-a)$ itself. If you already knew that $f_7'(x) = 7x^6$ you should recognise A as $f_7'(a)$ computed via (7.23). These values of A and B appear as the only choice[9] for these two coefficients which makes the resulting remainder term[10] contain a factor $(x-a)^2$.

The prefactor in the remainder term contains higher powers that can be estimated if we assume that x and ξ are contained in a fixed interval $[-r, r]$. In other words, if

$$|x| \leq r \quad \text{and} \quad |\xi| \leq r,$$

the prefactor is estimated by

$$(1 + 2 + 3 + 4 + 5 + 6)\, r^5.$$

You will not be surprised that (7.25) and its splitting in a linear term and such an remainder term generalise to general $n \in \mathbb{N}$ as

$$x^n = a^n + na^{n-1}(x-a) + R_{a,n}(x), \tag{7.26}$$

in which you recognise

$$f_n'(a) = na^{n-1}$$

when computed from (7.23) with

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

Exercise 7.12. A nice expression for $R_{a,n}(x)$ you guess from (7.25). Prove what you rightly guessed for all $n \in \mathbb{N}$.

Exercise 7.13. For $n \in \mathbb{N}$ and $x, a \in \mathbb{R}$ let $R_{a,n}(x)$ defined by (7.26). Let $r > 0$. Show that

$$|R_{a,n}(x)| \leq \frac{n(n-1)}{2} r^{n-2} (x-a)^2$$

if $x, a \in [-r, r]$. Note that for $n = 0$ and $n = 1$ there is no remainder term.

[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.

7.7 From polynomials to a proof for power series

Let $\alpha_0, \alpha_1, \alpha_2, \ldots$ be a sequence of real coefficients. Then for the polynomials

$$p_k(x) = \sum_{n=0}^{k} \alpha_n x^n = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_k x^k$$

of degree $k \geq 2$ the story is quite the same as in Section 7.6. Simply multiply both sides of (7.26) by α_n and take the sum over n. With some care for $n = 0, 1, 2$ it follows that

$$p_k(x) = p_k(a) + \underbrace{\sum_{n=1}^{k} n\alpha_n a^{n-1}(x-a)}_{A} + \underbrace{\sum_{n=2}^{k} \alpha_n R_{a,n}(x)}_{remainder\ term} \tag{7.27}$$

in which for all $x, a \in [-r, r]$ the remainder term satisfies

$$|\sum_{n=2}^{k} \alpha_n R_{a,n}(x)| \leq \underbrace{\sum_{n=2}^{k} |\alpha_n| \frac{n(n-1)}{2} r^{n-2}}_{C_{r,k}} (x-a)^2. \tag{7.28}$$

As before only the choice

$$A = \sum_{n=1}^{k} n\alpha_n a^{n-1} \tag{7.29}$$

leads to such an estimate for the remainder term, which is why, among all linear approximations

$$p_k(a) + A(x-a),$$

the best linear approximation of $p_k(x)$ near $x = a$ is

$$\underbrace{p_k(a) + \sum_{n=1}^{k} n\alpha_n a^{n-1}}_{p_k'(a)} (x-a),$$

which thereby recovers $p_k'(a)$ computed via the limit (7.23) of the differential quotient for p_k.

But now, if for some $r > 0$ it holds that

$$C_r = \sum_{n=2}^{\infty} |\alpha_n| \frac{n(n-1)}{2} r^{n-2} < \infty, \tag{7.30}$$

we can let $k \to \infty$ in (7.27). Indeed, it then follows from Theorem 7.2 that the sums

$$\sum_{n=0}^{\infty} \alpha_n x^n, \quad \sum_{n=1}^{\infty} \alpha_n a^n, \quad \sum_{n=1}^{\infty} n\alpha_n a^{n-1}$$

exist for all $x, a \in [-r, r]$ because

$$1 \leq n \leq \frac{n(n-1)}{2}$$

for $n \geq 2$, and so does the sum

$$R_a(x) = \sum_{n=2}^{\infty} \alpha_n R_{a,n}(x).$$

Thus (7.30) allows to take the limit $k \to \infty$ in (7.27) and (7.28) to obtain[11]

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n = p(a) + \underbrace{\sum_{n=1}^{\infty} n\alpha_n a^{n-1}}_{A}(x-a) + R_a(x) \tag{7.31}$$

with

$$|R_a(x)| \leq C_r(x-a)^2 \tag{7.32}$$

for all $x, a \in [-r, r]$.

As before we observe that

$$A = \sum_{n=1}^{\infty} n\alpha_n a^{n-1} \tag{7.33}$$

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 (7.32) with a constant depending only on r. Evaluating the difference quotient in (7.23) with f replaced by p then yields

$$\frac{p(x) - p(a)}{x-a} = A + \frac{R_a(x)}{x-a},$$

and (7.32) is sufficient to conclude that

$$\lim_{x \to a} \frac{p(x) - p(a)}{x-a} = A \tag{7.34}$$

[11] The convergence is in fact uniform on $[-r, r]$, why?

as given by (7.33).

The r-values for which (7.30) holds form an interval

$$\{r \geq 0 : \sum_{n=1}^{\infty} n^2 |\alpha_n| r^n < \infty\}$$

which contains $r = 0$. The only possibilities are

$$\{0\},\ [0, R),\ [0, R],\ \mathbb{R},$$

with $R > 0$ in the second and third case, and $R = \infty$ and $R = 0$ in the extreme fourth and first cases. This concludes the proof of Theorem 7.11, except for the statement about $|x| > R$ when $R < \infty$.

Exercise 7.14. 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 $|x| < |x_0|$ implying that $R \geq |x_0| > R$.

The limit statement (7.34) is equivalent to saying that

$$\lim_{x \to a} \frac{R_a(x)}{x - a} = 0, \tag{7.35}$$

meaning that for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\left|\frac{R_a(x)}{x - a}\right| < \varepsilon \quad \text{if} \quad 0 < |x - a| < \delta,$$

i.e.

$$|R_a(x)| < \varepsilon |x - a| \quad \text{if} \quad |x - a| < \delta. \tag{7.36}$$

Exercise 7.15. If we define

$$J_k = \{r \geq 0 : \sum_{n=1}^{\infty} n^k |\alpha_n| r^n \text{ exists}\}$$

for $k \in \mathbb{N}$. These intervals don't change much as we vary k. It is clear that

$$J_1 \supset J_2 \supset J_3 \supset \cdots,$$

Prove the existence of $R \in [0, \infty]$ with for every $k \in \mathbb{N}$ either $J_k = [0, R)$ or $J_k = [0, R]$ and give examples of $R = 0$, $R = 1$ and $R = \infty$.

7.8 Taylor's formula for power series

We substitute $x = x_0 + h$ in

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n. \tag{7.37}$$

Changing[12] the order of summation we find

$$\begin{aligned} p(x_0+h) &= \sum_{n=0}^{\infty} \alpha_n (x_0+h)^n = \sum_{n=0}^{\infty} \alpha_n \sum_{k=0}^{n} \binom{n}{k} x_0^{n-k} h^k \\ &= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \alpha_n \frac{n(n-1)\dots(n-k+1)}{k!} x_0^{n-k} h^k \\ &= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \alpha_n \frac{n(n-1)\dots(n-k+1)}{k!} x_0^{n-k} h^k \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} \underbrace{\sum_{n=k}^{\infty} \alpha_n n(n-1)\dots(n-k+1) x_0^{n-k}}_{p^{(k)}(x_0)} h^k \\ &= \sum_{k=0}^{\infty} \frac{p^{(k)}(x_0)}{k!} h^k, \end{aligned}$$

i.e.

$$p(x_0+h) = \sum_{n=0}^{\infty} \frac{p^{(n)}(x_0)}{n!} h^n. \tag{7.38}$$

Do note the special case $x_0 = 0$ and $h = x$:

$$p(x) = \sum_{n=0}^{\infty} \alpha_n x^n = \sum_{n=0}^{\infty} \frac{p^{(n)}(0)}{n!} x^n.$$

7.9 Laurent series

Everything we did for the differentiation of power series in (7.14) also works for

$$L(x) = \sum_{n=-\infty}^{\infty} \alpha_n x^n = \cdots + \frac{\alpha_{-2}}{x^2} + \frac{\alpha_{-1}}{x} + \alpha_0 + \alpha x + \alpha_2 x^2 + \cdots,$$

[12] By (7.13) in Theorem 7.5 this is allowed if $|x_0| + |h| < R$.

with $|x|$ not too large for the positive exponents and not too small for the negative exponents. Start with e.g.

$$\frac{1}{x^7} = \frac{1}{a^7} - \frac{7}{a^8}(x-a) + R_a(x),$$

and figure out what $R_a(x)$ is.

7.10 Power series solutions of differential equations

We can easily solve simple linear differential equations for power series (7.37), for instance

$$p'(x) = p(x), \tag{7.39}$$

with boundary condition $p(0) = 1$. The more visual term by term representation with perhaps mathematically unsatisfactory dots clarifies the proceedings. We have

$$p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 x^4 + \alpha_5 x^5 + \alpha_6 x^6 + \alpha_7 x^7 + \cdots,$$

making sense for $|x| < R$ with some hopefully positive R, possibly for all $x \in \mathbb{R}$, or x in any other suitable Banach algebra for that matter.

For $x \in \mathbb{R}$ the derivative is given by the algebra that started in Section 7.6 resulting in Theorem 7.11. We don't know R yet, but provided $|x| < R$ it follows that

$$p'(x) = \alpha_1 + 2\alpha_2 x + 3\alpha_3 x^2 + 4\alpha_4 x^3 + 5\alpha_5 x^4 + 6\alpha_6 x^5 + 7\alpha_7 x^6 + 8\alpha_8 x^7 + \cdots,$$

and so

$$p'(x) - p(x) = (\alpha_1 - \alpha_0) + (2\alpha_2 - \alpha_1)x + (3\alpha_3 - \alpha_2)x^2 + (4\alpha_4 - \alpha_3)x^3 + \cdots,$$

which can only be zero for all $x \in \mathbb{R}$ if

$$0 = \alpha_1 - \alpha_0 = 2\alpha_2 - \alpha_1 = 3\alpha_3 - \alpha_2 = 4\alpha_4 - \alpha_3 = \cdots,$$

as the next exercise will confirm.

Exercise 7.16. Let $r > 0$ and assume that $p(x)$ is convergent for all $x \in \mathbb{R}$ with $|x| < r$, and that $p(x) = 0$ for all those x. Explain why $\alpha_0 = 0$. Then factor out x and repeat the argument. And so one[13].

[13] Domino principle.

Applied to $p'(x) - p(x)$ Exercise 7.16 implies, in view of $\alpha_0 = p(0) = 1$, that

$$\alpha_1 = 1,\ \alpha_2 = \frac{1}{2},\ \alpha_3 = \frac{1}{2}\frac{1}{3},\ \alpha_4 = \frac{1}{2}\frac{1}{3}\frac{1}{4},\ \cdots,$$

so

Theorem 7.17. *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 $x \in \mathbb{R}$.

Exercise 7.18. Prove this theorem.

Remark 7.19. *We still have to show that there are no other functions $f(x)$ that satisfy $f(0) = 1$ and $f'(x) = f(x)$ for all x in some open interval $(-r, r)$. In fact we still have to show the analogous statement for $f(0) = 0$ and $f'(x) = 0$. Think about it.*

Definition 7.20. *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 7.21. Show for every fixed $n \in \mathbb{N}$ that

$$\frac{x^n}{\exp(x)} \to 0 \quad \text{as} \quad x \to \infty.$$

This is the standard limit that says that $\exp(x)$ beats every power of x as $x \to \infty$.

Theorem 7.22. *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 $x \in \mathbb{R}$.

Exercise 7.23. Write $p(x)$ using the sum notation and prove Theorem 7.22. Let $\cos x = p'(x)$. What is the derivative of $\cos$?

At this point we don't know yet that $\exp(x)$, $\sin x$, $\cos x$ are what they should be. One way to verify what is and will ever be is to check all the formulas by brute force calculation. For instance:

Exercise 7.24. Show for all $x \in \mathbb{R}$ that

$$\cos^2 x + \sin^2 x = 1,$$

by substituting the power series for $\cos x$ and $\sin x$ and working out the squares.

Exercise 7.25. A bit smarter: prove the equality in Exercise 7.24 by noting that $\cos^2 x$ and $\sin^2 x$ are squares of power series and thereby powerseries themselves, and so is their sum. Use Theorem 7.22 and Exercise 7.23 to conclude.

Exercise 7.26. 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 $\alpha \in \mathbb{N}_0$.

8 Differentiability via linear approximation

In this section we formulate the linearisation approach to differentiation, first for a real valued function f defined on 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.

Definition 8.1. *Let x_0 be an interior point of D_f and $f : D_f \to \mathbb{R}$. Then for some $\delta_0 > 0$ the equality*

$$f(x_0 + h) = f(x_0) + A_0 h + R_0(h) \tag{8.1}$$

defines a remainder term $R_0(h)$ for all $h \in \mathbb{R}$ with $|h| < \delta_0$. Depending on A_0 it may happen that for every $\varepsilon > 0$ a $\delta > 0$ can be chosen such that

$$|R_0(h)| < \varepsilon |h| \quad \text{if} \quad 0 < |h| < \delta. \tag{8.2}$$

If so then the function f is called differentiable in x_0.

To be crystal clear we have used subscripts in A_0 and R_0 to indicate the x_0-dependence in (8.1). Note that the limitation $|h| < \delta_0$ applies to (8.1) as a whole, but not to the term $A_0 h$ which is defined for all $h \in \mathbb{R}$.

Dropping some subscripts, we can write the expansion (8.1) as

$$f(x_0 + h) = f(x_0) + \phi(h) + R(h), \tag{8.3}$$

in which ϕ is as in Exercise 6.13, i.e.

$$\phi(h) = Ah.$$

In fact we may just as well speak about $D_f \subset X$, X a normed space, x_0 in the interior of D_f, $f : D_f \to \mathbb{R}$ and $\phi : X \to \mathbb{R}$ linear and preferably[1] continuous. The ε-δ statement (8.2), often written as

$$R(h) = o(|h|) \quad \text{for} \quad h \to 0,$$

then becomes

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{h\in X} : \quad 0 < |h|_X < \delta \implies |R(h)| < \varepsilon |h|.$$

[1] Otherwise f will be discontinuous in x_0.

Theorem 8.2. *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 8.1 holds, and $f'(x_0) = A_0$ is called the derivative of f in x_0.*

The proof can be given in the more general context of $D_f \subset X$, and the formulation in (8.3) in which A_0h has been replaced by $\phi(h)$ with $\phi \in X^*$. So suppose there is another A_0 that does the job, say B_0 instead of A_0 in (8.1), or $\psi \in X^*$ in (8.3), with remainder term $S(h)$, also satisfying $S(h) = o(|h|)$, instead of $R(h)$. Subtraction then gives

$$(A_0 - B_0)h + R(h) - S(h) = 0 \quad \text{or} \quad \phi(h) - \psi(h) + R(h) - S(h) = 0.$$

Exercise 8.3. Show that $A_0 = B_0$ in the case that $X = \mathbb{R}$. In the general case show that $\chi = \phi - \psi \in X^*$ has $|\chi|_{X^*} = 0$ whence $\chi = 0$ so $\phi = \psi$ in X^*.

If you understand what's going on you see that everything also works for maps ϕ from $D_\phi \subset X$ to Y, X and Y normed spaces. Denoting the continuous linear maps as in Theorem 6.11 we have the same definition and the same theorem, in which we write A_0h instead of $A_0(h)$ for $A_0 \in L(X, Y)$ and $h \in X$. Note that the inequality

$$|A_0h|_Y \leq |A_0|_{L(X,Y)} \, |h|_X \tag{8.4}$$

replaces the equality

$$|A_0h| = |A_0| \, |h| \tag{8.5}$$

that we have for $X = Y = \mathbb{R}$, the only change in the proceedings in fact. Definition 8.1 and Theorem 8.2 are a special case of the following definition and theorem, in which we chose to keep the subscripts.

Definition 8.4. *Let X, Y be real normed spaces, $D_\phi \subset X$, $\Phi : D_\Phi \to Y$, x_0 an interior point of D_Φ and $A_0 \in L(X, Y)$. Then*

$$\Phi(x_0 + h) = \Phi(x_0) + A_0h + R_0(h) \tag{8.6}$$

defines a remainder term $R_0(h)$ for $h \in X$ with $|h|_X < \delta_0$ for some $\delta_0 > 0$. Depending on A_0 it may happen that for every $\varepsilon > 0$ a $\delta > 0$ can be chosen such that

$$|R_0(h)|_Y < \varepsilon |h|_X \quad \text{if} \quad 0 < |h|_X < \delta. \tag{8.7}$$

If so then the map Φ is called differentiable in x_0.

Theorem 8.5. *Let X, Y be real normed spaces, $D_\Phi \subset X$, $\Phi : D_\Phi \to Y$, x_0 an interior point of D_Φ, and suppose that f is differentiable in x_0. Then there is precisely one $A_0 \in L(X, Y)$ for which the statement in Definition 8.4 holds, and $\Phi'(x_0) = A_0$ is called the derivative of Φ in x_0.*

8.1 The simple rules of differential calculus

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, but just as in Section 8 the results generalise.

Theorem 8.6. *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

$$(fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0).$$

The proofs are easy. Writing expansions with $x - x_0$ instead of h, and the remainder term as $R_0(x)$, we expand $f(x)$ as

$$f(x) = f(x_0) + A_0(x - x_0) + R_0(x), \tag{8.8}$$

in which

$$A_0 = f'(x_0)$$

if

$$R_0(x) = o(|x - x_0|) \quad \text{as} \quad x \to x_0,$$

meaning that

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, \forall_{x\in\mathbb{R}} \quad 0 < |x - x_0| < \delta \implies |R_0(x)| < \varepsilon|x - x_0|. \tag{8.9}$$

Note that we still write R_0 for the remainder term, but now as a function of x. Using the alphabetic shift convention this becomes

$$g(x) = g(x_0)+B_0(x-x_0)+S_0(x), \quad B_0 = g'(x_0), \quad S_0(x) = o(|x-x_0|) \tag{8.10}$$

for g.

Adding (8.8) to (8.10) gives

$$(f + g)(x) = f(x) + g(x) =$$

$$f(x_0) + g(x_0) + A_0(x - x_0) + B_0(x - x_0) + R_0(x) + S_0(x) =$$

$$(f + g)(x_0) + \underbrace{(A_0 + B_0)}_{(f+g)'(x_0)}(x - x_0) + \underbrace{R_0(x) + S_0(x)}_{remainder\ term}$$

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

$$(f+g)'(x_0) = A_0 + B_0 = f'(x_0) + g'(x_0). \tag{8.11}$$

The argument sees no difference between $D_f \cap D_g \subset \mathbb{R}$ and $D_f \cap D_g \subset X$, and applies equally well to $\Phi : D_\Phi \to Y$ and $\Psi : D_\Psi \to Y$ as in Definition 8.4 and Theorem 8.5.

Next we consider the product function fg defined by

$$(fg)(x) = f(x)g(x)$$

for all $x \in D_f \cap D_g$ and multiply (8.8) and (8.10) to get

$$(fg)(x) = f(x)g(x) = (f(x_0)+A_0(x-x_0)+R_0(x))(g(x_0)+B_0(x-x_0)+S_0(x))$$

$$= f(x_0)g(x_0) + \underbrace{A_0(x-x_0)g(x_0) + f(x_0)B_0(x-x_0)}_{(fg)'(x_0)(x-x_0)?} + T_0(x). \tag{8.12}$$

The remainder term $T_0(x)$ consists of the 6 other combinations of the 3 terms in (8.8) and (8.10). To conclude that fg is differentiable in x_0 you must check that each of these 6 terms is $o(|x-x_0|)$ as $x \to x_0$. This is formulated as an exercise below. We have indeed that

$$T_0(x) = o(|x-x_0|_X) \quad \text{as} \quad x \to x_0. \tag{8.13}$$

We then read of from (8.12) that

$$(fg)'(x_0) = g(x_0)A_0 + f(x_0)B_0 = g(x_0)f'(x_0) + f(x_0)g'(x_0). \tag{8.14}$$

The argument sees 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 (8.14).

The argument also applies equally well if Φ and Ψ map to a normed algebra Y and are as in Definition 8.4 and Theorem 8.5. If the multiplication is commutative we have

$$\underbrace{A_0(x-x_0)}_{in\ Y}\underbrace{\Psi(x_0)}_{in\ Y} = \underbrace{\Psi(x_0)A_0}_{in\ L(X,Y)}(x-x_0) \in Y$$

and (8.14) remains unaltered. If multiplication in Y is not commutative we have that $(\Phi\Psi)'(x_0)$ is defined by

$$((\Phi\Psi)'(x_0))(h) = (\Phi'(x_0)(h))\Psi(x_0) + \Phi(x_0)(\Psi'(x_0)(h)). \tag{8.15}$$

It belongs to $L(X,Y)$ because, using[2] $|yz|_Y \le |y|_Y\,|z|_Y$, we have

$$\begin{aligned}|(\Phi\Psi)'(x_0)(h)|_Y &\le |(\Phi'(x_0)(h))\Psi(x_0)|_Y + |\Phi(x_0)(\Psi'(x_0)(h))|_Y\\ &\le |\Phi'(x_0)(h)|_Y\,|\Psi(x_0)|_Y + |\Phi(x_0)|_Y\,|(\Psi'(x_0)(h))|_Y\\ &\le |\Phi'(x_0)|_{L(X,Y)}\,|h|_X\,|\Psi(x_0)|_Y + |\Phi(x_0)|_Y\,|\Psi'(x_0)|_{L(X,Y)}\,|h|_X,\end{aligned}$$

whence

$$|(\Phi\Psi)'(x_0)|_{L(X,Y)} \le |\Phi'(x_0)|_{L(X,Y)}\,|\Psi(x_0)|_Y + |\Phi(x_0)|_Y\,|\Psi'(x_0)|_{L(X,Y)}.$$

Next we look at the remainder term $T_0(x)$, which is the sum of

$$\Phi(x_0)S_0(x) + R_0(x)\Psi(x_0),$$

$$A_0(x-x_0)B_0(x-x_0),$$

$$A_0(x-x_0)S_0(x) + R_0(x)B_0(x-x_0),$$

and

$$R_0(x)S_0(x).$$

Exercise 8.7. Prove in the general setting of normed spaces X and Y that (8.13) holds. That is, use

$$\forall_{\varepsilon>0}\,\exists_{\delta>0}\,\forall_{x\in X}\quad 0<|x-x_0|_X<\delta \implies |R_0(x)|_Y<\varepsilon|x-x_0|_X \tag{8.16}$$

and the same statement for $S_0(x)$ to prove the same statement for each of the above 6 terms in $T_0(x)$.

8.2 Differential calculus: the chain rule

This concerns a rule which is in fact easier than (8.14), easier because it only needs *linear* algebra. It could replace the rules in Section 8.1 as this exercise should help you to reflect:

Exercise 8.8. 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?

[2] In $|x+y|\le|x|+|y|$ and $|xy|\le|x|\,|y|$ we did not put constants but we might.

Now 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

$$y_0 = f(x_0) \in D_g.$$

There is no difference in the arguments below between

$$D_f \subset \mathbb{R}, \quad f : D_f \to \mathbb{R}, \quad D_g \subset \mathbb{R}, \quad g : D_g \to \mathbb{R},$$

and

$$D_\Phi \subset X, \quad \Phi : D_\Phi \to Y, \quad D_\Psi \subset Y, \quad \Psi : D_g \to Z,$$

with X, Y, Z normed spaces.

A common notation is

$$\Psi(\Phi(x) = (\Psi \circ \Phi)(x),$$

with $\Psi \circ \Phi$ standing for the composition of Ψ and Φ. In other words,

$$x \xrightarrow{\Phi} \Phi(x) \xrightarrow{\Psi} \Psi(\Phi(x)) = (\Psi \circ \Phi)(x)$$

is long for the short notation

$$x \xrightarrow{\Psi \circ \Phi} \Psi(\Phi(x)),$$

a map we want to linearise around x_0.

To do so,

$$\Psi(y) = \Psi(y_0) + B_0(y - y_0) + S_0(y), \quad B_0 = \Psi'(y_0)$$

has to be combined 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 $S_0(y)$ to have the properties we have seen before, namely (8.7,8.16) for Φ, i.e.

$$\forall_{\varepsilon>0}\, \exists_{\delta>0} \quad 0 < |x - x_0|_X < \delta \implies |R_0(x)|_Y < \varepsilon |x - x_0|_X,$$

and

$$\forall_{\varepsilon>0}\, \exists_{\delta>0}\, 0 < |y - y_0|_Y < \delta \implies |S_0(y)|_Z < \varepsilon |y - y_0|_Y \tag{8.17}$$

for Ψ. In particular 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.

The properties of the remainder terms $R_0(x)$ and $S_0(y)$ should carry over to the remainder term $T_0(x)$ in

$$\Psi(\Phi(x)) = \Psi(y) = \Psi(y_0) + B_0(\underbrace{\Phi(x) - \Phi(x_0)}_{y - y_0}) + S_0(y) =$$

$$\Psi(y_0) + B_0 A_0(x - x_0) + \underbrace{B_0 R_0(x) + S_0(y)}_{T_0(x)}.$$

The first term in $T_0(x)$ exists for all $x \in X$ and is immediately estimated via

$$|B_0 R_0(x)|_Z \leq |B_0|_{L(Y,Z)} \, |R_0(x)|_Y,$$

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

$$|S_0(y)|_Z < \varepsilon |y - y_0|_Y \quad \text{if} \quad 0 < |y - y_0|_Y < \delta,$$

with $\delta > 0$ as in (8.17). What we want is an estimate in terms of a multiple of $\varepsilon |x - x_0|_X$ if $0 < |x - x_0|_X < \tilde{\delta}$ for some other $\tilde{\delta} > 0$ chosen depending on the positive ε we started with.

If by chance $y = y_0$ there's no work to be done. If not, then we need

$$0 < |y - y_0|_Y < \delta$$

if we want to conclude via (8.17). We actually have

$$\begin{aligned} |y - y_0|_Y = |\Phi(x) - \Phi(x_0)|_Y = |A_0(x - x_0) + R_0(x)|_Y \\ \leq |A_0|_{L(X,Y)} \, |x - x_0|_X + |R_0(x)|_Y \\ < (|A_0|_{L(X,Y)} + 1) \, |x - x_0|_X \quad \text{if} \quad 0 < |x - x_0|_X < \delta_R, \end{aligned}$$

in which $\delta_R > 0$ is provided by (8.2) applied with $\varepsilon = 1$. So we indeed conclude via (8.17) if

$$0 < |x - x_0|_X < \frac{\delta}{|A_0|_{L(X,Y)} + 1} = \tilde{\delta},$$

which then implies that the second term in $T_0(x)$ exists so that x is actually in the domain of $\Psi \circ \Phi$. Moreover the second term is estimates by

$$|S_0(y)|_Z < \varepsilon |y - y_0|_Y < \underbrace{\varepsilon(|A_0|_{L(X,Y)} + 1)}_{\tilde{\varepsilon}} |x - x_0|_X.$$

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 proved the chain rule.

Theorem 8.9. *Let x_0 be an interior point of the domain of Φ, assume Φ differentiable in x_0, let $y_0 = \Phi(x_0)$ be an interior point of the domain of Ψ, and assume that Ψ differentiable in y_0. Then x_0 is in the interior of the domain of $\Psi \circ \Phi$ and $\Psi \circ \Phi$ is differentiable in x_0 with*

$$(\Psi \circ \Phi)'(x_0) = \Psi'(y_0)\Phi'(x_0). \tag{8.18}$$

Exercise 8.10. Derive and prove the differentiation rules for fg and $\frac{g}{f}$ if f and g are real valued functions from Exercise 8.8 and Theorem 8.9. Hint: use also $y \to \frac{1}{y}$.

8.3 Critical points and the mean value theorem

A critical point[3] of a differentiable function $f : \mathcal{O} \to \mathbb{R}$ is by definition a point $\xi \in \mathcal{O}$ where $f'(\xi) = 0$. This statement makes sense for $\mathcal{O} \subset X$ open and X any real normed space. The following theorem is formulated for the case that $\mathcal{O} = (a, b) \subset \mathbb{R} = X$ and $f : (a, b) \to \mathbb{R}$ differentiable, but generalises to $f : \mathcal{O} \to \mathbb{R}$.

Theorem 8.11. *Let $f : (a, b) \to \mathbb{R}$ and assume that $\xi \in (a, b)$ is such that $f(x) \leq f(\xi)$ for all $x \in (a, b)$. Then $f'(\xi) = 0$ provided f is differentiable in ξ.*

Theorem 8.12. *The mean value theorem: if $f \in C([a, b])$ is differentiable on (a, b) then for at least one ξ in (a, b) it holds that*

$$\frac{f(b) - f(a)}{b - a} = f'(\xi).$$

Remember this theorem as stating that the difference quotient on the left is equal to a differential quotient in some point ξ strictly between a and b.

[3] Also: a stationary point.

In the special case that $f(a) = f(b)$ the point ξ appears as maximizer or minimizer of f on $[a, b]$. Such a maximizer and minimizer must exist in $[a, b]$ in view of Theorem 4.29.

If the maximizer ξ lies in (a, b) then $f'(\xi) = 0$ in view of Theorem 8.11, which is exactly what Theorem 8.12 asserts in the case that $f(a) = f(b)$. The same conclusion holds if a 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 $f(x) = f(a) = f(b)$ for all $x \in [a, b]$. Then the statement in the theorem is trivially true. This completes the proof for the case that $f(a) = f(b)$, which is also called Rolle's Theorem[4].

Exercise 8.13. Reduce the general case in Theorem 8.12 to the special case $f(a) = f(b)$ and prove Theorem 8.12.

For $x, y \in X$ the function $t \to \xi(t) \in X$ defined by

$$\xi(t) = (1 - t)x + ty, \quad \xi : [0, 1] \to X, \tag{8.19}$$

parameterizes the interval

$$[x, y] = \{\xi(t) = (1 - t)x + ty;\ 0 \leq t \leq 1\}. \tag{8.20}$$

We also write

$$(x, y) = \{\xi(t) = (1 - t)x + ty;\ 0 < t < 1\}. \tag{8.21}$$

Theorem 8.14. *Let $x, y \in X$, X a normed space, $\mathcal{O} \subset X$ open, $[x, y] \subset \mathcal{O}$, and $f : \mathcal{O} \to \mathbb{R}$ differentiable. Then there exists $\xi \in (x, y)$ such that*

$$f(y) - f(x) = f'(\xi)(y - x).$$

Exercise 8.15. For the proof apply the chain rule to

$$t \to f(\xi(t)) \in \mathbb{R}, \tag{8.22}$$

and use Theorem 8.12 on the t-interval $[0, 1]$

[4] Read about Rolle in wikipidia.

8.4 Differentiability of inverse functions

Consider the functions f and g in Theorem 4.25. 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

$$f'(x_0)g'(y_0) = 1, \tag{8.23}$$

a statement which is symmetric in f and g.

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 $f'(x_0) = 1$, meaning that

$$f(x) = x + o(x) \quad \text{as} \quad x \to 0, \tag{8.24}$$

i.e.

$$\forall_{\varepsilon>0}\, \exists_{\delta>0} \quad 0 < |x| < \delta \implies |f(x) - x| < \varepsilon|x|. \tag{8.25}$$

The inequality for $|f(x) - x|$ means that

$$(1-\varepsilon)x < y < (1+\varepsilon)x \quad \text{if} \quad 0 < x < \delta \quad \text{and} \quad y = f(x), \tag{8.26}$$

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

$$g(y) = y + o(y) \quad \text{as} \quad y \to 0. \tag{8.27}$$

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 (8.25) implies that

$$\forall_{\varepsilon\in(0,1)}\, \exists_{\delta>0} \quad \begin{matrix} 0 < x < \delta \\ y = f(x) \end{matrix} \implies (1-\varepsilon)x < y < \frac{1}{1-\varepsilon}x, \tag{8.28}$$

and likewise[5] for $-\delta < x < 0$.

[5] With the same δ given $0 < \varepsilon < 1$, and with reversed inequalities for y.

Next observe that (8.28) and its version for $x < 0$ in turn imply

$$\forall_{\varepsilon\in(0,1)}\,\exists_{\delta>0}\quad 0<|x|<\delta \implies |f(x)-x|<\frac{\varepsilon}{1-\varepsilon}|x|, \tag{8.29}$$

since

$$\frac{1}{1-\varepsilon}=1+\frac{\varepsilon}{1-\varepsilon}.$$

But (8.29) and (8.25) are equivalent, by setting

$$\tilde{\varepsilon}=\frac{\varepsilon}{1-\varepsilon},$$

and thus (8.28) and its version for $x < 0$ make up for an equivalent definition of (8.24) stating that

$$\forall\varepsilon\in(0,1)\,\exists\delta>0:\quad G_\delta=\{(x,y)\in\mathbb{R}^2:\,0<|x|<\delta,\,y=f(x)\}\subset S_\varepsilon, \tag{8.30}$$

in which

$$S_\varepsilon=\{(x,y)\neq(0,0):\,\frac{1}{1-\varepsilon}<\frac{y}{x}<1-\varepsilon\} \tag{8.31}$$

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.

8.5 Examples of inverse functions

Exercise 8.16. In view of Section 8.4 and Theorem 4.25 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 7.17 and explain why $f = \exp$.

Exercise 8.17. Show that $\exp(x + y) = \exp(x)\exp(y)$ for all $x, y \in \mathbb{R}$, and that with $e = \exp(1)$ defined by

$$\ln e=\int_1^e\frac{1}{x}\,dx=1,$$

it follows that

$$\exp(\frac{p}{q}) = e^{\frac{p}{q}} = \sqrt[q]{e^p}$$

for all $p \in \mathbb{Z}$ and all $q \in \mathbb{N}$. By general agreement we define $e^x = \exp(x)$ for all other $x \in \mathbb{R}$ as well.

Likewise for x^α with $x > 0$. Via

$$x^n = (e^{\ln x})^n = e^{n \ln x}$$

for $n \in \mathbb{N}$, but also with $n \in \mathbb{N}$ replaced by $r = \frac{p}{q} \in \mathbb{Q}$ and finally by general agreement:

$$x^\alpha = e^{\alpha \ln x} \quad \text{for} \quad x > 0 \quad \text{and} \quad \alpha \in \mathbb{R} \tag{8.32}$$

Exercise 8.18. 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 8.19. Show that

$$x \to \sin x$$

s strictly increasing on $(-\frac{\pi}{2}, \frac{\pi}{2})$ and has an inverse function

$$y \to \arcsin y$$

on $(-1, 1)$ with derivative

$$\frac{1}{\sqrt{1-y^2}}$$

and derive a power series expression for $\arcsin y$ for $|y| < 1$.

Exercise 8.20. Consider

$$x \to \cos x$$

on $(0, \pi)$. Show for the inverse that $\arccos y + \arcsin y$ is constant on $(-1, 1)$. Which constant?

Exercise 8.21. Show that $\sin$ is a periodic function. Its period is by definition 2π. Show that $-\sin(-x) = \sin x = \sin(\pi - x) > 0$ for $0 < x < \pi$.

Exercise 8.22. Solve the differential equation in Exercise 7.26 via

$$\frac{f'_\alpha(x)}{f'_\alpha(x)} = \frac{\alpha}{1+x}$$

and integration from 1 to x. Prove that

$$(1+x)^\alpha = 1 + \alpha x + \frac{\alpha(\alpha-1)}{2}x^2 + \frac{\alpha(\alpha-1)(\alpha-2)}{3\cdot 2}x^3 + \cdots = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k \quad (8.33)$$

for all $x \in \mathbb{R}$ with $|x| < 1$.

Exercise 8.23. Take $\alpha = \frac{1}{2}$ and square the series in (8.33). 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 $|x| < 1$, which was perhaps known to extent known by the Babylonians.

Exercise 8.24. Write out the first few terms of

$$\sqrt[n]{1+x} = 1 + \frac{x}{n} + \cdots \quad \text{en} \quad \frac{1}{\sqrt[n]{1+x}} = 1 - \frac{x}{n} + \cdots$$

8.6 Asymptotic formulas

The notation

$$f(x) \sim g(x) \quad \text{for} \quad x \to a \tag{8.34}$$

means that

$$\frac{f(x)}{g(x)} \to 1 \quad \text{if} \quad x \to a,$$

in which usually a is 0 or ∞. Similarly for $n \in \mathbb{N}$ and $n \to \infty$, for instance

$$n! \sim (\frac{n}{e})^n \sqrt{2\pi n} \quad \text{als} \quad n \to \infty. \tag{8.35}$$

Exercise 8.25. Investigate $f : x \to x^x$ with $x \in \mathbb{R}^+$ using (8.32). 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 8.26. 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 definedby $x \ln x = \ln y$, 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 $y \to \infty$.

8.7 Some strange examples

Exercise 8.27. 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 Section 8.

For g in Exercise 8.27 you can take a strange function like for instance g defined by $g(x) = 0$ for $x \in \mathbb{Q}$ and $g(x) = x$ for $x \notin \mathbb{Q}$. Then $f : \mathbb{R} \to \mathbb{R}$ is discontinuous in every $x \neq 0$ while differentiable in $x = 0$.

Exercise 8.28. Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(0) = 0$ and

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

for $x \neq 0$. Show that f is differentiable in every $x \in \mathbb{R}$ but that $f'(x)$ is unbounded on $[0, 1]$.

Exercise 8.29. 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.

Exercise 8.30. Let R be the radius of convergence of the power series $P(x)$. Then (7.38) holds for all x_0 and h in $\mathbb{R}$ with $|x_0| + |h| < R$, as the sum of an absolutely convergent series. Hint: this has to do with unconditional convergence, which will be discussed somewhere in these notes (not translated yet).

9 From integral- to differential calculus

We have now seen both integral and differential calculus. The presentation so far has avoided the direct question about the relation between the two, but do take note of Sections 7.3 and 7.5. The present chapter was conceived to be independent of Chapter 7.4 and basically builds on Section 5.5 only. We ask about the properties of the function F defined by[1]

$$F(x) = \int_a^x f(s)\,ds. \tag{9.1}$$

Think of

$$\int_1^x \frac{1}{s}\,ds$$

for example, an integral that in fact makes sense for all real x with $x > 0$.

9.1 The fundamental theorem of calculus

Working towards (8.1) we write

$$F(x) = F(x_0) + \int_{x_0}^x f(s)\,ds = F(x_0) + \int_{x_0}^x f(x_0)\,ds + \int_{x_0}^x (f(s) - f(x_0))\,ds,$$

so with $h = x - x_0$ it follows that

$$F(x) = F(x_0) + f(x_0)h + R_0(h), \tag{9.2}$$

in which

$$R_0(h) = \int_{x_0}^{x_0+h} (f(s) - f(x_0))\,ds. \tag{9.3}$$

Thus F is differentiable in x_0 with $F'(x_0) = f(x_0)$ if $R_0(h) = o(|h|)$ as $h \to 0$. Since the integral in (9.3) is over an interval of length h, continuity of f in x_0 will suffice to conclude. 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

$$|R_0(h)| \le |\int_{x_0}^{x_0+h} |f(s) - f(x_0)|\,ds| < \varepsilon|h| \quad \text{if} \quad 0 < |h| \le \delta \tag{9.4}$$

and $x = x_0 + h \in [a, b]$. This proves the following theorem.

[1] We already introduced the map $f \xrightarrow{A} F$ in Section 5.7.

Theorem 9.1. *Let $a, b \in \mathbb{R}$ with $a < b$. Define for $f \in \mathcal{RI}([a,b])$ the function $F \in C([a,b])$ by*

$$F(x) = \int_a^x f(s)\,ds.$$

Then F is differentiable in every $x_0 \in [a,b]$ where f is continuous, with derivative $F'(x_0) = f(x_0)$.

We recall that we see $F'(x_0) = f(x_0)$ as a linear map

$$h \to F'(x_0)h \tag{9.5}$$

for every such x_0. Note that x_0 is now also allowed to be one of the boundary points, for which case the obvious one-sided statement[2] that F is differentiable was not given yet. If F is differentiable in every $x \in [a,b]$ with $F'(x) = f(x)$, then F is called a primitive (functie) of f, or anti-derivative.

Theorem 9.1 thus says that every continuous function $f : [a,b] \to \mathbb{R}$ has a primitive on $[a,b]$. For this particular primitive we have that[3]

$$\int_a^b f(x)\,dx = F(b) - F(a), \tag{9.6}$$

because $F(a) = 0$, and if we add a constant to F this does not change. But does (9.6) hold for every primitive of F of f? To put it differently, is every primitive of f of the form (9.1), up to an additive constant? Applied to the difference of two primitive functions Theorem 8.12 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.

Theorem 9.2. *The fundamental theorem of calculus: for every $f \in C([a,b])$ it holds that*

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

in which F is any primitive of f. Such a primitive exists in view of (9.1). If G is any other primitive than the primitive defined by (9.1), then $F - G$ is constant on $[a,b]$.

The formula in Theorem 9.2 is often written as

$$\int_{[a,b]} dF = F(x)\Big|_a^b \quad \text{with} \quad dF = F'(x)dx = f(x)dx \tag{9.7}$$

[2] Formulate this statement for $x_0 = a$ and $x_0 = b$.

[3] Have look at Exercise 5.13 now.

and

$$F(x)\big|_a^b = [F(x)]_a^b = F(b) - F(a).$$

A formal notation with the d of F which will get a deeper and more general meaning in vector calculus and expressions like $dF = f(x,y)dx + g(x,y)dy$ and also 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 a 1-form can be the d of a 0-form. In Theorem 9.2 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.

9.2 A mean value theorem in integral form

Exercise 9.3. Theorem 9.2 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

$$F(b) - F(a) = \int_0^1 F'((1-t)a+tb)(b-a)\,dt = \int_0^1 F'((1-t)a+tb)\,dt\,(b-a), \quad (9.8)$$

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 $[a,b]$, with Lipschitz constant L, the first integral in (9.8) 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

$$F(x) = F(a) + F'(a)(x-a) + R(x;a) \tag{9.9}$$

with

$$R(x;a) = \int_0^1 (F'((1-t)a + tx) - F'(a))(x-a)\,dt,$$

and therefore

$$|R(x;a)| \leq \int_0^1 Lt|x-a|^2\,dt = \frac{L}{2}|x-a|^2. \tag{9.10}$$

In (9.9) we have a linear approximation with a remainder term estimated in (9.10) 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 $x \to a$. Note that $O(|x-a|^2)$ implies $o(|x-a|)$ but in general it is not true that $o(|x-a|)$ implies $O(|x-a|^2)$.

9.3 The generalised mean value formula

Theorem 9.4. *Let X be Banach space. For $f : [a,b] \to X$ continuous let the function $F : [a,b] \to X$ be defined 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 9.4 says that F is a primitive of f, and that for this primitive

$$\int_a^b f(s)\,ds = F(b) - F(a), \tag{9.11}$$

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 9.5. Show that for every $\psi \in X^*$ the real vaued 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 8.12.

We conclude that $\psi(G(b)) - \psi(G(a)) = 0$ for every $\psi \in X^*$. For $y = G(b) - G(a)$ it thus holds that $\psi(y) = 0$ for every $\psi \in X^*$. We say that X^* separates the points of X if this implies that $y = 0$. If so 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 9.6. *Let X be a Banach space on which X^* separates the points. 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 9.4.

Summing up, the mean value integral formula (9.8) 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 8.12, under a much stronger assumption than Theorem 8.12, exclusively for $\mathbb{R}$-valued functions.

For continuously differentiable $F : \mathcal{O} \to Y$, Y a Banach space, $\mathcal{O}, x, y$ as in Theorem 8.14, we apply Theorem 9.6 with $a = 0$ and $b = 1$ to the function defined by (**??**), and conclude that

$$F(y) - F(x) = \int_0^1 F'((1-t)x + ty)(y-x)\,dt, \tag{9.12}$$

as a Y-valued integral, which we can write as

$$F(y) - F(x) = \int_0^1 F'((1-t)x + ty)\,dt(y-x), \tag{9.13}$$

a $B(X,Y)$-valued integral acting on $y - x \in X$. This version of the mean value theorem will be used in the proof of Theorem 10.4.

[4] $F'(x) = f(x)$ for all $x \in [a,b]$.

9.4 More on exp and ln

Exercise 9.7. 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 9.8. Same question as in Exercise 9.7 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 9.9. 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 9.8. 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[5] just computed is equal to $F_\alpha(x)$?

9.5 Integrals with parameters

Let's consider

$$j(t) = \int_0^1 f(x, t)\, dx$$

in which, for each t in a t-interval $[0, 1]$, the function $x \to f(t, x)$ is continuous on the x-interval $[0, 1]$. Then $j(t)$ is well defined. What do we need to have j differentiable? Let's examine a follow your nose argument for what (the one-sided) derivative $j'(0)$ should be and see what we need to prove it.

If we use the mean value theorem in the form of Theorem 8.12 itself[6], for every fixed $x \in [0, 1]$ applied to $t \to f(x, t)$, it follows that

$$f(t, x) = f(0, x) + f_t(\tau, x)\, t,$$

[5] NB Take note of $\alpha = -1$, but also of $\alpha = \pm\frac{1}{n}$.

[6] The integral form would require the use of not yet discussed double integrals.

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

$$f(t,x) = f(0,x) + f_t(0,x)t + \underbrace{f_t(\tau(x),x) - f_t(0,x)}_{R(t,x)}. \tag{9.14}$$

This defines $R(t,x)$ and if in (9.14), with t fixed, everything is continuous in x, it follows that

$$\begin{aligned} j(t) = \int_0^1 f(t,x)\,dx &= \int_0^1 (f(0,x) + f_t(0,x)t + R(t,x))\,dx \\ &= j(0) + t\int_0^1 f_t(x,0)\,dx + \int_0^t R(t,x)\,dx \\ &= j(0) + t\int_0^1 f_t(x,0)\,dx + r(t), \end{aligned} \tag{9.15}$$

in which

$$r(t) = \int_0^t R(t,x)\,dx \quad \text{with} \quad R(t,x) = \underbrace{(f_t(\tau(x),x) - f_t(0,x))}_{<\varepsilon\,?}t.$$

If we assume that

$$x \to f(t,x)$$

and

$$x \to f_t(0,x) = g(t,x)$$

are continuous on $[0,1]$ we don't have to worry about existence of the integrals. The integral $r(t)$ of $R(t,x)$ in (9.15) is then also continuous. The second expression with $\tau(x) \in (0,t)$ above (9.14) can now be used to establish $r(t) = o(t)$ as $t \to 0$.

Indeed, for the remainder term $r(t)$ we need $|r(t)| < \varepsilon t$ for t sufficiently small. Thus, if for $f_t(t,x) = g(t,x)$ it holds that

$$|g(t,x) - g(0,x)| < \varepsilon \tag{9.16}$$

if $t \in (0,\delta)$ for all $x \in [0,1]$ simultaneously for some $\delta > 0$, we will be happily done.

How can this uniform ε-statement fail to be true? Only if for some $\varepsilon > 0$ there exists a sequence of points (t_n, x_n) with $0 < t_n \to 0$ for which

$$|g(t_n,x_n) - g(0,x_n)| \geq \varepsilon.$$

But then the sequence x_n has a convergent subsequence x_{n_k} with limit $\bar{x} \in [0,1]$ and both sequences of points (t_n, x_n) and of points $(0, x_n)$ converge to $(0, \bar{x})$ preventing $(t,x) \to g(t,x)$ from being continuous in every point $(0,x)$ with $x \in [0,1]$.

Theorem 9.10. *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 9.11. *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 9.12. To prove the continuity of the derivative you need to prove: $t \to \int_a^b g(t,x)\,dx$ is continuous on I if $(t,x) \to g(t,x)$ is continuous on $I \times [a,b]$. Hint: use a uniform ε-argument.

9.6 Partial integration and Taylor polynomials

Via Theorem 9.2 the Leibniz rule in Theorem 8.6 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

$$\int_\alpha^\beta x(t)y'(t)\,dt = [x(t)y(t)]_\alpha^\beta - \int_\alpha^\beta x'(t)y(t)\,dt. \tag{9.17}$$

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 9.2 applied to a product of two continuously differentiable functions[7].

We apply this rule to a simple so-called boundary value problem. For given $f \in C([0,1])$ we ask for a function u such that

$$-u''(x) = f(x) \quad \text{for all} \quad 0 \le x \le 1, \quad \text{and} \quad u(0) = u(1) = 0. \tag{9.18}$$

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

$$\underbrace{1\,F(s)}_{G'(s)F(s)} = \underbrace{(s-a)'}_{G'(s)}F(s) = (\underbrace{(s-a)}_{G(s)}F(s))' - \underbrace{(s-a)}_{G(s)}F'(s)$$

$$= \underbrace{((s-a)F(s))'}_{(G(s)F(s))'} - \underbrace{(s-a)f(s)}_{G(s)F'(s)},$$

in which $1 = G'(s)$ with $G(s) = s - a$ and a free to choose.

The primitive of $F(x)$ then rewrites as

$$\int_0^x F(s)\,ds = [(s-a)F(s)]_0^x - \int_0^x (s-a)f(s)\,ds = \int_0^x (x-s)f(s)\,ds. \tag{9.19}$$

With $a = x$ it follows that

$$u(x) = u'(0)x - \int_0^x (x-s)f(s)\,ds$$

and $x = 1$ gives

$$u'(0) = \int_0^1 (1-s)f(s)\,ds.$$

[7] And in a much more general setting in fact.

Therefore

$$u(x) = \int_0^1 (1-s)f(s)\,ds\,x - \int_0^x (x-s)f(s)\,ds$$

$$= x\int_x^1 (1-s)f(s)\,ds + (1-x)\int_0^x sf(s)\,ds = \int_0^1 A(x,s)f(s)\,ds.$$

The expression

$$A(x,s) = \begin{cases} (1-x)s & \text{for } 0 \le s \le x \\ (1-s)x & \text{for } x \le s \le 1 \end{cases} \tag{9.20}$$

is called the kernel for the solution operator, which gives u in terms of f as

$$u(x) = \int_0^1 A(x,s)f(s)\,ds. \tag{9.21}$$

You may prefer to memorize the integration by parts formula as

$$\int_a^b F(x)G'(x)\,dx = \left[F(x)G(x)\right]_a^b - \int_a^b F'(x)G(x)\,dx. \tag{9.22}$$

It's handy for computing integrals, but also for taking primitives of primitives, as we just saw and see again below.

Exercise 9.13. 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 (9.22) 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 9.14. In the context of Exercise 9.13 let

$$F_{n+1}(x) = \int_a^x F_n(s)\,ds \quad (n = 1, 2, 3\ldots).$$

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 9.15. Modify the scheme in Exercise 9.14 as

$$F_0(x) = f(x), \quad F_n(x) = b_n + \int_a^x F_{n-1}(s)\,ds \quad (n = 1, 2, 3\ldots), \tag{9.23}$$

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 $m = m_n(x,a)$ are 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

$$f(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}(0)}{k!}(x-a)^k + \underbrace{\frac{f^{(n)}(\sigma)}{n!}(x-a)^n}_{\frac{1}{(n-1)!}\int_a^x (x-s)^{n-1} f^{(n)}(s)\,ds.} \tag{9.24}$$

for some σ bewteen a and x.

The result in (9.24) holds in fact without the assumption that $f^{(n)}$ is continuous, with σ strictly between a and x, as a clever application of Theorem 8.12) shows. The case $n = 1$ reduces to Theorem 8.12,

9.7 Substitution rule for integrals

Write some text to go with these formulas:

$$\int_\alpha^\beta x(t)\,\underbrace{y'(t)\,dt}_{dy} = [x(t)y(t)]_\alpha^\beta - \int_\alpha^\beta y(t)\,\underbrace{x'(t)\,dt}_{dx},$$

$$\int_\alpha^\beta \underbrace{F'(x(t))}_{f(x(t))} x'(t)\,dt = F(x(\beta)) - F(x(\alpha)) = \int_{x(\alpha)}^{x(\beta)} F'(x)\,dx = \int_a^b \underbrace{F'(x)}_{f(x)}\,dx,$$

$$\int_a^b f(x)\,dx = \int_\alpha^\beta f(x(t))x'(t)\,dt. \tag{9.25}$$

9.8 Stirling's formule via scalings and limits

Exercise 9.16. 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 $n!$

NB These are improper integrals, defined via

$$\int_0^\infty = \lim_{R\to\infty}\int_0^R.$$

Exercise 9.17. Sketch the graph $y = x^n e^{-x}$ (for n not too large) in the x, y-plane. Where's the top of the moumtain?

Exercise 9.18. 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 9.19. Write

$$g_n(x) = e^{-\psi_n(x)} \quad \text{met} \quad \psi_n(x) = -\ln(g_n(x)),$$

and verify that

$$\psi_n(x) = x - n\ln(1+\frac{x}{n}) = n(\frac{x}{n} - \ln(1+\frac{x}{n})) = n\psi_1(\frac{x}{n}).$$

Put $x = s\sqrt{n}$ to conclude that

$$n! = (\frac{n}{e})^n \sqrt{n} \int_{-\sqrt{n}}^{\infty} e^{-n\Psi(\frac{s}{\sqrt{n}})} ds \tag{9.26}$$

and show that

$$\int_{-\sqrt{n}}^{\infty} e^{-n\Psi(\frac{s}{\sqrt{n}})} ds \to \int_{-\infty}^{\infty} e^{-\frac{1}{2}s^2} ds \tag{9.27}$$

as $n \to \infty$.

10 Locally defined implicit functions

If a function of two real variables, say

$$\mathbb{R}^2 = \{(x, y) : x, y \in \mathbb{R}\} \xrightarrow{F} \mathbb{R},$$

satisfies $F(0, 0) = 0$, then the equation

$$F(x, y) = 0 \tag{10.1}$$

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 Banach spaces X, Y and Z.

A special case is

$$F(x, y) = g(y) - x,$$

when the question concerns a possible inverse function f of a given function g, see Section 4.3. 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 8.4, where we already had a global inverse, we also asked about behaviour in a single point.

In this chapter we ask both about the existence of an implicit function f, as well as its properties, but only near a given point. Thus we want to solve $F(x, y) = 0$ for given x close to $x = 0$, hoping that near $y = 0$ precisely one solution $y = f(x)$ can be shown to exist.

Before we formulate a local implicit function theorem we discuss Newton's method for solving equations[1]. We assume that for fixed x near $x = 0$ the function

$$y \to F(x, y)$$

is differentiable near $y = 0$. The derivative is denoted by $F_y(x, y)$. The special case $F(x, y) = g(y) - x$ with partial derivative $F_y(x, y) = g'(y)$ is not really different, and will lead to a local *inverse function* theorem.

For fixed x we take $y_0 = 0$ as starting value for Newton's method. Thus we put the linear expansion of $F(x, y)$ around $y = 0$ equal to 0, solve for $y = y_1$, and use the linear the linear expansion of $F(x, y)$ around $y = y_1$ to find y_2, and so on. In every step we need $F_y(x, y_{n-1})$ to be invertible[2]. The next y_n is uniquely defined by

$$F(x, y_{n-1}) + F_y(x, y_{n-1})(y_n - y_{n-1}) = 0.$$

[1] Fast convergence of this method will be shown in Section 10.7.

[2] 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

$$y_n = y_{n-1} - F_y(x, y_{n-1})^{-1} F(x, y_{n-1}), \quad \text{starting from} \quad y_0 = 0. \tag{10.2}$$

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

$$x \to y = f(x). \tag{10.3}$$

We then expect/hope that

$$F(x, f(x)) = 0, \tag{10.4}$$

and that $y = f(x)$ is the only solution of (10.1) near $y = 0$. If so we also ask which conditions will make f continuous and differentiable in $x = 0$.

10.1 A simpler version of Newton's method

A direct proof of (fast) convergence of the sequence y_n defined by (10.2) will be given in Section 10.7 via an estimate of the form

$$|y_{n+1} - y_n| \le C|y_n - y_{n-1}|^2 \tag{10.5}$$

and requires a condition on the second derivative[3] of $y \to F(x, y)$. Here we avoid second derivatives of $y \to F(x, y)$ by simplifying the scheme: the derivative $F_y(x, y_{n-1})$ that has to be inverted in every step of Newton's scheme is replaced by $F_y(0, 0)$. The modified scheme reads

$$y_n = y_{n-1} - F_y(0, 0)^{-1} F(x, y_{n-1}), \tag{10.6}$$

and we look for an estimate which is very much like the estimate (3.3) for Heron's sequence: we lose the square in (10.5) 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 10.1. *Let $\bar\delta > 0$, $\bar\varepsilon > 0$,*

$$B = \{x \in \mathbb{R} : |x| < \bar\delta\}, \quad C = \{y \in \mathbb{R} : |y| < \bar\varepsilon\},$$

[3] In fact Lipschitz continuity of $y \to F_y(x, y)$ will suffice, see Section 10.7.

and suppose that $F : B \times C \to \mathbb{R}$ *has the properties that*

$F(0,0) = 0$*;*
$x \to F(x,0)$ *is continuous in* $x = 0$*;*
$(x,y) \to F_y(x,y)$ *is continuous in* $(0,0)$*;*
$F_y(0,0)$ *is invertible;*
$y \to F_y(x,y)$ *is continuous on* C *for every* $x \in B$*.*

Then there exists $\delta_0 > 0$ *and* $\varepsilon_0 > 0$ *for which the statement*

$$\forall (x,y) \in \bar{B}_{\delta_0} \times \bar{B}_{\varepsilon_0} : \quad F(x,y) = 0 \iff y = f(x)$$

holds, in which

$$B_{\delta_0} = \{x \in X : |x| \le \delta_0\}, \quad B_{\varepsilon_0} = \{y \in Y : |y| < \varepsilon_0\},$$

and $f : \bar{B}_{\delta_0} \to B_{\varepsilon_0}$ *is constructed via (10.6) 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 3.24, 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 (10.6) 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 $F(x,y) = 0$.

10.2 Estimating the steps: convergence

How large can y_1 be if $F(x,y_0) = F(x,0) \neq 0$? If we set

$$M_0 = |F_y(0,0)^{-1}| > 0. \tag{10.7}$$

then[4]

$$|y_1| = |F_y(0,0)^{-1} F(x,0)| \le M_0 \, |F(x,0)|. \tag{10.8}$$

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 (10.6) with $n = 2$. The trick however is to use (10.6) with both $n = 1$ *and* $n = 2$ via

$$y_2 - y_1 = y_1 - F_y(0,0)^{-1} F(x,y_1) - y_0 + F_y(0,0)^{-1} F(x,y_0)$$

[4] 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 (9.8), the mean value theorem in integral form[5], to $y \to F(x,y)$ with $a = y_1$ and $b = y_0$, x fixed. Combined with the third and fourth term the whole large factor equals[6]

$$\int_0^1 \left(F_y(x, ty_0 + (1-t)y_1) - F_y(0,0)\right)\, dt\, (y_0 - y_1),$$

in which we brought the other two terms inside the integral. We conclude that

$$y_2 - y_1 = F_y(0,0)^{-1} \int_0^1 \left(F_y(x, ty_0 + (1-t)y_1) - F_y(0,0)\right)\, dt\, (y_0 - y_1)$$

if $y \to F_y(x,y)$ is continuous on[7]

$$[y_0, y_1] = \{ty_0 + (1-t)y_1 :\ 0 \le t \le 1\} \tag{10.9}$$

for fixed x. Therefore

$$|y_2 - y_1| \le M_0 \int_0^1 |(F_y(x, ty_0 + (1-t)y_1) - F_y(0,0)|\, dt\, |y_0 - y_1|. \tag{10.10}$$

We now ask that $(x,y) \to F_y(x,y)$ is continuous[8] in $(0,0)$. In particular this continuity requires the existence of $F_y(x,y)$ for (x,y) close to $(0,0)$. To be precise we assume that for every $\eta > 0$ an $\varepsilon > 0$ can be found such that for all x and y the implication

$$|x| \le \varepsilon \text{ en } |y| \le \varepsilon \implies |F_y(x,y) - F_y(0,0)| < \eta \tag{10.11}$$

holds. Note that instead of an ε, δ-statement we used an η, ε-statement of continuity, with nonstrict inequalities on the left hand side of the implication arrow. In the end we want to have that $y = f(x)$, the limit of the x-dependent sequence y_n, satisfies $|y| < \varepsilon$ for all x with $|x| \le \delta$, for some $\delta > 0$ depending

[5] Which will also do for $F : X \times Y \to Y$.

[6] Look at (9.13), this argument is not restricted to $F : \mathbb{R}^2 \to \mathbb{R}$!

[7] This notation for $[y_0, y_1]$ does not require $y_0 < y_1$.

[8] For $F(x,y) = g(y) - x$ this means g' continuous in 0.

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 (10.10) and (10.11) we have that $|x|, |y_0|, |y_1| \leq \varepsilon$ implies

$$|y_2 - y_1| < M_0\eta\,|y_1 - y_0| \quad \text{in which} \quad M_0 = |F_y(0,0)^{-1}| > 0.$$

The inequality is strict unless $y_0 = y_1$, which is why we assumed $y_0 \neq y_1$. Thus the second step has

$$|y_2 - y_1| < \theta\,|y_1 - y_0| = \theta\,|y_1| \quad \text{with} \quad \theta = M_0\eta.$$

By the same reasoning we have

$$|y_3 - y_2| \leq \theta\,|y_2 - y_1|,$$

provided $|y_2| < \varepsilon$, and so on.

Any $\theta < 1$ is now fine for our purposes[9]: as long as $|y_n| < \varepsilon$ it holds that[10]

$$|y_{n+1}| = |y_{n+1} - y_0| \leq \underbrace{|y_{n+1} - y_n|}_{\leq\theta|y_n - y_{n-1}|} + \cdots + \underbrace{|y_2 - y_1|}_{<\theta|y_1|} + |y_1| <$$

$$(\theta^n + \cdots + 1)\,|y_1| < \frac{|y_1|}{1-\theta} \leq \frac{M_0|F(x,0)|}{1-\theta},$$

so

$$|y_{n+1}| < \frac{M_0|F(x,0)|}{1-\theta} < \frac{M_0\tilde{\varepsilon}}{1-\theta} = \frac{M_0\tilde{\varepsilon}}{1 - M_0\eta} \tag{10.12}$$

if $|x| \leq \tilde{\delta}$. Here $\tilde{\varepsilon} > 0$ is still to be chosen and $\tilde{\delta} > 0$ corresponds to $\tilde{\varepsilon}$ via the definition[11] of continuity of $x \to F(x,0)$ in $x = 0$.

Now choose

$$\eta_0 < \frac{1}{M_0}, \tag{10.13}$$

and then, given the corresponding ε_0 as in (10.11), 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[12] of continuity of $x \to F(x,0)$ in $x = 0$.

[9] In (3.3) we chose $\theta = \frac{1}{2}$ for the sake of simplicity only.

[10] In view of $1 + \theta + \theta^2 + \cdots = \frac{1}{1-\theta}$, see Section 1.5.

[11] With $\leq \tilde{\delta}$ instead of $< \tilde{\delta}$.

[12] With $\leq \tilde{\delta}_0$ 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

$$y \xrightarrow{\Phi} y - F_y(0,0)^{-1}F(x,y), \tag{10.14}$$

and the estimate

$$|\Phi(x,y) - \Phi(x,\tilde{y})| \le \theta\,|y - \tilde{y}| \tag{10.15}$$

with $\theta < 1$ and strict inequality if $y \neq \tilde{y}$. Equation $F(x,y) = 0$ is via (10.14) equivalent to $y = \Phi(x,y)$ because $F_y(0,0)^{-1}$, being the inverse of $F_y(0,0)$, is invertible. For the limit $y = f(x)$ the continuity[13] of $y \to \Phi(x,y)$ implies

$$y = \lim_{n\to\infty} y_{n+1} = \lim_{n\to\infty} \Phi(x, y_n) = \Phi(x,y).$$

Thus

$$\forall (x,y) \in \bar{B}_{\delta_0} \times \bar{B}_{\varepsilon_0} : \quad F(x,y) = 0 \iff y = f(x), \tag{10.16}$$

and Theorem 10.1 is proved.

10.3 Differentiable implicit functions

The implicit function in Theorem 10.1 satisfies

$$|f(x)| \le \frac{M_0|F(x,0)|}{1 - M_0\eta_0}, \tag{10.17}$$

in which η_0 was chosen at the beginning of Section 10.2, see (10.13). Estimate (10.17) 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?

[13] Continuity follows from differentiability.

Use (9.8) to write

$$0 = F(x, f(x)) = F(x,0) + F(x, f(x)) - F(x,0)$$

$$= F(x,0) + \int_0^1 F_y(x, tf(x))f(x)\,dt = F(x,0) + F_y(0,0)f(x) + R(x),$$

$$\text{with} \quad R(x) = \int_0^1 (F_y(x, tf(x)) - F_y(0,0))f(x)\,dt. \tag{10.18}$$

Clearly $x \to F(x,0)$ differentiable in $x = 0$ is the natural additional assumption, because then

$$0 = F(x, f(x)) = F_x(0,0)x + r(x) + F_y(0,0)f(x) + R(x), \tag{10.19}$$

with $r(x) = o(|x|)$ as $x \to 0$.

Theorem 10.2. *Let f be as in Theorem 10.1. If $x \to F(x,0)$ is differentiable in $x = 0$ then also f is differentiable in $x = 0$ and*

$$f'(0) = -F_y(0,0)^{-1}F_x(0,0).$$

The proof now follows the nose. Isolating $f(x)$ in (10.19) we have

$$f(x) = \underbrace{-F_y(0,0)^{-1}F_x(0,0)}_{f'(0)?}\, x \underbrace{-F_y(0,0)^{-1}r(x) - F_y(0,0)^{-1}R(x)}_{remainder}. \tag{10.20}$$

Since

$$|F_y(0,0)^{-1}r(x)| \le M_0|r(x)| \quad \text{and} \quad |F_y(0,0)^{-1}R(x)| \le M_0|R(x)|$$

it remains to be proved that $R(x) = o(|x|)$ as $x \to 0$. Given an arbitrary[14] $\varepsilon > 0$ we need to conclude that

$$|R(x)| < \varepsilon|x| \quad \text{if} \quad 0 < |x| < \delta$$

for some $\delta > 0$. Since $R(x)$ is given by (10.18) we use (10.11) again to conclude that

$$|R(x)| < \tilde{\eta}\,|f(x)| \quad \text{if} \quad |x| < \tilde{\varepsilon} \quad \text{and} \quad |f(x)| < \tilde{\varepsilon}. \tag{10.21}$$

The latter inequality will hold if $|x| < \tilde{\delta}$, $\tilde{\delta}$ corresponding to $\tilde{\varepsilon}$ in the established statement, via the construction and (10.17), that f is continuous in 0.

[14]Earlier we only took one fixed ε_0 corresponding to one fixed η_0 as in (10.13).

Restricting also to $|x| \le \delta_0$ we have

$$|R(x)| < \tilde{\eta}\,|f(x)| \le \frac{M_0\tilde{\eta}}{1 - M_0\eta_0}\,|F(x,0)|,$$

while

$$|F(x,0)| < (|F_x(0,0)| + \varepsilon_r)\,|x|$$

if $|x| < \delta_r$, where δ_r corresponds to some arbitrarily chosen but then fixed $\varepsilon_r > 0$ in the definition of $r(x) = o(|x|)$.

For given $\varepsilon > 0$ we then choose $\tilde{\eta} > 0$ such

$$\frac{M_0\tilde{\eta}}{1 - M_0\eta_0}(|F_x(0,0)| + \varepsilon_r) = \varepsilon,$$

take the corresponding $\tilde{\varepsilon}$ and $\tilde{\delta}$ as in and below (10.21). 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 (10.20) is $o(|x|)$ as $x \to 0$. This then completes the proof of Theorem 10.2.

Exercise 10.3. Actually the continuity of f in $x = 0$ follows directly from (10.20) and (10.18) if we assume that $|y| = |f(x)| \le \varepsilon_0$ with ε_0 chosen via (10.11) for (10.13). Use (10.19) in the form

$$0 = F(x,y) = F(x,0) + F_y(0,0)y + \underbrace{\int_0^1 (F_y(x,y) - F_y(0,0))y\,dt}_{\text{in norm less than } \eta_0|y| \quad \text{if } |x|,|y| \le \varepsilon_0}, \tag{10.22}$$

and derive that for solutions (x,y) of $F(x,y) = 0$ it holds that

$$|y| \le \frac{M_0|F(x,0)|}{1 - M_0\eta_0} \quad \text{if} \quad |x| \le \varepsilon_0 \text{ and } |y| \le \varepsilon_0. \tag{10.23}$$

Thus the existence of a solution of $F(x,y) = 0$ with $|y| \le \varepsilon_0$ for every x with $|x| < \delta_0 \le \varepsilon$ implies that $y \to 0$ if $F(x,0) \to 0$. Except for the choice of ε_0 this statement is independent of the construction of f and the uniqueness of the solution.

What about the other x-values in the domain $\bar{B}_{\delta_0}$ of f? We should have that f is differentiable in every x with $|x| \le \tilde{\delta}_0$ for some $0 < \tilde{\delta}_0 < \delta_0$, and

$$f'(x) = -F_y(x, f(x))^{-1}F_x(x, f(x)). \tag{10.24}$$

For every $x \in \bar{B}_{\delta_0}$ the validity of (10.24) 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 10.4. *The Implicit Function Theorem. Let X, Y and Z be Banach 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 10.1 and 10.2, which also hold in the general context of Banach spaces. The proofs can be copy-pasted replacing absolute values by norms in X, Y, Z and provide us with δ_0 and ε_0. The existence and continuity of $f'(x)$ requires restriction to a possibly smaller $\bar{B}_{\tilde{\delta}_0}$, as explained above and formulated in the final theorem.

10.4 Application to integral equations

This concerns smooth dependence of the solution of (6.13) on ξ, and

$$x(t) = \xi + \int_0^t f(x(s))\,ds$$

as the integral equation corresponding to the differential equation $x' = f(x)$ with initial condition $x(0) = \xi$ for X-valued functions $t \to x(t)$. Assume the existence and uniform continuity of f'. Let $x = x(\xi)$ be the solution of (10.25). Then

$$\xi \to x(\xi)$$

is continuously differentiable, and x_ξ is the solution of the integral equation corresponding to

$$y'(t) = f'(x(t))y(t) \quad \text{with} \quad y(0) = 1.$$

This is a bit of a project[15]. The first steps are sketched below.

For $a, b \in \mathbb{R}$ met $0 \in [a, b]$ and $\xi \in \mathbb{R}$ introduce

$$x = \xi + \Phi(x) \quad \text{with} \quad (\Phi(x))(t) = \int_0^t f(x(s))\,ds, \tag{10.25}$$

defining a new $\Phi(x) \in C([a, b])$ given and ("old") function $x \in C([a, b])$. Theorem 10.4 is applicable if

$$\Phi : C([a, b]) \to C([a, b])$$

is continuously differentientable.

To see why and how, take $h \in C([a, b])$ and write

$$(\Phi(x + h))(t) = \int_0^t f(x(s) + h(s))\,ds = \int_0^t [f(x(s) + \tau h(s))]_0^1\,ds$$

$$= \int_0^t \int_0^1 f'(x(s) + \tau h(s))h(s)\,d\tau\,ds$$

$$= \int_0^t \int_0^1 f'(x(s))h(s)\,d\tau\,ds + \underbrace{\int_0^t \int_0^1 (f'(x(s) + \tau h(s)) - f(x(s))h(s)\,d\tau\,ds}_{R(h;x)(t)}$$

$$= (\Phi'(x)h)(t) + R(h; x)(t),$$

[15] We shall also deal with pararameters in f, e.g. $f(x, \mu, \varepsilon)$ or so, see Section 12.5.

in which

$$h \xrightarrow{\Phi'(x)} \Phi'(x)h \quad \text{with} \quad (\Phi'(x)h)(t) = \int_0^t f'(x(s))h(s)\,ds, \tag{10.26}$$

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'.

10.5 For later: partial differentiability $\Longrightarrow$?

Exercise 8.8 contained an example of a differentiable function $F : \mathbb{R}^2 \to \mathbb{R}$. Differentiability of F in (x_0, y_0) via linear expansion rewrites as

$$F(x,y) = F(x_0,y_0) + a(x-x_0) + b(y-y_0) + R_0(x,y),$$

with

$$|R_0(x,y)| < \varepsilon \max(|x-x_0|, |y-y_0|) \quad \text{if} \quad \max(|x-x_0|, |y-y_0|) < \delta,$$

$\delta > 0$ depending on ε.

Exercise 10.5. 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 8.14 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

$$F(x, y) = F(x_0, y_0) + F_x(x_0, y_0)(x - x_0) + F_y(x_0, y_0)(y - y_0) + R_0 \quad (10.27)$$

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 10.6. *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 10.7. For X, Y, Z normed spaces and $\Phi : X \times Y \to Z$ the method via the mean value theorem fails. Write

$$\Phi(x, y) = \Phi(x_0, y_0) + \underbrace{\Phi(x, y) - \Phi(x_0, y)}_{\text{vary } x} + \underbrace{\Phi(x_0, y) - \Phi(x_0, y_0)}_{\text{vary } x}.$$

Assume Z is complete, $x \to \Phi(x, y_0)$ is continuously differentiable for $x \in X$ with $|x - x_0| < \delta_x$. If for each of these x the partial function $y \to \Phi(x, y)$ is continuously differentiable in $y \in Y$ with $|y - y_0| < \delta_y$, $\delta_x, \delta_y > 0$, and if $(x, y) \to \Phi_y(x, y)$ is continuous in (x_0, y_0), then Φ is differentiable in (x_0, y_0). Use (9.12) to prove this statement.

Exercise 10.8. If X, Y, Z are normed spaces, Z complete, and $\Phi : X \times Y \to Z$ has partial functions with partial derivatives Φ_x and Φ_y continuous on an open set O in $X \times Y$, then Φ is differentiable in every point of O and $\Phi' : O \to L(X \times Y, Z)$ is continuous and defined in every $(x_0, y_0) \in O$.

10.6 Stationary under a constraint

Suppose Φ and F are functions of x and y differentiable in $(x, y) = (0, 0)$, and f is a function of x differentiable in $x = 0$, for which it holds that

$$F_x(0, 0) + F_y(0, 0)f'(0) = 0. \tag{10.28}$$

In practice, f is the implicit function in Theorems 10.1 and 10.2. Then $y = f(x)$ describes the solution set of $F(x, y) = 0$ near $(0, 0)$, and we are interested in the restriction of Φ to the zero set of F. Clearly

$$x \xrightarrow{\phi} \phi(x) = \Phi(x, f(x))$$

is differentiable in $x = 0$, with

$$\phi'(0) = \Phi_x(0,0) + \Phi_y(0,0)f'(0). \tag{10.29}$$

If $F_y(0,0)$ is invertible it follows from (10.28) and (10.29) that

$$\phi'(0) = 0 \iff \Phi_x(0,0) = \Phi_y(0,0)F_y(0,0)^{-1}F_x(0,0). \tag{10.30}$$

Invertibility of $F_y(0,0) \in \mathbb{R}$ means that $F_y(0,0) \neq 0$, whence

$$\phi'(0) = 0 \iff \Phi_x(0,0)F_y(0,0) = \Phi_y(0,0)F_x(0,0),$$

equivalent to the existence of $\lambda \in \mathbb{R}$ for which it holds that

$$\begin{pmatrix} \Phi_x(0,0) \\ \Phi_y(0,0) \end{pmatrix} = \lambda \begin{pmatrix} F_x(0,0) \\ F_y(0,0) \end{pmatrix}.$$

This is a special case of the statement in Lagrange multiplier theorem which will be discussed elsewhere, starting from (10.30).

10.7 Convergence of Newton's method

Consider Newton's method without the x-dependence in $F(x,y)$, replace y by x and F by f: so $f(x) = 0$ is the equation to be solved.

For $f : \mathbb{R} \to \mathbb{R}$ differentiable on

$$\bar{B}_\varepsilon = \{x \in \mathbb{R} : \ |x| \leq \varepsilon\}, \quad \varepsilon_0 > 0,$$

with $x \to f'(x)$ Lipschitz continuous on $\bar{B}_{\varepsilon_0}$, and a sequence x_n in $\bar{B}_{\varepsilon_0}$ we write (9.9) as

$$f(x_n) = \underbrace{f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1})}_{\text{lineaire approximation}} + R(x_n; x_{n-1}), \tag{10.31}$$

in which

$$|R(x_n; x_{n-1})| \leq \frac{L}{2}|x_n - x_{n-1}|^2,$$

L the Lipschitz constant of f' on $\bar{B}_{\varepsilon_0}$. Assume for all $x \in \bar{B}_{\varepsilon_0}$ that

$$|(f'(x))^{-1}| \leq C,$$

$C > 0$ a positive constant.

Let

$$p_n = |x_n - x_{n-1}| \ \text{ and } \ q_n = |f(x_n)|, \tag{10.32}$$

take x_n defined by

$$x_n = x_{n-1} - (f'(x_{n-1}))^{-1} f(x_{n-1}) \quad (n \in \mathbb{N}), \tag{10.33}$$

with $x_0 = 0$. Then $x_n \in \bar{B}_{\varepsilon_0}$ as long as

$$p_1 + p_2 + \cdots + p_n \leq \varepsilon_0, \tag{10.34}$$

in which case it follows that

$$p_n \leq C q_{n-1} \ \text{ en } \ q_n \leq \frac{1}{2} L p_n^2, \tag{10.35}$$

because (10.33) puts the linear approximation de linear approximation in (10.31) equal to zero.

The inequalities in (10.35) are now used beginning with

$$q_0 = |f(0)| \quad \text{en} \quad p_1 \leq C q_0 = C|f(0)|, \tag{10.36}$$

combining (10.35) and (10.36) as

$$p_n \leq \mu p_n^2 \quad \text{met} \quad \mu = \frac{1}{2} LC \quad \text{and} \quad p_1 \leq C|f(0)|. \tag{10.37}$$

The question is for which $\bar{P} = \bar{P}(\mu)$ we can conclude that de implication

$$C|f(0)| \leq \bar{P} \implies \sum_{n=1}^{\infty} p_n \leq \varepsilon_0 \tag{10.38}$$

holds.

The larger $\bar{P}$, the stronger the statement in the sense that larger values of $|f(0)|$ allowed to find a solution $x \in B_{\varepsilon_0}$ of $f(x) = 0$ via (10.33) starting from $x_0 = 0$. The largest $\bar{P}$ is found by taking equalities in (10.37) and (10.38). This gives

$$p_n = \mu p_{n-1}^2 \ \text{ voor } \ n \in \mathbb{N}; \quad p_1 = \bar{P}\,; \quad \sum_{n=1}^{\infty} p_n = \varepsilon_0. \tag{10.39}$$

Via $\xi_n = \mu p_n$ en $\xi_n = \xi_{n-1}^2$ this is equivalent to

$$\mu\varepsilon_0 = G(\mu\bar{P}) \quad \text{with} \quad G(\xi) = \xi + \xi^2 + \xi^4 + \xi^8 + \xi^{16} + \cdots, \tag{10.40}$$

but no simple formula $\bar{P} = \bar{P}(\mu, \varepsilon_0)$ is obtained.

Exercise 10.9. Use

$$G(\xi) \leq \frac{\xi}{1-\xi}$$

to formulate a simpler condition $|f(0)| \leq \dots$ in terms of C and L, which guarantees $x_n \to x \in \bar{B}_{\varepsilon_0}$ with $f(x) = 0$.

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 10.10. Put $y = 1 + z$ and see what you get for the sequence z_n to understand why the convergence is so fast.

Exercise 10.11. Put $e = x^2 - 2$, rewrite (2.1) in terms of e and $\tilde{e}$, examine the sequence e_n, and compare to Exercise 10.10.

11 Quadratic functions and Morse's Lemma

Let X be a Banach space. By a quadratic function Q on X we mean a function of the form[1]

$$X \ni x \xrightarrow{Q} (Sx)(x) = \langle Sx, x\rangle \in \mathbb{R} \tag{11.1}$$

in which $S \in L(X, X^*)$ is a continuous linear map

$$X \ni x \xrightarrow{S} S(x) = Sx \in X^*$$

from X to X^*. Recall that

$$\langle \phi, x\rangle = \phi(x) \quad \text{for} \quad \phi \in X^* \quad \text{and} \quad x \in X.$$

Exercise 11.1. 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 11.7. 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 11.2. 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 (11.1)? If so we say that F and Q are conjugate functions.

11.1 Intermezzo: second order partial derivatives

Theorem 11.3. *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

$$g_{yx}(x_0, y_0) = \frac{\partial}{\partial x}\frac{\partial g}{\partial y} = \frac{\partial}{\partial y}\frac{\partial g}{\partial x} = g_{xy}(x_0, y_0).$$

[1] See Remark 6.7 and Theorem 6.11 for notation.

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 8.12 applied to

$$y \to g(x, y) - g(0, y)$$

for $x \neq 0$ and $y \neq 0$ small imply that for some x-dependent η between 0 and y we have

$$g(x, y) - g(0, y) - g(x, 0) + g(0, 0) = (g_y(x, \eta) - g_y(0, \eta))y$$

$$= (g_y(0, 0) + g_{yx}(0, 0)x + g_{yy}(0, 0)\eta + R(x, \eta) - g_y(0, 0) - g_{yy}(0, 0)\eta - R(0, \eta))y$$

$$= (g_{yx}(0, 0)x + R(x, \eta) - R(0, \eta))y,$$

in which

$$R(x, \eta) = o(\sqrt{x^2 + \eta^2}) \quad \text{and so also} \quad R(0, \eta) = o(\eta) \quad \text{as} \quad \sqrt{x^2 + \eta^2} \to 0.$$

The differentiability of

$$(x, y) \to g_y(x, y)$$

in $(0, 0)$ has been used twice, with the same "remainder function" R. Since $|\eta| \leq |y|$ it follows that

$$g(x, y) - g(0, y) - g(x, 0) + g(0, 0) = g_{yx}(0, 0)xy + y\, o(r)$$

$$= g_{yx}(0, 0)xy + o(r^2) = g_{xy}(0, 0)xy + o(r^2) \tag{11.2}$$

for $r = \sqrt{x^2 + y^2} \to 0$. The second version under (11.2) follows by interchanging the roles of x and y and implies $g_{yx}(0, 0) = g_{xy}(0, 0)$.

11.2 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,

$$f'(t) = F'(tx)(x) = \langle F'(tx), x\rangle, \tag{11.3}$$

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

$$F(x) = f(1) = \int_0^1 (1 - t)f''(t)\, dt. \tag{11.4}$$

The differentiability of $t \to F'(tx)x = f'(t)$ will follow from differentiability of

$$x \to F'(x) \in X^*$$

in points ξ near 0, which means that

$$F'(x) = F'(\xi) + F''(\xi)(x - \xi) + R(x;\xi), \tag{11.5}$$

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 (11.5) this becomes

$$F'(tx) = F'(t_0 x) + F''(t_0 x)(tx - t_0 x) + R(tx; t_0 x)$$

$$= F'(t_0 x) + (t - t_0)F''(t_0 x)x + R(tx; t_0 x)$$

in X^*, and (11.3) then gives

$$f'(t) = \langle F'(tx), x\rangle = \underbrace{\langle F'(t_0 x, x\rangle}_{f'(t_0)} + (t - t_0)\langle F''(t_0 x)x), x\rangle + \langle R(tx; t_0 x), x\rangle.$$

We conclude that f' is differentiable in every $t \in [0, 1])$ for which F' is differentiable in tx, with

$$f''(t) = \langle F''(tx)x, x\rangle. \tag{11.6}$$

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 $\mathcal{O}$.

11.3 The second derivative as symmetric bilinear form

Theorem 11.4. *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*

$$(h, k) \xrightarrow{F''(\xi)} (F''(\xi)h)k = \langle F''(\xi)h, k\rangle \in \mathbb{R} \tag{11.7}$$

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 11.4 is proved by Exercise 11.5 and Theorem 11.3.

Exercise 11.5. 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

$$S \in S(X, X^*) \subset L(X, X^*). \tag{11.8}$$

With also $F''(tx) \in S(X, X^*)$ we have from (11.4) 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

$$F(x) = \langle \int_0^1 (1-t)F''(tx)\, dt\, x, x\rangle = \langle \Phi_x x, x\rangle, \tag{11.9}$$

in which

$$\Phi_x = \int_0^1 (1-t)F''(tx)\, dt \in S(X, X^*). \tag{11.10}$$

Here we use a subscript to denote the x-dependence of the operator Φ_x which acts on X.

It follows that

$$F(x) = \frac{1}{2}\langle F''(0)x, x\rangle + \langle \int_0^1 (1-t)(F''(tx) - F''(0))\, dt\, x, x\rangle$$

$$= \langle \Phi_0 x, x\rangle + o(|x|_X^2), \tag{11.11}$$

as $|x|_X \to 0$ if F'' is continuous in $x = 0$. The quadratic function defined by

$$Q_0(x,x) = \langle \Phi_0 x, x\rangle = \frac{1}{2}\langle F''(0)x, x\rangle \tag{11.12}$$

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 11.6. Check that ontinuity of F'' in 0 means that for every $\varepsilon > 0$ a $\delta > 0$ exists such that

$$0 < |x|_X < \delta \implies |(F''(x) - F''(0))y|_{X^*} < \varepsilon |y|_X$$

for all $0 \neq y \in X$.

Exercise 11.7. 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 11.8. 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 $|x| < r$.

11.4 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

$$\langle \Phi_x x, x\rangle = (\Phi_x x)\, x = (\Phi_0 y)\, y = \langle \Phi_0 y, y\rangle \tag{11.13}$$

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 (11.13) holds if the map

$$h \to \Phi h \quad \text{is equal to the map} \quad h \to \Phi_0 T h \circ T = \kappa_0(T,T)\, h. \tag{11.14}$$

This is an $L(X, X^*)$-valued "quadratic" equation for $T \in L(X,X)$.

Abstractly we may write (11.14) as

$$\kappa_0(T,T) = \Phi, \tag{11.15}$$

in which

$$X \times X \xrightarrow{\kappa_0} L(X, X^*)$$

is the bilinear form defined by

$$h \to \kappa_0(T,U)\, h = \Phi_0 T h \circ U.$$

Clearly $T = I$ is a solution of (11.14) when $\Phi = \Phi_0$. We want a solution $T = T_x$ for $\Phi = \Phi_x$ given by (11.10) close to Φ_0. If you like you can skip Section 11.5 and jump to (11.26), or even Exercise 11.12. Just put $T = I + H$ in (11.14) and see what you can get[2].

11.5 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 (11.15) rewrites as

$$\underbrace{\Phi_0 H h + \Phi_0 h \circ H + \Phi_0 H h \circ H}_{\chi_0(H)h} = (\Phi_x - \Phi_0)h \tag{11.16}$$

for all $h \in X$. The left hand side defines an X^*-valued function

$$H \xrightarrow{\chi_0} \chi_0(H)$$

[2] But that's not how I found equation (11.27).

quadratic in H, with Φ_0 in the "coefficients" of the two linear terms and one quadratic term. Writing (11.16) as

$$\chi_0(H) = \Phi_x - \Phi_0, \tag{11.17}$$

the right hand side is in $S(X, X^*)$.

Look at (11.16). Clearly the derivative of χ_0 in $H = 0$ is given by

$$h \xrightarrow{\chi_0'(I)H} \Phi_0 Hh + \Phi_0 h \circ H.$$

Since $\chi_0'(0)H \in L(X, X^*)$ is characterised by

$$\langle \chi_0'(0)Hh, k\rangle = \langle \Phi_0 Hh, k\rangle + \langle \Phi_0 Hk, h\rangle, \tag{11.18}$$

we have that $\chi_0'(0)H \in S(X, X^*)$. Thus the invertibility condition cannot be that

$$\forall_{h\in X} : \chi_0'(I)H = \Phi_0 Hh + \Phi_0 h \circ H = Ch \tag{11.19}$$

is solvable for every $C \in L(X, X^*)$, while (11.19) is underdetermined for $C \in S(X, X^*)$.

A handy[3] extra condition on H is that $\Phi_0 H \in S(X, X^*)$. Then (11.18) reduces to

$$\langle \chi_0'(0)Hh, k\rangle = 2\langle \Phi_0 Hh, k\rangle, \tag{11.20}$$

and the invertibility condition (11.19) becomes

$$2\Phi_0 H = C, \tag{11.21}$$

which is solvable for H as

$$H = \frac{1}{2}\Phi_0^{-1} C \tag{11.22}$$

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[4]

$$2Kh + Kh \circ \Phi_0^{-1} K) = (\Phi_x - \Phi_0)h \tag{11.23}$$

for all $h \in X$, which amounts to the equation

$$2K + T_0(K) = C_x = \Phi_x - \Phi_0 \tag{11.24}$$

[3] As it turns out is how Duistermaat and Kolk put it.

[4] Equation (11.23) follows directly from (11.16).

for $K \in S(X, X^*)$, in which the quadratic term is given by

$$T_0 : S(X, X^*) \to S(X, X^*), \quad T_0(K)h = Kh \circ (\Phi_0^{-1} K) \tag{11.25}$$

for all $h \in X$, and

$$X \ni x \to C_x \in S(X, X^*)$$

is continuous in $x = 0$ with $C_0 = 0$.

11.6 Yes, but main result via power series instead

Theorem 11.9. *Let X be a Banach 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 11.10. Prove Theorem 11.9 by applying the implicit function theorem to (11.23).

Remark 11.11. *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 space in disguise because

$$x \to \sqrt{\langle F''(0)(x), x \rangle}$$

then defines an equivalent 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 11.7.

In fact there's a direct way to solve (11.15) in the space

$$\{T \in L(X, X^*) : \Phi_0 T \in S(X, X^*)\}. \tag{11.26}$$

Via $T = I + H$ and (11.16) equation (11.15) was equivalent to (11.23) for

$$K = \Phi_0 H \in S(X, X^*).$$

We now return to an equation for H. Write (11.23) as

$$2Kh + Kh \circ (H) = (\Phi_x - \Phi_0)h$$

and apply it to $k \in X$. Then

$$\langle 2Kh, k\rangle + \underbrace{\langle Kh \circ (H), k\rangle}_{\langle Kh, Hk\rangle = \langle KHk, h\rangle} = \langle (\Phi_x - \Phi_0)h, k\rangle$$

for all $h, k \in X$. The first and the third term are symmetric in h and k. It follows that

$$2K + KH = \Phi_x - \Phi_0,$$

and applying Φ_0^{-1}, the equation to solve for H, still under the assumption that $\Phi_0 H \in S(X, X^*)$, is

$$2H + HH = \Phi_0^{-1}\Phi - I = P, \tag{11.27}$$

in which $P \in L(X, X^*)$ also has $\Phi_0 P \in S(X, X^*)$.

Exercise 11.12. Derive (11.27) directly from (11.15), 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

$$I + 2H + H^2 = I + P \quad \text{in} \quad L(X) = L(X, X). \tag{11.28}$$

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 8.23. The same power series tricks[5] give

$$T = I + H = I + \frac{1}{2}\,P - \frac{1}{2!}\,\frac{1}{2}\,\frac{1}{2}\,P^2 + \frac{1}{3!}\,\frac{1}{2}\,\frac{1}{2}\,\frac{3}{2}\,P^3 - \frac{1}{4!}\,\frac{1}{2}\,\frac{1}{2}\,\frac{3}{2}\,\frac{5}{2}\,P^4 + \cdots \quad (11.29)$$

if $|P| < 1$, and so $y = T_x x$ with

$$T_x = I + E_x - \frac{1}{2!}\,E_x^2 + \frac{1 \cdot 3}{3!}\,E_x^3 - \frac{1 \cdot 3 \cdot 5}{4!}\,E_x^4 + \cdots \quad (11.30)$$

and

$$E_x = \int_0^1 (1-t)(F''(0)^{-1}F''(tx) - I)\,dt, \quad (11.31)$$

which allows a more general setting[6]. 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 $L(X)$, which could be $\mathbb{R}$ itself via Exercise 6.13 in case $X = \mathbb{R}$ and I is just the number 1.

Exercise 11.13. See if you can give a direct derivation of (11.30) and (11.31) 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?

11.7 Bilinear forms and the Lax-Milgram theorem

Section 11.6 is not restricted to the case that X is a Hilbert space in disguise[7]. In particular (11.31) does not require a Hilbert spaces setting. In this section we do require a Hilbert space setting.

Theorem 11.14. *Let H be a Hilbert space and $B : H \times H \to \mathbb{R}$ be a bounded coercive bilinear form, meaning that*

(a) for every $u \in H$ fixed $v \to B(u,v)$ is linear;
(b) for every $v \in H$ fixed $u \to B(u,v)$ is linear;
(c) $\exists_{\alpha \geq 0}\, \forall_{u,v \in H} : |B(u,v)| \leq \alpha\,|u|\,|v|$.
(d) $\exists_{\beta > 0}\, \forall_{u \in H} : B(u,u) \geq \beta\,|u|^2$.

[5] Copy/paste what you know by now for the case that $P, H \in \mathbb{R}$.
[6] Think of examples in which $F''(0)$ is not invertible in $L(X)$.
[7] A Banach space which allows an equivalent inner product norm.

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

$$\langle \phi, v \rangle = \langle Au, v \rangle = B(u, v) \quad \text{for all} \quad v \in H, \tag{11.32}$$

which has $|A| \leq \alpha$.

For the proof we observe that (11.32) and assumption (c) imply that

$$|\langle Au, v \rangle| = |B(u, v)| \leq \alpha \, |u| \, |v|$$

for all u and v in H, and that for u fixed assumption (a) says that

$$Au : H \to \mathbb{R}$$

is linear. It follows that $Au \in H^*$ and

$$|Au| \leq \alpha \, |u|.$$

Assumption (b) implies that the map

$$A : H \to H^*$$

is linear, and assumption (d) gives

$$\beta |u|^2 \leq B(u, u) = \langle Au, u \rangle \leq |Au| \, |u|$$

for all $u \in H$, whence

$$|Au| \geq \beta \, |u|.$$

We conclude that

$$H \xrightarrow{A} R(A) = \{Au : \, u \in H\}$$

is a linear bijection, continuous in both directions, because

$$\beta |u| \leq |Au| \leq \alpha |u| \tag{11.33}$$

for all $u \in H$. Thus $R(A)$ is complete because H is. In particular $R(A)$ is closed in H^*. It remains to show that $R(A) = H^*$.

Now let Φ be as in Theorem 6.31 and $L = \Phi^{-1}(R(A) \subset H$. If $L \neq H$ then

$$M = \{v \in H : v \cdot w = 0 \text{ for all } w \in L\} \neq \{0\}.$$

Choose $v \in M$ with $v \neq 0$. Then

$$\langle \Phi(w), v \rangle = w \cdot v = 0$$

for all $w \in R(A) = \{Au : u \in H\}$, whence $\langle Av, v \rangle = 0$, a contradiction with assumption (d). Thus $L = H$, whence $R(A) = H^*$. This completes the proof of Theorem 11.14.

If we start from the Banach 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 11.15. *Let X be a normed space. A map $(u, v) \to B(u, v)$ from $X \times X$ to $\mathbb{R}$ is called a bounded bilinear form if*

(a) for every $u \in X$ fixed $v \xrightarrow{\phi} B(u, v)$ is linear;
(b) for every $v \in X$ fixed $u \xrightarrow{\psi} B(u, v)$ is linear;
(c) $\exists_{\alpha \geq 0} \forall_{u,v \in X} : |B(u, v)| \leq \alpha |u| |v|$.

If in addition

$$\exists_{\beta > 0} \forall_{u \in X} : B(u, u) \geq \beta |u|^2,$$

then B is called coercive.

Remark 11.16. *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 11.17 below.

Theorem 11.17. *Every bounded bilinear form on a normed space X is of the form*

$$(u, v) \to B(u, v) = \langle Au, v\rangle \in \mathbb{R} \tag{11.34}$$

with $A \in L(X, X^*)$, and[8]

$$\sup_{u,v\in X\setminus\{0\}} \frac{|B(u, v)|}{|u|\,|v|} = |A|. \tag{11.35}$$

If X is complete and B is coercive then X is a Hilbert space in disguise, and A is a bijection[9] *between X and X^* with*

$$\beta\,|u| \le |Au| \le \alpha|u|$$

for all $u \in X$, $0 < \beta \le \alpha$, as in Definition 11.15.

For the proof we use (a) again to define A by $Au = \phi$, so (11.34) holds by definition. In particular Au is a linear functional on X for every $u \in X$. By (c) we have

$$|\langle Au, v\rangle| = |B(u, v)| \le \alpha\,|u|\,|v|$$

for all $v \in X$ whence $Au \in X^*$ with

$$|Au| \le \alpha\,|u|, \tag{11.36}$$

and (b) implies that $A : X \to X^*$ is linear. Thus $A \in L(X, X^*)$ with $|A| \le \alpha$.

Exercise 11.18. Prove (11.35) by showing that

$$\sup_{u,v\in X\setminus\{0\}} \frac{|\langle Au, v\rangle|}{|u|\,|v|} = |A|.$$

Hint: choose u with $|u| = 1$ and $|Au|$ close to $|A|$, and then v with $|v| = 1$ and $|\langle Au, v\rangle|$ close to $|Au|$.

Finally assume that X is complete and B is coercive. Then

$$\beta\,|u|^2 \le B(u, u) = \langle Au, u\rangle \le |\langle Au, u\rangle| \le |Au|,$$

whence (11.33) holds and

$$X \xrightarrow{A} R(A) = \{Au :\ u \in X\}$$

[8] The norms have subscripts that we omit in this section.

[9] Lax-Milgram: $\forall_{\phi\in X^*}\,\exists_{u\in X}\,\forall_{v\in X} :\ B(u, v) = \phi(v) = \langle\phi, v\rangle$, u is unique for ϕ.

is a linear bijection, continuous in both directions. Thus $R(A)$ is a Banach 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[10] that closed subspaces $M \subset X^*$ with $M \neq X^*$ have $^0M \neq \{0\}$. Hilbert spaces (complete inner product spaces) have this property[11], and thus so does X. This completes the proof of Theorem 11.17.

11.8 The method of Lagrange

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 (10.30) 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 12.6, for now we write it as (12.25), 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

$$\Phi' = \Lambda \circ F'. \tag{11.37}$$

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 11.7. Details to follow.

[10] This property holds for reflexive spaces.

[11] See Section 6.4 and also the proof of Theorem 11.14.

12 Analysis unpacked: more variables

In this chapter we are concerned with differential and integral calculus for functions from X to Y in which X and Y are Euclidean spaces. We begin with $X = Y = \mathbb{R}^2$, with (rectangular) coordinates $x, y \in \mathbb{R}$ for $X = \mathbb{R}^2$ and coordinates $u, v \in \mathbb{R}$ for $Y = \mathbb{R}^2$. Later we shall perhaps prefer $x_1, x_2 \in \mathbb{R}$ for $x = (x_1, x_2) \in X = \mathbb{R}^2$ and $y_1, y_2 \in \mathbb{R}$ for $y = (y_1, y_2) \in Y = \mathbb{R}^2$.

We frequently use polar coordinates r, θ and the transformation

$$x = r\cos\theta;$$

$$y = r\sin\theta,$$

to describe points $(x, y) \neq (0, 0)$ in the plane via their distance $r = \sqrt{x^2 + y^2}$ to the origin $(0, 0)$ and the angle θ between the halfline

$$\{(tx, ty) : t \geq 0\}$$

and the positive x-axis. Whenever convenient we identify $\mathbb{R}^2$ with the set $\mathbb{C}$ of *complex numbers*

$$z = x + iy,$$

and call $|z| = r$ the *absolute value* of z, the distance from z to the origin $z = 0$. The angle $\theta = \arg z$ is called the *argument* of z, uniquely determined modulo 2π for every $z \neq 0$.

Next to complex addition

$$w + z = (u + iv) + (x + iy) = u + x + i(v + y) = (u + x, v + y) = (u, v) + (x, y)$$

we also have complex mulitplication

$$wz = (u+iv)(x+iy) = ux-vy+i(uy+vx) = (ux-vy, uy+vx) = (u, v)(x, y),$$

based on the rule $i^2 = -1$, for $w = u + iv = (u, v)$ and $z = x + iy = (x, y) \in \mathbb{R}^2 = \mathbb{C}$. The rules for addition and multiplication in $\mathbb{C}$ are the same as the rules for addition and multiplication in $\mathbb{R}$. We also have

$$|w + z| \leq |w| + |z| \quad \text{and} \quad |wz| = |w|\,|z|.$$

Very important is the rule formulated in this exercise.

Exercise 12.1. The summation rules for cos and sin imply that

$$z_1 z_2 = r_1 r_2(\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)) \quad \text{for} \quad z_j = r_j(\cos\theta_j + i\sin\theta_j),\ j = 1, 2.$$

This rule is one many reasons to write

$$\cos\theta + i\sin\theta = \exp(i\theta) \quad \text{and} \quad \exp(z) = \exp(x)\exp(iy)$$

We note that polar coordinates are not needed to prove that for every nonzero γ the map

$$z \to \gamma z \tag{12.1}$$

is a rotation[1] around 0 followed by a point multiplication with 0 as fixed point, see (12.8) and Exercise 12.5.

12.1 Intermezzo: algebra's main theorem

The set $\mathbb{C}$ is algebraically closed: every polynomial

$$P(z) = \sum_{k=0}^{n-1} \alpha_k z^k + z^n \tag{12.2}$$

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

$$P(z) = (z - z_1)\cdots(z - z_n) \quad \text{with} \quad z_1, \ldots, z_n \in \mathbb{C}. \tag{12.3}$$

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

$$Q(w) = 1 + \sum_{k=1}^{n} \gamma_k w^k \tag{12.4}$$

with $\gamma_1, \gamma_2, \ldots, \gamma_n \in \mathbb{C}$, $\gamma_n \neq 0$, 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 12.1 we have

$$Q(w) = 1 + \sum_{k=1}^{n} c_k r^k \exp(i(\phi_k + k\theta)), \tag{12.5}$$

an expression[2] in which the ϕ_k are parameters and $r > 0$ can be taken as small as we want. Exercise 12.2 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 (12.2).

Exercise 12.2. 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 (12.6) below.

So why would f have a global minimum? Observe that f is continuous[3]. By Theorem 4.51 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.

[2] Ptolemaeus would have liked this.

[3] Prove this using Definition 3.34 with $X = D_r$.

Indeed, with $z = x + iy$ and $x^2 + y^2 = r^2$ we have for $|P(z)| = f(x, y)$ that

$$|P(z)| = |\sum_{k=0}^{n-1} \alpha_k z^k + z^n| \geq |z^n| - |\sum_{k=0}^{n-1} \alpha_k z^k| \geq r^n - \sum_{k=0}^{n-1} |\alpha_k| r^k. \qquad (12.6)$$

On the circle defined by $x^2 + y^2 = r^2$ it then follows that

$$f(x, y) \geq r^{n-1}(r - \underbrace{\sum_{k=0}^{n-1} |\alpha_k|}_{r_0}) = \underline{M}_r,$$

a lower bound which is positive for r larger than

$$r_0 = \sum_{k=0}^{n-1} |\alpha_k|.$$

For $r = r_0$ we have $\underline{M}_{r_0} = 0 < m_{r_0}$. Clearly $\underline{M}_r$ increases to ∞ as r increases from r_0 to ∞. Thus for some $r_1 > r_0$ we have

$$\underline{M}_{r_1} > m_{r_0} \geq m_{r_1},$$

and then also

$$f(x, y) > m_{r_1} \quad \text{for all} \quad (x, y) \notin D_{r_1}.$$

It follows that m_{r_1} is the global minimum of f on the whole of $\mathbb{R}^2$ and the contradiction arises as explained above. This completes this truly remarkable proof in which elegant algebra, basically algebraic estimates, and rock solid analysis combine.

12.2 Complex and multivariate differential calculus

In Section 7.7 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 7.11 and its proof carry over by copy-paste to complex valued powerseries with complex coefficients and variables. Also, differentiability via (8.1) or (9.2) becomes complex differentiability for functions[4]

$$H : \mathbb{C} \to \mathbb{C},$$

but now via

$$\begin{aligned} w = H(z) &= H(z_0) + \gamma(z - z_0) + T(z; z_0) \\ &= H(z_0) + H'(z_0)(z - z_0) + o(|z - z_0|) \end{aligned} \tag{12.7}$$

as $z \to z_0$.

If we unpack (12.7), 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 (12.7) unpacks as

$$\gamma(z - z_0) = (\alpha + i\beta)(h + ik) = \alpha h - \beta k + i(\beta h + \alpha k).$$

This corresponds to

$$\begin{pmatrix} \alpha h - \beta k \\ \beta h + \alpha k \end{pmatrix} = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} h \\ k \end{pmatrix}, \tag{12.8}$$

in which the matrix describes the map (12.1).

The complex expansion (12.7) rewrites as

$$u = F(x, y) = F(x_0, y_0) + a(x - x_0) + b(y - y_0) + R(x, y; x_0, y_0);$$

$$v = G(x, y) = G(x_0, y_0) + c(x - x_0) + d(y - y_0) + S(x, y; x_0, y_0),$$

with remainder terms R and S defined via $T = R + iS$, and a special form of the 2×2 matrix A in the linear expansion around (x_0, y_0), namely

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = A = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}.$$

[4] For convenience we assume H is globally defined.

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 12.3. *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 12.4. Prove Theorem 12.3.

Exercise 12.5. Examine (12.1) using (12.8).

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

$$Ah = \begin{pmatrix} (Ah)_1 \\ (Ah)_2 \end{pmatrix} = \begin{pmatrix} A_{11}h_1 + A_{12}h_2 \\ A_{21}h_1 + A_{22}h_2 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \tag{12.9}$$

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[5] to on the one hand the notation in (8.4) with A_0 acting on h and the norm of A_0 in $L(X, Y)$, and on the other hand (8.5) with A_0 algebraically multiplying h. In the latter context we can estimate

$$|(Ah)_1| = |A_{11}h_1 + A_{12}h_2| \leq \sqrt{A_{11}^2 + A_{12}^2}\,\sqrt{h_1^2 + h_2^2};$$

$$|(Ah)_2| = |A_{21}h_1 + A_{22}h_2| \leq \sqrt{A_{21}^2 + A_{22}^2}\,\sqrt{h_1^2 + h_2^2},$$

[5] We dropped the zero-subcripts.

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[6]

$$|Ah|_2 \leq |A|_2 \, |h|_2 . \tag{12.10}$$

In (12.10) the "Euclidean" lengths of $h = x - a$, Ah and A appear[7], in each case the square root of the sum of the squared entries. You may well prefer here to forget[8] all about the norm of

$$h \xrightarrow{A} Ah$$

in $L(\mathbb{R}^2, \mathbb{R}^2)$: going back to

$$F(x) = F(a) + A(x-a) + R(x;a) \quad \text{and} \quad |R(x;a)|_2 = o(|x-a|_2) \tag{12.11}$$

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

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(a_1, a_2) & \frac{\partial F_1}{\partial x_2}(a_1, a_2) \\ \frac{\partial F_2}{\partial x_1}(a_1, a_2) & \frac{\partial F_2}{\partial x_2}(a_1, a_2) \end{pmatrix} = F'(a) = DF(a) \tag{12.12}$$

in every point $x = (x_1, x_2) = (a_1, a_2) = a$ where F is differentiable.

We often identify the linear map[9] $F'(a) = DF(a)$ 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 10.5 for $F : \mathbb{R}^2 \to \mathbb{R}$ and $F : X \times Y \to \mathbb{R}$.

Exercise 12.6. State and prove a theorem for $F : \mathbb{R}^2 \to \mathbb{R}$ by specializing Theorem 10.6 to $X = Y = \mathbb{R}$ and generalise to $F : \mathbb{R}^m \to \mathbb{R}$ and $F : \mathbb{R}^m \to \mathbb{R}^n$.

---

[6] This generalises, see (12.32).

[7] Actually this 2-norm of A is called the Frobenius norm of A.

[8] If not note that (12.10) says that this operator norm of A is atmost equal to $|A|_2$.

[9] Both notations are widely used.

12.3 Cauchy-Riemann equations, harmonic functions

Have another look at Theorem 12.3 and let H be complex differentiable[10] 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

$$H'(z_0) = \alpha + i\beta.$$

Exercise 12.7. Show that α and β are then given by

$$\alpha = \frac{\partial u}{\partial x} \quad \text{and} \quad \beta = -\frac{\partial u}{\partial y}, \tag{12.13}$$

evaluated in $(x, y) = (x_0, y_0)$.

Thus Theorem 12.3 says that u and v, as functions of x and y, must satisfy the so-called Cauchy-Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y} \tag{12.14}$$

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

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = \Delta v, \tag{12.15}$$

in which the differential operator Δ, the *Laplacian*, occurs. This Δ is a feast to study, but not now. Here we want to be sure under what conditions (12.15) makes sense. We copy Theorem 11.3 from Section 11.1.

[10] We now prefer a notation with (x, y) and (x_0, y_0).

Theorem 12.8. *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 (12.15) 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 $n \in \mathbb{N}$.

Refering to Section 11.3, 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

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \tag{12.16}$$

without any v interfering[11].

There are many non-constant solutions of (12.16). Indeed, you should have noticed

$$x, y, x^2 - y^2, 2xy, x^3 - 3xy^2, 3x^2y - y^3, x^4 - 6x^2y^2 + y^4, 4x^3y - 4xy^3, \ldots \tag{12.17}$$

above.

Exercise 12.9. 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[12] functions

$$w \to \log w = \ln |w| + i \arg w.$$

[11] Not a priori.

[12] Which are thereby not functions.

12.4 Monomials and power series again

This should speak for itself. Recalling (7.26) we have with

$$H = \frac{|x-a|}{r}$$

that

$$x^m = a^m + ma^{m-1}(x-a) + R_{a,m}(x), \quad |R_{a,m}(x)| \le \frac{m(m-1)r^m}{2}H^2.$$

Likewise for $|y|, |b| \le s$, we have

$$y^n = b^n + nb^{n-1}(y-b) + R_{b,n}(y), \quad |R_{b,n}(y)| \le \frac{n(n-1)s^m}{2}K^2, \quad K = \frac{|y-b|}{s}.$$

Multiplication then gives[13]

$$x^m y^n = a^m b^n + \underbrace{ma^{m-1}b^n(x-a) + na^m b^{n-1}(y-b)}_{linear\ part} + \underbrace{R_2 + R_{21} + R_{12} + R_{2,2}}_{R_{a,b,m,n}(x,y)},$$

in which we identify

$$R_2 = b^n R_{a,m}(x) + mna^{m-1}b^{n-1}(x-a)(y-b) + a^m R_{b,n}(y),$$

$$R_{21} = ma^{m-1}(x-a)R_{b,n}(y),$$
$$R_{12} = nb^{n-1}R_{a,m}(x)(y-b),$$
$$R_{2,2} = R_{a,m}(x)R_{b,n}(y).$$

With rough but obvious estimates

$$|R_{2,2}| \le \frac{1}{4}m^2n^2r^m s^n H^2 K^2,$$

$$|R_{21}| \le \frac{1}{2}m^2 n r^m s^n H^2 K \le \frac{1}{2}m^2n^2r^m s^n H^2 K,$$
$$|R_{12}| \le \frac{1}{2}mn^2 r^m s^n H K^2 \le \frac{1}{2}m^2n^2r^m s^n H K^2,$$

and also, a little less obvious maybe,

$$|R_2| \le \frac{1}{4}(m^2+n^2)r^m s^n (H^2+K^2),$$

[13] This is a bit like (8.12).

we conclude that

$$x^m y^n = a^m b^n + m a^{m-1} b^n (x-a) + n a^m b^{n-1} (y-b) + \underbrace{R_{a,b,m,n}(x,y)}_{R}, \quad (12.18)$$

in which

$$|R| \leq \frac{r^m s^n}{4} \left(m^2 n^2 HK(HK + 2H + 2K) + (m^2 + n^2)(H^2 + K^2) \right). \quad (12.19)$$

The perhaps less obvious estimate for R_2 follows via

$$|R_2 \leq |s^n R_{a,m}(x)| + |mnr^{m-1} s^{n-1} (x-a)(y-b)| + |r^m R_{b,n}(y)| \leq$$
$$= \frac{m(m-1) r^m s^n}{2} H^2 + mnr^m s^n HK + \frac{n(n-1) r^m s^n}{2} K^2 =$$
$$\frac{r^m s^n}{4} \begin{pmatrix} m(m-1) & mn \\ mn & n(n-1) \end{pmatrix} \begin{pmatrix} H \\ K \end{pmatrix} \cdot \begin{pmatrix} H \\ K \end{pmatrix},$$

and the 2-norm of the matrix in this expression being less than $m^2 + n^2$.

We now multiply (12.18) by coefficients α_{mn} and the estimates for $R_{2,21,12,22}$ in

$$R = R_{a,b,m,n}(x,y) = R_2 + R_{21} + R_{12} + R_{2,2}$$

by coefficients $|\alpha_{mn}|$, and take the sum over $m, n \in \mathbb{N}_0$. Clearly a sufficient condition to conclude that on the rectangle

$$R_{rs} = \{(x,y) \in \mathbb{R}^2 : \, |x| < r, |y| < s\}$$

the power series

$$P(x,y) = \sum_{m,n \in \mathbb{N}_0} \alpha_{mn} x^m y^n$$

exists as a differentiable function, with

$$P_x(x,y) = \sum_{m,n \in \mathbb{N}_0} m \alpha_{mn} x^{m-1} y^n \quad \text{and} \quad P_y(x,y) = \sum_{m,n \in \mathbb{N}_0} n \alpha_{mn} x^m y^{n-1},$$

is that the series

$$\sum_{m,n \in \mathbb{N}_0} (m^2 + n^2) \, |\alpha_{mn}| r^m s^n \quad \text{and} \quad \sum_{m,n \in \mathbb{N}_0} m^2 n^2 \, |\alpha_{mn}| r^m s^n \qquad (12.20)$$

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 (12.20).

Exercise 12.10. Fill in the details of the above proof. Show in addition that the convergence of

$$\sum_{m,n\in\mathbb{N}_0} (m^2+n^2)\,|\alpha_{mn}|R^{m+n} \quad \text{and} \quad \sum_{m,n\in\mathbb{N}_0} m^2n^2\,|\alpha_{mn}|R^{m+n} \tag{12.21}$$

suffices to have $P(x,y)$ exist as a differentiable function on the disk

$$\{(x,y)\in\mathbb{R}^2 :\, x^2+y^2<R\}.$$

12.5 Application: the Hopf bifurcation

We examine the system of differential equations

$$\frac{dx}{dt} = \mu x - y + p_2(x,y) + p_3(x,y) + \cdots = P(x,y);$$

$$\frac{dy}{dt} = x + \mu y + q_2(x,y) + q_3(x,y) + \cdots = Q(x,y),$$

for real valued function $x(t)$ and $y(t)$, in which the functions

$$p_n(x,y) = a_{n0}x^n + a_{n1}x^{n-1}y + \cdots + a_{0n}y^n$$

and

$$q_n(x,y) = b_{n0}x^n + b_{n1}x^{n-1}y + \cdots + b_{0n}y^n$$

have real coefficients for every

$$n\in\mathbb{N}_2 = \{n\in\mathbb{N} :\, n\geq 2\},$$

and $\mu\in\mathbb{R}$ is a parameter. We shall call this family of systems the μ-systems.

In the special case that all the coefficients are zero the μ-systems reduce to

$$\frac{dx}{dt} = \mu x - y;$$

$$\frac{dy}{dt} = x + \mu y.$$

The reduced μ-system has nontrivial periodic solutions[14] if and only if $\mu = 0$. The plane defined by $\mu = 0$ and the line defined by $x = y = 0$ in μxy-space

[14] Namely $x = \varepsilon\cos t, y = \varepsilon\sin t$, in which $\varepsilon > 0$ is not necessarily small.

together form the set of all bounded solution orbits of the reduced μ-systems. We wish show that near $x = y = 0$ this family of periodic orbits persists as we add the nonlinear terms. Under the basic assumption that the coefficients are bounded we will show that there exists a locally defined smooth function $f(x, y)$ with $f_x(0, 0) = f_y(0, 0) = 0$ such that the graph $\mu = f(x, y)$ describes all the periodic solutions of the full system. In particular every level set

$$\Gamma_\mu = \{(x, y) \in \mathbb{R}^2,\ f(x, y) = \mu\}$$

is a periodic orbit of the full μ-system.

Exercise 12.11. Assume that the coefficients a_{mn} and b_{mn} are bounded. Use Section 12.4 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

$$|a_{mn}| \le 1 \quad \text{and} \quad |b_{mn}| \le 1 \quad \text{for all} \quad m, n \in \mathbb{N} \quad \text{with} \quad m + n \ge 2, \tag{12.22}$$

and introduce polar coordinates $x = r\cos\theta, y = r\sin\theta$ to transform solutions of the μ-systems to solutions of

$$\frac{dr}{dt} = \mu r + \alpha_2(\theta)r^2 + \alpha_3(\theta)r^3 + \cdots ;$$
$$\frac{d\theta}{dt} = 1 + \beta_2(\theta)r + \beta_3(\theta)r^2 + \cdots .$$

Exercise 12.12. Use the chain rule[15] and Section 8.4 to determine the expressions for α_n and β_n expressed in terms of $c = \cos\theta, s = \sin\theta, p_n(c, s), q_n(c, s)$. Show that

$$|\alpha_n| \le n \quad \text{and} \quad |\beta_n| \le n \quad \text{for all} \quad n \in \mathbb{N}_2,$$

and denoting the r-dependent part of the right hand side of the θ-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 $0 < r < 2 - \sqrt{2}$.

[15] Figure out how to use only the version with $X = Y = Z = \mathbb{R}$ from Section 8.2.

Exercise 12.13. Use the chainrule and Section 8.4 again to show that, for

$$0 < r < 2 - \sqrt{2},$$

solutions can be seen as functions $r = r(\theta)$ of θ, and that

$$\frac{dr}{d\theta} = r_\theta = \mu r + A_3(\theta, \mu) r^2 + A_4(\theta, \mu) r^3 + A_5(\theta, \mu) r^4 + \cdots, \tag{12.23}$$

with $A_3, A_4, \ldots$ polynomials in $\cos\theta$ en $\sin\theta$ in which also μ appears. Hint:

$$\frac{1}{1-\rho} = 1 + \rho + \rho^2 + \rho^3 + \cdots = \sum_{n=0}^{\infty} \rho^n.$$

Exercise 12.14. Show directly from the differential equations for $r(t)$ and $\theta(t)$ that

$$\left|\frac{dr}{d\theta}\right| = \left|\frac{\frac{dr}{dt}}{\frac{d\theta}{dt}}\right| \leq \frac{r}{1-2r}(|\mu|(1-r^2) + r(2-r))$$

for $0 < r < \frac{1}{2}$.

Exercise 12.15. Show that

$$\int_0^{2\pi} A_3(\theta, \mu) d\theta = 0.$$

Exercise 12.16. Consider the truncated differential equation

$$r_\theta = \mu r + A_3(\theta, \mu) r^2$$

and do the Kepler trick: introduce $w = \frac{1}{r} > 0$ as a function of θ. Why can this equation have no 2π-periodic solutions? Hint: you should get an equation in which only $\frac{dw}{d\theta}$, w and A_3 appear. Integrate from 0 to 2π to derive a contradiction if $w(\theta)$ is a (positive) 2π-periodic solution.

Consider (12.23) with $r(0) = \varepsilon > 0$ as initial value. For the original μ-system this corresponds to the solution with $x(0) = \varepsilon, y(0) = 0$. Now scale r by setting $r = \varepsilon R$. Then (12.23) becomes

$$\frac{dR}{d\theta} = R_\theta = \mu R + \varepsilon A_3(\theta, \mu)R^2 + \varepsilon^2 A_4(\theta, \mu)R^3 + \varepsilon^3 A_5(\theta, \mu)R^4 + \cdots , \quad (12.24)$$

and we look for solutions with $R(0) = 1$. Note that the explicit estimate in Exercise 12.14 carries over. We have

$$|\frac{dR}{d\theta}| \leq \frac{R}{1 - 2\varepsilon R}(|\mu|(1 - \varepsilon^2 R^2) + \varepsilon R(2 - \varepsilon R))$$

for $0 < \varepsilon R < \frac{1}{2}$.

If this initial value problem has a solution $R(\theta; \mu, \varepsilon)$ for small μ and small ε, then we set

$$F(\mu, \varepsilon) = R(2\pi; \mu, \varepsilon) - 1$$

and examine the equation

$$F(\mu, \varepsilon) = 0.$$

Clearly we have $F(0, 0) = 0$. Can we apply Theorems 10.1 and 10.2? The answer is yes, via what we already started in Section 10.4.

12.6 Stationary under boundary conditions

This topic was started in Section 10.6 with the remarkable formula

$$\Phi_x = \Phi_y (F_y)^{-1} F_x \quad (12.25)$$

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 10.2 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

$$S = \{(x,y) :\ F(x,y) = F(0,0)\} \tag{12.26}$$

is described as the graph of an implicitly definined continuously differentiable function f.

With this f the level set S is locally parameterised by

$$x \to X(x) = (x, f(x)),$$

which has a 2×1 Jacobi matrix $\frac{\partial X}{\partial x}$. The parameterisation is locally a bijection bewteen S and a neighbourhood of $x = 0$, which is due to the invertability of the 1×1 matrix

$$A = F_y \tag{12.27}$$

in $(0,0)$. Differentiability of

$$(x,y) \to \Phi(x,y)$$

sufficed to have (12.25) as both necessary and sufficient for $\phi'(x) = 0$, not only in $x = 0$ but as long as $F_y(x, f(x))$ is invertertible on a whole neighbourhood of $x = 0$ in which $f(x)$ was constructed.

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 are the same as in Chapter 10, also for (12.25) as the characterisation for

$$x \xrightarrow{\phi} \Phi(x, f(x))$$

being stationary.

Let's see how all this unpacks to give the method of Lagrange mulitpliers when we read (12.25) as a statement for Jacobi matrices and the corresponding linear maps. If you are familiar with matrices you can jump[16] to (12.33), the transposed form of (12.25).

[16] But do read on for a perhaps different perspective.

12.7 Intermezzo: matrices and matrix norms

In general[17] an $m \times n$ real matrix A is a block with real entries a_{ij}. The vertical index i runs from 1 up to m, the horizontal index j from 1 up to n. Considered as linear map, 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 lineair maps from $\mathbb{R}^n$ to $\mathbb{R}^m$.

If B is a real $n \times p$ matrix with entries b_{jk}, the vertical index j running from 1 up to n, the horizontal index k from 1 up to p, then AB is by definition the $m \times p$ matrix with entries

$$\sum_{j=1}^{n} a_{ij} b_{jk}, \tag{12.28}$$

with the corresponding linear map[18]

$$A \circ B : \mathbb{R}^p \stackrel{B}{\to} \mathbb{R}^n \stackrel{A}{\to} \mathbb{R}^m.$$

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 (12.28) reads as 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 $AB = I$, the $n \times n$ matrix with all diagonal entries equal to 1, and all off-diagonal entries equal to 0. This matrix I 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*[19] is that the map $A \circ B$ being the same map as the map I_n is equivalent to $AB = I$ for the matrices involved. It makes 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. Statements that only concern square matrices, and solutions of $Ax = y$ with A a square matrix.

[17] With m and n in $\mathbb{N}$.

[18] A preceded by B.

[19] A proof should be given in one of the first hours of any course in Linear Algebra.

If a third $p \times r$ matrix C has entries c_{kl} then $(AB)C$ is the matrix with entries

$$\sum_{k=1}^{p}(\sum_{j=1}^{n} a_{ij}b_{jk})c_{kl} = \sum_{k=1}^{p}\sum_{j=1}^{n} a_{ij}b_{jk}c_{kl}, \tag{12.29}$$

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 use in Section 12.8 for (12.25).

The series

$$I + A + A^2 + A^3 + \cdots, \tag{12.30}$$

with A a square matrix, a 2×2 matrix as in (12.9) for instance, is important for the implicit function theorem with $F : \mathbb{R}^{n+m} \to \mathbb{R}^m$ in Section 13.1. You should also compare[20] this to (11.29), and estimates required to justify those manipulations. Estimates which may start with a 2×2 matrix as in (12.9) and an estimate $A : \mathbb{R}^2 \to \mathbb{R}^2$ of the form (12.10).

Indeed you easily check[21] that for every $n \times n$ matrix and every real n-vector h it is true that

$$|Ah|_2 \leq M|h|_2, \tag{12.31}$$

in which $M \geq 0$ is defined by

$$M^2 = \sum_{i,j=1}^{n} a_{ij}^2.$$

If you like this defines the Pythagoras length of A, notation

$$M = |A|_2,$$

but recall that the smallest M for which (12.31) holds is called the operator norm of A, notation $|A|_{op}$, in which we again see A as a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n$. This notion was introduced in Theorem 6.11, in which the norm of A is the largest possible ratio between the norm of Ah and the norm of h when

$$A \in L(\mathbb{R}^n, \mathbb{R}^n) = L(\mathbb{R}^n).$$

We noted that $L(\mathbb{R}^n)$ is not only a vector space over $\mathbb{R}$, but also a normed algebra, because also the product operation

$$(A, B) \to AB$$

[20] See also Section 17.5.

[21] Using proof by induction it it makes you feel happy.

behaves as it should with respect to the norm

$$A \to |A|_{op}.$$

Indeed, it is easy to check that 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 $\lambda \in \mathbb{R}$, we also have

$$|AB|_{op} \leq |A|_{op}\,|B|_{op}.$$

As a vector space $L(\mathbb{R}^n)$ is equal to[22] $\mathbb{R}^{n^2}$, with the standard Pythagoraen norm[23]

$$A \to |A|_2$$

given by

$$|A|_2^2 = \sum_{i,j=1}^{n} a_{ij}^2,$$

which also has the pleasant property that

$$|AB|_2 \leq |A|_2\,|B|_2 \tag{12.32}$$

holds[24], but since $|A|_{op} \leq |A|_2$ for all $A \in L(\mathbb{R}^n)$ we prefer to use the smaller norm $|\cdot|_{op}$ in Exercise 1.4.

.

Exercise 12.17. 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[25]: if not then on

$$\{A \in L(\mathbb{R}^n) : |A|_{op} = 1\}$$

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 $|A|_{op} = 1$.

[22] Entries in a block or in a column, what's the difference really?
[23] In the literature it is called the Frobenius norm.
[24] Verify this!
[25] Section 18.5: $|A|_{op}$ is the square root of the largest eigenvalue of $A^T A$.

Exercise 12.18. If $A \in L(\mathbb{R}^n)$ has $|A|_{op} < 1$ then it holds for the series in (12.30) 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 12.19. *It should by now be clear that the whole machinery of power series carries over to Banach algebra's.*

12.8 The Lagrange multiplier method

We write (12.25) in transposed form as

$$\nabla_x F \, (\nabla_y F)^{-1} \, \nabla \Phi_y = \nabla_x \Phi, \tag{12.33}$$

in which

$$\nabla_x F, \nabla_y F, \nabla_x \Phi, \nabla_y \Phi$$

are the transposes of the "partial" Jacobi matrices

$$\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y}, \frac{\partial \Phi}{\partial x}, \frac{\partial \Phi}{\partial y}$$

corresponding to F_x, F_y, Φ_x, Φ_y.

Unpacking[26] 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 3×3 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[27] to the base vectors e_1, e_2, e_3 in $\mathbb{R}^3$.

Now write $\nabla_y \Phi \in \mathbb{R}^3$ as linear combination[28]

$$\nabla_y \Phi = \lambda_1 \nabla_y F_1 + \lambda_2 \nabla_y F_2 + \lambda_3 \nabla_y F_3 \tag{12.34}$$

[26] It is really no more than that, check it!

[27] Since the column vectors of a matrix A are the images under A of the e's.

[28] This is possible in view of the invertibility condition imposed on F_y in $(0, 0, 0, 0, 0)$.

with $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$. It follows that $\nabla_x F \left(\nabla_y F\right)^{-1}$ in the left hand side of (12.33) acts on (12.34) as

$$\nabla_y \Phi \xrightarrow{(\nabla_y F)^{-1}} \lambda_1 e_1 + \lambda_2 e_2 + \lambda_3 e_3 \xrightarrow{\nabla_x F} \lambda_1 \nabla_x F_1 + \lambda_2 \nabla_x F_2 + \lambda_3 \nabla_x F_3 = \nabla_x \Phi$$

by (12.33) again. With (12.34) this combines as

$$\nabla \Phi = \lambda_1 \nabla F_1 + \lambda_2 \nabla F_2 + \lambda_3 \nabla F_3, \tag{12.35}$$

simply[29] because it holds for ∇_x and ∇_y separately! The stationarity of

$$\Phi : S \to \mathbb{R}$$

in $(0,0)$ is thus equivalent with the existence of multiplicators $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$ for which (12.35) holds in $(0,0,0,0,0)$.

12.9 Application: Hölder's inequality

In (12.10) we had

$$|Ah|_2 \leq |A|_2 \, |h|_2$$

as a special case of

$$|AB|_2 \leq |A|_2 \, |B|_2 .$$

With $A = a$ a row matrix with entries a_i and $B = b$ a column matrix with entries b_i, this is the Cauchy-Schwarz inequality

$$|\sum_{i=1}^{n} a_i b_i| \leq \left(\sum_{i=1}^{n} |a_i|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} |b_i|^2 \right)^{\frac{1}{2}} .$$

This inequality is proved[30] in every linear algebra course and then used to prove the triangle inequality for the Euclidean norm.

We now ask for which values of $p > 1$ and $q > 1$ we can also have that

$$|\sum_{i=1}^{n} a_i b_i| \leq |a|_p \, |b|_q , \tag{12.36}$$

if $|a|_p$ and $|b|_q$ are defined by

$$|a|_p^p = \sum_{i=1}^{n} |a_i|^p \quad \text{and} \quad |b|_q^q = \sum_{i=1}^{n} |b_i|^q . \tag{12.37}$$

Note that (12.36) is the Cauchy-Schwarz inequality of $p = q = 2$.

[29] No 3×3 matrix inverted here.

[30] We come back to the proof in Exercise 18.1!

Exercise 12.20. Since (12.36) 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

$$\phi(a_1, \ldots, a_n) = |a_1|^p + \cdots + |a_n|^p = 1;$$

$$\psi(b_1, \ldots, b_n) = |b_1|^q + \cdots + |b_n|^q = 1,$$

and max- and minimise

$$(a_1, \ldots, a_n, b_1, \ldots, b_n) \xrightarrow{F} a_1 b_1 + \cdots + a_n b_n.$$

Exercise 12.21. Explain why the maximum and the minimum of F under the restriction $|a|_p = |b|_q = 1$ exist.

Exercise 12.22. Show that the functions ϕ and ψ are continuously differentiable if $p > 1$ and $q > 1$. Hint: if we redefine $x \to x^r$ to be odd for every $r > 0$ then the derivative of $x \to |x|^p$ is $x \to px^{p-1}$.

With two Lagrange multipliers λ en μ we arrive at $2n$ equations

$$b_i = \lambda p a_i^{p-1}; \quad a_i = \mu q b_i^{q-1} \quad (i = 1, \ldots, n)$$

to solve, together with

$$\sum_{i=1}^{n} |a_i|^p = \sum_{i=1}^{n} |b_i|^q = 1.$$

Exercise 12.23. Assume that $(p-1)(q-1) \neq 1$. Show that solutions have all $|a_i|$ equal and all $|b_i|$ equal, and therefore

$$\sum_{i=1}^{n} |a_i b_i| = n(\frac{1}{n})^{\frac{1}{p}+\frac{1}{q}} = n^{1-\frac{1}{p}-\frac{1}{q}}. \tag{12.38}$$

Deduce that (12.36) holds for $p > 1$ and $q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$.

13 Varieties in Euclidean space

Think of lines and planes as nontrivial examples in $\mathbb{R}^3$ of linear varieties $\mathcal{M}$. Along $\mathcal{M}$ something varies, and the variations are linear: by definition linear varieties in $\mathbb{R}^N$ are solution sets of systems[1] of linear equations, which upon solving these systems are described as graphs of linear functions[2]. The typical example[3] of $\mathcal{M}$ is the graph defined by[4]

$$y = Ax + b, \tag{13.1}$$

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 13.1. 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 (13.1).

If we see x and y as column vectors then (13.1) reads as

$$(A \ -I) \begin{pmatrix} x \\ y \end{pmatrix} = b \in \mathbb{R}^m,$$

with $C = (A \ -I)$ a somewhat special matrix with m rows and N columns. The first n columns form the matrix A, the last m columns the diagonal matrix with entries -1. The matrix C acts on column vectors

$$z = \begin{pmatrix} x \\ y \end{pmatrix}$$

in $\mathbb{R}^N$. Thus (13.1) is a system of m linear equation for Nunknowns $z_1, z_2, \ldots, z_N$:

$$C_{11}z_1 + C_{12}z_2 + \cdots + C_{1N}z_N = b_1;$$

$$C_{21}z_1 + C_{22}z_2 + \cdots + C_{2N}z_N = b_2;$$

$$\vdots$$

$$C_{m1}z_1 + C_{m2}z_2 + \cdots + C_{mN}z_N = b_m.$$

[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.

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 (13.1). We have

$$Cz = b \iff y = Ax + b \tag{13.2}$$

as equivalent descriptions of non-trivial linear varieties in $\mathbb{R}^{\mathrm{N}}$, under the assumption that C hax maximal rank.

13.1 Implicit function theorem in Euclidean spaces

Refering to Theorem 10.4 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 (10.24) 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[5]: F is continuously differentiable in a neighbourhood of $(x, y) = (a, b)$.

Theorem 13.2. *(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

$$f'(x) = -(F_y(x, f(x)))^{-1} F_x(x, f(x)) \quad \text{for all} \quad x \in \bar{B}_{\delta_0}(a).$$

The proof can be copied from the proofs of Theorems 10.1 and 10.2. Recall that the function $x \to F(x, f(x))$ is never differentiated tot derive the expression for $f'(x)$ but differentiation of this function does help to remember the result. The construction of $y = f(x)$ requires first a choice of $0 < \varepsilon_0 \leq r$ and then a choice of $\delta_0 > 0$ sufficiently small, which in the end has to be chosen even smaller to also have $f'(x) = -(F_y(x, f(x)))^{-1} F_x(x, f(x))$ for

[5] Combine Theorem 10.6 and Theorem 6.11.

$|x| \leq \delta_0$. In general it will not be the case that $\delta_0 > \varepsilon$. Thus Theorem 13.2 can be read as stating the existence of $0 < \delta_0 \leq \varepsilon_0 \leq r$ for which the assertions hold.

Applying Theorem 13.2 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 13.3. *(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 13.2 gives $f : \bar{B}_{\delta_0}(a) \to B_{\varepsilon_0}(b)$ in Theorem 13.3 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 13.3 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

$$\mathbb{R}^n \supset \bar{B}_{\delta_0}(a) \xrightarrow{f} \mathbb{R}^n \tag{13.3}$$

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 13.4. Derive 13.2 from Theorem 13.3. 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.

13.2 General subvarieties

For in general nonlinear subvarieties[6] we ask about an equivalence simliar to (13.2), starting from the nonlinear version $Cz = b$, written in Theorem 13.2 as[7]

$$F(z) = F(x, y) = 0,$$

with $F : \mathbb{R}^{\mathrm{N}} \to \mathbb{R}^m$ continuously differentiable. We use the nonlinear version of (13.1) to agree what we mean by a subvariety $\mathcal{M} \subset \mathbb{R}^{\mathrm{N}}$:

Definition 13.5. *Let $n \in \{1, \ldots, \mathrm{N} - 1\}$. 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 13.2: possibly after renumbering the coordinates it must be that every point $p \in \mathcal{M}$ has*

$$p = (a, b) \in \mathcal{M} \cap \bar{B}_{\delta_0}(a) \times \bar{B}_{\varepsilon_0}(b) = \{(x, f(x)) : x \in \bar{B}_{\delta_0}(a)\}.$$

[6] Not defined yet!

[7] 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 $n = \mathrm{N} - 1$ then $\mathcal{M}$ is called a hypersurface.

Exercise 13.6. Let $F : \mathbb{R}^{\mathrm{N}} \to \mathbb{R}^m$ be continuously differentiable. Assume that for all $z \in \mathbb{R}^{\mathrm{N}}$ with $F(z) = 0$ the derivative $F'(z)$, seen as matrix, has maximal rank. Prove that $\{z \in \mathbb{R}^{\mathrm{N}} : F(z) = 0\}$ is an n-dimensional subvariety of $\mathbb{R}^{\mathrm{N}}$, with $n + m = \mathrm{N}$.

Exercise 13.7. Give an example of an n-dimensional subvariety $\mathcal{M} \subset \mathbb{R}^{\mathrm{N}}$ which is not given by a function F as in Exercise 13.6.

The standard example for Exercise 13.6 is the boundary of a ball in $\mathbb{R}^n$ with center $(a_1, a_2, \ldots, a_n)$ and radius $\delta > 0$:

$$(x_1 - a_1)^2 + \cdots + (x_n - a_n)^2 - \delta^2 = 0. \tag{13.4}$$

There are three equivalent ways to say that $\mathcal{M} \subset \mathbb{R}^{\mathrm{N}}$ is an n-dimensional subvariety:

(A) $\mathcal{M}$ is locally the graph of a continuously differentiable function

$$f : \mathbb{R}^n \to \mathbb{R}^m \quad (n + m = \mathrm{N}),$$

given by $y = f(x)$ after renumbering $z = (x, y)$.

(B) $\mathcal{M}$ is locally the zero level set of

$$F : \mathbb{R}^{\mathrm{N}} \to \mathbb{R}^m \quad (n + m = \mathrm{N}),$$

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[8] of a continuously differentiable function

$$\Phi : \mathbb{R}^n \to \mathbb{R}^{\mathrm{N}},$$

which is injective and has Φ' of maximal rank.

Theorem 13.2 showed that (B) $\Longrightarrow$ (A), and (A) $\Longrightarrow$ (B) because (A) is a special case of B with $F(x, y) = g(y) - x$. Likewise (A) is a special case of

[8] The inverse map of Φ is called a chart on M.

C with $\Phi(x) = (x, f(x))$. To complete the circle with a proof that (C) $\Longrightarrow$ (A) we use Theorem 13.3 and the chain rule.

To wit, consider Φ as

$$\Phi : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^m,$$

with, after renumbering, $\Phi(x) = (\Psi(x), \chi(x))$, $\Psi : \bar{B}_r(a) \to \mathbb{R}^n$ and $\chi : \bar{B}_r(a) \to \mathbb{R}^m$ continuously differentiable, and $\Psi'(a)$ invertible in a. This is possible because we assumed that $\Phi'(x)$ is of maximal rank in $x = a$. Theorem 13.3, applied to $g = \Psi$ with $y = x$, provided us with a continuously differentiable injective function f renamed here as ϕ, $\phi : \bar{B}_{\delta_0}(\Psi(a)) \to \mathbb{R}^n$, with $\phi'(\xi)$ invertible[9] 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

$$\text{(A)} \Longleftrightarrow \text{(B)} \Longleftrightarrow \text{(C)}$$

is now complete.

Exercise 13.8. 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 13.9. As Exercise 13.8, but with $f : \mathbb{R}^n \to \mathbb{R}^m$ and $f'(x)$ of maximal rank in every $x \in \mathcal{M}$.

13.3 Images of ball boundaries

With $0 < \varepsilon_1 \leq \delta_1 \leq \delta_0 \leq \varepsilon_0$ Theorem 13.3 provided us with a chain

$$\bar{B}_{\varepsilon_1}(b) \xrightarrow{g} \bar{B}_{\delta_1}(a) \xrightarrow{f} \bar{B}_{\varepsilon_0}(b)$$

[9] Not used here.

in which both links are injective but not surjective as every mage is contained in the open ball. The smaller δ_1 and ε_1 were needed for the injectivity of g on the smaller closed ball $\bar{B}_{\varepsilon_1}(b)$.

The images of the boundaries $\partial B_{\varepsilon_1}(b)$ and $\partial B_{\varepsilon_0}(a)$ are the subvarieties $g(\partial B_{\varepsilon_1}(b))$ and $f(\partial B_{\varepsilon_0}(a))$. In case g and f are linear maps and $a = b = 0$, it is easy to see that these images are graphs over the unit sphere

$$S^{n-1} = \{x \in \mathbb{R}^n : |x| = 1\}.$$

If $A : \mathbb{R}^n \to \mathbb{R}^n$ is such an invertible linear map, then a height function $h : S^{n-1} \to \mathbb{R}^+$ can be constructed to make that the image of $\partial B_1(0)$ under A is of the form

$$S_h = \{h(x)x : x \in S^{n-1}\}. \tag{13.5}$$

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 13.10. 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 13.10 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 13.3 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 exrcise is a small project that also requires Theorem 13.2, to be expanded on.

Exercise 13.11. Prove the statement in Exercise 13.10 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.

13.4 Coordinate transformations

If a point P on an n-dimensional subvariety $\mathcal{M}$ of $\mathbb{R}^N$ lies in the image of a Φ and a Ψ as in (C) in Section 13.2, say with $\Phi(\xi)$ and $\Psi(\eta)$, and $P = \Phi(0) = \Psi(0)$, with 0 an interior point of the domains of Φ and Ψ, then ξ are η are related by statements as in Theorem 13.3 in a neighbourhood of 0.

13.5 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.

14 Applications in Biology

This is joint work with BioBob and the SysBio group. We consider cellular chemical networks of reacting metabolites with enzymes catalyzing the reactions. Let $\boldsymbol{x} = (\boldsymbol{x}_E, \boldsymbol{x}_I)$ be a set of cellular metabolite concentrations, with E an index set of external concentrations and I an index set of internal ones, and let e_j with j in some other index set J denote the enzyme concentrations. Each enzyme drives a reaction with reaction rate $v_j(e_j, \boldsymbol{x}) = e_j f_j(\boldsymbol{x})$, and the dynamics of the network are specified by

$$\dot{x}_i = \sum_{j \in J} N_{ij} v_j(e_j, \boldsymbol{x}) \quad (i \in I), \tag{14.1}$$

in which N is a stoichiometry matrix. We shall be interested in the question as to if and how the cell is capable of tuning its enzyme concentrations to maximise certain output flows.

The external concentrations are often considered to be prescribed and constant. In particular we may then append the equations

$$\dot{x}_i = 0 \quad (i \in E) \tag{14.2}$$

to (14.1).

The simplest of such networks are linear chains

$$X_0 \overset{E_1}{\longleftrightarrow} X_1 \overset{E_2}{\longleftrightarrow} X_2 \overset{E_3}{\longleftrightarrow} \ldots \overset{E_N}{\longleftrightarrow} X_N \overset{E_{N+1}}{\longleftrightarrow} X_{N+1},$$

with

$$E = \{0, N+1\}, \quad I = \{1, 2, \ldots, N\}, \quad J = \{1, 2, \ldots, N+1\},$$

$$N_{ii} = 1, \quad N_{ij} = -1 \quad \text{for} \quad j = i+1, \quad N_{ij} = 0$$

for $j < i$ and $j > i+1$.

The function $f_j(\boldsymbol{x})$ is often referred to as the saturation level of enzyme E_j. It is derived from mass action kinetics involving different time scales, or more heuristically from fitting such functions to experimental data.

14.1 Henry-Michaelis-Menten kinetics

Enzyme reaction rates are derived from mass action kinetics for reaction blocks which in the simplest case are of the form

$$S + E \rightleftarrows ES \rightleftarrows EP \rightleftarrows P + E, \tag{14.3}$$

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

$$s = [S],\ p = [P],\ c_0 = [E],\ c_1 = [ES],\ c_2 = [EP],$$

mass action kinetics for each of the six arrows in (14.3) gives a coupled system of differential equations

$$\frac{ds}{dt} = -k_1 s c_0 + k_2 c_1; \quad \frac{dc_0}{dt} = -k_1 s c_0 + k_2 c_1 + k_5 c_2 - k_6 p c_0;$$

$$\frac{dc_1}{dt} = k_1 s c_0 - k_2 c_1 - k_3 c_1 + k_4 c_2;$$

$$\frac{dc_2}{dt} = k_3 c_1 - k_4 c_2 - k_5 c_2 + k_6 p c_0; \quad \frac{dp}{dt} = k_5 c_2 - k_6 p c_0,$$

with reaction constants $k_{1,3,5}$ for the forward and $k_{2,4,6}$ for the backward reactions in (14.3).

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, which typically involves some small molecular groups consumed or produced.

The right hand sides of these equations are linear in c_0, c_1, c_2, so we can write

$$\begin{pmatrix} \dot{s} \\ \dot{c}_0 \\ \dot{c}_1 \\ \dot{c}_2 \\ \dot{p} \end{pmatrix} = \begin{pmatrix} -k_1 s & k_2 & 0 \\ -k_1 s - k_2 p & k_2 & k_5 \\ k_1 s & -k_2 - k_3 & k_4 \\ k_6 p & k_3 & -k_4 - k_5 \\ -k_6 p & 0 & k_5 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix},$$

in which we recognise that the total enzyme concentration

$$c_0 + c_1 + c_2 = e_{tot} = \varepsilon$$

is constant, a positive constant denoted by ε and assumed to be small in what follows. We also have that

$$\dot{s} + \dot{c}_1 + \dot{c}_2 + \dot{p} = 0.$$

The system is of the form

$$\begin{array}{l} \dot{x} = A(x)c \\ \dot{c} = B(x)c \end{array} \quad \text{for} \quad x = \begin{pmatrix} s \\ p \end{pmatrix} \quad \text{and} \quad c = \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix},$$

with $A(x)$ and $B(x)$ matrices depending on x. If we scale the free and bounded enzyme concentrations with ε setting

$$c = \varepsilon\gamma,$$

then a splitting of time-scales appears when ε is small:

$$\dot{x} = \varepsilon A(x)\gamma; \quad \dot{\gamma} = B(x)\gamma.$$

Introducing a new time variable $\tau = \varepsilon t$, we write

$$\dot{x} = \frac{dx}{dt} = \varepsilon\frac{dx}{d\tau} = \varepsilon x', \quad \dot{\gamma} = \frac{d\gamma}{dt} = \varepsilon\frac{d\gamma}{d\tau} = \varepsilon\gamma'$$

and conclude that

$$x' = A(x)\gamma,\ \varepsilon\gamma' = B(x)\gamma.$$

Exercise 14.1. For $\varepsilon = 0$ this reduces to

$$x' = B(x)\gamma \quad \text{with} \quad A(x)\gamma = 0 \quad \text{and} \quad \gamma_0 + \gamma_1 + \gamma_2 = 1.$$

Explain how this leads to

$$\dot{p} = \frac{\varepsilon(k_1k_3k_5s - k_2k_4k_6p)}{k_2k_4 + k_2k_5 + k_3k_5 + k_1(k_3 + k_4 + k_5)s + (k_2 + k_3 + k_4)k_6p} = -\dot{s} \qquad (14.4)$$

as the modelling equation for

$$S \overset{E}{\longleftrightarrow} P$$

Hint: the rank of the matrix $A(x)$ 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

$$k_2k_4 + k_2k_5 + k_3k_5 + k_1s(k_3 + k_4 + k_5) + (k_2 + k_3 + k_4)k_6p.$$

Exercise 14.2. Explain why

$$k_2k_4 + k_2k_5 + k_3k_5 : k_1s(k_3 + k_4 + k_5) + (k_2 + k_3 + k_4)k_6p$$

is the ratio between free enzyme and bonded enzyme.

Exercise 14.3. The resulting differential equation (14.4) is of the form

$$\dot{p} = \frac{\varepsilon(k_{135}s - k_{246}p)}{k_{2345} + k_{1345}s + k_{2346}p} = -\dot{s}, \tag{14.5}$$

but often[1] written as

$$\dot{p} = \frac{\frac{V^+}{K_s}(s - \frac{p}{K_{eq}})}{1 + \frac{s}{K_s} + \frac{p}{K_p}} = -\dot{s}.$$

Verify that

$$K_s = \frac{k_{2345}}{k_{1345}} = \frac{k_2k_4 + k_2k_5 + k_3k_5}{k_1(k_3 + k_4 + k_5)}, \quad K_p = \frac{k_{2345}}{k_{2346}} = \frac{k_2k_4 + k_2k_5 + k_3k_5}{(k_2 + k_3 + k_4)k_6},$$

$$K_{eq} = \frac{k_{135}}{k_{246}} = \frac{k_1k_3k_5}{k_2k_4k_6}, \quad V^+ = \frac{k_{135}}{k_{1345}} = \frac{\varepsilon k_3k_5}{k_3 + k_4 + k_5}.$$

14.2 More complicated reactions

14.3 Optimisation problems

14.4 Self-steering networks

[1] See Appendix 1 of Teusink's FEBS 2000 paper.

15 In grondverf: integrals in several variables

Integraalrekening voor continue functies

$$u : [a, b] \times [c, d] \to \mathbb{R},$$

met $[a, b] \times [c, d]$ een rechthoek in $\mathbb{R}^2$ gaat met partities P als in (5.6) voor $[a, b]$, en

$$c = y_0 \leq y_1 \leq \cdots \leq y_{\mathrm{M}} = b \tag{15.1}$$

voor $[c, d]$.

Om kort te gaan, met onder- en bovensommen, of met tussensommen van de vorm

$$\sum_{k=1}^{\mathrm{N}} \sum_{l=1}^{\mathrm{M}} u(\xi_k, \eta_l)(x_k - x_{k-1})(y_l - y_{l-1}),$$

met tussenwaarden $\xi_k \in [x_{k-1}, x_k]$ en $\eta_l \in [y_{l-1}, y_l]$, volgt het bestaan van één unieke $J \in \mathbb{R}$ die de integraal van u over $[a, b] \times [c, d]$ genoemd wordt, met notatie (en stelling over het vrij mogen kiezen van de integratievolgorde)

$$J = \int_{[a,b]\times[c,d]} u = \int_a^b \int_c^d u(x, y)\, dy\, dx = \int_c^d \int_a^b u(x, y)\, dx\, dy. \tag{15.2}$$

De herhaalde integralen worden uitgewerkt met de technieken en regels voor gewone integralen.

We maken daar hier nu verder geen woorden aan vuil maar wijzen wel nog even op de details in de hier gebruikte notaties in relatie tot de discussie na Stelling 9.2 en Opgave **??** onder Definitie 5.26.

15.1 Notationele kwesties met die d-tjes

Anders dan de notatie suggereert is (15.2) geen uitdrukking met de onder Stelling 9.2 aangekondigde 2-vormen. In (15.2) is de herhaalde integraal immers te lezen als

$$\int_c^d \left\{ \int_a^b u(x, y)\, dx \right\} dy,$$

een integraal van een integraal, net als de herhaalde integraal direct daarvoor, met de afsluitende accolade stevig tussen dx en dy in. Beide herhaalde integralen zijn gelijk aan

$$\int_{[a,b]\times[c,d]} u,$$

gedefinieerd via Riemannsommen en een limietovergang, met een notatie waarin $dxdy$ of $dydx$ als produkt van dx en dy vermeden is.

Exercise 15.1. Je kunt je je nu drie assen (nu in de ruimte in plaats van in het vlak) voorstellen die loodrecht op elkaar staan in hun gemeenschappelijke snijpunt $x = y = z = 0$, met richtingen nog te kiezen, maar dat pas nadat je op de x-as a en b gemarkeerd hebt, op de y-as c en d, en de grafiek van zo maar een continue functie zonder nulpunten en tekenwisselingen hebt gekozen. Nu zijn er acht gevallen waarin je de integraal kunt relateren aan de inhoud ingesloten door $x = a$, $x = b$, $y = c$, $y = d$, $z = 0$ en $z = u(x, y)$. Kijk nu naar de tekens van $b - a$ en $d - c$, zeg maar van dx en dy, in de twee gevallen dat zowel u als de integraal positief zijn, en laat zien dat ze hetzelfde teken hebben: zeg maar $dxdy > 0$.

Exercise 15.2. Neem nu eventueel voor het gemak $a = c = 0$, $|b| = |d| = 1$. In het geval dat $dxdy$ en $u(x, y)$ allebei positief zijn is er een orienterende ordening tussen de drie assen die overeenkomt met de correpondentie

positieve x-as $\leftrightarrow$ middelvinger
positieve y-as $\leftrightarrow$ wijsvinger
positieve z-as $\leftrightarrow$ duim

op je linkerhand. Overtuig jezelf daarvan en ook van het volgende: teken je een x-as en een y-as op het schoolbord die elkaar loodrecht snijden in $x = y = 0$ dan vertellen de positieve x-as en de positieve y-as als geordend assenpaar welke kant van het bord $z > 0$ heeft; bij de standaardkeuze met x naar rechts en y naarboven komt de positieve z-as naar voren het bord uit. Het bijbehorende assenkruis heet dan rechtsdraaiend[1].

Als je wil rekenen met formules waarin produkten van dx en dy voorkomen dan kun je in het licht van Opgave 15.1 en Opgave 15.2 tot de conclusie komen dat het handig is om af te spreken dat

$$dxdy = -dydx,$$

zonder de afsluitende accolade tussen dx en dy, in de zin dat

$$J = \int_{[a,b]\times[c,d]} u = \iint_{[a,b]\times[c,d]} u(x, y)\, dxdy,$$

gezien als

$$\iint_{[a,b]\times[c,d]} \text{ werkend op } \quad u(x, y)\, dxdy,$$

gelijk is aan

$$-\iint_{[a,b]\times[c,d]} u(x, y)\, dydx,$$

[1] Dit is moeilijk voor jongens.

gezien als

$$-\iint_{[a,b]\times[c,d]} \quad \text{werkend op} \quad u(x,y)\,dydx,$$

waarbij die werking dan precies gedefinieerd moet worden op zo'n manier dat onder coördinatentransformaties de integralen transformeren op een manier die consistent is met wat er gebeurt met de inhouden waar het verhaal mee begonnen is. Maar dat hebben we hier nog niet behandeld.

15.2 Formele d-algebra zonder betekenis?

Ondertussen spreekt de algebra voor differentiaalvormen deels voor zich. Na $du = u'(x)dx$ voor $u = u(x)$ hebben we natuurlijk ook

$$du = u_x dx + u_y dy = \frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy \tag{15.3}$$

als de d van de 0-vorm $u = u(x,y)$, een uitdrukking van de vorm

$$fdx + gdy = f(x,y)dx + g(x,y)dy,$$

een 1-vorm waarvan de d via

$$d(f(x,y)dx + g(x,y)dy)$$

vraagt[2] om een som en produktregel met d van $f(x,y)$ en $g(x,y)$, en een nog te definiëren d van dx en dy.

Voor de hand als som- en daarna twee keer de produktregel ligt

$$d(fdx + gdy) = d(fdx) + d(gdy) =$$

$$dfdx + fddx + dgdy + gddy =$$

$$f_x dxdx + f_y dydx + g_x dxdy + g_y dydy + fddx + gddy.$$

Als $dxdy = -dydx$ volgt dan

$$d(fdx + gdy) = f_x dxdx + g_y dydy + (g_x - f_y)dxdy + fddx + gddy.$$

Met $y = x$ leest $dxdy = -dydx$ nu toch echt als $dxdx = -dxdx$, dus $dxdx = 0 = dydy$ als consistent voor te schrijven tweede rekenregel ligt dan ook voor de hand. Daarmee reduceert het bovenstaande tot

$$d(fdx + gdy) = (g_x - f_y)dxdy + fddx + gddy,$$

[2] Wat d is dat kun je straks weten.

en blijft de vraag wat ddx en ddy dan wel mogen zijn.

Keine blasse Ahnung zou ik zeggen dus doe dan maar NUL. Dan wordt de werking van d op 1-vormen

$$f(x,y)dx + g(x,y)dy \xrightarrow{d} (g_x(x,y) - f_y(x,y))dxdy,$$

en dus

$$u(x,y) \xrightarrow{d} u_x(x,y)dx + u_y(x,y)dy \xrightarrow{d} (u_{yx}(x,y) - u_{xy}(x,y))dxdy = 0.$$

Der blasse Ansatz $ddx = ddy = 0$ gibt $ddu = 0$ voor alle $u = u(x,y)$.

Ik zou zeggen: machen wir. Dus $d^2 = 0$[3], en met $f = f(x,y), g = g(x,y)$ hebben we in de notatie met partiële differentiaalquotiënten nu dat

$$f \xrightarrow{d} \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy, \quad fdx + gdy \xrightarrow{d} (\frac{\partial g}{\partial x} - \frac{\partial f}{\partial y})dxdy, \quad gdxdy \xrightarrow{d} 0 \quad (15.4)$$

Exercise 15.3. Doe de algebra voor

$$f \xrightarrow{d} \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz,$$

$$fdx + gdy + hdz \xrightarrow{d} (\frac{\partial h}{\partial y} - \frac{\partial g}{\partial z})dydz + (\frac{\partial f}{\partial z} - \frac{\partial h}{\partial x})dzdx + (\frac{\partial g}{\partial x} - \frac{\partial f}{\partial y})dxdy,$$

$$fdydz + gdzdx + hdxdy \xrightarrow{d} (\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} + \frac{\partial h}{\partial z})dxdydz,$$

$$hdxdydz \xrightarrow{d} 0,$$

met $f = f(x,y,z), g = g(x,y,z), h = h(x,y,z)$, gebruikmakend van de som- en produktregel voor d, $dxdy = -dydx, dxdz = -dzdx, dydz = -dzdy$, en $ddx = ddy = ddz = 0$. Verifieer dat niet alleen $ddf = 0$ maar ook $dd(fdx + gdy + hdz) = 0$: dd maakt alles nul en zo hoort dat kennelijk.

Exercise 15.4. Doe de algebra nog een keer voor $F \xrightarrow{d} F'(x)dx$ en $f(x)dx \xrightarrow{d} 0$ met $f(x)$ en $F(x)$.

[3] Daar kan i een puntje aan zuigen.

15.3 Transformaties en parametrisaties

Als $x \to f(x) = F'(x)$ en $t \to x'(t)$ continu zijn met bijvoorbeeld $x(0) = a$ en $x(1) = b$, dan is

$$\int_a^b f(x)\,dx = \int_a^b F'(x)\,dx = \int_a^b dF = F(b) - F(a) = F(x(1)) - F(x(0)) =$$

$$[F(x(t))]_0^1 = \int_0^1 F'(x(t))x'(t)\,dt = \int_0^1 f(x(t))x'(t)\,dt,$$

waarin we

$$dx = x'(t)dt \tag{15.5}$$

als een stukje van de formele d-algebra in Sectie 15.2 herkennen.

Een 1-vorm $f(x)dx$ in x wordt via $t \to x(t)$ zo teruggetrokken tot een 1-vorm

$$f(x)dx = f(x(t))x'(t)dt \tag{15.6}$$

in t. Evenzo wordt de 1-vorm

$$f(x,y)dx + g(x,y)dy$$

via $t \to x(t)$ en $t \to y(t)$ met dezelfde algebra teruggetrokken tot een 1-vorm

$$f(x,y)dx + g(x,y)dy = (f(x(t),y(t))x'(t) + g(x(t),y(t))y'(t))dt \tag{15.7}$$

in t.

De gelijktekens in (15.6) en zeker (15.7) zijn wat dubieus, het is wellicht beter om (15.6) en (15.7) nu als afbeeldingen te lezen, bijvoorbeeld

$$f(x,y)dx + g(x,y)dy \xrightarrow{terugtrekken} (f(x(t),y(t))x'(t) + g(x(t),y(t))y'(t))dt,$$

waarbij het terugtrekken gebeurd via de afbeelding in kwestie, in dit voorbeeld $t \to (x(t),y(t))$, die van 1 naar 2 variabelen gaat.

Nemen we in plaats van een 1-vorm een 2-vorm, zeg $dxdy$ zelf, dan geeft terugtrekken via dezelfde $t \to (x(t),y(t))$ dat

$$dxdy \to x'(t)dt\,y'(t)dt = x'(t)y'(t)dtdt = 0.$$

Weinig zinvol wellicht, maar met

$$(r,\theta) \to (x(r,\theta),y(r,\theta))$$

krijgen we

$$dxdy \to (\frac{\partial x}{\partial r}dr + \frac{\partial x}{\partial \theta}d\theta)(\frac{\partial y}{\partial r}dr + \frac{\partial y}{\partial \theta}d\theta) = (\frac{\partial x}{\partial r}\frac{\partial y}{\partial \theta} - \frac{\partial y}{\partial r}\frac{\partial x}{\partial \theta})drd\theta, \quad (15.8)$$

met de determinant[4] van de Jacobimatrix.

In het geval dat $x = r\cos\theta, y = r\sin\theta$ wordt dit

$$dxdy \to rdrd\theta,$$

op grond van een algebra die zich (ook) voor coördinatentransformaties genereert uit de regeltjes die we in de d-algebra hebben geponeerd.

Merk op dat (15.7) niet met een coördinatentransformatie kan corresponderen maar (15.8) wel, en lees de afleiding van (15.7) nog eens terug met t vervangen door ϕ in bijvoorbeeld $[0, 2\pi]$. Met $x(0) = x(2\pi)$ en $y(0) = y(2\pi)$ kun je dit voorbeeld nu goed vergelijken met (15.8) waarin r vast is genomen, alvorens verder te ontdekken hoe de terugtrekalgebra werkt voor

$$(\theta, \phi) \to (x(\theta, \phi), y(\theta, \phi), z(\theta, \phi) \quad (15.9)$$

en 2-vormen in x, y, z.

Exercise 15.5. Trek $f(x,y,z)dx+g(x,y,z)dy+h(x,y,z)dz$ terug naar een 2-vorm in θ en ϕ.

15.4 Een transformatiestelling

Voor $R \subset \mathbb{R}^2$ en $f : R \to \mathbb{R}$ geldt dat

$$\iint_R f(x,y)\,dxdy = \iint_Q f(x(r,\theta), y(r,\theta))\,(\frac{\partial x}{\partial r}\frac{\partial y}{\partial \theta} - \frac{\partial y}{\partial r}\frac{\partial x}{\partial \theta})drd\theta,$$

als

$$Q \xrightarrow{(r,\theta)\to(x,y)} R$$

aan wat aannamen voldoet. Voorbeeld: Sectie 16.1. Hoe doen we dit? Onderstaande is om de gedachten te bepalen.

Met een bijectie

$$(x,y) \xrightarrow{\Phi} (u,v)$$

[4] Plus of min de oppervlakte van de ruit opgespannen door de twee kolomvectoren.

tussen $R \subset \mathbb{R}^2_{x,y}$ en $A \subset \mathbb{R}^2_{u,v}$ is wat we willen een uitspraak die de integraal

$$\iint_A g(u,v)\,dudv$$

relateert aan een integraal met $g(u(x,y),v(x,y))$ en $dxdy$ over R, liefst meteen al met de conventie dat $dudv = -dvdu$ en $dxdy = -dydx$.

Als de bijectie niet-lineair is dan zijn R en A zeker niet allebei een rechthoek, laat staan van de vorm zoals in Sectie 15, dus neem in eerste instantie alleen R van die vorm. We nemen R gesloten, dus met de rand erbij, en Φ gedefinieerd[5] op heel $\mathbb{R}^2$, met continue partiële afgeleiden. Voor het gemak nemen we $R = [0,1] \times [0,1]$.

Zie nu Opgave **??** en lees i.p.v

$$F(x,y) = g(y) - x \quad \text{nu} \quad F(x,y,u,v) = \begin{pmatrix} \Phi_1(x,y) - u \\ \Phi_2(x,y) - v \end{pmatrix}.$$

De uitgepakte versie van de stelling die je daar bewezen hebt is de inverse functiestelling die we nu nodig hebben. Die stelling zegt dat als in (x_0,y_0) de Jacobimatrix

$$J(x,y) = \begin{pmatrix} \frac{\partial \Phi_1}{\partial x} & \frac{\partial \Phi_1}{\partial y} \\ \frac{\partial \Phi_2}{\partial x} & \frac{\partial \Phi_2}{\partial y} \end{pmatrix}$$

inverteerbaar is, in een omgeving van $(u_0,v_0) = (\Phi_1(x_0,y_0),\Phi_2(x_0,y_0))$ de inverse functie

$$(u,v) \xrightarrow{\Phi^{-1}} (x,y)$$

bestaat en continu differentieerbaar is. De Jacobimatrix van de inverse afbeelding is de inverse van de Jacobimatrix van Φ.

Voor de te formuleren transformatiestelling nemen we nu aan dat de Jacobimatrix $J(x,y)$ in elk punt van R inverteerbaar is. Daarmee is A een gebied in $\mathbb{R}^2_{u,v}$ met vier randen geparametriseerd[6] door

$$x \to \Phi(x,0), \quad y \to \Phi(1,y), \quad x \to \Phi(x,1), \quad y \to \Phi(0,y).$$

Bij partities

$$(P) \quad 0 = x_0 \le x_1 \le \cdots \le x_{\mathrm{N}} = 1 \quad \text{met} \quad \mathrm{N} \in \mathbb{N},$$

[5] Zonder beperking der algemeenheid, maar dat moet nog blijken.

[6] Het alternatief dat later nog komt wellicht is:

$$t \to \Phi(t,0), \quad t \to \Phi(1,t), \quad t \to \Phi(1-t,1), \quad t \to \Phi(1-t,0),$$

$$(Q) \quad 0 = y_0 \leq y_1 \leq \cdots \leq y_{\mathrm{M}} = 1 \quad \text{met} \quad \mathrm{M} \in \mathbb{N},$$

horen $(\mathrm{M}+1)(\mathrm{N}+1)$ parametrisaties

$$x \to \Phi(x, y_j) \quad \text{en} \quad y \to \Phi(x_i, y) \quad (i = 0, \ldots, \mathrm{M},\, j = 0 \ldots, \mathrm{N}),$$

die een rooster van vervormde rechthoekjes S_{ij} maken in A.

De goede definitie van Riemann integreerbaarheid van $g : A \to \mathbb{R}$ moet[7] nu geven dat met

$$M_{ij} = \sup_{S_{ij}} g \quad \text{en} \quad m_{ij} = \inf_{S_{ij}} g$$

geldt dat

$$\sum_{ij} m_{ij}|S_{ij}| \leq \iint_A g \leq \sum_{ij} M_{ij}|S_{ij}|,$$

waarin S_{ij} de oppervlakte is van S_{ij}, hetgeen we herschrijven als

$$\sum_{ij} m_{ij}\frac{|S_{ij}|}{|R_{ij}|}|R_{ij}| \leq \iint_A g \leq \sum_{ij} M_{ij}\frac{|S_{ij}|}{|R_{ij}|}|R_{ij}|,$$

en opmerken dat ook geldt dat

$$M_{ij} = \sup_{R_{ij}} f \quad \text{en} \quad m_{ij} = \inf_{R_{ij}} f$$

met $f = g \circ \Phi$.

Het is niet zo moeilijk[8] om precies te maken dat

$$\frac{|S_{ij}|}{|R_{ij}|} \sim |\mathrm{det}J(x_i, y_i)| \tag{15.10}$$

als $\mathrm{M}, \mathrm{N} \to \infty$ om op de Riemann integreerbaarheid van

$$(x, y) \to f(x, y)|J(x, y)|$$

over R uit te komen MET gelijkheid van de integralen:

$$\iint_R f|\mathrm{det}J| = \iint_A g.$$

[7] Moet nog blijken!

[8] Zie e.g. Sectie 5 van Hoofdstuk III in het Advanced Calculus boek van Edwards.

16 Toepassingen

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 de functional calculus waarmee voor z in $f(z)$ ook iets heel anders mag worden ingevuld, bijvoorbeeld een vierkante matrix.

16.1 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

$$\iint_{\mathbb{R}^2} u = \lim_{R\to\infty} \underbrace{\iint_{[-R,R]\times[-R,R]} u(x,y)d(x,y)}_{J(R)} = \lim_{R\to\infty} J(R) \tag{16.1}$$

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

$$\iint_{x^2+y^2\leq R^2} u(x,y)d(x,y) = \int_0^{2\pi}\int_0^R u(r\cos\theta, r\sin\theta)\, r dr\, d\theta$$

$$= \int_0^R \int_0^{2\pi} u(r\cos\theta, r\sin\theta)\, d\theta\, r dr \tag{16.2}$$

gebruiken we

$$0 = r_0 \leq r_1 \leq \cdots \leq r_{\mathrm{M}} = R \quad \text{met} \quad \mathrm{M} \in \mathbb{N} \tag{16.3}$$

en

$$0 = \theta_0 \leq \theta_1 \leq \cdots \leq \theta_{\mathrm{N}} = 2\pi \quad \text{met} \quad \mathrm{N} \in \mathbb{N}, \tag{16.4}$$

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 (16.2) in de volgende stelling over harmonische functies.

Exercise 16.1. 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 9.11 als je de integraal van Δu over $\bar B_R$ hebt vertaald naar een integraal met alleen maar $d\theta$.

Ook leuk is dat voor niet-negatieve continue functies $u:\mathbb{R}^2\to\mathbb{R}$ de integraal

$$\iint_{\mathbb{R}^2}u=\lim_{R\to\infty}\iint_{x^2+y^2\le R^2}u(x,y)d(x,y) \tag{16.5}$$

nu net zo natuurlijk gedefinieerd is in $[0,\infty]$ als door (16.1). Alleen een wiskundige vraagt zich dan af dit consistent is. Dat moet en dat mag hoor:

Exercise 16.2. Voor niet-negatieve continue $u:\mathbb{R}^2\to\mathbb{R}$ geldt

$$\lim_{R\to\infty}\iint_{x^2+y^2\le R^2}u(x,y)d(x,y)=\lim_{R\to\infty}\iint_{[-R,R]\times[-R,R]}u(x,y)d(x,y).$$

Zoek maar uit waarom en onthoud dat

$$\iint_{\mathbb{R}^2}u$$

op twee (eigenlijk vier) manieren is uit te rekenen als dat nodig is, zie (15.2) en (16.2).

In de formule van Stirling, zie (**??**), stond nog een integraal die we nu netjes kunnen uitrekenen met behulp van Opgave 16.2 en de functie

$$(x,y)\xrightarrow{u}e^{-\frac{1}{2}(x^2+y^2)}$$

Kort door de bocht opgeschreven concluderen we dat

$$\left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx\right)^2 = \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\,dy = \iint_{\mathbb{R}^2} e^{-\frac{1}{2}(x^2+y^2)}\,d(x,y)$$

$$= \int_0^{\infty}\int_0^{2\pi} e^{-\frac{1}{2}r^2}\,rd\theta dr = \int_0^{\infty} 2\pi e^{-\frac{1}{2}r^2}\,rdr = 2\pi[-e^{-\frac{1}{2}r^2}]_0^{\infty} = 2\pi.$$

Exercise 16.3. Laat met Opgave 16.2 zien dat

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx = \sqrt{2\pi}.$$

Bijgevolg hebben

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} \quad \text{en} \quad u(x,y) = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)} \tag{16.6}$$

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 (16.6) worden de 1-en 2-dimensionale standaard *normale verdeling* genoemd. Is de functie g hetzelfde als de functie f in (16.6), dan zijn X en Y allebei standaard normaal verdeeld. De twee stochastische grootheden X en Y kunnen op elkaar gedeeld worden. De kans op

$$Q = \frac{Y}{X} \in [a, b]$$

is dan gelijk aan de integraal van $u(x, y)$ over het gebied ingesloten door de lijnen $y = ax$ en $y = bx$.

In het geval dat X en Y standaard normaal verdeeld en onderling onafhankelijk zijn, bestaat die integraal uit twee identieke stukken waarvan er één gegeven wordt door

$$\{(x, y) : x \geq 0,\, ax \leq y \leq bx\},$$

een gebied dat in poolcoördinaten beschreven wordt door θ in een deelinterval van $(-\frac{\pi}{2}, \frac{\pi}{2})$.

We willen concluderen dat

$$P(Q \in [a, b]) = 2 \int_0^\infty \int_{ax}^{bx} \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}\, dy\, dx$$

$$= \frac{1}{\pi} \int_0^\infty \int_{\arctan a}^{\arctan b} e^{-\frac{1}{2}r^2}\, r d\theta dr = \frac{1}{\pi}(\arctan b - \arctan a) = \int_a^b \frac{1}{\pi} \frac{1}{1+q^2}\, dq.$$

De stochastische grootheid Q heeft dan een kansdichtheid gegeven door de functie

$$q \to \frac{1}{\pi} \frac{1}{1+q^2}.$$

Exercise 16.4. 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.

16.2 Gradient, kettingregel, coördinatentransformaties

De *kettingregel* generaliseert de regel in Opgave 8.2, zie ook Sectie **??**, Opgave **??** en later Opgave **??**. Met de opmerking dat (**??**) 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

$$x = r\cos\theta \quad \text{en} \quad y = r\sin\theta \tag{16.7}$$

zien als gegeven door poolcoördinaten $r, \theta \in \mathbb{R}$ voor $(x, y) \neq (0, 0)$.

Een differentieerbare scalaire functie $F(x, y)$ van x en y is zo automatisch ook een differentieerbare functie van r en θ. In wat volgt zien we (16.7) 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.

16.2.1 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

$$F(x_0 + h, y_0 + k) = F(x_0, y_0) + ah + bk + R(h, k; x_0, y_0), \tag{16.8}$$

met $a, b \in \mathbb{R}$ en

$$\frac{R(h, k; x_0, y_0)}{\sqrt{h^2 + k^2}} \to 0 \quad \text{als} \quad \sqrt{h^2 + k^2} \to 0, \tag{16.9}$$

zie ook Sectie **??** en vergelijk met het uitpakken van (**??**) hierboven. De volgende opgave is misschien nu wat dubbelop, maar dat kan geen enkel kwaad.

Exercise 16.5. 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 (16.8) geldt met (16.9). 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.

[1] Beter: lineaire afbeeldingen, in dit hele hoofdstuk de facto matrices.

Meerdere notaties worden gebruikt, zoals

$$a = F_x(x_0, y_0) = \frac{\partial F}{\partial x}(x_0, y_0) = (\delta_x F)(x_0, y_0) = (D_1 F)(x_0, y_0); \qquad (16.10)$$

$$b = F_y(x_0, y_0) = \frac{\partial F}{\partial y}(x_0, y_0) = (\delta_y F)(x_0, y_0) = (D_2 F)(x_0, y_0), \qquad (16.11)$$

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

$$e_x = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{en} \quad e_y = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad (16.12)$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \nabla F = \frac{\partial F}{\partial x} e_x + \frac{\partial F}{\partial y} e_y = e_x \frac{\partial F}{\partial x} + e_y \frac{\partial F}{\partial y}, \qquad (16.13)$$

de *gradient* van F in (x_0, y_0), geschreven zonder (x_0, y_0). Merk op dat het lineaire gedeelte in (16.8) te schrijven is als

$$ah + bk = \begin{pmatrix} a \\ b \end{pmatrix} \cdot \begin{pmatrix} h \\ k \end{pmatrix} = \begin{pmatrix} h \\ k \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = h \frac{\partial F}{\partial x} + k \frac{\partial F}{\partial y}, \qquad (16.14)$$

het inprodukt[2] van ∇F en de verschilvector $\binom{h}{k}$.

We zien dus hoe de gradiënt de vector is die de lineaire afbeelding $DF : \mathbb{R}^2 \to \mathbb{R}$ via het inprodukt representeert als

$$\begin{pmatrix} h \\ k \end{pmatrix} \xrightarrow{DF} \nabla F \cdot \begin{pmatrix} h \\ k \end{pmatrix},$$

maar ook dat (16.14) te lezen is als de *differentiaaloperator*

$$h \frac{\partial}{\partial x} + k \frac{\partial}{\partial y} \quad \text{werkend op} \quad F.$$

Evenzo zien we ∇ als

$$\nabla = e_x \frac{\partial}{\partial x} + e_y \frac{\partial}{\partial y} \quad \text{werkend op} \quad F \quad \text{geeft} \quad \nabla F, \qquad (16.15)$$

[2] Het inprodukt van twee vectoren in $\mathbb{R}^2$ wordt gegeven door $\binom{a}{b} \cdot \binom{h}{k} = ah + bk$.

een vectorwaardige differentiaaloperator.

Middels het inprodukt kan ∇ ook werken op een vectorwaardige differentieerbare functie

$$(x,y) \to \begin{pmatrix} V_x(x,y) \\ V_y(x,y) \end{pmatrix} = \begin{pmatrix} V_x \\ V_y \end{pmatrix} = V_x e_x + V_y e_y,$$

en wel als

$$\nabla \cdot V = (e_x \frac{\partial}{\partial x} + e_y \frac{\partial}{\partial y}) \cdot (V_x e_x + V_y e_y) = \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y}, \tag{16.16}$$

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 (16.12) die samen de standaardbasis van $\mathbb{R}^2$ vormen. Merk wel op dat V_x en V_y van x en y afhangen maar e_x en e_y niet. De indices x en y staan voor de x-richting en de y-richting, en die richtingen zijn overal in het x, y-vlak hetzelfde.

Elk van de twee termen in ∇ werkt nu alleen op V_x en V_y, en omdat

$$e_x \cdot e_x = e_y \cdot e_y = 1 \quad \text{en} \quad e_x \cdot e_y = e_y \cdot e_x = 0, \tag{16.17}$$

blijven er maar twee termen over in (16.16). 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 (16.16) 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 (16.17) en de andere twee vereenvoudigen en blijven in die vorm over in (16.16).

[3]Niet te verwarren met het gebruik van subscripten voor partiële afgeleiden!

Als $V = \nabla F$ differentieerbaar is dan volgt zo dat

$$\nabla \cdot \nabla F = (e_x \frac{\partial}{\partial x} + e_y \frac{\partial}{\partial y}) \cdot (\frac{\partial F}{\partial x} e_x + \frac{\partial F}{\partial y} e_y) = \frac{\partial}{\partial x}\frac{\partial F}{\partial x} + \frac{\partial}{\partial y}\frac{\partial F}{\partial y} = \Delta F, \quad (16.18)$$

de Laplaciaan van F, die weer gezien kan worden als

$$\Delta F \quad \text{is} \quad \Delta = \frac{\partial}{\partial x}\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\frac{\partial}{\partial y} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \quad \text{werkend op} \quad F. \quad (16.19)$$

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, zie ook Sectie **??** waar dat voor $t \to F(x(t), y(t)$ al is gedaan.

16.2.2 Kettingregel uitgeschreven voor transformaties

We weten dat we de kettingregel toe mogen passen op

$$(r, \theta) \to (r\cos\theta, r\sin\theta) = (X(r,\theta), Y(r,\theta)) = (x, y) \to F(x, y) = G(r, \theta),$$

door de lineaire benadering van

$$(r, \theta) \to (r\cos\theta, r\sin\theta)$$

rond (r_0, θ_0) in te vullen in de lineaire benadering van

$$(x, y) \to F(x, y)$$

rond (x_0, y_0). We doen dit nu met $\tilde{h} = r - r_0$ en $\tilde{k} = \theta - \theta_0$, met weglating van (r_0, θ_0) in de partiële afgeleiden.

Omdat we in deze sectie $F(x, y) = G(r, \theta)$ als onbekende afhankelijke grootheid willen zien, bijvoorbeeld de oplossing van een partiële differentiaalvergelijking, kiezen we nu eerst voor de schrijfwijze zoals rechts in (16.14). 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 (16.14) 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

$$\tilde{h}\frac{\partial G}{\partial r} + \tilde{k}\frac{\partial G}{\partial \theta}.$$

Er volgt dus dat

$$\frac{\partial G}{\partial r} = \frac{\partial X}{\partial r}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial r}\frac{\partial F}{\partial y} \tag{16.20}$$

$$\frac{\partial G}{\partial \theta} = \frac{\partial X}{\partial \theta}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial \theta}\frac{\partial F}{\partial y}, \tag{16.21}$$

in vector-matrixnotatie te schrijven als

$$\begin{pmatrix} \frac{\partial G}{\partial r} \\ \frac{\partial G}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \frac{\partial X}{\partial r}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial r}\frac{\partial F}{\partial y} \\ \frac{\partial X}{\partial \theta}\frac{\partial F}{\partial x} + \frac{\partial Y}{\partial \theta}\frac{\partial F}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{\partial X}{\partial r} & \frac{\partial Y}{\partial r} \\ \frac{\partial X}{\partial \theta} & \frac{\partial Y}{\partial \theta} \end{pmatrix} \begin{pmatrix} \frac{\partial F}{\partial x} \\ \frac{\partial F}{\partial y} \end{pmatrix}, \tag{16.22}$$

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 (16.7) 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. Precies andersom als in (**??**) dus, omdat we de schrijfwijze rechts in (16.14) hebben gebruikt. Transponeren geeft natuurlijk de vorm die consistent is met (**??**).

De kolomvectoren in (16.22) zien er uit als gradiënten, maar dat is slechts misleidende schijn, zoals we in Sectie 16.2.4 zullen zien.

[4] We gaan er nu niet echt vanuit dat de lezer al met matrices heeft leren rekenen.

16.2.3 Kettingregel met Jacobimatrices

Mooie voorbeelden van matrixprodukten als in (12.29) zien we als we in (16.22) aan beide kanten links $(\tilde{h}\ \tilde{k})$ erbij zetten. Dan is

$$(\tilde{h}\ \tilde{k})\begin{pmatrix}\frac{\partial G}{\partial r}\\ \frac{\partial G}{\partial \theta}\end{pmatrix} = (\tilde{h}\ \tilde{k})\begin{pmatrix}\frac{\partial X}{\partial r} & \frac{\partial Y}{\partial r}\\ \frac{\partial X}{\partial \theta} & \frac{\partial Y}{\partial \theta}\end{pmatrix}\begin{pmatrix}\frac{\partial F}{\partial x}\\ \frac{\partial F}{\partial y}\end{pmatrix}, \tag{16.23}$$

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 (16.20) en (16.21) te komen.

Via links en rechts transponeren is (16.23) equivalent met

$$(\frac{\partial G}{\partial r}\ \frac{\partial G}{\partial \theta})\begin{pmatrix}\tilde{h}\\ \tilde{k}\end{pmatrix} = (\frac{\partial F}{\partial x}\ \frac{\partial F}{\partial y})\begin{pmatrix}\frac{\partial X}{\partial r} & \frac{\partial X}{\partial \theta}\\ \frac{\partial Y}{\partial r} & \frac{\partial Y}{\partial \theta}\end{pmatrix}\begin{pmatrix}\tilde{h}\\ \tilde{k}\end{pmatrix}, \tag{16.24}$$

waarin we de *Jacobimatrices* van G, F en Z herkennen, waarin de beeldvariabelen niet horizontaal maar verticaal genummerd worden, zoals we in (**??**) al gezien hebben. Ook zien we dat de volgorde in (16.24) nu net is als in (**??**), hetgeen 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 (**??**) voor beide componenten in één keer opgeschreven worden als

$$\begin{pmatrix}\frac{\partial G_1}{\partial r} & \frac{\partial G_1}{\partial \theta}\\ \frac{\partial G_2}{\partial r} & \frac{\partial G_2}{\partial \theta}\end{pmatrix}\begin{pmatrix}\tilde{h}\\ \tilde{k}\end{pmatrix} = \begin{pmatrix}\frac{\partial F_1}{\partial x} & \frac{\partial F_1}{\partial y}\\ \frac{\partial F_2}{\partial x} & \frac{\partial F_2}{\partial y}\end{pmatrix}\begin{pmatrix}\frac{\partial X}{\partial r} & \frac{\partial X}{\partial \theta}\\ \frac{\partial Y}{\partial r} & \frac{\partial Y}{\partial \theta}\end{pmatrix}\begin{pmatrix}\tilde{h}\\ \tilde{k}\end{pmatrix}, \tag{16.25}$$

en zien we hoe de kettingregel toegepast op

$$\mathbb{R}^2 \xrightarrow{Z} \mathbb{R}^2 \xrightarrow{F} \mathbb{R}^2$$

de Jacobimatrix van G produceert via het matrixprodukt van de Jacobimatrices van F en Z.

Deze notatie suggereert om de afhankelijke grootheid $F(x,y) = G(r,\theta)$ als 2-vector te zien, dus

$$F(x,y) = \begin{pmatrix}F_1(x,y)\\ F_2(x,y)\end{pmatrix} \quad \text{en} \quad G(r,\theta) = \begin{pmatrix}G_1(r,\theta)\\ G_2(r,\theta)\end{pmatrix},$$

en dus ook x, y en r, θ als componenten van de 2-vectoren

$$\begin{pmatrix}x\\ y\end{pmatrix} \quad \text{en} \quad \begin{pmatrix}r\\ \theta\end{pmatrix}.$$

We blijven echter $F = F(x,y)$ en $G = G(r,\theta)$ schrijven.

16.2.4 Omschrijven van differentiaaloperatoren

De notatie (16.23) 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 (16.22) en de kettingregel voor coördinatentransformaties te schrijven als

$$\begin{pmatrix} \frac{\partial}{\partial r} \\ \frac{\partial}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \frac{\partial X}{\partial r} & \frac{\partial Y}{\partial r} \\ \frac{\partial X}{\partial \theta} & \frac{\partial Y}{\partial \theta} \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix}, \tag{16.26}$$

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 16.6. 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 16.6 zijn we nog niet klaar als we in (16.15)

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix} = e_x \frac{\partial}{\partial x} + e_y \frac{\partial}{\partial y}$$

willen omschrijven naar r en θ. De vraag is ook hoe we e_x en e_y omschrijven naar e_r en e_θ, en daarvoor komt de vraag wat e_r en e_θ eigenlijk zijn.

Een natuurkundige zal hier niet lang over nadenken Teken maar een plaatje en het is evident dat

Teken plaatje!

$$e_r = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \quad \text{en} \quad e_\theta = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix},$$

en

$$\nabla = e_x \frac{\partial}{\partial x} + e_y \frac{\partial}{\partial y} = e_r \frac{\partial}{\partial r} + \frac{1}{r} e_\theta \frac{\partial}{\partial \theta} \tag{16.27}$$

de gradiënt in poolcoördinaten geeft. Daar had'ie de hele kettingregel überhaupt niet voor nodig. Omdat

$$e_r \cdot e_r = e_\theta \cdot e_\theta = 1 \quad \text{en} \quad e_r \cdot e_\theta = e_\theta \cdot e_r = 0,$$

staan de vectoren e_r en e_θ in ieder punt onderling loodrecht[5], met elk lengte 1, en wijzen in de richtingen waarin het punt $(x, y) = (r\cos\theta, r\sin\theta)$ loopt als je r respectievelijk θ varieert. De voorfactor $\frac{1}{r}$ compenseert de met r evenredige snelheid bij gelijkmatige toename van θ.

Exercise 16.7. In (16.27) staan twee representaties van dezelfde operator. Door e_x en e_y in e_r en e_θ uit te drukken en Opgave 16.6 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 16.8. Laat zien dat de divergentie in poolcoördinaten wordt gegeven door

$$\nabla \cdot V = (e_r \frac{\partial}{\partial r} + \frac{1}{r} e_\theta \frac{\partial}{\partial \theta}) \cdot (V_r e_r + V_\theta e_\theta) = \frac{\partial V_r}{\partial r} + \frac{V_r}{r} + \frac{1}{r} \frac{\partial V_\theta}{\partial \theta}.$$

Hint: Omdat e_r en e_θ van θ afhangen werkt met de produktregel van Leibniz de $e_\theta \frac{\partial}{\partial \theta}$ in de factor links nu ook op e_r en e_θ in de factor rechts, en één van die twee geeft na inprodukt met de voorfactor e_θ een bijdrage.

Exercise 16.9. Pas de regel in Opgave 16.8 nu toe op ∇ zelf en laat zien dat

$$\Delta = \nabla \cdot \nabla = \underbrace{\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}}_{\Delta_r} + \frac{1}{r^2} \underbrace{\frac{\partial^2}{\partial \theta^2}}_{\Delta_S}$$

Hint: wellicht eerst Opgave 16.8 toepassen op ∇ als werkend op de afhankelijke grootheid $G = F$, waarvoor de natuurkundige dezelfde letter gebruikt en de wiskundige

[5] Wiskundig is dit per definitie en consistent met wat je ziet als je pijltjes tekent.

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 16.9 zien we

$$\Delta = \Delta_r + \frac{1}{r^2}\Delta_S, \tag{16.28}$$

waarin Δ_r de radiële Laplaciaan is, die ook werkt op functies $R = R(r)$, en Δ_S de Laplace-Beltrami operator op

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

uitgedrukt in de hoekvariabele θ als

$$\Delta_S = \frac{\partial^2}{\partial\theta^2}.$$

Het aardige nu is dat de integraal van de Laplaciaan van een nette functie

$$u(x,y) = u(r\cos\theta, r\sin\theta)$$

over een disk B_R met straal $R > 0$ in poolcoördinaten meteen tot een belangrijke conclusie leidt, maar daarvoor moeten we eerst weten wat meervoudige integralen zijn.

16.3 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 16.9 los te laten op

$$u(x,y) = R(r)\Theta(\theta), \tag{16.29}$$

en het resultaat gelijk aan nul te stellen. Dit geeft

$$0 = (\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2})R(r)\Theta(\theta)$$

$$= \Theta(\theta)(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r})R(r) + \frac{R(r)}{r^2}\frac{\partial^2}{\partial\theta^2}\Theta(\theta)$$

$$= \Theta(\theta)(R''(r) + \frac{1}{r}R'(r)) + \frac{R(r)}{r^2}\Theta''(\theta).$$

Als Θ'' een veelvoud is van Θ, zeg

$$-\Theta'' = \mu\Theta \tag{16.30}$$

dan volgt Euler's vergelijking

$$R''(r) + \frac{1}{r}R'(r) = \mu\frac{R(r)}{r^2} \tag{16.31}$$

voor $R(r)$.

Merk op dat (16.30) gezien kan worden als een (eigenwaarde)probleem voor

$$-\Delta_S = -\frac{d^2}{d\theta}$$

op de eenheidscirkel waar bij Θ een 2π-periodieke functie moet zijn om een functie op de cirkel

$$S = \{(x, y) : x^2 + y^2 = 1\}$$

te definiëren.

Exercise 16.10. Welke μ zijn toegestaan in (16.30) voor oplossingen (16.29) 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 (16.30) te kennen. Schrijf die harmonische polynomen in gescheiden variabelen r en θ als $R(r)\Theta(\theta)$ en verifieer dat $R(r)$ een oplossing is van (16.31) met de bijbehorende μ.

Exercise 16.11. Voor elke $\mathrm{N} \in \mathbb{N}$ en $a_0, \ldots, a_\mathrm{N}$, $b_1, \ldots, b_\mathrm{N}$ 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 $(0, 0)$.

[6] In $\mathbb{R}^3$ eigenwaarden en -functies van Laplace-Beltrami operator ook via polynomen.

Opgave 16.11 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

$$u(\cos\theta, \sin\theta) = f(\theta) = \frac{a_0}{2} + \sum_{k=1}^{\infty}(a_k \cos k\theta + b_k \sin k\theta), \tag{16.32}$$

een zogenaamde *Fourierreeks*[7] voor een 2π-periodieke functie $\theta \to f(\theta)$. Ook deze $u(x,y)$ is dan in $(x,y) = (0,0)$ het gemiddelde van $u(x,y)$ op elke disk met middelpunt $(0,0)$ en straal voldoende klein, kleiner dan 1 in dit geval.

Exercise 16.12. In $\mathbb{R}^3$ gebruiken we *bolcoördinaten*

$$x = r\sin\theta\cos\phi;$$
$$y = r\sin\theta\sin\phi;$$
$$z = r\cos\theta,$$

en

$$e_r = \sin\theta\cos\phi\, e_x + \sin\theta\sin\phi\, e_y + \cos\theta\, e_z$$
$$e_\theta = \cos\theta\cos\phi\, e_x + \cos\theta\sin\phi\, e_y - \sin\theta\, e_z$$
$$e_\phi = -\sin\phi\, e_x + \cos\phi\, e_y.$$

Schrijf e_r, e_θ, e_ϕ al of niet als kolomvectoren, en verifieer dat

$$e_r \cdot e_r = e_\theta \cdot e_\theta = e_\phi \cdot e_\phi = 1; \quad e_r \cdot e_\theta = e_r \cdot e_\phi = e_\theta \cdot e_\phi = 0.$$

Overtuig jezelf van

$$\nabla = e_r \frac{\partial}{\partial r} + \frac{1}{r} e_\theta \frac{\partial}{\partial \theta} + \frac{1}{r\sin\theta} e_\phi \frac{\partial}{\partial \phi}, \tag{16.33}$$

en gebruik (16.33) om voor

$$V = V_r e_r + V_\theta e_\theta + V_\phi e_\phi$$

eerst af te leiden dat

$$\nabla \cdot V = \frac{\partial V_r}{\partial r} + \frac{2}{r} V_r + \frac{1}{r}\left(\frac{\partial V_\theta}{\partial \theta} + \frac{\cos\theta}{\sin\theta} V_\theta + \frac{1}{\sin\theta}\frac{\partial V_\phi}{\partial \phi}\right),$$

[7] Uitgebreid behandeld in de mamanotes van vorig jaar.

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 (16.9), maar nu met Δ_S gedefinieerd op

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

De formules in $\mathbb{R}^n$ laten zich nu raden, afgezien wellicht van de exacte vorm van Δ_S in de hoekvariabelen $\theta_1, \ldots, \theta_{n-1}$, maar met

$$\Delta_r = \frac{\partial^2}{\partial r^2} + \frac{n-1}{r}\frac{\partial}{\partial r}$$

voor het radiële gedeelte.

Exercise 16.13. In $\mathbb{R}^3$ gebruiken we bolcoördinaten

$$x = r\sin\theta\cos\phi;$$

$$y = r\sin\theta\sin\phi;$$

$$z = r\cos\theta,$$

en

$$e_r = \sin\theta\cos\phi\, e_x + \sin\theta\sin\phi\, e_y + \cos\theta\, e_z$$

$$e_\theta = \cos\theta\cos\phi\, e_x + \cos\theta\sin\phi\, e_y - \sin\theta\, e_z$$

$$e_\phi = -\sin\phi\, e_x + \cos\phi\, e_y.$$

Schrijf e_r, e_θ, e_ϕ al of niet als kolomvectoren en verifieer dat

$$e_r \cdot e_r = e_\theta \cdot e_\theta = e_\phi \cdot e_\phi = 1; \quad e_r \cdot e_\theta = e_r \cdot e_\phi = e_\theta \cdot e_\phi = 0.$$

Overtuig jezelf van

$$\nabla = e_r\frac{\partial}{\partial r} + \frac{1}{r}e_\theta\frac{\partial}{\partial \theta} + \frac{1}{r\sin\theta}e_\theta\frac{\partial}{\partial \phi} \tag{16.34}$$

en gebruik (16.34) 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 (16.9), maar nu met Δ_S gedefinieerd op

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

De formules in $\mathbb{R}^n$ laten zich nu raden, afgezien wellicht van de exacte vorm van Δ_S in de hoekvariabelen $\theta_1, \ldots, \theta_{n-1}$, maar met

$$\Delta_r = \frac{\partial^2}{\partial r^2} + \frac{n-1}{r}\frac{\partial}{\partial r}$$

voor het radiële gedeelte.

16.4 Intermezzo: het waterstofatoom

Met

$$V(r) = -\frac{e^2}{r}$$

is de stationaire Schrödinger vergelijking voor het waterstofatoom

$$\frac{\hbar^2}{2m}\Delta\psi - \frac{e^2}{r}\psi = E\psi, \tag{16.35}$$

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 (16.35) 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 16.14. Corrigeer eventuele typo's hierboven. De machtreeksoplossing[8]

$$u(x) = 1 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots$$

breekt af voor een n die van l afhangt. Welke n is dat?

[8]Instructief om eerst $\frac{d^2R}{dx^2} + \frac{2}{x}\frac{dR}{dx} = R$ op te lossen.

17 Functional calculus

17.1 Lijnintegralen over polygonen en Coursat

We beginnen met formule (9.8) uit Sectie 8.3, 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 $z, z(t), z_0, z_1, f(z)$,

$$t \to z(t) = (1-t)z_0 + tz_1 \tag{17.1}$$

het interval $[z_0, z_1]$ parametriseert, en

$$\int_0^1 f(z(t))z'(t)dt =$$

$$\int_0^1 f((1-t)z_0 + tz_1)dt\,(z_1 - z_0) = \int_{z_0}^{z_1} f(z)\,dz. \tag{17.2}$$

Dit is een formule die we, zonder dat (9.8) 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 (17.1), tenzij het nodig[1] is zou ik zeggen. Merk op dat $[z_0, z_1]$ voor alle $z_0, z_1 \in \mathbb{C}$ is gedefinieerd, dus $[1, 0]$ heeft nu ook betekenis. Je moet er even aan wennen maar het spreekt vanzelf. Het ligt voor de hand om aan z_0 als het begin- en z_1 als het eindpunt van $[z_0, z_1]$ te denken. Daarmee wordt $[z_0, z_1]$ meer dan alleen een verzameling: $[z_0, z_1]$ is zo een georienteerd lijnstuk.

Exercise 17.1. Laat zien dat

$$\int_{z_0}^{z_1} z\,dz = \frac{1}{2}(z_1^2 - z_0^2).$$

Evalueer vervolgens $\int_{z_0}^{z_1} z^n\,dz$ voor alle $n \in \mathbb{N}$.

[1] Quod non.

Exercise 17.2. 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 17.3. Laat zien dat

$$|\int_{z_0}^{z_1} f(z)\,dz| \leq |z_0 - z_1| \max_{z\in[z_0,z_1]} |f(z)|.$$

Integralen gedefinieerd als hierboven door (17.2) kunnen we rijgen tot een integraal over een *polygonaal pad*

$$z_0 \to z_1 \to \cdots \to z_n$$

middels

$$\int_{z_0}^{z_1} f(z)\,dz + \int_{z_1}^{z_2} f(z)\,dz + \cdots + \int_{z_{n-1}}^{z_n} f(z)\,dz = \int_{z_0,\ldots,z_n} f(z)\,dz, \quad (17.3)$$

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 17.4. Ga dit na. Hint: neem eerst $z_0 = z_2, z_1 \in \mathbb{R}$ om te zien hoe het moet werken.

Wat deze opgave zegt is dat heen en weer lopen geen bijdrage aan een keten als in (17.3) geeft omdat vrijwel per definitie

$$\int_{z_0,z_1,z_0} f(z)\,dz = \int_{z_0}^{z_1} f(z)\,dz + \int_{z_1}^{z_0} f(z)\,dz = 0,$$

voor *iedere* $[z_0, z_1] \xrightarrow{f} \mathbb{C}$ continu. We kunnen integralen dus vereenvoudigen door heen en weer stukjes weg te laten, ook als die niet achter elkaar zitten in het polygonale pad.

[2] De eersten zullen de laatsten zijn.

Neem nu in je complexe vlak z_0, z_1, z_2, not all on a line[3], en neem aan dat

$$\Delta = \Delta_{z_0,z_1,z_2} = \{t_0 z_0 + t_1 z_1 + t_2 z_2 : t_0, t_1, t_2 \geq 0, t_0 + t_1 + t_2 = 1\} \xrightarrow{f} \mathbb{C}$$

continu is. De verzameling Δ bestaat dus uit alle convexe combinaties van z_0, z_1, z_2 gezien als punten in het complexe vlak. Dat is een driehoekige tegel waarvan de rand een driehoek[4] is.

Teken plaatje!

Laat z_3, z_4, z_5 de middens zijn van $[z_0, z_1]$, $[z_1, z_2]$, $[z_2, z_0]$, die je krijgt door achtereenvolgens t_2, t_0, t_1 nul, en steeds de andere twee t-tjes $\frac{1}{2}$ te kiezen. Dan is

$$\int_{z_0,z_1,z_2,z_0} f(z)\,dz = \int_{z_0,z_3,z_5,z_4,z_3,z_1,z_4,z_2,z_5,z_0} f(z)\,dz =$$

$$\int_{z_0,z_3,z_5,z_0} f(z)\,dz + \int_{z_3,z_4,z_5,z_3} f(z)\,dz + \int_{z_3,z_1,z_4,z_3} f(z)\,dz + \int_{z_5,z_4,z_2,z_5} f(z)\,dz.$$

De integraal over het gesloten[5] pad

$$z_0 \to z_3 \to z_5 \to z_4 \to z_3 \to z_1 \to z_4 \to z_2 \to z_5 \to z_0$$

is zo enerzijds gelijk aan de integraal over het gesloten pad

$$z_0 \to z_1 \to z_2 \to z_0$$

rond de grote driehoek, en anderzijds de som van vier integralen over de vier gesloten paden

$$z_0 \to z_3 \to z_5 \to z_0, \quad z_3 \to z_4 \to z_5 \to z_3,$$

$$z_3 \to z_1 \to z_4 \to z_3, \quad z_5 \to z_4 \to z_2 \to z_5$$

rond de vier kleinere driehoeken.

Als de oorspronkelijke integraal niet nul was, zeg gelijk[6] aan 1, dan is tenminste één van de vier integralen in absolute waarde minstens gelijk aan $\frac{1}{4}$, en kan daarna met die integraal het argument herhaald worden om een rij geneste driehoeken

$$\Delta = \Delta_{z_0,z_1,z_2} \supset \Delta_{z_0^{(1)},z_1^{(1)},z_2^{(1)}} \supset \Delta_{z_0^{(2)},z_1^{(2)},z_2^{(2)}} \supset \Delta_{z_0^{(3)},z_1^{(3)},z_2^{(3)}} \supset \cdots$$

te maken met

$$\left|\int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} f(z)\,dz\right| \geq \frac{1}{4^k}$$

voor $k = 0, 1, 2, 3, \ldots$.

[3] Sommigen horen hierbij de stem van Erdös.

[4] Vereniging van 3 lijnstukjes: $[z_0, z_1] \cup [z_1, z_2] \cup [z_2, z_0]$, met zo gewenst een orientatie.

[5] Teken het in je plaatje.

[6] Door $f(z)$ te delen door zijn integraal over de rand van Δ_{z_0,z_1,z_2}.

Exercise 17.5. Bewijs dat de rijen $z_0^{(k)}, z_1^{(k)}, z_2^{(k)}$ convergeren naar een limiet in Δ_{z_0,z_1,z_2} als $k \to \infty$.

Zonder beperking der algemeenheid mogen we nu wel aannemen[7] dat deze limiet gelijk is aan 0, i.e.

$$z_0^{(k)}, z_1^{(k)}, z_2^{(k)} \to 0.$$

Kan het zo zijn dat f complex differentieerbaar is in 0? Zo ja, dan geldt voor $z \in \Delta$ dat

$$f(z) = f'(0)z + R(z)$$

met $R(z) = o(|z|)$ als $|z| \to 0$.

Maar dan is

$$\int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} f(z)\,dz = \int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} f'(0)z\,dz + \int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} R(z)\,dz$$

$$= \int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} R(z)\,dz,$$

omdat de eerste integraal nul is vanwege (17.1) toegepast op $\int_{z_0}^{z_1} z\,dz$, $\int_{z_1}^{z_2} z\,dz$, $\int_{z_2}^{z_0} z\,dz$. Dus volgt

$$\frac{1}{4^k} \leq |\int_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} R(z)\,dz| \leq \frac{|z_0 - z_1| + |z_1 - z_2| + |z_2 - z_0|}{2^k} \max_{z \in \delta\Delta^{(k)}} |R(z)|,$$

waarin $\delta\Delta^{(k)}$ de rand is van

$$\Delta^{(k)} = \Delta_{z_0^{(k)},z_1^{(k)},z_2^{(k)},z_0^{(k)}} \ni 0.$$

Omdat $|z|$ op zijn hoogste gelijk is aan de grootste afstand tussen twee punten in $\Delta^{(k)}$ geldt

$$|z| \leq \frac{d}{2^k} \quad \text{met} \quad d = \max_{z,w \in \Delta} |z - w|,$$

en samen met de 2^k die we al hadden krijgen nu ook een bovengrens voor de integraal, met een 4^k in de noemer en een $\varepsilon > 0$ naar keuze in de teller:

Exercise 17.6. Gebruik (de definitie van) $|R(z)| = o(|z|)$ als $|z| \to 0$ om met de laatste twee ongelijkheden een tegenspraak te forceren.

[7] Schuif de boel anders op.

Theorem 17.7. *Voor iedere $f : \Delta_{z_0,z_1,z_2} \to \mathbb{C}$ die complex differentieerbaar is op de gesloten*[8] *driehoek Δ_{z_0,z_1,z_2} met hoekpunten $z_0, z_1, z_2 \in \mathbb{C}$ geldt dat*

$$\oint_{z_{012}} f(z)\,dz = \int_{z_0,z_1,z_2,z_0} f(z)\,dz = 0,$$

nu ook met een notatie[9] *die wellicht hierboven al voor de hand lag.*

Eenzelfde uitspraak geldt voor iedere n-hoek ($n > 3$) gemaakt uit n driehoeken

$$\Delta_{w_0,z_0,z_1}, \Delta_{w_0,z_1,z_2}, \ldots, \Delta_{w_0,z_{n-1},z_0},$$

met

$$\Delta_{w_0,z_0,z_1} \cap \Delta_{w_0,z_1,z_2} = [w_0, z_1], \quad \Delta_{w_0,z_1,z_2} \cap \Delta_{w_0,z_2,z_3} = [w_0, z_2],$$

$$\ldots$$

$$\Delta_{w_0,z_{n-2},z_{n-1}} \cap \Delta_{w_0,z_{n-1},z_n} = [w_0, z_{n-1}], \quad \Delta_{w_0,z_{n-1},z_0} \cap \Delta_{w_0,z_0,z_1} = [w_0, z_0].$$

Als de lijnstukjes waarmee (17.3) is gemaakt de zijden zijn van zo'n door

$$z_n = z_0, \ldots, z_{n-1} \tag{17.4}$$

gemaakte n-hoek met een punt w_0 in de n-hoek waarvoor alle $[w_0, z_k]$ in de veelhoek liggen[10], dan is

$$\oint_{z_0,\ldots,z_n} f(z)\,dz = \sum_{k=1}^{n} \oint_{w_0,z_{k-1},z_k,w_0} f(z)\,dz.$$

Als $z \to f(z)$ complex differentieerbaar is in elk punt van

$$P_{z_0,\ldots,z_{n-1}} = \cup_{k=1}^{n} \Delta_{w_0,z_{k-1},z_k},$$

dan volgt zo dat

$$\oint_{z_n=z_0,\ldots,z_n} f(z)\,dz = 0. \tag{17.5}$$

Bovendien kan ieder punt z_k dan met de andere punten vastgehouden naar binnen geschoven worden naar een $\tilde{z}_k$ in de door (17.4) gemaakte veelhoek, waarbij (17.5) niet verandert met eenzelfde argument waarin driehoeken met

[8] Voor w op de rand $f(z) = f(w) + f'(w)(z-w) + R(z;w)$ met $z \in \Delta_{z_0,z_1,z_2}$ lezen.
[9] Zie het rondje door de integraalslang heen maar als driehoekje.
[10] De veelhoek is dan convex.

hoekpunten $z_{k-1}, z_k, \tilde{z}_k, z_{k+1}$ voorkomen. Kortom, (17.5) 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 (17.4), ook als de n-hoek niet gemaakt is zoals boven (17.4) 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 (17.5) gelijk is aan de som van de integralen over de randen van kleine driehoekjes

$$\Delta^{(j)} = \Delta_{z_0^{(j)},z_1^{(j)},z_2^{(j)},z_0^{(j)}}$$

in $P_{z_0,\ldots,z_{n-1}}$ waar ζ_j echt in ligt, als die allemaal maar met dezelfde orientatie genomen worden. In dat geval is dus

$$\oint_{z_n=z_0,\ldots,z_n} f(z)\,dz = \sum_{j=1}^{p} \oint_{z_2^{(j)}=z_0^{(j)},z_1^{(j)},z_2^{(j)},} f(z)\,dz, \tag{17.6}$$

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 (17.6) een verrassend simpele vorm krijgen.

17.2 Machtreeksen via een Cauchy integraalformule

Neem nu aan dat

$$D = \{z \in \mathbb{C} : |z| < 1\} \xrightarrow{f} \mathbb{C}$$

een complex differentieerbare functie is op de open eenheidsdisk. Neem een $\zeta \in \mathbb{C}$ met $|\zeta| = \rho < 1$. We laten nu rechtstreeks zien dat

$$f(\zeta) = \sum_{n=0}^{\infty} a_n \zeta^n,$$

met integraalformules voor de coëfficiënten $a_n \in \mathbb{C}$. De integralen zijn daarbij over regelmatige polygonen in D, met hoekpunten dicht bij de cirkelvormige[11] rand van D.

We leiden de formules of door Stelling 17.7 toe te passen op integralen van

$$z \to \frac{f(z) - f(\zeta)}{z - \zeta}$$

[11] Denk aan deze contouren: http://www.fi.uu.nl/publicaties/literatuur/7244.pdf

over geschikt gekozen driehoeken. Voor iedere $r \in (0,1)$ en $n \in \mathbb{N}$ met $n \geq 3$ definiëren de punten

$$z_k = r \exp(2\pi i \frac{k}{n})$$

de hoekpunten van een regelmatig n-hoek $C_{r,n}$ in D met middelpunt 0. Maak een plaatje met $0 < |\zeta| < r$ en $n = 8$ of zo. Laat k van 0 tot en met n lopen om het kringetje rond[12] te maken.

Teken plaatje!

Op dezelfde manier maken

$$w_k = \zeta + \rho \exp(2\pi i \frac{k}{n})$$

de hoekpunten van een regelmatig n-hoek $\zeta + C_{\rho,n}$, met middelpunt ζ dat binnen $C_{r,n}$ ligt als $\rho < r - |\zeta|$. Teken beide polygonen in je plaatje en merk op dat voor iedere complex differentieerbare functie

$$\{z \in D : z \neq \zeta\} \xrightarrow{g} \mathbb{C}$$

nu geldt dat

$$\oint_{z_{0-n}} g(z)\,dz =$$

$$\int_{z_0,z_1,\ldots,z_n=z_0} g(z)\,dz = \int_{w_0,w_1,\ldots,w_n=w_0} g(z)\,dz \tag{17.7}$$

$$= \oint_{w_{0-n}} g(z)\,dz,$$

voor elke $n \geq 3$, waarbij we ook hier de voor de hand liggende notatie[13] met $\oint$ gebruiken voor de integralen van $g(z)$ over de linksom[14] doorlopen n-hoeken.

Exercise 17.8. Bewijs (17.7) 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 Coursat toe te passen op in totaal $2n$ driehoekjes.

[12] Nou ja, rond...

[13] Voor grote n is de veelhoek bijna een rondje.

[14] Dat is maar een woord hier, de punten bepalen de richting.

Nu nemen we voor $g(z)$ het differentiaalquotiënt

$$\frac{f(z)-f(\zeta)}{z-\zeta}$$

en concluderen dat

$$\oint_{w_{0-n}} \frac{f(z)-f(\zeta)}{z-\zeta}\,dz = \oint_{z_{0-n}} \frac{f(z)-f(\zeta)}{z-\zeta}\,dz. \tag{17.8}$$

Exercise 17.9. De integraal in het linkerlid van (17.8) hangt van ρ af. Gebruik de differentieerbaarheid van f in ζ om te laten zien dat

$$\oint_{w_{0-n}} \frac{f(z)-f(\zeta)}{z-\zeta}\,dz \to 0$$

als $\rho \to 0$.

Maar de integraal in het linkerlid van (17.8) is gelijk aan de integraal in het rechterlid en hing niet van ρ af als $\rho < r - |\zeta|$. Hij ging[15] dus niet naar 0 want hij was al 0. Zo reduceert (17.8) tot

$$0 = \oint_{z_{0-n}} \frac{f(z)}{z-\zeta}\,dz - \oint_{z_{0-n}} \frac{f(\zeta)}{z-\zeta}\,dz,$$

en volgt

$$f(\zeta) \oint_{z_{0-n}} \frac{1}{z-\zeta}\,dz = \oint_{z_{0-n}} \frac{f(z)}{z-\zeta}\,dz.$$

Exercise 17.10. 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 17.8 en de tweede integraal hangt niet van ζ of ρ af. In het linkerlid kan dus $\zeta = 0$ genomen worden. Op elke $[z_{k-1}, z_k]$ heeft $\frac{1}{z}$ een primitieve: de meerwaardige functie gedefinieerd in Opgave 12.9 die je als het goed is in Opgave 17.2 als laatste hebt gebruikt.

[15] Letterlijk gesproken.

Theorem 17.11. *Als ζ ligt binnen een n-hoek zoals hierboven in de open eenheidsdisk D, en $f : D \to \mathbb{C}$ complex differentieerbaar is, dan is*

$$f(\zeta) = \frac{1}{2\pi i} \oint_{z_{0-n}} \frac{f(z)}{z - \zeta}\, dz.$$

This is the Cauchy Integral Formula, maar dan met n-hoeken in plaats van de gebruikelijke cirkels met middelpunt 0 en straal $r < 1$ groot genoeg.

Tenslotte volgt na het invullen van[16] de meetkundige reeksontwikkeling

$$\frac{1}{z - \zeta} = \frac{1}{z} \frac{1}{1 - \frac{\zeta}{z}} = \frac{1}{z} + \frac{\zeta}{z^2} + \frac{\zeta^2}{z^3} + \frac{\zeta^3}{z^4} + \cdots$$

de machtreeksontwikkeling in de vorm als aangekondigd, via

$$f(\zeta) = \frac{1}{2\pi i} \oint_{z_{0-n}} f(z)(\frac{1}{z} + \frac{\zeta}{z^2} + \frac{\zeta^2}{z^3} + \frac{\zeta^3}{z^4} + \cdots)\, dz =$$

$$\frac{1}{2\pi i} \oint_{z_{0-n}} \frac{f(z)}{z}\, dz + \frac{1}{2\pi i} \oint_{z_{0-n}} \frac{f(z)}{z^2}\, dz\, \zeta + \frac{1}{2\pi i} \oint_{z_{0-n}} \frac{f(z)}{z^3}\, dz\, \zeta^2 + \cdots,$$

met

$$a_j = \frac{1}{2\pi i} \oint_{z_{0-n}} \frac{f(z)}{z^{j+1}}\, dz \tag{17.9}$$

voor alle $j \in \mathbb{N}_0$.

Theorem 17.12. *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 (17.9), waarin de integraal nu door $z_k = r \exp(2\pi i \frac{k}{n})$ $(k = 0, 1, \ldots, n)$ wordt gedefinieerd. En achteraf zijn dan zowel $r \in (0, 1)$ als $n \geq 3$ arbitrair.

Het rechtvaardigen van het verwisselen van $\oint$ en $\sum$ is wezen niets anders dan opmerken dat voor elke α en β in $\mathbb{C}$ de verzameling

$$\{f : [\alpha, \beta] \to \mathbb{C} : \ f \text{ is continu}\}$$

een (complexe) Banachruimte is, net zoals $C([a, b])$ een reële Banachruimte is. Het wordt dus tijd voor het volgende hoofdstuk.

[16] We prefereren weer de notatie met puntjes natuurlijk.

Voor hier is het nog de vraag of we de limiet $n \to \infty$ willen nemen in Stelling 17.11 en (17.9) teneinde de $\oint$ te nemen over de cirkel geparametriseerd door

$$z = r\exp(i\theta) \quad \text{met} \quad dz = ir\exp(i\theta)d\theta \tag{17.10}$$

und so weiter.

Dat laatste kan komen na de observatie dat voor $0 \leq \rho < R$ en complex differentieerbare

$$A_{\rho,R} = \{z \in \mathbb{C} : \rho < |z| < R\} \xrightarrow{f} \mathbb{C}$$

geldt dat $f(z)$ te schrijven is als een zogenaamde *Laurentreeks*, i.e.

$$f(z) = \sum_{j=-\infty}^{\infty} a_j z^j = \sum_{j=0}^{\infty} a_j z^j + \sum_{j=1}^{\infty} \frac{a_{-j}}{z^j}, \tag{17.11}$$

met a_j gegeven door (17.9) voor *alle* $j \in \mathbb{Z}$, maar $n \geq 3$ en $r \in (\rho, R)$ wel zo gekozen dat met de punten $z_k = r\exp(2\pi i\frac{k}{n})$ de n-hoek in de annulus $A_{\rho,R}$ ligt.

Je bewijst dit met drie veelhoeken in $A_{\rho,R}$, waar ζ dan tussen twee van de drie in moet liggen, en in de kleinste. Als je eenmaal op het idee[17] bent gekomen wijst het zich vanzelf. De integraal over de nieuwe veelhoek wordt ook weer via een net iets andere meetkundige reeks in een machtreeks vertaald, nu met $\frac{1}{\zeta}$ die naar buiten gehaald wordt uit $\frac{1}{z-\zeta}$.

Deze zo verkregen spectaculaire uitspraak wordt gewoonlijk bewezen *na* het invoeren van lijnintegralen over echte krommen[18] zoals gegeven door (17.10) en de hele machinerie die nodig is om netjes te beschrijven wat krommen[19] eigenlijk zijn, waarbij vaak ook de continuiteit van $z \to f'(z)$ wordt gebruikt om de nog te bespreken stellingen van Green te kunnen gebruiken.

Die laatste stellingen zijn dus hier niet nodig. En de veelhoeken bieden veel meer mogelijkheden. Bijvoorbeeld voor functies die gedefinieerd en complex differentieerbaar zijn op gebieden ingesloten door twee geneste veelhoeken. Dat is misschien nog leuk om uit te zoeken.

Exercise 17.13. Natuurlijk geldt Stelling 17.12 niet alleen voor de eenheidsdisk, en kan r zowel zo klein als zo groot mogelijk gekozen worden voor het polygon waarmee de coefficienten worden berekend. Gebruik dit om te bewijzen dat er geen niet-constante begrensde complex differentieerbare functies $f : \mathbb{C} \to \mathbb{C}$ zijn.

[17] Gauss en Cauchy gingen ons voor.....

[18] Denk aan die goal van nummer 14 in het Zuiderpark en de enige echte Kromme.

[19] Denk ook aan hoe Murre dit woord uitspreekt.

17.3 De Cauchy Integraal Transformatie

De formule in Stelling 17.11 schrijven we met $1 = I$ en $\zeta = A$ als

$$f(A) = \frac{1}{2\pi i} \oint_{z_{0-n}} f(z)(zI - A)^{-1}\, dz, \tag{17.12}$$

nu voor een willekeurig polygon waar A binnen ligt en waarop

$$z \to (zI - A)^{-1} \tag{17.13}$$

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 (17.12) 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 17.5.

De vraag is natuurlijk wel eerst wat we precies onder *A binnen het polygon gedefinieerd door*

$$z_0 \to z_1 \to \cdots \to z_n = z_0$$

moeten verstaan, als we (17.12) zomaar overschrijven met ζ vervangen door een lineaire operator $A : X \to X$. Voor de hand ligt dat A zo moet zijn dat met een groter polygon de zigzagtruc weer werkt, en de integrand als $L(X)$-waardige functie complex differentieerbaar is op het gebied tussen de twee polygonen, en ook op de twee polygonen zelf, en dat weer voor ieder groter polygon.

[20] Via een zigzagintegraal als in Opgave 17.8 volgt de geldigheid van (17.12).

[21] Een Banachalgebra zelfs, zie Charlie's teleurstelling in Flowers for Algernon.

Daartoe moeten X en ook $L(X)$ zelf eerst complex uitgebreid worden, hetgeen abstract een constructie vereist maar in voorbeelden automatisch[22] gaat. En daarna is dan de natuurlijke eis dat (17.13) op het polygon en zijn buitengebied moet bestaan in de complexe versie van $L(X)$. Lees wat dit betreft verder in Sectie 17.5.

17.4 Kromme lijnintegralen

Met behulp van (17.2) is in (17.3)

$$\int_{z_0,\ldots,z_n} f(z)\,dz = \sum_{k=1}^{n} \int_{z_{k-1}}^{z_k} f(z)\,dz \tag{17.14}$$

gedefinieerd voor een rij punten die we (nog) niet als partitie zien, waarvoor we ook (nog) niet Riemanntussensommen als

$$\sum_{k=1}^{n} f(\zeta_k)(z_k - z_{k-1}) \quad \text{met} \quad \zeta_k \in [z_{k-1}, z_k] \tag{17.15}$$

hebben ingevoerd, zie Sectie 6.5. Maar als de "incrementen" $z_k - z_{k-1}$ klein zijn ligt gezien (17.2) iedere term in (17.15) voor de hand als benadering voor de overeenkomstige term in het rechterlid van (17.14) via

$$\int_{z_{k-1}}^{z_k} f(z)\,dz = \int_0^1 f((1-t)z_{k-1} + tz_k)dt\,(z_k - z_{k-1}) \approx f(\zeta_k)(z_k - z_{k-1}).$$

De vraag wat er gebeurt als $n \to \infty$ is echter nog niet goed gesteld, want de rij "partities" kan in principe willekeurig zijn.

In iedere schatting die het limietgedrag onder controle moet krijgen zal, behalve het klein worden van de incrementen, ook het gedrag van

$$\sum_{k=1}^{n} |z_k - z_{k-1}|$$

een rol spelen, met

$$z_k = z_k^{(n)}$$

zinvol afhankelijk van n gekozen, maar wat is zinvol? Hieronder wat overwegingen en een aanzet tot een uitgewerkt antwoord.

[22] Denk hier even over na.

Het stuksgewijs lineaire pad P_n van $z_0^{(n)}$ via $z_1^{(n)}, \ldots, z_{n-1}^{(n)}$, naar $z_n^{(n)}$ voor $n \to \infty$ moet een nog te formuleren limietgedrag hebben, waarmee in ieder geval voor continue $z \to f(z)$ volgt dat

$$\int_{P_n} f(z)\,dz = \sum_{k=1}^{n} \int_{z_{k-1}^{(n)}}^{z_k^{(n)}} f(z)\,dz \quad \text{en} \quad \sum_{k=1}^{n} f(\zeta_k^{(n)})(z_k^{(n)} - z_{k-1}^{(n)}) \tag{17.16}$$

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

$$\mu_n = \max_{k=1,\ldots,n} |z_k^{(n)} - z_{k-1}^{(n)}| \to 0 \tag{17.17}$$

en

$$L_n = \sum_{k=1}^{n} |z_k^{(n)} - z_{k-1}^{(n)}| \quad \text{begrensd} \tag{17.18}$$

is als $n \to \infty$ voldoende om het verschil tussen de termen in (17.16) 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 (17.17) en (17.18) gelden, en

$$z_0^{(n)} = a \quad \text{en} \quad z_n^{(n)} = b \tag{17.19}$$

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[23] 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[24] dat

$$|z(t_1) - z(t_2)| \leq |t_1 - t_2|L, \tag{17.20}$$

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 (17.1) is. Een kromme waarvan de lengte nog niet gedefinieerd maar wel gelijk aan L is, als alles goed is[25], 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] Zie Sectie **??**.

[24] Wel even nagaan!

[25] 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 $z \to f(z)$.

17.5 Calculus in Banachalgebras van operatoren

Deze sectie is nog wat schetsmatig maar niettemin precies. We willen (17.12) uitwerken voor $A \in L(X)$ en schrijven met z vervangen door λ

$$f(A) = \frac{1}{2\pi i} \oint_P f(\lambda)(\lambda - A)^{-1}\, d\lambda, \tag{17.21}$$

nu voor een willekeurig polygon[26] met hoekpunten $\lambda_1, \ldots, \lambda_n = \lambda_0$, waarop en waarbuiten[27]

$$\lambda \to (\lambda - A)^{-1} = (\lambda I - A)^{-1} \tag{17.22}$$

gedefinieerd is. Het complement van het domein van (17.22) 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 (17.22) heet de resolvente van A.

Exercise 17.14. 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 (17.22) complex differentieerbaar is op $\rho(A)$. Wat is de afgeleide?

Exercise 17.15. 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 (17.22) 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 17.13.

In deze opgaven heb je niet gebruikt dat met $AB = BA = I$ en $A \in L(X)$ ook volgt dat $B \in L(X)$, een wat diepere stelling voor Banachruimten, die ook maar eens heel kort en clean moet worden uitgelegd. Dat komt nog wel een keer. Denk in het vervolg voorlopig bijvoorbeeld eerst aan $X = \mathbb{C}^2$ als complexe uitbreiding van $\mathbb{R}^2$ en A een lineaire afbeelding gegeven door een 2×2 matrix, met complexe of reële entries. In dat geval bestaat $\sigma(A)$ meestal uit 2 punten, en met twee disjuncte driehoekjes Δ_1 en Δ_2 om die punten heen kunnen we al aan de slag met ieder paar complex differentieerbare functies

$$f_1 : \Delta_1 \to \mathbb{C} \quad \text{en} \quad f_1 : \Delta_1 \to \mathbb{C}$$

[26] Of een vereniging daarvan.
[27] Wat bedoelen we daarmee?

die samen één functie

$$f : \Delta_1 \cup \Delta_2 \to \mathbb{C}$$

maken waarvan de twee stukken elkaar niet zien. Maar ook het rechterlid van (9.21) 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

$$\sigma(A) \subset K_{int} \subset K = C_0 \cap U_1 \cap \cdots \cap U_n, \tag{17.23}$$

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

$$f(A) = \frac{1}{2\pi i} \oint_{\delta K} f(\lambda)(\lambda - A)^{-1}\, d\lambda =$$

$$\frac{1}{2\pi i} \left(\oint_{P_0} f(\lambda)(\lambda - A)^{-1}\, d\lambda - \sum_{j=1}^{n} \oint_{P_j} f(\lambda)(\lambda - A)^{-1}\, d\lambda \right) \tag{17.24}$$

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

$$f(A) = \frac{1}{2\pi i} \sum_{j=1}^{m} \oint_{\delta K_j} f(\lambda)(\lambda - A)^{-1}\, d\lambda \tag{17.25}$$

met iedere term in de som gedefinieerd als in (17.24) 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[28] uitdrukkingen voor K moet daarbij graag op de koop toe worden genomen, en als $K_j \neq C_j$ kunnen de bijbehorende buitenkanten ook genest liggen. In het simpele geval dat $\sigma(A)$ een eindige discrete puntverzameling is kunnen we natuurlijk toe met $K = C_j = \Delta_j$, met de driehoekjes Δ_j zo klein als we maar willen en

$$\sigma(A) \subset K^{int} \subset K = \Delta_1 \cup \cdots \cup \Delta_m.$$

Wat we nu sowieso in alle gevallen willen is dat, als we de hoekpunten van de polygonen een beetje naar binnen schuiven, K in dus, de integralen die in (17.24) en (17.25) de nieuwe lineaire afbeelding $f(A) : X \to X$ moeten maken, niet veranderen. En hetzelfde als we K groter maken door de punten naar buiten te schuiven, zolang we maar niet uit het definitiegebied van de continu differentieerbare complexwaardige f lopen. Bij het verder kleiner of groter maken kan de structuur van K versimpelen als twee polygonen elkaar ontmoeten en vervolgens samen één polygon vormen. Strict genomen hebben we niet nodig hoe dat precies kan gaan, maar het is toch aardig om daar even over na te denken.

Exercise 17.16. 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

[28] Als we alle zijden van alle poygonen dicht bij $\sigma(A)$ willen hebben.

andere scenarios er nog zijn, zoals ondermeer polygonen tot een punt laten krimpen en verdwijnen.

Via de inmiddels vertrouwde zigzagkrommen vernandert bij het geschuif met de hoekpunten (17.25) niet, mits de Stelling van Coursat geldt voor driehoekjes waarop en waarbinnen (17.22) complex differentieerbaar is. De betreffende integralen bestaan weer uit integralen over lijnstukjes. Integralen die gedefinieerd zijn dankzij het werk in Sectie **??**, waarvoor het voorwerk al in Sectie 6.5 was gedaan: 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 Coursat met bewijs en al worden overgeschreven[29] en is (17.24) een goede definitie van $f(A)$. Voorlopig houden we nu A vast en kijken naar nog zo'n f, een g dus, waarbij we eerst aannemen dat we het allersimpelste geval hebben, één polygon rond $\sigma(A)$ waarmee de berekeningen gedaan worden. In dat geval is de samenstelling van de afbeeldingen $f(A)$ en $g(A)$ te schrijven als

$$f(A)g(A) = \frac{1}{2\pi i} \oint_{\lambda_{0-n}} f(\lambda)(\lambda - A)^{-1}\, d\lambda\, \frac{1}{2\pi i} \oint_{\mu_{0-n}} g(\mu)(\mu - A)^{-1}\, d\mu,$$

met in de Cauchyintegraal voor $g(A)$ de hoekpunten μ_l een klein beetje naar binnen geschoven hebben, niet omdat het moet, maar omdat het kan, iets minder ver naar binnen dan de hoekpunten λ_k. Het μ-polygon komt zo binnen het λ-polygon te liggen.

Omdat

$$f(A) = \frac{1}{2\pi i} \sum_{k=1}^{n} \int_{\lambda_{k-1}}^{\lambda_k} f(\lambda)(\lambda - A)^{-1}\, d\lambda,$$

$$g(A) = \frac{1}{2\pi i} \sum_{l=1}^{n} \int_{\mu_{l-1}}^{\mu_l} g(\mu)(\mu - A)^{-1}\, d\mu,$$

[29] Op detail nog te bespreken.

wordt $f(A)g(A)$ afgezien van de voorfactoren dankzij overwegingen als bij (15.2) 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

$$(fg)(A) = \frac{1}{2\pi i} \oint_{\lambda_{0-n}} f(\lambda)g(\lambda)(\lambda - A)^{-1}\, d\lambda = f(A)g(A) = g(A)f(A), \tag{17.26}$$

en daar is nog veel mee te spelen.

Exercise 17.17. Ga na dat in het algemene geval (17.25), wanneer $f(A)$ en $g(A)$ de som zijn van een eindig aantal integralen over links- dan wel rechtsom[30] 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 (17.26) 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 (17.24) en (17.25) 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 17.18. *Laat voor $A \in L(X)$ en een complexwaardige f de operator $f(A)$ gedefinieerd zijn via (17.24) en (17.25). 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 (17.25).

De driehoekjes in I_0 zijn niet relevant voor (17.25), maar iedere $\Delta \in J$ heeft 12 buren[31] waarvan er tenminste één ook in J ligt, zeg $\tilde{\Delta}$, gekarakteriseerd door

$$\delta\tilde{\Delta} \cap \delta\Delta \cap \sigma(A) \neq \emptyset,$$

en in dat geval noemen we Δ en $\tilde{\Delta}$ fijne buren in J. Twee zulke fijne buren die verder geen andere fijne buren hebben vormen samen een fijn duo verenigd in

$$\Delta \cup \tilde{\Delta},$$

[30] Lees: linksom, maar met een min voor het integraalteken.

[31] Waarvan er drie een zijde met Δ gemeen hebben en de rest alleen een hoekpunt.

en I_2 is per definitie de verzameling van zulke verder geisoleerde fijne buren, die verenigd steeds een ruit vormen, een ruit die als een K_j kan worden meegenomen in (17.25).

Een paar niet geisoleerde fijne buren kan nog 1 of meerdere fijne buren hebben, en als het maar 1 is, zeg $\hat{\Delta}$, dan kan het zijn dat die verder zelf geen fijne buren meer heeft. Dan vormen ze een fijn triootje waarbij verschillende standjes denkbaar zijn. Dit definieert de verzameling I_3, alle driehoeken Δ die onderdeel vormen van een fijn trio verenigd in

$$\Delta \cup \tilde{\Delta} \cup \hat{\Delta},$$

dat een parallelogram of een halve zeshoek is.

En zo gaat dat door met fijne quatrootjes, fijne quintootjes, etc totdat J op is, waarbij het aantal standjes flink maar niet oneindig toe kan nemen. Kortom, met I gepartioneerd als

$$I = I_0 \cup I_1 \cup I_2 \cup \cdots \cup I_p$$

is het nu nog de vraag wat de mogelijke onderlinge standjes zijn: als $\Delta_1 \in I_k$ met $k-1$ andere driehoeken in I_k een fijn k-stel vormt hoe kan de vereniging

$$\Delta_1 \cup \Delta_2 \cup \cdots \cup \Delta_k$$

er dan uitzien?

Antwoord: als een binnengebied van een polygon, of als het rechterlid van (17.23). Dat moet dus nog door iemand[32] bewezen worden, als dat niet al eens gebeurd is. Maar verder zijn we nu wel klaar met de beschrijving van $f(A)$. Dat kan altijd met eindig veel polygonen die willekeurig dicht bij $\sigma(A)$ liggen door ρ klein te kiezen. Hoe dichter bij $\sigma(A)$ hoe meer je er nodig hebt en hoe wilder de standjes kunnen worden.

We zijn nu klaar voor het bewijs van Stelling 17.18. 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

$$g(A)(\mu - f(A) = (\mu - f(A)g(A) = I,$$

[32] Ik pas, maar dat is voor even.

waarmee $\mu \in \rho(f(A))$. Dus $\sigma(f(A)) \subset f(\sigma(A))$.

Kan de inclusie strict zijn? In dat geval is er een $\mu_0 = f(\lambda_0) \in \sigma(f(A))$ waarvoor $\mu_0 - f(A)$ inverteerbaar is terwijl $\lambda_0 - A$ het niet is. Door schuiven en schalen van f, en schuiven van A en λ kunnen we zonder beperking der algemeenheid wel aannemen dat $\lambda_0 = 0 = \mu_0$ en dat de machtreeks van f begint met λ^n voor zekere $n \in \mathbb{N}$ omdat $f(0) = 0$. In dat geval is

$$f(\lambda) = \lambda^n g(\lambda) \quad \text{met} \quad g(\lambda) = 1 + b_1\lambda + b_2\lambda^2 + \cdots$$

en dus is $g(A)$ inverteerbaar, net als $f(A)$. Maar de algebra geeft

$$f(A) = A^n g(A).$$

Voor $n = 1$ is de tegenspraak onmiddellijk. Voor $n > 1$ niet helemaal. Pas daarom het argument hierboven aan en concludeer eerst dat $\mu_0 = f(\lambda_0)$ zo gekozen kan worden dat $f'(\lambda_0) \neq 0$. Hiermee is het bewijs van de stelling wel klaar. Als g een andere functie is die complex differentieerbaar is op een omgeving van $\sigma(f(A)) = f(\sigma(A))$ dan volgt ook vrij direct uit de definities dat

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

Exercise 17.19. Bewijs dit.

Nog een expliciet voorbeeld. Als

$$\lambda = \lambda \sum_{j=1}^{\mathrm{N}} \chi_j(\lambda) = \sum_{j=1}^{\mathrm{N}} \lambda\chi_j(\lambda),$$

met

$$\chi_j(\lambda) = \delta_{ij} \quad \text{voor} \quad \lambda \in K_i,$$

dan

$$I = \sum_{j=1}^{\mathrm{N}} I\chi_j.$$

Definieer de "spectraalprojecties"

$$P_j = \chi_j(A).$$

Exercise 17.20. Laat zien dat $P_iP_j = \delta_{ij}P_j$, $AP_j = P_jA$, $\sigma(AP_j) = \sigma(A) \cap K_j$,

$$I = \sum_{j=1}^{\mathrm{N}} P_j \quad \text{en} \quad A = \sum_{j=1}^{\mathrm{N}} AP_j = \sum_{j=1}^{\mathrm{N}} P_jA.$$

Zo wordt

$$X = R(P_1) \oplus \cdots \oplus R(P_n),$$

en beeld A iedere $X_i = R(P_i)$) op zich zelf af, en volgt voor

$$A_j : X_i \xrightarrow{AP_j} X_i$$

dat $\sigma(A_j) = \sigma(A) \cap K_j$.

Zo, en dat alles met een beetje lijnintegreren.

18 Multilinear algebra and integration

This chapter is at the crossroads of analysis and algebra[1]. Let us first state the main analytical result, which concerns a bounded open set $\Omega \subset \mathbb{R}^N$ with $\partial\Omega \in C^1$ and $v \in C^1(\overline{\Omega})$. We will establish that

$$\int_\Omega v_{x_i} = \int_{\partial\Omega} \nu_i \, v. \tag{18.1}$$

The integral on the left in (18.1) is a Riemann integral of the continuous partial derivative

$$v_{x_i} = \frac{\partial v}{\partial x_i} = D_i v = \partial_i v$$

of v with respect to the i^{th} variable. It is defined if $\partial\Omega$ is a set with zero N-dimensional measure. This is the case if the bounded closed set $\partial\Omega$ is the zero level set of a function $F \in C^1(\mathbb{R}^N)$ with $\nabla F(x) \neq 0$ for every[2]

$$x \in \{x \in \mathbb{R}^N : \, F(x) = 0\} = \partial\Omega.$$

If the bounded open set Ω is given by

$$\Omega = \{x \in \mathbb{R}^N : \, F(x) < 0\}, \tag{18.2}$$

the vector

$$\nu = \frac{\nabla F}{|\nabla F|_2}$$

is defined in every point of $\partial\Omega$, and thereby we have a unit vector field on $\partial\Omega$ normal[3] to $\partial\Omega$, which is then a compact C^1-hypersurface[4] in $\mathbb{R}^N$, and the integral on the right in (18.1) is the integral of the continuous function

$$x \to \nu_j(x) \, v(x)$$

over $\partial\Omega$, as defined for continuous functions on compact n-dimensional manifolds $M \subset \mathbb{R}^N$ such as $M = \partial\Omega$ in Section 19 below.

This analytical result is later restated for continuously differentiable vector valued functions

$$\Omega \ni x \xrightarrow{V} \mathbb{R}^N$$

at the beginning of Section 19.4 as the Gauss Divergence Theorem in (19.25), while it is first proved in Section 18.1 for functions v that vanish outside an

[1] Written while teaching from Edwards' book Advanced Calculus of Several Variables.
[2] Zero is then called a regular value of F, google Sard's Theorem.
[3] To the local tangent space in every point of.
[4] See Opgave 13.6.

N-dimensional block $[a, b]$ in which $\partial\Omega \cap [a, b]$ is the graph of a C^1-function as explained at the beginning of that section, with $\Omega \cap [a, b]$ on one[5] side of the graph. In the 2-dimensional case N = 2 this formulation avoids the Jordan Curve Theorem[6].

The local statement under (19.11) becomes a global statement dropping the tildes and skipping the intermediate term. It then reads

$$\int_{\partial\Omega} \nu \cdot v \; dS_1 = \iint_{\Omega} \left(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} \right) dx_1 \, dx_2, \tag{18.3}$$

which putting $v_2 = -p$ and $v_1 = q$ rewrites as (19.12), the 2-dimensional of the Stokes Curl Theorem, in which the line integral is the integral over $\partial\Omega$ of the inner product of the vectorfield with components p and q and the tangent[7] vectorfield τ with components $-\nu_2$ and ν_1. Renaming p and q as v_1 and v_2 the result

$$\int_{\partial\Omega} \tau \cdot v \; dS_1 = \iint_{\Omega} \left(\frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \right) dx_1 \, dx_2 \tag{18.4}$$

$$= \int_{\partial\Omega} v_1(x_1, x_2) dx_1 + v_2(x_1, x_2) dx_2,$$

is known as Green's Theorem, the second formulation being valid if $\partial\Omega$ is parameterised by continuously differentiable parameterisations

$$t \to x(t) = (x_1(t), x_2(t))$$

defining normalised tangent and normal vectors

$$\tau(t) = \frac{1}{\sqrt{x_1'(t)^2 + x_2'(t)^2}} \begin{pmatrix} x_1'(t) \\ x_2'(t) \end{pmatrix} \quad \text{and} \quad \nu(t) = \frac{1}{\sqrt{x_1'(t)^2 + x_2'(t)^2}} \begin{pmatrix} x_2'(t) \\ -x_1'(t) \end{pmatrix}$$

with ∇F in $x(t)$ being a positive multiple of $\nu(t)$ if Ω is defined by (18.2).

The local statement in Section 18.1 does indeed lead to the global statement via arguments which involve cut-off functions[8] and partitions of unity, which are discussed as an independent topic in Section 20. These are also needed for a proper definition of integrals over manifolds in Section 19, the other analytical tool being the transformation theorem discussed in Section 15.4 applied in Section 20.2 to the transformations in Section 20.4.

[5] Which easily leads to mistakes in the direction of ν so be careful.

[6] Which I should discuss in the context of Section 17.1 and polygons, to do.

[7] Edwards writes $\nu = \mathbf{N}$ and $\tau = \mathbf{T}$.

[8] A misnomer, I would call them fading functions.

So far for analysis. In Section 19.1 we rewrite (18.1) as stated in (18.11) in the language of the differential forms introduced in Section 15.2 and below formula (18.4) just above. This language is re-presented in the language also used by Edwards in Section 19.3. In Section 19.4 we spell out the algebra to explain how the general Stokes Theorem for integrals of differential forms then follows[9]. This general theorem then has the Stokes Curl Theorem in (19.31) for certain[10] closed closed curves in $\mathbb{R}^3$ as a reformulation without differential forms as a second most commonly occuring example.

18.1 Local integrals with the normal at the boundary

This section is influenced by all the local boundary flattening arguments in Chapter 5 on Sobolev spaces (see Chapter 23 below) in Evans' PDE book, which involve applications of the Gauss Divergence Theorem (in essence integration by parts) stated in an appendix, without proof. The flattening argument introduced here effectively puts that horse in front of the wagon again.

Let Ω be a bounded open set in $\mathbb{R}^{\mathrm{N}} = \mathbb{R}^{n+1}$ and $M = \partial\Omega$ the union of finitely many patches $P = M \cap (a, b)$, each of which, after renumbering, comes with a description of $[a, b] \cap \Omega$ as given by

$$a_{_N} \leq x_{_N} < f(x_1, \ldots, x_{_{N-1}}) < b_{_N} \tag{18.5}$$

or

$$a_{_N} < f(x_1, \ldots, x_{_{N-1}}) \leq x_{_N} < b_{_N}. \tag{18.6}$$

Then there exist finitely many patches as above such that

$$M \subset P_1 \cup \cdots \cup P_I$$

If $P_i \cap P_j \neq \emptyset$, the transformations used in Section 20.4 respect the local descriptions of Ω, which we can use to compute a unique normal ν which points out of Ω in every patch on $M = \partial\Omega$. Overlapping patches then have the same outward normal vector ν.

In the simplest[11] relevant case with $n < \mathrm{N}$ and $\mathrm{N} > 1$, which is $n = 1$ and $\mathrm{N} = 2$, and $\Omega \subset \mathbb{R}^2$ bounded and open, $M = \partial\Omega$ is a 1-dimensional compact manifold covered by finitely many graphs

$$\{(x, f_i(x)) : a_i < x < b_i\} \text{ and } \{(g_j(y), y) : c_j < y < d_j\},$$

[9] Still to be done: integrals over manifolds with boundaries.

[10] The ones obtained as the image of $\partial\Omega$ under injective C^1 maps $\Phi : \Omega \to \mathbb{R}^3$.

[11] What follows will generalise to $\mathrm{N} = n + 1 > 1$.

with $f_i \in C^1([a_i, b_i])$, $g_j \in C^1([c_j, d_j])$. The question we ask is how the integrals

$$\int_\Omega v_x(x,y)\,dxdy \quad \text{and} \quad \iint_\Omega w_y(x,y)\,dxdy \tag{18.7}$$

evaluate in terms of the values of v and w on the boundary $M = \partial\Omega$ for v and w in $C^1(\Omega)$. Here I use the common notation $dxdy = dydx$. With x_1 and x_2 replacing x and y the notation

$$\int_\Omega v_{x_1} = \int_\Omega v_{x_1}(x)\,dx = \iint_\Omega v_{x_1}(x_1,x_2)\,d(x_1,x_2)$$

(and likewise for the other integral) is more to my liking, but everybody writes $dx_1\,dx_2$ and $dx], dy$, rather than $d(x,y)$.

To answer the question we first consider a piece of the boundary described by $y = f(x)$, with $f \in C^1([a,b])$ and $c < f(x) < d$ for all $x \in [a,b]$, such that

$$\tilde\Omega = \Omega \cap ([a,b]\times[c,d]) = \{(x,y) :\, a \le x \le b,\, f(x) < y \le d\}, \tag{18.8}$$

and multiply w by a function $\zeta \in C^1(\mathbb{R}^2)$ as in Section 19 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 $\tilde w = \zeta w$ we now evaluate

$$\int_\Omega \tilde w_y = \iint_{\tilde\Omega} \tilde w_y(x,y)\,dx\,dy = \int_a^b \left(\int_{f(x)}^d \tilde w_y(x,y)dy\right) dx$$

$$= -\int_a^b \tilde w(x,f(x))\,dx$$

$$= \int_a^b \underbrace{\frac{-1}{\sqrt{1+f'(x)^2}}}_{\nu_y}\, \tilde w(x,f(x))\, \underbrace{\sqrt{1+f'(x)^2}dx}_{ds=dS_1}$$

$$= \int_\Phi \nu_y \tilde w\,ds,$$

in which we recognised the y-component of the normal vector

$$\nu = \frac{1}{\sqrt{1+f'(u)^2}}\begin{pmatrix} -f'(u) \\ 1 \end{pmatrix}$$

and

$$dS_1 = ds = |\Phi'(u)|^2\,du = \sqrt{1+f'(u)^2}\,du$$

evaluated via the parameterisation $\Phi(u) = (u, f(u))$.

For the integral of $\tilde{v} = \zeta v$ we use the new coordinates

$$\xi = x,\ \eta = y - f(x) \quad \text{whence} \quad x = \xi,\ y = \eta - f(\xi) \quad \text{and} \quad dx\,dy = d\xi\,d\eta$$

when transforming the integral over $(x, y) \in \tilde{\Omega}$ to an integral over

$$(\xi, \eta) \in D = \{(x, y - f(x)) :\ a \le x \le b,\ f(x) \le y \le d\}.$$

Defining $\phi(\xi, \eta)$ by

$$\phi(\xi, \eta) = \tilde{v}(x, y) \quad \text{we have} \quad v_x(x, y) = \phi_\xi(\xi, \eta) - f'(\xi)\phi_\eta(\xi, \eta)$$

via the chainrule, whence[12]

$$\begin{aligned}
\int_\Omega \tilde{v}_x &= \iint_{\tilde{\Omega}} \tilde{v}_x(x, y)\,dx\,dy = \iint_D \left(\phi_\xi(\xi, \eta) - f'(\xi)\phi_\eta(\xi, \eta)\right)\,d\xi\,d\eta \\
&= \int_0 \left(\int_a^b \phi_\xi(\xi, \eta)\,d\xi \right) d\eta - \int_a^b \left(\int_0 f'(\xi)\phi_\eta(\xi, \eta)\,d\eta \right) d\xi \\
&= \int_0 \left(\phi(b, \eta) - \phi(a, \eta)\right) d\eta - \int_a^b f'(\xi) \int_0 \phi_\eta(\xi, \eta)\,d\eta\,d\xi \\
&= \int_a^b f'(\xi)\phi(\xi, f(\xi))\,d\xi = \int_\Phi \nu_x \tilde{v}\,ds,
\end{aligned}$$

just as before, after inserting $\sqrt{1 + f'(\xi)^2}$ and recognising $ds = dS_1$ as well as the x-component of the normal vector ν. In conclusion we have

$$\int_\Omega \tilde{v}_x = \int_{\partial\Omega} \nu_x \tilde{v}\,dS_1 \quad \text{and} \quad \int_\Omega \tilde{w}_y = \int_{\partial\Omega} \nu_y \tilde{w}\,dS_1 \tag{18.9}$$

Now look at (19.2), and choose[13] functions $\zeta_0 \in C_c^1(\Omega)$ and $\zeta_1, \cdots, \zeta_n$, such that $0 \le \zeta_i \le 1$ for $i = 0, \ldots, n$, $\zeta_1, \ldots, \zeta_n$ as ζ above, such that

$$\zeta_0 + \zeta_1 + \cdots + \zeta_n \equiv 1 \quad \text{on} \quad \Omega, \tag{18.10}$$

to conclude, via $v = \zeta_0 v + \zeta_1 v + \cdots + \zeta_n v$ that

$$\int_\Omega v_{x_i} = \int_{\partial\Omega} \nu_i v\,dS_{_N} \tag{18.11}$$

for $v \in C^1(\Omega)$ with $M = \partial\Omega$ and ν a globally defined[14] normal vector field on M such that locally a description as above holds.

[12] We drop a conveniently chosen fixed upper bound in the η-integrals from the notation.
[13] This requires a technique we take for granted now.
[14] This is no additional assumption on Ω and its boundary.

In (18.11) we stated the formula for $\Omega \subset \mathbb{R}^N$ open and $M = \partial\Omega$ a compact $(N-1)$-dimensional manifold. The local proof involves

$$dS_{N-1} = \sqrt{1 + (\frac{\partial f}{\partial u_1})^2 + \cdots + (\frac{\partial f}{\partial u_{N-1}})^2}\; du_1 \cdots du_{N-1}$$

and the normal vector ν with

$$\nu_1 = \frac{1}{\sqrt{1+|\nabla f(u)|^2}}\,\frac{\partial f}{\partial u_1}\,, \ldots,\, \nu_N = \frac{1}{\sqrt{1+|\nabla f(u)|^2}}\,\frac{\partial f}{\partial u_{N-1}},$$

$$\nu_N = \nu_N = \frac{-1}{\sqrt{1+|\nabla f(u)|^2}}.$$

Formula (18.11) is often called Green's Theorem. Applying it to the product of v and some other function $\zeta \in C^1(\Omega)$ we obtain the integration by parts formula

$$\int_\Omega v_{x_i}\zeta = \int_{\partial\Omega} \zeta v\, \nu_i dS_{N-1} - \int_\Omega \zeta_{x_i} v. \tag{18.12}$$

18.2 The length of a curve

In the 1-dimensional case I now follow Edwards and write $x = \gamma(t)$ with $t \in [a,b]$ and $\gamma : [a,b] \to \mathbb{R}^N$. For any such γ the natural definition of the length would be the smallest upper bound on the set of numbers obtained via

$$\sum_{j=1}^{m} |\gamma(t_j) - \gamma(t_{j-1})|_2$$

with

$$a = t_0 < t_1 < \cdots < t_m = b.$$

Clearly this definition of length is invariant under reparameterisation of γ via strictly monotone bijections $\phi : [a,b] \to [c,d]$ as in Section 4.3. It's not a very hard exercise to show that for continuously differentiable $\gamma : [a,b] \to \mathbb{R}^N$ the length is given by

$$s(\gamma) = \int_a^b |\gamma'(t)|_2\, dt,$$

and the change of variables formula applied to $u = \phi(t)$ with $\phi \in C^1([a,b])$ with $\phi'(t) \neq 0$ confirms that

$$\Phi :\, u \to \gamma(\phi^{-1}(u)) \tag{18.13}$$

is in $C^1([c,d])$ with $[c,d] = \phi([a,b])$, and has the same length[15]. Also, if $f = f(x)$ is continuous on $\gamma([a,b]) = \Phi([c,d])$, it follows that

$$\int_\gamma f = \int_\gamma f ds = \int_a^b f(\gamma(t))|\gamma'(t)|_2 \, dt = \int_c^d f(\Phi(u))|\Phi'(u)|_2 \, du = \int_\Phi f ds. \tag{18.14}$$

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.

18.3 Line integrals of vector fields along curves

Besides (18.14) as a 1-dimensional example of what is to come in (18.33) we can also define an integral for $F = F(x) \in \mathbb{R}^n$ continuous on $\gamma([a,b])$, namely

$$\int_\gamma F \cdot ds = \int_a^b F(\gamma(t)) \cdot \gamma'(t) \, dt = \int_a^b F(\gamma(t)) \cdot \underbrace{\frac{\gamma'(t)}{|\gamma'(t)|_2}}_{T(t)} |\gamma'(t)|_2 \, dt, \tag{18.15}$$

but Edwards avoids the commonly used notation in the left hand side of (18.15), and instead writes

$$\int_\gamma F \cdot T \, ds,$$

with T the unit tangent vector[16] defined by

$$T(t) = \frac{\gamma'(t)}{|\gamma'(t)|_2}.$$

For reparametrisations $u = \phi(t)$ with $\phi \in C^1([a,b])$ and $\phi'(t) > 0$ and Φ defined as in (18.13) above you easily verify that the work

$$W = \int_\gamma F \cdot ds = \int_\gamma F \cdot T \, ds = \int_\Phi F \cdot T \, ds = \int_\Phi F \cdot ds.$$

[15] The condition that $\gamma'(t) \neq 0$ also carries over to $\Phi'(u) \neq 0$.

[16] 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

$$W = \int_\gamma F \cdot ds = \int_\gamma F \cdot T\, ds = \int_a^b \left(F_1(\gamma(t))\gamma_1'(t) + \cdots + F_{_N}(\gamma(t))\gamma_{_N}{}'(t)\right) dt$$

$$= \int_a^b F_1(\gamma(t))\underbrace{\gamma_1'(t)\, dt}_{dx_1} + \cdots + \int_a^b F_{_N}(\gamma(t))\underbrace{\gamma_{_N}{}'(t)\, dt}_{dx_{_N}}$$

leads to the notational convention

$$\int_\gamma F \cdot ds = \int_\gamma F_1 dx_1 + \cdots + \int_\gamma F_{_N} dx_{_N} = \int_\gamma F_1 dx_1 + \cdots + F_{_N} dx_{_N}. \quad (18.16)$$

If $F = \nabla f$ it is then 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 (9.7), after which d was seen[17] as acting on f to produce $df = f'(x)dx$. Here we have d acting on f as

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_{_N}} dx_{_N}, \quad (18.17)$$

producing what is called a 1-form in Section 1 of Chapter V in Edwards.

These 1-forms act on vectors. Whereas the x-dependent vector

$$F(x) = F_1(x)e_1 + \cdots + F_{_N}(x)e_{_N} \quad (18.18)$$

and the vector

$$v = v_1 e_1 + \cdots + v_{_N} e_{_N}$$

have and x-dependent inner product

$$F(x) \cdot v = F_1(x)v_1 + \cdots + F_{_N}(x)v_{_N},$$

the 1-form

$$\omega = F_1(x)dx_1 + \cdots + F_{_N}(x)dx_{_N} \quad (18.19)$$

[17] Writing f instead of F again.

assigns to the same vector v the same x-dependent scalar

$$F_1(x)v_1 + \cdots + F_{\rm N}(x)v_{\rm N},$$

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

$$\int_a^b (F_1(\gamma(t))\gamma_1'(t) + \cdots + F_{\rm N}(\gamma(t))\gamma_{\rm N}'(t))\,dt = \int_\gamma \omega.$$

Thus, ω evaluated in $x = \gamma(t)$ acts on $\gamma'(t)$ and is integrated from $t = a$ to $t = b$ to define $\int_\gamma \omega$. Note a reparameterisation of γ with $u = \phi(t)$ and $\phi'(t) < 0$ changes the sign of the integral.

The notation for ω hides the x-dependence and may look like the abuse of notation in $f = f(x)$, but (13) and (14) in Section 1 of Chapter V in Edwards are mathematically precise[18]. 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 $\omega = F_1(x)dx_1 + \cdots + F(x)dx_{\rm N}$.

18.4 Surface area

We need some linear algebra for integrals over more general surface patches then the ones encountered in Section 18.1 , a surface patch being a set in $\mathbb{R}^3$ parameterised by a continuously differentiable injective map

$$\Phi : [0,1] \times [0,1] \to \mathbb{R}^3, \tag{18.20}$$

with

$$\nabla\Phi = (\nabla\Phi_1\ \nabla\Phi_2\ \nabla\Phi_3) = \begin{pmatrix} \frac{\partial\Phi_1}{\partial u_1} & \frac{\partial\Phi_2}{\partial u_1} & \frac{\partial\Phi_3}{\partial u_1} \\ \frac{\partial\Phi_1}{\partial u_2} & \frac{\partial\Phi_2}{\partial u_2} & \frac{\partial\Phi_3}{\partial u_2} \end{pmatrix}$$

denoting the matrix of which the columns are the gradients of the N = 3 components Φ_1, Φ_2, Φ_3 of Φ with respect to the $n = 2$ variables[19] u_1, u_2 in $\Phi = \Phi(u) = \Phi(u_1, u_2)$, consistent with the notation in Section (12.6).

Momentarily switching to a notation with Φ_1, Φ_2, Φ_3 as functions of u, v, $\nabla\Phi$ is the transpose of the Jacobian matrix

$$(\ \frac{\partial\Phi}{\partial u}\ \frac{\partial\Phi}{\partial v}\ ),$$

which has column vectors $\frac{\partial\Phi}{\partial u}$, $\frac{\partial\Phi}{\partial v}$.

[18] We also write $\int_a^b f$ for the integral of a function.

[19] Everything that follows should generalise or trivialise to $1 \le n \le$ N.

In the special linear case with

$$\Phi_i(u,v) = a_i u + b_i v \tag{18.21}$$

the Jacobian matrix is the transpose of

$$\nabla\Phi = A = \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix},$$

the matrix example in (18.30) starting the discussion in Section 18.7 below on

the area $\mathcal{M}_2(a,b)$ of a parallelogram spanned by two vectors a and b

with entries a_1, a_2, a_3 and b_1, b_2, b_3 respectively. This parallelogram is then the image of $[0,1] \times [0,1]$ under Φ defined by (18.21), and its area is then equal to

$$\int_0^1 \int_0^1 \mathcal{M}_2(\frac{\partial\Phi}{\partial u}, \frac{\partial\Phi}{\partial v})\, dudv, \tag{18.22}$$

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 (18.21).

It will be no surprise that (18.22) will also be used to define the area of the surface patch defined by Φ if Φ is not a linear map from $[0,1]^2$ to $\mathbb{R}^3$, and that everything generalises to $\Phi : [0,1]^n \to \mathbb{R}^{\mathrm{N}}$ with $1 \leq n < \mathrm{N}$. We expand on the linear case of this generalisation next.

18.5 Transpose, quadratic forms and operator norms

In (12.28) 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

$$S = AA^T \in L(\mathbb{R}^m, \mathbb{R}^m) \quad \text{with entries} \quad s_{ik} = \sum_{j=1}^{n} a_{ij} a_{kj} = s_{ki}. \tag{18.23}$$

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

$$|A^T|^2_{op} = \max_{|z|_2=1} \underbrace{|A^T z|^2_2}_{A^T z \cdot A^T z} = \max_{|z|_2=1} AA^T z \cdot z = \max_{|z|_2=1} Sz \cdot z = \max_{0 \neq z \in \mathbb{R}^m} \frac{Sz \cdot z}{z \cdot z}, \tag{18.24}$$

and we note that the bilinear mapping

$$(z, w) \to Sz \cdot w$$

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 18.1. 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

$$S(z + w) \cdot (z + w) + S(z - w) \cdot (z - w) = 2\,Sz \cdot z + 2\,Sw \cdot w, \tag{18.25}$$

an identity to play with, with $S = AA^T$as above, but also with $S = I$ the identity:

Exercise 18.2. The Cauchy-Schwarz inequality and the definition of the operator norm in Section ?? immediately imply that $M \leq |S|_{op}$. Write

$$4\, Sz \cdot w = S(z+w) \cdot (z+w) - S(z-w) \cdot (z-w)$$

and estimate the right hand side in terms of M to obtain that in particular for all $z, w \in \mathbb{R}^m$ with $|z|_2 = |w|_2 = 1$ it holds that $|Sz \cdot w| \leq M$. Conclude that $|S|_{op} = M$.

The map

$$z \to Q(z) = Sz \cdot z$$

defined by the symmetric matrix S is called a quadratic form. Observe that in Exercise 18.2 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.

18.6 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

$$|S|_{op} = \sup_{0 \neq z \in H} \frac{|Sz|}{|z|} = \sup_{0 \neq z \in H} \sqrt{\frac{Sz \cdot Sz}{z \cdot z}} = \sup_{z \cdot z = 1} \sqrt{Sz \cdot Sz}, \tag{18.26}$$

and

$$M = \sup_{z \cdot z = 1} Sz \cdot z, \tag{18.27}$$

it suffices to have that S is bounded on the unit ball in H to have

$$M = |S|_{op} < \infty. \tag{18.28}$$

Ignoring the trivial case that $M = 0$ we now observe that the Cauchy-Schwarz inequality in Exercise 18.1 also holds with S replaced by $M - S = MI - S$, I being the identity map, and it thus holds that

$$|(M-S)z \cdot w| \leq \sqrt{(M-S)z \cdot z}\, \sqrt{(M-S)w \cdot w}, \tag{18.29}$$

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 18.3. *Let H be an inner product space and $S : H \to H$ linear, symmetric with $Sz \cdot z \geq 0$ for all $z \in H$, $Sz \neq 0$ for at least one $z \in H$. If for every bounded sequence z_n in H it holds that Sz_n has a convergent subsequence, then*

$$\lambda_1 = \max_{0\neq z\in H} \frac{Sz\cdot z}{z\cdot z} > 0$$

exists, and λ_1 is an eigenvalue of S whose eigenvectors are the maximizers[20] *of the quotient under consideration.*

Remark 18.4. *In fact we only need one single sequence z_n with $z_n \cdot z_n = 1$ such that Sz_n converges and*

$$Sz_n \cdot z_n \to \sup_{0\neq z\in H} \frac{Sz\cdot z}{z\cdot z}$$

to conclude that λ_1 exists, and is an eigenvalue of S whose eigenvectors are the maximizers. In particular this is the case when the supremum is a maximum.

Given an eigenvector w_1 with $|w_1| = 1$ it easily follows that S maps

$$H_1 = \{z \in H : \, z \cdot w_1 = 0\}$$

to itself. Unless H_1 is[21] the null space of S it then follows that

$$\lambda_2 = \max_{z\cdot w_1 = 0 \neq z\in H} \frac{Sz\cdot z}{z\cdot z} > 0$$

[20] Typically only multiples of one eigenvector.

[21] This includes the possibility that $H_1 = \{0\}$.

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[22], 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

$$w_1, w_2, \ldots,$$

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 (18.29) still holding[23] while the version in Exercise 18.1 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.

[22] If the range of H is spanned by $v_1, \ldots, v_N$ for some $N \in \mathbb{N}$.

[23] I first saw this Cauchy-Schwarz trick in the appendix of the PDE book of Craig Evans.

18.7 Singular values and measures of parallelotopes

In the case that $H = \mathbb{R}^m$ the subsequence argument is not needed as the maximizer w for the maximum in Theorem 18.3 exists in view of the compactness of the unit ball in $\mathbb{R}^m$. In particular, in relation to (18.23), we have achieved that $\mathbb{R}^m$ has an orthonormal basis basis of eigenvectors $w_1, \ldots, w_m$ corresponding to eigenvalues $\sigma_1 \geq \cdots \geq \sigma_m \geq 0$ of AA^T. If $n \geq m$ these eigenvalues are called the singular values of A.

As an example consider the case that

$$A = \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} \tag{18.30}$$

and

$$AA^T = \begin{pmatrix} a_1^2 + a_2^2 + a_3^2 & a_1b_1 + a_2b_2 + a_3b_3 \\ b_1a_1 + b_2a_2 + b_3a_3 & b_1^2 + b_2^2 + b_3^2 \end{pmatrix} = \begin{pmatrix} a \cdot a & a \cdot b \\ b \cdot a & b \cdot b \end{pmatrix} \tag{18.31}$$

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

$$|a \times b|^2 = (a_2b_3 - a_3b_2)^2 + (a_3b_1 - a_1b_3)^2 + (a_1b_2 - a_2b_1)^2 = \det(\mathrm{AA^T}),$$

as you should verify. That is to say, $\det(\mathrm{AA^T})$ is the sum of all the squares of all the 2×2-determinants of 2×2 submatrices of A. Here we count these 2×2 submatrices modulo the column permutations in (18.30).

As you may know, the length of the outer product $a \times b$ of a and b equals the area of the parallellogram spanned by a and b. Thus this area is the square root of the sum of the squares of the three 2×2-determinants in (18.30). It is precisely this statement that generalises to the n-dimensional measure of a parallelotope spanned by n vectors $x_1, \ldots, x_n$ in $\mathbb{R}^{\mathrm{N}}$.

Theorem 18.5. *Let $1 \leq n \leq \mathrm{N}$. Consider the parallelotope P spanned by the vectors $x_1, \ldots, x_n$ in $\mathbb{R}^{\mathrm{N}}$. After putting these vectors in the columns[24] of a matrix A, the n-dimensional measure $\mathcal{M}_n(x_1, \ldots, x_n)$ of P is the square root of the determinant of A^TA, and this determinant in turn is the sum of all the squares of the determinants of all $n \times n$ submatrices, and also equals the product $\sigma_1 \cdots \sigma_n$ of the singular values of A.*

[24] NB! Compared to (18.30) we switch from A to A^T in the notation.

Let us sketch a proof of this statement, first for (18.30), without using the outer product, using the invariance of the area under shear transformations. That is to say, the area of the parallelogram spanned by the vectors a and b is the same as that of the parallelogram spanned by the vectors $a + tb$ and b with $t \in \mathbb{R}$ arbitrary. The same statement holds for the determinant of $S = A^T A$ and the determinant of $S_t = A_t^T A_t$ where A_t is the matrix with column vectors $a + tb$ and b. Indeed, writing $A_t = A + tB$ we have

$$A_t^T A_t = (A + tB)^T (A + tB) = A^T A + tA^T B + tB^T A + t^2 B^T B$$
$$= \underbrace{A^T A + tA^T B}_{C_t} + t(\underbrace{B^T A + tB^T B}_{D_t}) = S_t$$

The matrix C_t is the matrix obtained from $S = A^T A$ 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 (18.23) is also equal to the sum of the squares of the determinants of all the maximal square submatrices of A. These are also invariant under the shear transformations used above. Rather than using these transformations to reduce the statement to be proved to the case that the column vectors satisfy $x_i \cdot x_j = 0$ for $i \neq j$ we now use them diagonalise a maximal square part of the matrix A. Note that if the matrix A has no $n \times n$ submatrix with nonzero determinant, then the sum of the squared $n \times n$ determinants is zero, while also it cannot be the case that the column vectors are independent. Then our reduction to the case that the column vectors satisfy $x_i \cdot x_j = 0$ leads to one of these vectors being zero making the n-dimensional measure of P, and thereby the determinant of $A^T A$ zero as well.

Thus we may as well assume that the upper $n \times n$ part of A has nonzero determinant. It is a straightforward linear algebra exercise to show that, most

likely after relabeling the first n coordinates, shear transformations bring A in the form

$$A = \begin{pmatrix} \Lambda \\ B \end{pmatrix}$$

where Λ is an $n \times n$ diagonal matrix with nonzero entries $\lambda_1, \ldots, \lambda_n$. Here we already assumed that $n < \mathrm{N}$ because otherwise there was nothing to prove[25] 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

$$s_{ij} = \sum_{k=1}^{m} b_{ik} b_{jk}.$$

We therefore have, writing $B = [B_1, \ldots, B_n]$ with $B_1, \ldots, B_n$ the column vectors of B and using product notation, that

$$\det(\mathrm{A^T A}) = \underbrace{\Pi_j \lambda_j^2}_{\lambda_\mathrm{j}^1 \cdots \lambda_\mathrm{j}^\mathrm{n}} + \mathrm{s}_{11} \underbrace{\Pi_{j \neq 1} \lambda_j^2}_{\lambda_\mathrm{j}^2 \cdots \lambda_\mathrm{j}^\mathrm{n}} + \cdots + \mathrm{s}_{\mathrm{nn}} \Pi_{\mathrm{j} \neq \mathrm{n}} \lambda_\mathrm{j}^2$$

$$+\det \begin{pmatrix} \mathrm{s}_{11} & \mathrm{s}_{12} \\ \mathrm{s}_{12} & \mathrm{s}_{22} \end{pmatrix} \Pi_{\mathrm{j} \neq 1,2} \lambda_\mathrm{j}^2 + \cdots + \det \mathrm{S} =$$

$$\Pi_j \lambda_j^2 + (B_1 \cdot B_1) \Pi_{j \neq 1} \lambda_j^2 + \cdots + \det \begin{pmatrix} \mathrm{B}_1 \cdot \mathrm{B}_1 & \mathrm{B}_1 \cdot \mathrm{B}_2 \\ \mathrm{B}_1 \cdot \mathrm{B}_2 & \mathrm{B}_2 \cdot \mathrm{B2} \end{pmatrix} \Pi_{\mathrm{j} \neq 1,2} \lambda_\mathrm{j}^2 + \cdots,$$

in which we wrote the term of degree n and only the first terms of degree $2n-2$ and degree $2n-4$ in $\lambda_1, \ldots, \lambda_n$. It should be obvious what the remaining terms are.

On the other hand, the sum of the squared determinants of the $n \times n$ submatrices of A is

$$\Pi_j \lambda_j^2 + (b_{11}^2 + b_{21}^2 + \cdots + b_{m1}^2) \Pi_{j \neq 1} \lambda_j^2 + \cdots + \left(\det \begin{pmatrix} \mathrm{b}_{11} & \mathrm{b}_{12} \\ \mathrm{b}_{12} & \mathrm{b}_{22} \end{pmatrix}^2 + \cdots \right) \Pi_{j \neq 1,2} \lambda_j^2 + \cdots$$

It remains to show that

$$B_1 \cdot B_1 = b_{11}^2 + b_{21}^2 + \cdots + b_{m1}^2,$$

which is clearly the case, and then that

$$\det \begin{pmatrix} \mathrm{B}_1 \cdot \mathrm{B}_1 & \mathrm{B}_1 \cdot \mathrm{B}_2 \\ \mathrm{B}_1 \cdot \mathrm{B}_2 & \mathrm{B}_2 \cdot \mathrm{B2} \end{pmatrix} = \det \begin{pmatrix} \mathrm{b}_{11} & \mathrm{b}_{12} \\ \mathrm{b}_{12} & \mathrm{b}_{22} \end{pmatrix}^2 + \cdots + \det \begin{pmatrix} \mathrm{b}_{(\mathrm{m}-1)1} & \mathrm{b}_{(\mathrm{m}-1)2} \\ \mathrm{b}_{\mathrm{m}2} & \mathrm{b}_{\mathrm{m}2} \end{pmatrix}^2,$$

[25] If you know your determinants.

etcetera. These are the statements we set out to prove for A, before applying shear transformations, but with shorter columnvectors, namely of length $\mathrm{N}-n$, respectively for two such vectors, up to m such vectors. We can thus systematically reduce the statement we want to proof to lower dimensions of the matrix under consideration, until we reach the easy case that $m = 1$.

18.8 Surface integrals

The treatment of Theorem 18.5 above is somewhat different from Edwards' exposition. I now return to (18.22). Generalising to $1 \leq n \leq \mathrm{N}$ we consider

$$\int_{[0,1]^n} \mathcal{M}_n(\frac{\partial\Phi}{\partial u})\,du = \int_0^1 \cdots \int_0^1 \mathcal{M}_n(\frac{\partial\Phi}{\partial u_1}, \ldots, \frac{\partial\Phi}{\partial u_n})\,du_1 \cdots du_n \qquad (18.32)$$

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 18.5. 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 in Edwards[26] after a change of coordinates defined by

$$\frac{ds}{du} = \sqrt{\Phi_1'(u)^2 + \cdots + \Phi_n'(u)^2}.$$

While not corresponding to a change of coordinates the notation

$$dS = \mathcal{M}_2(\Phi_u, \Phi_v)dudv,$$

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 (18.32), i.e.

$$\int_\Phi dS_n = \int_0^1 \cdots \int_0^1 \mathcal{M}_n(\frac{\partial\Phi}{\partial u_1}, \ldots, \frac{\partial\Phi}{\partial u_n})\,du_1 \cdots du_n,$$

[26] In Section V.1 his $\gamma(t)$ would correspond $\Phi(u)$.

and, for a function $f = f(x) = f(x_1, \cdots, x_n)$ which is continuous on

$$\{x = \Phi(u) : \, u \in [0,1]^n\},$$

write

$$\int_\Phi f dS_n = \int_0^1 \cdots \int_0^1 f(\Phi(u_1, \ldots, u_n)) \mathcal{M}_n(\frac{\partial \Phi}{\partial u_1}, \ldots, \frac{\partial \Phi}{\partial u_n}) \, du_1 \cdots du_n, \tag{18.33}$$

the subscript Φ on the integral being at least consistent with the case $n = 1$ and $ds = dS_1$, and in fact coinciding with the notation in the second part of (18.14). 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 $\mathbb{R}^N$. Below we will also allow general closed blocks

$$[a, b] = [a_1, b_1] \times \cdots \times [a_n, b_n].$$

19 Integrating functions over manifolds

Section 13.2 and Section 20.3 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) just below Exercise **??**, 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 20.3 above[1], such that

$$M = \Phi_1((a_1, b_1)) \cup \cdots \cup \Phi_m((a_n, b_n)) = \Phi_1([a_1, b_1]) \cup \cdots \cup \Phi_m([a_n, b_n]), \quad (19.1)$$

and moreover that there exist corresponding smooth functions

$$\zeta_i : \mathbb{R}^N \to [0, 1]$$

with

$$\zeta_1 + \cdots + \zeta_n \equiv 1 \text{ on } M \text{ and } \operatorname{supp} \zeta_i \circ \Phi_i \subset (a_i, b_i)$$

for every $i = 1, \ldots, I$. Here $\operatorname{supp} \zeta_i \circ \Phi_i$ is the support of the function $u \to \zeta_i(\Phi_i(u))$, defined as the closure of the set

$$\{u \in (a_i, b_i) : \zeta_i(\Phi_i(u)) \neq 0\}.$$

We say that $u \to \zeta_i(\Phi_i(u))$ belongs to $C_c^1((a_i, b_i))$, the class of C^1-functions with support contained in the open set (a_i, b_i).

You can think of each function ζ_i as fading the patch $\Phi_i((a_i, b_i))$, making it fade away completely near its boundary where $\zeta_i \equiv 0$, while together the ζ_i leave the whole of M as bright as it was before. Such *brightness* functions ζ_i can be chosen to vanish outside a neighbourhood in $\mathbb{R}^N$ of the image $\Phi_i(K_i)$, and the collection $\zeta_1, \ldots, \zeta_m$ is called a finite partition of unity on M, which is then (turning[2] a theorem around which says that such partitions exist if M is compact) a closed and bounded subset of $\mathbb{R}^N$.

If $f : M \to \mathbb{R}$ is continuous we now wish to define

$$\int_M f dS_n = \int_{\Phi_1} f\zeta_1 dS_n + \cdots + \int_{\Phi_m} f\zeta_m dS_n, \quad (19.2)$$

which requires a theorem that says this is independent of the choice of patches and brightness 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 13.4. 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 $\Phi_i : [a_i, b_i] \to \mathbb{R}^{\mathrm{N}}$ with $\Phi'(u)$ of maximal rank can be chosen such that[4]

$$M = \Phi_1([a_1, b_1]) \cup \cdots \cup \Phi_m([a_n, b_n]) \quad \text{with} \quad \Phi_i((a_i, b_i)) \cap \Phi_j((a_j, b_j)) = \emptyset \tag{19.3}$$

for $i \neq j$, then

$$\int_M f dS_n = \int_{\Phi_1} f \, dS_n + \cdots + \int_{\Phi_m} f dS_n \tag{19.4}$$

is the obvious definition which Edwards uses, and which is what you do in examples.

19.1 More integration of differential forms

We look again at the right hand side of (18.11) with $\mathrm{N} = n+1$, evaluated for $\tilde{v}_i = \zeta v_i$ with ζ a cut-off function vanishing outside and near the boundary of some window

$$[a, b] = [a_1, b_1] \times \cdots \times [a_{\mathrm{N}}, b_{\mathrm{N}}],$$

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

$$\Phi(u_1, \ldots, u_n) = (u_1, \ldots, u_n, f(u_1, \ldots, u_n)) \tag{19.5}$$

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

$$\begin{vmatrix} \frac{\partial\Phi_1}{\partial u_1} & \frac{\partial\Phi_2}{\partial u_1} & \frac{\partial\Phi_3}{\partial u_1} \\ \frac{\partial\Phi_1}{\partial u_2} & \frac{\partial\Phi_2}{\partial u_2} & \frac{\partial\Phi_3}{\partial u_2} \\ e_1 & e_2 & e_3 \end{vmatrix} = \begin{vmatrix} \frac{\partial\Phi_1}{\partial u_1} & \frac{\partial\Phi_2}{\partial u_1} \\ \frac{\partial\Phi_1}{\partial u_2} & \frac{\partial\Phi_2}{\partial u_2} \end{vmatrix} e_3 + \begin{vmatrix} \frac{\partial\Phi_2}{\partial u_1} & \frac{\partial\Phi_3}{\partial u_1} \\ \frac{\partial\Phi_2}{\partial u_2} & \frac{\partial\Phi_3}{\partial u_2} \end{vmatrix} e_1 + \begin{vmatrix} \frac{\partial\Phi_3}{\partial u_1} & \frac{\partial\Phi_1}{\partial u_1} \\ \frac{\partial\Phi_3}{\partial u_2} & \frac{\partial\Phi_1}{\partial u_2} \end{vmatrix} e_2 \tag{19.6}$$

[4] Edwards: a hard theorem says this can be done.

[5] Like (18.5), whereas I did most local arguments using (18.6).

is commonly called the cross product of the vectors Φ_{u_1} and Φ_{u_2}, and for $\Phi(u_1, u_2) = (u_1, u_2, f(u_1, u_2))$ it evaluates as[6]

$$e_3 - \frac{\partial f}{\partial u_1} e_1 - \frac{\partial f}{\partial u_2} e_2 = -\frac{\partial f}{\partial u_1} e_1 - \frac{\partial f}{\partial u_2} e_2 + e_3, \tag{19.7}$$

which is a positive multiple of the unit vector ν characterised by having its last component positive and being perpendicular to the graph defined by $u_3 = f(u_1, u_2)$. For any continuously differentiable

$$\Phi : [a_1, b_1] \times [a_2, b_2] \to \mathbb{R}^3$$

with Φ_{u_1} and Φ_{u_2} linearly independent, the vector defined by (19.6) is perpendicular to the plane spanned by Φ_{u_1} and Φ_{u_2}, and can be normalised by dividing it by its length, which we recognise as

$$\mathcal{M}_2(\Phi_{u_1}, \Phi_{u_2})$$

in view of Theorem 18.5. If we call this normalised vector ν, which in case of (19.7) is simply[7]

$$\nu = \frac{1}{\sqrt{1 + f_{u_1}^2 + f_{u_2}^2}} \left(-\frac{\partial f}{\partial u_1} e_1 - \frac{\partial f}{\partial u_2} e_2 + e_3 \right), \tag{19.8}$$

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

$$\int_M \nu \cdot \tilde{v} \; dS_2 =$$

$$\iint_{[a_1,b_1]\times[a_2,b_2]} \left(\tilde{v}_1 \begin{vmatrix} \frac{\partial \Phi_2}{\partial u_1} & \frac{\partial \Phi_3}{\partial u_1} \\ \frac{\partial \Phi_2}{\partial u_2} & \frac{\partial \Phi_3}{\partial u_2} \end{vmatrix} + \tilde{v}_2 \begin{vmatrix} \frac{\partial \Phi_3}{\partial u_1} & \frac{\partial \Phi_1}{\partial u_1} \\ \frac{\partial \Phi_3}{\partial u_2} & \frac{\partial \Phi_1}{\partial u_2} \end{vmatrix} + \tilde{v}_3 \begin{vmatrix} \frac{\partial \Phi_1}{\partial u_1} & \frac{\partial \Phi_2}{\partial u_1} \\ \frac{\partial \Phi_1}{\partial u_2} & \frac{\partial \Phi_2}{\partial u_2} \end{vmatrix} \right) \underbrace{du_1 \, du_2}_{du},$$

which we may be inclined to write as

$$\int_\Phi \tilde{v}_1 \, dx_2 dx_3 + \tilde{v}_2 \, dx_3 dx_1 + \tilde{v}_3 \, dx_1 dx_2 = \int_\Phi \omega, \tag{19.9}$$

with

$$\omega = \tilde{v}_1 \, dx_2 dx_3 + \tilde{v}_2 \, dx_3 dx_1 + \tilde{v}_3 \, dx_1 dx_2,$$

[6] Denoting the partial 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

$$dx_2 dx_3 = \begin{vmatrix} \frac{\partial x_2}{\partial u_1} & \frac{\partial x_3}{\partial u_1} \\ \frac{\partial x_2}{\partial u_2} & \frac{\partial x_3}{\partial u_2} \end{vmatrix} \underbrace{du_1\, du_2}_{du} = \begin{vmatrix} \frac{\partial \Phi_2}{\partial u_1} & \frac{\partial \Phi_3}{\partial u_1} \\ \frac{\partial \Phi_2}{\partial u_2} & \frac{\partial \Phi_3}{\partial u_2} \end{vmatrix} \underbrace{du_1\, du_2}_{du}.$$

We then have that (19.9) 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

$$d\omega = \left(\frac{\partial \tilde{v}_1}{\partial x_1} + \frac{\partial \tilde{v}_2}{\partial x_2} + \frac{\partial \tilde{v}_3}{\partial x_3} \right) \underbrace{dx_1 dx_2 dx_3}_{\neq dx},$$

in which $dx_1 dx_2 dx_3$ is part of a 3-form and not be read as $dx = dx_1\, dx_2\, dx_3$.

All of the above generalises[8] to arbitrary $\mathrm{N} = n + 1$, e.g. we also have

$$\int_\Omega \left(\frac{\partial \tilde{v}_1}{\partial x_1} + \frac{\partial \tilde{v}_2}{\partial x_2} + \frac{\partial \tilde{v}_3}{\partial x_3} + \frac{\partial \tilde{v}_4}{\partial x_4} \right) dx \tag{19.10}$$

$$= \int_\Phi \tilde{v}_1\, dx_2 dx_3 dx_4 + \cdots \text{ (cyclicly permutated terms) } \cdots = \int_\Phi \omega,$$

using rules like

$$dx_2 dx_3 dx_4 = \begin{vmatrix} \frac{\partial x_2}{\partial u_1} & \frac{\partial x_3}{\partial u_1} & \frac{\partial x_4}{\partial u_1} \\ \frac{\partial x_2}{\partial u_2} & \frac{\partial x_3}{\partial u_2} & \frac{\partial x_4}{\partial u_2} \\ \frac{\partial x_2}{\partial u_3} & \frac{\partial x_3}{\partial u_3} & \frac{\partial x_4}{\partial u_3} \end{vmatrix} \underbrace{du_1\, du_2\, du_3}_{du},$$

and (19.10) should be the integral of the 4-form

$$d\omega = \left(\frac{\partial \tilde{v}_1}{\partial x_1} + \frac{\partial \tilde{v}_2}{\partial x_2} + \frac{\partial \tilde{v}_3}{\partial x_3} + \frac{\partial \tilde{v}_4}{\partial x_4} \right) dx_1 dx_2 dx_3 dx_4.$$

Clearly such a d-calculus requires rules such as $dx_i dx_j = -dx_j dx_i$. I played with the formal rules that one might like to have in Section 15, see also the discussion after Stelling 9.2. This 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 $dx_i dx_j$.

[8] This is why we put the unit vectors in the last row of the determimant in (19.6).

We conclude with the simplest but slightly confusing case, $n = 1$ and $\mathrm{N} = 2$, when (19.6) should be replaced by

$$\begin{vmatrix} \frac{\partial \Phi_1}{\partial u} & \frac{\partial \Phi_2}{\partial u} \\ e_1 & e_2 \end{vmatrix} = \frac{\partial \Phi_2}{\partial u}\, e_1 - \frac{\partial \Phi_1}{\partial u}\, e_2, \tag{19.11}$$

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

$$\int_{\partial\Omega} p(x,y)dx + q(x,y)dy = \int_\Omega (q_x - p_y)dxdy, \tag{19.12}$$

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 18.3. Section 19.2 below explores what's going on here.

Note that in all these examples the N-form $\omega = f(x)dx_1 \cdots dx_\mathrm{N}$ integrated over the domain Ω should sensibly be agreed to give[9]

$$\int_\Omega \omega = \int_\Omega f(x)dx_1 \cdots dx_\mathrm{N} = \int_\Omega f.$$

[9] Don't confuse this f with f in the local description above.

19.2 From Green's to Stokes' curl theorem

Now consider (19.5) as a local description of a manifold M and forget about Ω as being a domain with $M = \partial\Omega$. Instead let Ω be as in (18.1) with N = 2 and let M be the graph of $f : \Omega \to \mathbb{R}$. Assume for simplicity that $\partial\Omega$ is parameterised by a 1-periodic continuously differentiable function $t \to u(t) = (u_1(t), u_2(t))$. Then

$$t \xrightarrow{\gamma} (u_1(t), u_2(t), f(u_1(t), u_2(t))) \tag{19.1}$$

parameterises the "boundary"

$$\partial M = \{\underbrace{(u, f(u))}_{\Phi(u)} : u \in \partial\Omega\},$$

and

$$u \xrightarrow{\Phi} (u, f(u)) \tag{19.2}$$

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 (18.18) and (18.19) and consider the integral

$$\int_{\partial M} \omega$$

as in (18.16). It evaluates as

$$\int_{\partial M} \omega = \int_0^1 (F_1(\gamma(t))\gamma_1'(t) + F_2(\gamma(t))\gamma_2'(t)) + F_3(\gamma(t))\gamma_3'(t))) \, dt$$

$$= \int_0^1 (F_1(u(t), f(u(t))u_1'(t) + F_3(u(t), f(u(t)))f_{u_1}(u(t)u_1'(t)) \, dt$$

$$+ \int_0^1 (F_2(u(t), f(u(t))u_2'(t) + F_3(u(t), f(u(t)))f_{u_2}(u(t)u_2'(t)) \, dt =$$

$$\int_{\partial\Omega} \zeta = \int_\Omega d\zeta, \tag{19.3}$$

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

$$d\zeta = \phi(u_1, u_2) du_1 du_2 \tag{19.4}$$

with $\phi(u_1, u_2)$ given by

$$\phi = -\underbrace{\left(\frac{\partial F_3}{\partial x_2} - \frac{\partial F_2}{\partial x_3}\right)}_{G_1}\frac{\partial f}{\partial u_1} - \underbrace{\left(\frac{\partial F_1}{\partial x_3} - \frac{\partial F_3}{\partial x_1}\right)}_{G_2}\frac{\partial f}{\partial u_2} + \underbrace{\left(\frac{\partial F_2}{\partial x_1} - \frac{\partial F_1}{\partial x_2}\right)}_{G_3} \tag{19.5}$$

$$= -G_1\frac{\partial f}{\partial u_1} - G_2\frac{\partial f}{\partial u_2} + G_3.$$

You should note that the *second order derivatives* of (19.2) are dropouts in the calculations that lead to (19.5).

Now compare (19.5) to ν in (19.8) and recall that for Φ given by (19.2) 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 (19.5), and commonly denoted as $G = \nabla \times F$, i.e.

$$\int_{\partial M} (F \cdot \tau)\, dS_1 = \int_{M} (G \cdot \nu)\, dS_2 \quad \text{with} \quad G = \nabla \times F, \tag{19.6}$$

using the parameterisations as indicated[10]. But don't say goodbye:

19.3 Pullbacks and the action of d

We already saw in the reasoning from (18.16) to (18.17) that d acting on a C^1-function $f = f(x_1, \ldots, x_{\rm N})$ produces a 1-form

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_{\rm N}} dx_{\rm N} = \frac{\partial f}{\partial x_i} dx_i, \tag{19.7}$$

using the convention that we sum over repeated indices. With $f(x_1, \ldots, x_{\rm N})$ replaced by $u(x, y)$ this is (15.3) in Section 15.2. There I played with the d-algebra that emerges whenever you do integration using formal notations such as (9.7), which is just (19.7) with $n = 1$ and $f(x_1, \cdots, x_{\rm N})$ replaced by $F(x)$.

Now consider a parameterisation $x = \Phi(u)$ as in (C) in Section 13.2. We use Φ to pull back expressions with x and $dx_1, \ldots, dx_{\rm N}$ back to expressions with u and $du_1, \ldots, du_n$, in a way that is consistent with the discussion leading to (19.9) and the formal rules that emerge in the calculations to do so. Thus we certainly want to deal with

$$f(x) = \phi(u) \quad \text{via} \quad x = \Phi(u). \tag{19.8}$$

A mathematician's way to do so is to introduce

$$\phi = \Phi^*(f) = f \circ \Phi, \tag{19.9}$$

the pullback of f via Φ, which then also provides us with

$$d\phi = \frac{\partial \phi}{\partial u_1} du_1 + \cdots + \frac{\partial \phi}{\partial u_n} du_n. \tag{19.10}$$

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

$$\Phi^*(f_i dx_i) = \underbrace{\Phi^*(f_i)}_{\phi_i} \, \Phi^*(dx_i), \tag{19.11}$$

[10] 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

$$\Phi^*(df) = \underbrace{\frac{\partial f}{\partial x_i}((\Phi(u))}_{\Phi^*(D_i f)(u)}\ \Phi^*(dx_i). \tag{19.12}$$

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 (19.10) implies via the chain rule that

$$d(\Phi^*(f)) = \frac{\partial}{\partial u_j}(f(\Phi(u))\, du_j = \frac{\partial f}{\partial x_i}(\Phi(u))\frac{\partial \Phi_i}{\partial u_j}\, du_j, \tag{19.13}$$

and comparing to (19.12) we see that, if we define the pullback of dx_i under Φ to be

$$\Phi^*(dx_i) = \frac{\partial \Phi_i}{\partial u_j}\, du_j, \tag{19.14}$$

it follows that

$$\Phi^*(df) = \Phi^*(df). \tag{19.15}$$

The definition of $\Phi^*(dx_i)$ by (19.14) 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 (19.8) is formalised by (19.9). It implies that the pullback of the 1-form in (19.11) evaluates as

$$\underbrace{\Phi^*(f_i dx_i) = \phi_i\, \frac{\partial \Phi_i}{\partial u_j}\, du_j}_{with\ \ \phi_i(u)=f_i(\Phi(u))} = f_i(\Phi(u))\, \frac{\partial \Phi_i}{\partial u_j}\, du_j = f_i(\Phi(u))\, D_j\Phi_i(u)\, du_j. \tag{19.16}$$

Next we observe that d acting on the resulting 1-form in (19.16) may be evaluated, using the chain rule and $du_k du_j = -du_j du_k$, as

$$d(\Phi^*(f_i dx_i)) = d(f_i(\Phi(u))\, \frac{\partial \Phi_i}{\partial u_j}\, du_j) = \frac{\partial}{\partial u_k}(f_i(\Phi(u))\, \frac{\partial \Phi_i}{\partial u_j})\, du_k du_j$$

$$= (\frac{\partial}{\partial u_k}(f_i(\Phi(u)))\, \frac{\partial \Phi_i}{\partial u_j}\, du_k du_j + f_i(\Phi(u))\ \underbrace{\frac{\partial^2 \Phi_i}{\partial u_k \partial u_j}\, du_k du_j}_{zero\ the\ hero!} =$$

$$\frac{\partial f_i}{\partial x_k}(\Phi(u))\, \frac{\partial \Phi_k}{\partial u_k}\, \frac{\partial \Phi_i}{\partial u_j}\, du_k\, du_j = \Phi^*(D_k f_i)\, \frac{\partial \Phi_k}{\partial u_k}\, \frac{\partial \Phi_i}{\partial u_j}\, du_k\, du_j, \tag{19.17}$$

in which we used

$$d(f_i dx_i) = \frac{\partial f_i}{\partial x_k}\, dx_k\, dx_i \tag{19.18}$$

in the u-variables. Recall that this was the definition[11] in Section 15.2 of the action of d on 1-forms. With

$$\Phi^*(f_{ij} dx_i dx_j) = \underbrace{\Phi^*(f_{ij})}_{\phi_{ij}}\, \Phi^*(dx_i dx_j) \tag{19.19}$$

as the obvious defining analog of (19.11), we have that

$$\Phi^*(d(f_i dx_i)) = \Phi^*(\frac{\partial f_i}{\partial x_k}\, dx_k dx_i) = \Phi^*(D_k f_i)\, \Phi^*(dx_k dx_i). \tag{19.20}$$

Comparing to (19.20) to (19.17) it follows that

$$\Phi^*(d(f_i dx_i)) = d(\Phi^*(f_i dx_i)), \tag{19.21}$$

provided we define

$$\Phi^*(dx_k dx_i) = \underbrace{\frac{\partial \Phi_k}{\partial u_k}\, \frac{\partial \Phi_i}{\partial u_j}\, du_k\, du_j}_{sum\ over\ 1\le k,j\le n} = \underbrace{(\frac{\partial \Phi_k}{\partial u_k}\, \frac{\partial \Phi_i}{\partial u_j} - \frac{\partial \Phi_k}{\partial u_k}\, \frac{\partial \Phi_i}{\partial u_j})\, \underline{du_k\, du_j}}_{sum\ over\ 1\le k<j\le n}$$

$$= \frac{\partial(\Phi_k, \Phi_i)}{\partial u_k \partial u_j}\, \underline{du_k\, du_j}, \tag{19.22}$$

in which the underline indicates that we sum over all k, j with $1 \le k < j \le n$. Just as in (19.15) we see that the actions of d and Φ^* commute.

Note that the second order derivatives have disapperared in (19.17). The derivation is typically done under the assumption that $\Phi \in C^2$, also in Edwards, and an additional analysis argument is needed[12] to give meaning to the results if Φ is only in C^1, because the determinants in (19.22) are exactly the determinants that showed up in (19.6) and the subsequent derivation of (19.9), where effectively $dx = dx_1\, dx_2\, dx_3$ is first replaced by a 3-form $dx_1 dx_2 dx_3$ pulled back to a 2-form $du_1 du_2$, which in turn is replaced by $du = du_1\, du_2$ again.

The step by step generalisation to the action of d and Φ^* on k-forms of any order k is easily made once the reasoning above is understood. For any k-form

$$\omega = f_{i_1,\ldots,i_k} dx_{i_1} \cdots dx_{i_k}$$

[11] Recall the choice to set $ddx_i = 0$, leading to $dd\omega = 0$ for any form ω.

[12] Using approximation arguments.

we have

$$\Phi^*(d\omega) = d(\Phi^*(\omega)) \tag{19.23}$$

Every such form may be written as

$$\omega = f_{i_1,\ldots,i_k} dx_{i_1} \cdots dx_{i_k} = \tilde{f}_{i_1,\ldots,i_k} \underline{dx_{i_1} \cdots dx_{i_k}}, \tag{19.24}$$

where in the second expression we sum only over those $i_1, \ldots, i_k$ for which $1 \leq i_1 < \cdots < i_k \leq \mathrm{N}$. For instance

$$\omega = f_{ij}\, dx_i dx_j = \underbrace{(f_{ij} - f_{ji})}_{\tilde{f}_{ij}}\, \underline{dx_i dx_j},$$

but this is not compulsory, as the examples

$$\omega = f_1\, dx_1 + f_2\, dx_2 + f_3\, dx_3$$

with cyclic notation for

$$d\omega = \underbrace{(\frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3})}_{g_1}\, dx_2 dx_3 + \underbrace{(\frac{\partial f_1}{\partial x_3} - \frac{\partial f_3}{\partial x_1})}_{g_2}\, dx_3 dx_1 + \underbrace{(\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2})}_{g_3}\, dx_1 dx_2$$

and

$$\zeta = g_1\, dx_2 dx_3 + g_2\, dx_3 dx_1 + g_3\, dx_1 dx_2$$

with

$$d\zeta = (\frac{\partial g_1}{\partial x_1} + \frac{\partial g_2}{\partial x_2} + \frac{\partial g_3}{\partial x_3})\, dxdydz$$

in Section 19.4 show.

Finally we observe that if we put the coefficients f_1, f_2, f_3 of this ω in a vector $F = f_1 e_1 + f_2 e_2 + f_3 e_3$ and the coefficients g_1, g_2, g_3 in this cyclic representation of $d\omega$ in a vector $G = g_1 e_1 + g_2 e_2 + g_3 e_3$, we obtain that

$$G = \nabla \times F,$$

the curl of F, whereas with the coefficients of η we obtain the coefficient of $d\zeta$ as

$$\frac{\partial g_1}{\partial x_1} + \frac{\partial g_2}{\partial x_2} + \frac{\partial g_3}{\partial x_3} = \nabla \cdot G,$$

the divergence of G. These appear in the Gauss divergence and the Stokes curl theorems for vectorfields in $\mathbb{R}^3$ in Section 19.4 below[13]. The general statement (19.34) is also called Stokes Theorem. It has both theorems in $\mathbb{R}^3$ and Green's Theorem in $\mathbb{R}^3$ as special cases.

[13] The statement that $dd\omega = 0$ corresponds to the div of a curl being always zero:

$$\nabla \cdot \nabla \times F = 0.$$

19.4 From Gauss' to general Stokes' Theorem

From Section 19.1 and partitions of unity arguments we have that for $\Omega \subset \mathbb{R}^{\mathrm{N}} = \mathbb{R}^{n+1}$ 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 $\nu : \partial\Omega \to \mathbb{R}^{\mathrm{N}}$ with $\nu(p)$ pointing out of Ω in every patch as above. For every continuously differentiable $V : \Omega \to \mathbb{R}^{\mathrm{N}}$ it now holds that

$$\int_\Omega \nabla \cdot V = \int_{\partial\Omega} \nu \cdot V \, dS_{_{N-1}}, \tag{19.25}$$

and this statement is called the Gauss Divergence Theorem.

We now use the reformulation with differential forms and pullbacks of forms with $\Phi : \mathbb{R}^{n+1} \to \mathbb{R}^{\mathrm{N}}$ with $\mathrm{N} > n+1$ to formulate Stokes' Theorem for integral n-forms over $\Phi(M)$ considered as the boundary of $\Phi(\Omega)$, first for $n+1 = 2$ and $\mathrm{N} = 3$. So let

$$\omega = f_1(x)dx_1 + f_2(x)dx_2 + f_3(x)dx_3 \tag{19.26}$$

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 (19.12) applies, whence

$$\int_{\partial\Omega} \zeta = \int_{\partial\Omega} p_1(u_1,u_2)du_1 + p_2(u_1,u_2)du_2 = \int_\Omega (\frac{\partial p_2}{\partial u_1} - \frac{\partial p_1}{\partial u_2})du_1du_2 = \int_\Omega d\zeta. \tag{19.27}$$

Observe that the second equality in (19.27) holds in view of (19.12), which is a rewritten version of (19.25) 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 (19.18).

We need

$$\int_{\partial\Omega} \zeta = \int_{\partial\Omega} \Phi^*\omega = \int_{\Phi(\partial\Omega)} \omega, \tag{19.28}$$

and

$$\int_{\Omega} d\zeta = \int_{\Omega} d\,\Phi^*\omega = \int_{\Omega} \Phi^*(d\omega) = \int_{\phi(\Omega)} d\omega \tag{19.29}$$

to conclude for ω given by (19.26) that

$$\int_{dS} \omega = \int_{dS} f_1 dx_1 + f_2 dx_2 + f_3 dx_3 = \int_S d\omega, \tag{19.30}$$

in which $S = \Phi(\Omega)$. It is the last equality in each of (19.28) and (19.29) that has to be checked, the other equalities follow from our d-algebra and the commutation of d and Φ^*.

Let us once more spell out the d-algebra by which (19.18) evaluates as

$$d\omega = (\frac{\partial f_1}{\partial x_1}dx_1 + \frac{\partial f_1}{\partial x_2}dx_2 + \frac{\partial f_1}{\partial x_3}dx_3)dx_1$$

$$+(\frac{\partial f_2}{\partial x_1}dx_1 + \frac{\partial f_2}{\partial x_2}dx_2 + \frac{\partial f_2}{\partial x_3}dx_3)dx_2$$

$$+(\frac{\partial f_3}{\partial x_1}dx_1 + \frac{\partial f_3}{\partial x_2}dx_2 + \frac{\partial f_3}{\partial x_3}dx_3)dx_3 =$$

$$\frac{\partial f_1}{\partial x_2}dx_2dx_1 + \frac{\partial f_2}{\partial x_1}dx_1dx_2 + \frac{\partial f_1}{\partial x_3}dx_3dx_1 + \frac{\partial f_3}{\partial x_1}dx_1dx_3 + \frac{\partial f_2}{\partial x_3}dx_2 + \frac{\partial f_3}{\partial x_2}dx_2dx_3$$

$$= (\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2})dx_1dx_2 + (\frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3})dx_2dx_3 + (\frac{\partial f_1}{\partial x_3} - \frac{\partial f_3}{\partial x_1})dx_3dx_1$$

$$(\frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3})dx_2dx_3 + (\frac{\partial f_1}{\partial x_3} - \frac{\partial f_3}{\partial x_1})dx_3dx_1 + (\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2})dx_1dx_2$$

$$= g_1 dx_2dx_3 + g_2 dx_3dx_1 + g_3 dx_1dx_2.$$

Comparing to (19.9) we recognise for $F(x) = f_1(x)e_1 + f_2(x)e_2 + f_3(x)e_3$ that

$$\int_{dS} f_1 dx_1 + f_2 dx_2 + f_3 dx_3 =$$

$$\int_S (\frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3})dx_2dx_3 + (\frac{\partial f_1}{\partial x_3} - \frac{\partial f_3}{\partial x_1})dx_3dx_1 + (\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2})dx_1dx_2$$

$$= \int_S G \cdot \nu = \int_S (\nabla \times F) \cdot \nu, \tag{19.31}$$

in which $g_1(x)e_1 + g_2(x)e_2 + g_3(x)e_3 = G(x) = \nabla \times F$ and ν is the normal vector on $S = \Phi(\Omega)$ defined by (19.6).

It thus remains to check the two analytical statements

$$\int_{\partial\Omega} \Phi^*\omega = \int_{\Phi(\partial\Omega)} \omega \quad \text{and} \quad \int_\Omega \Phi^*(d\omega) = \int_{\Phi(\Omega)} d\omega, \tag{19.32}$$

which complement the d-algebra presented above, and which are both of the form

$$\int_M \Phi^*\omega = \int_{\Phi(M)} \omega, \tag{19.33}$$

with respectively $M = \partial\Omega$ and $M = \Omega$. For this we need again Section 15.4 combined with the usual localisations via partitions of unity. Not very hard but still to be done.

It will be convenient here to have $\Phi(M)$ described by compositions of Φ and patches of M, see the remark at the end of Section 20.4. Also, we still have to deal with integrals over manifolds with boundaries, to obtain

$$\int_{\Phi(\partial\Omega)} \omega = \int_{\Phi(\Omega)} d\omega, \tag{19.34}$$

as the final result in which $M = \Phi(\Omega)$ is a manifold with boundary $\partial M = \Phi(\partial\Omega)$, with $\Omega \in \mathbb{R}^n$ as described at the beginning of this section, Φ a continuously differentiable injective map from Ω to $\mathbb{R}^N$ with Jacobian matrix of rank n throughout Ω, and ω an n-form with continuously differentiable coefficients. Generalisations to piecewise C^1-boundaries then still have to be discussed.

19.5 More exercises

Laat

$$\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}. \tag{19.35}$$

Op $\partial\Omega$ is ν de normaalvector met componenten $\nu_x = x$ en $\nu_y = y$ in de notatie zoals in de uitleg in Sectie 18.1 en de route van (18.7) naar (18.9). Met $y = \sqrt{1-x^2}$, $y = -\sqrt{1-x^2}$, $x = \sqrt{1-y^2}$, en $x = -\sqrt{1-y^2}$ hebben we vier grafieken van functies f zoals in (18.5) and (18.6), op te kiezen domeinen, bijvoorbeeld $[-a, a]$ met $\frac{1}{\sqrt{2}} < a < 1$ if we think in terms of (19.1), or $a = \frac{1}{\sqrt{2}}$ if we think in terms of (19.3).

The graph parameterisations

$$x \to (x, \sqrt{1-x^2}), \quad x \to (x, -\sqrt{1-x^2}), \qquad (19.36)$$
$$y \to (\sqrt{1-y^2}, y), \quad y \to (-\sqrt{1-y^2}, y)$$

may look uglier than

$$\phi \to (\cos\phi, \sin\phi), \qquad (19.37)$$

which requires only two ϕ-domains, e.g. $[-\frac{3\pi}{4}, \frac{3\pi}{4}]$ and $[\frac{\pi}{4}, \frac{7\pi}{4}]$ for (19.1), and $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and $[\frac{\pi}{2}, \frac{3\pi}{2}]$ for (19.3).

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 the two semicircles on the left and on the right, and likewise $x = ty$ for the upper and lower semicircle, with t running from $-\infty$ to $+\infty$, and you can put

$$t = \frac{s}{1-s} \quad \text{or} \quad t = -\frac{s}{1-s} \qquad (19.38)$$

in each of them to obtain.

$$x = \frac{1-s}{\sqrt{1-2s+2s^2}};\ y = \frac{s}{\sqrt{1-2s+2s^2}}$$

parameterising $\{(x, y) \in \mathbb{R}^2 : x \geq 0,\ y \geq 0,\ x^2 + y^2 = 1\}$ with $s \in [0, 1]$.

Exercise 19.1. Use the t-parameterisations above to calculate the area of the unit disk via integrals such as $\int x dy$ of $\int y dx$ over $\partial\Omega$. You should get and evaluate integrands[14] like

$$\frac{t^2}{(1+t^2)^2} = \frac{1}{1+t^2} - \frac{1}{(1+t^2)^2}.$$

Exercise 19.2. Refering to (18.4) and the subsequent 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 (19.11) defines a vector pointing out of Ω.

[14] Recall $\int_{-\infty}^{\infty} \frac{1}{1+t^2}\, dt = \pi$, $\int_{-\infty}^{\infty} \frac{1}{(1+t^2)^2}\, dt = \frac{\pi}{2}$, $\int_{-\infty}^{\infty} \frac{1}{(1+t^2)^3}\, dt = \frac{3\pi}{8}$, ...

Exercise 19.3. Same as Exercise 19.2 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.

Exercise 19.4. In physics results like (18.3) 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 (18.3) holds for $v : [a, b] \to \mathbb{R}^2$ continuously differentiable.

Note that Ω in Exercise 19.4 is given by

$$F(x_1, x_2) = (x_1 - a_1)(x_1 - b_1)(x_2 - a_2)(x_2 - b_2) < 0,$$

which has a zero gradient in the corners of Ω, just as (19.41) for Exercise 19.6 below. Here is an example without an F to define Ω.

Exercise 19.5. Suppose that the boundary of a bounded open set $\Omega \subset \mathbb{R}^2$ is given by a periodic solution of a system of differential equations $\dot{x} = P(x, y)$ and $\dot{y} = Q(x, y)$, with $P, Q : \mathbb{R}^2 \to \mathbb{R}$ continuously differentiable on Ω. Show that

$$\iint_\Omega (\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y})\, dxdy = 0.$$

Edwards has a nice exercise about Descartes' Folium from which I lifted the $y = tx$-trick above. It allows to find the solutions of

$$F(x, y) = x^3 + y^3 - 3xy = 0, \tag{19.39}$$

in the form

$$x = x(t) = \frac{3t}{1 + t^3};\ y = y(t) = \frac{3t^2}{1 + t^3}, \tag{19.40}$$

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

[15] https://www.quora.com/What-is-the-plural-of-axis

by (19.40) with $t \in (-1, 1)$, the other by (19.40) with x and y interchanged. Exercise 2.3 in Chapter V of Edwards is about

$$\Omega = \{(x, y) \in \mathbb{R}^2 : x > 0,\, y > 0,\, F(x, y) = x^3 + y^3 - 3xy < 0\}. \quad (19.41)$$

with $\partial\Omega$ given by (19.40) and $t \in [0, \infty)$. You should examine the graphs of x and y as functions of t in (19.40). You can get the area of Ω as

$$-\int_0^\infty y(t)x'(t)\,dt = \int_0^\infty x(t)y'(t)\,dt, \quad (19.42)$$

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 (19.40) 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 19.6. Substitute $y = t^{\frac{1}{3}}x$ in the equation for the folium to get x and y in terms of t and evaluate (19.42) 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 19.7. As Exercise 19.6 but use (19.38) to get the boundary parameterised with $0 \le s \le 1$.

In the last exercise you see that the boundary of (19.41) 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 18.1 fails. The same issue occurs in the trivial case of Exercise 19.4.

Note that (19.41) 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

$$x = s^{\frac{1}{p(p-2)}}(1 - s)^{\frac{p-1}{p(p-2)}}\,;\; y = s^{\frac{p-1}{p(p-2)}}(1 - s)^{\frac{1}{p(p-2)}}\,, \quad (19.43)$$

parameterises the loop in the solution set of $x^p + y^p = xy$, with

$$x\frac{dy}{ds} - y\frac{dx}{ds} = \frac{1}{p}\,s^{\frac{1}{p-2}-1}(1 - s)^{\frac{1}{p-2}-1}, \quad (19.44)$$

[16] See Exercise 19.14.

which looks much better than the individual terms $x\frac{dy}{ds}$ and $y\frac{dx}{ds}$. With the β-function[17] defined by

$$B(x,y) = \int_0^1 s^{x-1}(1-s)^{y-1}\,ds,$$

the area surrounded by $[0,1] \ni s \to (x(s), y(s))$, the loop in

$$x^p + y^p = xy \tag{19.45}$$

is thus equal to

$$A_p = \frac{1}{2p}\,B(\frac{1}{p-2}, \frac{1}{p-2}), \tag{19.46}$$

which gives $\frac{1}{6}$ for $p = 3$ and differs from Exercise 19.6 by a factor 3^2, consistent with (19.40).

Note that in deriving (19.44) from (19.43) you may get lost if you don't introduce

$$\alpha = \frac{1}{p(p-2)} \quad \text{and} \quad \beta = \frac{p-1}{p(p-2)} = (p-1)\alpha$$

and continue your calculations with α and β. I also suggest to write derivatives such as

$$\frac{d}{ds}\,s^\alpha(1-s)^\beta = (\frac{\alpha}{s} - \frac{\beta}{1-s})\,s^\alpha(1-s)^\beta = (\alpha - (\alpha+\beta)s)\,s^{\alpha-1}(1-s)^{\beta-1},$$

which will help you to factor out common factors when such expressions have to be combined later on, as you will notice if you tackle this question: how about the volume V_p in $\{(x,y,z) \in \mathbb{R}^3 : x \geq 0,\, y \geq 0,\, z \geq 0\}$ surrounded by $x^p + y^p + z^p = xyz$ when $p > 3$?

Exercise 19.8. Substitute $y = s^{\frac{1}{p}}x$ and $z = t^{\frac{1}{p}}x$ in $x^p + y^p + z^p = xyz$ to obtain a parameterisation of the solutions with $x, y, z > 0$ in the form

$$x = s^\alpha t^\alpha (1+s+t)^{-p\alpha}, \quad y = s^{(p-2)\alpha} t^\alpha (1+s+t)^{-p\alpha}, \quad z = s^\alpha t^{(p-2)\alpha}(1+s+t)^{-p\alpha},$$

and evaluate

$$xdydz = x(\frac{\partial y}{\partial s}\frac{\partial z}{\partial t} - \frac{\partial y}{\partial t}\frac{\partial z}{\partial s})$$

as xyz times a factor that you have to computer carefully, to find the correct double integral in s and t that gives the desired volume. The integral is the difference of two similar terms each of which is st to some power times $(1+s+t)$ to some power. Substituting $t = (1+s)x$ both integrals reduce to products of single integrals that reduce to β-functions again.

[17] More on the β-function in [HM].

Just in case, I arrived via

$$xyz = \frac{(st)^{\frac{1}{p-3}}}{(1+s+t)^{\frac{3}{p-3}}}$$

and

$$\frac{1}{yz}(\frac{\partial y}{\partial s}\frac{\partial z}{\partial t} - \frac{\partial y}{\partial t}\frac{\partial z}{\partial s}) = \frac{1}{p^2(p-3)st}\left(\frac{p}{1+s+t} - 1\right)$$

at

$$\frac{1}{p(p-3)}\underbrace{\int_0^\infty\int_0^\infty \frac{(st)^{\frac{1}{p-3}-1}\,dsdt}{(1+s+t)^{\frac{p}{p-3}}}}_{S(\frac{1}{p-3},\frac{p}{p-3})} - \frac{1}{p^2(p-3)}\underbrace{\int_0^\infty\int_0^\infty \frac{(st)^{\frac{1}{p-3}-1}\,dsdt}{(1+s+t)^{\frac{3}{p-3}}}}_{S(\frac{1}{p-3},\frac{3}{p-3})}.$$

These integrals are known. With

$$B(a,b) = \int_0^1 s^{a-1}(1-s)^{b-1}\,ds$$

we have[18]

$$T(a,b) = \int_0^\infty \frac{s^{a-1}\,ds}{(1+s)^b} = B(a,b-a)$$

and[19]

$$S(a,b) = \int_0^\infty\int_0^\infty \frac{(st)^{a-1}\,dsdt}{(1+s+t)^b} = T(a,b)T(a,b-a),$$

so V_p can be expressed in p via β-functions. It should lead to what we get in Exercise 19.13, which is really nice[20].

Exercise 19.9. 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.

[18] Via $s = \frac{t}{1-t}$, a substitution I avoided for (19.8).

[19] Via $t = (1+s)\tau$.

[20] There were mistakes in an earlier version and then it did not, but now it does.

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 19.10. Verify that computing the area of (19.41) using polar coordinates is even a bigger pain than using the $y = tx$-trick.

Exercise 19.11. In Exercise 19.9 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 19.6 from the formula for $x(t)$ in (19.40).

Exercise 19.12. Verify (19.46) by putting $y = tx$ in (19.45), solve for x, and set $t^p = s$ in the integral you get from Exercise 19.11 and convert to β-functions.

Exercise 19.13. See Exercise 19.11. 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

$$x = x(s, t) = \left(\frac{st}{1 + s^p + t^p} \right)^{\frac{1}{p-3}}$$

[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.

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 19.14. Let $p > 4$. The 4-dimensional measure of the bounded open set in $\mathbb{R}^4$ with all coordinates positive and bounded by

$$x_1^p + x_2^p + x_3^p + x_4^p = x_1 x_2 x_3 x_4$$

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[24] 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....

[24] Exponents and dimensions, another story: Chapter 23.

20 Cut-off functions and partitions of unity

This chapter explains the basic tools for cutting up functions in smaller parts which are localised. This involves two tricks, each of which you can play with.

The first trick concerns an open set $O \subset \mathbb{R}^N$ and a compact subset $K \subset O$ which should be non-empty[1]. Then every $a \in K$ is contained in an open ball B centered at a such that the closed ball with the same center but twice the radius is contained in O. We denote this larger ball by $2B$. Thus we have

$$K \ni a \in B \subset 2B \subset O.$$

These balls cover K and the (sequential) compactness[2] of K implies that K is covered bij finitely many of such balls, i.e.

$$K \subset B_1 \cup \cdots \cup B_k \subset 2B_1 \cup \cdots \cup 2B_k \subset O.$$

On each such ball $2B_i$ we choose a smooth function $\eta_i \in C_c^\infty(2B)$ with $0 \leq \eta_i \leq 1$ and $\eta_i \equiv 1$ on B_i, and we extend these functions[3] to the whole of $\mathbb{R}^N$ by setting $\eta_i \equiv 0$ outside $2B_i$. Then $\eta_i \in C_c^\infty(\mathbb{R}^N)$ for $i = 1, \ldots, k$ and a new function $\chi \in C_c^\infty(\mathbb{R}^N)$ may be defined by[4]

$$1 - \chi(x) = (1 - \eta_1(x)) \cdots (1 - \eta_k(x)). \tag{20.1}$$

Indeed, if x is not contained in the union of the supports of $\eta_1, \ldots, \eta_k$ then all factors in the right hand side of (20.1) are equal to 1 and hence $\chi(x) = 0$. On the other hand, if x is contained in one of the balls B_i then the corresponding factor in the right hand side of (20.1) is equal to zero making the right hand side vanish whence $\chi(x) = 1$. In particular $\chi(x) \equiv 1$ on K. Moreover, since all factors take values in $[0, 1]$ the same holds for $\chi(x)$, for any $x \in \mathbb{R}^N$. We conclude that

$$\chi \in C_c^\infty(O), \quad \forall x \in \Omega \quad \chi(x) \in [0, 1], \quad \forall x \in K \quad \chi(x) = 1, \tag{20.2}$$

and this is why χ is called a cut-off function for K in O.

The second trick applies the first trick to a finite collection of such sets

$$\emptyset \neq K_1 \subset O_1, \ldots, \emptyset \neq K_m \subset O_m \quad \text{with} \quad \eta_j \in C_c^\infty(O_j) \tag{20.3}$$

cut-off functions as in (20.2). We define $\zeta_j \in C_c^\infty(O_j)$ by

$$\zeta_j(x) = \frac{\chi_j(x)}{\chi_1(x) + \cdots + \chi_m(x)} \tag{20.4}$$

[1] The set K could be the closure of a bounded domain Ω, or its boundary.

[2] This characterisation of compactness was not discussed yet in these notes.

[3] We can use the p-norm to our liking, the choice $p = 2$ allows radially symmetric η_i.

[4] I first saw this elegant trick in Folland's Real Analysis book.

and extend ζ_j to $\mathbb{R}^N$ via $\zeta_j(x) \equiv 0$ outside O_i. Note that below we don't really use the last part of (20.2) as $\chi_j(x) > 0$ for all $x \in K_i$ suffices to obtain the essential properties of the collection $\zeta_1, \ldots, \zeta_m$, which is called a partition of unity. For every x for which one of the $\chi_j(x) > 0$ it follows that

$$\zeta_1(x) + \cdots + \zeta_m(x) = 1. \tag{20.5}$$

Certainly this holds for x in $K_1 \cup \cdots \cup K_m$. On the other hand, outside the union of $O_1, \ldots, O_m$ this sum is by definition equal to zero.

Any function

$$f : K_1 \cup \cdots \cup K_m \to \mathbb{R}$$

splits up via

$$f(x) = f_1 + \cdots + f_m = \zeta_1(x)f(x) + \cdots + \zeta_m(x)f(x),$$

with the smaller parts $f_j = \zeta_j f$ compactly supported in O_j, and ζ_j not harming any smoothness the original function f may enjoy. Adding more K_j to the collection changes the functions ζ_j only via (20.4), with (20.5) remaining valid.

20.1 Partitions of compact manifolds

This section was written before Section 20.3. For Ω and $M = \partial\Omega$ you can jump to the end of this section. We now apply the techniques in Section 20 to a non-empty compact set $M \subset \mathbb{R}^N$ for which (C) in Section 13.2 applies in a sense we made more precise in Section 20.3 specifying blocks $[\tilde{a}, \tilde{b}] \subset \mathbb{R}^N$ in which the description (A) of Section 13.2 can be given, see (20.13). Below we rather choose blocks $[\tilde{a}_i, \tilde{b}_i] \subset \mathbb{R}^n$, given Φ_i as in (19.1). Thus for each $p \in M$ there exists a continuously differentiable injective

$$\Phi_i : [a, b] = [a_1, b_1] \times \cdots \times [a_n, b_n] \to \mathbb{R}^N$$

with $\mathcal{M}(\frac{\partial\Phi}{\partial u}) > 0$ such that $p \in \Phi((\tilde{a}, \tilde{b}))$ for some $[\tilde{a}, \tilde{b}] \subset (a, b)$, and in some open neighbourhood O of the compact set $K = \Phi([\tilde{a}, \tilde{b}])$ it holds that

$$x \in M \iff x \in \Phi((a, b)) \tag{20.6}$$

We would now like to consider the sets $\Phi((\tilde{a}, \tilde{b}))$ as open sets covering M, so that by compactness

$$M \subset \Phi_1((\tilde{a}_1, \tilde{b}_1)) \cup \cdots \cup \Phi_m((\tilde{a}_m, \tilde{b}_m)), \tag{20.7}$$

for some finite collection Φ_j, but clearly the sets $\Phi((\tilde{a},\tilde{b}))$ are not open[5] in $\mathbb{R}^N$, unless $n = N$. Nevertheless such a finite subcover exists.

To see this first choose $[\underline{a},\underline{b}] \subset (\tilde{a},\tilde{b})$ with $p \in \Phi((\underline{a},\underline{b}))$ and a suitable open neighbourhood $\underline{O}$ of $\underline{K} = \Phi([\underline{a},\underline{b}])$ with $\underline{O} \subset O$ to have the characterisation in (20.6) hold for all $x \in \underline{O}$ as well, and such that $\underline{O}$ does not intersect the (compact) image under Φ of the compact set $[a,b]\backslash(\tilde{a},\tilde{b})$. It then follows that $M \cap \underline{O} \subset \Phi((\tilde{a},\tilde{b}))$ because Φ is injective.

Varying $p \in M$ the open sets $\underline{O}$ cover M and by compactness there exists a finite collection $O_1,\dots,O_m$ such that

$$M \subset \underline{O}_1 \cup \cdots \cup \underline{O}_m \subset \Phi_1((\tilde{a}_1,\tilde{b}_1)) \cup \cdots \cup \Phi_m((\tilde{a}_m,\tilde{b}_m)),$$

which is the desired finite covering (20.7) consisting of patches.

We can now put $K_j = \Phi_j([\tilde{a}_j,\tilde{b}_j])$ and the corresponding open neighbourhoods O_j of K_j in which (20.6) characterises the elements of M. The description following (19.1) in Section 19 with unit blocks then results from Section 20.

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 (20.7), but it is then slightly more complicated to formulate (20.4), because each χ_j is then a function of u. This allows us to deal with manifolds which are not necessarily embedded in $\mathbb{R}^N$.

Finally we observe that any Ω and $M = \partial\Omega$ as in Section 18.1 allow a choice of functions $\zeta_1,\dots,\zeta_n \in C_c^\infty((a_i,b_i))$ with $0 \le \zeta_i \le 1$ and $\zeta_1+\cdots+\zeta_n \equiv 1$ on a neighbourhood of Ω such that every for every i either $[a_i,b_i] \subset \Omega$ holds, or $P_i = M \cap (a_i,b_i)$ is a patch such as in Section 18.1.

20.2 Changing partitions

We still have to check that the integrals do not depend on the choice of the partitioning functions $\zeta_1,\dots,\zeta_n$. We observe that (19.2) defines a linear map

$$f \xrightarrow{L} \int_M f\,dS_n \tag{20.8}$$

from $X = C(M)$, the space of continuous real valued functions on M, to $\mathbb{R}$. Note that L is bounded in the sense that $|Lf| \le C|f|_\infty$, just as in Section 5.7, but we will not be using this below[6].

The partition naturally defines linear subspaces

$$X_i = \{\zeta_i f :\ f \in C(M)\},$$

[5] Of course they should be open in M.

[6] But we will need it to get rid of the annoying assumption $\Phi \in C^2$ in Section 19.3.

and the same holds for any other partition of M, given by say $\eta_1, \ldots, \eta_J$, which also defines a linear map

$$f \xrightarrow{K} \int_M f \, dS_n \tag{20.9}$$

via (19.2), 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

$$f = f \sum_{i=1}^{I} \zeta_i = \sum_{i=1}^{I} \zeta_i f = \sum_{i=1}^{I} \zeta_i f \sum_{j=1}^{J} \eta_j = \sum_{i=1}^{I} \sum_{j=1}^{J} \zeta_i \eta_j f, \tag{20.10}$$

whence

$$Lf = L(\sum_{i=1}^{I} \zeta_i f) = \sum_{i=1}^{I} \int_{\Phi_i} \zeta_i f \, dS_n = \sum_{i=1}^{I} \sum_{j=1}^{J} \int_{\Phi_i} \zeta_i \eta_j f \, dS_n,$$

and likewise

$$Kf = \sum_{j=1}^{J} \sum_{i=1}^{I} \int_{\Psi_j} \eta_j \zeta_i f \, dS_n,$$

and thus it remains to show that

$$\int_{\Phi_i} \zeta_i \eta_j f \, dS_n = \int_{\Psi_j} \eta_j \zeta_i f \, dS_n \tag{20.11}$$

The integral on the left is defined via (18.33) 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 (18.33) 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 20.4. These were defined in neighbourhoods of a given points $p \in \Phi_i((a,b)) \cap \Psi_j((c,d))$ only. Therefore we need another localisation argument[7] before we can apply Section 15.4 to conclude that the two integrals are the same.

[7] Try this one by yourself.

20.3 Again: local descriptions of a manifold

Let us be very precise in what we established for the local descriptions as in (A), (B) and (C) of Section 13.2, 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

$$[a, b] = [a_1, b_1] \times \cdots \times [a_n, b_n] \times [a_{_N}, b_{_N}] \times \cdots \times [a_{\mathrm{N}}, b_{\mathrm{N}}] \subset \mathbb{R}^{\mathrm{N}}$$

and that for some $p \in (a, b)$ the derivative $F'(p)$ is of maximal rank. Renaming and relabeling the variables in $z = (x, y)$ we can then arrange for the "partial" derivative $F_y(p)$ to be invertible. Theorem 13.2 then implies that there exists $(\tilde{a}, \tilde{b}) \subset (a, b)$ with $p \in (\tilde{a}, \tilde{b})$ and a continuously differentiable function

$$f : [\tilde{a}_x, \tilde{b}_x] \to (\tilde{a}_y, \tilde{b}_y)$$

such that $p = (p_x, p_y) \in (\tilde{a}, \tilde{b})$ and

$$F^{-1}(p) \cap [\tilde{a}, \tilde{b}] = \{(x, f(x)) : x \in [\tilde{a}_x, \tilde{b}_x]\} \subset [\tilde{a}_x, \tilde{b}_x] \times (\tilde{a}_y, \tilde{b}_y), \qquad (20.12)$$

with subscripts indicating the x and the y-parts of p, $\tilde{a}$ and $\tilde{b}$. Thus in the smaller block $[\tilde{a}, \tilde{b}]$ the level set of $F(p)$ coincides with the graph of f, and in the same block $[\tilde{a}, \tilde{b}]$ this graph then coincides with the zero-level set of $\tilde{F}(z) = \tilde{F}(x, y) = y - f(x)$.

As for (C), if we have, with subscripts denoting the x and the y-parts of Φ, that $\Phi(u) = (\Phi_x(u), \Phi_y(u))$ is continuously differentiable on $[a, b]$ with $0 \in (a, b)$ and $p = \Phi(0)$, then via Theorem 13.3 the invertibility of $\Phi'_x(0)$ is sufficient for the existence of $[\underline{a}_x, \underline{b}_x]$ with $p_x \in (\underline{a}_x, \underline{b}_x)$ and a continuously differentiable function $\phi : [\underline{a}_x, \underline{b}_x] \to (a, b)$ such that $\Phi_x(\phi(x)) = x$ for all $x \in [\underline{a}_x, \underline{b}_x]$. Moreover[8], we can choose $[\underline{a}_x, \underline{b}_x]$ such that $\phi((\underline{a}_x, \underline{b}_x))$ is an open set as the inverse image of $(\underline{a}_x, \underline{b}_x)$ under Φ_x.

The function f defined by $f(x) = \Phi_y(\phi(x))$ now defines a graph

$$\{(x, f(x)) : x \in [\underline{a}_x, \underline{b}_x]\}$$

which is a subset of $\Phi([a, b])$. If in addition Φ is injective on $[a, b]$ then the image under Φ of the closed bounded set $[a, b] \backslash \phi((\underline{a}_x, \underline{b}_x))$ is bounded and closed, and does not contain p. Thus there exists a block $[\tilde{a}, \tilde{b}]$ with $p \in (\tilde{a}, \tilde{b})$ such that $[\tilde{a}_x, \tilde{b}_x] \subset (\underline{a}_x, \underline{b}_x)$ with

$$\Phi([a, b] \backslash \phi((\underline{a}_x, \underline{b}_x))) \cap [\tilde{a}, \tilde{b}] = \emptyset.$$

[8] See the discussion after Theorem 13.3.

The continuity of f implies that we can restrict $\tilde{a}_x$ and $\tilde{b}_x$ a bit further to ensure that $f([\tilde{a}_x, \tilde{b}_x]) \subset (\tilde{a}_y, \tilde{b}_y)$. We note we also have that

$$\Phi([a,b]\backslash\phi((\tilde{a}_x, \tilde{b}_x))) \cap [\tilde{a}, \tilde{b}] = \emptyset,$$

since the additional points in the larger image $\Phi([a,b]\backslash\phi((\tilde{a}_x, \tilde{b}_x)))$ are on the graph of f outside $[\tilde{a}_x, \tilde{b}_x]$. Thus we have arrived from (C) to exactly the same formulation of (A) as above starting from (B): $p \in (\tilde{a}, \tilde{b})$ and

$$\Phi([a,b] \cap [\tilde{a}, \tilde{b}] = \{(x, f(x)) : x \in [\tilde{a}_x, \tilde{b}_x]\} \subset [\tilde{a}_x, \tilde{b}_x] \times (\tilde{a}_y, \tilde{b}_y). \qquad (20.13)$$

The two statements (20.12) and (20.13) should be compared to the definition Edwards gives in Section 4 of his Chapter III for $M \subset \mathbb{R}^N$ to be an n-dimensional manifold. Every $p \in M$ should, after relabeling and renaming in $z = (x, y)$, be contained in an open set O in which

$$P = O \cap M = \{(x, f(x)) : x \in U\},$$

with $U \subset \mathbb{R}^n$ open and $f : U \to \mathbb{R}^m$ continuously differentiable, is called a C^1-patch of M. Of course it is then clear that $U \supset [\tilde{a}_x, \tilde{b}_x] \supset (\tilde{a}_x, \tilde{b}_x)$ and $O \supset [\tilde{a}, \tilde{b}] \supset (\tilde{a}, \tilde{b})$ for some $(\tilde{a}, \tilde{b}) \ni p$, and thus it is completely equivalent to ask that $p \in (\tilde{a}, \tilde{b})$ and

$$p \in M \cap [\tilde{a}, \tilde{b}] = \{(x, f(x)) : x \in [\tilde{a}_x, \tilde{b}_x]\} \subset [\tilde{a}_x, \tilde{b}_x] \times (\tilde{a}_y, \tilde{b}_y) \qquad (20.14)$$

for some closed block $[\tilde{a}, \tilde{b}]$ with $(\tilde{a}_x, \tilde{b}_x) \ni p_x$, and some continuously differentiable $f : [\tilde{a}_x, \tilde{b}_x] \to (\tilde{a}_y, \tilde{b}_y)$, exactly as in (20.12,20.13), the patch being

$$M \cap (\tilde{a}, \tilde{b}) = \{(x, f(x)) : x \in (\tilde{a}_x, \tilde{b}_x)\} \ni p = (p_x, f(p_x)). \qquad (20.15)$$

In the closed N-block $[\tilde{a}, \tilde{b}]$ there are no other points of M than the points on the graph of $f : [\tilde{a}_x, \tilde{b}_x] \to (\tilde{a}_y, \tilde{b}_y)$.

20.4 Coordinate transformations

By definition every $p \in M$ is in such a patch as above and typically patches overlap. If p is in two such patches, say with functions f and g, it may happen that f and g are functions of the x-part of z. In that case the patches are parameterised by

$$u \to \Phi(u) = (u, f(u)) \quad \text{and} \quad v \to \Psi(v) = (v, g(v)) \qquad (20.16)$$

defined on overlapping blocks with p_x in the interior of the intersection of the blocks, which is an open block itself. The common part of M is then contained in the intersection of the two N-blocks.

Viewing the n-tuples u and v as local coordinates on M near p, a transformation of these coordinates is simply given by $v = u$. In all other cases, we may renumber the variables of $\mathbb{R}^{\mathrm{N}}$ 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 $\mathrm{N} = n + 1$.

To rewrite Ψ in the form Φ we need the invertibility of respectively

$$\frac{\partial g_2}{\partial v_3} \quad \text{and} \quad \frac{\partial g_1}{\partial v_2}, \tag{20.17}$$

in which case we we obtain respectively

$$w \to \tilde{\Phi}(w) = (w_1, w_2, h_3(w_1, w_2), h_4(w_1, w_2))$$

and

$$w \to \tilde{\Phi}(w) = (w_1, h_2(w_1), h_3(w_1))$$

as local descriptions of the Ψ-patches near p. The definition of what a manifold is then implies that

$$\tilde{\Phi} \equiv \Phi$$

on an open block containing (p_1, p_2) in the first case and p_1 in the second case. It then follows as above that $u = w$ is a coordinate transformation just as $u = v$ for (20.16) while w is obtained from v via a coordinate transformation just as x from u in the proof of (A) from (C) above.

It thus remains to establish the invertibility of the partial Jacobian matrices in (20.17) in p to conclude there exists a local C^1-transformation from u to v near p. Note that these are also the conditions for solving part[9] of $\Phi(u) = \Psi(v)$ via

$$v_1 = u_1,\ v_3 = f_3(u_1, u_2) \quad \text{and} \quad u_1 = v_1,\ u_2 = g_2(v_1, v_3) \tag{20.18}$$

in the first case, and

$$u_1 = g_1(v_2) \quad \text{and} \quad v_2 = f_2(u_1) \tag{20.19}$$

[9] All equations but the last one, which then requires some argument to hold as well.

in the second case. The invertiblity of the partial Jacobian matrices in (20.17) in p follows because otherwise the Ψ-patch cannot achieve all respectively (u_1, u_2)-directions and u_1-directions that occur in the Φ-patch, contradicting the assumption that the Ψ-patch covers all of M in its defining neighbourhood.

The restriction to patches of the form (20.15) looks like an obvious choice for simplicity, but may bother us later when dealing with (19.33), we'll see.

21 Standing at the crossroads of PDE and FA

This chapter relates to the courses in Functional Analysis and Partial Differential Eqations as given in the bachelor pogrammes in Amsterdam, as well as courses run under the same name in the national mastermath programme. Have a look at Section 9.6, in particular at the solution method for (9.18), which clearly does not generalise to the problem of solving

$$-\Delta u = f \quad \text{in} \quad \Omega \quad \text{with} \quad u = 0 \quad \text{on} \quad \partial\Omega \tag{21.1}$$

for given $f : \Omega \to \mathbb{R}$ and $\Omega \subset \mathbb{R}^N$ on a bounded open set with boundary $\partial\Omega$.

Moreover, even is $\partial\Omega$ is smooth, it does not in general hold that (21.1) has a twice differentiable solution which solves the partial differential equation $-\Delta u = f$ for the given $f \in C(\bar{\Omega})$. Below we show another way of solving (9.18) which does generalise to a large class of problems including (21.1). This technique is based on integration by parts[1] and the theory of Hilbert spaces, mainly the unique and obvious generalisation[2] of the inner product space $\mathbb{R}^2$ with the dimension 2 replaced by the first infinite cardinal.

Als we de vergelijking in (9.18) vermenigvuldigen met een functie $v \in C^1[0,1]$ dan bestaan onder de aanname dat $u \in C^2[0,1]$ en $f \in C^0[0,1]$ beide integralen

$$-\int_0^1 u''(x)v(x)\,dx = \int_0^1 f(x)v(x)\,dx$$

en kan de linkerkant partieel geintegreerd worden. Het resultaat is

$$-[u'(x)v(x)]_0^1 + \underbrace{\int_0^1 u'(x)v'(x)\,dx}_{\text{symmetrisch in } u,v} = \underbrace{\int_0^1 f(x)v(x)\,dx}_{\text{symmetrisch in } f,v}, \tag{21.2}$$

waarbij de lelijke eerste term verdwijnt als we de extra aanname maken dat $v(0) = v(1) = 0$.

De oplossing $u \in C^2[0,1]$ van (9.18) heeft dus de eigenschap dat $u(0) = u(1) = 0$ en

$$\int_0^1 u'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx \quad \forall v \in C_0^1[0,1], \tag{21.3}$$

waarin

$$C_0^1[0,1] = \{v \in C^1[0,1] : \, v(0) = v(1) = 0\},$$

[1] Now have a look at (18.12).

[2] In different guises.

en de gelijkheid in (21.3) kan voor elke $u \in C_0^1[0,1]$ geverifieerd worden. Kortom, we zouden dus kunnen afspreken om $u \in C_0^1[0,1]$ een oplossing van (9.18) te noemen als aan (21.3) voldaan is.

Rechts in (21.3) zien we een integraal die te zien is als een inwendig product van f en v, dat we kunnen noteren als

$$f \cdot v = \int_0^1 f(x)v(x)\,dx, \tag{21.4}$$

waarmee (9.18) zich uiteindelijk herschrijft als

$$u \in C_0^1[0,1], \quad \underbrace{u' \cdot v'}_{((u,v))} = \underbrace{f \cdot v}_{(f,v)} \quad \forall v \in C_0^1[0,1], \tag{21.5}$$

een uitdrukking waarin twee inwendige producten voorkomen en $u \in C_0^1[0,1]$ en $f \in C^0[0,1]$ vast zijn, en $v \in C_0^1[0,1]$ willekeurig.

Exercise 21.1. Waarom is $(u,v) \to u' \cdot v'$ wel een inprodukt op $C_0^1[0,1]$ en niet op $C^1[0,1]$?

Exercise 21.2. Laat $L > 0$, bijvoorbeeld $L = 2\pi$ of $L = 1$. De vectorruimte van continue L-periodieke functies noemen we $C(\mathbb{R}_L) = C^0(\mathbb{R}_L)$, en de deelruimte van k keer ($k \in \mathbb{N}$) continu differentieerbare functies noemen we $C^k(\mathbb{R}_L)$. Voor welke $f \in C(\mathbb{R}_L)$ is de vergelijking $-u'' = f$ oplosbaar met u in $C^2(\mathbb{R}_L)$? Is de oplossing uniek? Hint: neem eerst $L = 1$ natuurlijk.

Exercise 21.3. Neem $L = 1$ en los de vergelijking $-u'' = f$ voor f continu, 1-periodiek met $\int_0^1 f(x)\,dx = 0$: geef een uitdrukking van de vorm (9.21) voor de oplossing u die (ook) voldoet aan $\int_0^1 u(x)\,dx = 0$. Hint: zonder deze laatste conditie is de oplossing niet uniek bepaald en evenzo is de (wederom symmetrische!) kern $A(x,s)$ niet uniek bepaald. Maar wel onder de conditie dat $\int_0^1 A(x,s)\,dx = \int_0^1 A(x,s)\,ds = 0$.

Exercise 21.4. Laat

$$\bar{C}^k(\mathbb{R}_L) = \{u \in C^k(\mathbb{R}_L) : \int_0^L u(x)dx = 0\}$$

en geef een herformulering van $-u'' = f$ voor $f \in \bar{C}^0(\mathbb{R}_L)$ en $u \in \bar{C}^1(\mathbb{R}_L)$ zoals in (21.5).

Hopelijk is duidelijk dat de twee laatste opgaven kwa bewerkelijkheid nogal uiteenliepen. Formuleringen als in (21.5) gebaseerd op *integration by parts*, zonder dat het doel daarvan het uitrekenen van getallen is, bieden een ander en vaak algemener perspectief om eigenschappen van oplossingsoperatoren te begrijpen dan expliciete oplossingsmethoden gebaseerd op primitiveren. In wat volgt zullen we daartoe v als een variabele zien en

$$v \to ((u, v)) \quad \text{en} \quad f \to (f, v)$$

als dezelfde lineaire functie, maar anders gepresenteerd.

Lineaire functies en inprodukten, hoe zit dat? Hoe weet je dat er bij f via dezelfde lineaire functie van v een u hoort? En kan dat algemener? Voor later[3]. Dit hoofdstuk besluiten we met een paar opgaven die laten zien hoe intrinsiek de herformulering als (21.5) verbonden is met de eigenschappen van de oplossingsoperator

$$A : f \xrightarrow{\forall v((u,v))=(f,v)} u, \tag{21.6}$$

waarbij het van het specifieke probleem afhangt welke inproducten en welke functieruimten gedefinieerd moeten worden om een A te maken die in simpele gevallen samenvalt met expliciet uitgerekende integraaloperatoren als in (9.21).

Exercise 21.5. Laat zien dat een solver als A in (21.6) de eigenschap heeft dat $((Au, v)) = ((u, Av))$ en $(Af, g) = (f, Ag)$ voor alle u, v en f, g in de nog te kiezen ruimten $V \subset H$ en H waarop de inproducten zijn gedefinieerd met de eigenschappen die we nodig hebben om alles precies te maken.

De operator A is dus symmetrisch[4] met betrekking tot twee inproducten, waaronder het 'gewone' inproduct (21.4) dat in eerste instantie was opgeschreven onder verschillende aannames voor f en v.

Exercise 21.6. Gebruik de symmetrie van A om te laten zien dat eigenvectoren[5] van A bij verschillende eigenwaarden van A loodrecht op elkaar staan.

[3] See Chapter 24.

[4] We praten nog niet over complexwaardige functies hier.

[5] $A\phi = \lambda\phi$, $\lambda \in \mathbb{R}$, ϕ een (eigen)functie.

Exercise 21.7. Gebruik de vorige opgave en Opgave 21.4 om zonder rekenwerk te laten zien dat

$$\int_{-\pi}^{\pi} \sin nx \sin mx\, dx = 0 = \int_{-\pi}^{\pi} \cos nx \cos mx\, dx \quad (m, n \in \mathbb{N},\ m \neq 0)$$

$$\int_{-\pi}^{\pi} \sin nx \cos mx\, dx = 0 \quad (m, n \in \mathbb{N})$$

Natuurlijk wist je dit al, waarschijnlijk via $\exp(inx) = \cos nx + i \sin nx$ en de gebruikelijke rekenregeltjes gebaseerd op de somformules[6] voor $\cos(a + b)$ en $\sin(a + b)$ die niet meer tot de tegenwoordig zelden precies gerechtvaardigde basiskennis van de gemiddelde β-student horen. De functie sin kan gedefinieerd worden als de unieke oplossing van het beginwaardeprobleem

$$u'' + u = 0; \quad u(0) = 0; \quad u'(0) = 1, \tag{21.7}$$

en cos als de afgeleide van sin. Alle eigenschappen van cos en sin, i.h.b. de somformules volgen uit deze definities en kunnen gebruikt worden voor de volgende opgave.

Exercise 21.8. Bepaal alle $\lambda > 0$ waarvoor $u'' + \lambda u = 0$ oplossingen van periode 2π heeft en bepaal alle even en oneven oplossingen voor die waarden λ.

De even oplossingen die je zo vindt zijn veelvouden van $c_1 : x \to \cos x$, $c_2 : x \to \cos 2x$, $c_3 : x \to \cos 3x$, ..., en de oneven oplossingen zijn veelvouden van $s_1 : x \to \sin x$, $s_2 : x \to \sin 2x$, $s_3 : x \to \sin 3x$, ..., en ieder tweetal van deze functies staat loodrecht op elkaar, zoals we in Opgave 21.7 gezien hebben. En ze zijn gemiddeld allemaal nul, hetgeen betekent dat ze loodrecht staan op de functie $\mathbf{1} : x \to 1$, bijvoorbeeld

$$(\mathbf{1}, s_1) = \mathbf{1} \cdot s_1 = \int_{-\pi}^{\pi} 1 \sin x\, dx = \int_{-\pi}^{\pi} \sin x\, dx = 0.$$

Wat we in vervolg gaan doen is $1, c_1, c_2, c_3, \ldots, s_1, s_2, s_3, \ldots$ zien als vectoren die lijnen door de oorsprong[7] definiëren. En die lijnen zien we als een assenkruis waarmee we een oneindig-dimensionale ruimte opspannen, een ruimte waarin we willen werken zoveel mogelijk als we dat in het platte vlak doen.

[6] Zie Wiskunde in je Vingers, sectie 10.4.

[7] Die oorsprong is de nulfunctie $\mathbf{0} : x \to 0$.

22 Lebesgue spaces

If you are already familiar with Lebesgue spaces you may like to jump to Section 22.3 and flip back when needed. The following definitions usually come at the end of a course on measure theory and Lebesgue integration.

Definition 22.1. *Let $U \subset \mathbb{R}^N$ be open and $p \geq 1$. A measurable function $u : U \to \mathbb{R}$ is said to be in $L^p_{loc}(U)$ if and only if*

$$\int_B |f|^p < \infty$$

for every open ball $B \subset U$, and in $L^p(U)$ if the p-norm of u defined by

$$|f|_p^p = \int_U |f|^p < \infty \tag{22.1}$$

exists.

Modulo hassle needed to deal with $|f|_p = 0$ not implying that $f(x) = 0$ for all $x \in U$, but only that

$$\{x \in \mathbb{R}^N : f(x) \neq 0\}$$

is a set of zero measure[1], the normed space $L^p(U)$ is Banach space with its norm defined by (22.1).

Remark 22.2. *Every $f \in L^p(U)$ extends to $f \in L^p(\mathbb{R}^N)$ by setting $f(x) = 0$ for $x \notin U$. No such general[2] statement holds for $f \in L^p_{loc}(U)$ and $L^p_{loc}(\mathbb{R}^N)$.*

We recall from the discussion in Section 12.9 about (12.36,12.37) that

$$|\sum_{i=1}^n a_i b_i| \leq |a|_p \, |b|_q \quad \text{for} \quad p, q > 1 \quad \text{with} \quad \frac{1}{p} + \frac{1}{q} = 1, \tag{22.2}$$

Hölder's inequality for finite sums of real numbers. Memorise that

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

and convince yourself that via any definition of the integral it also holds that

$$|\int_U fg| \leq |f|_p \, |g|_q \tag{22.3}$$

with the norms defined by (22.1).

[1] Here $|A|$ denotes the Lebesgue measure of a Lebesgue measurable subset $A \subset \mathbb{R}^N$.

[2] Example: $p = N = 1$, $f(x) = \frac{1}{x}$, $U = \mathbb{R}_+$.

Exercise 22.3. Explain why the spaces $L^p_{loc}(U)$ are nested: $L^p_{loc}(U) \subset L^q_{loc}(U) \subset L^1_{loc}(U)$ if $p \geq q \geq 1$. Hint: use (22.3) with $g \equiv 1$ to show that the spaces $L^p(U)$ are nested if U is bounded.

Exercise 22.4. No estimate of the type

$$|u|_p \leq C_{pqN}|u|_q$$

with $p > q \geq 1$ can hold for all $u \in C_c(\mathbb{R}^N)$. Why? Show that the spaces $L^p(\mathbb{R}^N)$ are not nested.

Exercise 22.5. Apply (22.3) to f^a and f^b to show that

$$|f|^{a+b}_{a+b} \leq |f|^a_{ap}|f|^b_{bq}$$

and solve the equations $1 \leq ap = r < a + b = s < bq = t$ and $(p-1)(q-1) = 1$ to obtain an (interpolation) inequality for $|f|_s$ in terms of $|f|_r$ and $|f|_t$. This shows that

$$L^r(U) \cap L^t(U) \subset L^s(U) \quad \text{for} \quad r < s < t.$$

Discuss the limit case $t = \infty$.

22.1 The Lebesgue's Differentiation Theorem

Since Lebesgue we see every $f \in L^1(\mathbb{R}^N)$ as an equivalence class[3] F of integrable measurable functions $f : \mathbb{R}^N \to \mathbb{R}$ for the equivalence relation

$$f \sim g \iff |\{x \in \mathbb{R}^N : f(x) \neq g(x)\}| = 0 \iff \int_{\mathbb{R}^N} |f - g| = 0.$$

In this chapter we shall also explore and perhaps prefer the alternative approach, similar tot the construction[4] of $\mathbb{R}$ out of $\mathbb{Q}$, namely as equivalence classes of Cauchy sequences f_n in $C_c(\mathbb{R}^N)$ with respect to the p-norm. This will completely avoid the notion of Lebesgue measure and all integrals will be limits of integrals of continuous functions.

[3] Often denoted as $[f]$, so $f \in [f] = F$.

[4] Which for $\mathbb{R}$ may obscure what was already obvious.

Remark 22.6. *We have that $|A| = 0$ if and only if for every $\varepsilon > 0$ there exist a sequence of open balls $B_n = B(x_n, r_n)$ indexed by $n \in \mathbb{N}$ such that*

$$A \subset \cup_{n\in\mathbb{N}} B_n \quad \text{and} \quad \sum_{n\in\mathbb{N}} |B_n| < \varepsilon,$$

in which

$$|B_n| = \omega_{_N} r_n^N.$$

Note that this zero measure concept does not even involve the Lebesgue measure of the covering countable union of (open) balls, but it does contain the fundamental idea that measure theory should deal with countable unions.

Exercise 22.7. Prove that a countable union of zero measure sets is again a zero measure set. Hint: every small $\varepsilon > 0$ is the sum of countably many smaller positive epsilons; this goes back to Zeno and Section 1.5.

Remark 22.8. *No matter how we define $L^1(\mathbb{R}^{\mathrm{N}})$, a fundamental truth is that for every $f \in L^1(\mathbb{R}^{\mathrm{N}})$ and every open ball $B(x, r)$ the integral*

$$\int_{B(x,r)} f$$

is independent of the choice of $f \in F = [f]$ for whatever concept of equivalence and equivalence classes used to define $L^1(\mathbb{R}^{\mathrm{N}})$, and varies[5] continuously with $x \in \mathbb{R}^{\mathrm{N}}$ and $r \geq 0$. Moreover

$$|\int_{B(x,r)} f| \leq \int_{B(x,r)} |f| \to |f|_{_1} \quad \text{as} \quad r \to \infty,$$

$$B(x, r) \subset B(y, s) \implies \int_{B(x,r)} |f| \leq \int_{B(y,s)} |f|,$$

and[6]

$$\int_{B_1\cup\cdots\cup B_n} |f| = \int_{B_1} |f| + \cdots + \int_{B_n} |f|$$

for balls $B_1, \ldots, B_n$ with $B_i \cap B_j = \emptyset$ if $i \neq j$.

[5] This follows from the dominated convergence theorem for Lebesgue integrals in fact.

[6] The finite additivity of the integral over disjoint unions of open balls.

As a consequence also the average

$$A_f(x,r) = \fint_{B(x,r)} f = \frac{1}{|B(x,r)|}\int_{B(x,r)} f = A_{x,r}f \tag{22.4}$$

of f over $B(x,r)$ varies continuously with $x \in \mathbb{R}^N$ and $r > 0$. The function

$$(x,r) \xrightarrow{A_f} \fint_{B(x,r)} f$$

is continuous from $\mathbb{R}_+ \times \mathbb{R}^N$ to $\mathbb{R}$, and independent of the choice of $f \in F$. For fixed $x \in \mathbb{R}^N$ and $r > 0$ the map

$$f \xrightarrow{A_{x,r}} \fint_{B(x,r)} f$$

is linear and continuous from $L^1(\mathbb{R}^N)$ to $\mathbb{R}$, and the estimate

$$|A_{x,r}f| \le A_{x,r}|f|$$

holds for every $f \in L^1(\mathbb{R}^N)$.

Definition 22.9. *The good set of a function $f \in L^1_{loc}(\mathbb{R}^N)$ is defined by*

$$G_f = \{x \in \mathbb{R}^N : \lim_{r\downarrow 0} A_f(x,r) = f(x)\}. \tag{22.5}$$

Clearly the existence and value of the limit does not rely on the choice of $f \in F$. If we set

$$\mathcal{N}_f = \{x \in \mathbb{R}^N : \lim_{r\downarrow 0} A_f(x,r) \text{ does not exist}\},$$

then the complement of the set $\mathcal{N}_f$ contains the good set G_f of every $f \in F$.

Theorem 22.10. *For every $f \in L^1_{loc}(\mathbb{R}^N)$ the good set G_f has a complement with zero measure*[7]*. This complement contains the set $\mathcal{N}_f$, which is therefore not that bad: it also has zero measure, and it is natural to choose the unique $f \in F$ for which*

$$f(x) = \lim_{r\downarrow 0} A_f(x,r)$$

for all $x \notin \mathcal{N}_f$, and $f(x) = 0$ for all $x \in \mathcal{N}_f$. In that case $\mathbb{R}^N$ is the disjoint union of G_f and $\mathcal{N}_f$. For every $x \in G_f$ the value of $f(x)$ is what it should be, and for every $x \in \mathcal{N}_f$ the value of $f(x)$ is irrelevant as far as integrals are concerned, and chosen to be 0.

[7] This is the Lebesgue Differentiation Theorem, a name we do not explain yet.

Exercise 22.11. Explain why it suffices to prove Theorem 22.10 for $f \in L^1(\mathbb{R}^N)$.

Exercise 22.12. Prove the statements in Theorem 22.10 for $f \in C_c(\mathbb{R}^N)$ by showing that $G_f = \mathbb{R}^N$: the limit exists for every $x \in \mathbb{R}^N$ and is what it should be.

Remark 22.13. *Check out the literature to understand why this theorem is called the Lebesgue Differentation Theorem. We prove the theorem in Section 22.2 the way Folland does it, and provide an alternative proof starting from Section 22.3, avoiding measure theory and topology.*

22.2 The proof of the good set theorem

The proof of Theorem 22.10 invokes the Hardy-Littlewood function, defined for $f \in L^1(\mathbb{R}^N)$ by

$$H_f(x) = \sup_{r>0} A_{|f|}(x,r) = \sup_{r>0} \frac{1}{|B(x,r)|} \int_{B(x,r)} |f| \in [0,\infty],$$

the largest possible average of $|f|$ on balls centered in x. Since for every fixed $x \in \mathbb{R}^N$ we have

$$0 \le A_{|f|}(x,r) \le \frac{|f|_1}{\omega_N r^N},$$

the supremum $H_f(x)$ is finite unless $A_{|f|}(x,r) \to \infty$ as $r \to 0$. Note that

$$|A_f(x,r)| \le A_{|f|}(x,r) \le H_f(x) \le \infty. \tag{22.6}$$

We examine A_f via H_{f-g} with $g \in C_c(\mathbb{R}^N)$ and small $|f-g|_1 > 0$. Writing

$$A_f(x,r) - f(x) = \underbrace{A_f(x,r) - A_g(x,r)}_{A_{f-g}(x,r)} + \underbrace{A_g(x,r) - g(x)}_{\to 0 \ as \ r\to 0} + g(x) - f(x),$$

we use (22.6) with $f-g$ and observe that in the resulting inequality

$$|A_f(x,r) - f(x)| \le H_{f-g}(x) + \underbrace{|A_g(x,r) - g(x)|}_{\to 0 \ as \ r\to 0} + |g(x) - f(x)| \tag{22.7}$$

the role of r disappears as $r \to 0$. So if the left hand side is not small then the first or the third term is not small. Or both.

We therefore consider the sets[8]

$$O^\varepsilon = \{x \in \mathbb{R}^N : H_{f-g}(x) > \varepsilon\} \quad \text{and} \quad S^\varepsilon = \{x \in \mathbb{R}^N : |f(x) - g(x)| > \varepsilon\},$$

[8] O is for open, as we shall see.

and let W_ε be the set of all points $x \in \mathbb{R}^N$ for which the statement

$$\exists_{\delta>0}\, \forall_{r\in(0,\delta)} : \, |A_f(x,r) - f(x)| \le 2\varepsilon$$

fails. Then it must be that

$$W^\varepsilon \subset O^\varepsilon \cup S^\varepsilon. \tag{22.8}$$

These sets are nested,

$$0 < \delta < \varepsilon \implies O^\varepsilon \subset O^\delta, \quad S^\varepsilon \subset S^\delta \quad \text{and} \quad W^\varepsilon \subset W^\delta,$$

and via

$$\int_{\mathbb{R}^N} |f-g| \ge \int_{S^\varepsilon} |f-g| \ge \varepsilon |S^\varepsilon|$$

we have

$$\int_{\mathbb{R}^N} |f-g| > \varepsilon |S^\varepsilon|$$

unless both sides are zero. Note that this statement requires to have the Lebesgue measure of S^ε be well defined. Refering to Exercise 22.12 we may as well assume that $|f-g|_{_1} > 0$ and conclude from (22.8) that

$$|W^\varepsilon| < |O^\varepsilon| + \frac{1}{\varepsilon} |f-g|_{_1}. \tag{22.9}$$

Now suppose that

$$|O^\varepsilon| \le \frac{C_N}{\varepsilon} |f-g|_{_1} \tag{22.10}$$

for some universal N-dependent constant C_N. We can then choose $|f-g|_{_1}$ as small we like and thereby establish that

$$|W^\varepsilon| = 0$$

for every $\varepsilon > 0$. This will complete the proof because G_f is the complement of the union of the sets

$$W_1, W_{\frac{1}{2}}, W_{\frac{1}{3}}, W_{\frac{1}{4}}, W_{\frac{1}{5}}, W_{\frac{1}{6}}, \ldots,$$

and thereby, see Exercise 22.7, the complement of a set of measure zero.

It remains to estimate $H_{f-g}(x)$ and establish (22.10), but this argument will not depend on the choice of g. So we take $g \equiv 0$ and note that the set

$$O^\varepsilon = \{x \in \mathbb{R}^N : \, H_f(x) > \varepsilon\}$$

is open because

$$x \in O^\varepsilon \iff \exists r > 0 : \underbrace{\int_{B(x,r)} |f|}_{\substack{continuous \\ in\ r\ en\ x}} > \varepsilon |B(x,r)|. \tag{22.11}$$

How big can O^ε be? Every compact $K \subset O^\varepsilon$ is covered[9] by only finite many balls as in (22.11), say

$$K \subset B_1 \cup \cdots \cup B_m.$$

If, for every such $K \subset O^\varepsilon$, these balls were disjoint, then

$$\varepsilon |K| \leq \varepsilon(|B_1 + \cdots + |B_m|) < \int_{B_1} |f| + \cdots + \int_{B_m} |f| = \int_{B_1 \cup \cdots \cup B_m} |f| \leq \int_{\mathbb{R}^N} |f|.$$

and (22.10) would follow with $C_N = 1$.

Remark 22.14. *So in addition to (22.10) we need the statement that the measure of an open set O is the supremum of all the measures of compact subsets K of O. I will come back to get this rid of this issue in Section 22.3.*

It is of course highly unlikely that such disjoint coverings are possible, but with an extra N-dependent factor the estimate does indeed hold. We show below that

$$\varepsilon |O^\varepsilon| \leq 3^N |f|_1, \tag{22.12}$$

as a consequence of what is known as Vitali's covering lemma. This gives (22.10) with $C_N = 3^N$ and will complete the proof.

To wit, choose the[10] largest ball, say B_{j_1}, take it out of the collection and make it the first ball in a new collection. The balls B_i in the old collection for which $B_i \cap B_{j_1} \neq \emptyset$ are all contained in $3B_{j_1}$, the ball with the same center as B_{j_1} but 3 times it radius. Take these B_i out of the old collection and throw them away. If there are any balls left, let B_{j_2} be the largest of these remaining balls in the collection and take it as second ball in the new collection. Repeat the procedure until, say after choosing B_{j_k} and having thrown away all the remaining balls intersecting it, there are no more balls left in the old collection. Then[11]

$$B_{j_1}, \ldots, B_{j_k}$$

[9] Both compactness and measurability rely on definitions with coverings.
[10] Better: a ball which maximizes the radius in the collection.
[11] This is Vitali's covering lemma.

are disjoint, most likely don't cover K of course, but we do have

$$B_1 \cup \cdots \cup B_m \subset 3B_{j_1} \cup \cdots \cup 3B_{j_k},$$

whence

$$|f|_1 \geq \int_{B_1 \cup \cdots \cup B_m} |f| \geq \int_{B_{j_1} \cup \cdots \cup B_{j_k}} |f| = \int_{B_{j_1}} |f| + \cdots + \int_{B_{j_k}} |f|$$

$$> \varepsilon(|B_{j_1}| + \cdots + |B_{j_k}|) = 3^{-\mathrm{N}}\varepsilon(|3B_{j_1}| + \cdots + |3B_{j_k}|)$$

$$\geq 3^{-\mathrm{N}}\varepsilon|3B_{j_1} \cup \cdots \cup 3B_{j_k}| \geq 3^{-\mathrm{N}}\varepsilon|B_1 \cup \cdots \cup B_m| \geq 3^{-\mathrm{N}}\varepsilon|K|$$

for all compact $K \in O^\varepsilon$. It follows that (22.12) holds and this then completes the proof that the complement of the good set (22.5) has zero measure.

Theorem 22.15. *Let $f \in L^1(\mathbb{R}^{\mathrm{N}})$. Then for almost all x it holds that*

$$⨍_{B(x,r)} |f - f(x)| \to 0 \quad \text{as} \quad r \to 0.$$

Exercise 22.16. Apply Theorem 22.10 to the function $s \to |f(s) - q|$ for every $q \in \mathbb{Q}$ to prove Theorem 22.15. Hint: with the integration variable in

$$|f(s) - f(x)| \leq |f(s) - q| + |q - f(x)|$$

being s, it follows that

$$⨍_{B(x,r)} |f - f(x)| \leq ⨍_{B(x,r)} |f - q| + |q - f(x)|.$$

Given x you can take $|q - f(x)|$ as small as you like. Show that the complement of the intersection of all the good sets $G_{|f-q|}$ is a set of measure zero and conclude.

22.3 Vitali coverings and Hardy-Littlewood's again

Both compactness and measurability were defined in terms of properties of coverings of the sets under consideration. In this and later sections we avoid these notions but do use countable coverings with open balls. For convenience we restrict the attention to $f \in L^1(\mathbb{R}^{\mathrm{N}})$.

Theorem 22.17. *For $f \in L^1(\mathbb{R}^N)$ and $\varepsilon > 0$ let*

$$H_f(x) = \sup_{r>0} \fint_{B(x,r)} |f| \quad \textit{and} \quad O^\varepsilon = \{x \in \mathbb{R}^N : H_f(x) > \varepsilon\}.$$

Then there exists an atmost countable family of balls B_i indexed by a subset I of $\mathbb{N}$ such that

$$O^\varepsilon \subset \cup_{i\in I} B_i \quad \text{with} \quad \sum_{i\in I} |B_i| \leq \frac{6^N}{\varepsilon} |f|_1.$$

We use Remark 22.8 to prove Theorem 22.17. The set O^ε is open because

$$x \in O^\varepsilon \iff \exists r > 0 : \underbrace{\int_{B(x,r)} |f|}_{\textit{continuous in } r \textit{ en } x} > \varepsilon |B(x,r)|.$$

Thus O^ε is contained in the union of all such balls $B(x,r)$ and close to every $B(x,r)$ there is a ball B with rational center and rational radius such that

$$\int_B |f| > \varepsilon |B| \quad \text{and} \quad x \in B.$$

We conclude that there is a countable family of open balls B_n such that

$$O^\varepsilon \subset \cup_{n\in\mathbb{N}} B_n \quad \text{with} \quad \int_{B_n} |f| > \varepsilon |B_n|,$$

and we may of course assume that non of these balls are concentric. We will show that a subcollection of enlarged balls will do the job.

To see how let r_n be the corresponding sequence of radii and denote the distances between the centers of the balls B_m and B_n by $d_{mn} > 0$. Since

$$\varepsilon |B_n| < |f|_1,$$

the sequence r_n is bounded. Let R_1 be its supremum, choose $n_1 \in \mathbb{N}$ with

$$r_{n_1} > \frac{R_1}{2},$$

and let $\tilde{B}_1 = B_{n_1}$. Every ball B_n with $d_{nn_1} \leq 2R_1$ is contained in the ball concentric with $\tilde{B}_1$ with six times its radius, because the radius of this ball $6\tilde{B}_1$ is larger than $3R_1$, and the distance from any point in B_n to the center of $\tilde{B}_1$ is atmost $R_1 + 2R_1 = 3R_1$. Throw all these balls away. If there are

any balls left consider the supremum R_2 of the remaining radii and choose n_2 with[12]

$$r_{n_2} > \frac{R_2}{2},$$

and let $\tilde{B}_2 = B_{n_2}$, and throw away all B_n with $d_{nn_2} \leq 2R_2$. And so on. This gives a possibly infinite sequence of disjoint[13] open balls $\tilde{B}_k$ indexed by k, and for every finite sum indexed by a finite subset K of $\mathbb{N}$ we have

$$\varepsilon \sum_{k\in K} |B_k| < \sum_{k\in K} \int_{B_k} |f| = \int_{\cup_{k\in K} B_k} |f| \leq |f|_1 .$$

If the process to choose the balls $\tilde{B}_k$ did not stop at some $k = n \in \mathbb{N}$ it follows that $R_k \to 0$, and thus every ball not chosen as a $\tilde{B}_k$ is eventually thrown away, whence

$$O^\varepsilon \subset \cup_{k\in\mathbb{N}} 6\tilde{B}_k,$$

which we view as $n = \infty$ in

$$O^\varepsilon \subset \cup_{k=1}^{n} 6\tilde{B}_k \tag{22.13}$$

for the case that the process does stop, at some $k = n \in \mathbb{N}$.

For every $m \in \mathbb{N}$ with $m \leq n$ we now have that

$$\varepsilon \sum_{k=1}^{m} |6\tilde{B}_k| = 6^N \varepsilon \sum_{k=1}^{m} |\tilde{B}_k| < 6^N \sum_{k=1}^{m} \int_{\tilde{B}_k} |f| = 6^N \int_{\cup_{k=1}^{m} \tilde{B}_k} |f| \leq 6^N |f|_1 ,$$

so we conclude that

$$O^\varepsilon \subset \cup_{k=1}^{n} 6\tilde{B}_k \quad \text{with} \quad \sum_{k=1}^{n} |6\tilde{B}_k| \leq \frac{6^N}{\varepsilon} |f|_1 \quad \text{and} \quad n \in \mathbb{N} \cup \{\infty\}. \tag{22.14}$$

This completes the proof of Theorem 22.17.

Remark 22.18. *Note that the number* 6 *appears as* $2 \cdot 3$. *Choosing* $p > 1$ *instead of* 2 *it may be replaced by any number*[14] *larger than* 3. *Thus we have shown that the Lebesgue outer measure of* O^ε *is atmost*

$$\frac{3^N}{\varepsilon} |f|_1 .$$

[12] The $2 > 1$ in the denominator leads to $3 \cdot 2 = 6 > 3$, any other 2 will also do.

[13] Nontouching because $d_{kl} > 2R_k \geq R_k + R_l$ for $l > k \geq 1$.

[14] If 6 was π...

22.4 Via Cauchy sequences instead?

Observe that Theorem 22.10 identified the in some sense unique best choice $f \in F$ when F is an equivalence class of functions. In hindsight this would justify the sloppy notation

$$f \underbrace{\in [f] = F}_{skipped} \in L^1(\mathbb{R}^N),$$

properly taking into account that $f \in L^1(\mathbb{R}^N)$ is not a space of functions but a space of equivalence classes of functions. From here on we skip F as the notation for the equivalence class $[f]$, as we will be needing a symbol in an alternative approach for introducing the space $L^1(\mathbb{R}^N)$, as consisting of equivalence classes of Cauchy sequences of compactly supported continuous functions f_n.

Definition 22.19. *Two sequences f_n and g_n in $C_c(\mathbb{R}^N)$ are called equivalent if $|f_n - g_n|_1 \to 0$ as $n \to \infty$.*

Cauchy sequences are characterised by the property that

$$|f_n - f_m|_1 = \int_{\mathbb{R}^N} |f_n - f_m| \to 0$$

as $m, n \to \infty$. If $f_n \in C_c(\mathbb{R}^N)$ a Cauchy sequence with respect to the 1-norm and $f_n \sim g_n$ then also g_n is a Cauchy sequence with respect to the 1-norm. Moreover, for every $x \in \mathbb{R}^N$ and every $r > 0$ the sequences

$$\int_{B(x,r)} f_n \quad \text{and} \quad \int_{B(x,r)} g_n$$

are (equivalent) Cauchy sequences in $\mathbb{R}$ with the same limit. Writing

$$F = [f_n]$$

for such am equivalence class we see that

$$\int_{B(x,r)} F = \lim_{n\to\infty} \int_{B(x,r)} f_n, \tag{22.15}$$

is the natural definition of the integral of F over the ball $B(x, r)$, and thus

$$A_F(x, r) = \fint_{B(x,r)} F = \frac{1}{|B(x, r)|} \int_{B(x,r)} F \tag{22.16}$$

is well defined for every equivalence class.

Theorem 22.20. *Let F be an equivalence class of Cauchy sequences in $C_c(\mathbb{R}^N)$ with respect to the 1-norm, and let $\mathcal{N}_F$ be the set of points x for which*

$$\lim_{r\to 0} A_F(x,r) \tag{22.17}$$

does not exist. Then $\mathcal{N}_F$ is a zero measure set.

For the proof we examine $\mathcal{N}_F$ again using

$$H_F(x) = \sup_{r>0} A_{|F|}(x,r) = \sup_{r>0} \frac{1}{|B(x,r)|}\int_{B(x,r)} |F| \in [0,\infty],$$

in which

$$\int_{B(x,r)} |F| = \lim_{n\to\infty}\int_{B(x,r)} |f_n| \geq |\lim_{n\to\infty}\int_{B(x,r)} f_n| = |\int_{B(x,r)} F|.$$

Note that (22.16) defines a quantity which is continuous as a function of $r>0$ and $x\in\mathbb{R}^N$. This is because

$$|\int_{B(x,r)} F - \int_{B(y,s)} F| \leq$$

$$\underbrace{|\int_{B(x,r)} F - \int_{B(x,r)} f_n|}_{\leq\varepsilon} + |\int_{B(x,r)} f_n - \int_{B(y,s)} f_n| + \underbrace{|\int_{B(y,s)} f_n - \int_{B(y,s)} F|}_{\leq\varepsilon}$$

$$\leq \underbrace{\int_{B(x,r)} |F-f_n|}_{\leq\varepsilon} + |\int_{B(x,r)} f_n - \int_{B(y,s)} f_n| + \underbrace{|\int_{B(y,s)} |f_n - F|}_{\leq\varepsilon}$$

for $n\geq N$, if N corresponds to $\varepsilon>0$ via the definition of f_n being a Cauchy sequence with respect to the 1-norm.

Clearly the definition of the integral of the class F in (22.15) as the limit of the Cauchy sequence

$$\int_{B(x,r)} f_n,$$

has that same N doing the job for $\varepsilon>0$ for all ball $B(x,r)$ simultaneously. In the second middle term we then fix $n=N$ and ask for that difference to be atmost $\varepsilon>0$. Since $f_N\in C_c(\mathbb{R}^N)$, this can be done uniformly in terms of the smallness of $|r-s|$ and $|x-y|$. As a result the function

$$(x,r)\to\int_{B(x,r)} F$$

is (uniformly) continuous, just as in (22.11).

We can now consider the existence issue for the limit in (22.17), before we have even identified what its limit value should be, for x in the good set of F, the set for which the limit exists. Most of these limit values will come from Theorem 22.23 in Section 22.5 below, but first we reason as in Theorem 22.10, replacing the basic estimate (22.7) via

$$|A_F(x,r) - A_F(x,s)| \leq$$

$$|A_F(x,r) - A_{f_m}(x,r)| + |A_{f_m}(x,r) - A_{f_m}(x,s)| + |A_{f_m}(x,s) - A_F(x,s)|$$
$$\leq A_{|F-f_m|}(x,r) + |A_{f_m}(x,r) - A_{f_m}(x,s)| + A_{|f_m-F|}(x,s)$$

for $0 < s < r$ by

$$|A_F(x,r) - A_F(x,s)| \leq 2H_{F-f_m}(x) + \underbrace{|A_{f_m}(x,r) - A_{f_m}(x,s)|}_{\to 0 \ as \ r \to 0}. \qquad (22.18)$$

The first term on the right hand side of (22.18) is twice the upper bound $H_{F-f_m}(x)$ for

$$A_{|F-f_m|}(x,r) = A_{|f_m-F|}(x,s),$$

in which $|F - f_m|$ with m fixed denotes the equivalence class of the Cauchy sequence[15] $|f_n - f_m|$.

Let W_ε be the set of all points $x \in \mathbb{R}^N$ for which the statement

$$\exists_{\delta>0} \forall_{r,s\in(0,\delta)} : |A_F(x,r) - A_F(x,s)| \leq 2\varepsilon$$

fails. Then (22.18) gives

$$W_\varepsilon \subset O_m^\varepsilon = \{x \in \mathbb{R}^N : H_{F-f_m}(x) > \varepsilon\},$$

and similar to (22.11) we have

$$x \in O_m^\varepsilon \iff \exists r > 0 : \underbrace{\int_{B(x,r)} |F - f_m|}_{\substack{continuous \\ in \ r \ en \ x}} > \varepsilon |B(x,r)|. \qquad (22.19)$$

Exercise 22.21. Modify the proof of Theorem 22.17 to show that W_ε is set of zero measure for every $\varepsilon > 0$. Thus the limit in (22.17) exists outside a set of measure zero. Hint: use (22.14).

[15] Indexed by n.

Remark 22.22. *In view of Theorem 22.20 the function f defined by*

$$f(x) = \lim_{r\to 0} A_F(x,r) \quad \text{for} \quad x \notin \mathcal{N}_F \quad \text{and} \quad f(x) = 0 \quad \text{for} \quad x \in \mathcal{N}_F \quad (22.20)$$

has to be examined next. Is it in $L^1(\mathbb{R}^N)$ and does it coincide with the f chosen in Theorem 22.10?

22.5 Pointwise limits of the Cauchy sequence?

In relation to Remark 22.20 we first try to extract a function from the Cauchy sequence f_n. It remains to be seen if this can be avoided, and give a direct formulation and proof of the desired interpretation of the function f defined in Remark 22.20.

Theorem 22.23. *Given a sequence $f_n \in C_c(\mathbb{R}^N)$ with $|f_n - f_m|_1 \to 0$ and a number $\eta > 0$, we can extract a subsequence along which the sequence converges uniformly on the complement of a union of open balls*

$$U = \cup_{k\in\mathbb{N}} B_k \quad \text{with} \quad \sum_{k\in\mathbb{N}} |B_k| < \eta.$$

For the proof we observe that

$$|f_n(x) - f_m(x)| \leq$$

$$\underbrace{|f_n(x) - A_{f_n}(x,r)|}_{\to 0 \ \text{ as } \ r\to 0} + \underbrace{|A_{f_n}(x,r) - A_{f_m}(x,r)|}_{\leq H_{f_n - f_m}(x)} + \underbrace{|A_{f_m}(x,r) - f_m(x)|}_{\to 0 \ \text{ as } \ r\to 0},$$

so

$$|f_n(x) - f_m(x)| \leq H_{f_n - f_m}(x), \quad (22.21)$$

and with

$$O^\varepsilon_{mn} = \{x \in \mathbb{R}^N : \ H_{f_n - f_m}(x) > \varepsilon\} \quad (22.22)$$

we have

$$|f_n(x) - f_m(x)| \leq \varepsilon \quad \text{for} \quad x \notin O^\varepsilon_{mn}.$$

We now cover O^ε_{mn} with finitely many open balls such that sum of the measures of these balls is bounded as in (22.10), with $|f - g|_1$ replaced by $|f_n - f_m|_1$, and this norm we can make as small as we like by taking m, n larger then some $N \in \mathbb{N}$. This is an argument that only uses the function $g = f_n - f_m \in C_c(\mathbb{R}^N)$, and is independent of how we arrived at the particular choice of g.

So consider $g \in C^c(\mathbb{R}^N)$ and let

$$x \in O^\varepsilon = \{x \in \mathbb{R}^N : \ H_g(x) > \varepsilon\},$$

in which for every x the supremum

$$H_g(x) = \sup_{r>0} A_{|g|}(x,r) \in [0,\infty]$$

is possibly realised as a maximum by some $r > 0$. If not then the continuity of g in x implies that $H_g(x) = |g(x)|$ because

$$A_{|g|}(x,r) \to |g(x)| \quad \text{as} \quad r \to 0 \quad \text{and} \quad A_{|g|}(x,r) \to 0 \quad \text{as} \quad r \to \infty,$$

the latter being a consequence of $|g|_1 = \int |g|$ existing as the Riemann integral of the continuous function $x \to |g(x)|$ over some large ball.

That being said we only use that $x \in O^\varepsilon$ means that for some radius $r = r_x > 0$ it must be that

$$\int_{B(x,r)} |g| > \varepsilon |B(x,r_x)|.$$

This property is classifying for $x \in O^\varepsilon$. It follows that

$$O^\varepsilon = \{x \in \mathbb{R}^N : \int_{B(x,r)} |g| > \varepsilon |B(x,r)| \quad \text{for some } r > 0\}$$

is open and bounded, and its closure is compact. For every boundary point $\bar{x}$ of O^ε there is a sequence of points $x_n \in O^\varepsilon$ with $x_n \to \bar{x}$ and the sequence[16] r_n of corresponding radii converging to either 0 or to a positive limit $\bar{r} > 0$. In both cases it follows that $H_g(\bar{x}) \geq \varepsilon$, possibly as $H_g(\bar{x}) = |g(\bar{x})|$, but it cannot be that $H_g(\bar{x}) > \varepsilon$, so $H_g(\bar{x}) = \varepsilon$. We can thus cover the closure of O^ε with finitely many balls open balls $\tilde{B}_1, \ldots, \tilde{B}_m$ such that

$$\int_{\tilde{B}_i} |g| > \frac{\varepsilon}{2} |\tilde{B}_i|, \tag{22.23}$$

and using Vitali's covering lemma again we choose $j_1, \ldots, j_k$ such that

$$O^\varepsilon \subset \tilde{B}_1 \cup \cdots \cup \tilde{B}_m \subset 3\tilde{B}_{j_1} \cup \cdots \cup 3\tilde{B}_{j_k} = B_1 \cup \cdots \cup B_k,$$

with $\tilde{B}_{j_1}, \ldots, \tilde{B}_{j_k}$ disjoint, implying that

$$O^\varepsilon \subset B_1 \cup \cdots \cup B_k \quad \text{with} \quad |B_1| + \cdots + |B_k| < \frac{2 \cdot 3^N}{\varepsilon} |g|_1,$$

and likewise for O^ε_{mn}.

[16] Taking a subsequence.

Given $\varepsilon > 0$ and $m, n \in \mathbb{N}$ we have a finite collection of open balls $B_1, \ldots, B_k$ such that

$$\{x \in \mathbb{R}^{\mathrm{N}} : |f_n(x) - f_m(x)| > \varepsilon\} \subset B_1 \cup \cdots \cup B_k \tag{22.24}$$

with

$$|B_1| + \cdots + |B_k| < \frac{2 \cdot 3^N}{\varepsilon} |f_n - f_m|_1 < \eta \tag{22.25}$$

if we choose $m, n \geq \mathrm{N}$, N depending on $\frac{1}{2}\eta\varepsilon 3^{-\mathrm{N}}$ via the Cauchy-property of the sequence f_n.

Now choose $\varepsilon_k \downarrow 0$ and $\eta_k \downarrow 0$ with

$$\sum_{k=1}^{\infty} \varepsilon_k = \varepsilon, \quad \sum_{k=1}^{\infty} \eta_k = \eta,$$

and corresponding N_k as above. Then f_{N_k} converges uniformly on the complement of the countable union U of all the balls used in (22.24) with $m = N_{k+1}$ and $n = N_k$ for $k = 1, 2, 3, \ldots$.

This completes the proof, but we can do slightly better. After applying the theorem with $\eta > 0$ we obtain a countable union U of open balls. Outside U the constructed subsequence converges uniformly. On the open set U we repeat the proof, which still concerns functions defined on the whole of $\mathbb{R}$ with compact supports. These supports most likely are not subsets of U, but this is of no importance. The required modifications are only minor. We look for coverings

$$B_1 \cup \cdots \cup B_k \supset \{x \in U : |f_n(x) - f_m(x)| > \varepsilon\} \tag{22.26}$$

via coverings of

$$\{x \in U : H_g(x) > \varepsilon\} = O^{\varepsilon} \cap U.$$

For boundary points of $O^{\varepsilon} \cap U$ in O^{ε} the adjustment in (22.23) with the factor $\frac{1}{2}$ is not necesssary, and for the other boundary points we reason as before to conclude that U contains an open union $\tilde{U}$ of balls $\tilde{B}_1, \tilde{B}_2, \ldots$, for which $|\tilde{B}_1| + |\tilde{B}_2| + \cdots < \tilde{\eta}$, outside of which there is uniform convergence. With $\delta_1 = \eta$, $U_1 = U$, $\delta_2 = \tilde{\eta} < \delta_1$, $U_2 = \tilde{U} \subset U = U_1$, we have the first numbers and unions in the sequences in the following theorem.

Theorem 22.24. *Given a sequence $f_n \in C_c(\mathbb{R}^{\mathrm{N}})$ with $|f_n - f_m|_1 \to 0$ as $m, n \to \infty$, and a sequence of positive real numbers*

$$\delta_1 > \delta_2 > \delta_3 \cdots \to 0,$$

there exists a sequence

$$U_1 \supset U_2 \supset U_3 \supset \cdots$$

of countable unions

$$U_j = \cup_{k\in\mathbb{N}} B_{jk} \quad \text{with} \quad \sum_{k\in\mathbb{N}} |B_{jk}| < \delta_j,$$

such that for some sequence

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

of natural numbers it holds that $f_{n_l}(x)$ converges to a limit for every x in the complement of the intersection

$$\mathcal{N} = \cap_{j\in\mathbb{N}} U_j,$$

uniformly on the complement of every U_j. The limit is denoted by

$$f(x) = \lim_{l\to\infty} f_{n_l}(x).$$

The set $\mathcal{N}$ has measure zero and empty interior.

We still have to relate this limit function f to the equivalence class f we started with.

23 Sobolev spaces

This chapter is roughly based on the section about mollifiers in the appendix of Evans' PDE book and the chapter about Sobolev spaces. We let U be an open set in $\mathbb{R}^{\mathrm{N}}$ and consider functions u and v rather than f and g. For $1 \leq p < \infty$ the Lebesgue p-norm of a function $v \in C_c(U)$ is defined by

$$|v|_p^p = \int_{\mathbb{R}^{\mathrm{N}}} |v|^p, \tag{23.1}$$

and the Sobolev $W^{1,p}$-norm of a function $v \in C_c^1(U)$ by

$$|v|_{1,p}^p = |v|_p^p + |v_{x_1}|_p^p + \cdots + |v_{x_{\mathrm{N}}}|_p^p, \tag{23.2}$$

The spaces $L^p(\mathbb{R}^{\mathrm{N}})$ and $W_0^{1,p}(U)$ are the closures of $C_c^1(U)$ with respect to the p-norm and the $W^{1,p}$-norm. In case of the $W^{1,p}$-norm the closure is either taken in a not yet defined larger space, or in the abstract sense with equivalence classes of Cauchy sequences. In what follows partial derivatives will be taken in the distributional sense[1], with the functions in $C_c^1(U)$, the space of continuously differentiable functions with compact support in U, acting as test functions.

23.1 Mollifiers and density tricks

We first restrict the attention to $L^p(\mathbb{R}^{\mathrm{N}})$. We note that in case of the p-norm the closure of $C_c^1(U)$ is the same as the closure of $C_c(U)$, and it may also be taken in the abstract sense with equivalence classes of Cauchy sequences. One way or another, Theorem **??** allows for every $u \in L^1_{loc}(\mathbb{R}^{\mathrm{N}})$ the introduction of its ε-mollified version

$$u^\varepsilon(x) = (\eta_\varepsilon * u)(x) = \int_{\mathbb{R}^{\mathrm{N}}} \eta_\varepsilon(x-y)u(y)\,dy = \int_{\mathbb{R}^{\mathrm{N}}} \eta_\varepsilon(y)u(x-y)\,dy$$

$$= \int_{|y|\leq\varepsilon} \eta_\varepsilon(y)u(x+y)\,dy,$$

in which

$$\eta_\varepsilon(x) = \frac{1}{\varepsilon^{\mathrm{N}}}\eta(\frac{x}{\varepsilon}), \quad \eta(x) = \eta(|x|) \geq 0, \quad \eta \in C_c^\infty(B), \quad \int_{\mathbb{R}^{\mathrm{N}}} \eta = 1,$$

B denoting the open unit ball. In practice we prove statements about u^ε and u via calculations with v^ε, $v \in C_c^1(U)$, but we do note that u^ε is smooth.

[1] The definition of (unique) weak derivatives relies on Theorem 22.15.

Exercise 23.1. Use the techniques presented in Section 9.5 to show that u^ε is smooth.

I try to restrict the use of mollifiers to globally defined locally integrable functions u for which

$$u^\varepsilon(x) = \int_{\mathbb{R}^N} \eta_\varepsilon(x-y)u(y)\,dy = \int_{\mathbb{R}^N} \eta_\varepsilon(y)u(x-y)\,dy = \int_{|y|\le\varepsilon} \eta_\varepsilon(y)u(x+y)\,dy.$$

For $u \in L^p(\mathbb{R}^N)$ Hölder's inequality with

$$\frac{1}{p} + \frac{1}{q} = 1$$

gives

$$|u^\varepsilon(x)| \le \int_{|y|\le\varepsilon} \eta_\varepsilon(y)^{\frac{1}{q}+\frac{1}{p}} |u(x+y)|\,dy \le \left(\int_{|y|\le\varepsilon} \eta_\varepsilon(y)|u(x+y)|^p\,dy\right)^{\frac{1}{p}},$$

whence, taking the p-th power, integration gives[2] the following result for $u \in L^p(\mathbb{R}^N)$, which is complemented by a convergence result in case $v \in C_c^1(U)$.

Theorem 23.2. *Let $u \in L^p(\mathbb{R}^N)$. Then*

$$\int_{\mathbb{R}^N} |u^\varepsilon(x)|^p\,dx \le \int_{\mathbb{R}^N} |u(x)|^p\,dx, \text{ i.e. } \quad |u^\varepsilon|_p \le |u|_p. \tag{23.3}$$

Theorem 23.3. *Let $v \in C_c^1(U)$ and $1 \le p < \infty$. Then*

$$|v^\varepsilon(x) - v(x)| \le \left(\int_{\mathbb{R}^N} |v(x+\varepsilon y) - v(x)|^p\,\eta(y)dy\right)^{\frac{1}{p}}. \tag{23.4}$$

for $v \in C_c^1(U)$, and

$$|v^\varepsilon - v|_p \le \varepsilon\,|\nabla v|_p, \tag{23.5}$$

in which the right hand side is the p-norm of the Euclidean length of ∇v.

Before we prove Theorem 23.3 we note that it is via these two theorems, namely (23.3) with $u - v$, the splitting

$$|u - u^\varepsilon|_p \le |u - v|_p + |v - v^\varepsilon|_p + |v^\varepsilon - u^\varepsilon|_p,$$

the density of $C_c^1(U)$ in $L^p(\mathbb{R}^N)$ and (23.5), that the following theorem follows.

[2] Changing the order of integration, same trick as in Exercise 23.22.

Theorem 23.4. *Let $u \in L^p(\mathbb{R}^N)$, then $|u^\varepsilon|_p \le |u|_p$ and $|u^\varepsilon - u|_p \to 0$.*

Exercise 23.5. Prove Theorem 23.4 using Theorems 23.2 and 23.3.

Remark 23.6. *We shall also use (23.5) in Section 23.3 for a direct proof that every bounded sequence u_n in $W_0^{1,p}(U)$ has a subsequence which, considered as a sequence in $L^p(U)$, is convergent. In fact this could already be an exercise here.*

Exercise 23.7. Let u_n be a bounded sequence in $W_0^{1,p}(U)$. Prove that there is a subsequence of u_n which is Cauchy with respect to the p-norm. Hint: use the splitting

$$|u_n - u_m|_p \le |u_n - v_n|_p + |v_n - v_n^\varepsilon|_p + |v_n^\varepsilon - v_m^\varepsilon|_p + |v_m^\varepsilon - v_m|_p + |v_m - u_m|_p,$$

deal with the second and fourth term by (23.5), with the first and fifth term by density of $C_c^1(U)$ in $W_0^{1,p}(U)$, and finally with the third term by Theorem 4.44 (with $[0,1]$ replaced by a large closed box) and ε-dependent bounds for ε fixed. A diagonal argument completes the proof.

For the proof of Theorem 23.3 write

$$v^\varepsilon(x) - v(x) = \int_{\mathbb{R}^N} \eta_\varepsilon(y) v(x+y)\, dy - v(x) = \int_{\mathbb{R}^N} \eta_\varepsilon(y)\, (v(x+y) - v(x))\, dy$$

$$= \int_{\mathbb{R}^N} \eta(y)\, (v(x+\varepsilon y) - v(x))\, dy = \int_{\mathbb{R}^N} \eta(y)^{\frac{1}{q}} \eta(y)^{\frac{1}{p}}\, (v(x+\varepsilon y) - v(x))\, dy.$$

Hölder's inequality gives

$$|v^\varepsilon(x) - v(x)| \le \underbrace{\left(\int_{\mathbb{R}^N} \eta(y)\, dy \right)^{\frac{1}{q}}}_{=1} \left(\int_{\mathbb{R}^N} \eta(y) |v(x+\varepsilon y) - v(x)|^p\, dy \right)^{\frac{1}{p}}.$$

whence (23.4) follows.

We next use the mean value theorem in integral form, see (9.8), and the Hölder estimate

$$|\int_0^1 f|^p \le |\int_0^1 |f|^p \qquad (1 \le p < \infty).$$

The x-integral in the right hand side of (23.4) is estimated for $|y| \le \varepsilon$ as

$$\int_{\mathbb{R}^N} |v(x+\varepsilon y) - v(x)|^p\, dx = \int_{\mathbb{R}^N} |[v(x+t\varepsilon y)]_{t=0}^{t=1}|^p\, dx$$

$$= \int_{\mathbb{R}^N} |\int_0^1 \nabla v(x+\varepsilon t y)\cdot \varepsilon y\, dt|^p\, dx$$

$$\le \varepsilon^p \int_{\mathbb{R}^N} \left(\int_0^1 |\nabla v(x+\varepsilon t y)|\, dt\right)^p dx \le \varepsilon^p \int_{\mathbb{R}^N}\int_0^1 |\nabla v(x+\varepsilon t y)|^p\, dt\, dx$$

whence (23.4) gives

$$\int_{\mathbb{R}^N} |v^\varepsilon(x)-v(x)|^p\, dx \le \int_{\mathbb{R}^N} \eta(y) \int_{\mathbb{R}^N} |v(x+\varepsilon t y)-v(x)|^p\, dx\, dy \le$$

$$\le \int_{\mathbb{R}^N} \eta(y)\, \varepsilon^p \int_{\mathbb{R}^N}\int_0^1 |\nabla v(x+\varepsilon t y)|^p\, dt\, dx\, dy,$$

and (23.5) follows by changing the order of integration from $dt\, dx\, dy$ to $dx\, dt\, dy$. This completes the proof.

Theorem 23.8. *Let $u \in L^p(U)$, $U \subset \mathbb{R}^N$ open, and $V \subset\subset U$. Then $u^\varepsilon \to u$ in $L^p(V)$.*

Theorem 23.8 follows by extending u to $\tilde{u}$ defined on the whole of $\mathbb{R}^N$ via $\tilde{u}(x) = u(x)$ for $x \in U$ and $\tilde{u}(x) = 0$ for $x \notin U$. Then Theorem 23.4 implies $\tilde{u}^\varepsilon \to \tilde{u}$ in $L^p(\mathbb{R}^N)$ and thus also in $L^p(V)$, but on V we have $\tilde{u} = u$ and $\tilde{u}^\varepsilon = u^\varepsilon$ if ε is small.

In Appendix C.5 Theorem 7 Evans proves a similar result for functions $u \in L^p_{loc}(U)$ via basically[3] $v \in C_c(U)$. Such functions are uniformly continuous, i.e

$$\forall_{\delta>0}\, \exists_{\eta>0}\, \forall_{x,y\in U} : |x-y| < \eta \implies |v(x)-v(y)| < \delta.$$

Then

$$|v^\varepsilon(x)-v(x)| = |\int_{|y|\le\varepsilon} \eta_\varepsilon(y)(v(x+y)-v(x))\, dy| < \delta,$$

provided $\varepsilon \le \eta$. This proves that $v^\varepsilon \to v$ uniformly, i.e $|v^\varepsilon - v|_\infty \to 0$ as $\varepsilon \to 0$.

Theorem 23.9. *Let $v \in C_c(\mathbb{R}^N)$, then $v^\varepsilon \in C_c(\mathbb{R}^N)$, $|v^\varepsilon|_\infty \le |v|_\infty$ and $|v^\varepsilon - v|_\infty \to 0$.*

Remark 23.10. *Evans' proof that $u \in L^1_{loc}(\mathbb{R}^N)$ implies $u^\varepsilon(x) \to u(x)$ for almost every x relies on Theorem 22.15.*

[3] He has u defined on a subset only.

23.2 Sobolev spaces of functions with weak derivatives

The concept of weak derivative comes as a theorem.

Theorem 23.11. *Suppose that*

$$\int_U v\phi = -\int_U u\phi_{x_i} \tag{23.6}$$

for every $\phi \in C_c^1(U)$, for some given u and v in $L^1_{loc}(U)$, $U \subset \mathbb{R}^N$. Then v is unique in $L^1_{loc}(U)$ for $u \in L^1_{loc}(U)$. We say that v is the weak derivative of u with respect to its i^{th} variable, notation $v = D_i u = u_{x_i}$.

For the proof suppose that some other v, say $\tilde{v} \in L^1_{loc}(U)$ also satisfies this condition then the difference $w = v - \tilde{v}$ is in $L^1_{loc}(U)$ and satisfies

$$\int_U w\phi = 0$$

for every $\phi \in C_c^1(U)$. Take an open ball $B \subset U$ and redefine $w(x) = 0$ for $x \notin B$. The mollifier w^ε is then identically zero on $\mathbb{R}^N$ for all $\varepsilon > 0$. By Theorem 23.8 we have

$$|w|_1 = |w^\varepsilon - w|_1 \to 0$$

as $\varepsilon \to 0$. But $|w|_1$ doesn't go anywhere. So $|w|_1 = 0$, and Theorem 22.10 tells us what w is, as a function: zero! It follows that $v = \tilde{v}$ in B outside a set of measure zero, for every $B \subset U$ open. Thus $v = \tilde{v}$ in U outside a set of measure zero.

Definition 23.12. *For $1 \leq p < \infty$ and $U \subset \mathbb{R}^N$ the space $W^{1,p}(U)$ is defined as the space of functions $u \in L^p(U)$ for which the weak derivatives $u_{x_1}, \ldots, u_{x_N}$ exist and are in $L^p(U)$.*

Exercise 23.13. Prove that $W^{1,p}(U)$ is a Banach space with the norm defined by

$$|u|^p_{1,p} = |u|^p_p + |u_{x_1}|^p_p + \cdots + |u_{x_N}|^p_p. \tag{23.7}$$

Hint: use that $L^p(U)$ is a Banach space.

Remark 23.14. *You may now prefer to define $W^{1,p}_0(U)$ as the closure of $C_c^1(U)$ in $W^{1,p}(U)$. As a closure in a Banach space the space $W^{1,p}_0(U)$ is itself then also complete.*

Exercise 23.15. Every $u \in W_0^{1,p}(U)$ extends to a $u \in W_0^{1,p}(\mathbb{R}^N)$ by setting u equal to zero outside U. Hint: a similar statement holds for $u \in C_c^1(U)$ and $C_c^1(\mathbb{R}^N)$.

Remark 23.16. *Every $C_c^k(U)$ with $k \in \mathbb{N}$ is dense in $W_0^{1,p}(U)$ and so is $C_c^\infty(U)$, the space of test functions used throughout in the literature.*

Exercise 23.17. A bit harder and to do somewhere along the road in this chapter:

$$W_0^{1,p}(\mathbb{R}^N) = W^{1,p}(\mathbb{R}^N).$$

Remark 23.18. *The definitions of $W^{2,p}(U)$ and of $W_0^{2,p}(U)$ (the closure of $C_c^k(U)$ with $k \geq 2$) should be obvious, starting from the definition of weak second order derivatives $u_{x_ix_j}$ and $u_{x_jx_i}$, the inevatible observation that $u_{x_ix_j} = u_{x_jx_i}$ under assumptions you should figure out, and the norm defined by*

$$|u|_{2,p}^p = |u|_p^p + \sum_{1\leq i\leq N} |u_{x_i}|_p^p + \sum_{1\leq i\leq j\leq N} |u_{x_ix_j}|_p^p.$$

Exercise 23.19. Fill in the details of Remark 23.18 and generalise to $W^{k,p}(U)$ and $W_0^{k,p}(U)$ with $k \geq 2$.

23.3 Compactness for $W_0^{1,p}(U)$

In Section 23.8 we use calculus to derive the Gagliardo-Nirenberg-Sobolev[4] and Morrey estimates and identify $p = N$ as a critical exponent. We will have

Theorem 23.20. *Let $p > N$ and $u \in W_0^{1,p}(\mathbb{R}^N)$. Then, after redefining u on a zero measure set,*

$$u : \mathbb{R}^N \to \mathbb{R}$$

is contininous, and, as a consequence of the uniform continuity,

$$\max_{\substack{x\in\mathbb{R}^N \\ |x|=R}} |u(x)| \to 0 \quad \text{as} \quad R \to \infty.$$

[4] GNS-estimates for short.

The Morrey estimates come with a uniform modulus of continuity which via the Ascoli-Arzelà theorem[5] *implies that the inclusion map* $W_0^{1,p}(U) \to C(\bar{U})$ *is compact if* U *is bounded*[6].

What follows is not restricted to $p > N$, and was announced in Remark 23.6 of Section 23.1. The proof of Theorem 23.21 below is also based on the AA Theorem, which states for compact metric spaces X that a sequence u_n in $C(X)$, the space of $\mathbb{R}$-valued continuous functions on X with norm

$$|u|_\infty = \max_{x \in X} |u(x)|,$$

has a convergent[7] subsequence u_{n_k}, provided the sequence is bounded in $C(X)$ and has the property[8] that

$$\forall \varepsilon > 0 \, \exists \delta > 0 \, \forall n \in \mathbb{N} \, \forall x, y \in X : \, d(x,y) < \delta \implies |f_n(x) - f_n(y)| < \varepsilon.$$

The other ingredient in the proof of Theorem 23.21 is the use of mollifiers.

Theorem 23.21. *We have for all* $p \in [1, \infty)$ *that*

$$W_0^{1,p}(U) \to L^p(U) \quad \textit{is compact if} \quad U \quad \textit{is bounded.}$$

NB: the weaker statement that the embedding is bounded will be characterised by the Poincaré inequality: $|u|_p \leq C_{pU} |\nabla u|_p$ *for all* $u \in W_0^{1,p}(U)$, C_{pU} *a constant depending on* p *and* U *only*[9].

For a large part the proof has already been done in Section 23.1. Consider a bounded sequence $u_n \in W_0^{1,p}(U)$, choose $v_n \in C_c^1(U)$ and extend v_n to $v_n \in C_c^1(\mathbb{R}^N)$, such that $|u_n - v_n|_{1,p} \to 0$, and consider mollified versions of v_n defined by

$$v_n^\varepsilon(x) = (\eta_\varepsilon * v)(x) = \int_{\mathbb{R}^N} \eta_\varepsilon(x-y) v_n(y) \, dy = \int_{\mathbb{R}^N} \eta_\varepsilon(y) v_n(x-y) \, dy$$

$$= \int_{|y| \leq \varepsilon} \eta_\varepsilon(y) v_n(x+y) \, dy,$$

in which

$$\eta_\varepsilon(y) = \frac{1}{\varepsilon^N} \eta(\frac{y}{\varepsilon}) \quad \text{with} \quad 0 \leq \eta \in C_c^\infty(B) \quad \text{radial}, \quad \int_B \eta = 1,$$

[5] See Section 4.6, AA Theorem for short.

[6] For $U = \mathbb{R}^N$ compactness fails for the same reason as Theorem 4.44. Which reason?

[7] Here convergence means uniform convergence.

[8] This is called the (uniform) equicontinuity of the sequence u_n.

[9] And thereby also on the dimension N.

where B is the open unit ball. Then split $|u_n - u_m|_p$ as

$$|u_n - u_m|_p \le |u_n - v_n|_p + |v_n - v_n^\varepsilon|_p + |v_n^\varepsilon - v_m^\varepsilon|_p + |v_m^\varepsilon - v_m|_p + |v_m - u_m|_p.$$

The first and fifth term converge to zero in view of $|u_n - v_n|_{1,p} \to 0$, the second and fourth are controlled by, with $v = v_n$ and $v = v_m$, the estimate

$$\int_{\mathbb{R}^N} |v^\varepsilon(x) - v(x)|^p \, dx \le \varepsilon^p \int_{\mathbb{R}^N} |\nabla v(x)|^p \, dx, \tag{23.8}$$

and therefore bounded by $C\varepsilon$, with C depending only on the bound for the sequence $|v_n|_{1,p}$, and thus only on the original bound for the sequence u_n in $W_0^{1,p}(U)$,

For every $\varepsilon > 0$ fixed the third term converges to zero along a subsequence in view of the AA Theorem applied to the sequence v_n^ε in $C(\bar{V})$, with V bounded and slight larger than U so as to have all v_n^ε in $C_c^\infty(V)$. We thus have that $|u_n - u_m|_p$ is asymptotically controlled by $C\varepsilon$ along that subsequence, along which $|v_n^\varepsilon - v_m^\varepsilon|_\infty \to 0$. A standard diagonal argument now produces a subsequence which is a Cauchy sequence in $L^p(V)$. The following exercises[10] fill in the details and are all that's needed to conclude the proof of Theorem 23.21.

Exercise 23.22. Show that

$$\int_{\mathbb{R}^N} |v^\varepsilon(x) - v(x)|^p \, dx \le \int_{\mathbb{R}^N} \eta(y) \int_{\mathbb{R}^N} |v(x + \varepsilon y) - v(x)|^p \, dx \, dy. \tag{23.9}$$

Hint: write the difference $v^\varepsilon(x) - v(x)$ as a single y-integral over $|y| \le \varepsilon$, scale y to have the integral over the unit ball, use

$$\eta(y) = \eta(y)^{\frac{1}{p'}} \eta(y)^{\frac{1}{p}},$$

and apply Hölder's inequality. Then use this estimate for the integral of $|v^\varepsilon(x) - v(x)|^p$ and change the order of integration to conclude.

Exercise 23.23. Write the difference in the right hand side of (23.9) as an s-integral over the interval $[0, 1]$, use Hölder's inequality and $|y| \le 1$ to arrive at an integral over s and x only, and obtain (23.8).

[10] We already did most of them in Section 23.1.

Exercise 23.24. Show that for fixed $\varepsilon > 0$ the sequence v_n^ε has a convergent subsequence in $C(\bar{V})$.

We apply the AA Theorem. The bounds

$$|v_n^\varepsilon(x)| = |\int_{\mathbb{R}^N} \eta_\varepsilon(x-y)v_n(y)\,dy| \le |\eta_\varepsilon|_{p'}|v_n|_p$$

and

$$|\nabla v_n^\varepsilon(x)| = |\int_{\mathbb{R}^N} \nabla\eta_\varepsilon(x-y)v_n(y)\,dy| \le |\nabla\eta_\varepsilon|_{p'}|v_n|_p$$

imply that the conditions of the AA Theorem are satisfied if v_n is merely a bounded sequence in $L^p(U)$ extended by $v_n(x) = 0$ for $x \notin U$, whence v_n has a convergent subsequence on the compact closure of an ε-neighbourhood of U. Since v_n is in fact bounded in $W_0^{1,p}(U)$ we are done (also for $p = 1$).

23.4 The need for extension operators

To extend the results for $W_0^{1,p}(U)$ to $W^{1,p}(U)$ we need a well behaved extension operator that maps $W^{1,p}(U)$ into $W_0^{1,p}(\tilde{U})$ with $\tilde{U}$ slightly larger than U. Boundary straightenings and partitions of unity[11] will play a crucial role here, just like in the proof of the local version of the Gauss divergence or Green's theorem in (18.9), and the step to the global version in (9.21). The extension operator is first defined for $u \in C^1(\bar{U})$ and requires the boundary ∂U to be bounded and C^1 (locally the graph of a C^1-function).

Partitions of unity are first used to establish that $W^{1,p}(U)$ itself is the closure of $C^1(\bar{U})$ if ∂U is bounded and C^1. This is Theorem 23.33, which is pretty explicit in how the approximations are constructed. It shows that the assumptions on the boundary can be weakened. But we apply Theorem 23.33 to domains which are assumed to have ∂U bounded and C^1 for other reasons.

Remark 23.25. *The elements of the afore mentioned partitions are suitably chosen $\zeta \in C_c^\infty(\mathbb{R}^N)$ which, when multiplied by $u \in W^{1,p}(U)$, produce products $\zeta u \in W^{1,p}(U)$ to which the Leibniz rule applies*[12]*. Any statement that we want make about a function $u \in W^{1,p}(U)$ can be localised using partitions of unity, splitting u via $\zeta_0, \zeta_1, \ldots, \zeta_n \in C_c^1(\mathbb{R}^N) \subset C_c^\infty(\mathbb{R}^N)$ with $\zeta_0 + \zeta_1 + \cdots + \zeta_n \equiv 1$ on $\bar{U}$, writing*

$$u = u_0 + u_1 + \cdots + u_n = \zeta_0 u + \zeta_1 u + \cdots + \zeta_n u,$$

in which we take $\zeta_0 \in C_c^1(U)$ and $\zeta_1, \ldots, \zeta_n \in C_c^1(\mathbb{R}^N)$, just like in (18.10).

[11] Introduced in (18.10) and explained in Chapter 20.

[12] Evans' stronger assumption $\zeta \in C_c^\infty(U)$ for Leibniz' rule leads to cumbersome details.

23.5 Mollifiers and weak derivatives

The following remarks summarise what we have and what we don't have yet towards the density of $C^1(\bar{U})$ in $W^{1,p}(U)$.

Remark 23.26. *The basic estimates*[13]

$$|u^\varepsilon|_p \le |u|_p \quad \text{and} \quad |v^\varepsilon - v|_p \le \varepsilon\, |\nabla v|_p$$

for $u \in L^p(\mathbb{R}^N)$ and $v \in C^1_c(\mathbb{R}^N)$ sufficed to show that

$$u^\varepsilon \to u \quad \text{in} \quad L^p(\mathbb{R}^N) \quad \text{as} \quad \varepsilon \to 0 \tag{23.10}$$

for every $u \in L^p(\mathbb{R}^N)$ via

$$|u^\varepsilon - u|_p \le \underbrace{|u^\varepsilon - v^\varepsilon|_p}_{\le |u-v|_p} + |v^\varepsilon - v|_p + |v - u|_p, \tag{23.11}$$

in the proof of Theorem 23.4: first choose $v \in C^1_c(\mathbb{R}^N)$ such that $|u - v|_p$ is small, say less than δ, and then choose $\varepsilon > 0$ to make also $|v^\varepsilon - v|_p$ less than δ. The same statements hold with $\mathbb{R}^N$ replaced by an open set $U \subset \mathbb{R}^N$.

Remark 23.27. *If $u \in W^{1,p}(\mathbb{R}^N)$ and $v \in C^2_c(\mathbb{R}^N)$ then Remark 23.26 applies to*[14] *$w_i = u_{x_i} = D_i u \in L^p(\mathbb{R}^N)$ and $D_i v = v_{x_i} \in C^1_c(\mathbb{R}^N)$. Since*

$$D_i(u^\varepsilon) = (u^\varepsilon)_{x_i} = (u_{x_i})^\varepsilon = (D_i u)^\varepsilon \tag{23.12}$$

it follows that

$$u^\varepsilon \to u \quad \text{in} \quad W^{1,p}(\mathbb{R}^N) \quad \text{as} \quad \varepsilon \to 0.$$

The same statements do not hold with $\mathbb{R}^N$ replaced by an open set $U \subset \mathbb{R}^N$. We prove (23.12) below and then worry about what to do for U.

We have[15]

$$w_i^\varepsilon(x) = (u_{x_i})^\varepsilon(x) = (D_i u)^\varepsilon(x) = \int_{\mathbb{R}^N} \eta_\varepsilon(x-y)(D_i u)(y)\, dy \tag{23.13}$$

$$= \int_{\mathbb{R}^N} (D_i \eta_\varepsilon)(x-y) u(y)\, dy = \Big(\int_{\mathbb{R}^N} \eta_\varepsilon(x-y) u(y)\, dy\Big)_{x_i} = (D_i u^\varepsilon)(x),$$

which is and proves (23.12).

[13] Applied with u replaced by $u - v$.

[14] In practice: $u_{x_i} - v_{x_i}$.

[15] Note that D_i acts on u to give $D_i u$ which we can evaluate in x, $x - y$, y and so on.

Exercise 23.28. Explain why the inequalities in the above chain hold. Hint: you first need the techniques from Section 9.5, then the definition of weak derivatives in Theorem 23.11, and then again the techniques from Section 9.5.

We record the positive result in Remark 23.27 as

Theorem 23.29. *Let $u \in W^{1,p}(\mathbb{R}^N)$, $1 \le p < \infty$. Then*

$$u^\varepsilon \to u \quad \text{in} \quad W^{1,p}(\mathbb{R}^N) \quad \text{as} \quad \varepsilon \to 0.$$

Now what about $u \in W^{1,p}(U)$ if U is an open subset of $\mathbb{R}^N$? Note that we can extend not only u but also $w_1 = D_1 u, \dots, w_N = D_N u$ to $\mathbb{R}^N$ by setting $u(x) = w_1(x) = \cdots = w_N(x) = 0$ for $x \notin U$, but then we don't have that $w_i = D_i u$ in $L^p(\mathbb{R}^N)$ for $i = 1, \dots, N$. We can only conclude that

$$w_i^\varepsilon = D_i u^\varepsilon \quad \text{in} \quad L^p(U_\varepsilon), \quad U_\varepsilon = \{x \in U : B(x, \varepsilon) \subset U\}. \tag{23.14}$$

It is only on this issue that the reasoning in Remark 23.27 towards

$$u^\varepsilon \to u \quad \text{in} \quad W^{1,p}(U) \quad \text{as} \quad \varepsilon \to 0 \tag{23.15}$$

fails. We still have that

$$u^\varepsilon \to u \quad \text{and} \quad w_i^\varepsilon \to w_i \quad \text{for} \quad i = 1, \dots, N \quad \text{in} \quad L^p(\mathbb{R}^N) \quad \text{as} \quad \varepsilon \to 0,$$

and then also in $L^p(U)$. We next use localisations and translates to prove that, provided the boundary ∂U is sufficiently nice, there is a family[16] $u_\varepsilon \in C^1(\bar{U})$ with $u_\varepsilon \to u$ in $W^{1,p}(U)$. To do so we first establish equality in (23.14) for mollified translates of localised functions $\tilde{w}$ and $\tilde{u}$ obtained from $u \in W^{1,p}(U)$.

23.6 Shifts and localisation

The translation trick goes as follows. Given $h > 0$, a unit vector e and a function $u \in L^p(\mathbb{R}^N)$ we define $u_{he} \in L^p(\mathbb{R}^N)$ by[17]

$$u_h(x) = u(x + he)$$

and consider u_h^ε. Since clearly

$$(u_h)^\varepsilon = (u^\varepsilon)_h$$

[16] Do note ε is a subscript now, so $u_\varepsilon \neq u^\varepsilon$.

[17] Dropping e from the subscript notation.

it follows from (23.10) that

$$|u_h^\varepsilon - u_h|_p = |u^\varepsilon - u|_p \to 0 \quad \text{in} \quad L^p(\mathbb{R}^N) \quad \text{as} \quad \varepsilon \to 0.$$

But then

$$|u_h^\varepsilon - u|_p \le \underbrace{|u_h^\varepsilon - u_h|_p}_{=|u^\varepsilon - u|_p} + |u_h - u|_p = \underbrace{|u^\varepsilon - u|_p}_{\to 0} + \underbrace{|u_h - u|_p}_{\to 0}, \tag{23.16}$$

the latter because

$$|u_h - u|_p \le \underbrace{|u_h - v_h|_p}_{=|v-u|_p} + |v_h - v|_p + |v - u|_p \tag{23.17}$$

for every $v \in C_c(\mathbb{R}^N)$, similar to (23.11).

Exercise 23.30. Use (23.17) with $v \in C_c(\mathbb{R}^N)$ and $|v-u|_p$ as small as desired to prove that $u_h \to u$ in $L^p(\mathbb{R}^N)$. Hint: use the uniform continuity of each such v.

We conclude from (23.16) that

$$u_h^\varepsilon \to u \quad \text{in} \quad L^p(\mathbb{R}^N) \quad \text{as} \quad h \to 0 \quad \text{and} \quad \varepsilon \to 0, \tag{23.18}$$

which will be used in next both for $\tilde{u}$ and $\tilde{w}_i$ as functions extended by zero outside their original domain.

Recall that the limitation to U_ε in (23.14) kept us from concluding that $u^\varepsilon \to u$ in $W^{1,p}(U)$. We now localise u by multiplying it by a function $\zeta \in C_c^1(\mathbb{R}^N)$ and consider shifts $\tilde{u}_h$ of $\tilde{u} = \zeta u$ defined by the choice of a fixed unit vector e to be chosen in relation to ζ, and the local description of U and its boundary in and near the support of ζ. Note that

$$\tilde{u} \in W^{1,p}(\tilde{U}), \quad \tilde{U} = U \cup (\operatorname{supp}\zeta)^c,$$

and that

$$\tilde{u}_h \in W^{1,p}(\tilde{U}_h)$$

is defined by

$$\tilde{u}_h(x) = \tilde{u}(x + eh) \quad \text{for} \quad x \in \tilde{U}_h = \{x \in \mathbb{R}^N : x + eh \in \tilde{U}\}.$$

We extend $\tilde{u}$ and $\tilde{w}_i = D_i\tilde{u}$, defined in $L^p(\tilde{U})$, to $L^p(\mathbb{R}^N)$ by setting

$$\tilde{u}(x) = \tilde{w}_i(x) = 0 \quad \text{for} \quad x \notin \tilde{U} \quad \text{and} \quad i = 1, \dots, N$$

as before, and know from (23.18) that $u^\varepsilon, w_i^\varepsilon \in C_c^\infty(\mathbb{R}^N)$ have the property that

$$u_h^\varepsilon \to u \quad \text{and} \quad w_{ih}^\varepsilon \to w_i \quad \text{in} \quad L^p(\mathbb{R}^N) \quad \text{as} \quad h \to 0 \quad \text{and} \quad \varepsilon \to 0.$$

To conclude that

$$\tilde{u}_h^\varepsilon \to \tilde{u} \quad \text{in} \quad W^{1,p}(U)$$

we need

$$\tilde{w}_{ih}^\varepsilon = D_i \tilde{u}_h^\varepsilon \quad \text{in} \quad L^p(U), \tag{23.19}$$

and in view of (23.14) this will follow if

$$U \subset \tilde{U}_{h\varepsilon} = \{x \in \tilde{U}_h : B(x, \varepsilon) \subset \tilde{U}_h\}. \tag{23.20}$$

23.7 Global density of smooth functions

We use a partition of unity as in (18.10) with each $\zeta_1, \dots, \zeta_n$ taking care of some part of the boundary of U, and $\zeta_0 \in C_c^\infty(U)$.

Exercise 23.31. Use Theorem 23.29 to show that $(\zeta_0 u)^\varepsilon \to \zeta_0 u$ in $W^{1,p}(U)$ as $\varepsilon \to 0$.

Without loss of generality we continue the reasoning in the 2-dimensional setting, with

$$\text{supp } \zeta = [\tilde{a}, \tilde{b}] = [\tilde{a}_1, \tilde{b}_1] \times [\tilde{a}_2, \tilde{b}_2],$$
$$a_1 < \tilde{a}_1 < \tilde{b}_1 < b_1, \quad a_2 < \tilde{a}_2 < \tilde{b}_2 < b_2, \quad f : [a_1, b_1] \to (\tilde{a}_2, \tilde{b}_2)$$

Lipschitz continuous with Lipschitz constant L,

$$U \cap (a, b) = \{(x, y) : a_1 < x < b_1,\ f(x) < y < b_2\},$$

e the second unit vector, whence

$$\tilde{u}_h(x, y) = \tilde{u}(x, y + h)$$

and

$$\tilde{U}_h \supset \{(x, y) :\ a_1 < x < b_1,\ y > f(x) - h\}.$$

Now let $\lambda = \sqrt{1 + L^2}$ and $h = \lambda\varepsilon$. Then every point in $[\tilde{a}, \tilde{b}] \cap U$ with

$$x_{_N} \geq f(x_1, \dots, x_{_{N-1}}) + \lambda\varepsilon$$

is the center of an open ball with radius $\varepsilon > 0$ that is contained in $(a, b) \cap U$, provided ε is smaller than the distance from $[\tilde{a}, \tilde{b}]$ to the boundary of (a, b). This implies (23.20) holds with $h = \lambda\varepsilon$, whence

$$\tilde{u}_{\lambda\varepsilon}^\varepsilon \to \tilde{u} \quad \text{in} \quad W^{1,p}(U) \tag{23.21}$$

Exercise 23.32. To convince yourself of the statement preceding (23.21) draw a picture in the xy-plane with the line $y = Lx$ and find the point $P_\varepsilon = (0, \lambda\varepsilon)$ on the positive y-axis with distance ε to that line, and the point Q_ε on that line which realises this distance. Shift the origin $O = (0,0)$ to a point on the graph $y = f(x)$ contained in $[\tilde{a}, \tilde{b}]$, and pull the triangle $OP_\varepsilon Q_\varepsilon$ along. Specify the smallness condition on ε.

The above argument applies to every $\zeta_1, \ldots, \zeta_n$. Combined with Exercise 23.31 this allows to conclude that the following theorem has been proved.

Theorem 23.33. *Assume*[18] *that U allows a partition of unity $\zeta_0, \zeta_1, \ldots, \zeta_n$ such as used above. For every $u \in W^{1,p}(U)$ there exists a family $u_\varepsilon \in C_c^\infty(\mathbb{R}^N)$ with $u_\varepsilon \to u$ in $W^{1,p}(U)$. We can take*[19]

$$u_\varepsilon = (\zeta_0 u)^\varepsilon + \sum_{i=1}^{n} (\zeta_i u)^\varepsilon_{\lambda\varepsilon e_i},$$

in which λ is the largest $\sqrt{1+L^2}$ that occurs in the construction. Thus the result is valid under the assumption that ∂U is compact and uniformly Lipschitz continuous[20]*. If $u \in W^{k,p}(U)$ with $k \in \mathbb{N}$ and $1 \le p < \infty$ then $u_\varepsilon \to u$ in $W^{k,p}(U)$.*

23.8 Estimates and embeddings for $W_0^{1,p}(U)$

Estimates derived for functions in $C_c^1(U)$ carry over to functions in $W_0^{1,p}(U)$, and there are two basic estimates to which this principle is applied. The first Gagliardo-Nirenberg-Sobolev estimate is

$$|u|_q \le C_{p,N} |\nabla u|_p \quad \text{for} \quad \frac{1}{q} = \frac{1}{p} - \frac{1}{N} \quad \text{if} \quad 1 \le p < N, \tag{23.22}$$

in which the norm[21] of ∇u is evaluated via an integral over the whole of U, and over the whole of $\mathbb{R}^N$ in the derivation, via repeated application of the one-dimensional estimate[22]

$$|u|_\infty \le \frac{1}{2} |u|_1 \quad \text{for} \quad u \in C_c^1(\mathbb{R}), \tag{23.23}$$

[18] See Chapter 20 for C^1-boundaries. TO DO: non smooth boundaries!
[19] Bringing unit vectors e_i back into the notation, these do not have to the basis vectors.
[20] Give a definition of what this should mean.
[21] For the p-norm of ∇u the p-norm of any vector norm of $\nabla u(x)$ can be used.
[22] The case $p = N = 1, q = \infty$ in (23.22), does not generalise to $p = N > 1, q = \infty$.

first[23] in the special case that $p = 1$, with a clever use of the Hölder's inequality with exponents satisfying

$$\frac{1}{p_1} + \cdots + \frac{1}{p_{n-1}} = 1.$$

The general case in (23.22) follows from putting u^γ for u and a follow your nose estimate invoking Hölder's inequality for the integral of $\gamma|u|^{\gamma-1}u_{x_i}$, which involves a particular choice of γ to get the exponents right. The constant $C_{p,\mathrm{N}}$ blows up as $p \to \mathrm{N}$ (from below).

The second (Morrey) estimate[24] is usally stated as

$$|u(x_1) - u(x_2)| \le C_{p,\mathrm{N}} |\nabla u|_p |x_1 - x_2|^\alpha \quad \text{for} \quad \alpha = 1 - \frac{\mathrm{N}}{p} \quad \text{if} \quad p > \mathrm{N},$$

but the p-norm of ∇u may be restricted to the intersection of the two balls centered in x_1 and x_2 with radius $|x_1 - x_2|$. That is

$$|u(x_1) - u(x_2)| \le C_{p,N} |\nabla u|_{L^p(W_{x_1x_2})} |x_1 - x_2|^{1-\frac{N}{p}}, \tag{23.24}$$

in which

$$W_{x_1x_2} = B(x_1, |x_1 - x_2|) \cap B(x_2, |x_1 - x_2|).$$

This estimate is derived from the inequality

$$\int_{C_R} |u - u(0)| \le \frac{R^\mathrm{N}}{\mathrm{N}} \int_{C_R} \frac{|u_r|}{r^{\mathrm{N}-1}} \tag{23.25}$$

for cones described in polar coordinates as

$$C_R = \{x = r\omega : 0 \le r \le R,\, \omega \in A\},$$

with A a nice subset of the unit sphere, and u_r denoting the radial derivative. The r-part of the integral in (23.25) is in some sense the counter part of (23.23), and integral on the right hand side is estimated using a follow your nose estimate invoking Hölder's inequality.

The Morrey estimate (23.24) is then proved estimating

$$|u(x_1) - u(x_2)| \le |u(x_1) - u(x)| + |u(x) - u(x_2)|,$$

and integrating over the intersection of the two cones C_1 and C_2 centered in x_1 and x_2, chosen to have $C_1 \cup C_2$ equal to the union of the two balls mentioned earlier. Again the constant $C_{p,\mathrm{N}}$ blows up as $p \to \mathrm{N}$ (from above).

[23] See Section **??**.
[24] See Section **??**.

Exercise 23.34. Let $1 \le p < \mathrm{N}$. Prove that $W_0^{1,p}(U) \subset L^q(U)$ if $\frac{1}{q} \ge \frac{1}{p} - \frac{1}{\mathrm{N}}$ if U is bounded, and that in that case $|\nabla u|_p$ defines an equivalent norm on $W_0^{1,p}(U)$. Theorem 23.21 already stated the desired estimate. Make C_{pU} as explicit as possible in terms of p, N and the measure of U.

Exercise 23.35. Let $p > \mathrm{N}$. Prove that $W_0^{1,p}(U) \subset C^\alpha(U)$ for $\alpha = 1 - \frac{\mathrm{N}}{p}$, in which

$$C^\alpha(U) = \{u \in C(U) : [u]_\alpha < \infty\} \quad \text{where} \quad [u]_\alpha = \sup_{x_1 \neq x_2} \frac{|u(x_1) - u(x_2)|}{|x_1 - x_2|^\alpha},$$

the supremum taken over the whole of U.

Exercise 23.36. This is to convince you that it is better to rename $C^\alpha(U)$ and write $C^\alpha(\bar{U})$: show that for U bounded and $\alpha \in (0, 1]$, every $u \in C^\alpha(U)$ extends to a continuous function on $\bar{U}$, and that the space $C^\alpha(U)$ is a Banach space with norm defined by[25] $|u|_\alpha = |u|_\infty + [u]_\alpha$.

Exercise 23.37. Let $p > \mathrm{N}$, U bounded, $\alpha = 1 - \frac{\mathrm{N}}{p}$. Use the AA Theorem to prove that every bounded sequence u_n in $W_0^{1,p}(U)$ has a subsequence that, considered as a sequence in $C(\bar{U})$, converges uniformly to a limit u, which is also in $C_0^\alpha(U)$, the subspace consisting of functions $u \in C^\alpha(U)$ which vanish on ∂U. Verify that this subspace has the seminorm $[\cdot]_\alpha$ as an equivalent norm.

In view of Exercise 23.37 the embedding

$$W_0^{1,p}(U) \to C(\bar{U})$$

is compact for $p > \mathrm{N}$ if U is bounded. This is Theroem 23.20. Since $C(\bar{U}) \subset L^p(U)$, with the obvious bound on the norms, it then also holds that

$$W_0^{1,p}(U) \to L^p(U) \quad \text{is compact if} \quad U \quad \text{is bounded,} \tag{23.26}$$

but this holds for all $p \ge 1$ because of Theorem 23.21 via a different argument[26].

[25] If you don't use Greek letters for Lebesgue norms this will not confuse.

[26] But it is still the AA Theorem after all.

Exercise 23.38. Let $1 \leq p < \mathrm{N}$ and $U \subset \mathbb{R}^{\mathrm{N}}$ open and bounded. Prove that the embedding

$$W_0^{1,p}(U) \to L^q(U)$$

is compact if

$$\frac{1}{q} > \frac{1}{p} - \frac{1}{\mathrm{N}}.$$

Hint: use Theorem 23.21 and interpolation inequalities with the p-norms.

23.9 Statements for $W^{1,p}(U)$ via extension

Given a bounded domain U we look for a slightly larger domain $\tilde{U}$ such that every $u \in W^{1,p}(U)$ extends to a $\tilde{u} \in W_0^{1,p}(\tilde{U})$ in the sense that $\tilde{u}(x) = u(x)$ for (almost) all $x \in U$. The extension map

$$u \in W^{1,p}(U) \xrightarrow{E} \tilde{u} \in W_0^{1,p}(\tilde{U})$$

should be linear and bounded. The extensions are first defined for $u \in C^1(\bar{U})$ and require U bounded and $\partial U \in C^1$.

I usually followed Evans' approach in which the extensions are first defined locally for u and then glued together using a partition of unity, but it came to me that here too it is in fact simpler to first cut up u in smaller pieces $\zeta_i u$ and choose globally defined extensions of $\zeta_i u$ rather than locally defined extensions of u. This requires a suitable set of functions

$$\zeta_0 \in C_c^\infty(O_0), \zeta_1 \in C_c^\infty(O_1), \ldots, \zeta_n \in C_c^\infty(O_n),$$

with $0 \leq \zeta_i \leq 1$ and

$$\zeta_0(x) + \zeta_1(x) + \cdots + \zeta_n(x) = 1 \quad \text{for all} \quad x \in \bar{U}$$

with $O_0 \subset \bar{O}_0 \subset U$ and $O_1, \ldots, O_n$ chosen so as to allow globally defined extensions $\tilde{u}_1, \ldots, \tilde{u}_n$ of $u_1 = \zeta_1 u, \ldots, u_n = \zeta_n u$ which define

$$u \xrightarrow{E_i} \tilde{u}_i$$

as a linear map with

$$|\tilde{u}_i|_{1,p} \leq C_i \, |u|_{1,p},$$

allowing to define

$$u \in C_c^1(\bar{U}) \xrightarrow{E} C_c^1(\tilde{U}) \quad \text{by} \quad u \to \zeta_0 u + \tilde{u}_1 + \cdots + \tilde{u}_n.$$

In view of the Leibniz rule

$$(\zeta u)_{x_j} = \zeta u_{x_j} + \zeta_{x_j} u$$

it follows that

$$|Eu|_{1,p} \leq C\,|u|_{1,p},$$

with C some horrible constant depending on $\tilde{U}$ and U via the norms of ζ_i in C^1. The functions $\zeta_1, \ldots, \zeta_n$ are chosen to allow a C^1-coordinate transformation similar to the ones used by Evans. It all looks a bit cleaner if the O_i are taken, after a permutation of coordinates, in cylindrical form as $C_i = B_i \times I_i$, with B_i an open ball, I_i a bounded interval, and

$$U \cap C_i = \{x = (x_1, \ldots, x_{N-1}, x_N) \in C : \, x_N > \gamma_i(x_1, \ldots, x_{N-1})\},$$

in which $\gamma_i : \bar{B}_i \to I_i$ is C^1, and the supports of the ζ_i are contained in smaller cylinders $\tilde{C}_i = \tilde{B}_i \times \tilde{I}_i \subset\subset C_i$. The extensions of $\zeta_i u$ are then defined via the transformations in Appendix C.1 and the higher order reflection in Section 5.4.

Finally

$$u \in C^1(\bar{U}) \xrightarrow{E} \tilde{u} \in C_c^1(\tilde{U}),$$

as then defined with the desired $W^{1,p}$-estimates, has to be extended as a map from $W^{1,p}(U)$ to $W_0^{1,p}(\tilde{U})$ using the density of $C^1(\bar{U})$ in $W^{1,p}(U)$.

23.10 The extension and trace operators

We needed the boundedness of U and the C^1-regularity of ∂U to define an extension operator

$$C^1(\bar{U}) \xrightarrow{E} C_c^1(\tilde{U})$$

with the $W^{1,p}(\tilde{U})$-norm bound of Eu controlled by the $W^{1,p}(U)$-norm of u. The domain $\tilde{U}$ can be taken as close to U as desired by taking the cylinders C_i small, and E extends to

$$W^{1,p}(U) \xrightarrow{E} W_0^{1,p}(\tilde{U})$$

via the density result in Theorem 23.33. The other important operator is the bounded linear trace operator

$$W^{1,p}(U) \xrightarrow{T} L^p(\partial U)$$

in Section 5.4 of Evans, which extends

$$u \in C^1(\bar{U}) \to u|_{\partial U} \in C^1(\partial U).$$

Evans defines it locally, first under the assumption that ∂U is flat and $u \in C^1(\bar{U})$. The same splitting as in Theorem 23.33 can be used to first define $T(\zeta_i u)$ instead, for $u \in C^1(\bar{U})$, so

$$u \in C^1(\bar{U}) \to \zeta_i u = u_i \in C^1(\bar{U} \cap \bar{C}_i) \to u_i|_{\partial U} \in C^1(\partial U).$$

The local coordinate transformation flattening $\partial U \cap \bar{C}_i$ is not even needed, as u_i is defined for all $x_{_N} \geq \gamma(x_1, \ldots, x_{\mathrm{N}-1})$ with $(x_1, \ldots, x_{\mathrm{N}-1}) \in B_i$ and vanishes for x_{N} large. Thus

$$Tu_i(x_1, \ldots, x_{\mathrm{N}-1}) = u_i(x_1, \ldots, x_{\mathrm{N}-1}, \gamma(x_1, \ldots, x_{\mathrm{N}-1})) = -\int_\gamma^\infty (u_i)_{x_{\mathrm{N}}},$$

and the p-norm on B_i is estimated by the p-norm of ∇u_i, the factor

$$\left(1 + \gamma_{x_1}^2 + \cdots + \gamma_{x_{\mathrm{N}-1}}^2\right)^{\frac{1}{2}}$$

being irrelevant for the estimate. The characterisation of the kernel of T as in Theorem 2 of Section 5.4 is also done locally then, as Evans observes in (6), in which the flattening avoids cumbersome notation in the already technical proof that follows. Actually the proof is not so hard. It relies on this estimate, formulated in $\mathbb{R}^2$ without loss of generality for $u \in C_c^2(\mathbb{R}^2)$:

$$\int_{-\infty}^\infty |u(x,y)|^p \, dx \leq 2^p \left(\int_{-\infty}^\infty |u(x,0)|^p \, dx + y^{p-1} \int_0^y \int_{-\infty}^\infty |u_y|^p \right). \quad (23.27)$$

Exercise 23.39. Prove (23.27) and explain why it holds for $u \in W^{1,p}(\mathbb{R}^2_+)$ with compact support in $\mathbb{R} \times [0, \infty)$.

Exercise 23.40. If such a u has $Tu = 0$, then the functions u_m defined by $u_m(x,y) = (1 - \zeta(my))u(x,y)$ with $\zeta \in C_c^\infty([0,2))$ and $\zeta \equiv 1$ on $[0,1]$, $\zeta' \leq 0$ on $[0,2)$ are in $W_0^{1,p}(\mathbb{R}^2_+)$ and converge to u in $W^{1,p}(\mathbb{R}^2_+)$. Prove this and conclude that $u \in W_0^{1,p}(\mathbb{R}^2_+)$. Hint: you have to use Exercise 23.27.

23.11 More exercises that fill in details

Exercise 23.41. Let U be a bounded domain in $\mathbb{R}^2 = \{(x,y) : x, y \in \mathbb{R}\}$ and $\zeta \in C^1(\mathbb{R}^2)$. Prove that $\zeta u \in W^{1,p}(U)$ if $u \in W^{1,p}(U)$.

Exercise 23.42. Introduce new coordinates ξ, η by

$$x = x_0 + a\xi + b\eta, \quad y = y_0 + c\xi + d\eta,$$

with $ad \neq bc$, define V by $(\xi, \eta) \in V \iff (x, y) \in U$, and write $v(\xi, \eta) = u(x, y)$. Show that

$$u \in W^{1,p}(U) \iff v \in W^{1,p}(V)$$

and that this correspondence defines a linear homeomorphism between the two Sobolev spaces.

Exercise 23.43. Assume $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$ is C^1, injective on $\bar{U}$, with invertible Jacobian matrix in every $(x, y) \in \bar{U}$. Then $u \to u \circ \Phi^{-1} = v$ defines a bijective map from $C^1(\bar{U}) \to C^1(\bar{V})$ where $V = \Phi(U)$. Show that

$$\frac{1}{C}|v|_{W^{1,p(V)}} \leq |u|_{W^{1,p(U)}} \leq C|v|_{W^{1,p(V)}}$$

for some $C > 1$.

Exercise 23.44. Explain why this map uniquely extends to a bijection from $W^{1,p}(U)$ to $W^{1,p}(V)$ if $\partial U \in C^1$.

Exercise 23.45. The intersection of $W^{1,p}(U)$ and $C(\bar{U})$ is a Banach space with norm e.g. $|u|_\infty + |u_x|_p + |u_y|_p$. Explain why this Banach space is the closure of $C^1(\bar{U})$ with respect to this norm.

Exercise 23.46. Let $u \in C^1(\mathbb{R}^3)$. Write $u(x,y,z) = v(r,\theta,\phi)$, with

$$x = r\sin\theta\cos\phi,\ y = r\sin\theta\sin\phi,\ z = r\cos\theta,$$

and let $C_{R,\psi}$ be the region in $\mathbb{R}^3$ defined by $0 \le r \le R$, $0 \le \theta \le \psi$ and ϕ free. If $0 < \psi < \frac{\pi}{2}$ and $R > 0$, then ψ is called[1] the opening angle of the closed cone $C_{R,\psi}$, and R is called the radius of the cone. For $\psi > \frac{\pi}{2}$ we don't call $C_{R,\psi}$ a cone. It's a half ball with radius R if $\psi = \frac{\pi}{2}$ and a ball if $\psi = \pi$. Evans only integrates over balls. This is to clarify and improve the remark after the proof of Theorem 4 in his Section 5.6.2 on Morrey's inequality. Assuming that $u(0,0,0) = v(0,\theta,\phi) = 0$ we used

$$|v(r,\theta,\phi)| \le \int_0^r |v_r(\rho,\theta,\phi)|\,d\rho$$

to estimate

$$\begin{aligned}\int_{C_{R,\psi}} |u| &= \int_0^{\psi}\int_0^{2\pi}\int_0^R |v(r,\theta,\phi)|\,r^2\sin\theta\,dr\,d\phi\,d\theta \\ &\le \int_0^{\psi}\int_0^{2\pi}\int_0^R\int_0^r |v_r(\rho,\theta,\phi)|\,d\rho\,r^2\sin\theta\,dr\,d\phi\,d\theta\end{aligned}$$

(interchanging the order of the integrations with respect to r and ρ, throwing a away one negative term and replacing ρ 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 $u_r = xu_x + yu_y + zu_z = v_r$. Use generalised polar coordinates

$$x_1 = r\omega_1 = r\cos\theta_1 = rc_1,\ x_2 = r\omega_2 = r\sin\theta_1\cos\theta_2 = rs_1c_2,\ x_3 = r\omega_3 = rs_1s_2c_3,$$

$$\ldots, x_{N-2} = r\omega_{N-1} = rs_1\cdots s_{N-2}c_{N-1},\ x_N = r\omega_N = rs_1\cdots s_{N-2}s_{N-1}$$

to generalise and improve this estimate as

$$\int_{C_{R,\psi}} |u| + \frac{1}{N}\int_{C_{R,\psi}} r|u_r| \le \frac{R^N}{N}\int_{C_{R,\psi}} \frac{|u_r|}{r^{N-1}} \tag{23.1}$$

for $C_{R,\psi}$ in $\mathbb{R}^N$ defined by $0 \le r \le R$, $0 \le \theta_1 \le \psi$ and $\theta_2, \ldots, \theta_{N-1}$ free.

Exercise 23.47. (continued) Let $\omega \in \mathbb{R}^N$ with $|\omega| = 1$. Explain why estimate (23.1) holds with $C_{R,\psi}$ replaced by the closed cone

$$C_{R,\omega,\psi} = \{x \in \mathbb{R}^N :\ |x| \le R,\ x\cdot\omega \ge |x|\cos\psi\}, \tag{23.2}$$

a cone with direction[2] ω and opening angle $\psi \in (0, \frac{\pi}{2})$.

---

[1] And not 2ψ.

[2] Note $\omega = e_3$ in the 3-dimensional example, $\omega = e_1$ in the N-dimensional example.

Exercise 23.48. Explain why the measure $|C_{1,\omega,\psi}|$ of the cone $C_{1,\omega,\psi}$ defined by (23.2) is given by

$$\int_{C_{1,\omega,\psi}} 1 = \frac{2\pi}{N} \int_0^\psi \sin^{N-2}\theta_1 \, d\theta_1 \int_0^\pi \sin^{N-1}\theta_2 \, d\theta_2 \cdots \int_0^\pi \sin\theta_{N-2} \, d\theta_{N-2}. \quad (23.3)$$

if $R = 1$. Call this number $C_{N\psi}$. Show that

$$C_{N\psi} = \frac{\omega_{N-1}}{N} \int_0^\psi \sin^{N-2}\theta \, d\theta,$$

in which ω_{N-1} is the measure of the unit ball in $\mathbb{R}^{N-1}$. Correct my mistakes. Is there a quicker way? Does the integral simplify if $\psi = \frac{\pi}{3}$?

Exercise 23.49. We use Hölder's inequality[3] to estimate

$$\int_{C_{R,\omega,\psi}} \frac{|u_r|}{r^{N-1}} \le \left(\int_{C_{R,\omega,\psi}} \left(\frac{1}{r^{N-1}} \right)^{p'} \right)^{\frac{1}{p'}} \underbrace{\left(\int_{C_{R,\omega,\psi}} |u_r|^p \right)^{\frac{1}{p}}}_{|u_r|_{L^p(C_{R,\omega,\psi})}} \quad \text{with} \quad \frac{1}{p} + \frac{1}{p'} = 1.$$

Show that

$$\int_{C_{R,\omega,\psi}} \frac{|u_r|}{r^{N-1}} \le \left(C_{N\psi} \, \frac{p-1}{p-N} \right)^{1-\frac{1}{p}} |u_r|_{L^p(C_{R,\omega,\psi})} \, R^{1-\frac{N}{p}} \quad (23.4)$$

if $p > N$. Explain why the estimate holds for all $u \in C^1(C_{R,\omega,\psi})$. Why does the estimate fail for $p \le N$?

Combining (23.1) and (23.4) we have

$$\frac{N}{R^N} \int_{C_{R,\omega,\psi}} |u| \le \left(C_{N\psi} \, \frac{p-1}{p-N} \right)^{1-\frac{1}{p}} |u_r|_{L^p(C_{R,\omega,\psi})} \, R^{1-\frac{N}{p}}, \quad (23.5)$$

in which ψ can have any value in $[0, \pi]$. This estimate is hidden in Step 2 of the proof of Theorem 4 in Evans' Section 5.6.2, and only given there for $\psi = \pi$.

[3] Which for integrals follows from the inequality in Section 12.9.

Exercise 23.50. For $R > 0$, $x_1, x_2, \omega_1, \omega_2 \in \mathbb{R}^N$ with $|\omega_1| = |\omega_2| = 1$ and angles ψ_1, ψ_2, consider $C_1 = x_1 + C_{R,\omega_1,\psi_1}$ and $C_2 = x_2 + C_{R,\omega_2,\psi_2}$. Use

$$|C_1 \cap C_2|\,|u(x_1) - u(x_2)| \le \int_{C_1} |u(x_1) - u(x)|dx + \int_{C_2} |u(x) - u(x_2)|dx$$

and (23.5) to show that

$$|C_1 \cap C_2|\,|u(x_1) - u(x_2)| \le \frac{R^N}{N}\left(C_{N\psi_1}^{1-\frac{1}{p}} + C_{N\psi_2}^{1-\frac{1}{p}}\right)\left(\frac{p-1}{p-N}\right)^{1-\frac{1}{p}} |\nabla u|_{L^p(C_1 \cup C_2)}\, R^{1-\frac{N}{p}}.$$

Exercise 23.51. In Exercise 23.50 take[4]

$$R = |x_1 - x_2|,\ \omega_1 = \frac{x_2 - x_1}{R} = -\omega_2,\ \psi = \frac{\pi}{3},$$

and prove that

$$|u(x_1) - u(x_2)| \le C(N,p)|\nabla u|_{L^p(B_{|x_1-x_2|}(x_1) \cap (B_{|x_1-x_2|}(x_2))}\, |x_1 - x_2|^{1-\frac{N}{p}}, \qquad (23.6)$$

in which

$$C(N,p) = c_{_N} \frac{\left(\int_0^{\frac{\pi}{3}} \sin^{N-2}\theta\, d\theta\right)^{1-\frac{1}{p}}}{{\omega_{_{N-1}}}^{\frac{1}{p}}} \qquad (23.7)$$

with $c_{_N}$ to be specified by you. Hint: show first that the measure A_N of the set described by

$$x_1 \ge 0,\ x_2 = r\cos\theta_1,\ x_3 = r\sin\theta_1\cos\theta_2, \ldots,\ x_{_N} = r\sin\theta_1 \cdots \sin\theta_{_{N-2}},\ r \ge 0,$$

and

$$x_1 + \frac{r}{\sqrt{3}} \le \frac{1}{2}$$

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 $N > 3$.

[4] Sketch the balls with centers x_1 and x_2 and radius $|x_1 - x_2|$ to see what's going on.

Thus we have improved the estimate stated by Evans without proof in the remark following the proof of Theorem 4. Now recall the definitions of $W^{1,p}(U)$ and $W_0^{1,p}(U)$ for U in $\mathbb{R}^N$ bounded, open and connected,

$$u \in W^{1,p}(U) \iff u, u_{x_1}, \ldots, u_{x_N} \in L^p(U),$$

and, for $1 \leq p < \infty$, the space $W_0^{1,p}(U)$ being the closure of $C_c^1(U)$ in the Banach space $W^{1,p}(U)$.

Remark 23.52. *Every statement we will ever be able to make about $W^{1,p}(U)$ will be based on a statement about $W_0^{1,p}(\tilde{U})$ for a slightly larger $\tilde{U}$ and some extension $\tilde{u}$ of u from U to $\tilde{U}$, which will heavily depend on the properties of the boundary of U.*

Exercise 23.53. Let $u \in W_0^{1,p}(U)$, U in $\mathbb{R}^N$ bounded, open and connected, $N < p < \infty$ and let $\alpha = 1 - \frac{N}{p}$. Take a sequence $u_n \in C_c^\infty(U)$ with $u_n \to u$ in $W^{1,p}(U)$. Prove that u_n is a Cauchy sequence in[5] $C^\alpha(\bar{U})$, and that its limit $\bar{u}$ in $C^\alpha(\bar{U})$ has the property that $|u - \bar{u}|_p = 0$. Prove that the map $u \to \bar{u}$ is linear and continuous from $W_0^{1,p}(U)$ to $C^\alpha(\bar{U})$.

Exercise 23.54. (continued) A rough estimate for the seminorm

$$[u]_\alpha = \sup_{\substack{x,y \in U \\ x \neq y}} \frac{|u(x) - u(y)|}{|x - y|^\alpha}$$

with $\alpha = 1 - \frac{N}{p}$: show that

$$[u]_{1-\frac{N}{p}} \leq C(p, N) |\nabla u|_{L^p(U)},$$

and also show that

$$|u|_\infty \leq \tilde{C}(p, N, U) \, |\nabla u|_{L^p(U)}$$

for some constant $\tilde{C}(p, N, U)$ you can make as precise as you want. Give just one. Hint: first for $u \in C_c^1(U)$, reason as in Exercise 23.53 to get the estimate for all $u \in W_0^{1,p}(U)$ if $p > N$.

Exercise 23.55. Show that $C^\alpha(\bar{U})$ is a Banach space.

[5] See Exercise 23.36.

Exercise 23.56. Show that $W_0^{1,p}(U)$ is compactly embedded in $C_0(U)$. Hint: use the Ascoli-Arzéla theorem via Exercise 23.54.

Exercise 23.57. 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 α and β only. Hint: it is easy to estimate $[u]_\alpha$ by a product of powers of $[u]_\beta$ and $|u|_\infty$. Use Young's inequality

$$ab \le \frac{\varepsilon^p a^p}{p} + \frac{b^q}{q\varepsilon^q} \quad \text{for} \quad \varepsilon > 0,\ a, b \ge 0,\ p, q \ge 1 \quad \text{with} \quad \frac{1}{p} + \frac{1}{q} = 1$$

to conclude.

Exercise 23.58. Use Exercises 23.56 and 23.57 to conclude that $W_0^{1,p}(U)$ is compactly embedded in $C^\beta(\bar{U})$ is $0 < \beta < 1 - \frac{N}{p}$.

Exercise 23.59. Show that $W_0^{1,p}(U)$ is embedded in $h^\alpha(\bar{U})$, the closed subspace[6] of $C^\alpha(\bar{U})$ for which

$$\sup_{\substack{x,y\in U \\ 0<|x-y|\le\varepsilon}} \frac{|u(x) - u(y)|}{|x-y|^\alpha} \to 0 \quad \text{as} \quad \varepsilon \to 0,$$

if $\alpha = 1 - \frac{N}{p}$ and $p > N$.

It is easy to see that

$$|f(x)| \le \frac{1}{2}\int_{-\infty}^{\infty} |f'(x)|dx$$

for $f \in C_c^1(\mathbb{R})$ and apply it to $x \to u(x,y)$ and $y \to u(x,y)$ to derive an estimate for the 2-norm of $u \in C_c^1(\mathbb{R}^2)$ in terms of the 1-norms of u_x and u_y. The result is

$$\iint_{\mathbf{R}^2} |u|^2 \le \frac{1}{4}\iint_{\mathbf{R}^2} |u_x| \iint_{\mathbf{R}^2} |u_y| \quad \text{from which} \quad |u|_2 \le \frac{1}{2}\max_{i=1,2} |u_{x_i}|_1$$

[6] These are the so-called little Hölder spaces, unlike $C^\alpha(\bar{U})$ they are separable.

follows (I'm writing single bars with subscript p for the p-norm in L^p).

The same trick with $x \to u(x,y,z)$, $y \to u(x,y,z)$ and $z \to u(x,y,z)$ and Hölder's inequality applied 3 times with exponents $p_1 = p_2 = \frac{1}{2}$ applied to the successive integrations with respect to x, y, z gives

$$\begin{aligned}
\iiint_{\mathbf{R}^3} |u|^{\frac{3}{2}} &\leq \iiint_{\mathbf{R}^3} (\frac{1}{2}\int_x |u_x|)^{\frac{1}{2}} (\frac{1}{2}\int_y |u_y|)^{\frac{1}{2}} (\frac{1}{2}\int_z |u_z|)^{\frac{1}{2}} \\
&= (\frac{1}{2})^{\frac{3}{2}} \int_z \int_y \int_x (\int_x |u_x|)^{\frac{1}{2}} (\int_y |u_y|)^{\frac{1}{2}} (\int_z |u_z|)^{\frac{1}{2}} \\
&\leq (\frac{1}{2})^{\frac{3}{2}} \int_z \int_y (\int_x |u_x|)^{\frac{1}{2}} (\int_x \int_y |u_y|)^{\frac{1}{2}} (\int_x \int_z |u_z|)^{\frac{1}{2}} \\
&\leq (\frac{1}{2})^{\frac{3}{2}} \int_z (\int_y \int_x |u_x|)^{\frac{1}{2}} (\int_x \int_y |u_y|)^{\frac{1}{2}} (\int_y \int_x \int_z |u_z|)^{\frac{1}{2}} \\
&\leq (\frac{1}{2})^{\frac{3}{2}} (\int_z \int_y \int_x |u_x|)^{\frac{1}{2}} (\int_z \int_x \int_y |u_y|)^{\frac{1}{2}} (\int_y \int_x \int_z |u_z|)^{\frac{1}{2}}
\end{aligned}$$

(in each integration one of the 3 factors does not depend on the integration variable).

Exercise 23.60. Prove that

$$|u|_{\frac{3}{2}} \leq \frac{1}{2} |u_x|_1^{\frac{1}{3}} |u_y|_1^{\frac{1}{3}} |u_z|_1^{\frac{1}{3}} \leq \frac{1}{2} \max_{i=1,2,3} |u_{x_i}|_1$$

generalises to

$$|u|_{\frac{N}{N-1}} \leq \frac{1}{2} \max_{i=1,\dots,N} |u_{x_i}|_1$$

for $u \in C_c^1(\mathbb{R}^N)$ via N integrations and Hölder's inequality with $p_1 = \cdots = p_N = \frac{1}{N-1}$ applied in every step (in each integration one of the N factors does not depend on the integration variable).

Applied to $u^\gamma = |u|^{\gamma-1}u$ it follows via Hölder's inequality with

$$\frac{1}{p} + \frac{1}{p'} = 1$$

that

$$\begin{aligned}
|u|^\gamma_{\frac{\gamma N}{N-1}} = |u^\gamma|_{\frac{N}{N-1}} &\leq \frac{1}{2} \max_{i=1,\dots,N} |\gamma u^{\gamma-1} u_{x_i}|_1 \leq \frac{\gamma}{2} \max_{i=1,\dots,N} |u^{\gamma-1}|_{p'} |u_{x_i}|_p \\
&= \frac{\gamma}{2} |u|^{\gamma-1}_{\frac{(\gamma-1)p}{p-1}} \max_{i=1,\dots,N} |u_{x_i}|_p
\end{aligned}$$

in which γ can be chosen to have equal subscripts of $|u|$ in the first and last expression in this chain.

Exercise 23.61. For $1 \leq p < N$ you should check that this gives

$$q = \frac{\gamma N}{N-1} = \frac{(\gamma-1)p}{p-1} = \frac{pN}{N-p},$$

which you may prefer to memorise as

$$\frac{1}{q} = \frac{1}{p} - \frac{1}{N}.$$

What's the value of γ? Dividing by $|u|_q^{\gamma-1}$ on both sides you get

$$|u|_q \leq C_{Np} \max_{i=1,\dots,N} |u_{x_i}|_p$$

with an explicit constant C_{Np}. Give this value. Check again that $1 \leq p < N$ is the assumption to make here.

Exercise 23.62. In fact we have

$$|u|_q \leq C_{Np} |u_{x_1}|_p^{\frac{1}{N}} \cdots |u_{x_N}|_p^{\frac{1}{N}}$$

Prove that this estimate holds for all $u \in W_0^{1,p}(U)$ if $1 \leq p < N$ and

$$\frac{1}{q} = \frac{1}{p} - \frac{1}{N}.$$

Exercise 23.63. Show for $N > 2$ that[7]

$$|u|_{\frac{N}{N-2}} \leq \frac{1}{4} \max_{i \neq j} |u_{x_i x_j}|_1$$

for $u \in C_c^2(\mathbb{R}^N)$. Only the mixed derivatives are needed[8].

Exercise 23.64. Let $u \in W_0^{1,p}(U)$, U bounded, $1 \leq p < N$. Prove that

$$|u|_q \leq C_{p,q,N,|U|} |\nabla u|_p$$

[7] There are similar estimates for $u \in C_c^3(\mathbb{R}^N)$, $u \in C_c^4(\mathbb{R}^N)$,....

[8] Nice project: versions similar to Exercise 23.62 for $W_0^{2,p}$ with only mixed derivatives.

if

$$\frac{1}{q} > \frac{1}{p} - \frac{1}{N},$$

with a constant depending only on p, q, N and the measure $|U|$ of the domain. Hint: estimate the q-norm in terms of the 1-norm and the p-norm using Hölder's inequality[9].

Exercise 23.65. A special case in Exercise 23.64 is $q = p$, and the inequality for $p = q = 2$ is called Poincaré's inequality. For $1 \leq p < N$ it makes that

$$u \to |\nabla u|_p$$

is an equivalent norm on $W_0^{1,p}(U)$, which was defined as the closure of $C_c^1(U)$ in $W^{1,p}(U)$ with respect to the norm defined by

$$|u|_{1,p}^p = |u|_p^p + |u_{x_1}|_p^p + \cdots + |u_{x_N}|_p^p$$

Show that these norms are also equivalent for $N \leq p < \infty$.

[9] Interpolation between 1 and p similar to the interpolation in Exercise 23.57.

24 Riesz or no Riesz?

This chapter describes an example which in Chapter 25 will turn out to be the standard result for a large class of boundary value problems for elliptic partial differential equations which have a certain symmetry property. Remark 24.8 below puts the relation in perspective. The results are formulated in exercises that culminate in Remark 24.17, and are based on elementary Hilbert space theory.

Recall that in every Hilbert space the Riesz representation Theorem 6.31 is applicable, so also in $l^{(2)}$, the standard Hilbert space $H = l^{(2)} = L^2(\mathbb{N})$ with the counting measure on $\mathbb{N}$. Elements u in this H are functions

$$u : \mathbb{N} \to \mathbb{R}.$$

If we denote the values of u in $n \in \mathbb{N}$ by u_n then we can also think of $u \in H = l^{(2)}$ as a sequence $u_1, u_2, \ldots$. We can put these in a column vector

$$u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{pmatrix}$$

and then every such vector has length given by

$$|u| = \sqrt{u \cdot u} = \sqrt{u_1^2 + u_2^2 + \cdots},$$

defined via the inner product

$$u \cdot v = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \end{pmatrix} = u_1 v_1 + u_2 v_2 + u_3 v_3 + \cdots = \sum_{k=1}^{\infty} u_k v_k = (u, v)_{H}.$$

This inner product is the integral of the product function uv with respect to the counting measure on $\mathbb{N}$. Note that in general uv is not[1] in $l^{(2)} = L^2(\mathbb{N})$.

Exercise 24.1. Give a direct proof of Theorem 6.31 for $H = l^{(2)}$. Hint: take a fixed $\phi \in H^*$ and determine what the representing u should be.

Remark 24.2. *Every separable Hilbert space has an orthonormal basis via the Gramm-Schmidt procedure, and is therefore isometrically linearly isomorphic with $H = l^{(2)}$. Thus for separable Hilbert spaces Theorem 6.31 is immediate from Exericise 24.1.*

[1] Many function spaces are not algebra's and this is one of them.

24.1 Other standard Hilbert spaces

The example starts from the observation that there are other measures on $\mathbb{N}$: every sequence of positive numbers

$$0 < \lambda_1 \leq \lambda_2 \leq \lambda_2 \leq \cdots \tag{24.1}$$

definines a measure on $\mathbb{N}$ by assigning measure λ_n to the singleton $\{n\}$. The corresponding integral of the product of two functions $u, v : \mathbb{N} \to \mathbb{R}$ is

$$((u, v)) = (u, v)_V = \sum_{n=1}^{\infty} \lambda_n u_n v_n,$$

defined on a subspace V of our standard space $H = l^{(2)}$. This subspace is not closed in H if $\lambda_n \to \infty$.

Exercise 24.3. Why not? Assume that (24.1) holds. Show that V with $((\cdot, \cdot))$ is a Hilbert space, and that $V = H$ if and only if λ_n is a bounded sequence.

So $V \subset H$, and the norm on V is given by

$$|u|_V = ||u|| = \sqrt{\sum_{n=1}^{\infty} \lambda_n u_n^2},$$

and

$$||u||^2 \geq \lambda_1 |u|^2$$

for all $u \in V$. Here $|u|$ is the standard norm of u. The map

$$i : u \in V \to u \in H$$

is linear and continuous. For all $u \in V \subset H$ we have

$$\underbrace{|i(u)|}_{u \in V} = \underbrace{|u|}_{u \in H} \leq \frac{1}{\sqrt{\lambda_1}} \underbrace{||u||}_{u \in V},$$

in which we think of u in V as lying in both V and H. It follows that

$$|i| = \frac{1}{\sqrt{\lambda_1}}$$

is the norm of i in $L(V, H)$, see Theorem 6.11. There is no smaller constant L for which the bound $|u| \leq L||u||$ holds.

Exercise 24.4. Check that $\overline{i(V)} = \overline{V} = H$.

24.2 Double dealing with Riesz

We say that V is dense in H because $\overline{V} = H$. By the Riesz representation Theorem 6.31 every continuous linear function[2] $\phi : H \to \mathbb{R}$ is of the form

$$\phi(v) = (f, v)$$

with $f = R_H(\phi)$, and of course $\phi(v) = (f, v)$ is also defined for $v \in V$. The map

$$\phi \circ i : v \in V \xrightarrow{i} v \in H \xrightarrow{\phi} (f, v) \in \mathbb{R}$$

is thus continuous and linear, and represented by $u = R_V(\phi \circ i) \in V$. It follows that

$$\phi(v) = (f, v) = ((u, v)) \quad \forall v \in V.$$

The linear continuous functions

$$V \ni u \xrightarrow{f \in H} (f, u)_H \in \mathbb{R}$$

and

$$V \ni u \xrightarrow{u \in V} (u, u)_V \in \mathbb{R}$$

are exactly the same, but given by different (Riesz) representations: we have two different vectors u and f representing the same map via two different inner products.

Exercise 24.5. Assume $0 < \lambda_1 \leq \lambda_2, \ldots$ is unbounded. Why is not every continuous linear $\psi : V \to \mathbb{R}$ of the form $\phi(u) = (f, u)$ with $f \in H$?

It is easier[3] for a linear function on V to be continuous with respect to the norm on V than with respect to the norm on H: there are more continuous lineair functions on V then just the functions

$$v \in V \to (f, v)_H \in \mathbb{R}.$$

If we choose to identify H^* with H via R_H, then

$$V \subsetneq H = H^* \subsetneq V^*,$$

which conflicts with an identification of V^* and V via R_V.

[2] We do not yet use the notation in Remark 6.7 here.

[3] If $\lambda_n \to \infty$.

Nevertheless

$$H \ni f \xrightarrow{R_H^{-1}} \underbrace{\phi \in H^* \xrightarrow{i^*} \phi \circ i \in V^*}_{i^*(\phi)=\phi\circ i} \xrightarrow{R_V} u \in V,$$

is linear and continuous, because the first and third link in this chain are both isometries, and the second link, which is called the adjoint i^* of i, is continuous.

Exercise 24.6. Prove that $i^* : H^* \to V^*$ is linear and continuous. Hint: consider the norm of $i^*(\phi) = \phi \circ i$.

For V and H Hilbert spaces with $i : V \to H$ an injective, continuous linear map with $i(V) \subsetneq \overline{i(V)} = H$ the story is the same.

24.3 A more general abstract perspective

We do not need[4] to assume that $V \subset H$. It's instructive to see how the injectivity of $i : V \to H$ and the density of its range being dense come into play.

Exercise 24.7. Assume H and V are Hilbert spaces and that $i : V \to H$ is linear and continuous. Prove that $S : H \to V$ defined by $f \in H \to u = Sf \in V$ and

$$(u, v)_V = (f, i(v))_H$$

for all $v \in H$ is given by

$$S = R_V \circ i^* \circ (R_H)^{-1},$$

and has norm $|S| = |i^*|$.

Remark 24.8. *We think of S as a solution operator. See (25.10) for the more general case in which the inner product on V is replaced by a (in general) nonsymmetric coercive bilinear form. The Lax-Milgram theorem then replaces the Riesz representation theorem. In Section 11.7 we discuss why this approach still requires a Hilbert space setting. In Section 24.4 we consider the operator S in Exercise 24.7 as a solution operator as map from H to H and as a map from V to V. Look very carefully at the four quotients in Exercise 24.14 and how they are used in Exercise 24.16. They all relate to the solution operator, but one of them does not need the solution operator.*

[4] Note that in our applications to elliptic boundary values problems we do have $V \subset H$.

Exercise 24.9. (continued) Show that

$$|i^*|_{L(H^*,V^*)} = |i|_{L(V,H)}.$$

Hint: the notation in Remark 6.7 is handier now. We have that i^* is defined by $i^*(\phi) = \phi \circ i$ for every $\phi \in H^*$. This means that

$$\langle i^*(\phi), v\rangle = \langle \phi, i(v)\rangle \tag{24.2}$$

for every $v \in V$ and every $\phi \in H^*$. In case we identify H and H^* this reads

$$\langle i^*(\phi), v\rangle = (\phi, i(v))_H. \tag{24.3}$$

Now

$$|\langle i^*(\phi), v\rangle| = |\langle \phi, i(v)\rangle| \leq |\phi|_{H^*} |i(v)|_H \leq |\phi|_{H^*} |i|_{L(V,H)} |v|_V$$

means that

$$|\langle i^*(\phi)|_{V^*} \leq |\phi|_{H^*} |i|_{L(V,H)},$$

which in turn means that

$$|i^*|_{L(H^*,V^*)} \leq |i|_{L(V,H)}.$$

To bound $|i^*|$ from below take suitable choices of $\phi \in H^*$ and $v \in V$ with $|\phi|_{H^*} = 1$ and $|v|_V = 1$ in the chain

$$|i|_{L(V,H)} \geq |\langle i^*(\phi)|_{V^*} \geq |\langle i^*(\phi), v\rangle| = |\langle \phi, i(v)\rangle|.$$

To wit, take a sequence $v_n \in V$ with $|v_n|_V = 1$ and $|i(v_n)|_H \to |i|$, and then[5] $\phi_n \in H^*$ 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 24.10. Prove that S is injective if $\overline{i(V)} = H$. Hint: this concerns the second equivalence in S injective $\iff$ i^* injective $\iff$ $\overline{i(V)} = H$. Hint: use (24.2) to characterise the null space of i^* in H^*. We have $i^*(\phi) = 0$ if and only if

$$\langle i^*(\phi), v\rangle = \langle \phi, i(v)\rangle$$

for all $v \in V$.

[5] This is really the Hahn-Banach property, see Exercise 6.33.

Exercise 24.11. Assume H and V Hilbert spaces, $i : V \to H$ linear and continuous. Let $S : H \to V$ be given via Exercise 24.7 and $f \in H \to u = Sf \in V$ with

$$(u, v)_V = (f, i(v))_H$$

for all $v \in V$. Show that

$$N(i) = \{v \in V : i(v) = 0\} = S(H)^{\perp} = \{v \in V : (u, v)_V = 0 \text{ for all } u \in S(H)\}.$$

Thus the range of S is dense in V if and only if i is injective. Hint: use that $i(v) = 0$ in H if and only if $(f, i(v))_H = 0$ for all $f \in H$.

24.4 The operator remains the same?

Exercise 24.12. Assume $i : V \to H$ linear and continuous. Prove that

$$S_0 = i \circ S : H \to H$$

is symmetric, i.e.

$$(S_0 f_1, f_2)_H = (f_1, S_0 f_2)_H$$

for all $f_1, f_2 \in H$, and

$$(S_0 f, f)_H = |Sf|_V^2 .$$

Exercise 24.13. Assume $i : V \to H$ linear and continuous. Prove that

$$S_1 = S \circ i : V \to V$$

is symmetric, i.e.

$$(S_1 u_1, u_2)_V = (u_1, S_1 u_2)_V$$

for all $u_1, u_2 \in V$, and

$$(S_1 u, u)_V = |i(u)|_H^2 .$$

Exercise 24.14. Show that

$$\frac{(S_0 f, f)_H}{(f,f)_H} = \frac{(Sf, Sf)_V}{(f,f)_H} \quad \text{and} \quad \frac{(i(u), i(u))_H}{(u,u)_V} = \frac{(S_1 u, u)_V}{(u,u)_V}.$$

At the end of Section 18.5, see also (18.28), 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 via Theorem 6.11. 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 24.7 and 24.9.

Exercise 24.15. (continued) If the first supremum is a maximum then its maximizer ϕ is an eigenvector with eigenvalue $\lambda = |S_0|_{L(H,H)}$. You should give a direct proof of this, but see Remark 18.4. Same statement for S_1 and the second supremum of course.

Exercise 24.16. Any eigenvector ϕ of S_0 makes for an eigenvector $\psi = S\phi$ of S_1 with the same eigenvalue, unless $S\phi = 0$. Likewise, any eigenvector ψ of S_1 makes for an eigenvector $\phi = i(\psi)$ of S_0 with the same eigenvalue, unless $i(\psi) = 0$. Show that if one of the suprema in Exercise 24.14 for the norm of S_0 is a maximum, then so is the supremum for the norm of S_1 and vice versa.

Remark 24.17. *Each linear, injective, continuous*[6] *compact*

$$i : V \to H \quad \text{with} \quad \overline{i(V)} = H$$

defines via Exercises 24.7, 24.12 and 24.13 two strictly positive definite symmetric compact linear mappings $S_0 : H \to H$ and $S_1 : V \to V$ with the same eigenvalues, by dropping either the first or the last link in

$$V \xrightarrow{i} H \xrightarrow{(R_H)^{-1}} H^* \xrightarrow{i^*} V^* \xrightarrow{R_V} V \xrightarrow{i} H.$$

The triple

$$V \subset H = H^* \subset V^*$$

with V and H Hilbert spaces, $i : V \to H$ injective and $V = i(V)$ dense in H is the standard framework in the French PDE school, see the Brézis book on functional analysis.

[6] Follows from compactness of i.

24.5 Why?

All this relates to the problem stated in (9.18), (21.6) and Chapter 25. Note that the third quotient in Exercise 24.14 is the formula that will be used in the application to eigenvalue problems in Chapter 25. Our Hilbert spaces will be Sobolev spaces which come with a k and $p = 2$, see Chapter 23, which begins with $k = 1$ and general $1 \leq p < \infty$.

25 Evans' Chapter 6 and Navier-Stokes

This chapter is still under (rearranging) construction. The two large exercises in Section 25.4 relate to the Navier-Stokes equations. Problem (**N**) is like the first exercise to Exercise 6.6.4 in Evans. Exercises 3,4,5,6 in his Section 6.6 are variants on the general theme in Section 6.2. Note that 4 requires Theorem 1 in Section 5.8.1 of Evans. Each of these 4 exercises has $B(u,v)$ symmetric and the solution operator S compact and symmetric with respect to both the L^2-inner product, and an inner product defined by B which replaces the inner product with double brackets on V in Section 24.2. The space V depends on the problem. It may be $H_0^1(U)$, $H^1(U)$, or some other Sobolev-Hilbert space. Evans uses H in his formulation of the Lax-Milgram theorem. The space with which it is applied corresponds to V in Remark 24.17.

If S is the inverse of L, the formula's for the eigenvalues of L follow as in Theorem 2 of Section 6.5.1, see Exercise 24.14 and thereafter. Evaluate these eigenvalue formula's for the problems in Exercises 3,4,5,6. NB if numbers are not clickable they refer to Evans.

25.1 Existence of weak solutions via Lax-Milgram

Consider first the equation

$$Lu = -(a_{ij}u_{x_i})_{x_j} + cu = f \quad \text{with boundary condition} \quad u = 0 \tag{25.4}$$

for $u = u(x)$, $x \in U$, U a bounded domain in $\mathbb{R}^N$, ∂U at least continuous, i.e. locally the graph of a continuous function.

Compared to (1) in Section 6.1, I drop the summation signs, use subscripts for the coefficients, and omit the first order terms. Existence of classical solutions, i.e. solutions u with $u \in C^2(U)$ to have equation (25.4) make sense in U, and $u \in C(\bar{U})$ to have the homogeneous (Dirichlet) boundary condition $u = 0$ have a meaning, requires conditions on the coefficients $a_{ij} = a_{ij}(x)$ and $c = c(x)$, and on the right hand side f.

25.1.1 Weak solutions

In the weak solution approach we multiply (25.4) by a $v \in C^1(\bar{U})$, integrate over U and use integration by parts to rewrite the terms with a_{ij} as

$$-\int_U \underbrace{(a_{ij}u_{x_i})_{x_j}}_{w_{x_j}} v = -\int_{\partial U} \nu_j \underbrace{a_{ij}u_{x_i}}_{w} v + \int_U \underbrace{a_{ij}u_{x_i}}_{w} v_{x_j}, \tag{25.5}$$

in which ν_j is the j^{th} component of the outward normal on ∂U. This requires ∂U to be piecewise C^1, the coefficients $a_{ij} \in C^2(\bar{U})$, $c \in C(\bar{U})$, the right hand side $f \in C(\bar{U})$, and $u \in C^2(\bar{U})$. The boundary integral disappears if $v = 0$ on ∂U, leading to the identity

$$\underbrace{\int_U a_{ij} u_{x_i} v_{x_i} + \int_U cuv}_{B[u,v]} = \underbrace{\int_U fv}_{(f,v)} \tag{25.6}$$

for all $v \in C^1(\bar{U})$ with $v = 0$ on ∂U.

This identity that make sense for $u \in C^1(\bar{U})$, with u still required do have $u = 0$ on ∂U. The assumptions of a_{ij}, c, f can of course be weakened now. The weak solution approach works with weak solutions which have their first order derivatives in $L^2(U)$, the natural (Hilbert) space for u, v to live is $H_0^1(U)$ and $u \in H_0^1(U)$ is called a weak solution if (25.6) holds for all $v \in H_0^1(U)$.

This formulation requires $a_{ij}, c \in L^\infty(U)$ only, and $f \in L^2(U)$ then suffices for the right hand side of (25.6) to makes sense because $H_0^1(U) \subset L^2(U)$. Recall that we defined $H_0^1(U)$ as the closure of $C_c^1(U)$ in $H^1(U) = W^{1,2}(U)$. The right hand side of (25.6) is equal to the inner product of f and v in $L^2(U)$ and it defines a linear functional

$$v \in H_0^1(U) \xrightarrow{F} \int_U fv = (f,v)_{L^2(U)} = \underbrace{F(v) = \langle F, v \rangle}_{\substack{\text{different notations} \\ \text{for same functional } F}}, \tag{25.7}$$

the latter being the notation used in the Lax-Milgram Theorem in Section 6.2.1.

In (25.7) it is tempting to write f for F in $\langle F, v \rangle$, as it is really f that acts on v, but *not* via the $H^1(U)$ inner product, as the $H^1(U)$ inner product is defined by the left hand side of (25.6) with $a_{ij} = \delta_{ij}$ and $c = 1$, i.e.

$$(u,v)_{H^1(U)} = \underbrace{\int_U u_{x_i} v_{x_i}}_{\substack{\text{highest order} \\ \text{terms}}} + \int_U uv = \int_U \nabla u \cdot \nabla v + \int_U uv, \tag{25.8}$$

which is the bilinear form corresponding to the partial differential equation $\Delta u + u = f$.

25.1.2 The Lax-Milgram Theorem

The Lax-Milgram Theorem has already been done in Section 11.7. The symmetric case is also discussed in Section 24.2. The f in Theorem 1 in

Section 6.2.1 is really the F in (25.7) if the theorem is applied to the bilinear form in (25.6) with $V = H_0^1(U)$. The H below corresponds to V in Section 24.4.

If the bilinear form is bounded on H, i.e.

$$\forall u, v \in H \quad |B[u,v]| \leq \alpha |u|\,|v|,$$

them for each $u \in H$ the map

$$v \in H \xrightarrow{Au} B[Au, v] = \underbrace{(Au)(v) = \langle Au, v\rangle}_{\substack{\text{different notations} \\ \text{for same functional } Au}} \tag{25.9}$$

is linear and bounded, since

$$|\langle Au, v\rangle| = |B[u,v]| \leq \alpha\,|u|\,|v|,$$

implying that

$$|Au| \leq \alpha\,|u|$$

for all $u \in H$. Thus $A : H \to H^*$, where

$$H^* = \{f : H \to \mathbb{R} :\ f \text{ is linear and bounded}\}$$

normed by

$$|f| = \sup_{0 \neq v \in H} \frac{|\langle f, v\rangle|}{|v|},$$

is linear and bounded. This dual space H^* of H can be identified with H via the Riesz Representation Theorem and $\langle f, v\rangle = f \cdot v = (f, v)_H$, considering $f \in H = H^*$, but in the application to $H = H_0^1(U)$ this is not the inner product in the right hand side to (25.6).

If the bilinear form is also coercive on H, i.e.

$$\forall u \in H \quad B[u,u] \geq \beta\,|u|^2,$$

then

$$\beta\,|u|^2 \leq B[u,u] = \langle Au, u\rangle \leq |Au|\,|u| \quad \text{whence} \quad |Au| \geq \beta\,|u|$$

for all $u \in H$ and it follows that A is a bijection between H and $A(H)$, a subspace of H^*, bounded in both directions. Thus $A(H)$ is complete, and thereby a closed subspace of H^* which[7] coincides with H^*. The (linear) solution operator[8] is then defined by

$$F \in H^* \xrightarrow{S} u \in H \quad \text{defined by} \quad B[u,v] = \langle F, v\rangle \quad \text{for all} \quad v \in H \tag{25.10}$$

[7] Via the Riesz Representation or the Hahn-Banach Theorem and the reflexivity of H.

[8] In Section 24.4 I distinguised between S, S_0, S_1. Check which S this is about!

and has the property that

$$|u|_H = |SF|_H \leq \frac{1}{\beta}\,|F|_{H^*}$$

In the application to boundary value problems any right hand side of (25.4) that defines an F in the dual of the Sobolev space used is allowed. In the case of $H_0^1(U)$ this dual space is denoted by $H^{-1}(U)$ and may be viewed as the space consisting of functions in L^2 as well as their first order distributional derivatives, see Section 5.9.1 in Evans.

25.1.3 Lax-Milgram; boundedness condition

It is usually easy to show that the bilinear form derived from the boundary value problem formulation via integration by parts is bounded, also for other boundary conditions, such as the Neumann condition

$$\nu_j a_{ij} u_{x_i} = 0 \quad \text{on} \quad \partial U, \tag{25.11}$$

which is a special case of the Robin boundary condition

$$\nu_j a_{ij} u_{x_i} + bu = 0 \quad \text{on} \quad \partial U, \tag{25.12}$$

in which $b = b(x)$ is assumed to bounded and integrable for instance. See Exercises 6.6.4 and 6.6.5. The latter condition is called Newton's cooling law in the case that a_{ij} is a positive multiple of the identity matrix and u is the temperature in a body U with heat exchange at the boundary[9]. In (25.5) this condition gives the additional term

$$\int_{\partial U} buv$$

which should now be included in the left hand side of (25.6), and the natural Sobolev space to pose

$$\underbrace{\int_U a_{ij} u_{x_i} v_{x_i}}_{\text{highest order terms}} + \int_U cuv + \int_{\partial U} buv = \underbrace{\int_U fv}_{(f,v)} \tag{25.13}$$

in is now $H^1(U)$.

In the case of the Neumann condition (25.11) this extra term is not there and the only difference with the Dirichlet problem is the choice of the Sobolev

[9] This physical context forces the exchange coefficient to be positive.

space. Boundary conditions which are used in the integration by parts derivation of the weak formulation are sometimes called natural boundary conditions. The Dirichlet boundary condition is not such a natural boundary condition, it has to be forced on the solution by the choice of the smaller Sobolev space $H_0^1(U)$.

25.1.4 Lax-Milgram; coercivity

It is usually more delicate to show the coercivity of the bilinear form. The basic (ellipticity) assumption on the coefficients a_{ij} is (4) in Section 6.1.1 of Evans. With $v = u$ it bounds the highest order terms from below by the highest order terms in (25.8). In the case of $H_0^1(U)$ the Poincaré inequality

$$\int_U u^2 \leq C_{_U} \int_U |\nabla u|^2 \tag{25.14}$$

helps. In particular the bilinear form

$$B[u, v] = \int_U \nabla u \cdot \nabla v$$

used for solving

$$-\Delta u = f \quad \text{with boundary condition} \quad u = 0$$

is coercive on $H_0^1(U)$ considered as a subspace of $H^1(U)$ with the norm derived from (25.8).

The Neumann problem for $-\Delta u = f$ is very instructive. It requires a condition on f for solvability, as well as the same condition on u to have a unique condition, choosing

$$\tilde{H}^1(U) = \{u \in H^1(U) : \int_U u = 0\}$$

as the Sobolev space to be used in the weak formulation.

You should compare the role of b in the Robin boundary condition to that of c in the partial differential equation, as should be clear from (25.13). Coercivity requires some positivity. **It's easy to cook up exam questions on this theme.**

Also, the higher order problem for the bi-Laplacian in Exercise 6.6.3 is only one of the problems of this type. It has two "unnatural" boundary conditions, which are forced upon the solution by the choice of $H_0^2(U)$. Can you think of natural boundary conditions that lead to a formulation in $H^2(U)$, or a mix of natural and unnatural boundary conditions that require $H^2(U) \cap H_0^1(U)$ as the space to be used? Note that for the coercivity of the bilinear form you need the regularity theory in Section 6.3.

25.1.5 The general case with first order terms

The treatment of the Dirichlet problem in Section 6.2.2 should be easy to follow after the discussion above. The main issue is how to deal with the terms in $B[u, v]$ that come from the first order derivatives in the Lu. I did not discuss Section 6.2.3 with the adjoint operator and the Fredholm alternative but read Theorem 4. It is proved via an application of the Fredholm alternative to the solution operator S_μ for the bilinear form

$$B_\mu[u, v] = B[u, v] + \gamma \int_U uv$$

with γ chosen to make B_m coercive.

25.2 The selfadjoint case

See again Chapter 24. The first order terms in L typically prevent the bilinear form from being symmetric. Without these first order terms the symmetry of a_{ij} makes the bilinear form symmetric. This symmetry is usually assumed, see the opening statements in Section 6.5.1. In the case that $B[u, v]$ is a symmetric bounded coercive bilinear form, it defines an equivalent norm on the Sobolev space (used in in the weak formulation) via

$$|u| = \sqrt{B[u, u]}.$$

The solution operator

$$f \xrightarrow{S} u$$

then satisfies both

$$(Sf, g)_{L^2(U)} = (f, Sg)_{L^2(U)} \quad \text{and} \quad B(Su, v) = B(u, Sv)$$

as you should satisfy, and it is compact form $L^2(U)$ to $L^2(U)$ as well as from the Sobolev space to itself. The eigenvalue formula's I discussed for the solution operator using $B[u, v]$ then lead to eigenvalue formula's of which the first is stated in the remark following Theorem 2 in Section 6.5.1.

25.2.1 Second hand in homework set

I would have restricted a second homework set to Exercises like 6.6.3, 6.6.4, 6.6.5, 6.6.6 in the second edition of Evans. In all exercises also write down the eigenvalue formula for the first eigenvalue when f is replaced by λu.

25.2.2 Maximum principles

Evans Section 6.4. More on those principles in Chapters 5 and 10 in

`http://www.few.vu.nl/~jhulshof/NOTES/ellpar.pdf`

25.3 The Navier-Stokes equations

You can read about these equations in Dutch on a very introductionary and informal level in

`http://www.math.vu.nl/~jhulshof/handoutNS.pdf`

Consider the Navier-Stokes equations on a bounded domain $\Omega \subset \mathbb{R}^2$ for $t \geq 0$ with smooth boundary $\partial\Omega$, given initial data for the velocity

$$u = \begin{pmatrix} u_1(t, x_1, x_2) \\ u_2(t, x_1, x_2) \end{pmatrix}$$

at $t = 0$ and no-slip boundary conditions $u = 0$ on $\partial\Omega$ for all $t \geq 0$. For the exercises below you may restrict your attention to the case that

$$\Omega = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 < 1\} \quad \text{with outer normal} \quad n = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{in} \quad x \in \partial\Omega.$$

The Navier-Stokes equations read (with kinematic viscosity equal to unity)

$$u_t + (u \cdot \nabla)u + \nabla p = \Delta u, \quad \nabla \cdot u = 0.$$

The second zero divergence equation has to be imposed on the initial data for u at $t = 0$ as well. In view of the Laplacian in the equation and the boundary condition $u = 0$ on $\partial\Omega$ the natural spaces for solutions to live in as functions of t are

$$H_0^1(\Omega)^2 = \{ \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} : \, u_1, u_2 \in H_0^1(\Omega) \} \subset (L^2(\Omega))^2 = \{ \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} : \, u_1, u_2 \in L^2(\Omega) \},$$

but the zero divergence equation imposes an a priori restriction as explained next.

If $u \in (L^2(\Omega))^2$ satisfies $\nabla \cdot u \in L^2(\Omega)$ then the normal component $n \cdot u$ of the velocity is well defined in $L^2(\partial\Omega)$ by a theorem similar to the trace theorems in Evans, and the Gauss divergence formula

$$\int_\Omega \nabla \cdot u = \int_{\partial\Omega} n \cdot u$$

holds true for such u. Solutions with finite kinetic energy

$$E(u) = \frac{1}{2}\int_\Omega (u_1^2 + u_2^2)$$

actually live in

$$H = \{u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} : \, u_1, u_2 \in L^2(\Omega), \, \nabla \cdot u = 0 \text{ on } \Omega, \, n \cdot u = 0 \text{ on } \partial\Omega\}.$$

If also the first order spatial weak derivatives exist with

$$\mathcal{E}(u) = \int_\Omega |Du|^2 = \int_\Omega \left((\frac{\partial u_1}{\partial x_1})^2 + (\frac{\partial u_1}{\partial x_2})^2 + (\frac{\partial u_2}{\partial x_1})^2 + (\frac{\partial u_2}{\partial x_2})^2 \right) < \infty,$$

then $u \in H^1(\Omega)^2$ and it is possible to speak of u on $\partial\Omega$ as the trace of u and in particular of its tangential component $n \times u = n_1 u_2 - n_2 u_1$ in the usual sense.

25.4 Navier-Stokes related exercises

1. This exercise concerns the projection of

$$L^2_{div}(\Omega) = \{w \in (L^2(\Omega))^2 : \nabla \cdot w \in L^2(\Omega)\}$$

on the space H above (the subscript *div* stands for divergence). For $w \in L^2_{div}(\Omega)$ let $f = -\nabla \cdot w \in L^2(\Omega)$ and $g = n \cdot w \in L^2(\partial\Omega)$, and consider the Neumann problem

$$(\mathbf{N}) \qquad -\Delta p = f \quad \text{in} \quad \Omega \quad \text{with} \quad \frac{\partial p}{\partial n} = g \quad \text{on} \quad \partial\Omega.$$

You may think of p in (**N**) as related to the pressure in the Navier-Stokes equations.

   (a) What is the natural condition on arbitrary $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ to have a solution of (**N**)? Hint: use the divergence theorem, you may argue as if f, g and p are smooth. Does your condition hold for the particular choice of f and g above? If so, why? Explain why then the solution p is never unique and can be chosen to have $\int_\Omega p = 0$.

   (b) Explain why for $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ we say that $p \in H^1(\Omega)$ is a weak solution of (**N**) if

$$(\mathbf{N_{weak}}) \qquad \int_\Omega \nabla p \cdot \nabla \phi = \int_\Omega f\phi + \int_{\partial\Omega} g\phi \quad \text{for all} \quad \phi \in H^1(\Omega).$$

Check that this can only hold if your condition in (a) is satisfied, in which case it suffices to show that that the identity in ($\mathbf{N_{weak}}$) holds for every $\phi \in \tilde{H}^1(\Omega) = \{p \in H^1(\Omega) : \int_\Omega p = 0\}$.

(c) Let $\tilde{H}^1(\Omega)$ be as in (b). Show that

$$((p, \phi)) = \int_\Omega \nabla p \cdot \nabla \phi$$

defines an inner product on $\tilde{H}^1(\Omega)$ with an inner product norm that is equivalent on $\tilde{H}^1(\Omega)$ to the full H^1-norm defined by

$$(p, \phi)_{H^1(\Omega)} = \int_\Omega p\phi + \int_\Omega \nabla p \cdot \nabla \phi, \quad |p|^2_{H^1(\Omega)} = (p, p)_{H^1(\Omega)},$$

(d) Explain why for every $f \in L^2(\Omega)$ and every $g \in L^2(\partial\Omega)$ satisfying your condition in (a) there is a unique $p \in \tilde{H}^1(\Omega)$ that satisfies ($\mathbf{N_{weak}}$).

(e) Recall that

$$H = \{u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} : u_1, u_2 \in L^2(\Omega),\ \nabla \cdot u = 0 \text{ on } \Omega,\ n \cdot u = 0 \text{ on } \partial\Omega\}.$$

Explain why every $w \in L^2_{div}(\Omega)$ can be written as $w = \nabla p + u$ with $u \in H$ and $p \in H^1(\Omega)$, and that u is uniquely determined by w. This u is called the Leray projection of w.

2. In this exercise we consider smooth solutions of the Navier-Stokes equations with zero slip boundary conditions as above (so you can forget about weak derivatives and all that now).

(a) Write u_0 for the initial velocity field of a smooth solution u with pressure p: then $u(x, 0) = u_0(x)$ and u_0 must satisfy $\nabla \cdot u_0 = 0$. We write $u(t)$ for the function $x \to u(x, t)$. Integrate the inner product of

$$u_t + (u \cdot \nabla)u + \nabla p - \Delta u$$

with u over Ω and derive that

$$\frac{d}{dt} E(u(t) + \mathcal{E}(u(t)) = 0,$$

where $E(u)$ and $\mathcal{E}(u)$ are as in the introduction above. In other words, show that

$$\frac{1}{2}\frac{d}{dt}\int_\Omega |u|^2 + \int_\Omega |Du|^2 = 0.$$

Why does it follow that

$$\int_0^T \int_\Omega |Du|^2 \leq \frac{1}{2}\int_\Omega |u_0|^2 \,?$$

Hint: write terms out in coordinates, e.g.

$$u \cdot (u \cdot \nabla)u = \sum_{j,k=1}^{2} u_k u_j \frac{\partial u_k}{\partial x_j},$$

and use integration by parts (the boundary terms disappear, as well as $\nabla \cdot u$).

(b) For smooth solutions u and v with pressures p and q respectively let $w = u - v$. Subtract the equations for u and v, take the inner product with w and integrate over Ω to derive that

$$\frac{1}{2}\frac{d}{dt}\int_\Omega |w|^2 + \int_\Omega |Dw|^2 = -\int_\Omega w \cdot (w \cdot \nabla)v \leq \int_\Omega |Dv|\,|w|^2.$$

Hint: (i) use integration by parts for the equality, the boundary terms disappear, as well as $\nabla \cdot u$, $\nabla \cdot v$, $\nabla \cdot w$, if they show up. Check that the terms coming from the nonlinear terms in the equations may be rewritten as a term giving the integral with w and v, and another integral with u and w which disappears; (ii) for the subsequent inequality use $Ax \cdot x \leq |A|\,|x|^2$ for 2×2 matrices A and 2-vectors x, with $|A|^2 = A_{11}^2 + A_{12}^2 + A_{21}^2 + A_{22}^2$.

(c) Derive from (b) that

$$\frac{d}{dt}\int_\Omega |w|^2 + 2\int_\Omega |Dw|^2 \leq 2\Big(\int_\Omega |Dv|^2\Big)^{\frac{1}{2}}\Big(\int_\Omega |w|^4\Big)^{\frac{1}{2}}.$$

(d) Insert the inequality

$$\int_\Omega |w|^4 \leq \int_\Omega |w|^2 \int_\Omega |Dw|^2$$

in (c) to derive that

$$\frac{d}{dt}\int_\Omega |w|^2 \leq \frac{1}{2}\int_\Omega |Dv|^2 \int_\Omega |w|^2.$$

Hint: in the right hand side you get a product which contains the factor $a = \int_\Omega |Dw|^2$. Use the inequality $2ab \leq a^2 + b^2$ and observe that a^2 als appears on the left hand side.

(e) Derive from (d) and (a) with u replaced by v that

$$\int_\Omega |w(t)|^2 \leq \int_\Omega |w_0|^2 e^{\frac{1}{4}\int_\Omega |v_0|^2}.$$

(f) Prove the inequality used in (d) for compactly supported smooth vectorfields on $\mathbb{R}^2$ by first showing that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(x_1,x_2)^4\, dx_1 dx_2 \leq$$

$$(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u^2)\,(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u_{x_1}^2)^{\frac{1}{2}}\,(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u_{x_2}^2)^{\frac{1}{2}}$$

for compactly supported smooth functions u. In short

$$|u|_4^4 \leq |u|_2^2\, |u_{x_1}|_2\, |u_{x_2}|_2.$$

Hint: write $u(x_1,x_2)^4 = u(x_1,x_2)^2 u(x_1,x_2)^2$ and show first that

$$u(x_1,x_2)^2 \leq \int_{-\infty}^{\infty} u(\xi,x_2) u_{x_1}(\xi,x_2) d\xi$$

and likewise

$$u(x_1,x_2)^2 \leq \int_{-\infty}^{\infty} u(x_1,\eta) u_{x_2}(x_1,\eta) d\eta.$$

26 A very partial reader for Olver's PDE book

The introductionary chapter contains the basic example of the Navier-Stokes system, which I discussed in Dutch for high school teachers here:

`http://www.few.vu.nl/~jhulshof/handoutNS.pdf`

The system in Olver's (1.4) is the unpacked version of

$$(\mathbf{NS}) \qquad u_t + (u \cdot \nabla)u + \nabla p = \nu \Delta u \quad \text{with} \quad \nabla \cdot u = 0,$$

in which the fluid velocity $u = u(t, x)$ is the unknown 3-vector valued function of x in a domain Ω in 3-dimensional real space, and of time t, and the unknown pressure $p = p(t, x)$ a scalar function. The positive parameter ν is called the kinematic viscosity, which differs from the dynamic viscosity μ by a factor ρ, the density of the fluid, which for water is effectively a constant. With ρ normalised to 1 we have $\mu = \nu$, but people do get annoyed if you write (**NS**) with μ like I did in that handout.

Anyway, with $\rho = 1$ the velocity u is a mass flux, often denoted by $q = \rho v$ if v is the velocity. In (**NS**) the velocity is called u. Many PDE solvers (persons who solve PDE's) prefer u as the unknown function in their PDE's, whatever u is. With constant density the second equation says that

$$\nabla \cdot q = 0,$$

which is a special case of the (mass) conservation law

$$(\mathbf{CL}) \qquad \rho_t + \nabla \cdot q = 0,$$

an equation you should have seen in the context of the treatment of the Gauss Divergence Theorem in your vector calculus course. Such conservation laws may be supplemented with various flux laws, e.g.

$$q = -\nabla \rho,$$

leading to the linear diffusion equation

$$\rho_t = \nabla \cdot (\nabla \rho) = (\nabla \cdot \nabla)\rho = \Delta \rho$$

for $\rho = \rho(t, x)$ with x in n-dimensional space and t a 1-dimensional time variable.

Another, example, in dimension $n = 1$, is

$$F(\rho) = \frac{1}{2}\rho^2 \quad \text{leading to} \quad \rho_t + \rho \rho_x = 0,$$

sometimes called the inviscid Burgers' equation, the viscous Burgers' equation being

$$\rho_t + \rho\rho_x = \nu\rho_{xx}.$$

With ρ replaced by u all these equation are treated in Olver's book (with $\gamma = \nu$ in case of the Burgers' equation). The linear diffusion equation is invariably called the heat equation for temperature $u = T$, derived from a conservation for thermal energy and a heat flux law. Also of interest, but not mentioned in the book is the Porous Medium Equation

$$\rho_t = \Delta\rho^\alpha,$$

first derived in the context of gas flow in porous media using a flux law of the form $q = -\nabla\rho^\alpha$ with $\alpha > 1$ a positive parameter.

With $\rho = u$ the viscous Burgers' equation looks like a 1-dimensional version of (**NS**) but physically this is nonsense of course, even before you observe that the second equation in (**NS**) reduces to $u_x = 0$: (**NS**) a system to avoid in dimension 1. The 2-dimensional case is of much more interest, as is explained in handoutNS.pdf which I recommend you to read. Not explained but implicitly used in handoutNS.pdf is that the solution $u = u(t, x)$, once known, defines a time dependent vector field, and thereby a system of Ordinary Differential Equations (ODE's)

$$\dot{x}(t) = u(t, x(t))$$

for particle trajectories $t \to x(t)$ in the (flow of) the fluid. For any (smooth) function $(t, x) \to F(t, x)$ such a trajectory defines a function

$$t \to F(t, x(t))$$

whose (time) derivative is given by

$$F_t(t, x(t)) + \sum_{i=1}^{n} F_{x_i}(t, x(t))\, \dot{x}_i(t) = F_t(t, x(t)) + \sum_{i=1}^{n} u_i(t)\, F_{x_i}(t, x(t)),$$

i.e. by

$$\frac{\partial}{\partial t} + u \cdot \nabla \quad \text{acting on} \quad F,$$

and $x(t)$ substituted for x in $u(t, x)$ and $F(t, x)$. The quantity

$$\frac{\partial F}{\partial t} + u \cdot \nabla F$$

is called the convective derivative of F given the velocity field $u = u(t, x)$, and with $F = u$ it is the acceleration field of the fluid flow. This explains

the first two terms in (**NS**), which relates the acceleration field to the forces acting on the fluid (in zero gravity): the gradient of the pressure and the viscous (friction) forces.

PDE's can be classified. They can be linear in u with a right hand side independent of u. Such equations are called inhomegenous linear PDE's (homegenous in case of zero right hand side) or nonlinear (if they are not linear). Linear PDE's allow superposition of solutions, much like your linear equations from linear algebra do, as well as linear ODE's.

The order of a PDE refers to the order of the highest order derivative in the PDE.

PDE's come with initial conditions at $t = 0$ for u if they are of the form u_t equals a right hand side without time derivatives of u, much like you have seen this for ODE's, also for higher order (in t) equations. For instance, the wave equation $u_{tt} = u_{xx}$ comes with initial conditions for u and u_t.

If the spatial variables are confined to a domain Ω typically also boundary conditions are needed. The ones that make sense to impose come from physical considerations. Examples: prescribing temperature or density at the boundary, or the normal component of flux (in- or outflow of heat), or the flux in terms of the boundary values of u (Newton's cooling law for instance). These conditions are named after respectively Dirichlet, Neumann and Robin for various reasons you may be interested in to know.

26.1 First order equations

Chapter 2 in Olver's book is about the Cauchy initial value problem for first order PDE's. I will concentrate on 2 very simple PDE's for $u = u(t, x)$:

$$u_t + c(x)u_x = 0 \quad \text{and} \quad u_t + c(u)u_x = 0,$$

for example with $c(x) = x$ and $c(u) = u$. Both are first order evolution equations for the unknown function $u = u(t, x)$. We shall mainly consider the Cauchy initial value problem in which the PDE is to be solved with initial data

$$u(0, x) = f(x)$$

given for all $x \in \mathbb{R}$. Is there a unique solution and what does it do as t varies? Think of a movie in the x, u-plane (which we can also run backwards).

Both PDE's have as a special case the PDE

$$u_t + cu_x = 0$$

in which $c \in \mathbb{R}$ is a constant. If f is a C^1-function (a continuously differentiable function), then clearly a solution of the Cauchy problem for this PDE is given by

$$u(t,x) = f(x - ct).$$

This is a *classical* solution, meaning that $u(t,x)$ and all the derivatives which appear in the equation are jointly continuous in t and x. You should be able to use your calculus skills to show that the *only* classical solution of $u_t + cu_x = 0$ is $u(t,x) = f(x-ct)$. Even when f is not C^1, the solution formula still makes perfect sense, so we may like to consider also non-classical solutions, in particular as limits of classical solutions u_n by taking a sequence of smooth functions f_n with $f_n \to f$. However below I assume that f is C^1.

The above solution is constant along the lines $x - ct = \xi$ where $\xi \in \mathbb{R}$ is an arbitrary constant. These lines are called characteristics along which the initial data propagate and these characteristics are solutions of

$$\dot{x} = \frac{dx}{dt} = c,$$

a simple ODE with solutions $x(t) = ct + \xi$, in which ξ is a constant of integration that coincides with the initial value for $x = x(t)$.

26.1.1 The method of characteristics

All 3 PDE's above are special cases of the PDE

$$u_t + c(x,u)u_x = h(x,u).$$

This PDE can be solved using the method of characteristics. Introduce $x = x(t)$ and use the PDE with $u = u(t) = u(t, x(t))$. You may prefer to be more strict in the notation here and write $X(t), U(t)$ rather than $x(t), u(t)$ if you get confused, but I will write $\dot{x}$ and $\dot{u}$ for the time derivatives of $x(t)$ and $u(t)$ to distinguish between $u(t)$ and $u(x,t)$. The result is that we should have

$$\dot{u} = u_t + u_x\dot{x} = h(x,u) + (-c(x,u) + \dot{x})u_x.$$

If we impose that $\dot{x} = c(x,u)$ we conclude that we must have $\dot{u} = h(x,u)$. An ODE system to study in relation to the PDE is therefore

$$\frac{dx}{dt} = c(x,u), \quad \frac{du}{dt} = h(x,u).$$

Systems of this type are studied in every ODE course so you know that we can solve this system for given initial values of x and u at $t = 0$ if the

coefficients $c(x,u)$ and $h(x,u)$ are C^1. You also know that these systems come with many possible behaviours of solutions. Solutions may run away to infinity in finite time, they may converge to equilibrium, they may spiral, they may become periodic and what have you.

In relation to the Cauchy problem for the PDE we have to consider all solutions of the ODE system which start on the graph of the initial data. Thus, as t evolves we see the graph deform, since each point on the graph starts to move around in the x,u-plane following the flow of the ODE system. As long as the deformed graph is a graph without vertical tangent lines, we should have a unique classical solution of the Cauchy initial value problem for the PDE, but a priori there is no reason why this graph property should be preserved. Indeed, the second example, with $h \equiv 0$ and $c = c(u)$ makes that points move only in the horizontal x-direction, with a velocity that depends on the height u, so pulse/bell type initial data will always lose the graph property and give rise to multivalued solutions.

The first example, with $h \equiv 0$ and $c = c(x)$ does not have this feature, but depending on the solutions $x = x(t)$ (*the characteristics*) of the first order ODE

$$\frac{dx}{dt} = c(x),$$

there is a large variety of behaviours of solutions of the Cauchy problem for the PDE. One can solve this ODE using a primitive of $\frac{1}{c(x)}$ (if it exists) and a constant of integration, in the book denoted by ξ. WARNING: in general this ξ is *not* the initial value of $x(t)$ though sometimes it is. Olver denotes the initial value sometimes by y and sometimes by k, see Fig. 2.6 and below. Observe that a qualitative analysis (which most of you are familiar with, see also Section **??**) of the first order ODE for $x = x(t)$ based on the graph of the function $x \to c(x)$ already tells most of the story on the solution $u(t,x)$!

In this summary I have restricted the attention to autonomous PDE's: the time variable t does not occur explicitly in the examples above. Thus it is no restriction to assume that the initial data for $u(t,x)$ are given at $t = 0$. Once we are familiar with the solution methods for the Cauchy initial value problem for such autonomous PDE's we can also think of nonautonomous PDE's and other boundary conditions, say prescribing u on some curve in the t,x-plane, or on both of the positive coordinate axis.

Finally we note that the ODE system above in itself does not distinguish between the roles of x and u. On any point of the solution graph of u we can locally write $u = u(t,x)$ or $x = x(t,u)$. The PDE for $u(t,x)$ transforms to a PDE for $x = x(t,u)$. Below are some additional considerations in which I use this trick, with a more careful notation, to see what one can and should do if one insists on looking at the second PDE, e.g. with $c(u) = u$, as a (physical)

conversation law. This leads to classical solutions developing shocks which propogate obeying a so-called Rankine-Hugoniot condition. Such shocks have the property that in forward time our happy characteristics can only go into a shock and never come out of a shock. This property then also appears as an imposed *entropy* condition for the construction of a unique *entropy* solution starting from piecewise continuous initial data.

26.1.2 Shocks, mass conservation, Rankine-Hugoniot condition

Refering to Olver's book again consider Figure 2.16 where a solution of (2.31/2.48) constructed using the method of characteristics has become multivalued. Choose $x = a$ to the left and $x = b$ to the right of the multivalued part. The first point to make is that the integral

$$M(t) = M_{a,b}(t) = \int_a^b u(x,t)dx,$$

redefined in the appropriate sense after the solution has become multi-valued, still satisfies

$$\frac{dM}{dt} = \frac{1}{2}u(t,a)^2 - \frac{1}{2}u(t,b)^2.$$

Redefine M as the area of the region bounded by the solution curve and the lines $x = a$, $x = b$ and $u = 0$. We can write M as an integral with respect to u by inverting:

$$u = U(t,x) \quad \Leftrightarrow \quad x = X(t,u).$$

Momentarily I write upper cases for the unknown (possibly multivalued) solutions.

Exercise JH1. Check directly that

$$U_t + UU_x = 0 \quad \Leftrightarrow \quad X_t = u.$$

You already knew this from the method of characteristics of course.

The equation on the right is an easy PDE for $X(t,u)$, with initial data given by the possibly multivalued inverse function g of f as it appears in:

$$U(0,x) = f(x) \quad \Leftrightarrow \quad X(0,u) = g(u).$$

The solution is trivially given by

$$X(t,u) = g(u) + tu,$$

so it is easy to see that the multivaluedness of $g(u)$ is inherited by $X(t,u)$, and that different laps of the graph $u = U(t,x)$ do not overtake one another.

Below we reason from the assumption that $f(x)$ is positive, decreasing and continuous.

The redefined $M(t) = M_{a,b}(t)$ is given by

$$M = \int_{U(t,b)}^{U(t,a)} (X(t,u) - a)du + (b-a)U(b,t),$$

from which one easily recovers the above ODE for M by differentiating with respect to t. We then ask for a shock solution as indicated in Fig. 2.16 by writing an equation for the location of the vertical segment $x = \sigma(t)$ and asking the two areas to be the same. Why? Because such a shock solution has the same $M(t)$ as the multivalued solution. The equal area condition gives an equation involving the function $g(u)$, which supplements the two equations for the upper and lower values $u^-(t)$ and $u^+(t)$ at the shock:

$$g(u^-(t)) + tu^-(t) = g(u^+(t)) + tu^+(t) = \sigma(t).$$

Exercise JH2. Derive this third equation and verify (2.53) by applying the implicit function theorem. Then do the same for (2.64) and (2.59), replacing u by $c(u)$ and $\frac{u^2}{2}$ by a primitive $C(u)$ of $c(u)$.

Now the equation you found for $\dot{\sigma}(t)$ holds after the last moment $t = t_*$ that

$$x = g(u) + tc(u)$$

is invertible as $u = u(t,x)$, that is, as $u \to g(u) + tc(u)$ is no longer strictly monotone in u. Assuming that both $c(u)$ and $c'(u)$ are positive for the range of u-values given by $u = f(x)$ with x ranging over the real line, and $g'(u) < 0$ for those values, $t = t_*$ is characterised as the first t for which $g'(u)+tc'(u) = 0$ has a solution, say $u = u_*$, and $x = \sigma_* = g(u_*) + t_*c(u_*)$ in the point on the real line where the shock appears. If you understand the Implicit Function Theorem[1] here is a neat application for it:

26.1.3 Appearance of the shock

For $t > t_*$ but close to t_* there should be a small range of σ near σ^* for which $g(u) + tc(u) = \sigma$ has 3 solutions u in such a way that the equal area rule is satisfied for one particular σ in this small range. We write $t = t_* + \epsilon^2$ with ϵ small and try to solve for the 3 solutions u first. The reason to put $t = t_* + \epsilon^2$ and not $t = t_* + \epsilon$ is because the latter will lead to a sharp distinction between positive and negative ϵ, which makes the Implicit Function Theorem unlikely

[1] See Section 13.1.

to be applicable in $\epsilon = 0$. Since at $t = t_*$ the level u_* propagates with speed $c(u_*)$ to the right, the σ-range for the analysis to be carried out will be near $\sigma_* + c(u_*)\epsilon^2$, so it makes sense to put $\sigma = \sigma_* + c(u_*)\epsilon^2 + \nu\epsilon^3$ with ν a new unknown replacing σ. Finally we set $u = u_* + \epsilon v$.

Dividing the resulting equation by ϵ^3 you should get an equation for v in terms of ϵ and ν that replaces $g(u) + tc(u) = \sigma$ for u, and that has 3 distinct solutions for $\epsilon = \nu = 0$ around all 3 of which the Implicit Function Theorem can be applied to obtain 3 solutions $\underline{v} < v_0 < \overline{v}$ as implicit functions of ϵ and ν. The equal area rule to impose then becomes an integral from $\underline{v}$ to $\overline{v}$ which defines a new function to which the Implicit Function Theorem can be applied to find ν as an implicit function of ϵ, and thereby also the values of u at the top and bottom of the shock in terms of an expansion in ϵ.

26.1.4 ODE-system for the shock

From $g(u^-) + tc(u^-) = g(u^+) + tc(u^+) = \sigma$ and the equal area condition it follows, with Exercise JH2, that

$$(u^+ - u^-)\dot{\sigma} = C(u^+) - C(u^-), \quad (g'(u^-) + tc'(u^-))\dot{u}^- + c(u^-) = \dot{\sigma},$$

$$(g'(u^+) + tc'(u^+))\dot{u}^+ + c(u^+) = \dot{\sigma},$$

so that u^+, u^- solve

$$(g'(u^+) + tc'(u^+))\dot{u}^+ = \frac{C(u^+) - C(u^+)}{u^+ - u^-} - c(u^+)$$

$$= \frac{1}{u^+ - u^-}\int_{u^-}^{u^+} (c(s) - c(u^+))ds < 0,$$

$$(g'(u^-) + tc'(u^-))\dot{u}^- = \frac{C(u^+) - C(u^+)}{u^+ - u^-} - c(u^-)$$

$$= \frac{1}{u^+ - u^-}\int_{u^-}^{u^+} (c(s) - c(u^-))ds < 0,$$

if $c(u)$ is increasing in u. It's a very nice ODE exercise to examine how solutions $(u^-(t), u^+(t))$ of the system start from local minima of

$$t = -\frac{g'(u)}{c'(u)}$$

in the (t, u)-plane, and subsequently define $\sigma(t)$ via the shock condition, i.e. the equation for $\dot{\sigma}$. This produces curves

$$t \to (\sigma(t), u^+(t)) \quad \text{and} \quad t \to (\sigma(t), u^+(t))$$

in the (x,u)-plane from which we can trace back the solution of the PDE to $t=0$, thereby recovering the initial data.

For $c(u)=u$ the equations simplify to

$$(g'(u^+)+t)\dot{u}^+ = \frac{u^- - u^+}{2},$$

$$(g'(u^-)+t)\dot{u}^- = \frac{u^+ - u^-}{2}.$$

I did the example where the initial data are

$$f(x)=\frac{1}{1+x^2}, \quad g(u)=\sqrt{\frac{1}{u}-1}, \quad u_*=\frac{3}{4}, \quad x_*=\sqrt{3}, \quad t_*=\frac{8}{3\sqrt{3}},$$

put

$$t=\frac{8}{3\sqrt{3}}(1+s), \quad u^\pm=\frac{3+w^\pm}{4},$$

introduced

$$\phi(w)=(1+\frac{w}{3})^{\frac{3}{2}}(1+w)^{\frac{1}{2}}=\frac{1}{3}w^2+\frac{4}{27}w^3+\cdots$$

and arrived at

$$(s-\phi(w^+))\frac{dw^+}{ds}=\frac{w^- - w^+}{2}, \quad (s-\phi(w^-))\frac{dw^-}{ds}=\frac{w^+ - w^-}{2}.$$

Then I put $s=\tau^2$, $w^\pm=\tau z^\pm$ to get

$$(1-\frac{\phi(\tau z^+)}{\tau^2})(z^+ + \tau\frac{dz^+}{d\tau}) = z^- - z^+,$$

$$(1-\frac{\phi(\tau z^-)}{\tau^2})(z^- + \tau\frac{dz^-}{d\tau}) = z^+ - z^-,$$

in which $\tau=e^t$ (a new t) put this system into a form which to leading order (small $z^\pm$) is linear with a saddle point structure.

26.2 The one-dimensional wave equation

The last section of Chapter 2 concerns the one-dimensional inhomogeneous linear wave equation

$$u_{tt}-c^2u_{xx}=F(x,t)$$

which is solved explicitly in terms of its initial data for u and u_x at $t=0$. In the case that the initial data as well as F are 2π-periodic in x, the main result,

Theorem 2.18, must turn out to be consistent with the Fourier series approach in Chapters 3 and 4. Read and do the derivation of the homogeneous case ($F \equiv 0$) in Theorem 2.15 yourself. It shows how the solutions is constructed form signals to propagate along forward and backward characteristics. Here we will not attempt to take the concept of characteristics much further than that discovered in (2.71) and (2.74) for $u_{tt} - c^2 u_{xx} = 0$.

26.3 Fourier series

Olver's Chapter 3 is of independent interest, see also Hoofdstuk 36. Fourier series are not just important in the context op PDE's, but an introduction from PDE perspective is most appropriate. Section 3.1 rediscusses some of Chapter 1 and then derives (3.27) as the general form for 2π-periodic (in x) solutions $u(t, x)$ of the one-dimensional heat equation with unit diffusion coeffient:

$$u_t = u_{xx}$$

New and very important in relation to Chapter 1 is the eigenvalue perspective comparing (3.1) and (3.14), culminating in the table on page 68 and, in the 2π-periodic context, (3.27). Note that (3.18) is a solution of *separated variables*, the topic of Chapter 4.

The initial data of the solution in (3.27) is $f(x)$ given by (3.28). Section 3.2 explains how the left hand side $f(x)$ and the Fourier series on the right hand side correspond to one another. The starting point is (3.34) and (3.35) with

$$f(x) \sim \frac{a_0}{2} + \sum \cdots$$

This notation is introduced because a priori we do not know that or in what sense

$$f(x) = \frac{a_0}{2} + \sum \cdots,$$

except when f is a trigonometric polynomial, see Exercise 3.2.4.

Note that we interested in both directions. If we compute solutions $u(t, x)$ we find t-dependent coefficients of the solution and ask how well the solution is defined and how smooth it is. If we fit the general solution $u(t, x)$ to initial data f we do this by fitting the coefficients to the coefficients computed from f.

There are two equivalent forms of the Fourier series, real and complex. The real cos/sin form of the Fourier series allows a nice and useful distinction between even and odd functions $f(x)$: Prop. 3.14. The complex form (3.64) allows for smoother formula's and proofs and a much smoother transition

to the Fourier integral transform in Chapter 7. Both forms are intrinsically related to a suitable inner product for functions: (3.30)/(3.61).

One can prove that C^1 2π-periodic functions f are sums of *uniformly convergent* Fourier series computed through (3.35). A stronger localised statement is given in Thm 3.30. The uniform convergence result remains valid if $f'(x)$ has finitely many jump discontinuities (corners in the graph of f) on its interval of periodicity. However, if f itself has jump discontinuities the convergence *cannot be uniform*. It still holds pointwise, but only under the strong assumption that f' is piecewise continuous: Theorem 3.8, proved at the very end of Section 3.5. A slight variant of that proof, using the mean value theorem, shows the convergence is indeed uniform if f is continuous and f' is (piecewise) continuous, see e.g. The Way of Analysis by Strickarzt, or exercise JH4 below.

Remark: jump discontinuities may be removed by subtracting suitably scaled and shifted saw tooth functions. The resulting continuous piecewise C^1-function has a uniformly convergent Fourier series. For saw tooths we can examine the behaviour of their Fourier series by direct calculations which are somewhat reminiscent of the calculations in the convergence proof, and which also clarify the Gibbs overshoot phenomenon illustrated in Figure 3.7. The nicest saw tooth to illustrate what's going on is the odd 2π-periodic extension of

$$\pi - x = \frac{2\sin x}{1} + \frac{2\sin 2x}{2} + \frac{2\sin 3x}{3} + \frac{2\sin 4x}{4} + \frac{2\sin 5x}{5} + \dots \quad (0 < x < \pi).$$

If you differentiate the right hand side you get an expression which diverges! Its truncation after n terms is easily related to the outcome of (3.128) in the convergence proof.

Exercise JH3. Integrate the resulting expression to conclude that

$$\frac{2\sin x}{1} + \dots + \frac{2\sin nx}{n} = \int_0^x D_n(s)ds - x, \quad D_n(s) = \frac{\sin(n+\frac{1}{2})s}{\sin\frac{s}{2}}$$

This D_n is called the Dirichlet kernel. Examine the integral using the n-dependent scaling $y = (n+\frac{1}{2})x$ (and likewise for s). Identify π as a limit of the integral for $n \to \infty$ when $0 < x \leq \pi$ and

$$2\int_0^\pi \frac{\sin t}{t}dt > \pi = 2\int_0^\infty \frac{\sin t}{t}dt$$

as the limit of the largest maximum of $\int_0^x D_n(s)ds$, which occurs in $\frac{\pi}{n+\frac{1}{2}}$.

Basis general facts about uniform convergence are Thm 3.26 and 3.27 (which implies Thm 3.29), Prop. 3.28, and the discussion about interchanging sums and integrals just above Prop. 3.28. I will assume these facts

are known. Not mentioned here in the book is that uniform convergence, illustrated in Fig. 3.11, also comes with a norm, called the maximum or supremum norm,

$$||f||_m = \sup_x |f(x)|,$$

which is certainly defined if f is 2π-periodic and piecewise continuous in the sense of Def. 3.6. Note that the inner product norm is controlled by the maximum norm: if the truncation error goes to zero in the maximum norm, it certainly goes to zero in the inner product norm (but not the other way around!). It is in general not true that the partial sums s_n of the Fourier series of a 2π-periodic continuous function f satisfy

$$||s_n - f||_m \to 0,$$

as it is not even clear that $s_n(x) \to f(x)$ pointwise.

Section 3.5 should be studied in mathematical detail. It discusses not only uniform convergence, and pointwise convergence, but also convergence in the mean. The latter is by definition equivalent to convergence in the inner product norm (3.102) and allows one to think of and work with 2π-periodic functions (for which the inner product norm is defined) as column vectors, and the Fourier modes as unit base vectors. Think of straight angles and the Pythagorean theorem here, we will see expressions like

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots = \frac{\pi^2}{6}.$$

The relevant formula's are given in the complex notation (3.124), the real version is discussed in 3.5.38. The proof relies on the essential observation made in Thm 3.36 that in terms of the inner product norm the n-th partial sum s_n of the Fourier series is the best trigonometric approximation of degree n to f, and that the error goes to zero as $n \to \infty$, which amounts to convergence in the mean.

Convergence in the mean is in fact equivalent to the first Plancherel (Pythagoras) formula in (3.124). On page 115 the author avoids the consideration of different norms with a direct proof of the Plancherel formula for the sum function of a uniformly convergent Fourier series. For general square integrable functions the Plancherel formula then follows by approximation arguments (the details are not given in the book). Thus, although for continuous functions even pointwise convergence may fail, as complicated examples show, one can be content that convergence in the mean always holds.

Section 3.2 emphasises that Fourier series are *not* like power series. Therefore Section 3.3 on what is allowed in differentiation and integration of Fourier

series should be read with care. The statements are only about the coefficients, not about convergence of the Fourier series, and they follow from integration by parts. Indeed, the formula's for the Fourier coefficients can be integrated by parts whenever $f'(x)$ exists, giving similar formula's with prefactors $\frac{1}{k}$ and thus better (and faster) convergence of the Fourier series than expected from the defining formula's. The smoother f, the more times we can integrate by parts and gain a prefactor $\frac{1}{k}$, and thus the better and faster the convergence of the Fourier series. However, for uniform convergence we need (3.99) with $\alpha > 1$, which requires more than one derivative for f, so this trick cannot compete with the direct convergence proof which only needs the first order derivative.

Nevertheless, the larger we can choose n in (3.100), the smoother the sum function of the Fourier series: Thm 3.31. Solving PDE's, we compute Fourier series solutions

$$u(t,x) \sim \sum \dots$$

Solutions of the heat equation $u_t = u_{xx}$ are easily seen to be very well behaved in this respect as soon as $t > 0$, even with ugly initial data, because of the exponentials in (3.27). However, interesting (fractal) issues with respect to lack of smoothness will appear in Chapter 7 when we solve $u_t = u_{xxx}$ with e.g. piecewise constant initial data.

Read also Section 3.4 yourself. It discusses simple changes of scale needed to deal with l-periodic functions $f(x)$, $l > 0$. The limit $l \to \infty$ leads to the Fourier integral transform in Chapter 7.

Exercise JH4. Assume that f and f' are piecewise continuous, so that f' is bounded by a fixed constant M. Show that the function $g(y)$ defined immediately after (3.131) is bounded in terms of M. Then split the integral of $g(y)\sin(n+\frac{1}{2})y$ in $\int_0^\delta$ and $\int_\delta^\pi$. The first is bounded in terms of δ and M. Integrate the second one by parts and estimate in terms of δ, M and n which appears in the denominator. Conclude the pointwise convergence proof without using the Riemann-Lebesgue lemma and show that the convergence is uniform if f is continuous and f' is (piecewise) continuous.

26.4 The integral Fourier transform

Next I slightly modify the presentation in 7.1 of Olver. Let $f = f(x)$ be defined on the real line, smooth and $f(x) = 0$ for $|x| \geq l$. Then the (uniformly convergent) Fourier series of f on the interval $[-l, l]$ is obtained via scaling:

$$F(y) = f(x), \quad \frac{x}{l} = \frac{y}{\pi}.$$

We have

$$f(x) = F(y) = \sum_{n=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} F(\eta) e^{-in\eta} d\eta \, e^{iny}$$

$$= \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \frac{\pi}{l} \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-l}^{l} f(\xi) e^{-i\frac{n\pi}{l}\xi} d\xi}_{\widehat{f}(\frac{n\pi}{l}) = \widehat{f}(k_n)} \, e^{i\frac{n\pi}{l}x}.$$

Set

$$k_n = n\Delta k, \quad \Delta k = \frac{\pi}{l},$$

and define

$$\widehat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx} dx,$$

then

$$f(x) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \widehat{f}(k_n) e^{ik_n x} \, \Delta k.$$

In other words, $f(x)$ is equal to a uniformly convergent sum, which actually is a Riemann sum for

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \widehat{f}(k) e^{ikx} dk,$$

and likewise

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \int_{-l}^{l} |f(x)|^2 dx = \sum_{n=-\infty}^{\infty} |\widehat{f}(k_n)|^2 \, \Delta k,$$

which is a Riemann sum for

$$\int_{-\infty}^{\infty} |\widehat{f}(k)|^2 dk.$$

Note that these identities remain valid if we increase l. In particular both Riemann sums are independent of Δk in the limit $\Delta k \to 0$.

Can we conclude that both $f(x)$ and $\int |f(x)|^2$ are also equal to the integrals themselves? The answer will be yes if $\widehat{f}(k)$ is continuous and decays sufficiently fast as $|k| \to \infty$, to make the tails of both integrals and Riemann sums small. Then we can restrict the convergence argument to integrals and Riemann sums on bounded k-intervals. Since

$$\widehat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx} dx =$$

$$\frac{1}{\sqrt{2\pi}}\left(\left[f(x)\frac{e^{-ikx}}{-ik}\right]_{-\infty}^{\infty}-\int_{-\infty}^{\infty}f'(x)\frac{e^{-ikx}}{-ik}dx\right)=\frac{\widehat{f'}(k)}{ik}=\frac{\widehat{f^{(m)}}(k)}{(ik)^m}$$

and

$$|\widehat{f^{(m)}}(k)|\leq 2l\max_{-l\leq x\leq l}|f^{(m)}(x)|,$$

we can conclude that indeed

$$f(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\widehat{f}(k)e^{ikx}dk,\quad \int_{-\infty}^{\infty}|f(x)|^2dx=\int_{-\infty}^{\infty}|\widehat{f}(k)|^2dk$$

for all f in C_c^∞, the class of smooth compactly supported functions.

Denoting the class of measurable complex valued functions f with bounded $\int|f(x)|dx$ by L^1 and the class with bounded $\int|f(x)|^2dx$ by L^2 we have that C_c^∞ is dense in L^2, and that the map

$$\mathcal{F}:C_c^\infty\to L^1\cap L^2\cap C^\infty\subset L^2,\quad \mathcal{F}(f)=\widehat{f}$$

uniquely extends to an isometry

$$\mathcal{F}:L^2\to L^2.$$

Only if $f\in L^1\cap L^2$ we can write $\widehat{f}(k)$ as an absolutely convergent integral. The inverse map is the same, up to a reflection in x, that is, for $f=f(k)$,

$$(\mathcal{F}^{-1})f(x)=\widehat{f}(-x).$$

More to be done here from a different perspective (Folland' s Real Analysis book).

26.5 The fast Fourier transform

These are some notes I prepared for a class at AUC that I took over once from JB. Not related to Olver's book, but of the same flavor as the section just above. Condsider an L-periodic function $f=f(t)$. Write

$$f(t)=\sum_{k=-\infty}^{\infty}c_ke^{\frac{2\pi kit}{L}},\quad c_k=\frac{1}{L}\int_0^L f(t)e^{-\frac{2\pi kit}{L}}dt$$

If we only know $f(t)$ in the points $t=0,\frac{L}{N},\frac{2L}{N},\frac{3L}{N},\dots$ 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 $t = 0, \frac{L}{N}, \frac{2L}{N}, \frac{3L}{N}, \ldots$ as $f_1, f_2, \ldots, f_N$ and write $F_k = N\tilde{c}_{k-1}$. The relation between the N-vectors F and f is then given by

$$F = \Omega f, \quad \Omega_{jk} = \omega^{(j-1)(k-1)} \quad (j, k = 1, \ldots, N), \quad \omega = e^{-\frac{2\pi i}{N}}$$

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

$$\tilde{\omega} = \omega^2, \quad \tilde{\Omega}_{jk} = \tilde{\omega}^{(j-1)(k-1)} \quad (j, k = 1, \ldots, n)$$

and decompose the vector f in $f_o = (f_1, f_3, \ldots, f_{N-1})$ and $f_e = (f_2, f_4, \ldots, f_N)$. Reshuffling the rows of Ω accordingly we find for $F^+ = (F_1, F_2, \ldots, F_n)$ and $F^- = (F_{n+1}, F_{n+2}, \ldots, F_{2n})$ 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 ω^{j-1} with $j = 1, \ldots, n$. 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 $N \log_2 N$.

26.6 Prerequisites Sobolev spaces and PDE

In the national mastermath programme I teach a course on PDE's with Herman Jan Hupkes. My part is basically Chapter 23 on Sobolev spaces and the chapter on the weak solution approach for problems like (21.1) that follows. Essential is the use of the divergence theorem. You should be familiar with the analytic part of the theory of Chapter 18 from the viewpoint taken in Section 18.1: first local, flattening the boundary and then global using partitions of unity, with the boundary integrals defined using surface area as in Section 18.4. This requires basic linear algebra and matrix theory, see Section 18.5, that generalises to Theorem 18.3 in Section 18.6 and the discussion of the spectral theorem for compact symmetric operators that follows.

To understand what Sobolev spaces are you need to know what the L^p-spaces are, and the use statement of the Lebesque differentation theorem, see Chapter 22. You also have to be familiar with some analysis in Banach spaces, see Chapter 6, and the Arzéla-Ascoli Theorem, see Section 4.6. Of course you should now some basic about Hilbert spaces too, but not needed is a full blown functional analysis course.

27 Airy functions

I did the calculations below while reading Chapter 8 in Peter Olver's new PDE book with a little help of E.J. Hinch's nice little Cambridge Applied Math textbook on perturbation methods. Just goes to show how beautiful (also applied) complex analysis is.

The Airy function is defined by

$$\mathrm{Ai}(\xi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i(\xi x + \frac{x^3}{3})}dx,$$

an integral barely convergent. The Airy function plays the same role in the theory for $u_t + u_{xxx} = 0$ as the Gaussian $e^{-\frac{1}{2}x^2}$ for $u_t = u_{xx}$. Both functions define the spatial profile of the fundamental solution.

Replacing $\xi \in \mathbb{R}$ by $\zeta \in \mathbb{C}$ the Airy function is a complex analytic function of

$$\zeta = \xi + i\eta = \rho e^{i\psi}.$$

Replacing also $x \in \mathbb{R}$ by

$$z = x + iy \in \mathbb{C}$$

one may deform the "real" contour C defined by $z = z(t) = t$ with $-\infty < t < \infty$ to another contour γ_ζ that connects two points at infinity. This can be done (without changing the outcome) as long as the integrals of

$$e^{i(\zeta z + \frac{z^3}{3})} = e^{\Phi(z;\zeta)}$$

over the connecting arcs $|z| = R$ between C and the new contour γ_ζ go to zero as $R \to \infty$.

To answer the question

$$\mathrm{Ai}(\rho)e^{i\psi} \sim \quad ? \quad \text{as} \quad \rho \to \infty \quad (\text{for } \psi \text{ fixed}),$$

one chooses the new contour to be one along which the absolute value of the integrand, $e^{\mathrm{Re}\,\Phi(z;\zeta)}$, is peaked and has fast decay as $|z| \to \infty$, and along which $\mathrm{Im}\,\Phi(z;\zeta) = \phi_\zeta$ is a ζ-dependent constant, so that

$$\mathrm{Ai}(\zeta) = \frac{1}{2\pi}\int_{\gamma_\zeta} e^{i(\zeta z + \frac{z^3}{3})}dz = \frac{1}{2\pi}\int_{\gamma_\zeta} \underbrace{e^{\mathrm{Re}\,\Phi(z;\zeta)}}_{\text{real, positive}}\, dz\, e^{i\phi_\zeta},$$

in which the integrand is real, although $dz = dx + idy$ will typically still make the integral complex. The factor $e^{i\phi_\zeta}$ contains the "stationary phase"

ϕ_ζ. If M_ζ is the maximum of $\operatorname{Re}\Phi(z;\zeta)$ along γ_ζ, realised in some $z = m_\zeta$, one may also factor out e^{M_ζ} and write

$$\mathrm{Ai}(\zeta) = \frac{e^{M_\zeta + i\phi_\zeta}}{2\pi} \underbrace{\int_{\gamma_\zeta} e^{-f_0(z;\zeta)} dz}_{\to\, ?\,\text{as}\,|\zeta|\to\infty},$$

in which $f_0(z;\zeta) \geq 0$ along γ_ζ. Typically $f_0(z;\zeta)$ has a unique global minimum zero along γ_ζ and $f_0(z;\zeta) \to +\infty$ as $|z| \to \infty$ along γ_ψ. Note though that the integrand is likely to be ill-behaved as $\rho = |\zeta| \to \infty$, also because the contour γ_ζ may disappear in the limit. The resolution of this latter complication may be prepared by scaling x *before* going to complex variables and making the optimal choice of γ_ζ.

Thus, returning to the definition of $\mathrm{Ai}(\zeta)$ one writes $\mathrm{Ai}(\rho e^{i\psi}) =$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \underbrace{e^{i(\rho e^{i\psi}x + \frac{x^3}{3})}}_{\text{scale}\, x = \rho^{\frac{1}{2}}u} dx = \frac{\rho^{\frac{1}{2}}}{2\pi}\int_{-\infty}^{\infty} e^{i\rho^{\frac{3}{2}}(e^{i\psi}u + \frac{u^3}{3})} du = \frac{\rho^{\frac{1}{2}}}{2\pi}\int_{-\infty}^{\infty} e^{\rho^{\frac{3}{2}}\Psi(u)} du,$$

in which you should now view u as $u = \operatorname{Re} w$ with $w = u + iv \in \mathbb{C}$. The effect of this scaling is that the level lines of $\operatorname{Im}\Psi(w)$ are independent of ρ.

One has

$$\Psi(w) = i(e^{i\psi}w + \frac{w^3}{3}) = f(u,v;\psi) + ig(u,v;\psi),$$

with

$$f(u,v;\psi) = -v\cos\psi - u\sin\psi + v(-u^2 + \frac{v^2}{3})$$

and

$$g(u,v;\psi) = u\cos\psi - v\sin\psi + u(\frac{u^2}{3} - v^2).$$

These harmonic functions have mutually perpendicular level curves. It is convenient to think of the level curves of the imaginary part $g(u,v;\psi)$ as orbits of

$$\dot{u} = \frac{du}{dt} = f_u = \frac{\partial f}{\partial u} = -\sin\psi - 2uv$$
$$\dot{v} = \frac{dv}{dt} = f_v = \frac{\partial f}{\partial v} = -\cos\psi - u^2 + v^2,$$

a system of ordinary differential equations for $u = u(t)$ and $v = v(t)$. In fact, Cauchy-Riemann gives

$$\frac{df}{dt} = f_u\dot{u} + f_v\dot{v} = f_u^2 + f_v^2 > 0\,, \quad \frac{dg}{dt} = g_u\dot{u} + g_v\dot{v} = g_uf_u + g_vf_v = 0.$$

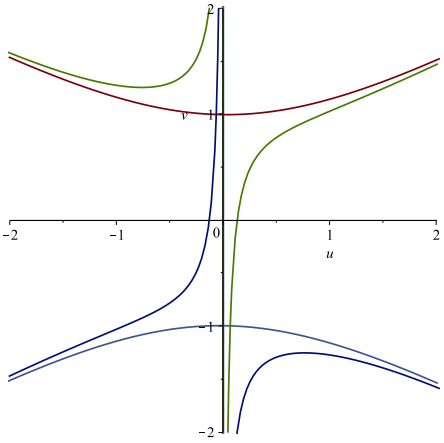


Figure 1: orbits for $\psi = \frac{\pi}{24}$

Thus the only orbits of interest as possible contours are the stable manifolds of saddle points, with the maximum of the real part f along the contour occurring in the saddle point.

This is illustrated for small positive ψ by Figure 1 which pictures the possibly relevant level curves of $g(u, v; \frac{\pi}{24})$. The red curve is the stable manifold of

$$m_\psi = (u_\psi, v_\psi) = (-\sin\frac{\psi}{2}, +\cos\frac{\psi}{2})$$

and asymptotes to $3v^2 = u^2$. In particular it has u ranging from $-\infty$ to $+\infty$, and may be written as the graph of a function $v = \varphi(u; \psi)$. The other stable manifold, that of

$$m_{\psi+2\pi} = (u_{\psi+2\pi}, v_{\psi+2\pi}) = (+\sin\frac{\psi}{2}, -\cos\frac{\psi}{2}),$$

the green curve on the right, fails this condition and has a vertical asymptote. The other branch of this level curve is the green curve on the left which is not a stable manifold of either two saddles. Neither of these two orbits is of direct use in relation to the Airy function, but this will change as ψ is taken larger. Note that although $\mathrm{Ai}(\rho e^{i\psi})$ is 2π-periodic in ψ, the parametrisation of the saddle point m_ψ is only 4π-periodic.

Deforming the contour as explained above,

$$\mathrm{Ai}(\rho e^{i\psi}) = \frac{\rho^{\frac{1}{2}}}{2\pi} \underbrace{\int_{-\infty}^{\infty} e^{\rho^{\frac{3}{2}}(f(u,\varphi(u;\psi);\psi)}(1 + i\varphi'(u;\psi))du}_{I(\rho,\psi)} e^{\rho^{\frac{3}{2}}(-\frac{2i}{3}(4\cos^2\frac{\psi}{2}-1)\sin\frac{\psi}{2})},$$

in which the phase factor has been made precise. It remains to examine the integral $I(\rho,\psi)$ in the limit $\rho\to\infty$. Clearly the most important information comes from the (second order) Taylor expansion

$$f = f(u,\varphi(u;\psi);\psi) = M_\psi - a_\psi^2(u-u_\psi)^2 + \cdots$$

with minor contributions coming from the higher order terms and the expansion of $\varphi'(u;\psi)$. Setting

$$u = u_\psi + p$$

one sees that to leading order the asymptotic expansion of the integral must be given by

$$I(\rho,\psi) \sim e^{M_\psi\rho^{\frac{3}{2}}}\int \underbrace{e^{-a_\psi^2\rho^{\frac{3}{2}}p^2}}_{\text{scale } s=a_\psi\rho^{\frac{3}{4}}p}\, dp \sim \frac{e^{M_\psi\rho^{\frac{3}{2}}}}{a_\psi\rho^{\frac{3}{4}}}\underbrace{\int e^{-s^2}ds}_{\sqrt{\pi}} + \cdots,$$

so that

$$\mathrm{Ai}(\rho e^{i\psi}) = \frac{1}{2\rho^{\frac{1}{4}}\sqrt{\pi}}\frac{e^{M_\psi\rho^{\frac{3}{2}}}}{a_\psi}\, e^{\rho^{\frac{3}{2}}(-\frac{2i}{3}(4\cos^2\frac{\psi}{2}-1)\sin\frac{\psi}{2})}(1+O(\rho^{-\frac{3}{2}}))$$

as $\rho\to\infty$. Notice the exponential decay combined with the increasingly rapid oscillations because of the phase factor.

At first sight you might expect an $O(\rho^{-\frac{3}{4}})$ error estimate but since the exponential function in the integrand expands as

$$e^{-s^2+b_3\frac{p^3}{\rho^{\frac{3}{4}}}+b_4\frac{p^4}{\rho^{\frac{6}{4}}}+b_5\frac{p^5}{\rho^{\frac{9}{4}}}+\cdots} = e^{-s^2}(1+(b_3\frac{p^3}{\rho^{\frac{3}{4}}}+\dots)+\frac{1}{2}(b_3\frac{p^3}{\rho^{\frac{3}{4}}}+\dots)^2+\cdots)$$

the higher order terms in the expansion of $\mathrm{Ai}(\rho e^{i\psi})$ involve the integrals

$$\int s^n e^{-s^2}ds \quad (n=3,4,\dots)$$

of which the odd ones vanish. Therefore a contribution of the first b_3-term appears only in combination with the first order term in the expansion of $\varphi'_\psi(u_\psi;\psi)$ (the second order term in the expansion of $\varphi_\psi(u_\psi;\psi)$). It is an exercise to make the expansion more precise.

One has

$$M_\psi = -\frac{2}{3}(4\cos^2\frac{\psi}{2}-3)\cos\frac{\psi}{2},$$

and by direct but tedious calculation the stable manifold is given by

$$v = \varphi_\psi(u) = \varphi_\psi(-s+p) = \frac{-sc + p\sqrt{1 - \frac{4}{3}sp + \frac{1}{3}p^2}}{-s+p} \quad (c = \cos\frac{\psi}{2},\ s = \sin\frac{\psi}{2}).$$

You should recognise a discriminant under the square root, which for the level curve going through the saddle has the property that it is everywhere positive, except in the saddle point m_ψ. Expansion gives

$$\varphi_\psi(-s+p) = c + \frac{-1+c}{s}p + \frac{1}{3}\frac{2c-1}{c+1}p^2 + \dots.$$

For $\psi = 0$ one has

$$v = \varphi_0(u) = 1 + \sqrt{1 + \frac{1}{3}u^2} = 1 + \frac{1}{6}u^2 - \frac{1}{72}u^4 + \cdots$$

and

$$f(u, \varphi_0(u)) = -2(1 + \frac{4}{9}u^2)\sqrt{1 + \frac{1}{3}u^2} = -\frac{2}{3} - u^2 - \frac{5}{36}u^4 + \cdots,$$

so that

$$a_0 = 1,\ M_0 = -\frac{2}{3},\ A_0 = 0,$$

and

$$\mathrm{Ai}(\xi) \sim \frac{e^{-\frac{2}{3}\xi^{\frac{3}{2}}}}{2\sqrt{\pi}\xi^{\frac{1}{4}}}$$

as $\xi \to +\infty$, give or take a mistake in the constants, without oscillations.

Increasing ψ there are changes as ψ crosses $\frac{\pi}{3}$ and $\frac{2\pi}{3}$. For all $0 \le \psi < \frac{2\pi}{3}$ it still holds that

$$\mathrm{Ai}(\rho e^{i\psi}) = \frac{\rho^{\frac{1}{2}}}{2\pi}\int_{-\infty}^{\infty} e^{\rho^{\frac{3}{2}}(f(u,\varphi(u;\psi);\psi)}(1 + i\varphi'(u;\psi))du\ e^{\rho^{\frac{3}{2}}(-\frac{2i}{3}(4\cos^2\frac{\psi}{2}-1)\sin\frac{\psi}{2})}.$$

Figure 2 shows the relevant orbits for $\psi = \frac{15\pi}{24}$, with the same stable manifold defining the contour, and the same asymptotics still valid, but with a different sign for M_ψ, as Figure 3 shows. The sign change occurs at $\frac{\pi}{3}$. Thus for $\frac{\pi}{3} < \psi < \frac{2\pi}{3}$ there is exponential growth of $\mathrm{Ai}(\rho e^{i\psi})$ as $\rho \to \infty$, while the nonzero phase factor accounts for increasingly rapid oscillations.

At $\psi = \frac{2\pi}{3}$, when the growth is maximal (and no oscillations, see Figure 8), the diagram (and the Maple automatic colour coding) changes. All orbits in Figure 4 are in the stable or unstable manifolds of the saddle points.

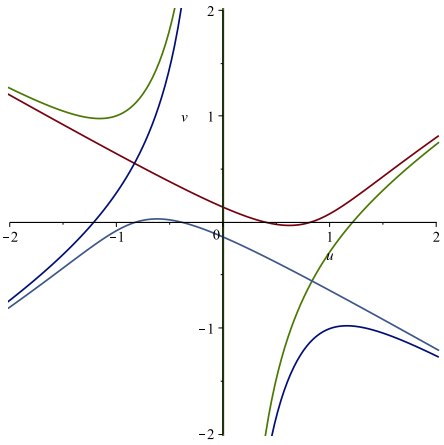


Figure 2: orbits for $\psi = \frac{15\pi}{24}$

The appropriate contour now consists of 3 orbits: the 2 orbits in the stable manifold of $m_{\frac{2\pi}{3}}$ (one of which is in the unstable manifold of $m_{\frac{8\pi}{3}}$), and one orbit in the stable manifold of $m_{\frac{8\pi}{3}}$. Can you see which one? You should convince yourself that $M_{\frac{8\pi}{3}}$ only enters the asymptotics beyond any relevant order.

Only as ψ is increased to $\psi = \pi$ both M_π and $M_{3\pi}$ are on par: $M_\pi = M_{3\pi} = 0$. The phases are then $\phi_\pi = \frac{2}{3}$ and $\phi_{3\pi} = -\frac{2}{3}$, and the two stable manifolds are given by

$$v = \frac{(u+1)\sqrt{u(u-2)}}{u\sqrt{3}} \quad (u < 0)\,, \quad v = \frac{(u-1)\sqrt{u(u+2)}}{u\sqrt{3}} \quad (u > 0).$$

You can now compute the expansion using both contours, with u running from $-\infty$ to 0 for the first integral and from 0 to ∞ for the second. Note the symmetry in Figure 7.

Observe that for $\frac{2\pi}{3} < \psi \leq \pi$ the contours are different. In Figures 5 you see the red curve turning blue after the turning point, and as it escapes to infinity along the negative v-axis it is joined by the green curve which is the stable manifold of the other saddle point. As in the case that $\psi = \pi$, the appropriate contour consists in fact of two contours: the sum of the integrals along both stable manifolds defines $\mathrm{Ai}(\rho e^{i\psi})$. For $\psi < \pi$ the main contribution comes from the contour on the left. Solving a cubic equation this contour can be written as a graph $u = \varphi(v)$, but the main contribution can be computed as above, still writing $v = \varphi(u)$ near the saddle point.

A similar program works for the solution of $u_t + \frac{1}{3}u_{xxx} = 0$ that starts

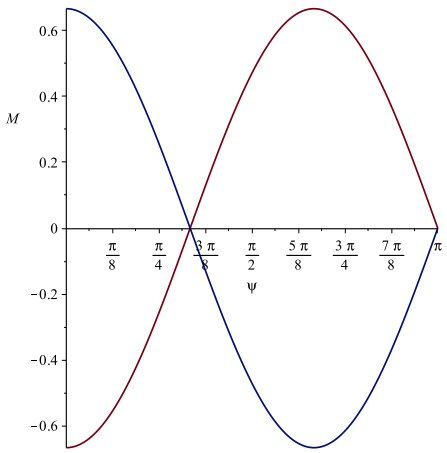


Figure 3: M_ψ and $M_{\psi+2\pi}$

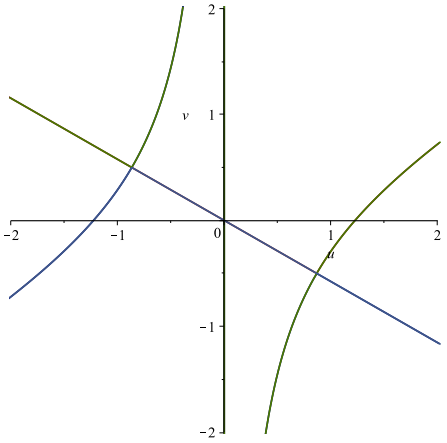


Figure 4: orbits for $\psi = \frac{16\pi}{24} = \frac{2\pi}{3}$

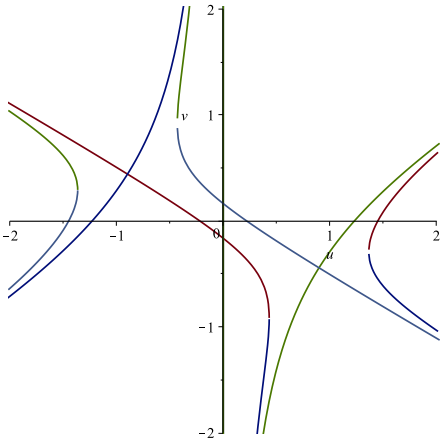


Figure 5: orbits for $\psi = \frac{17\pi}{24}$

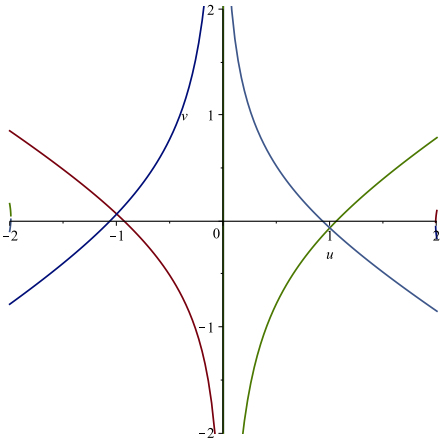


Figure 6: orbits for $\psi = \frac{23\pi}{24}$

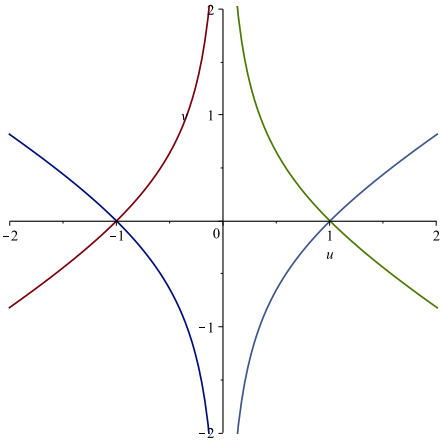


Figure 7: orbits (stable manifolds: blue curves) for $\psi = \frac{24\pi}{24} = \pi$

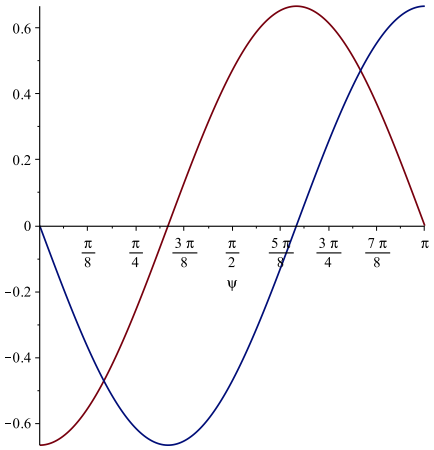


Figure 8: Amplitude M_ψ (red) and phase ϕ_ψ, changes at $\psi = 0, \frac{\pi}{3}, \frac{2\pi}{3}, \pi$.

from a "wave packet"

$$u_0(x) = e^{-\frac{x^2}{4a}} e^{ik_0 x},$$

along lines $x = ct + \xi$. One then has

$$u(t, ct+\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ik(c+\frac{1}{3}k^2)t + i\xi k - a(k-k_0)^2} dk = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t\Phi} dk.$$

Replace k by $z = x + iy$ (not the same x of course) and write

$$\Phi = \Phi(z) = \Phi(z;t) = \Phi(z;t,\xi,k_0,a) = f + ig$$

with

$$f = -ty(c + x^2 - \frac{1}{3}y^2) + a(-(x-k_0)^2 + y^2) - \xi y$$

and

$$g = tx(c + \frac{1}{3}x^2 - y^2) - 2a(x-k_0)y + \xi x.$$

Then as before on may rewrite

$$u(t, ct+\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\Phi(x;t)} dx = \frac{1}{\sqrt{2\pi}} \int_{\gamma} e^{\Phi(z;t)} dz$$

in which γ consists of orbits in stable manifolds of a suitable gradient flow of f, which is defined by

$$\dot{x} = \frac{dx}{d\tau} = \frac{1}{t}\frac{\partial f}{dx} = -2xy - \frac{2a}{t}(x-k_0),$$

$$\dot{y} = \frac{dy}{d\tau} = \frac{1}{t}\frac{\partial f}{dy} = -c - x^2 + y^2 + \frac{2ay - \xi}{t}.$$

Unlike in the analysis of the Airy function integral, there is now no need to scale x and y, because in the limit $t \to \infty$ the diagram in the x, y-plane is well defined. For $c = 1$ it is the same as in Figure 7 and for $c = -1$ it coincides with Figure 9 (with u, v replaced by x, y). Unlike the u, v-diagram the x, y-diagram varies with the parameter under consideration, as the role of ρ is now played by t. One computes the relevant unstable manifold(s) directly from solving $g = \phi$, which is a quadratic equation in y, asking that the discriminant

$$D = \frac{4}{3}x^2(x^2 + 3c)t^2 + 4x(\xi x - \phi)t + 4a^2(x-k_0)^2$$

of this equation is positive except in the saddle point, thus first determining simultaneously the saddle point and the phase ϕ by solving

$$D = \frac{dD}{dx} = 0.$$

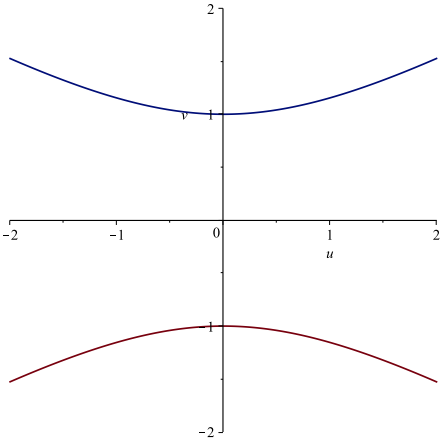


Figure 9: u, v-plane for $\psi = 0$, the vertical axis also consist of orbits

The phase ϕ then drops out of

$$x\frac{dD}{dx} - D = t^2x^4 + (a^2 + \xi t + ct^2)x^2 - k_0^2a^2 = 0,$$

which determines the square of the positive solution $x = x_c > 0$ uniquely in terms of the parameters t, c, ξ, a, the y-coordinate $y = y_c$. The phase ϕ_c and the value M_c of f in the saddle point (x_c, y_c) are then given by

$$y_c(t) = \frac{a(k_0 - x_c)}{x_c t}, \quad \phi_c(t) = -\frac{2tx_c^3}{3} - \frac{2a^2k_0(x_c - k_0)}{x_c t},$$

and

$$M_c(t) = -a(x_c - k_0)^2 - \frac{a^3(x_c + 2k_0)(x_c - k_0)^2}{3x_c^3t^2}.$$

For the values of y, ϕ, M in the other saddle point replace x_c by $-x_c$. Note that $x_c = x_c(t)$ and likewise for y_c, ϕ_c, M_c (the other dependencies are also suppressed in the notation). Observe the different behaviours as $t \to \infty$ for $c < 0$ and $c > 0$.

At this point I found it convenient to continue the calculations for the stable manifold with x_c implicitly defined by the quartic $x\frac{dD}{dx} - D$ and all other quantities explicitly in terms of x_c. With

$$x = x_c + u, \quad y = y_c + v,$$

the real and imaginary parts of Φ rewrite as

$$f = M_c + F_c, \quad F = F_c(t) = -t(2x_cuv + v(u^2 - \frac{v^2}{3})) - \frac{ak_0}{x_c}(u^2 - v^2),$$

$$g = \phi_c + G_c, \quad G = G_c(t) = t(x_c(u^2 - v^2) + u(\frac{u^2}{3} - v^2)) - \frac{2ak_0uv}{x_c},$$

the latter defining the stable manifold as the graph $v = \varphi_c(u) = \varphi_c(u;t)$ obtained from solving $G = 0$ for v, the discriminant having the desired behaviour: positive except for $u = 0$. For $c > 0$ this gives a globally defined function and deforming contours as before it follows that

$$u(t, ct + \xi) = \frac{e^{M_c(t) + i\phi_c(t)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{F_c(u,\varphi_c(u;t);t)}(1 + i\varphi_c'(u;t))du.$$

Note that c has disappeared completely from the formula's, except for the dependence through x_c.

The integral depends only on the formula's for F_c and G_c, φ_c being defined by solving $G = 0$, i.e.

$$(x_c + u)tv^2 + \frac{2ak_0u}{x_c}v - tu^2(x_c + \frac{u}{3}) = 0$$

and simplifying the discriminant using the quartic for x_c. This gives

$$v = \varphi_c(u) = \frac{u}{x_c(x_c + u)}\left(-\frac{ak_0}{t} + x_cR\right),$$

in which

$$R = \sqrt{\underbrace{c + 2x_c^2 + \frac{\xi}{t} + \frac{a^2}{t^2}}_{\text{positive}} + \frac{4x_cu}{3} + \frac{u^2}{3}}.$$

The derivative appears in the integral as

$$\varphi_c'(u) = -\frac{ak_0}{t(x_s + u)^2} + \frac{u(2x_c + u)}{3R(x_c + u)},$$

and the exponent in the integral rewrites as

$$F_c(u, \varphi_c(u;t);t) = -\frac{2}{9}\frac{tRu^2\,(3\,x_c + 2\,u)^2}{(x_c + u)^2} + \frac{2}{3}\frac{ak_0u^2\,(3\,x_c + 2\,u)}{(x_c + u)^2}$$

$$-\frac{2}{3}\frac{{k_0}^2a^2\,(Rx_ct - ak_0)\,u^2\,(3\,x_c + u)}{t^2{x_c}^3\,(x_c + u)^3}$$

Clearly these formula's suggest putting

$$u = x_cs, \quad k_0 = \frac{b}{a}, \quad t = b\tau, \quad \xi = b\eta$$

This scaling of u makes each term separate as far the integration variable and t are concerned, except for the R-terms. One now has to distinguish between $c < 0$ (the case discussed by Olver) when $u(t, ct + \xi)$ appears as the sum of 2 integrals involving the stable manifolds of both saddles, and $c > 0$, when $u(t, ct + \xi)$ appears as one single integral.

With x_c going to $\sqrt{-c}$ if $c < 0$, both integrals can be handled as in the Airy case, and the final expansion will depend on c. It may be handy to split the exponential in 3 separate exponentials before you proceed. From the c-dependence there should be a connection with the group velocity discussion by Olver, as we see below. On the other hand, when $c > 0$ (this case is not discussed by Olver) tx_c goes to a constant so $x_c \to 0$. Note that all 3 terms involve s^2, but with different signs. This is really an instructive example for understanding the method!

Now to back to WHY we did this analysis observe that the prefactor in the integral expression for

$$u(t, ct + \xi)$$

contains

$$e^{M_c}$$

which behaves very differently for $c < 0$ and $c > 0$.

For $c > 0$ it is the second term in the expression for $M_c(t)$ above that dominates and goes to infinity because tx_c goes to a constant, and this leading order term then goes to $-\infty$ linearly in t. Modulo the details of the analysis of the integral it follows that $u(t, ct + \xi) \to 0$ exponentially fast as $t \to \infty$.

On the other hand, if $c < 0$ then $x_c \to \sqrt{-c}$ and $M_c(t) \to -a(\sqrt{-c} - k_0)^2$ which is maximal and equal to zero for $c = -k_0^2$. Thus only for this value of c the solution is of order one along the line $x = ct + \xi$ as $t \to \infty$, with the more precise asymptotics following from a more detailed analysis of the integral, as in the Airy functions case, with contributions from both saddle points, and combining both phases and $\frac{2}{3}c^{\frac{3}{2}}t$ appearing in the imaginary part. Olver's point in the section about dispersion relations is that this *group velocity* $-c$ is 3 times larger as you would expect from looking at the single frequency solution with $a = 0$, and he did so by one single calculation starting from the dispersion relation. Read again what he did after the exam, and pay attention to the factor $\frac{1}{3}$ in the third order equation $u_t + \frac{1}{3}u_{xxx} = 0$ that I solved starting from a wave packet centered at $k = k_0$ rather than from a single wave with $k = k_0$.

1. This is an exercise about applying the Fourier transform to solve the equation $u_t + u_{xxx} = 0$ on the real line with initial data $u(0,x) = \delta(x)$, the Dirac δ-function, and to investigate the behaviour of the solution for $x \to -\infty$. The Fourier transform of a function $f = f(x)$ and the inverse transform are defined by

$$\hat{f}(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx} dx, \quad f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(k) e^{ikx} dk,$$

respectively whenever f and $\hat{f}$ are sufficiently nice. The improper integrals are to be understood in the principal value sense

$$\int_{-\infty}^{\infty} = \lim_{R \to \infty} \int_{-R}^{R}$$

and are often easiest evaluated using complex integration over appropriate contours.

   (a) Explain why $\hat{\delta}(k) = \frac{1}{\sqrt{2\pi}}$.

   (b) Show using integration by parts that $\widehat{(f')}(k) = ik\hat{f}(k)$.

   (c) Let u be a smooth solution of $u_t + u_{xxx} = 0$ which decays to zero sufficiently fast as $|x| \to \infty$ to have $(\hat{u})_t = \widehat{(u_t)}$. Here $\hat{u} = \hat{u}(t,k)$ denotes the Fourier transform of the function $x \to u(x,t)$. Denote the initial value of u by u_0, that is, $u_0(x) = u(0,x)$. Show that

$$\hat{u}(t,k) = \widehat{u}_0(k) e^{ik^3 t}$$

   (d) Show that the inversion formula formally applied to the case that $\widehat{u}_0(k) = \hat{\delta}(k) = \frac{1}{\sqrt{2\pi}}$ defines a solution formula

$$u(t,x) = \frac{1}{(3t)^{\frac{1}{3}}} \mathrm{Ai}\left(\frac{x}{(3t)^{\frac{1}{3}}}\right)$$

   in which

$$\mathrm{Ai}(\xi) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i(\xi x + \frac{x^3}{3})} dx.$$

   (e) Use the methods above to determine the asymptotic behaviour of $\mathrm{Ai}(\xi)$ for $\xi \to -\infty$.

28 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

$$I(u) = \int_a^b L(t, u(t), \dot{u}(t))\, dt \tag{28.1}$$

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:

$$\frac{d}{dt}\frac{\partial L}{\partial p^i} - \frac{\partial L}{\partial u^i} = 0 \quad (i = 1, \ldots, n) \tag{28.2}$$

We also saw the Jacobi equations, obtained from (1.3.6) and the linearised Lagrangian

$$\phi = \frac{\partial^2 L}{\partial p^i \partial p^j}\pi^i\pi^j + 2\frac{\partial^2 L}{\partial u^i \partial p^j}\pi^i\eta^j + \frac{\partial^2 L}{\partial u^i \partial u^j}\eta^i\eta^j \tag{28.3}$$

The Euler-Lagrange equations of (28.3) are the Jacobi equations

$$\frac{d}{dt}\frac{\partial \phi}{\partial \pi^i} - \frac{\partial \phi}{\partial \eta^i} = 0 \quad (i = 1, \ldots, n) \tag{28.4}$$

These Jacobi equations are the linearised Euler-Lagrange equations. Verify this!

For Lagrangians independent of t we noticed a conservation law. When you multiply (28.2) by $p^i(t) = \dot{u}^i(t)$ you get

$$0 = p^i(t)\frac{d}{dt}\frac{\partial L}{\partial p^i} - \dot{u}^i(t)\frac{\partial L}{\partial u^i} = \frac{d}{dt}\left(p^i\frac{\partial L}{\partial p^i}\right) \underbrace{-\dot{p}^i(t)\frac{\partial L}{\partial p^i} - \dot{u}^i(t)\frac{\partial L}{\partial u^i}}_{-\frac{dL}{dt}}$$

$$= \frac{d}{dt}\left(p^i\frac{\partial L}{\partial p^i} - L\right)$$

28.1 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

$$V = V^i x_i = V^i(u) x_i(u) = V^1(u) x_1(u) + \cdots + V^n(u) x_n(u), \tag{28.5}$$

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

28.2 Covariant differentiation

If we differentiate a vector field V as given by (28.5) 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 (28.5) with respect to u^j can be written as

$$\frac{\partial V}{\partial u^j} = \frac{\partial V^i}{\partial u^j} x_i + V^i x_{ij}, \quad x_{ij} = \frac{\partial x_i}{\partial u^j} = \frac{\partial^2 x}{\partial u^j \partial u^i} = \frac{\partial^2 x}{\partial u^i \partial u^j} = x_{ji} \tag{28.1}$$

In the case that $M = \mathbb{R}^n = \mathbb{R}^d$ with $x^i = u^i$, the tangent vectors x_i are the unit base vectors e_i so that $x_{ij} = 0$ and the covariant partial derivatives of V are just the partial derivatives V. The same holds if $x(u)$ in linear in u. In all other cases we decompose x_{ij} as

$$x_{ij} = \Gamma^l_{ij} x_l + \textit{ normal parts}$$

and take the inner product with x_k to get

$$\Gamma_{ijk} := x_{ij} \cdot x_k = \Gamma^l_{ij} x_l \cdot x_k = \Gamma^l_{ij} g_{lk}$$

Thus Γ_{ijk} is obtained from Γ^l_{ij} using g_{lk}. Introducing $g^{kl} = g^{lk}$ by

$$g_{lk} g^{km} = \delta^m_l,$$

we also obtain Γ^m_{ij} from Γ_{ijk}:

$$g^{mk} \Gamma_{ijk} = \Gamma^l_{ij} g_{lk} g^{km} = \Gamma^l_{ij} \delta^m_l = \Gamma^m_{ij}$$

The relation between both Γ-symbols is given by

$$\Gamma_{ijk} = \Gamma^l_{ij} g_{lk}, \quad \Gamma^m_{ij} = g^{mk} \Gamma_{ijk}$$

The metric coefficients are used to raise and lower the exponents[4].

Next we determine Γ_{ijk}. Differentiating g_{ij} with respect to u^k we get

$$g_{ij,k} = \frac{\partial g_{ij}}{\partial u^k} = \frac{\partial}{\partial u^k}(x_i \cdot x_j) = x_{ki} \cdot x_j + x_{jk} \cdot x_i = \Gamma_{kij} + \Gamma_{jki}$$

Note the two cyclic permutations kij and jki of ijk on the right. Using cyclic permutation, we have the following three equivalent forms of the resulting statement:

$$g_{ij,k} = \Gamma_{kij} + \Gamma_{jki}$$

[4]Just as with tensor coefficients, though the Γ's are not tensor coefficients

$$g_{jk,i} = \Gamma_{ijk} + \Gamma_{kij}$$
$$g_{ki,j} = \Gamma_{jki} + \Gamma_{ijk}$$

Multiplying by $-\frac{1}{2}$, $\frac{1}{2}$ and $\frac{1}{2}$ and adding up we get

$$\Gamma_{ijk} = \frac{1}{2}\left(g_{jk,i} + g_{ki,j} - g_{ij,k}\right)$$

Using the symmetry $g_{ij} = g_{ji}$ it follows that

$$\Gamma_{ijk} = \frac{1}{2}\left(g_{jk,i} + g_{ik,j} - g_{ij,k}\right), \quad \Gamma^m_{ij} = \frac{1}{2}g^{mk}\left(g_{jm,i} + g_{im,j} - g_{ij,m}\right) \tag{28.2}$$

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 (28.1) 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

$$D_{u^j}V := (\frac{\partial V}{\partial u^j})_T = \left(\frac{\partial V^l}{\partial u^j} + V^i\Gamma^l_{ij}\right)x_l, \quad V = V^i x_i \tag{28.3}$$

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

$$V^i \quad \text{and} \quad (D_{u^j}V)^l = \frac{\partial V^l}{\partial u^j} + V^i\Gamma^l_{ij}$$

28.3 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

$$\dot\gamma(t) = X(\gamma(t)), \quad \gamma(0) = x_0 \in M, \tag{28.1}$$

where X is a tangent vector field defined near x_0. The differential equation in (28.1) is called the *flow equation* for X. Using coordinates u, with $u = u_0$ corresponding to x_0, the expressions in (28.1) 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

$$\dot u^i = X^i(u), \quad u(0) = u_0. \tag{28.2}$$

The solution $u = u(t)$ exists locally and is unique. We have $\dot u^i(0) = X^i(u_0)$ and $X_0 := X(x_0) = \dot\gamma(0) = \dot u^i(0)x_i = X^i(u_0)x_i$. On scalar fields (functions) $\phi : M \to \mathbb{R}$, given in local coordinates as

$$\phi = \phi(u^1, \ldots, u^n),$$

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 ϕ in $u = u_0$, i.e. as the directional derivative

$$X_0^i\frac{\partial}{\partial u^i} \quad \text{corresponding to the direction vector} \quad X_0^i x_i$$

in $u = u_0$. The derivative only depends on the value of the vector field in x_0. Since the point $x_0 = x(u_0)$ was arbitrary we have

$$X = X^i\frac{\partial}{\partial u^i} \quad \text{corresponding to the tangent field} \quad X = X^i x_i = X^i\frac{\partial x}{\partial u^i}.$$

The two expressions above are merely different representations of the tangent vector field X (both in local coordinates):

The components

$$X^i\frac{\partial x^k}{\partial u^i}$$

of the *tangent field* X multiply

$$\frac{\partial\phi}{\partial x^k}$$

in the chain rule formula if ϕ is extended to a neighbourhood of M in $\mathbb{R}^d$. As *differential operator*

$$X = X^i\frac{\partial}{\partial u^i}$$

X acts on scalar fields like $\phi = \phi(u)$ and produces a scalar field $X\phi$, the derivative of ϕ in the direction of X. This directional derivative is denoted by

$$\nabla_X \phi = X\phi, \quad \text{replacing the notation} \quad \frac{\partial \phi}{\partial X}$$

in calculus texts. We already use the notation ∇_X customary for covariant differentiation. For reasons that should be clear, covariant differentiation of scalar fields is by definition the same as differentiation of scalar fields.

28.4 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 (28.1), 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 (28.2) around $u(t)$,

$$\dot{\xi}^i = \underbrace{(\frac{\partial X^i}{\partial u^j})}_{\text{in } (u(t)} \xi^j(t), \quad \xi^j(0) = Y^i(u_0) \tag{28.1}$$

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

$$[X, Y] = \mathcal{L}_X Y \tag{28.2}$$

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

$$[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0 \tag{28.3}$$

holds.

28.5 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 (28.3) 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

$$\nabla_X V = X^j \left(\frac{\partial V^l}{\partial u^j} + V^i \Gamma^l_{ij} \right) \frac{\partial}{\partial u^i} \tag{28.1}$$

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.

28.6 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

$$\bar{\nabla}_X (N \cdot Y) = \bar{\nabla}_X N \cdot Y + N \cdot \bar{\nabla}_X Y, \tag{28.1}$$

in which $\bar{\nabla}$ is the (standard covariant) derivative in $\mathbb{R}^d$. On M the left hand side of (28.1) 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

$$\bar{\nabla}_X N \cdot Y = -N \cdot \mathrm{I\!I}(X, Y) \quad \text{on } M. \tag{28.2}$$

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

$$h(X, Y) = N \cdot \mathrm{I\!I}(X, Y) = -\bar{\nabla}_X N \cdot Y = h_{ij} X^i Y^j \tag{28.3}$$

28.7 Curvature

The equality

$$\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z = \nabla_{[X,Y]} Z + R(X, Y) Z \tag{28.1}$$

[5] to M

[6] This can be done, certainly locally, why?

defines $R(X,Y)Z$ for tangent vector fields X, Y, Z. You may verify that $R(X,Y)Z$ is multilinear in X, Y, Z over the scalar fields on M. In the case $M = \mathbb{R}^n = \mathbb{R}^d$ you will find that $R(X,Y)Z \equiv 0$. The standard way to write $R(X,Y)Z$ in local coordinates u is

$$(R(X,Y)Z)^\alpha = R^\alpha_{ijk} Z^i X^j Y^k. \tag{28.2}$$

So Z comes first[7] and then X and Y. Using (28.1) and writing

$$\Gamma^\alpha_{ij,k} = \frac{\partial \Gamma_{ij}}{\partial u^k}$$

you should verify that[8]

$$R^\alpha_{ijk} = \Gamma^\beta_{ik}\Gamma^\alpha_{\beta j} - \Gamma^\beta_{ij}\Gamma^\alpha_{\beta k} + \Gamma^\alpha_{ik,j} - \Gamma^\alpha_{ij,k} \tag{28.3}$$

and the zero ijk and jk cyclic sums

$$R^\alpha_{ijk} + R^\alpha_{kij} + R^\alpha_{jki} = 0 = R^\alpha_{ijk} + R^\alpha_{ikj} \tag{28.4}$$

If W is another tangent field then[9]

$$Rm(X,Y,Z,W) = R(X,Y)Z \cdot W = R^\alpha_{ijk} Z^i X^j Y^k g_{\alpha l} W^l = R_{lijk} W^l Z^i X^j Y^k, \tag{28.5}$$

which has the symmetries

$$Rm(X,Y,Z,W) + Rm(Y,Z,X,W) + Rm(Z,X,Y,W) = 0,$$

$$Rm(X,Y,Z,W) + Rm(Y,X,Z,W) = 0 = Rm(X,Y,Z,W) + Rm(X,Y,W,Z)$$

(the second one obtained from $Rm(X,Y,Z,Z) = 0$), implying

$$Rm(X,Y,Z,W) = Rm(Z,W,X,Z)$$

In the 2-dimensional case $n = 2$ the only possible nonzero entries of R_{ijk} are

$$R_{1212} = R_{2121} = -R_{1221} = -R_{2112}$$

In the codimension 1 case

$$R_{lijk} = h_{ik}h_{lj} - h_{ij}h_{lk}$$

[7] As if we would have prefered the notation $ZR(X,Y)$

[8] note the order ijk in the minus terms and the $j \leftrightarrow k$ relation with the plus terms

[9] $lijk$ = dead body, as if we would have prefered the notation $W \cdot ZR(X,Y)$

consists of all the 2×2 determinants you can get from the matrix h_{ij}. Note that similarly

$$(W \cdot X)(Y \cdot Z) - (W \cdot Y)(X \cdot Z) = \underbrace{(g_{ik}g_{lj} - g_{ij}g_{lk})}_{G_{lijk}} W^l Z^i X^j Y^k, \qquad (28.6)$$

in which G_{lijk} has the same symmetry properties as R_{lijk} (and depends only on G_{1212} if $n = 2$).

For submanifolds you can verify from the definitions that

$$Rm(X, Y, Z, W) = \mathrm{I\!I}(X, W)\mathrm{I\!I}(Y, Z) - \mathrm{I\!I}(X, Z)\mathrm{I\!I}(Y, W), \qquad (28.7)$$

which in the codimension 1 case (28.3) reduces to

$$Rm(X, Y, Z, W) = h(W, X)h(Y, Z) - h(W, Y)h(X, Z),$$

so

$$Rm(X, Y, Z, W) = \underbrace{(h_{ik}h_{lj} - h_{ij}h_{lk})}_{R_{lijk}} W^l Z^i X^j Y^k, \qquad (28.8)$$

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 $\mathrm{I\!R}^d$ is the scalar ratio between (28.7) and (28.6). In $\mathrm{I\!R}^3$ this is the scalar ratio between (28.8) and (28.6).

28.8 Geodesic curves

A smooth curve $\gamma(t) \in M$ may require several coordinate patches to describe it. For the moment we assume that it can be described by one coordinate patch. If

$$\gamma : [a, b] \ni t \to u(t) \to x(u(t)) \in M$$

is such a curve in M, then its velocity is given by

$$\dot{\gamma} = \frac{\partial x}{\partial u^1}\dot{u}^1 + \cdots + \frac{\partial x}{\partial u^n}\dot{u}^n = \sum_{i=1}^{n} \dot{u}^i \frac{\partial x}{\partial u^i} = \sum_{i=1}^{n} \dot{u}^i x_i.$$

Think of $\dot{\gamma}$ as a vector at the point $x = \gamma(t)$ in M. For every t this vector is tangent to M, and written as a linear combination of the tangent vectors obtained from the parameterisation:

$$x_1 = \frac{\partial x}{\partial u^1}, \ldots, x_n = \frac{\partial x}{\partial u^n}.$$

Its length l is given by

$$l = \int_a^b |\dot\gamma(t)|\,dt = \int_a^b \sqrt{\dot\gamma(t)\cdot\dot\gamma(t)}\,dt = \int_a^b \sqrt{x_i\dot u^i\cdot x_j\dot u^j}\,dt$$

$$= \int_a^b \sqrt{\dot u^i\dot u^j g_{ij}(u)}\,dt$$

We will work with another quantity, called the energy, which involves an L as in Chapter 1. Since I prefer to have u in L, my u's are the γ's in the book. My $\gamma(t)$ is what is $c(t)$ in the book. The energy is defined by

$$E = \frac{1}{2}\int_a^b |\dot\gamma(t)|^2\,dt = \frac{1}{2}\int_a^b \dot\gamma(t)\cdot\dot\gamma(t)\,dt = \frac{1}{2}\int_a^b x_i\dot u^i\cdot x_j\dot u^j\,dt$$

$$= \frac{1}{2}\int_a^b \dot u^i\dot u^j g_{ij}(u)\,dt = \int_a^b L(u(t),\dot u(t))\,dt,$$

in which

$$L = L(u,p) = \frac{1}{2}p^i p^j g_{ij}(u). \tag{28.9}$$

Playing with the estimate

$$\int_a^b |\dot\gamma(t)|\,dt = \int_a^b 1\,|\dot\gamma(t)|\,dt \le \sqrt{\int_a^b 1^2\,dt}\sqrt{\int_a^b |\dot\gamma(t)|^2\,dt}$$

and reparameterisation of γ to make $|\dot\gamma|$ constant you should easily conclude that minimizers of l are minimizers of E and vice versa if we keep $[a,b]$ fixed.

The Euler-Lagrange equations for E involve the derivatives of g_{ij} and come out as

$$\ddot u^i + \Gamma^i_{\alpha\beta}\dot u^\alpha\dot u^\beta = 0 \tag{28.10}$$

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 (28.2). 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

$$g_{ij}(u) = g(|u|)\delta_{ij} \tag{28.11}$$

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 (28.11) 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 (28.11) in $\mathbb{R}^2$?

[10] implying $0 = g'(0) = g'''(0) = g''''''(0) = \cdots$

28.9 The Jacobi equations

Consider the Lagrangian (28.9).

- Show that the Jacobi equations (28.4) for (28.9) are

$$\ddot{\eta}^i + 2\Gamma^i_{jk}\dot{u}^j\dot{\eta}^k + \Gamma^i_{jk,l}\dot{u}^j\dot{u}^k\eta^l = 0 \tag{28.12}$$

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 (28.10) may be written as

$$D_t\dot{\gamma} = 0, \quad \dot{\gamma} = \dot{u}^i x_i$$

- Derive that (28.12) may be written as

$$(D_t^2\eta)^i + \dot{u}^\alpha R^i_{\alpha\beta k}\eta^\beta\dot{u}^k = 0, \quad \text{i.e.} \quad D_t^2\eta + R(\eta,\dot{\gamma})\dot{\gamma} = 0$$

29 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

$$u \to u - f'(u)^{-1} f(u) \tag{29.1}$$

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

$$u \to u + f'(0)^{-1}(v - f(u)) \tag{29.2}$$

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 (29.2) 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.

29.1 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

$$|f'(u)| \leq M_1 \quad \text{and} \quad |f''(u)| \leq M_2 \quad \forall u \in B. \tag{29.3}$$

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,

$$f(u_n) = \underbrace{f(u_{n-1}) + f'(u_{n-1})(u_n - u_{n-1})}_{\text{linear aproximation}} + Q_f(u_{n-1}, u_n), \tag{29.4}$$

in which

$$|Q_f(u_{n-1}, u_n)| \leq \frac{M_2}{2}|u_n - u_{n-1}|^2, \tag{29.5}$$

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 (29.3) which says that the linear map $f'(u) : X \to X$ satisfies

$$|f'(u)v| \leq M_1|v| \quad \forall u \in B \quad \forall u \in X, \tag{29.6}$$

but a similar bound

$$|L(u)| \leq C \tag{29.7}$$

for maps $L(u)$, that act as right inverses of $f'(u)$ in the sense that

$$f'(u_{n-1})L(u_{n-1})f(u_{n-1}) = f(u_{n-1}), \tag{29.8}$$

is essential. Writing

$$p_n = |u_n - u_{n-1}| \text{ and } q_n = |f(u_n)| \tag{29.9}$$

the Newton scheme

$$u_n = u_{n-1} - L(u_{n-1})f(u_{n-1}) \quad (n \in \mathbb{N}), \tag{29.10}$$

starting with $u_0 = 0$, then defines $u_n \in B$ as long as

$$p_1 + p_2 + \cdots + p_n < 1, \tag{29.11}$$

and the inequalities

$$p_n \leq Cq_{n-1} \text{ and } q_n \leq \frac{1}{2}M_2 p_n^2 \tag{29.12}$$

are immediate from (29.4,29.5,29.10). Note that (29.10) kills the linear approximation in (29.4). The inequalities in (29.12) are complemented by

$$q_0 = |f(0)| \quad \text{and} \quad p_1 \leq Cq_0 = C|f(0)|. \tag{29.13}$$

29.2 The optimal result

Clearly (29.12) and (29.13) combine as

$$p_n \le \mu p_n^2 \quad \text{with} \quad \mu = \frac{1}{2}MC \quad \text{and} \quad p_1 \le C|f(0)|, \tag{29.14}$$

and the condition to be stated is which $\bar{P} = \bar{P}(\mu)$ guarantees that the implication

$$C|f(0)| < \bar{P} \implies \sum_{n=1}^{\infty} p_n < 1 \tag{29.15}$$

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 (29.10) with $u_0 = 0$. Note that with $C|f(0)| \le \bar{P}$ the same conclusion will hold if only one of all the inequalites in the estimates below is strict, which will inevatibly be the case of course.

Obviously the smallest $\bar{P}$ we can get follows from replacing the three inequalities in (29.14) and (29.15) by inequalities. This leads to

$$p_n = \mu p_{n-1}^2 \ \text{ for } \ n \in \mathbb{N}; \quad p_1 = \bar{P}\,; \quad \sum_{n=1}^{\infty} p_n = 1. \tag{29.16}$$

Via $\xi_n = \mu p_n$ and $\xi_n = \xi_{n-1}^2$ this is easily seen to be equivalent to

$$\mu = G(\mu\bar{P}) \quad \text{with} \quad G(\xi) = \xi + \xi^2 + \xi^4 + \xi^8 + \xi^{16} + \cdots \tag{29.17}$$

but this does not yield a simple formula for $\bar{P} = \bar{P}(\mu)$.

29.3 A suboptimal result

A rough estimate

$$G(\xi) < \xi + \xi^2 + \xi^3 + \xi^4 + \xi^5 + \cdots = \frac{\xi}{1-\xi} \tag{29.18}$$

leads to a simple but suboptimal formula:

$$\bar{P} = \frac{1}{1+\mu} \quad \text{or} \quad \mu = \frac{1}{\bar{P}} - 1. \tag{29.19}$$

29.4 Alternative proof of convergence

The alternative approach to (29.12) and (29.13) in [KP] is not to solve the corresponding system with equalities but to derive an estimate of the form

$$p_n \le e^{-\gamma\lambda^n} \tag{29.20}$$

via induction starting from

$$p_1 \le C|f(0)| < \bar{P} = e^{-\gamma\lambda}, \tag{29.21}$$

with choices of γ and λ that guarantee both

$$\sum_{n=1}^{\infty} e^{-\gamma\lambda^n} \le 1 \tag{29.22}$$

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

$$\ln\mu \le \gamma\lambda^{n-1}(2-\lambda) \quad \forall n \ge 1.$$

29.5 The optimal alternative result

For a given μ this is equivalent to

$$\ln\mu \le \gamma\lambda(2-\lambda) \quad \text{and} \quad \lambda \le 2 \tag{29.23}$$

if we make the obvious restriction that γ and λ be positive. Conditions (29.21) and (29.23) suggest $\alpha = \gamma\lambda$ and λ as the more relevant parameter so we have to pick $\alpha > 0$ and $1 < \lambda \le 2$ with

$$\ln\mu \le \alpha(2-\lambda), \quad \sum_{n=0}^{\infty} e^{-\alpha\lambda^n} \le 1 \quad \text{and} \quad \bar{P} = e^{-\alpha} \quad \text{maximal.} \tag{29.24}$$

For $\mu > 1$ the inequalities define a region in the first quadrant of the λ, α-plane bounded by the two curves given by

$$\ln\mu = \alpha(2-\lambda) \ \text{ and } \ \sum_{n=0}^{\infty} e^{-\alpha\lambda^n} = 1, \tag{29.25}$$

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

$$1 = \sum_{n=0}^{\infty} \bar{P}^{(2+\frac{\ln\mu}{\ln\bar{P}})^n} \quad \text{and} \quad \mu \geq 1 \tag{29.26}$$

hits the curve defined by (29.17) in $\mu = 1$ and lies below (29.17) of course, but above (29.19) 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 (29.24) is given by

$$\sum_{n=0}^{\infty} \bar{P}^{2^n}.$$

29.6 A suboptimal alternative result

A more explicit formula is again obtained via a rough estimate

$$\sum_{n=1}^{\infty} e^{-\gamma\lambda^n} \leq \underbrace{\sum_{n=1}^{\infty} e^{-\gamma(1+n(\lambda-1))}}_{\text{a geometric series}} = \frac{e^{-\gamma\lambda}}{1-e^{-\gamma(\lambda-1)}} = \frac{e^{-\alpha}}{1-e^{\gamma}e^{-\alpha}} \tag{29.27}$$

and replacing (29.24) 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

$$1 \leq \mu = \frac{1}{\bar{P}^{2+\frac{\ln\bar{P}}{\ln(\frac{1}{\bar{P}}-1)}}} < \frac{1}{\bar{P}} - 1 \tag{29.28}$$

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 $\bar{P} \to 0$.

29.7 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.

29.8 A much better suboptimal alternative result

Actually the first rough estimate above works better with α dan with γ, as I only noticed May 21. Directly in terms of γ and λ we have

$$\sum_{n=1}^{\infty} e^{-\gamma\lambda^n} = \sum_{n=1}^{\infty} e^{-\gamma\lambda\lambda^{n-1}} \leq \sum_{n=1}^{\infty} e^{-\gamma\lambda(1+(n-1)(\lambda-1))} = \frac{e^{-\gamma\lambda}}{1-e^{-\gamma\lambda(\lambda-1)}} \leq 1 \tag{29.29}$$

if

$$2-\lambda \leq \frac{\ln(e^{\gamma\lambda}-1)}{\gamma\lambda} = \frac{\ln(e^{\alpha}-1)}{\alpha},$$

so that we arrive at

$$\ln\mu \leq \alpha(2-\lambda), \quad \alpha(2-\lambda) \leq \ln(e^{\alpha}-1) \quad \text{and} \quad \bar{P} = e^{-\alpha} \quad \text{maximal.} \tag{29.30}$$

This is the optimal estimate using the Bernoulli type inequality

$$\lambda^n \geq 1 + n(\lambda-1). \tag{29.31}$$

With equality in the final inequality in (29.29) we arrive at

$$\ln\mu \leq \ln(e^{\alpha}-1) = \ln(\frac{1}{\bar{P}}-1),$$

which for $\mu > 1$ coincides with (29.19) and we can forget about the annoying (29.28) above. Note that factoring out another λ in the exponent in (29.29)

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).

30 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

$$N = \frac{\bar{k} - k}{k - m} \tag{30.1}$$

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

$$|S(t)u|_k \leq K_{kl} t^{k-l} |u|_l \quad \text{and} \quad |(I - S(t))u|_l \leq \frac{K^{kl}}{t^{k-l}} |u|_k \tag{30.2}$$

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 (30.2) says that

$$|S(t)|_{kl} \leq K_{kl} t^{k-l} \quad \text{and} \quad |(I - S(t))|_{lk} \leq \frac{K^{kl}}{t^{k-l}}. \tag{30.3}$$

Besides (30.3) 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 $|u|_k \leq 1$.

30.1 The modified scheme

The idea of Nash was to modify Newton's scheme into

$$u_n = u_{n-1} - S(t_{n-1})L(u_{n-1})f(u_{n-1}), \tag{30.4}$$

with a suitable choice of $t_n \to \infty$ as $n \to \infty$. In (30.4) 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 (29.4) is modified when combined with (30.4). We have

$$f(u_n) = \underbrace{f(u_{n-1}) + f'(u_{n-1})(u_n - u_{n-1})}_{\text{vanishes with (29.10)}} + Q_f(u_{n-1}, u_n)$$

$$= \underbrace{f'(u_{n-1})(I - S_{n-1})L(u_{n-1})f(u_{n-1})}_{\text{because of (30.4)}} + Q_f(u_{n-1}, u_n),$$

so that, with

$$p_n = |u_n - u_{n-1}|_k \quad \text{and} \quad q_n = |f(u_n)|_l,$$

the estimate

$$q_n \leq \underbrace{M_1^{lk}|I - S_{n-1}|_{km}|L(u_{n-1})f(u_{n-1})|_m}_{\text{new error like term}} + \frac{1}{2}M_2^{lk}p_n^2 \tag{30.5}$$

holds.

30.2 The new error term

The third factor in the error like term in (30.5) 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 (29.10) and that should guarantee that quadratic term in (30.5) will still allow us to establish a conclusion like

(29.15). 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

$$|L(u)f(u)|_{\bar{m}} \leq A_{\bar{m}\bar{k}}(1 + |u|_{\bar{k}}) \tag{30.6}$$

with

$$\bar{k} - \bar{m} = k - m,$$

which requires an additional estimate for

$$r_n = 1 + |u_n|_{\bar{k}} \tag{30.7}$$

to be used in combination with

$$q_n \leq M_1^{lk} \underbrace{|I - S_{n-1}|_{k\bar{m}}}_{\text{controled by (30.3)}} r_{n-1} + \frac{1}{2} M_2^{lk} p_n^2 \tag{30.8}$$

and the estimate for p_n. Via (30.4) the latter now reads

$$p_n \leq |S_{n-1}|_{km} C_{ml} q_{n-1} \tag{30.9}$$

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 (30.4). 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

$$1 \leq r_n \leq 1 + A_{\bar{m}\bar{k}} \sum_{j=1}^{n} |S_{j-1}|_{\bar{k}\bar{m}} r_{j-1}. \tag{30.10}$$

The "error" terms accumulate but can be kept under control as we shall see below.

The system of inequalities (30.9,30.8,30.10) and initial inequalities for q_0, $r_0 = 1$ and r_1 allows again estimates of the form (29.20), provided $\bar{k} - k = \bar{m} - m$ is sufficiently large in terms of (30.1). The idea is to get the first term in (30.8) 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 (30.9) can be chosen not to large so as still to have (29.20) 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

$$t_{n-1} = e^{\beta\lambda^{n-1}} \tag{30.11}$$

with $\beta > 0$ to be chosen in terms of γ. Note that this gives λ^n in the exponents of the exponential bounds for S_n and $I - S_n$.

Here we choose to keep λ as a parameter in a range as large as possible, like we did in the analysis of (29.10). 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

$$p_n \leq e^{-\gamma\lambda^n} \quad \text{and} \quad r_n \leq e^{\delta\lambda^n} \tag{30.12}$$

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.

30.3 The system of inequalities

With (30.11) we have the system of inequalities

$$p_n \leq K_{km} e^{(k-m)\beta\lambda^{n-1}} C_{ml} q_{n-1}; \tag{30.13}$$

$$q_n \leq M_1^{lk} K^{k\bar{m}} e^{(k-\bar{m})\beta\lambda^{n-1}} A_{\bar{m}\bar{k}} \underbrace{r_{n-1}}_{\leq e^{\delta\lambda^{n-1}}} + \frac{1}{2} M_2^{lk} \underbrace{p_n^2}_{\leq e^{-2\gamma\lambda^n}}; \tag{30.14}$$

$$1 \leq r_n \leq 1 + \underbrace{A_{\bar{m}\bar{k}} K_{\bar{k}\bar{m}}}_{\mu_3} \sum_{j=1}^{n} e^{(\bar{k}-\bar{m})\beta\lambda^{j-1}} \underbrace{r_{j-1}}_{\leq e^{\delta\lambda^{j-1}}}, \tag{30.15}$$

and we aim for a proof of (30.12) via induction, using the underbraced estimates in the three inequalities above as induction hypothesis. In (30.14) 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

$$(\bar{m} - k)\beta \geq \delta + 2\gamma\lambda, \tag{30.16}$$

which says that in the λ, β-plane we must be above a line that comes down as $\bar{m}$ is increased.

Combining the first two inequalities we arrive at

$$p_n \le e^{(k-m)\beta\lambda^{n-1}}(\mu_1 e^{(k-\bar{m})\beta\lambda^{n-2}} r_{n-2} + \mu_2 p_{n-1}^2) \quad r_n \le 1 + \mu_3 \sum_{j=0}^{n-1} e^{(\bar{k}-\bar{m})\beta\lambda^j} r_j, \tag{30.17}$$

the constants μ_{123} given by

$$\mu_1 = \underbrace{K_{km}C_{ml}}_{C} M_1^{lk} \underbrace{K^{k\bar{m}}A_{\bar{m}\bar{k}}}_{A}, \quad \mu_2 = \frac{1}{2}\underbrace{K_{km}C_{ml}}_{C} M_2^{lk}, \quad \mu_3 = \underbrace{K_{\bar{k}\bar{m}}A_{\bar{m}\bar{k}}}_{\bar{A}}. \tag{30.18}$$

30.4 Estimating the increments

Under the assumption that (30.16) holds, the induction hypotheses for p_{n-1} and r_{n-2} produce the desired inequality for p_n from (30.17) 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

$$\ln(\mu_1 + \mu_2) \le \lambda(\gamma(2-\lambda) - (k-m)\beta) \quad \text{with} \quad (k-m)\beta \le \gamma(2-\lambda), \tag{30.19}$$

a sharp upper bound for β that we need to stay away from if we don't want to impose that $\mu_1 + \mu_2 \le 1$.

As sufficient condition for

$$\sum_{n=1}^{\infty} p_n < 1$$

we can use the optimal condition found using Bernoulli's inequality, namely

$$\lambda\gamma(2-\lambda) \le \ln(e^{\gamma\lambda} - 1). \tag{30.20}$$

30.5 Estimating the error terms

For the inductive construction of the upper bound for r_n we set

$$b = (\bar{k} - \bar{m})\beta = (k-m)\beta > 0 \tag{30.21}$$

and conclude from the inequality in (30.17) that (shifting the index)

$$r_n \le 1 + \mu_3 \sum_{j=0}^{n-1} e^{b\lambda^j} r_j \le 1 + \mu_3 \sum_{j=0}^{n-1} e^{b\lambda^j} e^{\delta\lambda^j}$$

in view of the induction assumption for (all) smaller n. Thus we need the inequality

$$1 + \mu_3 \sum_{j=0}^{n-1} e^{(b+\delta)\lambda^j} \le e^{\delta\lambda^n} \tag{30.22}$$

for all $n \ge 2$. Recall that we start with $r_0 = 1 \le e^\delta$ and

$$1 \le r_1 \le e^{\delta\lambda} \quad \text{(and also } p_1 \le e^{\gamma\lambda} \text{ of course)} \tag{30.23}$$

via a smallness assumptions on q_0 still to be discussed.

Dividing by the right hand side, (30.22) is equivalent to

$$e^{-\delta\lambda^n} + \mu_3 (e^{(b+\delta-\delta\lambda)\lambda^{n-1}} + e^{-\delta\lambda^n} \sum_{j=0}^{n-2} e^{(b+\delta)\lambda^j}) \le 1 \tag{30.24}$$

in which we have separated the probably dominant term with $j = n-1$ from the sum. Neglecting the sum in (30.24) a sufficient (and in any case necessary) condition for the induction step to work for all $n \ge 2$ would be that

$$\ln \mu_3 + (b + \delta - \delta\lambda)\lambda^{n-1} \le 0 \quad \text{with} \quad b \le \delta(\lambda - 1), \tag{30.25}$$

so that in particular we now need to impose two inequalities on b, namely

$$b < \delta(\lambda - 1) \quad \text{and} \quad b < \gamma(2 - \lambda), \tag{30.26}$$

the latter being the (strict) inequality from (30.19).

These two bounds severely restrict the bound in (30.16), which in terms of b becomes

$$\frac{\bar{m} - k}{k - m} b \ge \delta + 2\gamma\lambda, \tag{30.27}$$

and this does not really depend on how we turn the necessary condition (30.25) into a sufficient condition, which we do next, rewriting it as

$$e^{-\delta\lambda^n} + \mu_3 (e^{(b+\delta-\delta\lambda)\lambda^{n-1}} + \sum_{j=0}^{n-2} e^{(b+\delta-\delta\lambda^{n-j})\lambda^j}) \le 1.$$

In view of (30.26) and using Bernoulli's inequality (29.31) 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

$$e^{-\delta\lambda^2} + \mu_3 e^{(b+\delta-\delta\lambda)\lambda}(1 + \frac{e^{-(b+\delta)(\lambda-1)}}{1-e^{(\lambda-1)(b+\delta-\delta\lambda^2)}}) \leq 1 \tag{30.28}$$

Note that the first term on the right hand side of (29.31) is essential here. Without this first term the numerator, which is the first term ($j=0$) in the geometric series, would be 1 and we be stuck, as there would be no way to get a statement without an a priori bound on μ_3. We note that in [KP] the proof is without the 1 in (29.31) 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

$$e^{-\delta\lambda^2} + 3\mu_3 e^{(b+\delta-\delta\lambda)\lambda} \leq 1 \quad \text{suffices if} \quad \ln 2 \leq \delta\lambda(\lambda-1)^2, \tag{30.29}$$

[1] And it is not a proof of the theorem actually stated.

and the first inequality in (30.29) certainly holds if it holds with the first exponential replaced by the larger second exponential. Thus we arrive at

$$\ln(1+3\mu_3) \le \lambda(\delta(\lambda-1)-b) \quad \text{and} \quad \ln 2 \le \delta\lambda(\lambda-1)^2 \tag{30.30}$$

as the final condition needed.

30.6 Sufficient conditions for a convergence result

Summing up, with the condition on q_0 still to be imposed we arrive at

$$\lambda\gamma(2-\lambda) \le \ln(e^{\gamma\lambda}-1), \tag{30.31}$$

$$\ln 2 \le \delta\lambda(\lambda-1)^2, \tag{30.32}$$

$$(\bar{m}-k)\beta \ge \delta + 2\gamma\lambda, \tag{30.33}$$

$$(k-m)\beta < \gamma(2-\lambda) \quad \text{and} \quad (\bar{k}-\bar{m})\beta < \delta(\lambda-1) \tag{30.34}$$

as conditions on the parameters that we still have to choose.

The first inequality, (30.31), is to have the sum of the increments, and thereby the solution, bounded by 1 in the l-norm. Of course it can be replaced by just asking that

$$\sum_{n=1}^{\infty} e^{-\gamma\lambda^n} \le 1.$$

The second, (30.32), was a technical condition to bound the sum of the geometric series in (30.28) by 2. The third, (30.33), allows to bound the error term in estimate (30.14) 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 (30.12) for p_n and r_n. Thus, given μ_1, μ_2, μ_3, we need to choose $1 < \lambda < 2$ and $\gamma, \beta, \delta > 0$ such that

$$\lambda(\gamma(2-\lambda) - (m-k)\beta) \ge \ln(\mu_1+\mu_2); \tag{30.35}$$

$$\lambda(\delta(\lambda-1) - (\bar{m}-\bar{k})\beta) \ge \ln(1+3\mu_3). \tag{30.36}$$

After a simultaneous rescaling of γ, β, δ, this is always possible once the first 5 conditions are satisfied. The inequalities in (30.34) 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 (30.18) 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 $f : C^k \to C^l$.

30.7 Sufficient convergence condition on initial value

Finally we examine the initial inequalities we need. For p_1 we need, since $u_0 = 0$, that

$$p_1 = |u_1|_k = |S_0|_{km}|L(0)|_{ml}|f(0)|_l \leq e^{(k-m)\beta}\underbrace{K_{km}C_{ml}}_{C}|f(0)|_l \leq e^{-\gamma\lambda},$$

while via

$$|u_1|_{\bar{k}} \leq |S(0)|_{\bar{k}m}|L(0)|_{ml}|f(0)|_l \leq K_{\bar{k}m}e^{(\bar{k}-m)\beta}C_{ml}|f(0)|_l \leq e^{\delta\lambda}$$

we need

$$1 + \underbrace{K_{\bar{k}m}C_{ml}}_{\bar{C}}e^{(\bar{k}-m)\beta}|f(0)|_l \leq e^{\delta\lambda}$$

for r_1. Thus

$$Cq_0 \leq e^{-(k-m)\beta}e^{-\gamma\lambda} \quad \text{and} \quad \bar{C}q_0 \leq e^{-(\bar{k}-m)\beta}(e^{\delta\lambda} - 1) \tag{30.37}$$

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 30.6 to make the induction steps work in the proof of the desired estimates (30.12) for p_n and r_n.

30.8 The optimal choice of parameters

At this point we compare (30.35) and (30.36) to (29.23). The strict inequalities in (30.34) 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

$$\frac{\bar{m}-k}{k-m}=\frac{\bar{m}-m}{k-m}-1=\frac{\bar{m}-m}{\bar{k}-\bar{m}}-1=\frac{\bar{k}-k}{k-m}-1=N-1 \tag{30.38}$$

has to be sufficiently large for the set of allowable b, as defined by (30.21), to be nonempty. Note that in Nash' strategy to get around the ill-posedness of Newton's method, (30.1) 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 (30.34) equal to one another, so as to maximize the allowable upper bound for β. Thus we choose $1<\lambda<2$ such that

$$\gamma(2-\lambda)=\delta(\lambda-1)\quad\text{whence}\quad\lambda=\frac{2\gamma+\delta}{\gamma+\delta} \tag{30.39}$$

and (30.33,30.34) become

$$\frac{4\gamma^2+3\gamma\delta+\delta^2}{\gamma+\delta}\le(N-1)b<(N-1)\frac{\gamma\delta}{\gamma+\delta} \tag{30.40}$$

for

$$b=(k-m)\beta=(\bar{k}-\bar{m})\beta$$

in terms of γ,δ,N, subject to (30.31,30.32) which reduce to

$$e^{\frac{2\gamma+\delta}{\gamma+\delta}\frac{\delta}{\gamma+\delta}}+1\le e^{\gamma\frac{2\gamma+\delta}{\gamma+\delta}}\quad\text{and}\quad\ln 2\le\delta\,\frac{2\gamma+\delta}{\gamma+\delta}\left(\frac{\gamma}{\gamma+\delta}\right)^2. \tag{30.41}$$

In particular (30.40) requires

$$N>\frac{4\gamma}{\delta}+4+\frac{\delta}{\gamma}\ge 8, \tag{30.42}$$

the minimum 8 being realised by

$$\delta=2\gamma. \tag{30.43}$$

The further choice of parameters depends on the constants which are as indicated in (30.18), at the end of Section 30.6 and in (30.37):

$$C = K_{km}C_{ml};\ \bar{C} = K_{\bar{k}m}C_{ml};\ A = K^{k\bar{m}}A_{\bar{m}\bar{k}};\ \bar{A} = K_{\bar{k}\bar{m}}A_{\bar{m}\bar{k}}; \tag{30.44}$$

$$\mu_1 + \mu_2 = C(M_1 A + \frac{1}{2}M_2); \quad \mu_3 = \bar{A}. \tag{30.45}$$

We collect these constants in one single constant Θ as

$$\Theta = \frac{3}{4}\max(\ln C + \ln(M_1 A + \frac{1}{2}M_2), \ln(1 + 3\bar{A})) \tag{30.46}$$

and, depending on N, the remaining parameters γ, b have to be chosen to control these constants via

$$\Theta \leq \frac{2\gamma}{3} - b \tag{30.47}$$

and

$$\frac{14\gamma}{3} \leq (N-1)b < \frac{2\gamma}{3}(N-1), \quad e^{\frac{8}{9}} + 1 \leq e^{\frac{4\gamma}{3}}, \quad \ln 2 \leq \frac{8\gamma}{27}, \tag{30.48}$$

which is (30.40,30.41) with $\delta = 2\gamma$. The last inequality now implies the one preceding it.

For the initial condition q_0 we arrive via (30.39) 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

$$Cq_0 \leq e^{-\frac{4\gamma}{3}}e^{-b} \quad \text{and} \quad \bar{C}q_0 \leq (e^{\frac{8\gamma}{3}} - 1)e^{-(N+1)b}. \tag{30.49}$$

Setting

$$\rho = \frac{2\gamma}{3}$$

we arrive at

$$\Theta \leq \rho - b, \quad 7\rho \leq (N-1)b, \quad \rho \geq \frac{9}{4}\ln 2,$$

$$\ln C + \ln q_0 \leq -2\rho - b, \quad \ln \bar{C} + \ln q_0 \leq \ln 80 - (N+1)b,$$

as sufficient conditions. Note that we have used the lower bound for ρ to relax the bound on $\bar{C}q_0$.

Choosing

$$N > 8 \quad \text{and} \quad \rho = \frac{N-1}{7} b$$

and using the last lower bound for ρ we arrive at

$$b \geq \max(\frac{63}{4}\frac{\ln 2}{N-1}, \frac{7\Theta}{N-8}) \quad \text{and} \quad q_0 \leq \min(\frac{1}{C} e^{-\frac{2N+5}{7N} b}, \frac{80}{\bar{C}} e^{-(N-1)b}) \tag{30.50}$$

as sufficient conditions, to be used as: given Θ choose $N > 8$ and $b = (k-m)\beta$ sufficiently large to make the condition on q_0 follow and thereby obtain a solution of $f(u) = 0$ with $|u|_k < 1$.

30.9 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 (30.47,30.48). The conditions on γ rewrite as

$$\gamma \geq \max(\frac{3}{4}\ln(\frac{1}{\varepsilon} + e^{\frac{8}{9}}), 2^{\frac{27}{8}}) \sim \varepsilon^{-\frac{3}{4}}$$

as $\varepsilon \to 0$. This forces a larger choice of b and thereby via (30.49) a smaller (exponentially small in terms of ε in fact) bound on q_0 for the Nash scheme to converge within the ball of k-radius ε, as was to be expected of course. The fact that the limit u is a solution of $f(u) = 0$ is immediate from (30.14).

Note that for the standard Newton method the constructed solution of $f(u) = 0$ will have $|u| < \varepsilon$ if we take equalities in (29.30) 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 $q_0 \leq C\bar{P}_\varepsilon$.

31 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 28. Requires a deeper discussion of the smoothing operators used in the proof of Theroem 23.21 and the Fouriertransfrom.

32 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

$$k + \sum_{n=1}^{\infty} \frac{d_n}{10^n} \tag{32.1}$$

is. In (32.1) 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 Σ leidt tot controverses, alsmede het gebruik van het symbool ∞ boven op dat somteken. Wat het eerste betreft zou ik liever zoveel mogelijk over aftelbare sommen willen spreken, maar niet te vergeten dat de term "reeks" nu eenmaal door iedereen gebruikt wordt in zinsdelen als "de som van de reeks".

[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 ∞ 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 (32.1) niet altijd uniek is omdat getallen van de vorm

$$k + \sum_{n=1}^{m} \frac{d_n}{10^n} \tag{32.2}$$

nu eenmaal twee representaties hebben, bijvoorbeeld

$$1 = \frac{9}{10} + \frac{9}{100} + \frac{9}{1000} + \frac{9}{10000} + \frac{9}{100000} + \cdots = \sum_{n=1}^{\infty} \frac{9}{10^n}, \tag{32.3}$$

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?

32.1 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 ∞ 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 (32.1)? Een rekenvraag die we nu rekenend en redenerend willen beantwoorden zonder over verzamelingen van getallen te spreken.

Terzijde, helemaal consequent is de schrijfwijze in (32.1) 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, hondersten, duizendtallen, duizendsten, etc alleen te combineren met de "Arabische" cijfers

$$0, 1, 2, 3, 4, 5, 6, 7, 8, 9.$$

Met

$$\text{tien} = 10, \quad \text{honderd} = 10 \times 10 = 100, \quad \text{duizend} = 10 \times 10 \times 10 = 1000,$$

enzovoorts, en op zijn kop

$$\text{een tiende} = \frac{1}{10}, \quad \text{een honderste} = \frac{1}{100}, \quad \text{een duizendste} = \frac{1}{1000},$$

und so weiter, is het niet heel raar om bijvoorbeeld

$$8 \times 1000 + 7 \times 100 + 6 \times 10 + 5 \times 1 + 4 \times \frac{1}{10} + 3 \times \frac{1}{100} + 2 \times \frac{1}{1000}$$

als corresponderend met een punt op een lijn van hier tot ginder te zien. Bij zo'n punt hoort een getal dat we noteren als

$$8765{,}4321 \tag{32.4}$$

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 (32.4) hoeven verder geen misverstanden te bestaan. Links van de komma tellen de decimalen van rechts af de 1-tallen, 10-tallen, 10×10-tallen, $10 \times 10 \times 10$-tallen, en rechts van de komma vanaf links de $\frac{1}{10}$-tallen, $\frac{1}{10}$-tallen, $\frac{1}{10} \times \frac{1}{10}$-tallen, $\frac{1}{10} \times \frac{1}{10} \times \frac{1}{10}$-tallen, $\frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10}$-tallen. Aan beide kanten van de komma breekt het af, maar rechts is dat niet nodig. In principe kunnen we er nog een rij cijfers achter plaatsen, en een groter getal maken, bijvoorbeeld

$$8765{,}432143214321 \quad \text{of} \quad 8765{,}432199999999999999999999999,$$

[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 (32.4) 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

$$0{,}010101010101010101010101\ldots \tag{32.5}$$

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 (32.5) 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, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, en een kind kan al lerend voor het slapen gaan zien en begrijpen hoe dat zo altijd maar doorgaat als dromenland geen redding brengt. De eindeloze rij van de natuurlijke getallen, beginnend met 1, wordt zo ondubbelzinnig vastgelegd door de aftelling in het decimale stelsel.

Getallen uit die rij kunnen we optellen en vermenigvuldigen (in wezen herhaald optellen). Dat doen we cijferend met de getallen onder elkaar gezet, onhandig vanaf links of handig vanaf rechts per kolom. Die methodes[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

[10] In het PO heeft de onhandige methode veelal de voorkeur gekregen.

[11] Waar de laatste decimaal meestal met meetnauwkeurigheid te maken heeft.

beter vermeden worden, en de niet afbrekende kommagetallen zijn nu net de kommagetallen die we nog missen, en waar we zeker ook mee willen rekenen en cijferen. Dat *cijferen* moet dan wel vanaf links als je er even over nadenkt.

Tellen we bijvoorbeeld het getal dat gerepresenteerd wordt door (32.5) achtennegentig keer bij zichzelf op dan zien we dat 99 keer $\frac{1}{99}$ niet alleen gelijk is aan 1 maar ook aan

$$0{,}\underline{9} = 0{,}999999999999999999999999\ldots,$$

en deze decimale representatie kan de meestal toch onbereikbare eenheid 1 best vervangen, als we afspreken dat decimale representaties van positieve reële getallen naar links altijd, maar naar rechts nooit doorlopen met alleen maar 0-en. Een onhandige conventie wellicht, maar omdat het *cijferen* hier toch alleen maar onhandig vanaf links kan niet eens zo gek eigenlijk.

Bovendien is elk natuurlijk getal nu op een natuurlijke manier ook een echt kommagetal. Met $0{,}\underline{9}$, $1{,}\underline{9}$, $2{,}\underline{9}$, $3{,}\underline{9}$, $4{,}\underline{9}$, $5{,}\underline{9}$, $6{,}\underline{9}$, $7{,}\underline{9}$, $8{,}\underline{9}$, $9{,}\underline{9}$, $10{,}\underline{9}$, $11{,}\underline{9}$, $12{,}\underline{9}$, $13{,}\underline{9}$, $14{,}\underline{9}$, $15{,}\underline{9}$, $16{,}\underline{9}$, $17{,}\underline{9}$, $18{,}\underline{9}$, $19{,}\underline{9}$, $20{,}\underline{9}$, $21{,}\underline{9}$, $22{,}\underline{9}$, $23{,}\underline{9}$, $24{,}\underline{9}$, $25{,}\underline{9}$, $26{,}\underline{9}$, $27{,}\underline{9}$, $28{,}\underline{9}$, $29{,}\underline{9}$, $30{,}\underline{9}$, $31{,}\underline{9}$, $32{,}\underline{9}$, $33{,}\underline{9}$, $34{,}\underline{9}$, $35{,}\underline{9}$, $36{,}\underline{9}$, $37{,}\underline{9}$, $38{,}\underline{9}$, $39{,}\underline{9}$, $40{,}\underline{9}$, $41{,}\underline{9}$, $42{,}\underline{9}$, komen we ook op weg in de rekenkunde. Ook al bekt het niet zo lekker, we gaan rekenen met deze niet afbrekende kommagetallen zoals ze ons gegeven worden.

32.1.1 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

$$\begin{array}{r}
0,9999999\\
0,9999999\\
\hline +\\
\\
1,8000000\\
0,1800000\\
0,0180000\\
0,0018000\\
0,0001800\\
0,0000180\\
0,0000018\\
\hline +\\
\\
1,9999998
\end{array}$$

Immers, met doorlopende negens gaat dit onhandig *cijferen* precies hetzelfde en duurt nauwelijks langer dan hierboven:

$$\begin{array}{r}
0,99999\ldots\\
0,99999\ldots\\
\hline +\\
\\
1,80000\ldots\\
0,18000\ldots\\
0,01800\ldots\\
0,00180\ldots\\
0,000180..\\
0,000018..\\
.....\\
\hline +\\
\\
1,99999\ldots
\end{array} \tag{32.6}$$

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

$$\begin{array}{l} 0,99999\ldots \\ 0,99999\ldots \\ 0,99999\ldots \\ \text{————————} + \\ \\ 2,70000\ldots \\ 0,27000\ldots \\ 0,02700\ldots \\ 0,00270\ldots \\ 0,000270.. \\ 0,000027.. \\ \ldots\ldots\ldots\ldots\ldots \\ \text{————————} + \\ \\ 2,99999\ldots \end{array} \tag{32.7}$$

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

$$\begin{array}{r} 99 \\ 88 \\ 77 \\ \hline 264 \end{array}$$

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

$$\begin{array}{l} 0{,}909090909090909090... \\ \\ 0{,}222222222222222222... \\ \hline\hline \\ 0{,}686868686868686868... \end{array} \quad -$$

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:

$$\begin{array}{l} 0{,}12345... \\ 0{,}99999... \\ \\ \hline\hline \\ 1{,}00000... \\ 0{,}11000... \\ 0{,}01200... \\ 0{,}00130... \\ 0{,}000140.. \\ 0{,}000015.. \\ ..... \\ \hline\hline \\ 1{,}12345... \end{array} \quad + \qquad (32.8)$$

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 32.1.4.

32.1.2 Vermenigvuldigen?

Kunnen we ook vermenigvuldigen? Dit gemene[17] sommetje bijvoorbeeld?

$$
\begin{array}{r}
0,999999.....\\
0,999999.....\\
\hline \times\\
\\
??????
\end{array}
$$

In de vorige subsectie is het gelukt om de som van deze twee getallen kolomcijferend vanaf links zondere hogere wiskunde uit te werken. Kan dat met het produkt ook? We laten ons niet afschrikken en schrijven het produkt cijferend uit, waarbij we het *cijferen* symmetrisch houden in beide factoren, net zoals in (32.6) en (32.7) de uitwerking van de som symmetrisch in de bijdragen van de aparte termen was.

$$
\begin{array}{r}
0,999999.....\\
0,999999.....\\
\hline \times\\
\\
0,810000.....\\
0,081000.....\\
0,081000.....\\
0,008100.....\\
0,008100.....\\
0,008100.....\\
0,000810.....\\
0,000810.....\\
0,000810.....\\
0,000810.....\\
0,0000810...\\
0,0000810...\\
0,0000810...\\
0,0000810...\\
0,0000810...\\
.....\\
\hline \times\\
\\
????????
\end{array}
\qquad (32.9)
$$

[17] Denk nog niet meteen aan $0,\underline{6} \times 0,\underline{6}$.

Dat ziet er een stuk ingewikkelder uit dan (32.6). 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 (32.9) om de som van de getallen in het schema dat begint met

$$\begin{array}{cccccccccccc}
 & & & & & \frac{81}{10^2} & & & & & & \\
 & & & & \frac{81}{10^3} & & \frac{81}{10^3} & & & & & \\
 & & & \frac{81}{10^4} & & \frac{81}{10^4} & & \frac{81}{10^4} & & & & \\
 & & \frac{81}{10^5} & & \frac{81}{10^5} & & \frac{81}{10^5} & & \frac{81}{10^5} & & & \\
 & \frac{81}{10^6} & & \frac{81}{10^6} & & \frac{81}{10^6} & & \frac{81}{10^6} & & \frac{81}{10^6} & & \\
\frac{81}{10^7} & & \frac{81}{10^7} & & \frac{81}{10^7} & & \frac{81}{10^7} & & \frac{81}{10^7} & & \frac{81}{10^7} &
\end{array} \tag{32.10}$$

en naar beneden breder en breder doorloopt. Let wel, de volgorde waarin we cijferend optellen in (32.9) komt overeen met per regel optellen in (32.10) en leidt in de somnotatie tot

$$81 \times \sum_{n=1}^{\infty} \frac{n}{10^{n+1}} \tag{32.11}$$

[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

$$\begin{array}{cccccccccc} & & & & \frac{729}{10^3} & & & & & \\ & & & \frac{729}{10^4} & \frac{729}{10^4} & \frac{729}{10^4} & & & & \\ & & \frac{729}{10^5} & \frac{729}{10^5} & \frac{729}{10^5} & \frac{729}{10^5} & \frac{729}{10^5} & \frac{729}{10^5} & & \\ \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} & \frac{729}{10^6} \end{array} \tag{32.12}$$

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 (32.9) 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	1×9	3×9	5×9	7×9	9	11×9	13×9	15×9	...

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 (32.9) niet boven kan komen, een vraag vergelijkbaar met de minder ruw afgeleide vraag over (32.11). 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 (32.9) 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

$$\begin{array}{cccccccccccccccccc}
8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 & 9 & 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 & 9 \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 8 & 9 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 7 & 8 \\
 & & & & & & & & & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}$$

en alles loopt niet alleen naar rechts maar ook naar beneden door.

Opnieuw optellen per kolom geeft

$$\begin{array}{cccccccccccccccccc}
9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 8 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 8 & 8 \\
 & & & & & & & & 1 & & & & & & & & & & 1 & 1
\end{array}$$

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 (32.12), als de breedte maar niet te snel toeneemt. Dat idee verkennen we in de volgende subsectie, waarin we opnieuw afdwalen van het cijferen.

32.1.3 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

$$\sum_{n=1}^{\infty} \frac{a_n}{10^n} \tag{32.13}$$

[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 (32.13) te praten.

Een noodzakelijke voorwaarde is dat de eindige sommen

$$A_1 = \frac{a_1}{10}, \quad A_2 = \frac{a_1}{10} + \frac{a_2}{10^2}, \quad A_3 = \frac{a_1}{10} + \frac{a_2}{10^2} + \frac{a_3}{10^3}, \quad \ldots \tag{32.14}$$

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 $0{,}d_1\underline{9}$, $0{,}d_1d_2\underline{9}$, $0{,}d_1d_2d_3\underline{9}$, $\ldots$, en voor de bijbehorende

$$D_1 = \frac{d_1}{10}, \quad D_2 = \frac{d_1}{10} + \frac{d_2}{10^2}, \quad D_3 = \frac{d_1}{10} + \frac{d_2}{10^2} + \frac{d_3}{10^3}, \quad \ldots,$$

geldt dat

$$D_1 + \frac{1}{10}, \quad D_2 + \frac{1}{10^2}, \quad D_3 + \frac{1}{10^3}, \quad \ldots$$

bovengrenzen zijn. We kunnen niet uitsluiten dat na verloop van tijd alle d_n nul zijn, maar ze zijn zeker niet allemaal nul.

Kan het zo zijn dat de A_n-tjes niet boven de D_1 uitkomen? Wel, in dat geval zijn alle $A_n < D_1$ (want met $A_n = D_1$ komt een volgende A_n boven D_1), voor alle $n = 1, 2, 3, \ldots$, en was d_1 kennelijk niet minimaal gekozen om alle A_n onder $0{,}d_1\underline{9}$ te hebben. Dus A_n komt wel boven D_1 en blijft dan groter dan D_1. Hetzelfde geldt met het zelfde argument voor D_2, D_3, etcetera.

Als n_1 de eerste n is waarvoor $A_n > D_1$, n_2 de eerste n waarvoor $A_n > D_2$, n_3 de eerste n waarvoor $A_n > D_3$, etcetera, dan is n_2 minstens n_1, n_3 minstens n_2, n_4 minstens n_3, enzovoorts. We concluderen dat voor iedere k geldt dat

$$D_k < A_n < D_k + \frac{1}{10^k} \tag{32.15}$$

voor alle n vanaf $n = n_k$, en dat zou moeten betekenen dat

$$A_n \to D = 0{,}d_1d_2d_3d_4\ldots, \tag{32.16}$$

een nog niet precies gemaakte uitspraak voor een rij breuken A_n, breuken met noemers machten van 10 en $A_n \leq A_{n+1}$, met strikte ongelijkheid voor niet per se alle maar wel willekeurig grote n.

Elke A_n heeft decimalen genummerd door $j = 1, 2, 3, 4, \ldots$. De eerste decimaal kan niet kleiner worden met toenemende n. Dat betekent dat vanaf zekere $n = m_1$ de eerste decimaal van A_n niet meer verandert en gelijk is aan een vast cijfer α_1. Daarna geldt hetzelfde voor de tweede decimaal die vanaf zekere $n = m_2$ (waarbij we m_2 minstens gelijk aan m_1 kunnen nemen) niet meer verandert en gelijk is aan een vast cijfer α_2, enzovoorts.

Deze eigenschap moet voor de niet-dalende rij $A, A_2, A_3, \ldots$ toch wel de enige zinvolle definitie van

$$A_n \to 0{,}\alpha_1\alpha_2\alpha_3\alpha_4\ldots$$

zijn. Graag zouden[23] we nu uit (32.15) concluderen dat

$$0{,}d_1d_2d_3d_4\cdots = 0{,}\alpha_1\alpha_2\alpha_3\alpha_4\ldots,$$

waarbij we opmerken dat de ontwikkeling in het rechterlid bij constructie niet af kan breken maar de ontwikkeling in het linkerlid wel. Het kan dus gebeuren dat de eerste zoveel α_n en d_n hetzelfde zijn, daarna één keer $\alpha_n + 1 = d_n$, en vervolgens alle $d_n = 0$ en alle $\alpha_n = 9$. Hoe het ook zij, de uitdrukking in (32.13) definieert dus ondubbelzinnig een *nul komma minstens nog wat getal*, mits we weten dat alle eindige sommen in (32.14) kleiner zijn dan $0{,}\underline{9}$. Maar wie voor (32.10) en (32.12) meteen ziet dat dat inderdaad zo is mag het zeggen. We zijn er dus nog niet uit wat betreft produkten van positieve kommagetallen.

32.1.4 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 $0{,}9 \times 0{,}\underline{9}$:

$$
\begin{array}{l}
0,999999..... \\
0,9 \\
\text{————————} \times \\
0,810000..... \\
0,081000..... \\
0,008100..... \\
0,000810..... \\
0,0000810... \\
..... \\
\text{————————} \times \\
0,899999.....
\end{array}
$$

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.

32.1.5 Produkten van kommagetallen

Als het bovenstaande eenmaal in in één keer lukt als

$$
\begin{array}{l}
0,999999..... \\
0,9 \\
\text{————————} \times \\
0,899999.....
\end{array}
$$

dan kan daarna

$$
\begin{array}{l}
0,999999..... \\
0,999999..... \\
\text{————————} \times \\
\\
0,899999..... \\
0,089999..... \\
0,008999..... \\
0,000899..... \\
0,000089..... \\
.....
\end{array}
\tag{32.17}
$$

ook, en vervolgens kunnen we dan van boven af de kommagetallen term voor term optellen met wat we kolomcijferend geleerd hebben in sommetjes als (32.8).

De eerste stap is

$$
\begin{array}{l}
0,899999..... \\
0,089999..... \\
\text{————————} + \\
\\
0,899999..... \\
0,009999..... \\
0,080000..... \\
\text{————————} + \\
\\
0,909999..... \\
0,080000..... \\
\\
\text{————————} + \\
\\
0,989999.....
\end{array}
\tag{32.18}
$$

In (32.18) 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 (32.17), waarbij in dit voorbeeld de 8 eentje opgeschoven is naar rechts.

Zo gaat dat verder. Nu we met (32.17) zijn gevorderd tot

$$
\begin{array}{l}
0,999999..... \\
0,999999..... \\
\text{————————} \times \\
\\
0,989999..... \\
0,008999..... \\
0,000899..... \\
0,000089..... \\
.....
\end{array}
\tag{32.19}
$$

zien we dat het patroon zich herhaalt in

$$
\begin{array}{l}
0,989999..... \\
0,008999..... \\
\text{————————} + \\
\\
0,998999.....
\end{array}
$$

met als resultaat de som van de eerste drie kommagetallen in (32.17). De 8 is weer eentje opgeschoven en dat gaat zo door. In de volgende stap zien we

$$
\begin{array}{l}
0,999999..... \\
0,999999..... \\
\text{————————} \times \\
\\
0,998999..... \\
0,000899..... \\
0,000089..... \\
.....
\end{array}
\tag{32.20}
$$

met nu boven de drie nullen na de komma in (32.20) 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.

32.2 Kleinste bovengrenzen

Net als de aftelbare som in (32.1) met $k \geq 0$ is (32.3) een mooi voorbeeld van

$$\sum_{n=0}^{\infty} a_n \tag{32.21}$$

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 (32.21), notatie

$$S = \sum_{n=0}^{\infty} a_n.$$

In het geval van (32.1) 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 (32.21) een eindig getal kan komen is zo vanuit (32.1) vanzelfsprekend, ook al dacht ene Zeno daar destijds anders over. Mooie voorbeelden waarbij er uit de som geen eindig getal komt zijn

$$S = \sum_{n=0}^{\infty} 1 \quad \text{met} \quad S_N = N, \quad \text{en} \quad S = \sum_{n=0}^{\infty} \frac{1}{n}. \tag{32.22}$$

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 (32.22), waarover we opmerken dat

$$\underbrace{1 + \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{>\frac{1}{2}}}_{>1} + \underbrace{\underbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}_{>\frac{1}{2}} + \underbrace{\frac{1}{9} + \frac{1}{10} + \frac{1}{11} + \frac{1}{12} + \frac{1}{13} + \frac{1}{14} + \frac{1}{15} + \frac{1}{16}}_{>\frac{1}{2}}}_{>1},$$

enzovoorts, en zo komen de bijbehorende S_N boven elke $n \in \mathbb{N}$. Ook niet begrensd dus. Maar de som

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

heeft wel een uitkomst[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 32.3 te spreken.

[24] Overtuigd?

[25] Ik meen $\ln 2$.

[26] Kies positieve termen om boven S te komen, dan negatieve om onder S, dan ...

32.3 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 (32.1) een voorbeeld geweest van (32.21) zonder de a priori informatie dat $a_n \geq 0$ maar *wel* met de eigenschap dat

$$\sum_{n=0}^{\infty} |a_n| < \infty, \tag{32.23}$$

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

$$S = \sum_{n=0}^{\infty} a_n, \tag{32.24}$$

en hernummeren van de som verandert niets aan die uitkomst. Reeksen van de vorm (32.21) waarvoor (32.23) 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

$$|S| = |\sum_{n=0}^{\infty} a_n| \leq \sum_{n=0}^{\infty} |a_n|. \tag{32.25}$$

Wat betreft het bewijs van (32.24) gegeven (32.23), de invariantie onder hernummeren en de aftelbare 3-hoeksongelijkheid (32.25): 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 ε-definities van uitspraken als hierboven komen in deze cursus aan de orde op het moment dat dat nodig is. Want ze zijn nodig, bijvoorbeeld om precies te maken dat sommen als (32.24) 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.

32.4 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*

$$(P_N) \implies (P_{N+1}) \tag{32.26}$$

geldt voor alle N waarvoor je hem nodig hebt, namelijk om via herhaald toepassen van de inductiestap (32.26) 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 (32.26) 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 LaTeχ: $\emptyset$. Op het schoolbord liever $\not\emptyset$.
[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

$$(P_1) \wedge (P_2) \wedge \cdots \wedge (P_N) \implies (P_{N+1}) \tag{32.27}$$

is[35]. Zie verder ook Hoofdstuk IV [BE] .

Wat we hier precies met $(A) \implies (B)$ bedoelen moge duidelijk zijn: als de uitspraak (A) waar is dan is ook de uitspraak (B) waar. Hetgeen in onze wiskundige redenaties equivalent is met: als uitspraak (B) niet waar is dan kan uitspraak (A) ook niet waar zijn. Deze logica kan geformaliseerd worden met waarheidstabellen vol nullen en enen opgeleukt met bijzonder fraaie algebra, maar wat dat betreft laten we hier liever de Boole de Boole[36].

Du moment dat er over het bestaan van[37] $\mathbb{N}$ geen twijfel meer is, worden in de verzamelingsleer, bijvoorbeeld zoals in Hoofdstuk V en VI van [BE], $\mathbb{Z}$, $\mathbb{Q}$ en uiteindelijk $\mathbb{R}$ *wiskundig netjes* geconstrueerd. De constructie van $\mathbb{R}$ is in [BE] gebaseerd op de gedachte dat iedere manier om $\mathbb{Q}$ in twee stukken te knippen overeen zou moeten komen met een reëel getal, waarbij de rationale getallen dan wel met de schaar te maken krijgen en de overige getallen niet[38].

Het is instructief om de constructies van $\mathbb{Z}$ en $\mathbb{Q}$ uit $\mathbb{N}$ met elkaar te vergelijken zoals dat gebeurt in [BE]. Die van $\mathbb{Z}$ is inderdaad tamelijk kunstmatig. Die van $\mathbb{Q}$ is echter heel natuurlijk en gebaseerd op hoe je eigenlijk altijd al met de rationale getallen rekende, namelijk als breuken. Breuken met een teller en een noemer. Bijvoorbeeld

$$\frac{14}{333} = \frac{42}{999} = 0.\underline{042}$$

met een streep die aangeeft dat de decimale ontwikkeling van de breuk zich herhaalt. Anders dan gesuggereerd in de in

`http://www.few.vu.nl/~jhulshof/TAL.pdf`

besproken TAL-boekjes van het Freudenthal Instituut doe je echter het rekenen met rationale getallen bij voorkeur niet met zulke decimale ontwikkelingen, maar juist wel met de niet unieke representatie van rationale getallen als quotiënten van gehele getallen, dus in de vorm

$$q = \frac{t}{d}$$

[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

$$\frac{t_1}{d_1} = \frac{t_2}{d_2} \quad \text{als} \quad t_1 d_2 = t_2 d_1.$$

32.5 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

$$[r_n] = \{q_n \in \mathbb{Q} : \, q_n \sim r_n\} \tag{32.28}$$

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[46] 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?

[46] Zie echter ook Sectie 9.8.

32.6 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[47] 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[48] 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[49].

In deze cursus behandelen we ondermeer de analyse die de calculus onderbouwt voor functies $f : I \to \mathbb{R}$ met $I \subset \mathbb{R}$ een interval. Vaak, met $a, b \in \mathbb{R}$, is I daarbij een gesloten begrensd interval

$$I = [a, b] = \{x \in \mathbb{R} : a \leq x \leq b\},$$

of een open begrensd interval

$$I = (a, b) = \{x \in \mathbb{R} : a < x < b\}.$$

We beginnen met integraalrekening, eerst voor monotone functies, zonder over limieten te spreken, en daarna voor uniform continue functies $f : [a, b] \to \mathbb{R}$, waarbij we voor het eerst het limietbegrip tegenkomen en nodig hebben.

Voor zulke functies wordt

$$\int_a^b f(x)\, dx$$

via benaderende sommen gedefinieerd in relatie tot wat de oppervlakte van het gebied ingesloten door $x = a$, $x = b$, $y = 0$ en $y = f(x)$ in het x, y-vlak moet zijn in het geval dat f een positieve functie is. Je zou kunnen zeggen dat dit de eerste *probleemstelling* is in dit boek, geformuleerd in drie punten als:

Teken plaatje!

hoe definieer je de oppervlakte van niet meteen arbitraire verzamelingen;

en hoe reken je die vervolgens uit?

wat kun je vervolgens leren van de oplossing?

Dat laatste doe je dan wellicht zonder meteen een nieuw probleem te willen formuleren. Spelen met de verworven inzichten zonder een concreet doel op zich.

Bijvoorbeeld: met een variabele bovengrens in de integraal ontdekken we de opzet van de differentiaalrekening met behulp van lineaire benaderingen. Die werken we later uit voor *machtreeksen*

$$P(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \cdots ,$$

waarmee we een grote klasse van standaardfuncties tot onze beschikking krijgen, waarvoor "mag dat" vragen kort maar krachtig met "ja natuurlijk" te beantwoorden zijn. Binnen die klasse is de analyse namelijk ondergeschikt aan de algebra, en zodra je die algebra goed begrijpt, voor x^7 of zo, ben je wel klaar en daarmee dient een andere probleemstelling zich aan:

Hoe zit het met al die andere functies?

Als we buiten de klasse van machtreeksen treden verandert alles en moet er gewerkt worden. Dat werk halen we nu naar voren, waar we dat in [HM] zo lang mogelijk uitstelden.

De *middelwaardestelling* blijkt het belangrijkste hulpmiddel om ogenschijnlijk evidente uitspraken ook werkelijk te bewijzen. De uitspraak van die stelling is dat differentiequotiënten als

$$\frac{F(b) - F(a)}{b - a},$$

de richtingcoëfficiënt van

$$\text{de lijn door } (x, y) = (a, F(a) \text{ en } (x, y) = (b, F(b))$$

[47] Van het Latijnse fungor (deponens: een passieve vorm met actieve betekenis).
[48] Vaak slordig: de grafiek $y = f(x)$ in het x, y-vlak.
[49] Zoals wel eens voorgesteld in relatie tot $x \to \frac{1}{x}$ en het rekenonderwijs.

in het x, y-vlak, zelf doorgaans gelijk zijn aan de richtingscoëfficiënt van de raaklijn aan de grafiek van F in een punt met x-waarde tussen a en b, lees: aan de met de differentiaalrekening gedefinieerde

$$\text{afgeleide van } F'(x) \text{ van } F(x) \text{ in een tussenpunt.}$$

Is $F'(x)$ overal tussen a en b gelijk aan nul, dan is $F(x)$ kennelijk constant, aldus de NIET-TRIVIALE STELLING in [HM]. Die STELLING is niet zozeer de oplossing van een probleem, maar formuleert juist iets dat je zeker wil weten bij het oplossen van (bijvoorbeeld) differentiaalvergelijkingen. Het bewijs van de STELLING maakt essentieel gebruik van een fundamentele stelling over het bestaan van convergente deelrijen, die, triviaal[50] 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*[51] 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 Hoofdstuk 9 afgesloten met differentiaalrekening voor integralen met parameters, en partieel integreren en een stelling over Taylorbenaderingen met polynomen.

Daarna nemen we in Hoofdstuk 7.4 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 in Hoofdstuk 9 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[52] voor $n!$ als $n \to \infty$ een mooi voorbeeld[53] is.

In het tweede deel, dat begint met Hoofdstuk 12, 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

[50] Denk ook aan valsspelen met meetwaarden.

[51] Geen sport zonder *techniek*.

[52] De voorbeeldformule met $\sim$ een paar pagina's terug, uit te spreken als "twiddles".

[53] En buitengewoon relevant voor probleemstellingen in de natuurkunde.

maar wel zo snel bewijs van de hoofdstelling van de algebra. Daarna komen functies van $\mathbb{C}$ naar $\mathbb{C}$ en afbeeldingen en functies met meerdere variabelen. Lineaire 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[54] ligt hier voor de hand.

Gewone differentiaalvergelijkingen, al aan de orde geweest in de context van machtreeksen, motiveren het opnemen van 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 uit Hoofdstuk 9 kan dan weer worden overgeschreven. In Hoofdstuk 12 doen we dat al in één moeite door in combinatie de multiplicatorenmethode van Lagrange[55] 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[56]. De meer

[54] Door mijn medestudenten destijds ook wel Bananachalgebra's genoemd.

[55] De eerste stelling die ik ooit zelf aan anderen uitlegde, maar nu heel anders.

[56] Zie www.encyclo.nl/begrip/Submersie èn www.encyclo.nl/begrip/Immersie.

abstracte formulering van de methode van Lagrange in Hoofdstuk 11 is opgenomen for amusement. Ook wat pittiger is de behandeling van tweede orde afgeleiden die we pas in abstracte setting in meer detail doen. Het hoofdresultaat is het Lemma van Morse, Stelling 11.9, 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.

We besluiten met Hoofdstuk 37, eigenlijk een appendix, waarin we 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.

33 Terug naar het platte vlak

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.

33.1 Punten en vectoren in het platte vlak

Exercise 33.1. 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 $(0, 0)$.

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

[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.

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 33.2. 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 punt $\underline{x}$ heeft op elk moment een snelheid $\vec{v}$ die we ons vanwege de fysische

[4] In het *platte* vlak geen probleem maar google op Gauss en kromming.

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 33.3. De lijn l door $\underline{s} \in \mathbb{R}^2$ met richtingsvector $\vec{v} \in \mathbb{R}^2$ kan ook gegeven worden door een vergelijking van de vorm

$$a_1x_1 + a_2x_2 = c$$

voor de punten $\underline{x} = (x_1, x_2)$ op de lijn l. Voor welke lijnen kan dat met $c = 1$? Bepaal voor die lijnen de bijbehorende a_1 en a_2.

Naast de vectoroptelling is in de vectorvoorstelling van een rechte lijn met steunvector $\vec{s}$ en richtingsvector $\vec{v}$ ook de scalaire vermenigvuldiging gebruikt. Voor iedere $t \in \mathbb{R}$ en $\vec{v} \in \mathbb{R}^2$ is $t\vec{v}$ gedefinieerd zoals je zou verwachten. De formule voor $\vec{c} = \vec{a} + \vec{b}$ gaat via $\vec{c} = \vec{x}$, $\vec{a} = \vec{s}$ en $\vec{b} = t\vec{v}$ over in de vectorvoorstelling van de lijn, waarin $\vec{x}$ de met t variërende vector is bij het punt $\underline{x}$.

In de formules mogen alle punten in het platte vlak voorkomen. En alle punten dat zijn alle punten van de vorm $\underline{x} = (x_1, x_2)$ met $x_1, x_2 \in \mathbb{R}$. Het platte vlak past daarmee weliswaar niet in ons universum maar gelukkig wel in ons hoofd, waar het de naam $\mathbb{R}^2$ gekregen heeft, met de 2 van 2-dimensionaal.

Ieder element uit de verzameling $\mathbb{R}^2$ wordt gegeven door een geordend reëel getallenpaar dat we aan kunnen geven met de letters die we willen, en met de notatie die we willen. Nummerend met het alfabet of met indices 1 en 2, achter elkaar of boven elkaar als

$$v_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = v_1\vec{e}_1 + v_2\vec{e}_2$$

geschreven, of eventueel ook als

$$v_1 + iv_2,$$

[5] Vectorvoorstelling van een lijn.

als maar duidelijk is dat v_1 de eerste, en v_2 de twee coördinaat is. De laatste twee vormen suggereren alvast de correspondentie

$$1 \quad \leftrightarrow \quad \vec{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$i \quad \leftrightarrow \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

en de representatie van de complexe getallen $\mathbb{C}$ als het (complexe) vlak $\mathbb{R}^2$ met een wat rare notatie[6].

33.2 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 33.1. 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 33.4. Wat is de kortste afstand tussen $\underline{a} = (1, 1)$ en de lijn met vergelijking $3x_1 + x_2 = 1$?

Exercise 33.5. 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 33.6. Reflecteer[11] op wat het begrip loodrecht met het begrip afstand te maken heeft.

Exercise 33.7. 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 33.8. Een deelverzameling $K \subset \mathbb{R}^2$ heet convex als met elk tweetal punten $\underline{a}$ en $\underline{b}$ in K ook het lijnstuk

$$\{t\underline{a} + (1-t)\underline{b} : \, 0 \leq t \leq 1\}$$

dat $\underline{a}$ en $\underline{b}$ verbindt in K ligt. Kunnen er twee punten in K zijn die $f_O(\underline{x}) = |\underline{x}|$ minimaliseren op K? Maak een plaatje dat je helpt om de vraag te beantwoorden.

33.3 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 33.9. Bewijs dat $|\vec{a} \cdot \vec{b}| \leq |\vec{a}||\vec{b}|$, met andere woorden, dat

$$(a_1 b_1 + a_2 b_2)^2 \leq (a_1^2 + a_2^2)(b_1^2 + b_2^2).$$

Hint: breng alles naar de rechterkant, doe de algebra en herken het kwadraat. Doe vervolgens ook

$$(a_1 b_1 + a_2 b_2 + a_3 b_3)^2 \leq (a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2),$$

en overtuig jezelf ervan dat (even wat combinatoriek)

$$(\sum_{k=1}^{n} a_k^2)(\sum_{k=1}^{n} b_k^2) - (\sum_{k=1}^{n} a_k b_k)^2$$

de som is van $\frac{n(n-1)}{2}$ kwadraten.

Exercise 33.10. 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 33.11. Leid met Opgave 33.9 en Opgave 33.10 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 33.12. 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 33.13. 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?

33.4 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 33.14. 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 33.14 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 33.15. $A \subset \mathbb{R}^2$ 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 33.16. 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 33.17. 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 33.18. 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 33.16 nu een gelijkheid voor alle $\underline{x} \in l$? Gebruik dit en reken $P_l(\underline{b})$ gewoon uit voor gegeven $\underline{b}$.

Exercise 33.19. 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.

33.5 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 33.20. Laat zien dat als $B : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ voldoet aan

$$B(\vec{x}_1 + \vec{x}_2, \vec{y}) = B(\vec{x}_1, \vec{y}) + B(\vec{x}_2, \vec{y});$$

$$B(\vec{x}, \vec{y}_1 + \vec{y}_2) = B(\vec{x}, \vec{y}_1) + B(\vec{x}, \vec{y}_2);$$

$$B(t\vec{x}, \vec{y}) = B(\vec{x}, t\vec{y}) = tB(\vec{x}, \vec{y}),$$

voor alle $t \in \mathbb{R}$ en $\vec{x}, \vec{x}_1, \vec{x}_2, \vec{y}, \vec{y}_1, \vec{y}_2 \in \mathbb{R}^2$, dat B gegeven wordt door[15]

$$B(\vec{x}, \vec{y}) = \sum_{i,j=1}^{2} a_{ij}x_iy_j,$$

en dat $B(\vec{x}, \vec{y}) = B(\vec{y}, \vec{x})$ voor alle $\vec{x}, \vec{y} \in \mathbb{R}^2$ gelijkwaardig is met $a_{ij} = a_{ji}$ voor alle $i, j \in \{1, 2\}$.

Kortom, $B(\vec{x}, \vec{y})$ is van de vorm

$$B(\vec{x}, \vec{y}) = A\vec{x} \cdot \vec{y},$$

waarbij $A : \mathbb{R}^2 \to \mathbb{R}^2$ de lineaire afbeelding is gegeven is door

$$A\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

en de symmetrie van B equivalent is met de symmetrie van de lineaire afbeelding A en de bijbehorende matrix (a_{ij}):

$$B(\vec{x}, \vec{y}) = B(\vec{y}, \vec{x}) \quad \Leftrightarrow \quad A\vec{x} \cdot \vec{y} = \vec{x} \cdot A\vec{y} \quad \Leftrightarrow \quad a_{ij} = a_{ji}$$

[15] Let op: x_i en y_j zijn nu componenten van $\vec{x}$ en $\vec{y}$.

Een symmetrische bilineaire vorm definieert een inwendig produkt als de bijbehorende kwadratische vorm positief definiet is, dat wil zeggen

$$A\vec{x}\cdot\vec{x}>0 \quad \text{as} \quad \vec{x}\neq\vec{0}=\begin{pmatrix}0\\0\end{pmatrix},$$

en in dat geval heet A zelf ook positief[16] definiet[17]. Voorlopig zullen we in de notatie geen onderscheid maken tussen A als lineaire afbeelding en A als matrix. We schrijven dus ook

$$A=\begin{pmatrix} a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix}$$

en spreken over ook positief definiete (symmetrische) matrices.

Kwadratische vormen zijn homogene poynomen van graad twee in de variabelen. Een kwadratische vorm $Q:\mathbb{R}^2\to\mathbb{R}$ wordt dus gegeven door

$$Q(\vec{x})=Q(\underline{x})=Q(x_1,x_2)=q_{11}x_1^2+q_{12}x_1x_2+q_{22}x_2^2=\sum_{1\leq i\leq j=2}^{2}q_{ij}x_ix_j$$

en is altijd te schrijven als $Q(x_1,x_2)=B(\vec{x},\vec{x})=A\vec{x}\cdot\vec{x}$, met

$$a_{ii}=q_{ii} \quad \text{and} \quad a_{ij}=a_{ji}=\frac{1}{2}q_{ij} \quad (i<j).$$

Omdat

$$m=\min_{|\underline{x}|\leq 1}Q(\underline{x})=\min_{|\underline{x}|=1}Q(\underline{x}) \quad \text{and} \quad M=\max_{|\underline{x}|\leq 1}Q(\underline{x})=\max_{|\underline{x}|=1}Q(\underline{x})$$

bestaan als (op de rand aangenomen[18]) minimum en maximum van Q op de gesloten disk gegeven door

$$x_1^2+x_2^2\leq 1,$$

definieert een symmetrische A dus een (niet-standaard) inwendig produkt als $m>0$.

Exercise 33.21. 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 33.22. 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 33.23. Bewijs direct, dus zonder cosinussen en sinussen, dat voor elke symmetrische (hier nog twee bij twee) matrix A geldt dat het maximum μ van de *absolute waarde* van de bijbehorende kwadratische vorm Q op $|\vec{x}| = 1$ wordt aangenomen in een eigenvector, en dat iedere *maximizer* een eigenvector is, bij μ of bij $-\mu$ (of bij allebei in bijzonder gevallen).

Exercise 33.24. De eigenvector in Opgave 33.23 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 33.19 toe[22] op $l = l_1$ en noem $m = l_2$. Laat zien dat A deze l_2 op zichzelf afbeeldt.

33.6 Om te onthouden

Symmetrische twee bij twee matrices komen met paren onderling loodrechte lijnen die we, zo we willen, als nieuwe coördinaatassen kunnen gebruiken. Met in die lijnen (eigen)vectoren $\vec{v}_1$ en $\vec{v}_2$ die onderling loodrecht staan en lengte 1 hebben,

$$\vec{v}_1 \cdot \vec{v}_1 = \vec{v}_2 \cdot \vec{v}_2 = 1 \quad \text{and} \quad \vec{v}_1 \cdot \vec{v}_2 = 0,$$

bij eigenwaarden λ_1 en λ_2,

$$A\vec{v}_1 = \lambda_1\vec{v}_1 \quad \text{and} \quad A\vec{v}_2 = \lambda_2\vec{v}_2.$$

In het bijzondere geval dat A een diagonaalmatrix

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

is, krijgen we als eigenvectoren de standaardbasisvectoren $\vec{e}_1$ en $\vec{e}_2$.

Het andere belangrijke resultaat is dat we op (gesloten) convexe verzamelingen kunnen projecteren, Opgave 33.16. Niet benadrukt nog is wat de essentie was van het bewijs dat je in Opgave 33.23 hebt gegeven. Waar het resultaat in Opgave 33.16 via Opgave 33.14 en een convergente minimaliserende rij tot stand kwam, is in Opgave 33.23 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 33.19. 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?

33.7 Poolcoördinaten in het (complexe) vlak

We besluiten dit hoofdstuk met een korte herhaling van $\mathbb{R}^2$ gezien als de verzameling van complexe getallen $\mathbb{C}$. Het punt $(1,0)$ zien we als het getal 1 en het punt $(0,1)$ als het imaginaire getal i. We introduceren $\mathbb{C}$ door de correspondentie

$$(x,y) \in \mathbb{R}^2 \quad \leftrightarrow \quad z = x + yi = x + iy \in \mathbb{C}$$

met in $\mathbb{C}$ de gebruikelijke rekenoperaties: de complexe optelling en de complexe vermenigvuldiging. Die krijg je door te rekenen met uitdrukkingen als $z = x + iy$ en $c = a + bi$ alsof het eerstegraads polynomen in i zijn, met de afspraak dat $i^2 = -1$. De rollen van i en $-i$ zijn daarbij uitwisselbaar want ook $(-i)^2 = 1$. De coëfficiënten x, y, a, b zijn zelf reëel, en x en a heten de reële delen van respectievelijk z en c. De *imaginaire* delen zijn y en b en zijn net zo reëel als de reële delen.

We gaan ervan uit dat de lezer vertrouwd[23] is met deze complexe getallen en het waarom van de notatie en correspondentie

$$(\cos(t), \sin(t)) \quad \leftrightarrow \quad \exp(it) = \cos(t) + i\sin(t)$$

voor het over de eenheidscirkel bewegende punt $(\cos(t), \sin(t))$.

Die eenheidscirkel wordt gegeven door $|z| = 1$, waarbij de absolute waarde van $z = x + iy$ per definitie gelijk is aan

$$|z| = \sqrt{x^2 + y^2},$$

meestal r genoemd. Voor elke $r > 0$ doorloopt het punt

$$(r\cos(t), r\sin(t)) \quad \leftrightarrow \quad r\exp(it) = r(\cos(t) + i\sin(t))$$

een cirkel met straal r in het al of niet complexe vlak, en de (tijd) t is *per definitie* de hoek in radialen die de met dit punt corresponderende vector maakt met de positieve x-as. Ieder punt in het vlak wordt zo gegeven door een r en een t, en elke 2-vector is van de vorm

$$\vec{x} = \begin{pmatrix} r\cos(t) \\ r\sin(t) \end{pmatrix} = r\begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}, \quad \text{het scalaire product van } r \text{ en } \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix}.$$

Behalve de oorsprong heeft ieder punt $\underline{x}$ en iedere vector $\vec{x}$ een unieke r en een unieke t, waarbij je moet afspreken dat de t-waarden module 2π worden gerekend. En 2π per definitie het reële getal is waarvoor deze laatste karakterisatie correct is. In (tijd) $t = 2\pi$ ga je de cirkel rond.

[23] Zie anderes eventueel [HM, hoofdstuk 11].

Met behulp van deze *poolcoördinaten* volgt voor

$$\vec{c} = p\begin{pmatrix}\cos(s)\\ \sin(s)\end{pmatrix} \quad \text{en} \quad \vec{x} = r\begin{pmatrix}\cos(t)\\ \sin(t)\end{pmatrix}$$

dat

$$\vec{c}\cdot\vec{x} = pr(\cos(s)\cos(t) + \sin(s)\sin(t)) = pr\cos(s-t),$$

het product van de twee lengten en de cosinus van wat de *ingesloten hoek* wordt genoemd. Is die hoek gelijk aan $\pm\frac{\pi}{2}$ dan is het inproduct nul en staan de vectoren loodrecht op elkaar.

Exercise 33.25. 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 33.26. 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 $z = x + iy$.

34 Into Hilbert space

In $\mathbb{R}^3$, $\mathbb{R}^4$,.... we can do the same algebra as in Chapter 33 for $\mathbb{R}^2$. In $\mathbb{R}^3$ we have

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} = ax + by + cz \quad \text{or} \quad \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \sum_{i=1}^{3} a_i x_i,$$

in $\mathbb{R}^{42}$

$$\vec{a} \cdot \vec{x} = \sum_{i=1}^{42} a_i x_i \quad \text{for} \quad \vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{42} \end{pmatrix} \quad \text{and} \quad \vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{42} \end{pmatrix},$$

in $\mathbb{R}^\infty$, dropping the arrows,

$$a \cdot x = \sum_{i=1}^{\infty} a_i x_i \quad \text{for} \quad a = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{pmatrix}$$

Points or vectors, we maken het onderscheid in de notatie tussen x als $\vec{x}$ en $\underline{x}$ steeds vaker alleen als het echt nodig is[1].

De laatste uitdrukking definieert $a \cdot x$ soms wel en soms niet, want zonder restricties op $a, x \in \mathbb{R}^\infty$ kan het met

$$a \cdot x = \sum_{i=1}^{\infty} a_i x_i = \sum_{i \in \mathbb{N}} a_i x_i$$

alle kanten op. En het wordt nog spannender als we de $i \in \mathbb{N}$ in vervangen door bijvoorbeeld $t \in \mathbb{R} = (-\infty, \infty)$ of[2] $t \in [-\pi, \pi]$. In zulke gevallen is ook $\sum$ aan vervanging toe. Overaftelbare[3] sommen gaan niet werken en sommeren moet hier dus wel integreren worden, wat anders? Met de notatie $t \to a_t = a(t)$ en $t \to x_t = x(t)$ voor functies $a : \mathbb{R} \to \mathbb{R}$ en $x : \mathbb{R} \to \mathbb{R}$ wordt een voor de hand liggend inwendig produkt van de functies a en x nu gedefinieerd met behulp van de formule

$$a \cdot x = \int_{-\infty}^{\infty} a(t)x(t)dt,$$

[1] Als we niet meer recht kunnen praten wat krom is en rechte pijltjes niet passen.
[2] Denk ook aan de poolcoördinaten in het platte vlak.
[3] Waarom niet?

waarin *alle* $a(t)$ en $x(t)$ waarden gelijkwaardig voorkomen maar, paradoxaal wellicht, individueel geen invloed hebben op de uitkomst van de integraal die $a \cdot x$ definieert. Ook met die uitkomst kan het, bijvoorbeeld voor continue functies, alle kanten op, net als met $a \cdot x$ voor $a, x \in \mathbb{R}^\infty$.

Voor 2π-periodieke continue functies heeft deze integraalformule geen betekenis maar de formule

$$a \cdot x = \int_{-\pi}^{\pi} a(t)x(t)dt$$

vaak wel, het standaard inwendig produkt waarmee we werken in het geval van 2π-periodieke functies a en x, (goed) gedefinieerd voor continue functies als gewone Riemann integraal[4].

Exercise 34.1. Voor $n = 1, 2, 3, \ldots$ zijn de 2π-periodieke functies c_n en s_n gedefinieerd door $c_n(t) = \cos(nt)$ en $s_n(t) = \sin(nt)$. Bereken nog eens $c_n \cdot c_m$, $c_n \cdot s_m$, $s_n \cdot s_m$, voor $m, n = 1, 2, 3, \ldots$.

Je ziet het niet meteen, maar al deze cosinussen en sinussen staan "loodrecht" op elkaar, en ze hebben ook allemaal dezelfde "lengte", de wortel uit het inprodukt van de functie met zichzelf.

Exercise 34.2. Er is nog een functie die loodrecht staat op al deze cosinussen en sinussen. Welke functie?

34.1 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.

34.2 Symmetrische matrices

Net als in de twee-dimensionale context heeft iedere symmetrische $n \times n$ matrix

$$A = (a_{ij})_{i,j=1,\ldots,n}$$

(een basis van) eigenvectoren $\vec{v}_1, \ldots, \vec{v}_n \in \mathbb{R}^n$ met

$$\vec{v}_i \cdot \vec{v}_j = \delta_{ij} = \begin{cases} 1 & \text{als } i = j \\ 0 & \text{als } i \neq j, \end{cases}$$

bij (reële) eigenwaarden

$$\lambda_1, \ldots, \lambda_n,$$

die gemaakt worden door Opgave 33.23 herhaald toe te passen. Dit geeft in ieder stap zowel een nieuwe λ_k als een nieuwe $\vec{v}_k$ via

$$|\lambda_k| = \max_{\substack{\vec{v} \cdot \vec{v} \leq 1 \\ \vec{v} \cdot \vec{v}_1 = \cdots = \vec{v} \cdot \vec{v}_{k-1} = 0}} |A\vec{v} \cdot \vec{v}|,$$

waarbij $k = 1, \ldots, n$.

Details van deze constructie komen aan de orde in de context van de standaard aftelbaar oneindig-dimensionale Hilbertruimte die we in $\mathbb{R}^\infty$ maken. *Daarvóór* komen we *voor het eerst* over abstracte Hilbertruimten[8] H te praten die we zo snel mogelijk gelijk[9] praten aan het standaardvoorbeeld in $\mathbb{R}^\infty$, onder de aanname van separabiliteit van H: het bestaan van een rij $x_1, x_2, x_3, \ldots$ in H die als limieten van zijn convergente deelrijen *alle* elementen in H heeft.

Kortom, in termen van de dimensie van onze ruimten maken we in één keer de stap van $n = 2$ en concreet ($\mathbb{R}^2$) naar $n = \infty$ en abstract (niet concreet). Let wel, dat kan *alleen* voor ruimten met een inwendig produkt.

[8] Inprodukt ruimten waarin Cauchy rijtjes convergent zijn.

[9] Lees: isomorf.

34.3 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 34.3. 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 34.4. 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 33.14. 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 33.9.

Exercise 34.5. 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 34.4 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 34.6. 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 34.6. Het standaardinprodukt in $\mathbb{R}^2$ geeft via

$$\begin{pmatrix} a \\ b \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \underbrace{ax + cy = \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}}_{matrix\ notatie}$$

een representatie van de lineaire functie

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \underbrace{ax + cy = \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}}_{matrix\ notatie}$$

en omgekeerd is *iedere* lineaire[14] functie van deze vorm. De correspondentie

$$\begin{pmatrix} a \\ b \end{pmatrix} \quad \leftrightarrow \quad \begin{pmatrix} a & b \end{pmatrix}$$

[11] $x \cdot y = 0$ voor alle y in L wordt kort geschreven als: $x \cdot y = 0 \ \forall\, y \in L$.
[12] Het kwadratische stuk kan algemener, het lineaire stuk niet!
[13] K is i.h.a. *niet* begrensd, laat staan rijkompakt (iets met convergente deelrijen).
[14] De constante functie is NIET lineair, tenzij de constante nul is.

is evident bijectief en lineair. Links staat een 2-vector, rechts een 1 bij 2 matrix waarmee een lineaire afbeelding van $\mathbb{R}^2$ naar $\mathbb{R}$ gemaakt wordt.

In een willekeurige Hilbertruimte is er a priori geen matrixnotatie voor het maken van lineaire afbeeldingen. Welke lineaire afbeeldingen hebben we op zo'n Hilbertruimte H als van H verder niets gegeven is, behalve dan dat het een H is? Wel, in ieder geval is voor elke $y \in H$ de afbeelding $\phi_y : H \to \mathbb{R}$ gedefinieerd door[15]

$$x \to y \cdot x = \phi_y(x) = \phi_y x =< \phi_y, x > .$$

Kijk even goed, in de 1 na laatste notatie hebben we de haken weggelaten, zoals vaker bij lineaire afbeeldingen[16], en in de laatste staan ϕ_y en x zo te zien *gelijkwaardig* tussen strange brackets[17], waarbij net als in $y \cdot x$ de rollen van de tegenspelers verwisseld kunnen worden. Dualiteit heet dat met een mooi woord.

Voorlopig gebruiken we de notatie die het meest op de schoolnotatie lijkt. Een functie f van x, in dit geval ϕ_y, maak je expliciet[18] via $f(x)$, in dit geval $\phi_y(x) = y \cdot x$. Dat lijkt expliciet maar is het natuurlijk niet echt als we niet zeggen wat H is. Expliciet of niet, uit de ongelijkheid van Cauchy-Schwarz volgt nu dat

$$|\phi_y(x)| = |y \cdot x| \leq |y||x|$$

en dus ook, *vanwege de lineairiteit* , dat

$$|\phi_y(x_1) - \phi_y(x_2)| = |y \cdot (x_1 - x_2)| \leq L|x_1 - x_2| \quad \text{with} \quad L = |y|.$$

Een reëelwaardige functie f op een vectorruimte met een norm, die voldoet aan

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2|$$

voor alle x_1 en x_2 in die genormeerde vectorruimte, heet *Lipschitz continu*. Een mooi begrip, dat differentiaalrekening noch epsilons en delta's nodig heeft.

Exercise 34.7. 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 34.8. Voor elke genormeerde ruimte X vormen de Lipschitz continue functies $f : X \to \mathbb{R}$ een vectorruimte $Lip(X)$ met de vectorbewerkingen gedefinieerd door

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

Voor elke f is de kleinste L zoals boven per definitie een soort norm van f, die we noteren met $L = [f]_{Lip}$. Waarom definieert

$$f \to [f]_{Lip}$$

geen norm op $Lip(X)$? En waarom wel op

$$Lip_0(X) = \{f \in Lip(X) : \, f(0) = 0\}?$$

Bewijs dat met deze norm elke Cauchyrij $f_n \in Lip_0(X)$ convergent is. Hint: bewijs dit eerst voor $X = \mathbb{R}$ en schrijf je bewijs nog een keer over voor $X = X$.

Klein probleempje is natuurlijk dat er misschien maar weinig van die al of niet lineaire Lipschitz functies op X zijn, als je verder niks van X weet. Maar op H is dat probleempje er niet. Elke $y \in H$ geeft je een ϕ_y in $Lip_0(H)$ die nog lineair is ook, en je ziet meteen wat de kleinste L is: op zijn hoogst $|y|$ en kleiner kan niet, vul maar $x = y$ in. Dat betekent dat we met $y \to \phi_y$ een afbeelding

$$\Phi := H \to Lip_0(H)$$

hebben, en het beeld van Φ is bevat in H^*, de (genormeerde) ruimte van Lipschitz continue *lineaire* functies $f : H \to \mathbb{R}$, en Φ is zelf weer lineair[19]:

Exercise 34.9. Verifieer dat $\Phi : H \to H^*$ voldoet aan

$$\Phi(x_1 + x_2) = \Phi(x_1) + \Phi(x_2) \quad \text{and} \quad \Phi(tx) = t\Phi(x)$$

voor alle $t \in \mathbb{R}$ en $x, x_1, x_2 \in H$, en dat $[\Phi(x)]_{Lip} = |x|$.

De vraag nu is of Φ surjectief is: is elke $f \in H^*$ van de vorm ϕ_y? Bekijk daartoe[20]

$$N_f = \{x \in H : f(x) = 0\}.$$

[19] Nu maar eens de axioma's noemen en verifiëren.

[20] We schrijven nu N_f i.p.v. $N(f)$, t.b.v. het onderscheid tussen f en P_L.

Exercise 34.10. Bewijs dat $N_f \subset H$ een gesloten lineaire deelruimte is.

In het bijzonder bestaat nu dankzij Opgave 34.5 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 34.11. Bewijs dat $M = N(P_{N_f})$ een gesloten lineaire deelruimte is die gegeven wordt door $M = \{te : t \in \mathbb{R}\}$ waarin $e \in N_f^{\perp}$ met $|e| = 1$. Laat zien dat f een veelvoud is van ϕ_e: $f(x) = f(e)e \cdot x$.

Exercise 34.12. Leg uit waarom met het resultaat in Opgave 34.10 de afbeelding $\Phi : H \to H^*$ een lineaire isometrie is.

Lineaire isometriën zijn de mooiste continue afbeeldingen die er bestaan. De inverse van Φ wordt de Riesz representatie van H^* genoemd, en via deze isometrie erft H^* ook het inwendig produkt van H: de reële Hilbertruimten H en H^* zijn als Hilbertruimten hetzelfde, al is het in concrete situaties niet altijd even handig om hier de nadruk op te leggen.

Het resultaat geldt zonder enige verdere restrictie op H en het is ook niet nodig om aan te nemen dat H separabel is. We noteren de inverse van Φ als

$$R_H,$$

met de ruimte H als subscript aan $R = \Phi^{-1}$ gehangen. Het domein van R_H is zo de deelruimte

$$H^* \subsetneq Lip_0(H).$$

Exercise 34.13. Gebruik Opgave 34.4 om aan te tonen dat er plenty niet-lineaire functies in $Lip_0(H)$ zijn.

34.4 De standaard Hilbertruimte

De wat informeel geintroduceerde ruimte $\mathbb{R}^\infty$ bestaat uit alle functies $f, x, a : \mathbb{N} \to \mathbb{R}$, hoe je ze ook wil noemen[21]. We kunnen deze functies zien als kolomvectoren $\vec{f}$ met daarin de waarden van f, al protesteert LaTeχ daarbij zo te zien een beetje. Helemaal op dezelfde hoogte lukt typografisch niet,

$$f = \begin{pmatrix} f(1) \\ f(2) \\ f(3) \\ \vdots \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{pmatrix} = \vec{f},$$

en ook voor functies $f, x, a : \{1, 2, 3\} \to \mathbb{R}$ oogt

$$f = \begin{pmatrix} f(1) \\ f(2) \\ f(3) \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} = \vec{f}$$

niet echt lekker.

Wiskundig gezien praten we de facto over functies $f : A \to \mathbb{R}$, waarbij A hier een *discrete* verzameling is, en de verzameling van deze functies wordt ook wel genoteerd als $\mathbb{R}^A$. Als je $n \in \mathbb{N}$ gedefinieerd hebt als een wat rare verzameling, via[22]

$$1 = \{\emptyset\},\ 2 = \{\emptyset, \{\emptyset\}\},\ 3 = \{\emptyset, \{\emptyset, \{\emptyset\}\}\}, \ldots$$

of zoiets, dan is de notatie voor $\mathbb{R}^A$ consistent[23] met die voor $\mathbb{R}^n$.

Elke $f \in \mathbb{R}^A$ heeft als Pythagoras norm

$$|f| = \sqrt{\sum_{a \in A} |f(a)|^2},$$

hetgeen voor $A = \{1, 2, 3\}$ overeenkomt met de Euclidische lengte

$$\sqrt{f_1^2 + f_2^2 + f_3^3}$$

van de vector $\vec{f}$ hierboven.

Het ligt voor de hand om $\mathbb{R}^A$ en $\mathbb{R}^B$ als dezelfde ruimte te zien als er een bijectie bestaat tussen A en B. Voor eindige verzamelingen A is de Pythagoras norm natuurlijk op heel $\mathbb{R}^A$ gedefinieerd, maar als A oneindig veel elementen bevat[24] dan is dat niet meer het geval.

[21] Het zijn er meer dan 26.

[22] Ik schuif wat ik naief in Halmos las wellicht eentje op, boekje niet bij de hand.

[23] Let wel, $0 = \emptyset$ doet niet mee.

[24] We zeggen dan gemakshalve dat A oneindig is.

Exercise 34.14. 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 (Lebesque) 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 34.15. Bewijs dat $l^{(2)}$ *volledig* is. Dat wil zeggen, laat zien dat Cauchy rijtjes in $l^{(2)}$ convergent zijn met limiet in $l^{(2)}$.

[25] Dit heet onvoorwaardelijke convergentie.

[26] Ander voorbeeld: $\mathbb{R}$ modulo 2π, de facto de eenheidscirkel in $\mathbb{R}^2$.

[27] Zie "Wiskunde in je vingers" van H&M voor snelle intro maattheorie.

Als we $\mathbb{N}$ zien als maatruimte voorzien van de telmaat dan wordt kleine l weer groot. En met recht, want iedere separabele Hilbertruimte H is met $l^{(2)}$ te identificeren[28]. Hoe gaat dat? Wel, neem een rijtje $a_1, a_2, a_2, \ldots$ in H dat als limietpunten alle elementen van H heeft. Zet

$$e_1 = \frac{1}{|a_1|} a_1$$

als $a_1 \neq 0$ maar gooi a_1 weg als $a_1 = 0$. Hernummer in dat geval de rij en herhaal deze stap, net zolang[29] tot je een $a_1 \neq 0$ hebt. Stel vervolgens

$$y_2 = a_2 - (a_2, e_1)e_1 \quad \text{and} \quad e_2 = \frac{1}{|y_2|} y_2$$

als $y_2 \neq 0$, maar gooi a_2 weg als $y_2 = 0$ en hernummer in dat geval weer de rij. Herhaal deze stap, net zolang tot je een $y_2 \neq 0$ hebt en daarmee ook een e_2. Stel vervolgens

$$y_3 = a_3 - (a_3, e_2)e_2 - (a_3, e_1)e_1 \quad \text{and} \quad e_3 = \frac{1}{|y_3|} y_3,$$

als $y_3 \neq 0$, maar $\ldots$, enzovoorts. Dit produceert een rij $e_1, e_2, e_3, \ldots$ van vectoren waarvoor

$$(e_i, e_j) = \delta_{ij},$$

en deze vectoren spannen een lineaire deelruimte op in H.

Exercise 34.16. 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...

35 A function space for Fourier series

Under construction.

35.1 Een Hilbert ruimte voor (periodieke) functies?

De abstracte constructie van een Hilbertruimte H uit $C(\mathbb{R}_{2\pi})$ met zijn inprodukt is dat je een 2-Cauchyrijtje $f_1, f_2, \ldots$, een rijtje waarvoor geldt dat

$$|f_n - f_m|_2 \to 0$$

als $m, n \to \infty$, ziet als een benadering van een functie f in de te construeren H. Dit is analoog aan de vertrouwde gewoonte om decimale of binaire ontwikkelingen te zien als benaderingen van reële getalen, waarbij verschillende ontwikkelingen hetzelfde reële getal kunnen maken. Bij die reële getallen denk je doorgaans aan punten op een getallenlijn, al hoeft dat natuurlijk niet: als je het jezelf moeilijk wil maken kan de abstracte constructie van $\mathbb{R}$ prima zonder.

De standaardvisualisatie van een functie is de grafiek van die functie, in het geval $f : \mathbb{R} \to \mathbb{R}$ een deelverzameling G van

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

met de eigenschap dat

$$\forall x \in \mathbb{R}\ \exists_1 y \in \mathbb{R} : (x, y) \in G.$$

In deze notatie is $\forall$ *short* voor *voor alle* , en staat $\exists_1$ voor *er is precies één*[1]. Die ene y in de uitspraak wordt in deze context bij afspraak genoteerd als $f(x)$, en daarmee is de formele definitie van een grafiek in $\mathbb{R}^2$ de facto equivalent met de definitie van een functie van $\mathbb{R}$ naar $\mathbb{R}$. Omdat $\mathbb{R}^2$ als vlak in ons hoofd past, en de grafiek G van $f : \mathbb{R} \to \mathbb{R}$ daar weer een deelverzameling van is, zien we de grafiek G van f nu in het platte xy-vlak als de verzameling gegeven door $y = f(x)$.

Het ligt voor de hand om de abstracte constructie van H uit $C(\mathbb{R}_{2\pi})$ te zien als gebeurende in het platte vlak $\mathbb{R}^2$, waarbij de grafiek G van een functie $f : \mathbb{R}_{2\pi} \to \mathbb{R}$ dus de eigenschap moet hebben dat

$$\frac{x_1 - x_2}{2\pi} \in \mathbb{Z} \implies f(x_1) = f(x_2),$$

[1] $\exists$ staat voor *er bestaat een.*

hetgeen overeenkomt met het oprolbaar[2] zijn van het oneindige platte vlak tot een cylinder waarin de grafiek G keurig over zichzelf heen ligt.

Met of zonder voorstelling, twee verschillende 2-Cauchyrijtjes $f_1, f_2, \ldots$ en $g_1, g_2, \ldots$ in $C(\mathbb{R}_{2\pi})$ moeten hetzelfde element uit de te maken H zijn als geldt dat

$$|f_n - g_n| \to 0$$

voor $n \to \infty$. Opgave 36.1 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 35.1. Maak Opgave 36.1 nog een keer maar anders. Teken de grafieken van een rij functies $\tilde{Z}_n \in C(\mathbb{R}_{2\pi})$ waarvoor geldt dat $(\tilde{Z}_n - Z, \tilde{Z}_n - Z) \to 0$ als $n \to \infty$. Kies de rij functies $\tilde{Z}_1, \tilde{Z}_2, \ldots$ nu zo dat voor alle $n \in \mathbb{N}$ geldt dat $\tilde{Z}_n(0) = 1$.

Exercise 35.2. Maak Opgave 35.1 maar nu met $\tilde{Z}_n(0) = 0$.

Exercise 35.3. Maak Opgave 35.2 maar nu met $\tilde{Z}_n(0) = 2$.

Exercise 35.4. Maak Opgave 35.2 maar nu met $\tilde{Z}_n(0) = n$.

Exercise 35.5. Laat in Opgaven 35.1,35.2,35.3,35.4 hierboven zien dat $|Z_n - \tilde{Z}_n| \to 0$ als $n \to \infty$, waarbij Z_n is als in Opgave 35.1.

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.

[2]Stel je de problemen bij het oprollen even voor....

Exercise 35.6. 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 35.6, 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 35.7. 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

$$\sum_{n=1}^{\infty} a_n c_n + \sum_{n=1}^{\infty} b_n s_n \tag{35.1}$$

kunnen nemen, met coefficiënten als in Opgave 35.7, 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 35.8. Maak een Cauchyrij $f_1, f_2, \ldots$ die naar de nulfunctie **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 $C(\mathbb{R}_{2\pi})$.

35.2 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[3] 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[4]. 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 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 36.1. 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.

[3] Wat *bijna* betekent is de hamvraag.

[4] Waarom eigenlijk? Wel, we zijn uitgegaan van continue functies.

De functie Z heeft in ieder geheel veelvoud van 2π een sprong. De eveneens oneven blokfunctie $\mathit{blok} \in H$, gedefinieerd door

$$\mathit{blok}(x) = \begin{cases} 1 & \text{als } x \in (0,\pi) \ ; \\ -1 & \text{als } x \in [-\pi, 0), \end{cases}$$

heeft in ieder geheel veelvoud van π een sprong. De even *kartelrandfunctie* $\mathit{Ka} \in H$ daarentegen, gedefinieerd door

$$\mathit{Ka}(x) = \begin{cases} \frac{\pi}{2} - x & \text{als } x \in (0,\pi) \ ; \\ \frac{\pi}{2} + x & \text{als } x \in [-\pi, 0), \end{cases}$$

heeft geen sprongen als we de definitie van Ka uitbreiden met $\mathit{Ka}(2\pi n) = \frac{\pi}{2}$ in de gehele veelvouden $2\pi n$ van 2π ($n \in \mathbb{Z}$). Al deze functies zijn instructief als voorbeeld bij de vraag of ze te schrijven zijn als een oneindige som van de vorm (35.1). Met name de zaagtand is een bron van leerzaam vermaak zoals we zullen zien.

Exercise 35.9. 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

$$f = a_0 \frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} a_n \frac{c_n}{\sqrt{\pi}} + \sum_{n=1}^{\infty} b_n \frac{s_n}{\sqrt{\pi}} \tag{35.2}$$

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 (35.2) 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 (35.2) maken.

35.3 Fourierreeksen

De functie

$$f_7(x) = \sin x - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \frac{\sin 4x}{4} + \frac{\sin 5x}{5} - \frac{\sin 6x}{6} + \frac{\sin 7x}{7}$$

is periodiek met periode 2π. Op het interval $(-\pi, \pi)$ ligt de grafiek van f_7 vlakbij de grafiek van de oneven functie $f(x) = \frac{1}{2}x$. Kennelijk is

$$\frac{x}{2} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\sin kx}{k}, \tag{35.3}$$

maar alleen voor $-\pi < x < \pi$.

Exercise 35.10. Er is een verband met machtreeksen: het rechterlid in (35.3) is het imaginaire deel van

$$\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \zeta^k, \quad \zeta = e^{ix}.$$

Bepaal de som van deze machtreeks voor $|\zeta| < 1$. Hint: differentieer naar ζ, sommeer en primitiveer.

Exercise 35.11. Volgens de complexe versie van het criterium van Leibniz convergeert de machtreeks in in Opgave 35.10 voor alle ζ met $|\zeta| = 1$ behalve $\zeta = -1$. Aangenomen dat de somfunctie die je in Opgave 35.10 hebt gevonden ook geldig is voor alle ζ met $|\zeta| = 1$ waar de reeks convergeert, verifieer (35.3). Op de complexe versie van het criterium van Leibniz en de aanname komen we nog terug in Sectie **??**.

De reeks in (35.3) heet een Fouriersinusreeks. Maken we van de "minnen" plussen dan vinden we dat de grafiek van de functie

$$h_7(x) = \sin x + \frac{\sin 2x}{2} + \frac{\sin 3x}{3} + \frac{\sin 4x}{4} + \frac{\sin 5x}{5} + \frac{\sin 6x}{6} + \frac{\sin 7x}{7}$$

op het interval $(0, 2\pi)$ vlakbij de grafiek van de functie $f(x) = \frac{\pi - x}{2}$ ligt.

De functie

$$g_7(x) = \cos x - \frac{\cos 2x}{4} + \frac{\cos 3x}{9} - \frac{\cos 4x}{16} + \frac{\cos 5x}{25} - \frac{\cos 6x}{36} + \frac{\cos 7x}{49}$$

heeft een grafiek op het interval $(-\pi, \pi)$ vlakbij de grafiek van de even functie

$$g(x) = \frac{\pi^2}{12} - \frac{x^2}{4}$$

ligt. Kennelijk is

$$\frac{x^2}{4} = \frac{\pi^2}{12} + \sum_{k=1}^{\infty} (-1)^k \frac{\cos kx}{k^2}.$$

Het rechterlid, inclusief de constante term, heet een Fouriercosinusreeks. Invullen van $x = 0$ geeft

$$\frac{\pi^2}{12} = 1 - \frac{1}{4} + \frac{1}{9} - \frac{1}{16} + \frac{1}{25} - \cdots.$$

We zullen zien dat bij even 2π-periodieke functies Fouriercosinusreeksen horen, en bij oneven 2π-periodieke functies Fouriersinusreeksen. Omdat elke functie te splitsen is in een even en een oneven functie,

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

hoort zo bij een algemene 2π-periodieke functie de som van een Fouriercosinusreeks en een Fouriersinusreeks. Zo'n som heet een Fourierreeks. Schrijven we de cosinussen en sinussen uit in complexe e-machten,

$$\cos x = \frac{e^{ix} + e^{ix}}{2}, \quad \sin x = \frac{e^{ix} - e^{ix}}{2i},$$

dan wordt een algemene Fourierreeks een reeks van de vorm

$$\sum_{k=-\infty}^{\infty} c_k e^{ikx} = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx) \tag{35.4}$$

Uiteindelijk zullen we er voor kiezen om met de complexe vorm te werken, het linkerlid in (35.4) dus, een Laurentreeks in $\zeta = e^{ix}$.

Uit de Cauchy integraal formules hebben we Laurentreeksen van de vorm

$$\sum_{k=-\infty}^{\infty} c_k \zeta^k \tag{35.5}$$

zien verschijnen voor complex differentieerbare functies $f(\zeta)$ op een annulus

$$\{\zeta \in \mathbb{C} : R_1 < |\zeta| < R_2\},$$

waarbij het plusstuk

$$\sum_{k=0}^{\infty} c_k \zeta^k$$

convergent is voor $|\zeta| < R_2$ en het minstuk

$$\sum_{k=1}^{\infty} c_{-k} \zeta^{-k}$$

convergent is voor $|\zeta| > R_1$. Als $R_1 = 0$ dan heeft $f(\zeta)$ in $\zeta = 0$ een al dan niet ophefbare singulariteit. Omgekeerd, als we met een willekeurige Laurentreeks van de vorm (35.5) beginnen dan zijn er bijbehorende R_1 en R_2 zodat het plusstuk convergeert naar een complex differentieerbare functie op $\{\zeta \in \mathbb{C} : |\zeta| < R_2\}$ en het minstuk naar een complex differentieerbare functie op $\{\zeta \in \mathbb{C} : |\zeta| > R_1\}$. A priori kunnen zowel R_1 als R_2 ook 0 of ∞ zijn, en in het algemeen kan R_1 groter of kleiner zijn dan R_2. Laurentreeksen die verschijnen als Fourierreeksen van 2π-periodieke functies $f(x)$ via $\zeta = e^{ix}$ hebben meestal $R_1 = R_2 = 1$. In het vervolg zullen we zulke 2π-periodieke functies zien als functies $f : (-\pi, \pi) \to \mathbb{C}$.

Fourierreeksen gaan terug tot Daniel Bernouilli, die de golfvergelijking

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$$

voor bijvoorbeeld $0 < x < \pi$ en met randvoorwaarde $u = 0$ voor $x = 0$ en $x = \pi$, met Fouriersinusreeksen probeerde op te lossen. Later was Fourier de eerste die voor een gegeven functie f de coefficienten in integralen wist uit de drukken, toen hij Fouriersinusreeksen gebruikte om de warmtevergelijking

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

op te lossen.

Tegenwoordig zien we de functies

$$\frac{1}{2}, \cos x, \sin x, \cos 2x, \sin 2x, \ldots,$$

en

$$\ldots, e^{-3ix}, e^{-2ix}, e^{-ix}, e^{0ix} = 1, e^{ix}, e^{i2x}, e^{3ix}, \ldots$$

als orthonormale bases in een (Hilbert)ruimte van functies, en de Fouriercoefficienten als coordinaten ten opzichte van deze basis. Voor een grote klasse van functies $f : (-\pi, \pi) \to \mathbb{R}$ zijn de Fouriercoefficienten a_k, b_k en c_k als coordinaten van f ten opzichte van deze bases gedefinieerd. De N-de partiële som van de Fourierreeks van f is

$$S_{\mathrm{N}} f(x) = \sum_{k=-\mathrm{N}}^{\mathrm{N}} c_k e^{ikx} = \frac{a_0}{2} + \sum_{k=1}^{\mathrm{N}} (a_k \cos kx + b_k \sin kx), \tag{35.6}$$

met Fouriercoefficienten

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx \, dx, \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx \, dx, \tag{35.7}$$

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} \, dx. \tag{35.8}$$

Het volgende programma is bedoeld om vertrouwd te raken met Fourierreeksen. Gebruik Maple/Mathematica voor de plotjes. De integralen kun je beter met de hand doen.

Exercise 35.12. Bereken $\int_{-\pi}^{\pi} \cos nx \cos mx \, dx$ en $\int_{-\pi}^{\pi} \cos nx \sin mx \, dx$ voor gehele m en n.

Exercise 35.13. Laat $f : (0, \pi) \to \mathbb{R}$ gegeven zijn door $f(x) = 1$ en kies een 2π-periodieke uitbreiding $f : \mathbb{R} \to \mathbb{R}$ die even is (i.e. $f(x) = f(-x)$). Bereken alle Fouriercoefficienten a_n en b_n.

Exercise 35.14. Laat $f : (0, \pi) \to \mathbb{R}$ gegeven zijn door $f(x) = 1$ en kies een 2π-periodieke uitbreiding $f : \mathbb{R} \to \mathbb{R}$ die oneven is (i.e. $f(x) = -f(-x)$).

1. Bereken alle Fouriercoefficienten a_n en b_n.
2. Plot f en $S_{\mathrm{N}} f$ (voor een aantal waarden van N) in een grafiek.
3. Onderzoek numeriek wat er gebeurt met de grootte en plaats van het maximum van $S_{\mathrm{N}} f$ als $\mathrm{N} \to \infty$.
4. Vereenvoudig $S_{\mathrm{N}} f$ in $x = \frac{\pi}{2}$. Vergelijk dit met $f(\frac{\pi}{2})$. Van welke gewone reeks is, aangenomen dat $S_{\mathrm{N}} f(\frac{\pi}{2})$ naar $f(\frac{\pi}{2})$ convergeert, nu de som te bepalen?
5. Idem voor $x = \frac{\pi}{4}$.

Exercise 35.15. Laat $f : (0, \pi) \to \mathbb{R}$ gegeven zijn door $f(x) = \sin x$. Kies een even 2π-periodieke uitbreiding $f : \mathbb{R} \to \mathbb{R}$. Bereken alle Fouriercoefficienten a_n en b_n.

1. Bereken alle Fouriercoefficienten a_n en b_n.
2. Plot f en $S_{\mathrm{N}} f$ (voor een aantal waarden van N) in een grafiek.
3. Vereenvoudig $S_{\mathrm{N}} f$ in $x = 0$. Vergelijk dit met $f(0)$. Van welke gewone reeks is, aangenomen dat $S_{\mathrm{N}} f(0)$ naar $f(0)$ convergeert, nu de som te bepalen?
4. Idem voor $x = \frac{\pi}{2}$.
5. Idem voor $x = \frac{\pi}{4}$.

Exercise 35.16. Laat $f : (0, \pi) \to \mathbb{R}$ gegeven zijn door $f(x) = \cos x$ en kies een oneven 2π-periodieke uitbreiding $f : \mathbb{R} \to \mathbb{R}$.

1. Bereken alle Fouriercoefficienten a_n en b_n en plot f en $S_{\rm N}f$ (voor een aantal waarden van $\rm N$) in een grafiek.
2. Vergelijk het gedrag bij $x = 0$ voor $\rm N$ groot met dat in som 35.14.
3. Neem nu de oneven 2π-periodieke uitbreiding $f(x) = 1 - \cos x$ (het verschil van de functie in som 35.14 en de functie in deze som). Onderzoek numeriek het gedrag van $S_{\rm N}f$ bij $x = 0$ voor $\rm N$.

Exercise 35.17. Laat $f : (0, \pi) \to \mathbb{R}$ gegeven zijn door $f(x) = \pi - x$ en kies een oneven 2π-periodieke uitbreiding $f : \mathbb{R} \to \mathbb{R}$.

1. Bereken alle Fouriercoefficienten a_n en b_n en plot f en $S_{\rm N}f$ (voor een aantal waarden van $\rm N$) in een grafiek.
2. Differentieer $S_{\rm N}f(x)$ naar x en noem de afgeleide $d_{\rm N}(x)$. Zijn er waarden van x waarvoor $d_{\rm N}(x)$ convergeert als ${\rm N} \to \infty$?

Exercise 35.18. Laat $f : (0, \pi) \to \mathbb{R}$ gegeven zijn door $f(x) = x(\pi - x)$ en kies een oneven 2π-periodieke uitbreiding $f : \mathbb{R} \to \mathbb{R}$.

1. Bereken alle Fouriercoefficienten a_n en b_n en plot f en $S_{\rm N}f$ (voor een aantal waarden van $\rm N$) in een grafiek.
2. Vereenvoudig $S_{\rm N}f$ in $x = \frac{\pi}{2}$. Vergelijk dit met $f(\frac{\pi}{2})$. Van welke gewone reeks is, aangenomen dat $S_{\rm N}f(x)$ naar $f(x)$ convergeert, nu de som te bepalen?
3. Differentieer $S_{\rm N}f(x)$ naar x en noem de afgeleide $g_{\rm N}(x)$. Laat zien dat $g_{\rm N}(x)$ op $\mathbb{R}$ uniform convergeert naar een limietfunctie.
4. Bepaal die limietfunctie numeriek.
5. Vergelijk $g_{\rm N}(0)$ met zijn limietwaarde. Welke som van welke gewone reeks vinden we nu?

35.4 Convergentie van Fourierreeksen

Bij de vraag of, en in welke zin, de Fourierreeks convergeert, en ook als limiet f heeft, m.a.w. of

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx} = \frac{a_0}{2} + \sum_{k=1}^{\infty}(a_k \cos kx + b_k \sin kx),$$

speelt het begrip convolutie een belangrijke rol.

Exercise 35.19. Laat zien dat

$$S_{\mathrm{N}} f(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} D_{\mathrm{N}}(y) f(x-y) dy,$$

de convolutie van de functies D_n en f op $(-\pi, \pi)$, waarbij

$$D_{\mathrm{N}}(x) = \frac{\sin(\mathrm{N} + \frac{1}{2})x}{\sin \frac{1}{2}x} = \sum_{k=-\mathrm{N}}^{\mathrm{N}} e^{ikx}. \tag{35.9}$$

Maak plotjes van D_{N} voor een aantal waarden van N.

De functie D_{N} heet de Dirichlet kern. Voor grote N concentreert deze functie zich bij 0 met een steeds smallere piek waarvan de oppervlakte naar 2π gaat. Dat is de reden dat $S_{\mathrm{N}} f(x)$ naar $f(x)$ convergeert als f voldoende netjes is. Omdat D_{N} voor grotere N steeds meer tekenwisselingen heeft is dit lastig om te bewijzen. Het gemiddelde van D_0,..,D_{N},

$$F_{\mathrm{N}}(x) = \frac{1}{\mathrm{N}+1} \left(D_0(x) + \cdots + D_{\mathrm{N}}(x) \right) = \frac{1}{\mathrm{N}+1} \frac{\sin^2 \frac{(\mathrm{N}+1)x}{2}}{\sin^2 \frac{x}{2}}, \tag{35.10}$$

is een veel mooiere functie. Geen tekenwisselingen, integraal 2π, en gepiekt in 0.

Exercise 35.20. Leidt de laatste gelijkheid in (35.10) door

$$\sin \frac{x}{2} + \cdots + \sin \frac{(\mathrm{N}+1)x}{2}$$

als imaginair deel van een eindige meetkundige reeks te schrijven. Verifieer ook dat

$$\int_{-\pi}^{\pi} F_{\mathrm{N}}(x) dx = 2\pi,$$

en dat $F_{\mathrm{N}}(x) \to 0$ als $\mathrm{N} \to \infty$ behalve in veelvouden van 2π. Preciezer:

$$0 < \delta \leq x \leq \pi \implies 0 \leq F_{\mathrm{N}}(x) \leq \frac{1}{\mathrm{N}+1} \frac{1}{\sin^2 \frac{\delta}{2}}.$$

Voor vaste δ is de bovengrens klein te maken door N groot te maken. Merk op dat $F_{\mathrm{N}}(x)$ even en 2π periodiek is. Maak plotjes van F_{N} voor een aantal waarden van N.

Als een rij getallen a_n convergeert naar een limiet A, dan convergeren ook de gemiddelden

$$\frac{a_1 + a_2 + \cdots + a_n}{n}$$

naar A. De laatste limiet kan ook bestaan als de rij a_n zelf niet convergeert. Als we de rij a_n gebruiken om een A te benaderen dan kunnen we dus net zo goed eerst naar de gemiddelden kijken. Dat is het idee achter de Cesarosommen:

Exercise 35.21. Definieer

$$\sigma_{\mathrm{N}} f = \frac{1}{\mathrm{N}+1}(S_0 f + S_1 f + \cdots + S_{\mathrm{N}} f),$$

de Cesarosommen van f. Laat zien dat

$$\sigma_{\mathrm{N}} f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_{\mathrm{N}}(y) f(x-y) dy.$$

Exercise 35.22. Laat f een integreerbare (lees: begrensde stuksgewijs continue) 2π-periodieke functie zijn. Laat zien dat dan

$$\sigma_{\mathrm{N}} f(x) - f(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_{\mathrm{N}}(y)\left(f(x-y) - f(x)\right) dy$$

Exercise 35.23. Laat f een integreerbare 2π-periodieke functie zijn, met $|f(x)| \leq M$ voor alle x, waarbij $M \geq 0$ vast is. Neem aan dat voor x vast en $|y| \leq \delta$ geldt dat $|f(x-y) - f(x)| \leq \varepsilon$. Laat zien dat dan

$$|\sigma_{\mathrm{N}} f(x) - f(x)| = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_{\mathrm{N}}(y)|f(x-y) - f(x)| dy \leq \varepsilon + \frac{2M}{\mathrm{N}+1}\frac{1}{\sin^2 \frac{\delta}{2}}.$$

Hint: splits de integraal in 3 integralen.

Exercise 35.24. Laat f een 2π-periodieke continue functie zijn. Dan is f uniform continu en begrensd. Waarom? Bewijs dat $\sigma_{\mathrm{N}} f$ uniform naar f convergeert als $\mathrm{N} \to \infty$.

Exercise 35.25. Laat f een 2π-periodieke begrensde stuksgewijs continue functie zijn met de eigenschap dat in elk punt de linker en de rechterlimiet bestaan. Bewijs dat voor elke x de rij $\sigma_{\mathrm{N}} f(x)$ convergeert als $\mathrm{N} \to \infty$. Wat is de limiet? Hint: splits de integraal in 4 integralen.

Exercise 35.26. Laat $f : [-\pi, \pi] \to \mathbb{R}$ 2 keer continu differentieerbaar zijn met $f(\pm\pi) = f'(\pm\pi) = f''(\pm\pi) = 0$. Bewijs dat f in elk punt de som van zijn (uniform convergente) Fourierreeks is. Hint: laat met partieel integreren en schatten zien dat de Fouriercoefficienten a_n en b_n sommeerbare rijen vormen.

Opgave 35.24 laat zien dat in de ruimte van continue functies 2π-periodieke functies voorzien van de maximumnorm

$$||f||_\infty = \max_{x \in myR} |f(x)|$$

geldt dat de Cesarosommen van f naar f convergeren: $||\sigma_{\mathrm{N}} f - f||_\infty \to 0$ als $\mathrm{N} \to \infty$. Maar hoe zit het met $S_{\mathrm{N}} f$ zelf? Daartoe is een andere norm veel geschikter. Bij lineaire algebra of topologie zijn ongetwijfeld verschillende normen van 2-vectoren $x = \binom{x_1}{x_2}$ behandeld, bijvoorbeeld

$$||x||_\infty = \max(x_1, x_2), \quad ||x||_1 = |x_1| + |x_2|, \quad ||x||_2 = (|x_1|^2 + |x_2|^2)^{\frac{1}{2}}.$$

Als we in de laatste norm elke 2 door p vervangen (niet de subscript!) dan krijgen we de p-norm. De p-norm is een norm als $p \geq 1$. Je kunt er mee rederen zoals je dat gewend bent bij de absolute waarde. De normaxioma's zijn, voor vectoren x en y, en scalairen λ:

$$||x|| \geq 0; \quad ||x|| = 0 \Leftrightarrow x = 0; \quad ||\lambda x|| = |\lambda|\,||x||; \quad ||x + y|| \leq ||x|| + ||y||.$$

Alleen de 2-norm is een inprodukt norm. Met $x \cdot y = x_1 y_1 + x_2 y_2$ geldt

$$||x||_2^2 = x \cdot x.$$

Om ook voor functies de 2-norm in te voeren maken we eerst een inprodukt. Hieronder zijn f en g steeds 2π-periodieke stuksgewijs continue functies $f, g : \mathbb{R} \to \mathbb{R}$. We doen dus alles eerst nog reeel. Met stuksgewijs continu bedoelen we dat op elk begrensd interval er slechts eindig veel punten zijn waar de

functie niet continu is en dat in die punten linker- en rechterlimieten bestaan. De integraal

$$f \cdot g = \int_{-\pi}^{\pi} f(x)g(x)dx \tag{35.11}$$

heet het inprodukt van de functies f en g. We zien f en g als vectoren. Voor elke x zou je dan $f(x)$ en $g(x)$ als coordinaten van f en g kunnen zien. Daar heb er dan wel heel veel van, voor elke x één van f en één van g. Overeenkomstige coordinaten vermenigvuldigen en sommeren gaat niet, maar integreren wel. Vandaar de inproduktnotatie. Omdat we alles nog reeel doen kunnen we nu gebruik maken van onze intuitie voor gewone reele vectorruimten en inprodukten daarop.

Als $f \cdot g = 0$ dan zeggen we dat f en g *loodrecht op elkaar* staan. De 2-norm van f wordt gedefinieerd door

$$|f|_2 = \sqrt{f \cdot f}, \tag{35.12}$$

zeg maar de lengte van f gezien als vector. Er geldt de volgende implicatie (Pythagoras)

$$f \cdot g = 0 \quad \Rightarrow \quad |f+g|_2^2 = |f|_2^2 + |g|_2^2. \tag{35.13}$$

Hieronder schrijven we

$$S_N g(x) = \frac{c_0}{2} + \sum_{k=1}^{N} (c_k \cos kx + d_k \sin kx), \tag{35.14}$$

dus f heeft reele Fouriercoefficienten a_k en b_k, en g heeft reele Fouriercoefficienten c_k en d_k. Je kunt nu het volgende programma afwerken.

Exercise 35.27. De Cauchy-Schwartz ongelijkheid zegt dat $|f \cdot g| \leq |f|_2 |g|_2$.

1. Bewijs deze ongelijkheid voor functies f en g met $|f|_2 = |g|_2 = 1$ door $0 \leq \int_{-\pi}^{\pi} (f(x) - g(x))^2 dx = \ldots$ uit te werken.
2. Bewijs de Cauchy-Schwartz ongelijkheid. Hint: pas onderdeel 1 toe op $f(x)/|f|_2$ en $g(x)/|g|_2$.
3. Bewijs de driehoeksongelijkheid voor de 2-norm

$$|f+g|_2 \leq |f|_2 + |g|_2. \tag{35.15}$$

Exercise 35.28. Laat zien dat

$$|S_{\mathrm{N}} f|_2^2 = \pi \left(\frac{1}{2} a_0^2 + \sum_{k=1}^{\mathrm{N}} (a_k^2 + b_k^2) \right)$$

Exercise 35.29. Laat zien dat

$$S_{\mathrm{N}} f \cdot S_{\mathrm{N}} g = \pi \left(\frac{1}{2} a_0 c_0 + \sum_{k=1}^{\mathrm{N}} (a_k c_k + b_k d_k) \right)$$

Exercise 35.30. Definieer $R_{\mathrm{N}} f = f - S_{\mathrm{N}} f$ en, met

$$\sigma_{\mathrm{N}} f = \frac{1}{\mathrm{N}+1} (S_0 f + S_1 f + \cdots S_{\mathrm{N}} f),$$

ook $\rho_{\mathrm{N}} f = f - \sigma_{\mathrm{N}} f$.

1. Laat zien dat $R_{\mathrm{N}} f \cdot S_{\mathrm{N}} f = 0$.
2. Laat zien dat $R_{\mathrm{N}} f \cdot \sigma_{\mathrm{N}} f = 0$.
3. Laat zien dat
$$|S_{\mathrm{N}} f|_2^2 + |R_{\mathrm{N}} f|_2^2 = |f|_2^2,$$
zodat $|S_{\mathrm{N}} f|_2 \leq |f|_2$, en dat (Bessel's ongelijkheid)
$$\frac{1}{2} a_0^2 + \sum_{k=1}^{\mathrm{N}} (a_k^2 + b_k^2) \leq \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)^2 dx. \tag{35.16}$$
4. Laat zien dat
$$|R_{\mathrm{N}} f|_2^2 + |\sigma_{\mathrm{N}} f - S_{\mathrm{N}} f|_2^2 = |\rho_{\mathrm{N}} f|_2^2.$$
5. In Opgave 35.24 is bewezen dat als f continu en 2π-periodiek is dat dan $\sigma_{\mathrm{N}} f \to f$ uniform op $\mathbb{R}$ als $\mathrm{N} \to \infty$. Bewijs dat in dat geval ook $|R_{\mathrm{N}} f|_2 \to 0$ en dat dus (gelijkheid van Parceval)
$$\frac{1}{2} a_0^2 + \sum_{k=1}^{\infty} (a_k^2 + b_k^2) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)^2 dx. \tag{35.17}$$
Hint: gebruik onderdeel 4.

6. Laat zien dat
$$f \cdot g = (S_{\mathrm{N}} f + R_{\mathrm{N}} f) \cdot (S_{\mathrm{N}} g + R_{\mathrm{N}} g)$$
$$= S_{\mathrm{N}} f \cdot S_{\mathrm{N}} g + R_{\mathrm{N}} f \cdot R_{\mathrm{N}} g.$$

7. Bewijs dat als f en g continu en 2π-periodiek zijn, dat
$$\frac{1}{2} a_0 c_0 + \sum_{k=1}^{\infty} (a_k c_k + b_k d_k) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) g(x) dx = \frac{1}{\pi} f \cdot g. \tag{35.18}$$

   Hint: gebruik onderdeel 6, som 35.29 en pas de Cauchy-Schwartz ongelijkheid toe op $R_{\mathrm{N}} f \cdot R_{\mathrm{N}} g$.

Exercise 35.31. In deze som laten we zien dat de gelijkheid van Parceval (35.17) ook geldt voor 2π-periodieke stuksgewijs continue functies. Laat $f : \mathbb{R} \to \mathbb{R}$ zo'n functie zijn. Hint: laat zien dat er een rij 2π-periodieke continue functies $f_k : \mathbb{R} \to \mathbb{R}$ bestaat met
$$|f_k - f|_2^2 = \int_{-\pi}^{\pi} (f_k(x) - f(x))^2 dx \to 0$$
als $k \to \infty$. Als f discontinu is in x_0, vervang $f(x)$ dan op het interval $(x_0 - \frac{1}{k}, x_0 + \frac{1}{k})$ door een lineaire functie, zó dat de nieuwe functie f_k continu is en lineair op $(x_0 - \frac{1}{k}, x_0 + \frac{1}{k})$.

Exercise 35.32. Bewijs de gelijkheid van Parceval (35.17) voor f. Hint: de gelijkheid van Parceval is equivalent met $|R_{\mathrm{N}} f|_2 \to 0$. Schrijf
$$R_{\mathrm{N}} f = f - f_k + f_k - S_{\mathrm{N}} f_k + S_{\mathrm{N}} f_k - S_{\mathrm{N}} f = (f - f_k) + R_{\mathrm{N}} f_k + S_{\mathrm{N}}(f_k - f)$$
en gebruik de driehoeksongelijkheid (35.15) en som 3 voor $S_{\mathrm{N}}(f_k - f)$ om $|R_{\mathrm{N}} f|_2$ klein te krijgen. Kies hiertoe, gegeven een $\varepsilon > 0$, eerst een vaste k groot genoeg, en redeneer dan verder.

Exercise 35.33. (Het Gibbs verschijnsel) De Fouriersinusreeks van $f(x) = \pi - x$ heeft $b_n = \frac{2}{n}$. De oneven 2π-periodieke uitbreiding van f heeft $f(0^+) = \pi$ en $f(0^-) = -\pi$. De N-de Fouriersinussom is
$$S_{\mathrm{N}} f(x) = \sum_{n=1}^{\mathrm{N}} \frac{2}{n} \sin nx.$$

1. Laat zien dat
$$x + S_{\mathrm{N}} f(x) = \int_0^x D_{\mathrm{N}}(s) ds.$$

2. De integraal in het rechterlid heeft extrema in de nulpunten van D_{N}. Het eerste maximum M_{N} rechts van $x = 0$ is dus in
$$x = x_{\mathrm{N}} = \frac{\pi}{\mathrm{N} + \frac{1}{2}}.$$
Laat zien dat
$$M_{\mathrm{N}} = \int_0^{\frac{\pi}{\mathrm{N}+\frac{1}{2}}} \frac{\sin((\mathrm{N}+\frac{1}{2})s}{\sin\frac{1}{2}s} ds = \int_0^{\pi} \frac{\sin t}{\sin(\frac{1}{2}\frac{t}{\mathrm{N}+\frac{1}{2}})} \frac{1}{\mathrm{N}+\frac{1}{2}} dt$$
$$= 2\int_0^{\pi} \frac{\sin t}{t} \frac{\frac{t}{2\mathrm{N}+1}}{\sin(\frac{t}{2\mathrm{N}+1})} dt.$$

3. Laat zien dat
$$\frac{\frac{t}{2\mathrm{N}+1}}{\sin(\frac{t}{2\mathrm{N}+1})} \to 1,$$
uniform op $t \in [0, \pi]$.

4. Laat zien dat
$$M_{\mathrm{N}} \to 2\int_0^{\pi} \frac{\sin t}{t} dt$$
als $\mathrm{N} \to \infty$.

5. De functie $S_{\mathrm{N}} f(x)$ heeft in $x = x_{\mathrm{N}}$ een negatieve afgeleide. Leg uit waarom ook het eerste maximum van $S_{\mathrm{N}} f(x)$ rechts van $x = 0$ naar
$$2\int_0^{\pi} \frac{\sin t}{t} dt$$
convergeert als $\mathrm{N} \to \infty$. Dat is groter dan $\pi = f(0^+)$ met een factor 1.178979744.

Exercise 35.34. De integraal $\int_0^{\infty} \frac{\sin t}{t} dt$ is en wel uit te rekenen, met behulp van de complexe functie $\frac{e^{iz}}{z}$, zie Sectie 81 en 82 van Churchill & Brown.

35.5 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 24.3 niet voor alle x en y in $l^{(2)}$ gedefinieerd is. Informeel wordt V gegeven door

$$V = \{f \in H : f' \in L^2(-\pi, \pi)\},$$

waarbij met f ook f' steeds 2π-periodiek wordt uitgebreid tot een functie gedefinieerd op heel $\mathbb{R}$.

Dat uitbreiden is makkelijk, en komt in de opgaven hieronder eerst nog aan de orde, ook ter voorbereiding van wat een stuk lastiger is: *wat betekent het dat f' als meetbare en kwadratisch integreerbare functie bestaat?*

Exercise 35.35. 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 35.36. Zijn er functies f in $L^2(-\pi, \pi)$ waarvoor aan $f(0)$ geen betekenis[5] kan worden gegeven?

Exercise 35.37. Er is maar één 2π-periodieke oneven[6] functie Ka die voldoet aan $Ka(x) = 1$ voor $0 < x < \pi$. Schets de grafiek van Ka en maak een rij 2π-periodieke oneven continue functies $Ka_1, Ka_2 \ldots$ waarvoor geldt dat $|Ka_n - Ka|_2 \to 0$ als $n \to \infty$. Hint: schets eerst de grafieken van Ka_n.

Exercise 35.38. Bewijs dat een oneven 2π-periodieke functie wordt vastgelegd door zijn functiewaarden op het interval $(0, \pi)$. Hint: gebruik de regels $f(-x) = -f(x)$ en $f(x) = f(x + 2\pi)$. Wat is $f(0)$? En $f(\pi)$?

[5] Lees: een betekenisvolle waarde kan worden toegekend?

[6] $Ka(-x) = -Ka(x)$ voor alle $x \in \mathbb{R}$.

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 35.7. 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.

Exercise 35.39. Verifieer dat zo'n F een begrensde (2π-periodieke) functie is als $f \in H$, en dat het essentieel is dat in de definitie van H is opgenomen dat voor $f \in H$ moet gelden dat[7]

$$\int_{-\pi}^{\pi} f(x)dx = 0\,!$$

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 35.40. 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

$$V = \{F \in L^2_{loc}(\mathbb{R}) :\ f = F' \in H,\ \int_{-\pi}^{\pi} F(x)\,dx = 0\}$$

[7] $0! = 1$, maar hier roept het uitroepteken wel.

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[8]. En $L^2_{loc}(\mathbb{R})$ zelf is wel een vectorruimte maar geen genormeerde ruimte, althans niet met een natuurlijke[9] norm. Maar

$$H = \{f \in L^2_{loc}(\mathbb{R}) : \ (\forall x \in \mathbb{R}) \ f(x) = f(x + 2\pi); \ \int_{-\pi}^{\pi} f(x)dx = 0\}$$

wel, de ruimte van 2π-periodieke kwadratisch integreerbare[10] 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 33.22 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.

35.6 Blipfuncties

Het formulevoorschrift

$$x \xrightarrow{\textit{blip}} \begin{cases} \exp(-\frac{1}{x}) & \text{for } x > 0 \\ 0 & \text{for } x \leq 0, \end{cases}$$

definieert de functie $\textit{blip} : \mathbb{R} \to [0, 1)$ die met goed recht zowel oogverblindend mooi als gruwelijk lelijk genoemd[11] mag worden.

[8] Waarom niet?

[9] Bas, weet je nog, de discussie na je voordracht op de VU?

[10] D.w.z. f is meetbaar en $\int_{-\pi}^{\pi} f(x)^2 dx < \infty$.

[11] De naam *blip* heb ik gezien in een mooi boek, weet niet meer welk.

Exercise 35.41. Schets de grafiek van *blip* en onderzoek het gedrag van $blip'(x)$ als $x \downarrow 0$. En van $blip''(x)$. En van alle afgeleiden van *blip*. Concludeer dat *alle* afgeleiden van *blip* als continue functies van $\mathbb{R}$ naar $\mathbb{R}$ bestaan!

Exercise 35.42. Je kunt *blip* ook schalen. Definieer $blip_n$ door

$$blip_n(x) = blip(nx) = \exp(-\frac{1}{nx})$$

en bepaal $\lim_{n\to\infty} blip_n(x)$ voor elke $x \in \mathbb{R}$. De limietfunctie heet de Heaviside[12] functie, hier genoteerd als $He(x)$. Deze functie is niet continu in $x = 0$. De waarde van $He(0)$ als limietwaarde van $blip_n(0)$ is 0, maar net zo vaak wordt $He(0) = \frac{1}{2}$ of $He(0) = 1$ genomen. Of zelfs $He(0) = [0, 1]$.

Exercise 35.43. De functies *blip* en *He* zitten niet in $L^2(\mathbb{R})$. Waarom niet? Maar $He - blip$ wel. Waarom? En $He - blip_n$ ook. De affiene ruimte

$$He + L^2(\mathbb{R}) = \{f = He + g : \ g \in L^2(\mathbb{R})\}$$

is voorzien van de 2-metriek

$$d(f, g) = \left(\int_{-\infty}^{\infty} |f(x) - g(x)|^2 \, dx \right)^{\frac{1}{2}}$$

een volledige metrische ruimte[13]. Laat zien zien dat $d(blip_n, He) \to 0$ als $n \to \infty$.

Exercise 35.44. Definieer de functies $blok_n$ door

$$blok_n i(x) = blip_n(x) blip_n(\pi - x)$$

en laat zien dat

$$\lim_{n\to\infty} blok_n(x) = \chi_{(0,\pi)}(x)$$

voor alle $x \in \mathbb{R}$. Waarom geldt dat $blok_n \to \chi_{(0,\pi)}$ in 2-norm?

[12] Google Heaviside.
[13] Wat is dat?

Exercise 35.45. Dezelfde vragen als in Opgave 35.44 maar nu voor $blok_n$ gedefinieerd door

$$blok_n i(x) = blip_n(x - \frac{1}{n})blip_n(\pi - \frac{1}{n} - x),$$

Opgave 35.45 laat zien dat $\chi_{(0,\pi)}$, opgevat als

Exercise 35.46. 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

$$C_c^\infty(-\pi, \pi),$$

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.

35.7 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

$$|x|_p = \left(\sum_{j=1}^{\mathrm{N}} |x_j|^p \right)^{\frac{1}{p}},$$

de p-norm van $x = (x_1, \ldots, x_n) \in \mathbb{R}^{\mathrm{N}}$.

Exercise 35.47. Terug naar de overgeslagen calculussommetjes, bewijs (de ongelijkheid van Hölder)

$$|x \cdot y| \leq |x|_p |y|_q$$

voor $1 \leq p, q \leq \infty$ die voldoen aan

$$\frac{1}{p} + \frac{1}{q} = 1,$$

en $x = (x_1, \ldots, x_n)$ en $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^{\mathrm{N}}$. Hint: leg eerst uit waarom het *geen* beperking is om aan te nemen dat $|x|_p = |y|_q = 1$.

Exercise 35.48. Bewijs dat $x \to |x|_p$ een norm is op $\mathbb{R}^{\mathrm{N}}$.

Exercise 35.49. Bewijs dat $|x|_p \to |x|_\infty$ als $p \to \infty$.

Exercise 35.50. 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 $C[-\pi, \pi]$.

36 Functies op de cirkel

Als we afspreken dat twee getallen in $\mathbb{R}$ eigenlijk hetzelfde zijn als ze een geheel veelvoud van 2π verschillen dan maken $\mathbb{R}$ en 2π de verzameling $\mathbb{R}_{2\pi}$, een verzameling waarin op natuurlijke manier de optelling is gedefinieerd. Dat gaat net als in $\mathbb{Z}_n$, de verzameling die we krijgen uit de verzameling $\mathbb{Z}$ van gehele getallen en een vast getal $n \in \mathbb{N}$, door af te spreken dat twee gehele getallen gelijk zijn als ze een geheel veelvoud van n verschillen. Zoals vaak

$$\mathbb{Z}_n = \{0, 1, \ldots, n-1\}$$

wordt geschreven, met $0 = n$, kunnen we ook

$$\mathbb{R}_{2\pi} = [0, 2\pi)$$

schrijven, maar we geven er de voorkeur om $\mathbb{R}_{2\pi}$ in de schrijfwijze te laten corresponderen met $[-\pi, \pi)$, waarbij $-\pi = \pi$. Deze π is hier een positief reëel getal, waarvoor we op enig moment de π die we van de cirkel kennen zullen nemen, *maar dat hoeft nu nog even niet*. Net als $\mathbb{Z}_n$ is $\mathbb{R}_{2\pi}$ met de voor de hand liggende optelling een commutatieve groep[1].

Functies $f : \mathbb{R} \to \mathbb{R}$ die 2π-periodiek zijn kunnen we ook opvatten als functies $f : \mathbb{R}_{2\pi} \to \mathbb{R}$, en omgekeerd. De verzameling van continue 2π-periodieke functies noemen we

$$C(\mathbb{R}_{2\pi}).$$

Ieder tweetal functies gedefinieerd op dezelfde verzameling, dus ook f en g in $C(\mathbb{R}_{2\pi})$, kunnen we bij elkaar optellen[2] 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[3] over $\mathbb{R}$, waarop

$$f \cdot g = (f, g) = \int_{-\pi}^{\pi} f(x)g(x)\,dx$$

[1] Google: Abelian group.

[2] Evenzo is natuurlijk ook fg gedefinieerd via $x \xrightarrow{fg} f(x)g(x)$.

[3] En met de vermenigvulding een algebra.

een inwendig produkt[4] definieert, maar $C(\mathbb{R}_{2\pi})$ is met dit integraalinprodukt geen Hilbertruimte, zoals de volgende opgave laat zien.

Exercise 36.1. De zaagtandfunctie Z wordt gedefinieerd door

$$Z(x) = \begin{cases} \pi - x & \text{als } x \in (0, \pi] \ ; \\ -x - \pi & \text{als } x \in [-\pi, 0), \end{cases}$$

en door $Z(0) = 0$. Met deze keuze voor $Z(0)$ behoort Z tot $\mathcal{C}_0(\mathbb{R}_{2\pi})$, de ruimte[5] van functies $f : \mathbb{R}_{2\pi} \to \mathbb{R}$ die continu zijn in iedere $x \in \mathbb{R}_{2\pi}$, behalve eventueel in $x = 0$, maar waarvoor wel geldt dat

$$f(0) = \frac{1}{2}\left(\lim_{x \downarrow 0} f(x) + \lim_{x \uparrow 0} f(x)\right),$$

waarbij linker- en rechterlimiet dus allebei bestaan. Op $\mathcal{C}_0(\mathbb{R}_{2\pi})$ 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 $\mathcal{C}_0(\mathbb{R}_{2\pi})$ 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[6]. 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?

[4] Let op, met $(f, g) \dot{\to} f \cdot g = (f, g)$ is de haakjesnotatie soms verwarrend.
[5] De notatie $\mathcal{C}$ is alleen voor nu even.
[6] Waarom niet?

37 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:

37.1 Metrische ruimten; continue afbeeldingen

Aanvullend materiaal voor het college Analyse 1

J. Hulshof (destijds RUL)

Cursief wat opmerkingen van 24 jaar later ingevoegd tijdens het geven van het college Analyse 3 in het tweede jaar, met verwijzingen naar het boek Principles of Topology van Croom.

1. Inleiding. In deze syllabus behandelen we een aantal fundamentele onderwerpen uit de analyse. Uitgangspunt hierbij is de volgende algemene probleemstelling:

Laat X een puntverzameling zijn en A een niet-lege deelverzameling van X, eventueel X zelf. Zij $f: A \to \mathbb{R}$ een reëelwaardige functie. Hoe en onder wat voor veronderstellingen kunnen we dan concluderen dat de functie f op A een globaal maximum aanneemt, m.a.w. bestaat er een punt $x_0 \in A$, zo dat

$$f(x) \leq f(x_0) \quad \forall x \in A?$$

Om deze vraag te beantwoorden beschouwen we het supremum van f op A,

$$M = \sup\{f(x) : \ x \in A\} \ \in \ \mathbb{R} \cup \{+\infty\}.$$

Wat we van M nu willen weten is ten eerste of M eindig is, en zo ja, of de waarde M ook door de functie f wordt aangenomen. Omdat M het supremum is van alle door f aangenomen functiewaarden, kunnen we M benaderen met deze functiewaarden. We onderscheiden twee gevallen.

(i) $M < +\infty$. Dan bestaat er voor elk natuurlijk getal n een $x_n \in A$, zodat

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

(ii) $M = +\infty$. Dan bestaat er voor elk natuurlijk getal n een $x_n \in A$, zodat

$$f(x_n) > n.$$

In beide gevallen geeft dit ons een rij punten, genoteerd als $(x_n)_{n=1}^{\infty}$, in A. Als we nu kunnen concluderen dat, eventueel door een aantal van deze punten uit de rij weg te laten, deze punten voor grote waarden van $n \in \mathbb{N}$ steeds dichter komen te liggen bij een "limietpunt" dat zelf ook in A ligt, en dat de waarde van f in dat limietpunt gelijk is aan

$$\lim_{n\to\infty} f(x_n),$$

dan hebben we in één klap de beide bovenstaande vragen met ja beantwoord. In deze syllabus zullen we de voor bovenstaande probleemstelling relevante begrippen behandelen, met hier en daar een zijstapje.

In de vorige alinea staat de zinsnede "dichterbij". Aangezien we van X alleen maar hebben aangenomen dat X een puntverzameling is, heeft dit zonder verdere veronderstellingen geen betekenis. Een natuurlijke manier om dit de verhelpen is de invoering van een zogenaamde afstandsfunctie of metriek op X. Dit leidt dan tot de definitie van een metrische ruimte (Sectie 2). De deelverzamelingen waarvoor altijd een limietpunt van een rij bestaat blijken de zogenaamde rijkompakte verzamelingen te zijn (Sectie 3). Voor functies op (deelverzamelingen van) metrische ruimten kan het begrip "continu" worden gedefinieerd (Sectie 5), waarmee de bovenstaande limietovergang kan worden gerechtvaardigd. Als voorbeeld behandelen we de gevallen dat $X = \mathbb{R}$ en $X = \mathbb{R}^N$ (Sectie 6). In de appendix komen nog enige iets meer geavanceerde onderwerpen met betrekking tot kompaktheid aan de orde.

De schrijver van deze syllabus is van mening dat iedere student in de wiskunde of theoretische natuurkunde, onafhankelijk van wat hij/zij in de doctoraalfase als afstudeerrichting kiest, zich de basisstof in Sectie 1 tot en met 6 van deze syllabus moet eigen maken. In de meeste theoretische analyse boeken is deze stof, althans voor het geval $X = \mathbb{R}^N$, terug te vinden in de inleidende hoofdstukken. Zie bijvoorbeeld het boek "Mathematical Analysis, a modern approach to advanced calculus" van T.M. Apostol (Addison-Wesley 1957), waarin bijna alle analyse die in de eerste twee jaar van de studie aan de orde komt, is terug te vinden, of "Principles of Mathematical Analysis" van W. Rudin (McGraw Hill 1964). Voor meer algemene metrische (en topologische) ruimten zijn er o.a. de boeken "Topology and Normed Spaces" van G.J.O. Jameson (Wiley 1974) en "Introduction to Topology and Modern Analysis" van G.F. Simmons (McGraw Hill 1963).

Als voorkennis wordt verondersteld dat de lezer bekend is met de elementaire verzamelingsleer, begrippen als aftelbaar oneindig en overaftelbaar oneindig, en met de axioma's voor de natuurlijke getallen $\mathbb{N}$ en de reële getallen $\mathbb{R}$. Hoofdstuk 1 uit het boek "Calculus 1 2nd edition" van T.M. Apostol (Wiley 1967) is ruim voldoende.

2. Metrische ruimten. Laat X weer een puntverzameling zijn.

Definitie 2.1. Een functie $d: X \times X \to \mathbb{R}$ heet een *metriek* op X als
(i) $\forall x, y \in X$:

$$d(x,y) \geq 0 \text{ en } d(x,y) = 0 \iff x = y.$$

(ii) $\forall x, y \in X$:

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

(iii) $\forall x, y, z \in X$:

$$d(x,z) \leq d(x,y) + d(y,z) \quad \text{(driehoeksongelijkheid)}.$$

Als $d: X \times X \to \mathbb{R}$ een metriek is, dan heet het paar (X,d) een *metrische ruimte*.

Het is duidelijk dat op een verzameling X meerdere metrieken kunnen zijn gedefinieerd. Toch spreekt men vaak over de metrische ruimte X i.p.v. over de metrische ruimte (X,d). Aangenomen is dan dat over de stilzwijgend gemaakte keuze van de metriek d geen misverstand kan bestaan. In het vervolg is X nu steeds een metrische ruimte met metriek d.

Voorbeeld 2.2. $X = \mathbb{R}$ is de verzameling van de reële getallen, met $d(x,y) = |x-y|$.

Opmerking 2.3. Iedere deelverzameling van een metrische ruimte is met dezelfde metriek weer een metrische ruimte.

Definitie 2.4. Laat $A \subset X$.
(i) Een punt $a \in A$ heet een *inwendig punt* van A als

$$\varepsilon xists\delta > 0: \quad B_\delta(a) = \{x \in X: \ d(x,a) < \delta\} \ \subset \ A.$$

De verzameling $B_\delta(a)$ heet de open bol met straal δ en middelpunt a.
(ii) Een punt $a \in A$ heet een *geisoleerd punt* van A als

$$\exists \delta > 0: \quad B_\delta(a) \cap A = \{a\}.$$

(iii) Een punt $p \in X$ heet een *ophopingspunt* van A als

$$\forall \delta > 0 \ \exists a \in A: \quad a \neq p \text{ en } d(a,p) < \delta.$$

(iv) Als elk punt van A een inwendig punt van A is, dan heet A een *open deelverzameling* van X.
(v) Als het complement van A,

$$A^c = X - A = \{x \in X : x \notin A\},$$

open is, dan heet A een *gesloten deelverzameling* van X.

In het vervolg zullen we kortweg zeggen dat A open (gesloten) is als A een open (gesloten) deelverzameling van X is.

LET OP! *Definitie 2.4 benoemt eigenschappen van punten $a \in A$ en $p \in X$ met $A \subset X$ en X een metrische ruimte. Maar A is zelf ook weer een metrische ruimte en dus te zien in de rol van X hierboven. Je kunt de grotere X dan verder vergeten bij het doen van uitspraken over deelverzamelingen van A. Bijvoorbeeld over $A \subset A$, net zoals je uitspraken over X kunt doen, gezien als deelverzameling van de metrische ruimte X.*

Voorbeeld: *De gehele getallen vormen een verzameling waarvoor geldt dat $\mathbb{Z} \subset \mathbb{R}$. In $\mathbb{Z}$ is elk punt een geisoleerd punt en elke bol met straat $\frac{1}{2}$ een singleton, dus $\{0\}$ is een open deelverzameling van $\mathbb{Z}$ maar geen open deelverzameling van $\mathbb{R}$.*

Stelling 2.5. (i) Iedere open bol is een open verzameling.
(ii) Een verzameling A in X is open dan en slechts dan als A een vereniging van open bollen is.

Bewijs. Opgave.

Stelling 2.6. (i) X is open, de lege verzameling is open.
(ii) De vereniging van elke collectie open deelverzamelingen is open.
(iii) De doorsnede van elk eindig aantal open deelverzamelingen is open.

Bewijs. Opgave.

Stelling 2.7. (i) X is gesloten, de lege verzameling is gesloten.
(ii) De doorsnede van elke collectie gesloten deelverzamelingen is gesloten.
(iii) De vereniging van elk eindig aantal gesloten verzamelingen is gesloten.

Bewijs. Opgave.

Stelling 2.8. Laat $A \subset X$. Dan is A gesloten dan en slechts dan als A al zijn ophopingspunten bevat.

Bewijs. Neem eerst aan dat A gesloten is en laat p een ophopingspunt zijn van A. We moeten laten zien dat $p \in A$. Stel niet. Dan $p \in A^c$. Maar A^c is open, dus p is een inwendig punt van A^c. Zodoende is er een $\delta > 0$ waarvoor

$B_\delta(p) \subset A^c$, in tegenspraak met de veronderstelling dat p een ophopingspunt is van A. Dus p ligt wel in A.

Neem vervolgens aan dat A een verzameling is die al zijn ophopingspunten bevat. We tonen aan dat A gesloten is door te bewijzen dat A^c open is. Zij $p \in A^c$. Dan is p geen ophopingspunt van A, dus er is een $\delta > 0$ zo dat $B_\delta(p)$ geen punten van $A - \{p\}$ bevat, en wegens $p \in A^c$ betekent dit dat $B_\delta(p) \subset A^c$. m.a.w. p is een inwendig punt van A^c. Dit geldt voor elke $p \in A^c$, en dus is A^c open. Q.e.d.

Definitie 2.9. Een rij $(x_n)_{n=1}^\infty$ in X heet *convergent* als er een $x_0 \in X$ is zo dat

$$\forall \varepsilon > 0 \; \exists n_\varepsilon \in \mathbb{N} : \quad n \geq n_\varepsilon \implies d(x_n, x_0) < \varepsilon.$$

Het punt x_0 heet de *limiet* van de rij.

Stilzwijgend zijn n, n_ε, m en met andere letters in het midden van het alphabet genoteerde variabelen vaak elementen van $\mathbb{N}$. De definitie wordt vaak geschreven zonder de subindex ε:

$$\forall \varepsilon > 0 \; \exists N : \quad n \geq N \implies d(x_n, x_0) < \varepsilon.$$

Stelling 2.10. Als de rij $(x_n)_{n=1}^\infty$ in X convergent is, dan is de limiet x_0 eenduidig bepaald, notatie

$$\lim_{n\to\infty} x_n = x_0 \quad \text{of} \quad x_n \to x_0.$$

Bewijs. Neem aan dat de rij twee verschillende limieten heeft, zeg x_0 en x_0'. Omdat $x_0 \neq x_0'$, is $d(x_0, x_0') > 0$. Kies $0 < \varepsilon < \frac{1}{2} d(x_0, x_0')$. Dan is er een n_ε zo dat $d(x_n, x_0) < \varepsilon$ voor alle $n \geq n_\varepsilon$. Evenzo is er een n_ε' zo dat $d(x_n, x_0') < \varepsilon$ voor alle $n \geq n_\varepsilon'$. Met behulp van de driehoeksongelijkheid volgt nu

$$d(x_0, x_0') \leq d(x_0, x_n) + d(x_n, x_0') < \varepsilon + \varepsilon < d(x_0, x_0').$$

voor $n \geq \max(n_\varepsilon, n_e')$, tegenspraak. Q.e.d.

In plaats van "de rij $(x_n)_{n=1}^\infty$ is convergent", zegt men ook wel dat "$\lim_{n\to\infty} x_n$ bestaat".

Stelling 2.11. Als de rij $(x_n)_{n=1}^\infty$ in X convergent is, dan is de rij begrensd, d.w.z. bevat in een vaste (open) bol $B_\delta(a) \subset X$.

Bewijs. Opgave.

Stelling 2.12. Laat $A \subset X$ gesloten zijn. Als de rij $(x_n)_{n=1}^\infty$ in A convergent is in X, dan ligt de limiet in A.

Bewijs. Opgave.

Stelling 2.13. Laat $A \subset X$ en $p \in X$. De volgende drie uitspraken zijn equivalent:
(i) p is een ophopingspunt van A.
(ii) Er bestaat een rij $(a_n)_{n=1}^{\infty}$ in $A - \{p\} = \{x \in A : \ x \neq p\}$ die convergent is met p als limiet.
(iii) Iedere (niet-lege) open bol met middelpunt p bevat oneindig veel punten van A.

Bewijs. $(i \implies ii)$. Neem aan dat p een ophopingspunt van A is. We construeren de rij $(a_n)_{n=1}^{\infty}$ in A. Kies $\varepsilon_1 > 0$. Dan is er een $a_1 \in A - \{p\}$ met $d(a_1, p) < \varepsilon_1$. De rij $(a_n)_{n=1}^{\infty}$ wordt nu verder inductief gedefinieerd door voor $n = 1, 2, \ldots,$ nadat a_n gekozen is, $\varepsilon_{n+1} = \frac{1}{2}d(a_n, p)$ te stellen, en $a_{n+1} \in A - \{p\}$ met $d(a_{n+1}, p) < \varepsilon_{n+1}$ te kiezen. Opgave: laat zien dat $\varepsilon_{n+1} < \frac{1}{2}\varepsilon_n$ en dat de rij naar p convergeert.
$(ii \implies iii)$. Opgave.
$(iii \implies i)$. Triviaal.

Correctie! *In de 1993-versie stond in Stelling 2.13 en in het bewijs daarvan hier en daar nog een a waar een p moest staan.*
Het gebruik van definities met de quantor voor alle ($\forall$). *Belangrijk is om te onthouden dat een definitie die begint met $\forall \delta > 0$ geldt "iets" waarbij δ een rol speelt, pas echt gebruikt is als voor een rij $\delta_n \downarrow 0$ dat "iets" gebruikt is. En in plaats van δ's kunnen natuurlijk ook ε's gebruikt worden.*
Variaties op het bewijs. *Het dalend kiezen van de rij $\varepsilon_n > 0$ in $(i \implies ii)$ kan natuurlijk op vele manieren. Met $\varepsilon_1 = 1$ en vervolgens*

$$\varepsilon_{n+1} = \min(d(a_n, p), \frac{1}{n+1})$$

voor $\varepsilon_2, \varepsilon_3, \ldots$ werkt het bewijs net zo goed. Met die constructie volgt dan meteen dat $\varepsilon_n \leq \frac{1}{n}$. In het als opgave te geven bewijs van $a_n \to p$ moet je voor alle $\varepsilon > 0$ een N vinden zoals onder Definitie 2.9. Kies daartoe N met $\frac{1}{N} < \varepsilon$ (waarom kan dat?) en gebruik de bijbehorende ε_N gedefinieerd zoals hier direct boven.
De Archimedische eigenschap van de verzameling van de reële getallen. *Waarom bestaat er voor elke ε eigenlijk een n met $\frac{1}{n} < \varepsilon$? Wel, indien niet dan zou er een $\varepsilon > 0$ zijn met $\frac{1}{n} \geq \varepsilon$ en dus $n \leq \frac{1}{\varepsilon}$ voor alle $n \in \mathbb{N}$. De (niet-lege) verzameling $\mathbb{N}$ zou dan naar boven begrensd zijn in $\mathbb{R}$ en in $\mathbb{R}$ een kleinste bovengrens S hebben. Dan is $S - \frac{1}{2}$ geen bovengrens en dus bestaat er een $N \in \mathbb{N}$ met $S - \frac{1}{2} < N \leq S$ en bijgevolg $N + 1 > S + \frac{1}{2}$. Om dat $N + 1 \in \mathbb{N}$ kan S dus geen bovengrens zijn van $\mathbb{N}$, laat staan de kleinste.*

Zonder epsilons kan het dus ook. *De nu bewezen uitspraak dat onder elke $\varepsilon > 0$ altijd een $\frac{1}{n}$ zit met $n \in \mathbb{N}$ wordt de Archimedische eigenschap van $\mathbb{R}$ genoemd en maakt dat iedere definitie die begint met $\forall \varepsilon > 0$ en eindigt met $< \varepsilon$ kan worden vervangen door een definitie die begint met $\forall n \in \mathbb{N}$ en eindigt met $< \frac{1}{n}$. Alleen komen we dan al snel letters in het midden van het alphabet te kort.*

Om welk axioma voor, of eigenschap van de verzameling van de reële getallen ging het? *De Archimedische eigenschap geldt dankzij het axioma over het bestaan van kleinste bovengrenzen in $\mathbb{R}$ voor naar bovengrensde niet-lege deelverzamelingen van $\mathbb{R}$. In het boek van Croom wordt dit besproken in Sectie 2.1.*

Croom *gebruikt in zijn Sectie 3.2 het woord limietpunt als ander woord voor ophopingspunt. Ik reserveer de term limietpunt voor het gebruik zoals in Definitie 3.1 hieronder: limiet van een convergente deelrij van een gegeven rij. Dus rijen kunnen limietpunten hebben en verzamelingen ophopingspunten. Een rij in (A of) X is strict genomen ook geen deelverzameling van (A of) X maar een functie of afbeelding van $\mathbb{N}$ naar (A of) X. Croom's index verwijst voor limit point naar pagina 66 waar Stelling 3.6 komt na de definitie onderaan pagina 65, en daar zie je dat limietpunt bij Croom een andere naam is voor ophopingspunt. Een naamgeving wellicht verdedigbaar door de uitspraak dat bij een ophopingspunt p van A in X altijd een rij $(a_n)_{n=1}^{\infty}$ te vinden is waarvoor $d(a_n, p)$ strict dalend is in n en convergeert naar 0. Merk op dat voor $n \to \infty$ de equivalentie*

$$a_n \to p \iff d(a_n, p) \to 0$$

vanuit de definitie van convergentie vanzelfsprekend is.

Definitie A-1. *Laat weer $A \subset X$ met X een metrische ruimte. De afsluiting van A is de vereniging van A met al zijn ophopingspunten, notatie $\bar{A}$. Dus $p \in \bar{A}$ betekent dat de implicatie*

$$p \notin A \implies p \text{ is een ophopingspunt van } A$$

moet gelden.

Opgave A-2. *Bewijs dat $\bar{A}$ gesloten is in X. Hint: bewijs dat een ophopingspunt van $\bar{A}$ ook een ophopingspunt van A is gebruik Stelling 2.8.*

Opgave A-3. *De sterkere uitspraak is dat $\bar{A}$ de kleinste gesloten verzameling in X is die A bevat: bewijs dat $\bar{A}$ de doorsnede is van alle gesloten deelverzamelingen F van X met $A \subset F$. Hint: die doorsnede D is van vanwege Stelling 2.7 gesloten dus vanwege Opgave A-2 is $D \subset \bar{A}$. Kan A ophopingspunten hebben die niet in die doorsnede liggen?*

Croom noemt de verzameling van alle ophopingspunten de afgeleide verzameling van A, zonder verdere notatie, en schrijft $cl(A)$ voor $\bar{A}$.

We merken nog eens op dat de bovenstaande begrippen in principe afhangen van de keuze van de metriek op X. Toch kunnen verschillende metrieken tot hetzelfde leiden.

Definitie 2.14. Laat X een metrische ruimte zijn met metriek d, en laat $\bar{d}$ een andere metriek op X zijn. Dan heten d en $\bar{d}$ *equivalent* op X als er een reëel getal $\lambda > 0$ bestaat zo dat

$$\frac{1}{\lambda} d(x,y) \leq \bar{d}(x,y) \leq \lambda d(x,y) \quad \forall x, y \in X.$$

Opgave 2.15. Laat zien dat bij overgang op een equivalente metriek op X de in deze Sectie geintroduceerde begrippen (open, gesloten, convergent, etc.) niet veranderen. Hetzelfde geldt voor de begrippen die in de volgende drie secties worden behandeld (rijkompaktheid, volledigheid en continuiteit).

Op grond van het voorafgaande is het duidelijk dat voor het welslagen van de in de inleiding geschetste bewijsmethode, de geslotenheid van A vereist is. Dit staat echter los van de vraag of de in de inleiding geconstrueerde rij een limietpunt heeft.

3. Rijkompakte verzamelingen. We gaan nu de deelverzamelingen A karakteriseren waarvoor de in de inleiding geschetste bewijsmethode zal slagen.

Definitie 3.1. Laat $(x_n)_{n=1}^{\infty}$ een rij zijn in X.
(i) Als $(n_k)_{k=1}^{\infty}$ een strict stijgende rij natuurlijke getallen is, dan heet de rij $(x_{n_k})_{k=1}^{\infty}$ een *deelrij* van $(x_n)_{n=1}^{\infty}$.
(ii) Als $x_0 \in X$ de limiet is van een convergente deelrij van $(x_n)_{n=1}^{\infty}$, dan heet x_0 een *limietpunt* (ook wel: rijophopingspunt) van $(x_n)_{n=1}^{\infty}$.

Stelling 3.2. Een rij $(x_n)_{n=1}^{\infty}$ in X heeft een convergente deelrij met limiet x_0 dan en slechts dan als

$$\forall \varepsilon > 0 \ \forall n \ \exists k \geq n : \quad d(x_k, x_0) < \varepsilon.$$

Bewijs. Opgave.

Definitie 3.3. Laat $A \subset X$. A heet *rijkompakt* als iedere rij in A een limietpunt in A heeft.

Er is nog een andere definitie van kompaktheid die niet uitgaat van de metriek op X, maar van de open deelverzamelingen van X. In de appendix

zullen we deze definitie behandelen, en laten zien dat de beide definities voor deelverzamelingen van metrische ruimten hetzelfde betekenen.

Stelling 3.4. Gesloten deelverzamelingen van rijkompakte verzamelingen zijn rijkompakt.

Bewijs. Merk eerst op: als $G \subset A \subset X$, dan kan het gesloten zijn van G op twee manieren worden opgevat: gesloten in A of gesloten in X. Opgave: laat zien dat, als A gesloten is in X, dan

$$G \text{ is gesloten in } A \iff G \text{ is gesloten in } X.$$

Stel dat $(a_n)_{n=1}^{\infty}$ een rij is in G. Omdat A rijkompakt is, bestaat er een convergente deelrij met limiet in A. Daar G gesloten is ligt de limiet in G. Conclusie: elke rij in G heeft een convergente deelrij met limiet in G. Q.e.d.

Stelling 3.5. (i) Laat (X_1, d_1) en (X_2, d_2) twee metrische ruimten zijn. Dan is het Cartesisch produkt X van X_1 en X_2,

$$X = X_1 \times X_2 = \{(x_1, x_2) : \quad x_1 \in X_1,\ x_2 \in X_2\},$$

weer een metrische ruimte t.a.v. de metriek gedefinieerd door

$$d(x, y) = d_1(x_1, y_1) + d_2(x_2, y_2) \ \ \forall x = (x_1, x_2),\ y = (y_1, y_2) \in X = X_1 \times X_2.$$

(ii) Als A_1 rijkompakt is in X_1 en A_2 rijkompakt is in X_2, dan is $A = A_1 \times A_2$ rijkompakt in $X = X_1 \times X_2$.

Bewijs. Opgave.

Stelling 3.6. Laat A een rijkompakte deelverzameling zijn van X. Dan is A gesloten en begrensd.

Bewijs. Opgave.

4. Volledigheid.

Definitie 4.1. Een rij $(x_n)_{n=1}^{\infty}$ in X heet *een Cauchyrij* als

$$\forall \varepsilon > 0\ \exists n_\varepsilon : \quad m, n \geq n_\varepsilon \implies d(x_m, x_n) < \varepsilon.$$

Stelling 4.2. Iedere convergente rij is een Cauchyrij.

Bewijs. Zij $(x_n)_{n=1}^{\infty}$ een convergente deelrij met limiet x_0. Laat $\varepsilon > 0$, en gebruik de definitie van convergentie met $\frac{1}{2}\varepsilon$. Dan,

$$\forall m, n \geq n_{\frac{1}{2}\varepsilon} : \quad d(x_m, x_n) \leq d(x_m, x_0) + d(x_0, x_n) < \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon.$$

$\varepsilon > 0$ was willekeurig, dus de rij is een Cauchyrij. Q.e.d.

Stelling 4.3. Iedere Cauchyrij is begrensd.

Bewijs. Opgave.

Stelling 4.4. Als een Cauchyrij een limietpunt heeft, dan is de Cauchyrij convergent.

Bewijs. Laat $(x_n)_{n=1}^{\infty}$ een Cauchrij zijn die een convergente deelrij heeft. Laat $\varepsilon > 0$ en gebruik de definitie van Cauchyrij. Dus

$$\forall m, n \geq n_\varepsilon : \quad d(x_m, x_n) < \frac{1}{2}\varepsilon.$$

Anderzijds, omdat de rij een convergente deelrij heeft, zeg met limiet x_0, bestaat er een $k_\varepsilon > n_\varepsilon$ zodat

$$d(x_{k_\varepsilon}, x_0) < \frac{1}{2}\varepsilon.$$

Beide ongelijkheden combinerend vinden we

$$\forall n \geq n_\varepsilon : \quad d(x_n, x_0) \leq d(x_n, x_{k_\varepsilon}) + d(x_{k_\varepsilon}, x_0) < \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon.$$

Definitie 4.5. Als iedere Cauchyrij in X convergent is, dan heet X *volledig*.

Om na te gaan of een rij $(x_n)_{n=1}^{\infty}$ in een volledige metrische ruimte X convergent is, hoeft men slechts na te gaan dat de rij een Cauchyrij is. Daarbij is het niet nodig om a priori de limietwaarde te kennen. Het bewijs van een belangrijke stelling in de analyse, de Banach contractie stelling, berust op dit principe:

Stelling 4.6. Laat X een volledige metrische ruimte zijn, en $T : X \to X$ een contractie, d.w.z. een afbeelding met de eigenschap dat

$$\exists \theta \in [0, 1) \ \forall x, y \in X \quad d(T(x), T(y)) \leq \theta d(x, y).$$

Dan is er precies één vast punt van T, d.w.z. een punt $\bar{x}$ in X met de eigenschap dat $T(\bar{x}) = \bar{x}$. Bovendien geldt voor elke $x \in X$ dat de rij gedefinieerd door

$$x_1 = T(x), \quad x_{n+1} = T(x_n) \ (n = 1, 2, \dots),$$

convergent is met limiet $\bar{x}$.

Bewijs. Merk eerst op dat er hoogstens een vast punt kan zijn, immers als $\bar{x}$ en $\bar{y}$ vaste punten zijn, dan

$$d(\bar{x}, \bar{y}) = d(T(\bar{x}), T(\bar{y})) \leq \theta d(\bar{x}, \bar{y}) \implies d(\bar{x}, \bar{y}) = 0 \implies \bar{x} = \bar{y}.$$

Laat nu $x \in X$ willekeurig en laat de rij $(x_n)_{n=1}^{\infty}$ de in de stelling gedefinieerde rij zijn. We gaan bewijzen dat dit een Cauchyrij is. Neem hiertoe twee natuurlijke getallen m, n met $m < n$. Dan, door herhaald toepassen van de driehoeksongelijkheid,

$$d(x_m, x_n) \leq d(x_m, x_{m+1}) + d(x_{m+1}, x_{m+2}) + \cdots + d(x_{n-1}, x_n).$$

We gebruiken de notatie

$$T^0(x) = x, \quad T^1(x) = T(x), \quad T^2(x) = T(T(x)), \quad T^3(x) = T(T(T(x))), \quad \text{etc.}$$

Omdat

$$d(x_k, x_{k+1}) = d(T^k(x), T^{k+1}(x)) \leq \theta^k d(x, T(x)),$$

volgt nu dat

$$d(x_m, x_n) \leq (\theta^m + \theta^{m+1} + \cdots + \theta^{n-1}) d(x, T(x)) \leq \frac{\theta^m}{1-\theta} d(x, T(x)) \to 0 \quad \text{als } m \to \infty.$$

Dus $(x_n)_{n=1}^{\infty}$ is een Cauchyrij. Omdat X volledig is heeft deze rij een limiet $\bar{x}$, dus $T^n(x) \to \bar{x}$. Maar dan geldt ook dat $T^{n+1}(x) \to \bar{x}$, terwijl

$$d(T^{n+1}(x), T(\bar{x})) = d(T(T^n(x)), T(\bar{x})) \leq \theta d(T^n(x), \bar{x}) \leq d(T^n(x), \bar{x}) \to 0.$$

Dit impliceert dat $T^{n+1}(x) \to T(\bar{x})$. Omdat de limiet van een convergente rij uniek bepaald is, kunnen we dus concluderen dat $\bar{x} = T(\bar{x})$. Q.e.d.

Er is nog een andere karakterisatie van het begrip volledigheid, die wordt gegeven door de volgende stelling.

Stelling 4.7. (Cantor) Een metrische ruimte X is volledig dan en slechts dan als voor elke dalende rij gesloten deelverzamelingen

$$F_1 \supset F_2 \supset F_3 \ldots,$$

met de eigenschap dat

$$\operatorname{diam}(F_n) = \sup_{x,y \in F_n} d(x, y) \to 0 \text{ als } n \to \infty,$$

geldt dat de doorsnede

$$\cap_{n=1}^{\infty} F_n$$

precies één punt bevat.

5. Continue afbeeldingen.

Definitie 5.1. Laat X en Y metrische ruimten zijn, $A \subset X$, en $f : A \to Y$ een afbeelding. (i) f heet *continu in* $a \in A$, als

$$\forall \varepsilon > 0 \; \exists \delta > 0 \; \forall x \in A : \quad d(x,a) < \delta \implies d(f(x), f(a)) < \varepsilon.$$

(ii) f heet continu in A als f continu is in elke $a \in A$.

Stelling 5.2. Laat X en Y metrische ruimten zijn, $A \subset X$, $f : A \to Y$ een afbeelding, en $a \in A$. Dan is f continu in a dan en slechts dan als voor elke rij $(a_n)_{n=1}^{\infty}$ in A met $a_n \to a$ in X geldt dat $f(a_n) \to f(a)$ in Y.

Bewijs. Neem aan dat f continu is in a en laat $(a_n)_{n=1}^{\infty}$ een rij zijn in A met $a_n \to a$. Kies $\varepsilon > 0$ willekeurig, en laat $\delta > 0$ de bijbehorende δ uit de definitie van continuiteit van f in a zijn. Gebruik deze δ nu als ε in de definitie van convergentie. Dan

$$n \geq n_\delta \implies d(a_n, a) < \delta \implies d(f(a_n), f(a)) < \varepsilon.$$

Dus de rij $(f(a_n))_{n=1}^{\infty}$ convergeert naar $f(a)$.

Anderzijds, als f niet continu is in a, dan

$$\exists \varepsilon > 0 \; \forall \delta > 0 \; \exists a_\delta \in A : \quad d(a_\delta, a) < \delta \text{ en } d(f(a_\delta), f(a)) \geq \varepsilon.$$

Kies nu $\delta = \frac{1}{n}$, en noem de bijbehorende a_δ nu a_n, dan is $(a_n)_{n=1}^{\infty}$ een rij in A met $a_n \to a$, terwijl de rij $(f(a_n))_{n=1}^{\infty}$ niet convergeert naar $f(a)$. Q.e.d.

Stelling 5.3. Laat X en Y metrische ruimten zijn, en $f : X \to Y$ een afbeelding. Dan is f continu op X dan en slechts dan als het inverse beeld onder f van iedere open deelverzameling van Y weer een open deelverzameling van X is.

Bewijs. Neem aan dat f continu is, en zij B een open verzameling in Y. We moeten bewijzen dat het inverse beeld onder f van B open is. We laten zien dat elk punt van $A = f^{-1}(B)$ een inwendig punt is. Laat hiertoe $a \in A$ en $b = f(a) \in B$. Omdat B open is, is b een inwendig punt van B, dus er bestaat een $\varepsilon > 0$ met $B_\varepsilon(b) \subset B$. Kies de bij ε horende δ uit de definitie van continuiteit. Dan $f(B_\delta(a)) \subset B_\varepsilon(b)$ waardoor $B_\delta(a) \subset A = f^{-1}(B)$. Dit geldt voor elke $a \in A$, dus A is open.

Omgekeerd, als het inverse beeld van elke open verzameling open is, is te bewijzen dat f continu is in elk punt van X. Laat hiertoe $a \in X$ en $b = f(a)$, en zij $\varepsilon > 0$ willekeurig. Dan is $B_\varepsilon(b)$ open in Y, dus $A = f^{-1}(B_\varepsilon(b))$ is open in X, en omdat $a \in A$, is er een $\delta > 0$ zo dat $B_\delta(a) \subset A$. Met deze δ is dan aan de uitspraak in de definitie van continuiteit voldaan. Q.e.d.

Tot nu toe hebben we gesproken over afbeeldingen $f : X \to Y$. Vaak worden afbeeldingen ook functies genoemd, met name in het geval dat $Y = \mathbb{R}$. We keren nu terug naar de vraagstelling in de inleiding.

Stelling 5.4. Laat $A \subset X$, en $f : A \to \mathbb{R}$ een continue functie. Als A rijkompakt is, dan heeft f een maximum in A.

Bewijs. Met dit bewijs waren we al begonnen in de inleiding. Dus laat $(x_n)_{n=1}^{\infty}$ de rij zijn waarvan de functiewaarden het supremum M van f op A benaderen, zoals precies gemaakt in de inleiding. Omdat A rijkompakt is, heeft deze rij een convergente deelrij $(x_{n_k})_{k=1}^{\infty}$ met limiet $a \in A$. Vanwege de continuiteit van f in a is de rij $(f(x_{n_k}))_{k=1}^{\infty}$ convergent met limiet $f(a)$. Omdat een convergente rij begrensd is sluit dit de mogelijkheid $M = +\infty$ uit, zo dat

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

Dit impliceert dat de rij $(f(x_n))_{n=1}^{\infty}$ convergeert naar M. Maar een deelrij convergeert naar $f(a)$. Dus $f(a) = M$. Q.e.d.

Stelling 5.5. Laat X en Y metrische ruimten zijn, $A \subset X$, en $f : A \to Y$ een continue afbeelding. Als A rijkompakt is in X, dan is het beeld van A onder f,

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

rijkompakt in Y.

Bewijs. Opgave.

Definitie 5.6. Laat X en Y metrische ruimten zijn, $A \subset X$, en $f : A \to Y$ een afbeelding. Dan heet f *uniform continu in* A, als

$$\forall \varepsilon > 0 \; \exists \delta > 0 \; \forall x, y \in A : \quad d(x, y) < \delta \implies d(f(x), f(y)) < \varepsilon.$$

Stelling 5.7. Laat X en Y metrische ruimten zijn, $A \subset X$, en $f : A \to Y$ een continue afbeelding. Als A rijkompakt is in X, dan is f uniform continu in A.

Bewijs. Stel niet, dan

$$\exists \varepsilon > 0 \; \forall \delta > 0 \; \exists x, y \in A : \quad d(x, y) < \delta \;\text{ en }\; d(f(x), f(y)) \geq \varepsilon.$$

Kies $\delta = \frac{1}{n}$ en laat x_n en y_n de bijbehorende x en y zijn zoals in de regel hierboven. Omdat A rijkompakt is heeft de rij $(x_n)_{n=1}^{\infty}$ een convergente deelrij $(x_{n_k})_{k=1}^{\infty}$, zeg met limiet $a \in A$. Dan is ook de rij $(y_{n_k})_{k=1}^{\infty}$ convergent met dezelfde limiet. (Waarom?) Maar nu geldt

$$\lim_{k \to \infty} f(x_{n_k}) = \lim_{k \to \infty} f(y_{n_k}) = f(a),$$

terwijl $d(f(x_{n_k}), f(y_{n_k})) \geq \varepsilon$, tegenspraak. Q.e.d.

6. Het geval $X = \mathbb{R}$ en $X = \mathbb{R}^N$. We hebben gezien dat rijkompakte deelverzamelingen van X altijd gesloten en begrensd zijn. Als $X = \mathbb{R}$, geldt ook het omgekeerde. We laten dit eerst zien voor een gesloten begrensd interval.

Stelling 6.1. (Heine-Borel) Laat $a, b \in \mathbb{R}$. Dan is het gesloten begrensde interval

$$[a, b] = \{x \in \mathbb{R} : \quad a \leq x \leq b\}$$

rijkompakt.

Bewijs. Laat $(x_n)_{n=1}^{\infty}$ een rij zijn in $[a, b]$. We moeten bewijzen dat deze rij een convergente deelrij heeft met limiet in $[a, b]$. We delen hiertoe het interval in twee gelijke stukken, d.w.z. in

$$[a, \frac{a+b}{2}] \text{ en } [\frac{a+b}{2}, b].$$

Dan bevat tenminste één van deze twee intervallen voor oneindig veel waarden van n het rijelement x_n. Als we dit interval $[a_1, b_1]$ noemen, kunnen we dus een deelrij van $(x_n)_{n=1}^{\infty}$ kiezen die volledig bevat is in $[a_1, b_1]$. Noteer deze rij als $(x_n^1)_{n=1}^{\infty}$. Dit argument herhalende, krijgen we een dalende rij intervallen

$$[a, b] \supset [a_1, b_1] \supset [a_2, b_2] \supset [a_3, b_3] \supset \ldots,$$

en bijbehorende steeds verdere deelrijen, genoteerd als $(x_n^j)_{n=1}^{\infty}$, met dezelfde eigenschap, namenlijk dat elke deelrij $(x_n^j)_{n=1}^{\infty}$ steeds volledig bevat is in $[a_j, b_j]$.

Vervolgens nemen we de zogenaamde diagonaalrij, dat is de rij $(x_n^n)_{n=1}^{\infty}$. Dit is zelf weer een deelrij van de oorspronkelijke rij $(x_n)_{n=1}^{\infty}$, en er geldt dat $x_n^n \in [a_n, b_n]$.

Nu is $(a_n)_{n=1}^{\infty}$ een begrensde niet-dalende rij getallen, en $(b_n)_{n=1}^{\infty}$ een begrensde niet-stijgende rij getallen. Dus bestaan

$$\alpha = \sup_n a_n \text{ en } \beta = \inf_n b_n,$$

$$a \leq \alpha \leq \beta \leq b, \text{ en } \alpha, \beta \in [a_n, b_n] \text{ voor elk natuurlijk getal } n.$$

Bovendien is $|\beta - \alpha| \leq b_n - a_n = 2^{-n}(b - a)$ voor elk natuurlijk getal n, dus $\alpha = \beta$.

Zij $\varepsilon > 0$. Dan is er een n_ε zodat

$$\alpha - \varepsilon < a_{n_\varepsilon} \leq \alpha = \beta \leq b_{n_\varepsilon} < \beta + \varepsilon.$$

Maar dan geldt voor elke $n \geq n_\varepsilon$ dat

$$\alpha - \varepsilon < a_{n_\varepsilon} \leq a_n \leq x_n^n \leq b_n \leq b_{n_\varepsilon} < \beta + \varepsilon.$$

Dus

$$x_n^n \in (\alpha - \varepsilon, \alpha + \varepsilon), \text{ m.a.w. } |x_n^n - \alpha| < \varepsilon.$$

Omdat $\varepsilon > 0$ willekeurig was betekent dit dat de (deel)rij $(x_n^n)_{n=1}^\infty$ convergent is met limiet $\alpha \in [a, b]$. Q.e.d.

Gevolg 6.2. (Bolzano-Weierstrass) Iedere begrensde rij in $\mathbb{R}$ heeft een convergente deelrij.

Stelling 6.3. Laat $A \subset \mathbb{R}$. Dan geldt

$$A \text{ is rijkompakt } \iff A \text{ is gesloten en begrensd.}$$

Bewijs. Opgave.

Stelling 6.4. Laat A een gesloten begrensde deelverzameling van $\mathbb{R}$ zijn. Als $f : A \to \mathbb{R}$ een continue functie is, dan heeft f een maximum in A.

Bewijs. Opgave.

Stelling 6.5. Iedere Cauchyrij in $\mathbb{R}$ is convergent (m.a.w. $\mathbb{R}$ is volledig).

Bewijs. Opgave.

Stelling 6.6. (absoluut convergente reeksen zijn convergent) Als $(a_n)_{n=1}^\infty$ een rij is in $\mathbb{R}$, en

$$\sum_{n=1}^{\infty} |a_n| = \lim_{k\to\infty} \sum_{n=1}^{k} |a_n|$$

bestaat, dan bestaat ook

$$\sum_{n=1}^{\infty} a_n = \lim_{k\to\infty} \sum_{n=1}^{k} a_n.$$

Bewijs. Opgave. Hint: laat zien dat de rij $(s_k)_{k=1}^\infty$, gedefinieerd door

$$s_k = \sum_{n=1}^{k} a_n,$$

een Cauchyrij is.

De laatste vijf stellingen gelden ook voor de metrische ruimte

$$\mathbb{R}^N = \{x = (x_1, x_2, \ldots, x_N) : \ x_1, x_2, \ldots, x_N \in \mathbb{R}\}, \quad (N \in \mathbb{N})$$

met de standaard Euclidische metriek

$$d(x,y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \cdots + (x_N-y_N)^2}.$$

Dit kunnen we inzien door $\mathbb{R}^N$ op te vatten als het herhaald Cartesisch produkt van $\mathbb{R}$ met zich zelf. Echter, de produktmetriek zoals gedefinieerd in Sectie 3, is niet de Euclidische metriek, maar

$$d_{prod}(x,y) = |x_1-y_1| + |x_2-y_2| + \cdots + |x_N-y_N|.$$

Opgave 6.7. Laat zien dat deze twee metrieken equivalent zijn, en bewijs Stelling 6.2 tot en met 6.6 voor $\mathbb{R}^N$.

Appendix. Kompaktheid en rijkompaktheid.

Definitie. Laat $A \subset X$.
(i) Een door een verzameling I geindexeerde collectie (open) deelverzamelingen $\{O_i : \ i \in I\}$ van X heet een *(open) overdekking* van A als

$$A \ \subset \ \cup_{i\in I}\, O_i.$$

(ii) Als iedere open overdekking $\{O_i : \ i \in I\}$ van A een eindige deeloverdekking van A bevat, d.w.z.

$$\exists i_1, i_2, \ldots, i_m \in I : \quad A \ \subset \ \cup_{j=1,2,\ldots,m}\, O_{i_j},$$

dan heet A *kompakt* .

Stelling. Laat $A \subset X$. Dan

$$A \text{ is kompakt } \iff \ A \text{ is rijkompakt.}$$

Bewijs. ($\Longrightarrow$). Stel A is kompakt en laat $(a_n)_{n=1}^{\infty}$ en rij zijn in A. We moeten bewijzen dat A een convergente deelrij met limiet in A heeft. Stel niet, dan is er voor elke $p \in A$ een $\varepsilon_p > 0$ en een n_p zodanig dat $a_n \notin B_{\varepsilon_p}(p)$ voor alle $n > n_p$. Neem nu als open overdekking van A de zojuist gevonden open bollen, dus

$$\{B_{\varepsilon_p}(p) : \ p \in A\}.$$

Omdat A kompakt is heeft deze overdekking een eindige deeloverdekking, dus er zijn $p_1, p_2, \ldots, p_m$ in A zodat

$$A \subset B_{\varepsilon_{p_1}}(p_1) \cup B_{\varepsilon_{p_2}}(p_2) \cup \cdots \cup B_{\varepsilon_{p_m}}(p_m),$$

waardoor A op zijn hoogst $n_1 + n_2 + \cdots + n_m$ punten van de rij kan bevatten, m.a.w. de rij bevat slechts eindig veel verschillende punten van A,

en tenminste een punt moet dus oneindig vaak voorkomen. Dus kunnen we een deelrij maken waarin dit punt alleen maar voorkomt, en dit is dan de gezochte deelrij.

($\Leftarrow$) Dit heeft wat meer voeten in de aarde. Daartoe eerst het volgende.

Definitie. Laat A een deelverzameling zijn van een metrische ruimte X. Dan heet A *aftelbaar kompakt* als voor elke rij open verzamelingen $(O_n)_{n=1}^{\infty}$ met

$$A \subset \cup_{n=1}^{\infty} O_n,$$

er een index k is zo dat

$$A \subset \cup_{n=1}^{k} O_n.$$

Stelling. Als $A \subset X$ rijkompakt is, dan is A aftelbaar kompakt.

Bewijs. Neem aan dat A rijkompakt is maar niet aftelbaar kompakt. Dan is er een rij open verzamelingen $(O_n)_{n=1}^{\infty}$ van A die A geheel bedekt, en die geen eindige deeloverdekking heeft. Dat betekent dat er voor elk natuurlijk getal k een punt p_k in A is met

$$p_k \notin \cup_{n=1}^{k} O_n.$$

Omdat A rijkompakt is heeft de rij $(p_n)_{n=1}^{\infty}$ een convergente deelrij met limiet p in A. Dus moet er een m zijn waarvoor $p \in O_m$. Omdat O_m open is zijn er dus oneindig veel punten van de rij $(p_n)_{n=1}^{\infty}$ die in O_m liggen. Voor $n > m$ is dit in tegenspraak met de keuze van p_n. Q.e.d.

Definitie. Laat A een deelverzameling zijn van een metrische ruimte X. Dan heet A *totaal begrensd* als er voor elke $\varepsilon > 0$ een eindige deelverzameling $\{p_1, p_2, \ldots, p_n\}$ van A bestaat met

$$A \subset B_\varepsilon(p_1) \cup B_\varepsilon(p_2) \cup \cdots \cup B_\varepsilon(p_n).$$

Stelling. Als $A \subset X$ rijkompakt is, dan is A totaal begrensd.

Bewijs. Stel A is niet totaal begrensd. Dan is er een $\varepsilon > 0$ waarvoor geen eindige verzameling als in de definitie kan worden gevonden. Kies nu $p_1 \in A$. Inductief kiezen we nu voor $n = 1, 2, \ldots$ een punt $p_{n+1} \in A$ met de eigenschap dat

$$p_{n+1} \notin B_\varepsilon(p_1) \cup B_\varepsilon(p_2) \cup \cdots \cup B_\varepsilon(p_n).$$

Maar dan is $d(p_i, p_j) \geq \varepsilon$ voor alle $i \neq j$. Dus de rij $(p_n)_{n=1}^{\infty}$ kan geen convergente deelrij hebben, in tegenspraak met de rijkompaktheid van A. Q.e.d.

Definitie. Een metrische ruimte X heet *separabel* als er een rij is in X zo dat elk punt van X een limietpunt is van deze rij.

Stelling. Als $A \subset X$ totaal begrensd is, dan is A separabel.

Bewijs. Voor elke k bestaat er een eindige deelverzameling A_k van A zo dat elk punt van A dichter dan $\frac{1}{k}$ bij een punt van A_k ligt. Maak nu een rij door eerst de elementen van A_1 te kiezen, dan de elementen van A_2, A_3, enzovoort. Dan is elk punt van A een limietpunt van de aldus verkregen rij. Q.e.d.

Stelling. Als $A \subset X$ separabel is, dan heeft elke open overdekking van A een aftelbare deeloverdekking.

Bewijs. Laat $(p_n)_{n=1}^{\infty}$ een rij zijn in A met de eigenschap dat elk punt van A limietpunt is van deze rij, en laat $\{O_i : \ i \in I\}$ een open overdekking zijn van A. Dan is elk punt p in A bevat in een O_i. Kies nu een natuurlijk getal m zo groot dat $B_{\frac{2}{m}}(p) \subset O_i$, en daarna een n zo dat $d(p_n, p) < \frac{1}{m}$. Dan is

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

Laat J de verzameling zijn van de paren natuurlijke getallen (m, n) die we zo tegenkomen als p de verzameling A doorloopt. Dan is J aftelbaar en

$$A \subset \cup_{(m,n)\in J} B_{\frac{1}{m}}(p_n).$$

Maar voor elke $(m, n) \in J$ is er tenminste één O_i die $B_{\frac{1}{m}}(p_n)$ bevat. Kies er voor elke $(m, n) \in J$ precies één. Dit geeft een aftelbare deelcollectie die A overdekt. Q.e.d.

Uit bovenstaande stellingen volgt dat als $A \subset X$ rijkompakt is, dat A ook "gewoon" kompakt is.

Leiden, juni 1993.

37.2 Metrische ruimten

Onze genormeerde ruimten X, waaronder $\mathbb{R}, \mathbb{R}^2$ en ook $C([a,b])$ met de maximumnorm, maar helaas niet $R([a,b])$ met de 1-norm, zijn voorbeelden van metrische ruimten met het afstandsbegrip gedefinieerd door de metriek

$$(x,y) \stackrel{d}{\to} d(x,y) = |x-y|, \tag{37.1}$$

een afbeelding[1] d van $X \times X$ naar $[0,\infty)$ met de eigenschappen dat voor alle $x,y,z \in X$ geldt dat

$$(i) \quad d(x,y) = 0 \iff x = y; \quad (ii) \quad d(x,y) = d(y,x);$$

$$(iii) \quad d(x,y) \leq d(x,z) + d(z,x). \tag{37.2}$$

Iedere niet-lege deelverzameling A van X is zo een metrische ruimte, waarbij we de algebraische vectorruimte operaties nu vergeten.

Definition 37.1. *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 (37.2) hierboven gelden voor alle $x,y,z \in X$.*

Exercise 37.2. De ε, N-definitie van $d(x_n, x_m) \to 0$ als $m,n \to \infty$ definieert wat een Cauchyrij in X is. Geef die definitie. Geef ook de definitie van het convergent zijn van de rij x_n in X.

We gebruiken hieronder de notatie $x_n \to x$ voor $x_1, x_2, x_3, \ldots, x \in X$ zonder er steeds $n \to \infty$ bij te zetten en spreken over ook een rij x_n zonder te vermelden dat $n \in \mathbb{N}$ (of een andere deelverzameling van $\mathbb{Z}$ van de vorm $m + \mathbb{N}$ met $m \in \mathbb{Z}$, bijvoorbeeld $\mathbb{N}_0$).

Exercise 37.3. Een flauwe opgave om aan de de notaties, definities en axioma's te wennen: laat zien dat als $x_n \to x$ en $x_n \to y$ (alles in X) voor de limieten x en y geldt dat $x = y$. De limiet van een convergente rij is dus uniek.

Met convergente rijen kunnen we voor metrische ruimten X en Y zeggen wat het voor een afbeelding

$$F : X \to Y$$

betekent om continu te zijn in $a \in X$.

[1] De d van distance, a van afstand doen we maar niet.

Definition 37.4. *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 37.5. 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 37.6. *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 37.7. Bewijs dat rijkompakte metrische ruimten volledig zijn.

Exercise 37.8. 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 37.9. 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.

37.3 Omgevingen, open en gesloten verzamelingen

Continuiteit kunnen we ook met open verzamelingen beschrijven. In het standaardjargon heet een deelverzameling $G \subset X$ van een metrische ruimte X gesloten als voor iedere rij x_n in G met $x_n \to x \in X$ de limiet x in G zit (je kan G niet uit door limieten te nemen). Een verzameling $O \subset X$ heet open[2] als zijn complement gesloten is.

Exercise 37.10. 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 37.9 en Opgave 37.10 volgt dat de stelling over maxima en minima van continue functies op gesloten begrensde deelverzamelingen van $\mathbb{R}^{\mathrm{N}}$.

Theorem 37.11. *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 37.12. 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 37.13. 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 37.14. 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 37.12 leidt dat tot deze karakterisatie van equivalente normen op X.

Exercise 37.15. 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 37.16. 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 37.17. *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 37.18. Wel een kluifje: bewijs Stelling 37.17. 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 37.5.

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 37.19. 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 37.15 en de karakterisatie daaronder en boven van open verzamelingen met behulp bollen, gesloten of open, zoals $B^p_\varepsilon(\xi) = \{x \in \mathbb{R}^2 : |x - x_0|_p < \varepsilon\}$ met $\xi \in \mathbb{R}^2$ en $\varepsilon > 0$.

Exercise 37.20. De bollen B^1 en B^∞ zijn ook te beschrijven als doorsnijdingen van open halfvlakken van de form $K = \{x \in \mathbb{R}^2 : f(x) < b\}$ met $f : \mathbb{R}^2 \to \mathbb{R}$ lineair gegeven door $f(x) = a_1 x_1 + a_2 x_2$ en $a_1, a_2, b \in \mathbb{R}$. Laat dat zien.

Exercise 37.21. 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 37.20, 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 37.22. 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 37.13. 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 37.13. 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 37.23. 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 37.24. 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 37.25. 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 37.26. 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 $x \to \max(|x_1|, |x_2|)$.

38 Hartman-Grobman stelling

Some material prepared for this very enjoyable event (see also Section 10):

`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

$$f(x) = \frac{x}{(1-x)^{\frac{1}{7}}(1 + x + x^2 + x^3 + x^4 + x^5 + x^6 + x^7)^{\frac{1}{7}}} \quad \text{en} \quad F(x) = x^8,$$

maar dat terzijde. De afbeeldingen

$$x \to x^2 \quad \text{en} \quad x \to \frac{1}{2}(x + \frac{2}{x})$$

hebben vaste punten waar hun afgeleide 0 is en het is niet moeilijk jezelf ervan te overtuigen dat dit leidt tot snelle convergentie de rij x_n naar evenwicht, als je begint met x_0 in de buurt van een evenwicht.

Neem je zomaar een functie F om een dynamisch systeem te maken zoals hierboven, en is $F(0) = 0$, dan bepaalt de afgeleide $F'(0)$ in het algemeen of $x = 0$ een (lokaal) stabiel of onstabiel evenwicht is, zoals het voorbeeld

$$x \to \lambda x$$

met $\lambda \in \mathbb{R}$ bij inspectie meteen laat zien. Een voor de hand liggende vraag is dan of de dynamische systemen gedefinieerd door

$$x \to F(x) \quad \text{en} \quad \tilde{x} \to F'(0)\tilde{x}$$

niet eigenlijk hetzelfde zijn via een conjugatie:

$$\begin{array}{ccc} x & \stackrel{\phi}{\to} & \tilde{x} \\ F \downarrow & & \downarrow F'(0) \\ F(x) & \stackrel{\phi}{\to} & F'(0)\tilde{x} \end{array}$$

Dus is er een inverteerbare afbeelding ϕ waarmee

$$F'(0)\phi(x) = \phi(F(x))$$

voor x in een zo groot mogelijke buurt van $x = 0$? Als we $F(x)$ schrijven als

$$F(x) = \lambda x + a(x)$$

dan is de vraag dus of we gegeven $\lambda \in \mathbb{R}$ en $a : \mathbb{R} \to \mathbb{R}$ met $a'(0) = 0$ de functie $\phi : \mathbb{R} \to \mathbb{R}$ kunnen vinden zodanig dat

$$\lambda\phi(x) = \phi(\lambda x + a(x))$$

in de buurt van $x = 0$, en dit kunnen we proberen op te lossen door middel van

$$\phi_n(x) = \frac{\phi_{n-1}(\lambda x + a(x))}{\lambda} \quad \text{beginnend met} \quad \phi_0(x) = x.$$

Als we een a nemen met $a'(0) = 0$ en $a(x) = 0$ voor $|x|$ buiten een interval gedefinieerd door $|x| \leq \eta$ met η wellicht nog te kiezen, dan zien we dat de onbekende functie ϕ voor $|x| \geq \eta$ wel gegeven moet worden door $\phi(x) = x$. Hoewel? Is het duidelijk dat gegeven $\lambda \in \mathbb{R}$ uit

$$\lambda\phi(x) = \phi(\lambda x)$$

voor alle $x \in \mathbb{R}$ volgt dat $\phi(x) = x$ voor alle $x \in \mathbb{R}$? Niet meteen dus. Maar $\phi(x) = x$ doet het wel.

Terzijde, als ϕ differentieerbaar is volgt (als $\lambda \neq 0$) dat

$$\phi'(x) = \phi'(\lambda x)$$

voor alle $x \in \mathbb{R}$, en daarmee zijn heel veel ϕ' waarden gelijk aan elkaar, tenzij $|\lambda| = 1$. Als ϕ' continu is in 0 moet wel te bewijzen zijn dat $\phi'(x) = \phi'(0)$ voor alle x. En $\phi'(0) = 1$ ligt voor de hand als normaliserende voorwaarde.

Of we voor voor $a(x) \not\equiv 0$ zo'n differentieerbare ϕ wel maken is echter zeer de vraag. In het iteratieproces helpt de λ in de noemer wellicht als $|\lambda| > 1$ is. Weer terzijde, het voorbeeld met $a(x) = x^2$ laat zien dat zonder de aanname dat $a(x) \equiv 0$ voor $|x|$ groot er weinig hoop is, want we krijgen

$$\phi_1(x) = x + \frac{x^2}{\lambda},$$

$$\phi_2(x) = x + \left(\frac{1}{\lambda} + 1\right) x^2 + \frac{2x^3}{\lambda} + \frac{x^4}{\lambda^2},$$

$$\begin{aligned}\phi_3(x) = x &+ \left(\frac{1}{\lambda} + 1 + \lambda\right) x^2 + \left(\frac{2}{\lambda} + 2 + 2\,\lambda\right) x^3 + \left(\frac{1}{\lambda^2} + \frac{1}{\lambda} + 6 + \lambda\right) x^4 \\ &+ \left(\frac{6}{\lambda} + 4\right) x^5 + \left(\frac{2}{\lambda^2} + \frac{6}{\lambda}\right) x^6 + \frac{4x^7}{\lambda^2} + \frac{x^8}{\lambda^3},\end{aligned}$$

$$\begin{aligned}\phi_4(x) = x &+ \left(\frac{1}{\lambda} + 1 + \lambda + \lambda^2\right) x^2 + \left(\frac{2}{\lambda} + 2 + 4\,\lambda + 2\,\lambda^2 + 2\,\lambda^3\right) x^3 \\ &+ \left(\frac{1}{\lambda^2} + \frac{1}{\lambda} + 7 + 7\,\lambda + 6\,\lambda^3 + \lambda^4 + 7\,\lambda^2\right) x^4 + \cdots + \frac{x^{16}}{\lambda^4},\end{aligned}$$

enzovoorts.

De Stelling van Hartman Grobman gaat in het simpelste geval om de vraag of voor de afbeelding

$$(x, y) \xrightarrow{F} (\xi, \eta) = (\lambda x + a(x, y), \mu y + b(x, y))$$

het stelsel

$$\begin{aligned}\lambda\phi(x, y) &= \phi(\lambda x + a(x, y), \mu y + b(x, y)) \\ \mu\psi(x, y) &= \psi(\lambda x + a(x, y), \mu y + b(x, y))\end{aligned}$$

kunnen oplossen naar de functies ϕ, ψ onder de aanname dat

$$0 < |\mu| < 1 < |\lambda|,$$

teneinde de afbeelding F te conjugeren met de afbeelding

$$(\tilde{x}, \tilde{y}) \to (\tilde{\xi}, \tilde{\eta}) = (\lambda\tilde{x}, \mu\tilde{y}).$$

Dit zijn twee vergelijkingen, in de eerste is de onbekende de functie ϕ, in de tweede de functie ψ. Die voor ϕ lijkt op de vergelijking waarmee we begonnen en waarvoor de geschetste aanpak kans van slagen heeft als $|\lambda| > 1$. De aannamen op $a(x, y)$ en $b(x, y)$ zijn nu zoals die op $a(x)$ hierboven, dus

$$a_x(0, 0) = a_y(0, 0) = b_x(0, 0) = b_y(0, 0) = 0,$$

en $a(x, y)$ en $b(x, y)$ tenminste continu differentieerbaar. Met die conditie is het stelsel

$$\xi = \lambda x + a(x, y)$$

$$\eta = \mu y + b(x, y)$$

in de buurt van $(0, 0)$ op te lossen naar x, y in de vorm

$$x = \frac{1}{\lambda}\,\xi + \alpha(\xi, \eta)$$

$$y = \frac{1}{\mu}\,\eta + \beta(\xi, \eta)$$

met $\alpha(\xi, \eta)$ en $\beta(\xi, \eta)$ continu differentieerbaar in de buurt van $(0, 0)$ en

$$\alpha_\xi(0, 0) = \alpha_\eta(0, 0) = \beta_\xi(0, 0) = \beta_\eta(0, 0) = 0.$$

De eerste vergelijking houden we zoals die was, de tweede schrijven we in ξ, η. Beide vergelijkingen hebben dan dezelfde vorm:

$$\phi(x, y) = \frac{1}{\lambda}\phi(\lambda x + a(x, y), \mu y + b(x, y))$$

$$\psi(\xi, \eta) = \mu\psi(\frac{1}{\lambda}\xi + \alpha(\xi, \eta), \frac{1}{\mu}\eta + \beta(\xi, \eta))$$

Om deze vergelijkingen op te lossen moeten we dus eerst weten hoe we de eerdere vergelijking voor $\phi : \mathbb{R} \to \mathbb{R}$ oplossen.

Als $|a(x)| \leq \varepsilon|x|$ dan volgt

$$|\phi_1(x) - \phi_0(x)| = |\frac{1}{\lambda}(\lambda x + a(x)) - x| = |\frac{a(x)}{\lambda}| \leq \frac{\varepsilon}{\lambda}|x|,$$

en dan

$$|\phi_2(x) - \phi_1(x)| = |\frac{1}{\lambda}\phi_1(\lambda x + a(x)) - \frac{1}{\lambda}\phi_0(\lambda x + a(x))|$$

$$\leq \frac{\varepsilon}{|\lambda|^2}|\lambda x + a(x)| \leq \frac{\varepsilon(|\lambda| + \varepsilon)}{|\lambda|^2}|x|,$$

waarna

$$|\phi_3(x) - \phi_2(x)| = |\frac{1}{\lambda}\phi_2(\lambda x + a(x)) - \frac{1}{\lambda}\phi_1(\lambda x + a(x))|$$

$$\leq \frac{\varepsilon(|\lambda| + \varepsilon)}{|\lambda|^3}|\lambda x + a(x)| \leq \frac{\varepsilon(|\lambda| + \varepsilon)^2}{|\lambda|^3}|x|.$$

Zo wordt duidelijk dat

$$|\phi_n(x) - \phi_{n-1}(x)| \leq \frac{\varepsilon(\lambda + \varepsilon)^{n-1}}{|\lambda|^n}|x|,$$

niet genoeg om de rij $\phi_n(x)$ convergent te krijgen, maar als ook geldt dat $a(x) = 0$ voor $|x| \geq \eta$ dan kunnen we met $0 < \delta < 1$ de schattingen aanpassen als

$$|\phi_1(x) - \phi_0(x)| \leq \frac{\varepsilon}{|\lambda|}\eta^{1-\delta}|x|^\delta,$$

$$|\phi_2(x) - \phi_1(x)| \leq \frac{\varepsilon}{|\lambda|^2}\eta^{1-\delta}|\lambda x + a(x)|^\delta \leq \frac{\varepsilon}{|\lambda|^2}\eta^{1-\delta}(|\lambda| + \varepsilon)|x|^\delta,$$

en dan wordt duidelijk dat

$$|\phi_n(x) - \phi_{n-1}(x)| \leq \varepsilon\eta^{1-\delta}\frac{(|\lambda| + \varepsilon)^{\delta(n-1)}}{|\lambda|^n}|x|^\delta.$$

Uniforme convergentie volgt als

$$(|\lambda| + \varepsilon)^\delta < |\lambda|.$$

39 Wiskunde onder spanning

Ik bespreek hier de opgaven met een analysekarakter in het centraal examen Wiskunde B van 15 mei 2017.

1 is een kale som. De functies f en g zijn gegeven door

$$f(x) = \ln(x) \quad \text{en} \quad g(x) = \frac{1}{2e} \cdot x^2 = \frac{1}{2e}\, x^2,$$

ik neem aan met x in het domein $\mathbb{R}^+$ voor f en domein $\mathbb{R}$ voor g, en dat de punt voor vermenigvuldigen staat. Die punt laat ik dus liever weg. Gevraagd wordt een exacte berekening die laat zien dat de grafieken van f en g elkaar raken. Dan moet er dus een x_0 in het gemeenschappelijk domein $\mathbb{R}^+$ zijn waarvoor $f(x_0) = g(x_0)$ en $f(x) - g(x)$ geen tekenwisseling heeft in x_0. Bijgevolg heeft $f(x) - g(x)$ in x_0 een lokaal extremum en geldt dat $f'(x_0) - g'(x_0) = 0$ omdat $f(x) - g(x)$ differentieerbaar is in elke $x \in \mathbb{R}^+$. We vinden zulke x_0 door de vergelijkingen

$$f(x) = g(x) \quad \text{en} \quad f'(x) = g'(x)$$

gezamenlijk op te lossen, dus

$$\ln(x) = \frac{1}{2e}\, x^2 \quad \text{en} \quad \frac{1}{x} = \frac{1}{e}\, x.$$

De tweede vergelijking geeft $x^2 = e$ en heeft in $\mathbb{R}^+$ als enige oplossing $x = \sqrt{e}$. Als de gezochte x_0 bestaat dan is dus $x_0 = \sqrt{e}$.

De eerste vergelijking reduceert dan tot $\ln \sqrt{e} = \frac{1}{2e}\, e$ ofwel $\frac{1}{2} \ln(e) = \frac{1}{2e}\, e$ en dus $\ln(e) = 1$. Omdat $e = \exp(1)$ en ln de inverse functie van exp is geldt $\ln(e) = 1$. Met $x = \sqrt{e}$ is dus aan allebei de vergelijkingen voldaan. We moeten nog nagaan dat $f(x) - g(x)$ geen tekenwisseling heeft in $\sqrt{e}$. Daartoe berekenen we

$$f''(x) - g''(x) = -\frac{1}{x^2} - \frac{1}{e} \quad \text{en dus is} \quad f''(\sqrt{e}) - g''(\sqrt{e}) = -\frac{2}{e} < 0.$$

We concluderen dat $f(x) - g(x)$ in $\sqrt{e}$ een strict lokaal maximum heeft (dat ook het globale maximum blijkt bij nadere inspectie maar dat werd niet gevraagd). Klaar in 3 stappen: oplossen $f'(x) = g'(x)$, enige oplossing voldoet aan $f(x) = g(x)$ (die was niet exact op te lossen) en $f''(x) \neq g''(x)$.

NB. Leuker was geweest: voor welke $a > 0$ raken de grafieken van $\ln(x)$ en ax^2 elkaar. Verder zal het laatste stuk van de uitwerking wellicht niet verwacht worden.

2,3,4 horen bij elkaar en zijn bepaald niet kaal. De eerste formule wordt nu niet met een functie in verband gebracht en is

$$U(t) = 325\sin(100\pi t),$$

zonder de drie vermenigvuldigingspunten in de notatie. De speciale waarden $U = 230$ en $U = -230$ zijn in de grafiek van U versus t aangegeven, alsmede het tweede positieve nulpunt $t = \frac{1}{50} = 0{,}02$. Onduidelijk waarom de decimale notatie wordt gebruikt. Nog voor je verder leest komt de al of niet wiskundige gedachte op om $x = 100\pi t$ te stellen lijkt me, en U door 325 te delen. Kortom, via

$$x = 100\pi t \quad \text{en} \quad \frac{U(t)}{325} = \underbrace{u(x) = \sin(x) \quad \text{en} \quad \frac{46}{65}}_{dit\ \ dus} = \frac{230}{325}$$

oogt dit dus wat fijner.

Vraag 2 gaat nu over het meer dan 230 afwijken van $U(t)$ van 0, dus dat betreft de t-waarden waarvoor $U(t) > 230$ of $U(t) < -230$. In termen van x is dat $\sin(x) > \frac{46}{65}$ of $\sin(x) < -\frac{46}{65}$. Het eerste is op $[0, 2\pi]$ het geval voor $\arcsin(\frac{46}{65}) < x < \pi - \arcsin(\frac{46}{65})$, het tweede voor $\pi + \arcsin(\frac{46}{65}) < x < 2\pi - \arcsin(\frac{46}{65})$. Deze twee intervallen hebben samen een lengte van $2(\pi - 2\arcsin(\frac{46}{65}))$, en dat is een fractie

$$\frac{2(\pi - 2\arcsin(\frac{46}{65}))}{2\pi} = 1 - \frac{2}{\pi}\arcsin(\frac{46}{65})$$

van de lengte van het periodiciteits interval $[0, 2\pi]$, hetgeen je in een percentage kunt vertalen door intikken op een rekenmachine met een arcsin-knop en vermenigvuldigen met 100. Dat geeft 49.9 nog wat, zeg maar 50%, en dat is het bedoelde antwoord op de vraag hoeveel procent van de tijd de spanning meer dan 230 volt van 0 afwijkt. Want U was de spanning in volt, en $U(t)$ niet als ik goed lees, dus ook 0 niet maar toch weer wel. Storender: er staat niet bij welke tijd. De goede verstaander wordt geacht het inzicht te hebben dat het om een veelvoud van de periode gaat, en voor een veelvoud van een kwart periode is het dan ook goed. Voor andere tijdspannen niet.

3 gaat verder met wat op grond van het verhaaltje verwacht kan worden en introduceert U_{eff} door middel van (daar is de punt weer wel)

$$T \cdot U_{\text{eff}}^2 = \int_0^T U(t)^2 \, dt,$$

waarin T de periode is van de spanning U. Bedoeld wordt $T = \frac{1}{50}$, hetgeen voor $x = 100\pi t$ overeenkomt met periode 2π. Storend: er staat niet bij dat $U_{\text{eff}} > 0$ (al of niet in volt). Met

$$U = 325u, \quad U(t) = 325u(x) \quad \text{is} \quad U_{\text{eff}} = 325u_{\text{eff}}$$

en volgt zonder primitiveren natuurlijk dat (ik laat de punt weer weg)

$$2\pi \, u_{\text{eff}}^2 = \int_0^{2\pi} u(x)^2 \, dx = \int_0^{2\pi} \sin(x)^2 \, dx = \pi,$$

en dus

$$u_{\text{eff}} = \frac{1}{\sqrt{2}} \quad \text{en} \quad U_{\text{eff}} = \frac{325}{\sqrt{2}},$$

dat nog decimaal moet worden geschreven en afgerond de 230 uit Opgave 2 reproduceert.

Het was echter niet de bedoeling om de integraal exact uit te rekenen, hetgeen ook zonder de schaling van t naar x en U, U_{eff} naar u, u_{eff} een fluitje van misschien wel twee cent was geweest, want de GRM moest gebruikt worden. Terwijl het antwoord op de goede vraag exact is. Immers, bij spanning

$$U(t) = a \sin \omega t$$

met amplitude $a > 0$, golfgetal $\omega > 0$ en periode T gedefinieerd door $2\pi\omega = T$, is de effectieve U altijd

$$U_{\text{eff}} = \frac{a}{\sqrt{2}}$$

en daar zit $|U(t)|$ per periode de helft van de tijd boven. Zo wordt in deze twee opgaven zowel de wiskunde als de natuurkunde weer vertroebeld door het sturen naar zinloos gebruik van de GRM. En daar kan de politiek natuurlijk wel wat aan doen zeg ik na de discussie met Paul van Meenen op het BON-symposium en diens terechte punt over mobieltjes en rekenmachines.

4 gaat verder met drie U-tjes waarvan er (daar zijn de punten weer weg) twee gegeven worden,

$$U_1(t) = 325\sin(100\pi t) \quad \text{en} \quad U_2(t) = 325\sin(100\pi t - \frac{2}{3}\pi)\,,$$

in x en u

$$u_1(t) = \sin(x) \quad \text{en} \quad u_2(t) = \sin(x - \frac{2}{3}\pi)\,,$$

en vraagt naar de exacte maximale waarde van (de 325 wordt voor de zekerheid maar vast naar buitengehaald)

$$U_{\text{kracht}}(t) = U_1(t) - U_2(t) = 325(\sin(100\pi t) - \sin(100\pi t - \frac{2}{3}\pi).$$

Dat is natuurlijk 325 keer de maximale waarde van (mag de functie of formule nu weer f heten?)

$$f(x) = u_{\text{kracht}}(x) = u_1(x) - u_2(x) = \sin(x) - \sin(x - \frac{2}{3}\pi),$$

waarvoor geldt dat

$$f'(x) = \cos(x) - \cos(x - \frac{2}{3}\pi) = \cos(x) - \cos(x)\cos(\frac{2}{3}\pi) - \sin(x)\sin(\frac{2}{3}\pi)$$

$$= (1 + \frac{1}{2})\cos(x) - \frac{1}{2}\sqrt{3}\,\sin(x) = \frac{3}{2}\,\cos(x) - \frac{1}{2}\sqrt{3}\,\sin(x),$$

hetgeen 0 is als $\tan(x) = \sqrt{3}$. Dat is voor $x = \frac{\pi}{3}$ modulo π en de maximale waarde van $f(x)$ is $f(\frac{\pi}{3}) = \sin(\frac{\pi}{3}) - \sin(\frac{\pi}{3} - \frac{2}{3}\pi) = \sin(\frac{\pi}{3}) + \sin(\frac{\pi}{3}) = \sqrt{3}$. De maximizer $x = \frac{\pi}{3}$ ligt precies midden tussen de twee nulpunten en dat was wel te raden geweest. Voor U_{kracht} is het maximum dus $325\sqrt{3}$ en dat mag je zo laten staan voor deze ene keer.

NB Je kunt natuurlijk ook $f(x)$ als één sinusfunctie schrijven, maar die formules heb ik nooit uit mijn hoofd geleerd dus dit was sneller.

Waar waren alle getallen en verhaaltjes nu voor nodig? Joost mag het weten. Maar de wiskunde die getoetst wordt in de eerste vier opgaven is minimaal. Veel veredeld rekenen, dat wel. In Opgave 5 en 6 wordt dat vast goed gemaakt. We gaan verder met

7, 8, 9, 10, allemaal bijna kaal en weer zonder vermenigvuldigingspunten. De eerste gaat over

$$f(x) = \frac{1}{2}\sin(2x - \frac{2}{3}\pi) - \frac{1}{4}\sqrt{3} \quad \text{en} \quad g(x) = \sin(x - \frac{2}{3}\pi))$$

en vraagt om het maximum van $f(x) - g(x)$ waarbij x varieert tussen 0 en $\frac{2}{3}\pi$. Alleen heet x even p vanwege de nodeloos ingewikkeld uitleg. Dat lijkt dus heel erg op wat er net in Opgave 4 langskwam. Eigenlijk wordt het maximum van $|f(x)-g(x)|$ gevraagd maar omdat de grafieken er zoals altijd al bijstaan betreft het het maximum van $f(x) - g(x)$. Via

$$f'(x) - g'(x) = \cos(2x - \frac{2}{3}\pi) - \cos(x - \frac{2}{3}\pi))$$

gaat het nu wel anders verder dan in Opgave 4. Om $f'(x) - g'(x) = 0$ te krijgen moeten twee cosinussen aan elkaar gelijk zijn. Dat gaat hier alleen maar als

$$2x - \frac{2}{3}\pi = x - \frac{2}{3}\pi \quad \text{of} \quad 2x - \frac{2}{3}\pi = -(x - \frac{2}{3}\pi),$$

allebei modulo 2π, dus

$$x = 0 \mod 2\pi \quad \text{of} \quad 3x = \frac{4}{3}\pi \mod 2\pi.$$

In het gegeven interval is dat alleen $x = p = \frac{4}{9}\pi$

8 gaat over een functie die voor het eerst een domein heeft: voor $x \in [0, \pi]$ wordt de functie f gegeven door

$$f(x) = 3\sin(x) - 2\sin^2(x) = 3\sin(x) - 2(\sin(x))^2,$$

een functie van $\sin(x)$ eigenlijk. Met $s = \sin(x)$ gaat het over

$$g(s) = 3s - 2s^2 = f(x)$$

maar voor de vraagstelling is dat misschien niet handig als we verder lezen. Er staat weer een plaatje bij waaruit blijkt dat $f(x)$ een lokaal minimum 1 heeft precies in het midden, dus in $x = \frac{\pi}{2}$ (ofwel $s = 1$) en voor twee andere x-waarden ook gelijk is aan 1. In termen van $s = \sin(x)$ gaat het dus om de s-waarden waarvoor $3s - 2s^2 = 1$ of wel $2s^2 - 3s + 1 = 0$ en omdat $s = 1$ een oplossing is factoriseert dit, te weten als $(2s - 1)(s - 1)$ dus de andere oplossing is $s = \frac{1}{2}$. De bijbehorende x-waarden voldoen dus aan $\sin(x) = \frac{1}{2}$ en dat zijn $x = \frac{\pi}{6}$ en $-\frac{\pi}{6}$ in $[0, 2\pi]$. De afstand tussen de andere snijpunten van $y = f(x)$ en $y = 1$ is dus $\frac{\pi}{3}$. En dat werd exact gevraagd, dus $s = \sin(x)$ was wel handig (niet voor afstanden natuurlijk, maar het ging ook niet over afstanden eigenlijk).

9 vraagt vervolgens om

$$\int_0^\pi f(x)\,dx = 3\int_0^\pi \sin(x)\,dx - 2\int_0^\pi \sin^2(x)\,dx = 6 - \pi,$$

de eerste via primitiveren, de tweede via $\cos^2(x) + \sin^2(x) = 1$ (net als in Opgave 2, maar nu mag het zonder GR).

10 introduceert dan een functie g met

$$g(x) = ax^2 + bx$$

die voldoet aan $f(0) = g(0)$, $f'(0) = g('0)$, $f(\pi) = g(\pi)$, $f'(\pi) = g'(\pi)$, en vraagt om a en b. Beter geweest was de vraag naar een kwadratische functie met die eigenschap, want die moet natuurlijk een veelvoud zijn van $x(\pi - x)$ omdat $g(0) = f(0) = 0 = f(\pi) = g(\pi)$. Met $f'(0) = 3$ wordt dat dus $g(x) = \frac{3}{\pi}x(\pi - x) = 3x - \frac{3}{\pi}x^2$ en het antwoord volgt want in π klopt het ook.

Iets vrolijker geworden, maar nu komen de bosbranden.

11,12,13 gaan over branden en daar zijn de punten weer. Waar de context in 2,3,4 wel OK was betreft het in 11 een bepaald soort **natuurlijke brand** beschreven door het model

$$T_{\text{nat}}(t) = 20 + 1050 \cdot e^{-\ln^2(t)+6\ln(t)-9},$$

wie verzint dit vraagt een mens zich dan af. Een formule als model, je ziet het vaker in het modelleren, maar meestal is het onzin. Zo ook hier is mijn nulhypothese. De formule zelf ziet er vervaarlijk uit. De 20 zal wel kamertemperatuur zijn. Haal die er maar van af en deel dan ook maar door die uit de duim gezogen 1050. Dan blijft, zonder die punt,

$$e^{-\ln^2(t)+6\ln(t)-9} = e^{-\ln^2(t)}e^{6\ln(t)}e^{-9} = e^{-\ln(t)\ln(t)}t^6e^{-9}$$

over en haal die $e^{-9} \approx \frac{1}{8103}$ dan ook maar weg. Het gaat dan nog om

$$e^{-\ln(t)\ln(t)}t^6 = t^{6-\ln(t)},$$

je moet er maar opkomen. Omdat met

$$x = \ln(t) \quad \text{en} \quad T_{\text{nat}}(t) = 20 + 1050\, e^{-x^2+6x-9}$$

het eigenlijk gaat om

$$e^{-(x-3)^2}$$

hadden de makers (Nederlandse uitspraak hanteren hier) iets anders in gedachten. De natuurlijke brand heeft kennelijk een tijdsverloop waarin de Gauss curve is te herkennen als je de tijd op een logaritmische schaal uitzet. Nou misschien dat daar nog wel een kansmodel achter zit dan, ik zal het eens aan Ronald vragen. Hoe het ook zij, het maximum ligt bij $x = 3$ en dat geeft voor de maximale "temperatuur" exact en zonder rekenmachine $20 + 1050 = 1070$.

12 gaat over de brand in het lab:

$$T_{\text{lab}}(t) = 20 + 345 \cdot \log(8t + 1)$$

met log ipv ln, formule schijnt authentiek te zijn hoor ik van Gerard, al is onduidelijk waarvoor. Is dat de ${}^{10}\log$, of een andere ${}^{a}\log$? Gevraagd wordt algebraïsch (ja ja) naar de oplossing van

$$20 + 345 \cdot \log(8t + 1) = 300,$$

dat betreft dus de t waarvoor

$$\log(8t + 1) = \frac{300 - 20}{345} = \frac{280}{345} = \frac{56}{69},$$

een rationaal getal (of was het nu een breuk, mental note: navragen bij Jan Bergstra) met noemer 69. De gevraagde t-waarde is dus

$$t_{300} = \frac{1}{8}(a^{\frac{56}{69}} - 1)$$

met een a die je niet kunt weten. Ik gok op e of 10 en kijk naar de grafiek om mijn keus te maken. Maar die gaat over iets anders. Ik zie wel iets met 0,69. Met $a = 10$ volgt $t_{300} = 0{,}6850358$ nog wat. Dus het was de 10-log, hoe haal je het in je hoofd denk ik dan altijd.

13 gaat over de deur: als ik het goed zie gaat het om twee integralen die vergeleken moeten worden, om wat voor reden dan ook. De ene integraal is

$$I_{\text{lab}} = \int_{t_{300}}^{30} (T_{\text{lab}}(t) - 300)\, dt = \int_{t_{300}}^{30} (345\,{}^{10}\log(8t + 1) - 280)\, dt$$

en die is exact uit te rekenen maar niet gelijk aan

$$I_{\text{lab}} = \frac{625}{8}\, 10^{\frac{56}{69}} + \frac{83145}{8}\, \ln(241) - \frac{150625}{8} - 35\, 10^{\frac{56}{69}} \ln(10)$$

en dat is ongeveer 38160 maar het moest 11930 zijn, dank Gerard. Maple slikt immers de 10-log niet zomaar en moet via

$$\int_{t_{300}}^{30} (345 \frac{\ln(8t + 1)}{\ln 10} - 280)\, dt$$

gedwongen worden om het goed te doen en factoriseert om terug te pesten dan $\ln(10) = \ln(2) + \ln(5)$ om het antwoord zo onoverzichtelijk mogelijk te maken, maar er komt uiteindelijk afgerond 11930 uit.

Moet vergeleken worden met die andere integraal, en die is

$$I_{\text{nat}} = \int_{t_{300}}^{30} (T_{\text{nat}}(t) - 300)\, dt = \int_{\tau_{300}}^{30} (1050\, e^{-\ln^2(t)+6\ln(t)-9} - 280)\, dt$$

met τ_{300} de waarde van t in het eerste snijpunt van de grafiek van T_{nat} met $T = 300$. Daarvoor moet de e-macht zelf gelijk zijn aan $\frac{4}{15}$ en dat geeft voor $x = \ln(t)$ dat

$$x = 3 \pm \sqrt{\ln(\frac{15}{4})}$$

en kennelijk is

$$\tau_{300} = e^{3-\sqrt{\ln(\frac{15}{4})}}$$

want je moet de kleinste van de twee hebben. De integraal herschrijft met $t = e^x$ en $dt = e^x dx$ als

$$I_{\text{nat}} = \int_{3-\sqrt{\ln(\frac{15}{4})}}^{\ln(30)} (1050\, e^{-(x-3)^3} - 280)\, e^x dx$$

en dat wordt iets met errorfuncties, ongeveer gelijk aan 14242 en minder dan die 38160, het zal wel (maar meer dan die 11930, die Gerard en ik nu allebei hebben).

Blijft de vraag: wat hierboven was wel en wat was niet de bedoeling? Over branden ging het echt niet en de wiskunde zit verstopt onder een dikke laag van wat tegenwoordig gecijferdheid heet.

Ik pak als voorbeeld het herexamen VWO Wiskunde B van 2017. Van de 71 te verdienen punten betreffen 10 punten twee opgaven over een waterstraal uit een kraan. Figuur 1 geeft een duidelijke weergave van wat de makers in hun hoofd hebben maar strookt niet met de werkelijkheid. Het verloop van de diameter met de hoogte is immers veel kleiner dan het plaatje aangeeft. Formule (2) is juist onder de aanname dat het water recht naar beneden valt. Bij dit plaatje is dat evident (ook bij benadering) niet het geval en is het dus onjuist de snelheid v op deze manier in de formule te gebruiken, een formule die voor leerlingen die natuurkunde in hun pakket hebben vast wel betekenis heeft als voor v de verticale component van de snelheid wordt gelezen. Maar v is de snelheid zelve volgens de uitleg, en daarmee is de formule onjuist.

Formule (2) moet in Opgave 6 worden gecombineerd met Formule (1), waarin v juist wel de snelheid (dat is de lengte van de snelheidsvector) zelf is. Door beide formules te combineren kan Formule (3) worden afgeleid. Wie zich niets aantrekt van het verhaaltje zal daar met wiskunde uit de onderbouw geen moeite mee gehad hebben als het goed is. Het begeleidende verhaaltje is echter tamelijk omslachtig geformuleerd met een overbodige invoering van de spreekwoordelijke x, en de essentiële aanname dat het water verticaal een vrije val valt wordt niet geformuleerd.

In Opgave 7 wordt vervolgens om de inhoud van de waterstraal gevraagd, al wordt het anders geformuleerd. Daarvoor is alleen de net geverifieerde formule (3) nodig waarvan het kwadraat moet worden genomen en geintegreerd van $x = 0$ tot $x = h_0$. De waarden van r_0 en h_0 (straal en hoogte van de kraanopening) worden in cm (centimeters) gegeven, de uitstroomsnelheid v_0 in m/s (meters per seconde), en de valversnelling g was eerder al in m/s^2 gegeven. Een vals valstrikje alvorens het knoppendrukken kan beginnen: je moet eerst van centimeters meters maken. Alle getallen zijn decimaal gegeven. Of je van de examenmakers nu mag schrijven dat $h_0 = \frac{3}{10}$, $r_0 = \frac{1}{50}$, $v_0 = \frac{1}{2}$ weet ik niet, maar met die waarden gaat het om de eenvoudig exact uit te rekenen integraal

$$\pi r_0^2 \int_0^{h_0} \sqrt{\frac{v_0^2}{v_0^2 + 2gx}}\, dx = \frac{\pi}{2500} \int_0^{\frac{3}{10}} \frac{1}{\sqrt{1+8gx}}\, dx = \frac{\pi}{10000g}(\sqrt{1+\frac{12g}{5}} - 1),$$

waarin je $\pi = 3{,}14$ en $g = 9{,}81$ kunt invullen met je rekenmachine als je dat wilt. Helaas, @hetCvTE wil dat je *vanaf het begin de rekenmachine* gebruikt, met kommagetallen en numerieke benaderingen. Zo bestaat Opgave 7 alleen maar uit een oefening intoetsen op de rekenmachine en het niet in het valkuiltje van de fysische eenheden trappen.