The study of functions

Teo Banica

Department of Mathematics, University of Cergy-Pontoise, F-95000 Cergy-Pontoise, France. teo.banica@gmail.com

2010 *Mathematics Subject Classification.* 97I20

Key words and phrases. Functions, Derivatives

Abstract. This is an introduction to the study of real functions, $f : \mathbb{R} \to \mathbb{R}$. We first discuss motivations and examples, ways of representing functions, and with a detailed look into the basic functions, namely polynomials, and $\sin, \cos, \exp, \log$. Then we discuss continuity, with the standard results on the subject, and notably with the Weierstrass approximation theorem. We then discuss derivatives, again with the standard results on the subject, notably with the Taylor formula and its applications. Finally, we discuss integrals, with what can be done with Riemann sums, the relation with the derivatives, and with a look into more advanced functional analysis, and several variables too.

Preface

A function $f : \mathbb{R} \to \mathbb{R}$ is a device which to each real number $x \in \mathbb{R}$ associates another real number, $f(x) \in \mathbb{R}$. Basic examples of functions include $f(x) = 2x$, or $f(x) = x^2$. Further examples, which are more complicated, include $f(x) = \sin x$, or $f(x) = e^x$.

As the name indicates, a function functions. That is, once $f : \mathbb{R} \to \mathbb{R}$ is fixed, say $f(x) = 3x^2 + 1$, for having an example, give me any $x \in \mathbb{R}$, and even something quite complicated, like $x = 2\sqrt{5} - 1$, and me, or rather function f, will tell you right away that $f(x) = 64 - 2\sqrt{5}$. Which is something very satisfying, compared to the variety of things that can be purchased in stores, or on the internet, which do not necessarily function well. With our mathematical functions $f : \mathbb{R} \to \mathbb{R}$ we are into reliability, and beauty.

Needless to say, functions $f : \mathbb{R} \to \mathbb{R}$ are something useful too. Typically $x \in \mathbb{R}$ can be thought of as being the "input" for your problem, that is, the quantity that you can make vary, as a scientist or engineer, and $f(x) \in \mathbb{R}$ is your "output", that is, the quantity that you are interested in, and that you want for instance to minimize or maximize, in the context of your scientific or engineering business. And, the idea is that the abstract mathematical study of $f : \mathbb{R} \to \mathbb{R}$ will help you, in order to achieve your goals.

Very nice all this, and as a first question that you might have, given $f : \mathbb{R} \to \mathbb{R}$, what is the formula of f? Good point, and in answer, although billions and more of functions can be constructed by starting with the basic functions that we know well, and composing them, with a sample example here being $f(x) = \sin(100x) + 4e^{2x+5} + \tan(e^x + 7) + 9$, well, bad luck, we won't obtain in this way all the possible functions $f : \mathbb{R} \to \mathbb{R}$.

You will have to believe me here, and we will of course even prove this, in this book, as a theorem, once our knowledge of functions will be ripe. In short, this is how life and mathematics are, functions $f : \mathbb{R} \to \mathbb{R}$ can be quite wild, and for studying them, there is no formula allowed, and we will have to deal with them as such, $f : \mathbb{R} \to \mathbb{R}$.

Fortunately, there is an answer to this difficulty, coming from calculus, as developed by Newton, Leibnitz and others, a long time ago. Their idea was to say that, when thinking a bit, geometrically, any function $f : \mathbb{R} \to \mathbb{R}$ must be approximately linear, around each point $x \in \mathbb{R}$. Moreover, when looking at the error term, this must be approximately quadratic. And so on, and as a conclusion to this, called Taylor formula, any "reasonable"

function $f : \mathbb{R} \to \mathbb{R}$ must appear as some sort of "infinite polynomial", around each point $x \in \mathbb{R}$. Which is something extremely useful, because with this in hand, you can then go back to your scientific or engineering problems, such as the minimization or maximization problems for the output $f(x) \in \mathbb{R}$ evoked above, and eat them raw.

This book will be here for teaching you this, the theory of functions $f : \mathbb{R} \to \mathbb{R}$, developed along the above lines, following Newton and the others. That is, all basic knowledge, that you should perfectly master, as a scientist or engineer:

– Part I is introductory, with a look into polynomials, and $\sin, \cos, \exp, \log$.

– Part II discusses continuity, and approximation by continuous functions.

– Part III discusses derivatives, the Taylor formula, and basic applications.

– Part IV discusses integration, and with a look into spaces of functions too.

Most of the book will be theoretical, although we will talk about some applications too, to basic questions from physics. Also, we will include as well all sorts of useful formulae, for instance regarding all 24 basic trigonometric functions:

sin	cos	tan	sec	csc	cot
arcsin	arccos	arctan	arcsec	arccsc	arccot
sinh	cosh	tanh	sech	csch	coth
arcsinh	arccosh	arctanh	arcsech	arccsch	arccoth

Needless to say, this book will be just an introduction to the subject. For more, depending on needs and taste, you have my general calculus book [9], doing it all in a quick and intuitive way, by mixing mathematics and physics, my mesure and integration book [10], which is more rigorous and mathematical, and my physics book [11].

Many thanks to my cats, and to the other cats in this world. You have no idea how many functions and theorems must be used, in order to properly catch a mouse, and watching this daily display of mathematical knowledge has always been inspiring.

Cergy, December 2025
Teo Banica

Contents

Preface 3

Part I. Functions 9

Chapter 1. Real numbers 11
- 1a. Numbers 11
- 1b. Real numbers 17
- 1c. Sequences, limits 23
- 1d. Sums and series 28
- 1e. Exercises 32

Chapter 2. Polynomials 33
- 2a. Polynomials, roots 33
- 2b. Continuity basics 39
- 2c. Rational functions 44
- 2d. Complex numbers 50
- 2e. Exercises 56

Chapter 3. Sin and cos 57
- 3a. Angles, triangles 57
- 3b. Sine and cosine 63
- 3c. Pi, trigonometry 69
- 3d. Polar coordinates 76
- 3e. Exercises 80

Chapter 4. Exp and log 81
- 4a. The number e 81
- 4b. Euler formula 89
- 4c. The logarithm 96
- 4d. Poisson laws 100
- 4e. Exercises 104

Part II. Continuity 105

Chapter 5. Continuity, revised 107
5a. Continuity, jumps 107
5b. Topology methods 113
5c. Uniform continuity 119
5d. Complex functions 124
5e. Exercises 128

Chapter 6. Sequences, limits 129
6a. Fixed points 129
6b. Uniform convergence 136
6c. Weierstrass theorem 144
6d. Power series 148
6e. Exercises 152

Chapter 7. Special functions 153
7a. Fractionary powers 153
7b. The arcsine family 160
7c. Further trigonometry 166
7d. Hyperbolic functions 172
7e. Exercises 176

Chapter 8. Polynomials, again 177
8a. Multiple roots 177
8b. The discriminant 182
8c. Degree 3 equations 189
8d. Degree 4 equations 194
8e. Exercises 200

Part III. Derivatives 201

Chapter 9. Derivatives, rules 203
9a. Derivatives 203
9b. Theorems, rules 209
9c. Basic functions 213
9d. Local extrema 220
9e. Exercises 224

Chapter 10. Second derivatives 225
10a. Second derivatives 225
10b. Basic examples 230
10c. Taylor formula 236
10d. Convex functions 244
10e. Exercises 248

Chapter 11. Taylor formula 249
11a. Higher derivatives 249
11b. Taylor formula 254
11c. Arctangent, Leibnitz 262
11d. Bernoulli, Euler 268
11e. Exercises 272

Chapter 12. Differential equations 273
12a. Differential equations 273
12b. Parabolas, pendulum 281
12c. Harmonic oscillators 287
12d. Falls, waves, heat 292
12e. Exercises 296

Part IV. Integrals 297

Chapter 13. Integration theory 299
13a. Integration theory 299
13b. Riemann sums 305
13c. Primitives, rules 311
13d. Areas, volumes 316
13e. Exercises 320

Chapter 14. Heavy analysis 321
14a. Gauss, Fresnel 321
14b. Wallis, Stirling 328
14c. Normal variables 335
14d. Gamma, zeta, eta 340
14e. Exercises 344

Chapter 15. Function spaces 345

15a. Function spaces 345
15b. Fourier, Parseval 353
15c. Fourier transform 359
15d. Distributions 366
15e. Exercises 368

Chapter 16. Several variables 369
16a. Linear algebra 369
16b. Partial derivatives 375
16c. Kepler and Newton 383
16d. Spherical integrals 389
16e. Exercises 392

Bibliography 393

Index 397

Part I

Functions

Don't you know, things can change
Things will go your way
If you hold on
For one more day

CHAPTER 1

Real numbers

1a. Numbers

Welcome to functions, and mathematical analysis. We will be interested in this book in the functions $f : \mathbb{R} \to \mathbb{R}$, with such a function being a device which to each real number $x \in \mathbb{R}$ associates another real number, $f(x) \in \mathbb{R}$. As a basic example, $f(x) = x^2$.

In order to properly talk about functions, we first need to know more about the reals $x \in \mathbb{R}$ themselves. So, we will be talking about this, real numbers and their properties, in this chapter, and leave the functions for later, starting with chapter 2.

Getting started, you surely know about $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$, with $\mathbb{N}$ standing for "natural", and about $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ too, with $\mathbb{Z}$ standing for "zahlen", German for numbers. As a continuation of this, that you know too, we have:

DEFINITION 1.1. *The rational numbers are the quotients of type $r = a/b$, with $a, b \in \mathbb{Z}$ and $b \neq 0$, identified according to the usual rule for quotients, namely:*

$$\frac{a}{b} = \frac{c}{d} \iff ad = bc$$

These numbers add and multiply according to the usual rules for fractions, namely:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \quad , \quad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}$$

We denote the set of rational numbers by $\mathbb{Q}$, standing for "quotients".

As a first observation, we have inclusions of sets as follows, with each integer $a \in \mathbb{Z}$ being identified with the corresponding fraction $a/1 \in \mathbb{Q}$:

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

Moreover, these inclusions of sets agree with the respective addition and multiplication operations, so we have in fact inclusions of mathematical structures, as follows:

$$(\mathbb{N}, +, \cdot) \subset (\mathbb{Z}, +, \cdot) \subset (\mathbb{Q}, +, \cdot)$$

Observe that, a bit in the same way as $\mathbb{N} \to \mathbb{Z}$ appears as to be able to solve $a + x = b$, by setting $x = b - a \in \mathbb{Z}$, the passage $\mathbb{Z} \to \mathbb{Q}$ appears as to be able to solve $ax = b$, by setting $x = b/a \in \mathbb{Q}$. Thus, the constructions $\mathbb{N} \to \mathbb{Z} \to \mathbb{Q}$ are something very natural, starting with $(\mathbb{N}, +, \cdot)$, and with $(\mathbb{Q}, +, \cdot)$ appearing to be the end product.

As an important addition now to Definition 1.1, we have:

DEFINITION 1.2 (continuation). *We can talk about order on $\mathbb{Q}$, given by*

$$\frac{a}{b} > \frac{c}{d} \iff ad > bc$$

provided that both fractions are taken with positive denominators, $b, d > 0$.

Observe the similarity with the rule for addition. In fact, we have:

$$\frac{a}{b} > \frac{c}{d} \iff \frac{a}{b} - \frac{c}{d} > 0$$

Time for a first mathematical result, using these notions? Here that result is:

PROPOSITION 1.3. *We have the following inequality,*

$$a^2 + b^2 \geq 2ab$$

valid for any two rational numbers $a, b \in \mathbb{Q}$.

PROOF. We have indeed the following computation, based on Definition 1.1:

$$\begin{aligned}(a-b)^2 &= (a-b)(a-b)\\ &= a^2 - ab - ba + b^2\\ &= a^2 + b^2 - 2ab\end{aligned}$$

Now from $(a-b)^2 \geq 0$, coming from Definition 1.2, we get the result. □

The above result is quite interesting, and its proof suggests looking, more generally, at quantities of type $(a-b)^n$, who knows what these can teach us. However, this quickly leads into some sort of nightmare, so we have, as question to be solved:

$$(a-b)^n = \underbrace{(a-b)(a-b)\ldots(a-b)}_{n \text{ terms}} = ?$$

In order to discuss this, let us start with the following key result, which is something very useful, and that we will see in a moment to be related to the above question:

THEOREM 1.4. *The number of possibilities of choosing k objects among n objects is*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

called binomial number, where $n! = 1 \cdot 2 \cdot 3 \ldots (n-2)(n-1)n$, called "factorial n".

PROOF. This is something quite tricky, the idea being as follows:

(1) Imagine a set consisting of n objects. We have then n possibilities for choosing our 1st object, then $n-1$ possibilities for choosing our 2nd object, out of the $n-1$ objects

left, and so on up to $n-k+1$ possibilities for choosing our k-th object, out of the $n-k+1$ objects left. Since the possibilities multiply, the total number of choices is:

$$\begin{aligned} N &= n(n-1)\ldots(n-k+1) \\ &= n(n-1)\ldots(n-k+1)\cdot\frac{(n-k)(n-k-1)\ldots 2\cdot 1}{(n-k)(n-k-1)\ldots 2\cdot 1} \\ &= \frac{n(n-1)\ldots 2\cdot 1}{(n-k)(n-k-1)\ldots 2\cdot 1} \\ &= \frac{n!}{(n-k)!} \end{aligned}$$

(2) However, thinking a bit, this number N that we computed is in fact the number of possibilities of choosing k ordered objects among n objects. Thus, we must divide everything by the number M of orderings of the k objects that we chose:

$$\binom{n}{k}=\frac{N}{M}$$

(3) In order to compute now the missing number M, imagine a set consisting of k objects. There are k choices for the object to be designated #1, then $k-1$ choices for the object to be designated #2, and so on up to 1 choice for the object to be designated $\#k$. We conclude that $M=k(k-1)\ldots 2\cdot 1=k!$, so that we have, as claimed:

$$\binom{n}{k}=\frac{n!/(n-k)!}{k!}=\frac{n!}{k!(n-k)!}$$

(4) Finally, in order for everything to work fine, with our theorem, we must complement what was said in the statement with the following convention:

$$0!=1$$

Indeed, we obviously have $\binom{n}{n}=1$, and if we want to recover this via our general formula, we must declare that $0!=1$, as for the following computation to work:

$$\binom{n}{n}=\frac{n!}{n!0!}=\frac{n!}{n!\times 1}=1$$

And with this, our discussion on this subject is now complete. □

As a continuation of Theorem 1.4, solving our previous questions, we have:

THEOREM 1.5. *We have the binomial formula*

$$(a+b)^n=\sum_{k=0}^{n}\binom{n}{k}a^kb^{n-k}$$

valid for any two numbers $a,b\in\mathbb{Q}$.

PROOF. We have to compute the following quantity, with n terms in the product:

$$(a+b)^n=(a+b)(a+b)\ldots(a+b)$$

When expanding, we obtain a certain sum of products of a,b variables, with each such product being a quantity of type a^kb^{n-k}. Thus, we have a formula as follows:

$$(a+b)^n=\sum_{k=0}^{n}C_ka^kb^{n-k}$$

In order to finish, it remains to compute the coefficients C_k. But, according to our product formula, C_k is the number of choices for the k needed a variables among the n available a variables. Thus, according to Theorem 1.4, we have:

$$C_k=\binom{n}{k}$$

We are therefore led to the formula in the statement. □

Theorem 1.5 is something quite interesting, so let us doublecheck it with some numerics. At small values of n we obtain the following formulae, which are all correct:

$$(a+b)^0=1$$

$$(a+b)^1=a+b$$

$$(a+b)^2=a^2+2ab+b^2$$

$$(a+b)^3=a^3+3a^2b+3ab^2+b^3$$

$$(a+b)^4=a^4+4a^3b+6a^2b^2+4ab^3+b^4$$

$$(a+b)^5=a^5+5a^4b+10a^3b^2+10a^2b^3+5ab^4+b^5$$

$$\ldots$$

Now observe that in these formulae, say for memorization purposes, the powers of the a,b variables are something very simple, that can be recovered right away. What matters are the coefficients, which are the binomial coefficients $\binom{n}{k}$, which form a triangle. So, it is enough to memorize this triangle, and this can be done by using:

THEOREM 1.6. *The Pascal triangle, formed by the binomial coefficients* $\binom{n}{k}$,

$$1$$

$$1\ ,\ 1$$

$$1\ ,\ 2\ ,\ 1$$

$$1\ ,\ 3\ ,\ 3\ ,\ 1$$

$$1\ ,\ 4\ ,\ 6\ ,\ 4\ ,\ 1$$

$$1\ ,\ 5\ ,\ 10\ ,\ 10\ ,\ 5\ ,\ 1$$

$$\ldots$$

has the property that each entry is the sum of the two entries above it.

PROOF. In practice, the theorem states that the following formula holds:

$$\binom{n}{k}=\binom{n-1}{k-1}+\binom{n-1}{k}$$

There are many ways of proving this formula, all instructive, as follows:

(1) Brute-force computation. We have indeed, as desired:

$$\begin{aligned}\binom{n-1}{k-1}+\binom{n-1}{k}&=\frac{(n-1)!}{(k-1)!(n-k)!}+\frac{(n-1)!}{k!(n-k-1)!}\\&=\frac{(n-1)!}{(k-1)!(n-k-1)!}\left(\frac{1}{n-k}+\frac{1}{k}\right)\\&=\frac{(n-1)!}{(k-1)!(n-k-1)!}\cdot\frac{n}{k(n-k)}\\&=\binom{n}{k}\end{aligned}$$

(2) Algebraic proof. We have the following formula, to start with:

$$(a+b)^n=(a+b)^{n-1}(a+b)$$

By using the binomial formula, this formula becomes:

$$\sum_{k=0}^{n}\binom{n}{k}a^kb^{n-k}=\left[\sum_{r=0}^{n-1}\binom{n-1}{r}a^rb^{n-1-r}\right](a+b)$$

Now let us perform the multiplication on the right. We obtain a certain sum of terms of type a^kb^{n-k}, and to be more precise, each such a^kb^{n-k} term can either come from the $\binom{n-1}{k-1}$ terms $a^{k-1}b^{n-k}$ multiplied by a, or from the $\binom{n-1}{k}$ terms a^kb^{n-1-k} multiplied by b. Thus, the coefficient of a^kb^{n-k} on the right is $\binom{n-1}{k-1}+\binom{n-1}{k}$, as desired.

(3) Combinatorics. Let us count k objects among n objects, with one of the n objects having a hat on top. Obviously, the hat has nothing to do with the count, and we obtain $\binom{n}{k}$. On the other hand, we can say that there are two possibilities. Either the object with hat is counted, and we have $\binom{n-1}{k-1}$ possibilities here, or the object with hat is not counted, and we have $\binom{n-1}{k}$ possibilities here. Thus $\binom{n}{k}=\binom{n-1}{k-1}+\binom{n-1}{k}$, as desired. □

The binomial numbers, as constructed above, are quite fascinating objects, and the more you know about them, the better your mathematics will be. To start with, here are some basic formulae for binomial coefficients, that you should definitely memorize:

$$\binom{n}{2}=\frac{n(n-1)}{2}\quad,\quad\binom{n}{3}=\frac{n(n-1)(n-2)}{6}\quad,\quad\ldots$$

As a useful complement to these various particular cases, we have as well:

DEFINITION 1.7. *The central binomial coefficients are the following numbers,*

$$D_n = \binom{2n}{n}$$

which numerically are $1, 2, 6, 20, 70, 252, 924, 3432, 12870, 48620, \ldots$

Now if you are a number theory nerd, and hope so are you, one of the first things that you will discover is that these central binomial coefficients factorize as follows:

$$\begin{aligned} D_n &= 1 \times 1,\ 2 \times 1,\ 3 \times 2,\ 4 \times 5,\ 5 \times 14,\ 6 \times 42, \\ &\quad 7 \times 132,\ 8 \times 429,\ 9 \times 1430,\ 10 \times 4862,\ \ldots \end{aligned}$$

Thus, we are led in this way to the following conjecture:

CONJECTURE 1.8. *The central binomial coefficients factorize as*

$$D_n = (n+1)C_n$$

with $C_n = 1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, \ldots$ *being certain integers.*

However, this does not look easy to prove, and we will leave it for later. Moving away now from these difficulties, here is a bright application of our techniques:

THEOREM 1.9. *We have the following congruence, for any prime* p*,*

$$a^p = a(p)$$

called Fermat's little theorem.

PROOF. This does not look easy to prove directly, with bare hands, but we can establish this by recurrence on $a \in \mathbb{N}$, using the following computation:

$$\begin{aligned} (a+1)^p &= \sum_{k=0}^{p} \binom{p}{k} a^k \\ &= a^p + 1(p) \\ &= a + 1(p) \end{aligned}$$

Here we have used the fact that all non-trivial binomial coefficients $\binom{p}{k}$ are multiples of p, as shown by a close inspection of these binomial coeffients, given by:

$$\binom{p}{k} = \frac{p(p-1)\ldots(p-k+1)}{k!}$$

Thus, we have the result for any $a \in \mathbb{N}$, and with the case $p = 2$ being trivial, we can assume $p \geq 3$, and here by using $a \to -a$ we get it for any $a \in \mathbb{Z}$, as desired. □

1b. Real numbers

Getting now to the real numbers, you are certainly familiar with them, but let us review their definition, because who knows. As a first goal, we would like to construct a number $x = \sqrt{2}$ having the property $x^2 = 2$. But how to do this? Let us start with:

PROPOSITION 1.10. *There is no number $r \in \mathbb{Q}_+$ satisfying $r^2 = 2$. In fact, we have*

$$\mathbb{Q}_+ = \left\{p \in \mathbb{Q}_+ \middle| p^2 < 2\right\} \bigsqcup \left\{q \in \mathbb{Q}_+ \middle| q^2 > 2\right\}$$

with this being a disjoint union.

PROOF. In what regards the first assertion, assuming that $r = a/b$ with $a, b \in \mathbb{N}$ prime to each other satisfies $r^2 = 2$, we have $a^2 = 2b^2$, so $a \in 2\mathbb{N}$. But by using again $a^2 = 2b^2$ we obtain $b \in 2\mathbb{N}$, contradiction. As for the second assertion, this is obvious. □

It looks like we are a bit stuck. We can't really tell who $\sqrt{2}$ is, and the only piece of information about $\sqrt{2}$ that we have comes from the knowledge of the rational numbers satisfying $p^2 < 2$ or $q^2 > 2$. To be more precise, the picture that emerges is:

CONCLUSION 1.11. *The number $\sqrt{2}$ is the abstract beast which is:*

(1) *Bigger than all rationals satisfying $p^2 < 2$.*
(2) *Smaller than all positive rationals satisfying $q^2 > 2$.*

Which does not look very good, but you know what, instead of looking for more clever solutions to our problem, what about relaxing, and taking Conclusion 1.11 as a definition for $\sqrt{2}$. This leads to the following "lazy" definition for the real numbers:

DEFINITION 1.12. *The real numbers $x \in \mathbb{R}$ are formal cuts in the set of rationals,*

$$\mathbb{Q} = A_x \sqcup B_x$$

with such a cut being by definition subject to the following conditions:

$$p \in A_x \ , \ q \in B_x \implies p < q \qquad , \qquad \inf B_x \notin B_x$$

These numbers add and multiply by adding and multiplying the corresponding cuts.

This might look quite original, but believe me, there is some genius behind this definition. As a first observation, we have an inclusion $\mathbb{Q} \subset \mathbb{R}$, obtained by identifying each rational number $r \in \mathbb{Q}$ with the obvious cut that it produces, namely:

$$A_r = \left\{p \in \mathbb{Q} \middle| p \leq r\right\} \quad , \quad B_r = \left\{q \in \mathbb{Q} \middle| q > r\right\}$$

As a second observation, the addition and multiplication of real numbers, obtained by adding and multiplying the corresponding cuts, in the obvious way, is something very simple. To be more precise, in what regards the addition, the formula is as follows:

$$A_{x+y} = A_x + A_y$$

As for the multiplication, the formula here is similar, namely $A_{xy} = A_x A_y$, up to some mess with positives and negatives, which is quite easy to untangle, and with this being a good exercise. We can also talk about order between real numbers, as follows:

$$x \leq y \iff A_x \subset A_y$$

But let us perhaps leave more abstractions for later, and go back to more concrete things. As a first success of our theory, we can formulate the following result:

PROPOSITION 1.13. *The equation $x^2 = 2$ has two solutions over the reals, namely:*

(1) *The positive solution, denoted $\sqrt{2}$.*
(2) *And its negative counterpart, $-\sqrt{2}$.*

PROOF. By using $x \to -x$, it is enough to prove that $x^2 = 2$ has exactly one positive solution $\sqrt{2}$. But this is clear, because $\sqrt{2}$ can only come from the following cut:

$$A_{\sqrt{2}} = \mathbb{Q}_- \bigsqcup \left\{ p \in \mathbb{Q}_+ \middle| p^2 < 2 \right\} \quad , \quad B_{\sqrt{2}} = \left\{ q \in \mathbb{Q}_+ \middle| q^2 > 2 \right\}$$

Thus, we are led to the conclusion in the statement. □

More generally, the same method works in order to extract the square root $\sqrt{r}$ of any number $r \in \mathbb{Q}_+$, or even of any number $r \in \mathbb{R}_+$, and we have the following result:

THEOREM 1.14. *The solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{R}$ are*

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

provided that $b^2 - 4ac \geq 0$. In the case $b^2 - 4ac < 0$, there are no solutions.

PROOF. We can write our equation in the following way:

$$\begin{aligned} ax^2 + bx + c = 0 &\iff x^2 + \frac{b}{a}x + \frac{c}{a} = 0 \\ &\iff \left(x + \frac{b}{2a} \right)^2 - \frac{b^2}{4a^2} + \frac{c}{a} = 0 \\ &\iff \left(x + \frac{b}{2a} \right)^2 = \frac{b^2 - 4ac}{4a^2} \\ &\iff x + \frac{b}{2a} = \pm \frac{\sqrt{b^2 - 4ac}}{2a} \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Summarizing, we have a nice abstract definition for the real numbers, that we can certainly do some mathematics with. As a first general result now, which is something very useful, and puts us back into real life, and science and engineering, we have:

THEOREM 1.15. *The real numbers $x \in \mathbb{R}$ can be written in decimal form,*

$$x = \pm a_1 \ldots a_n . b_1 b_2 b_3 \ldots\ldots$$

with $a_i, b_i \in \{0, 1, \ldots, 9\}$, with the convention $\ldots b999 \ldots = \ldots (b+1)000 \ldots$

PROOF. This is something non-trivial, even for the rationals $x \in \mathbb{Q}$ themselves, which require some work in order to be put in decimal form, the idea being as follows:

(1) First of all, our precise claim is that any $x \in \mathbb{R}$ can be written in the form in the statement, with the integer $\pm a_1 \ldots a_n$ and then each of the digits $b_1, b_2, b_3, \ldots$ providing the best approximation of x, at that stage of the approximation.

(2) Moreover, we have a second claim as well, namely that any expression of type $x = \pm a_1 \ldots a_n . b_1 b_2 b_3 \ldots\ldots$ corresponds to a real number $x \in \mathbb{R}$, and that with the convention $\ldots b999 \ldots = \ldots (b+1)000 \ldots$, the correspondence is bijective.

(3) In order to prove now these two assertions, our first claim is that we can restrict the attention to the case $x \in [0, 1)$, and with this meaning of course $0 \leq x < 1$, with respect to the order relation for the reals discussed in the above.

(4) Getting started now, let $x \in \mathbb{R}$, coming from a cut $\mathbb{Q} = A_x \sqcup B_x$. Since the set $A_x \cap \mathbb{Z}$ consists of integers, and is bounded from above by any element $q \in B_x$ of your choice, this set has a maximal element, that we can denote $[x]$:

$$[x] = \max (A_x \cap \mathbb{Z})$$

It follows from definitions that $[x]$ has the usual properties of the integer part, namely:

$$[x] \leq x < [x] + 1$$

Thus we have $x = [x] + y$ with $[x] \in \mathbb{Z}$ and $y \in [0, 1)$, and getting back now to what we want to prove, namely (1,2) above, it is clear that it is enough to prove these assertions for the remainder $y \in [0, 1)$. Thus, we have proved (3), and we can assume $x \in [0, 1)$.

(5) So, assume $x \in [0, 1)$. We are first looking for a best approximation from below of type $0.b_1$, with $b_1 \in \{0, \ldots, 9\}$, and it is clear that such an approximation exists, simply by comparing x with the numbers $0.0, 0.1, \ldots, 0.9$. Thus, we have our first digit b_1, and then we can construct the second digit b_2 as well, by comparing x with the numbers $0.b_1 0, 0.b_1 1, \ldots, 0.b_1 9$. And so on, which finishes the proof of our claim (1).

(6) In order to prove now the remaining claim (2), let us restrict again the attention, as explained in (4), to the case $x \in [0, 1)$. First, it is clear that any expression of type $x = 0.b_1 b_2 b_3 \ldots$ defines a real number $x \in [0, 1]$, simply by declaring that the corresponding cut $\mathbb{Q} = A_x \sqcup B_x$ comes from the following set, and its complement:

$$A_x = \bigcup_{n \geq 1} \left\{ p \in \mathbb{Q} \middle| p \leq 0.b_1 \ldots b_n \right\}$$

(7) Thus, we have our correspondence between real numbers as cuts, and real numbers as decimal expressions, and we are left with the question of investigating the bijectivity of this correspondence. But here, the only bug that happens is that numbers of type $x = \ldots b999\ldots$, which produce reals $x \in \mathbb{R}$ via (6), do not come from reals $x \in \mathbb{R}$ via (5). So, in order to finish our proof, we must investigate such numbers.

(8) So, consider an expression of type $\ldots b999\ldots$ Going back to the construction in (6), we are led to the conclusion that we have the following equality:

$$A_{b999\ldots} = B_{(b+1)000\ldots}$$

Thus, at the level of the real numbers defined as cuts, we have:

$$\ldots b999\ldots = \ldots (b+1)000\ldots$$

But this solves our problem, because by identifying $\ldots b999\ldots = \ldots (b+1)000\ldots$ the bijectivity issue of our correspondence is fixed, and we are done. □

The above theorem was of course quite difficult, but this is how things are. Let us record as well the following result, coming as a useful complement to the above:

THEOREM 1.16. *A real number $r \in \mathbb{R}$ is rational precisely when*

$$r = \pm a_1\ldots a_m.b_1\ldots b_n(c_1\ldots c_p)$$

that is, when its decimal writing is periodic.

PROOF. In one sense, this follows from the following computation, which shows that a number as in the statement is indeed rational:

$$\begin{aligned} r &= \pm\frac{1}{10^n}\,a_1\ldots a_m b_1\ldots b_n.c_1\ldots c_p c_1\ldots c_p\ldots \\ &= \pm\frac{1}{10^n}\left(a_1\ldots a_m b_1\ldots b_n + c_1\ldots c_p\left(\frac{1}{10^p}+\frac{1}{10^{2p}}+\ldots\right)\right) \\ &= \pm\frac{1}{10^n}\left(a_1\ldots a_m b_1\ldots b_n + \frac{c_1\ldots c_p}{10^p-1}\right) \end{aligned}$$

As for the converse, given a rational number $r = k/l$, we can find its decimal writing by performing the usual division algorithm, k divided by l. But this algorithm will be surely periodic, after some time, so the decimal writing of r is indeed periodic, as claimed. □

Here are now some further philosophical results, regarding the passage from rationals to reals, with (1) being a true theorem, and with (2) being something less rigorous:

THEOREM 1.17. *The following happen, in relation with $\mathbb{Q} \to \mathbb{R}$:*

(1) *$\mathbb{Q}$ is countable, while $\mathbb{R}$ is not countable.*
(2) *The probability for a random $x \in \mathbb{R}$ to be rational is 0.*

PROOF. We have several things to be proved, the idea being as follows:

(1) We can count indeed the positive rationals, with some redundancies, by arranging them in a table, and snaking our way inside this table, as follows:

$$\begin{array}{cccccccc}
1/1 & \to & 1/2 & 1/3 & \to & 1/4 & 1/5 \to & 1/6 & \dots \\
\swarrow & \nearrow & & \swarrow & \nearrow & & \swarrow & & \\
2/1 & 2/2 & 2/3 & 2/4 & 2/5 & 2/6 & \dots \\
\downarrow & \nearrow & \swarrow & \nearrow & \swarrow & & & & \\
3/1 & 3/2 & 3/3 & 3/4 & 3/5 & 3/6 & \dots \\
\swarrow & \nearrow & \swarrow & & & & & & \\
4/1 & 4/2 & 4/3 & 4/4 & 4/5 & 4/6 & \dots \\
\downarrow & \nearrow & \swarrow & & & & & & \\
5/1 & 5/2 & 5/3 & 5/4 & 5/5 & 5/6 & \dots \\
\swarrow & & & & & & & & \\
6/1 & 6/2 & 6/3 & 6/4 & 6/5 & 6/6 & \dots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

Thus, after eliminating the redundancies, and then adding the negatives, which must be countable too, say via an alternating $+/-$ scheme, countability of $\mathbb{Q}$ proved.

(2) Regarding now the reals, assume by contradiction that $[0,1]$ is countable, listed as follows, and with the convention that the writings of type $\dots 999 \dots$ are avoided:

$$x_1 = 0.a_1a_2a_3\dots$$

$$x_2 = 0.b_1b_2b_3\dots$$

$$x_3 = 0.c_1c_2c_3\dots$$

$$\dots$$

Now pick digits $\sigma_1 \neq a_1$, $\sigma_2 \neq b_2$, $\sigma_3 \neq c_3$ and so on, again with a technical convention here, that these are different from 9, and define $x \in \mathbb{R}$ as follows:

$$x = 0.\sigma_1\sigma_2\sigma_3\dots$$

We have then $x \in [0,1]$, and since x is obviously not on the above list, this is a contradiction. Thus $[0,1]$ is not countable, and it follows that $\mathbb{R}$ is not countable either.

(3) Regarding the probability assertion, in order to avoid some troubles, we will prove instead that the probability for a real number $x \in [0,1]$ to be rational is 0. So, let us write the rational numbers $r \in [0,1]$ in the form of a sequence $r_1, r_2, r_3 \dots$ as follows:

$$\mathbb{Q} \cap [0,1] = \big\{r_1, r_2, r_3, \dots\big\}$$

Let us also pick a number $c > 0$. Since the probability of having $x = r_1$ is certainly smaller than $c/2$, then the probability of having $x = r_2$ is certainly smaller than $c/4$, then

the probability of having $x = r_3$ is certainly smaller than $c/8$ and so on, the probability for x to be rational satisfies the following inequality:

$$P \leq \frac{c}{2} + \frac{c}{4} + \frac{c}{8} + \ldots = c\left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots\right) = c$$

Now this being valid for any $c > 0$, we conclude that we have $P = 0$, as desired. □

As our first non-philosophical result now, generalizing Proposition 1.3, we have:

THEOREM 1.18. *We have the following inequality, for any $a_1, \ldots, a_n \geq 0$,*

$$\frac{a_1 + \ldots + a_n}{n} \geq \sqrt[n]{a_1 \ldots a_n}$$

telling us that the arithmetic mean is bigger than the geometric mean.

PROOF. This can be done in two steps, as follows:

(1) We know from Proposition 1.3 that this holds at $n = 2$, coming from:

$$(\sqrt{a} - \sqrt{b})^2 \geq 0 \implies a + b \geq 2\sqrt{ab}$$

But with this, we can prove our inequality at $n = 4$ too, as follows:

$$\begin{aligned}
\frac{a+b+c+d}{4} &= \frac{1}{2}\left(\frac{a+b}{2} + \frac{c+d}{2}\right) \\
&\geq \frac{1}{2}(\sqrt{ab} + \sqrt{cd}) \\
&\geq \sqrt{\sqrt{ab}\sqrt{cd}} \\
&= \sqrt[4]{abcd}
\end{aligned}$$

Next, we can prove our inequality, in the same way, at $n = 8$, then at $n = 16$, and so on. Thus, as a conclusion, we know how to prove the result at any $n = 2^s$.

(2) In general now, given numbers $a_1, \ldots, a_n \geq 0$, consider their arithmetic mean:

$$m = \frac{a_1 + \ldots + a_n}{n}$$

Now pick $s \in \mathbb{N}$ such that $n \leq 2^s$, and let us complete our series $a_1, \ldots, a_n$ with $2^s - n$ copies of m. The arithmetic mean stays the same, and by using (1) we obtain:

$$\begin{aligned}
m \geq \sqrt[2^s]{a_1 \ldots a_n m^{2^s - n}} &\implies m^{2^s} \geq a_1 \ldots a_n m^{2^s - n} \\
&\implies m^n \geq a_1 \ldots a_n \\
&\implies m \geq \sqrt[n]{a_1 \ldots a_n}
\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

1c. Sequences, limits

We already met, on several occasions, infinite sequences or sums, and their limits. Time now to clarify all this. Let us start with the following definition:

DEFINITION 1.19. *We say that a sequence* $\{x_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ *converges to* $x \in \mathbb{R}$ *when:*

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, |x_n - x| < \varepsilon$$

In this case, we write $\lim_{n\to\infty} x_n = x$*, or simply* $x_n \to x$*.*

This might look quite scary, at a first glance, but when thinking a bit, there is nothing scary about it. Indeed, let us try to understand, how shall we translate $x_n \to x$ into mathematical language. And here, things are quite straightforward, as follows:

– The condition $x_n \to x$ tells us that "when n is big, x_n is close to x".

– That is, this tells us that "when n is big enough, x_n gets arbitrarily close to x".

– But, "n big enough" obviously means $n \geq N$, for some $N \in \mathbb{N}$.

– And "x_n arbitrarily close to x" means $|x_n - x| < \varepsilon$, for some $\varepsilon > 0$.

– Thus, we are naturally led to the above definition, and end of the story.

But all this might sound quite theoretical, yes I know, so time perhaps for some concrete illustrations. As a basic example for all this, we have:

PROPOSITION 1.20. *We have* $1/n \to 0$*.*

PROOF. This is obvious, but let us prove it by using Definition 1.19. We have:

$$\begin{aligned}\left|\frac{1}{n} - 0\right| < \varepsilon &\iff \frac{1}{n} < \varepsilon\\ &\iff \frac{1}{\varepsilon} < n\\ &\iff \left[\frac{1}{\varepsilon}\right] < n\end{aligned}$$

Thus we can take $N = [1/\varepsilon] + 1$ in Definition 1.19, and we are done. □

There are countless other examples of limits, and more on this in a moment. Going ahead with more theory, let us complement Definition 1.19 with:

DEFINITION 1.21. *We write* $x_n \to \infty$ *when the following condition is satisfied:*

$$\forall K > 0, \exists N \in \mathbb{N}, \forall n \geq N, x_n > K$$

Similarly, we write $x_n \to -\infty$ *when the same happens, with* $x_n < -K$ *at the end.*

Again, this is something very intuitive, coming from the fact that $x_n \to \infty$ can only mean that x_n is arbitrarily big, for n big enough, and we will leave some further thinking here as an instructive exercise. As a basic illustration now for this, we have:

PROPOSITION 1.22. *We have $n^2 \to \infty$.*

PROOF. As before, this is obvious, but let us prove it via Definition 1.21. We have:

$$\begin{aligned} n^2 > K &\iff n > \sqrt{K} \\ &\iff n > [\sqrt{K}] \end{aligned}$$

Thus we can take $N = [\sqrt{K}] + 1$ in Definition 1.21, and we are done. □

We can unify and generalize Proposition 1.20 and Proposition 1.22, as follows:

PROPOSITION 1.23. *We have the following convergence,*

$$n^a \to \begin{cases} 0 & (a < 0) \\ 1 & (a = 0) \\ \infty & (a > 0) \end{cases}$$

with $n \to \infty$.

PROOF. This follows indeed by using the same standard method as in the proof of Proposition 1.20 and Proposition 1.22, first for a rational, and then for a real as well. □

Next, we have some general results about limits, summarized as follows:

THEOREM 1.24. *The following happen:*

(1) *The limit $\lim_{n\to\infty} x_n$, if it exists, is unique.*
(2) *If $x_n \to x$, with $x \in (-\infty, \infty)$, then x_n is bounded.*
(3) *If x_n is increasing or descreasing, then it converges.*
(4) *Assuming $x_n \to x$, any subsequence of x_n converges to x.*

PROOF. All this is elementary, coming from definitions:

(1) Assuming $x_n \to x$, $x_n \to y$ we have indeed, for any $\varepsilon > 0$, for n big enough:

$$|x - y| \leq |x - x_n| + |x_n - y| < 2\varepsilon$$

(2) Assuming $x_n \to x$, we have $|x_n - x| < 1$ for $n \geq N$, and so, for any $k \in \mathbb{N}$:

$$|x_k| < 1 + |x| + \sup\left(|x_1|, \ldots, |x_{n-1}|\right)$$

(3) By using $x \to -x$, it is enough to prove the result for increasing sequences. But here we can construct the limit $x \in (-\infty, \infty]$ in the following way:

$$\bigcup_{n\in\mathbb{N}} (-\infty, x_n) = (-\infty, x)$$

(4) This is clear indeed from definitions. □

Here are as well some general rules for computing limits:

THEOREM 1.25. *The following happen, with the conventions $\infty+\infty=\infty$, $\infty\cdot\infty=\infty$, $1/\infty=0$, and with the conventions that $\infty-\infty$ and $\infty\cdot 0$ are undefined:*

(1) $x_n\to x$ *implies* $\lambda x_n\to\lambda x$.
(2) $x_n\to x$, $y_n\to y$ *implies* $x_n+y_n\to x+y$.
(3) $x_n\to x$, $y_n\to y$ *implies* $x_ny_n\to xy$.
(4) $x_n\to x$ *with* $x\neq 0$ *implies* $1/x_n\to 1/x$.

PROOF. All this is again elementary, coming from definitions:

(1) This is something which is obvious from definitions.

(2) This follows indeed from the following estimate:

$$|x_n+y_n-x-y|\leq|x_n-x|+|y_n-y|$$

(3) This follows indeed from the following estimate:

$$\begin{aligned}|x_ny_n-xy| &= |(x_n-x)y_n+x(y_n-y)|\\ &\leq |x_n-x|\cdot|y_n|+|x|\cdot|y_n-y|\end{aligned}$$

(4) This is again clear, by estimating $1/x_n-1/x$, in the obvious way. □

As an application of the above rules, we have the following useful result:

THEOREM 1.26. *The $n\to\infty$ limits of quotients of polynomials are given by*

$$\lim_{n\to\infty}\frac{a_pn^p+a_{p-1}n^{p-1}+\ldots+a_0}{b_qn^q+b_{q-1}n^{q-1}+\ldots+b_0}=\lim_{n\to\infty}\frac{a_pn^p}{b_qn^q}$$

with the limit on the right being $\pm\infty$, 0, a_p/b_q, depending on the values of p,q.

PROOF. The first assertion comes from the following computation:

$$\begin{aligned}\lim_{n\to\infty}\frac{a_pn^p+a_{p-1}n^{p-1}+\ldots+a_0}{b_qn^q+b_{q-1}n^{q-1}+\ldots+b_0} &= \lim_{n\to\infty}\frac{n^p}{n^q}\cdot\frac{a_p+a_{p-1}n^{-1}+\ldots+a_0n^{-p}}{b_q+b_{q-1}n^{-1}+\ldots+b_0n^{-q}}\\ &= \lim_{n\to\infty}\frac{a_pn^p}{b_qn^q}\end{aligned}$$

As for the second assertion, this comes from Proposition 1.23. □

Getting back now to theory, some sequences which obviously do not converge, like for instance $x_n=(-1)^n$, have however "2 limits instead of 1". So let us formulate:

DEFINITION 1.27. *Given a sequence $\{x_n\}_{n\in\mathbb{N}}\subset\mathbb{R}$, we let*

$$\liminf_{n\to\infty}x_n\in[-\infty,\infty]\quad,\quad\limsup_{n\to\infty}x_n\in[-\infty,\infty]$$

to be the smallest and biggest limit of a subsequence of (x_n).

Observe that the above quantities are defined indeed for any sequence x_n. For instance, for $x_n = (-1)^n$ we obtain -1 and 1. Also, for $x_n = n$ we obtain ∞ and ∞. And so on. Of course, and generalizing the $x_n = n$ example, if $x_n \to x$ we obtain x and x.

Going ahead with more theory, here is a key result:

THEOREM 1.28. *A sequence x_n converges, with finite limit $x \in \mathbb{R}$, precisely when*

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall m, n \geq N, |x_m - x_n| < \varepsilon$$

called Cauchy condition.

PROOF. In one sense, this is clear. In the other sense, we can say for instance that the Cauchy condition forces the decimal writings of our numbers x_n to coincide more and more, with $n \to \infty$, and so we can construct a limit $x = \lim_{n\to\infty} x_n$, as desired. □

The above result is quite interesting, and as an application, we have:

THEOREM 1.29. *$\mathbb{R}$ is the completion of $\mathbb{Q}$, in the sense that it is the space of Cauchy sequences over $\mathbb{Q}$, identified when the virtual limit is the same, in the sense that:*

$$x_n \sim y_n \iff |x_n - y_n| \to 0$$

Moreover, $\mathbb{R}$ is complete, in the sense that it equals its own completion.

PROOF. There are several things going on here, the idea being as follows:

(1) Getting back to Definition 1.1, we know from there what the rational numbers are. But, as a continuation of that, we can talk about the distance between such rational numbers, as being given by the formula in the statement, namely:

$$d\left(\frac{a}{b}, \frac{c}{d}\right) = \left|\frac{a}{b} - \frac{c}{d}\right| = \frac{|ad - bc|}{|bd|}$$

(2) Very good, so let us get now into Cauchy sequences. We say that a sequence of rational numbers $\{r_n\} \subset \mathbb{Q}$ is Cauchy when the following condition is satisfied:

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, m, n \geq N \implies d(r_m, r_n) < \varepsilon$$

Here of course $\varepsilon \in \mathbb{Q}$, because we do not know yet what the real numbers are.

(3) With this notion in hand, the idea will be to define the reals $x \in \mathbb{R}$ as being the limits of the Cauchy sequences $\{r_n\} \subset \mathbb{Q}$. But since these limits are not known yet to exist to us, precisely because they are real, we must employ a trick. So, let us define instead the reals $x \in \mathbb{R}$ as being the Cauchy sequences $\{r_n\} \subset \mathbb{Q}$ themselves.

(4) The question is now, will this work. As a first observation, we have an inclusion $\mathbb{Q} \subset \mathbb{R}$, obtained by identifying each rational $r \in \mathbb{Q}$ with the constant sequence $r_n = r$. Also, we can sum and multiply our real numbers in the obvious way, namely:

$$(r_n) + (p_n) = (r_n + p_n) \quad , \quad (r_n)(p_n) = (r_n p_n)$$

We can also talk about the order between such reals, as follows:

$$(r_n) < (p_n) \iff \exists N, n \geq N \implies r_n < p_n$$

Finally, we can also solve equations of type $x^2 = 2$ over our real numbers, say by using our previous work on the decimal writing, which shows in particular that $\sqrt{2}$ can be approximated by rationals $r_n \in \mathbb{Q}$, by truncating the decimal writing.

(5) However, there is still a bug with our theory, because there are obviously more Cauchy sequences of rationals, than real numbers. In order to fix this, let us go back to the end of step (3) above, and make the following convention:

$$(r_n) = (p_n) \iff d(r_n, p_n) \to 0$$

(6) But, with this convention made, we have our theory. Indeed, the considerations in (4) apply again, with this change, and we obtain an ordered field $\mathbb{R}$, containing $\mathbb{Q}$. Moreover, the equivalence with the Dedekind cuts is something which is easy to establish, and we will leave this as an instructive exercise, and this gives all the results. □

Very nice all this, so have have now two equivalent definitions for the real numbers. Finally, getting back to the decimal writing, that can be recycled too, with some analysis know-how, and we have a third possible definition for the real numbers, as follows:

THEOREM 1.30. *The real numbers $\mathbb{R}$ can be defined as well via the decimal form*

$$x = \pm a_1 \ldots a_n . a_{n+1} a_{n+2} a_{n+3} \ldots\ldots$$

with $a_i \in \{0, 1, \ldots, 9\}$, with the usual convention for such numbers, namely

$$\ldots a999 \ldots = \ldots (a+1)000 \ldots$$

and with the sum and multiplication coming by writing such numbers as

$$x = \pm \sum_{k \in \mathbb{Z}} a_k 10^{-k}$$

and then summing and multiplying, in the obvious way.

PROOF. We can define the reals $x \in \mathbb{R}$ as being formal sums as follows, with the sum being over integers $k \in \mathbb{Z}$ assumed to be greater than a certain integer, $k \geq k_0$:

$$x = \pm \sum_{k \in \mathbb{Z}} a_k 10^{-k}$$

Now by truncating, we can see that what we have here are certain Cauchy sequences of rationals, and with a bit more work, we conclude that the $\mathbb{R}$ that we constructed is precisely the $\mathbb{R}$ that we constructed in Theorem 1.29. Thus, we get the result. □

1d. Sums and series

With the above understood, we are now ready to get into some truly interesting mathematics. Let us start with the following definition:

DEFINITION 1.31. *Given numbers $x_0, x_1, x_2, \ldots \in \mathbb{R}$, we write*

$$\sum_{n=0}^{\infty} x_n = x$$

with $x \in [-\infty, \infty]$ when $\lim_{k\to\infty} \sum_{n=0}^{k} x_n = x$.

As before with the sequences, there is some general theory that can be developed for the series, and more on this in a moment. As a first, basic example, we have:

THEOREM 1.32. *We have the "geometric series" formula*

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

valid for any $|x| < 1$. For $|x| \geq 1$, the series diverges.

PROOF. Our first claim, which comes by multiplying and simplifying, is that:

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

But this proves the first assertion, because with $k \to \infty$ we get:

$$\sum_{n=0}^{k} x^n \to \frac{1}{1-x}$$

As for the second assertion, this is clear as well from our formula above. □

Less trivial now is the following result, due to Riemann:

THEOREM 1.33. *We have the following formula:*

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots = \infty$$

In fact, $\sum_n 1/n^a$ converges for $a > 1$, and diverges for $a \leq 1$.

PROOF. We have to prove several things, the idea being as follows:

(1) The first assertion comes from the following computation:

$$\begin{aligned}
1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots &= 1+\frac{1}{2}+\left(\frac{1}{3}+\frac{1}{4}\right)+\left(\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}\right)+\ldots\\
&\geq 1+\frac{1}{2}+\left(\frac{1}{4}+\frac{1}{4}\right)+\left(\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}\right)+\ldots\\
&= 1+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\ldots\\
&= \infty
\end{aligned}$$

(2) Regarding now the second assertion, we have that at $a=1$, and so at any $a\leq 1$. Thus, it remains to prove that at $a>1$ the series converges. Let us first discuss the case $a=2$, which will prove the convergence at any $a\geq 2$. The trick here is as follows:

$$\begin{aligned}
1+\frac{1}{4}+\frac{1}{9}+\frac{1}{16}+\ldots &\leq 1+\frac{1}{3}+\frac{1}{6}+\frac{1}{10}+\ldots\\
&= 2\left(\frac{1}{2}+\frac{1}{6}+\frac{1}{12}+\frac{1}{20}+\ldots\right)\\
&= 2\left[\left(1-\frac{1}{2}\right)+\left(\frac{1}{2}-\frac{1}{3}\right)+\left(\frac{1}{3}-\frac{1}{4}\right)+\left(\frac{1}{4}-\frac{1}{5}\right)\ldots\right]\\
&= 2
\end{aligned}$$

(3) It remains to prove that the series converges at $a\in(1,2)$, and here it is enough to deal with the case of the exponents $a=1+1/p$ with $p\in\mathbb{N}$. We already know how to do this at $p=1$, and the proof at $p\in\mathbb{N}$ will be based on a similar trick. We have:

$$\sum_{n=0}^{\infty}\frac{1}{n^{1/p}}-\frac{1}{(n+1)^{1/p}}=1$$

Let us compute, or rather estimate, the generic term of this series. By using the formula $a^p-b^p=(a-b)(a^{p-1}+a^{p-2}b+\ldots+ab^{p-2}+b^{p-1})$, we have:

$$\begin{aligned}
\frac{1}{n^{1/p}}-\frac{1}{(n+1)^{1/p}} &= \frac{(n+1)^{1/p}-n^{1/p}}{n^{1/p}(n+1)^{1/p}}\\
&= \frac{1}{n^{1/p}(n+1)^{1/p}[(n+1)^{1-1/p}+\ldots+n^{1-1/p}]}\\
&\geq \frac{1}{n^{1/p}(n+1)^{1/p}\cdot p(n+1)^{1-1/p}}\\
&= \frac{1}{pn^{1/p}(n+1)}\\
&\geq \frac{1}{p(n+1)^{1+1/p}}
\end{aligned}$$

We therefore obtain the following estimate for the Riemann sum:

$$\begin{aligned}\sum_{n=0}^{\infty}\frac{1}{n^{1+1/p}} &= 1+\sum_{n=0}^{\infty}\frac{1}{(n+1)^{1+1/p}}\\ &\leq 1+p\sum_{n=0}^{\infty}\left(\frac{1}{n^{1/p}}-\frac{1}{(n+1)^{1/p}}\right)\\ &= 1+p\end{aligned}$$

Thus, we are done with the case $a=1+1/p$, which finishes the proof. □

Here is another tricky result, this time about alternating sums:

THEOREM 1.34. *We have the following convergence result:*

$$1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\ldots<\infty$$

However, when rearranging terms, we can obtain any $x\in[-\infty,\infty]$ as limit.

PROOF. Both the assertions follow from Theorem 1.33, as follows:

(1) We have the following computation, using the Riemann criterion at $a=2$:

$$\begin{aligned}1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\frac{1}{5}-\frac{1}{6}+\ldots &= \frac{1}{2}+\frac{1}{12}+\frac{1}{30}+\ldots\\ &< \frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}+\ldots\\ &< \infty\end{aligned}$$

(2) We have the following formulae, coming from the Riemann criterion at $a=1$:

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{6}+\frac{1}{8}+\ldots=\frac{1}{2}\left(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots\right)=\infty$$

$$1+\frac{1}{3}+\frac{1}{5}+\frac{1}{7}+\ldots\geq\frac{1}{2}+\frac{1}{4}+\frac{1}{6}+\frac{1}{8}+\ldots=\infty$$

Thus, both these series diverge. The point now is that, by using this, when rearranging terms in the alternating series in the statement, we can arrange for the partial sums to go arbitrarily high, or arbitrarily low, and we can obtain any $x\in[-\infty,\infty]$ as limit. □

Back now to the general case, we first have the following statement:

THEOREM 1.35. *The following hold, with the converses of* (1) *and* (2) *being wrong, and with* (3) *not holding when the assumption $x_n\geq 0$ is removed:*

(1) *If $\sum_n x_n$ converges then $x_n\to 0$.*
(2) *If $\sum_n|x_n|$ converges then $\sum_n x_n$ converges.*
(3) *If $\sum_n x_n$ converges, $x_n\geq 0$ and $x_n/y_n\to 1$ then $\sum_n y_n$ converges.*

PROOF. This is a mixture of trivial and non-trivial results, as follows:

(1) We know that $\sum_n x_n$ converges when $S_k = \sum_{n=0}^k x_n$ converges. Thus by Cauchy we have the following convergence, which gives the result:

$$x_k = S_k - S_{k-1} \to 0$$

As for the simplest counterexample for the converse, this comes from the following tricky formula of Riemann, that we know well from Theorem 1.33:

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots = \infty$$

(2) This follows again from the Cauchy criterion, by using:

$$|x_n + x_{n+1} + \ldots + x_{n+k}| \leq |x_n| + |x_{n+1}| + \ldots + |x_{n+k}|$$

As for the simplest counterexample for the converse, this comes from Theorem 1.34. Indeed, let us have a look at the formula established there, namely:

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots < \infty$$

So, definitely convergent series, but when passing to absolute values, the series diverges, due to $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots = \infty$. Thus, we have our conterexample.

(3) Again, the main assertion here is clear, coming from, for n big:

$$(1 - \varepsilon)x_n \leq y_n \leq (1 + \varepsilon)x_n$$

In what regards now the failure of the result, when the assumption $x_n \geq 0$ is removed, this is something quite tricky, the simplest counterexample being as follows:

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

To be more precise, we have $y_n/x_n \to 1$, so $x_n/y_n \to 1$ too, but according to the above-mentioned results from (1,2), modified a bit, $\sum_n x_n$ converges, while $\sum_n y_n$ diverges. □

Summarizing, we have some useful positive results about series, which are however quite trivial, along with various counterexamples to their possible modifications, which are non-trivial. Staying positive, here are some more positive results:

THEOREM 1.36. *The following happen, and in all cases, the situtation where $c = 1$ is indeterminate, in the sense that the series can converge or diverge:*

(1) *If $|x_{n+1}/x_n| \to c$, the series $\sum_n x_n$ converges if $c < 1$, and diverges if $c > 1$.*
(2) *If $\sqrt[n]{|x_n|} \to c$, the series $\sum_n x_n$ converges if $c < 1$, and diverges if $c > 1$.*
(3) *With $c = \limsup_{n\to\infty} \sqrt[n]{|x_n|}$, $\sum_n x_n$ converges if $c < 1$, and diverges if $c > 1$.*

PROOF. Again, this is a mixture of trivial and non-trivial results, as follows:

(1) Here the main assertions, regarding the cases $c < 1$ and $c > 1$, are both clear by comparing with the geometric series $\sum_n c^n$. As for the case $c = 1$, this is what happens for the Riemann series $\sum_n 1/n^a$, so we can have both convergent and divergent series.

(2) Again, the main assertions, where $c < 1$ or $c > 1$, are clear by comparing with the geometric series $\sum_n c^n$, and the $c = 1$ examples come from the Riemann series.

(3) Here the case $c < 1$ is dealt with as in (2), and the same goes for the examples at $c = 1$. As for the case $c > 1$, this is clear too, because here $x_n \to 0$ fails. □

Finally, generalizing the first assertion in Theorem 1.34, we have:

THEOREM 1.37. *If $x_n \searrow 0$ then $\sum_n (-1)^n x_n$ converges.*

PROOF. We have the $\sum_n (-1)^n x_n = \sum_k y_k$, where:

$$y_k = x_{2k} - x_{2k+1}$$

But, by drawing for instance the numbers x_i on the real line, we see that y_k are positive numbers, and that $\sum_k y_k$ is the sum of lengths of certain disjoint intervals, included in the interval $[0, x_0]$. Thus we have $\sum_k y_k \leq x_0$, and this gives the result. □

So long for convergence, and basic analysis over $\mathbb{R}$, in general. We will be back to this, with some further results, which are more specialized, whenever needed.

1e. Exercises

This was a standard introductory analysis chapter, and as exercises, we have:

EXERCISE 1.38. *Learn about the central binomial coefficients $D_k = \binom{2k}{k}$.*

EXERCISE 1.39. *Learn also about the Catalan numbers $C_k = \frac{1}{k+1}\binom{2k}{k}$.*

EXERCISE 1.40. *When is $\binom{n}{k}$ divisible by a given prime number p?*

EXERCISE 1.41. *Write a short essay, about what the real numbers really are.*

EXERCISE 1.42. *Find and fix the various potential bugs with $P(x \in \mathbb{Q}) = 0$.*

EXERCISE 1.43. *What are the allowed arithmetic operations involving $0, 1, \infty$?*

EXERCISE 1.44. *Try summing the series $\sum_n x^n$ geometrically, for x rational.*

EXERCISE 1.45. *Clarify what we said above, regarding rearranging $\sum_n (-1)^n/n$.*

As bonus exercise, find and solve 100 exercises featuring sequences and series.

CHAPTER 2

Polynomials

2a. Polynomials, roots

Welcome to functions. These are the basic objects of mathematical analysis, with their definition being something very simple and fundamental, as follows:

DEFINITION 2.1. *A real function is a correspondence as follows:*

$$f : \mathbb{R} \to \mathbb{R} \quad , \quad x \to f(x)$$

More generally, we can talk about functions $f : X \to \mathbb{R}$, with $X \subset \mathbb{R}$.

Here the first notion is indeed something very intuitive, with this covering countless functions that we already know, as for instance the usual power functions:

$$f : \mathbb{R} \to \mathbb{R} \quad , \quad f(x) = x^n$$

As for the second notion, this is something more general, which is useful too. As a basic example here, we have the inverse function, which cannot be defined at $x = 0$:

$$f : \mathbb{R} - \{0\} \to \mathbb{R} \quad , \quad f(x) = \frac{1}{x}$$

Still talking generalities, since we eventually allowed the domain to be an arbitrary set $X \subset \mathbb{R}$, why not doing the same for the image. We are led in this way into:

DEFINITION 2.2 (update). *More generally, we call real function any correspondence*

$$f : X \to Y \quad , \quad x \to f(x)$$

with $X \subset \mathbb{R}$ and $Y \subset \mathbb{R}$.

In practice, however, this update will not change much to what we already had, from Definition 2.1. Indeed, any function $f : X \to Y$ with $Y \subset \mathbb{R}$ can be regarded as a function $f : X \to \mathbb{R}$ in the obvious way, by composing it with the inclusion $Y \subset \mathbb{R}$, as follows:

$$f : X \to Y \qquad \rightsquigarrow \qquad f : X \to Y \subset \mathbb{R}$$

However, Definition 2.2 can be something useful, in relation with the notions of injectivity, or surjectivity. Consider for instance the usual square function:

$$f : \mathbb{R} \to \mathbb{R} \quad , \quad f(x) = x^2$$

This function is certainly not injective, but we can make it injective, as follows:

$$f : [0, \infty) \to \mathbb{R} \quad , \quad f(x) = x^2$$

Which is good, but this latter function is still not surjective. However, we can make it surjective, by using the framework of Definition 2.2, as follows:

$$f : [0, \infty) \to [0, \infty) \quad , \quad f(x) = x^2$$

Obviously, this latter trick, in relation with surjectivity, can work for any function, in obvious way, by setting $Y = f(X)$. Let us record this finding, as follows:

PROPOSITION 2.3. *Any function $f : X \to \mathbb{R}$ can be made into a function*

$$f : X \to Y$$

which is surjective, simply by setting $Y = f(X)$.

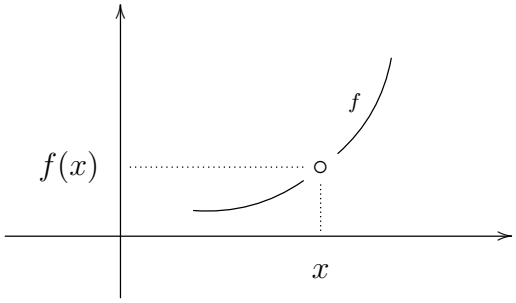
PROOF. This is indeed something clear from definitions, as explained above. □

With this done, you might perhaps ask at this point, why not pulling now a similar trick, for injectivity, a bit as we did before for $f(x) = x^2$, by restricting the domain. Well, the problem is that is not really possible, in a general way, convenient for all functions, because depending on the exact function $f : \mathbb{R} \to \mathbb{R}$ that we have in mind, restricting the domain to this or that $X \subset \mathbb{R}$, as to have f injective, remains something subjective.

Getting now to more concrete mathematics, as a first question, we have:

QUESTION 2.4. *How to suitably represent our functions $f : \mathbb{R} \to \mathbb{R}$?*

In answer to this, the graph of a function $f : \mathbb{R} \to \mathbb{R}$, which is something in 2D, drawn with the convention $y = f(x)$, is usually the best way to represent the function:



You are certainly familiar with this, drawing such graphs, so let us record here:

ANSWER 2.5. *The functions $f : \mathbb{R} \to \mathbb{R}$ are usually well represented by their graphs, drawn as usual in 2D, with the convention $y = f(x)$.*

As an illustration for the power of this method, representing functions by their graphs, we can invert quite easily the bijective functions, as follows:

THEOREM 2.6. *Given a bijective function $f : \mathbb{R} \to \mathbb{R}$, its inverse function*

$$f^{-1} : \mathbb{R} \to \mathbb{R}$$

is obtained by flipping the graph over the $x = y$ diagonal of the plane.

PROOF. This is something quite clear and intuitive, because by definition of the inverse function $f^{-1} : \mathbb{R} \to \mathbb{R}$, this is given by the following formula:

$$y = f(x) \iff f^{-1}(y) = x$$

Thus, in practice, drawing the graph of $f^{-1} : \mathbb{R} \to \mathbb{R}$ amounts in taking the graph of $f : \mathbb{R} \to \mathbb{R}$ and interchanging the coordinates, $x \leftrightarrow y$, as indicated. □

In what regards now the more general functions, $f : X \to Y$ with $X, Y \subset \mathbb{R}$, as in Definition 2.2, pretty much the same can be said here, and we have:

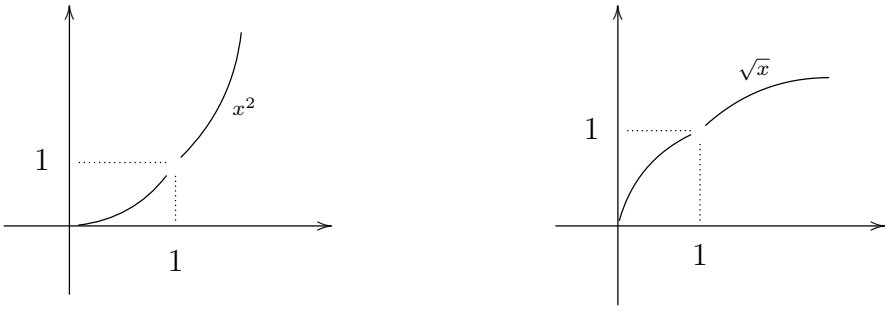
THEOREM 2.7 (update). *Given a bijective function $f : X \to Y$, its inverse function*

$$f^{-1} : Y \to X$$

is obtained by flipping the graph over the $x = y$ diagonal of the plane.

PROOF. This is indeed a straightforward generalization of Theorem 2.6. □

As a basic application, consider the square function $f(x) = x^2$, regarded as function $f : [0, \infty) \to [0, \infty)$. This function is certainly injective, being increasing, and is surjective too, with the existence of square roots coming from our Dedekind cut approach to $\mathbb{R}$. Thus the inverse function $f^{-1}(x) = \sqrt{x}$ can be obtained by flipping the graph, as follows:



We will see in what follows many other applications of the graphs of functions, for countless other questions that we can have, about them. However, and we should mention this, the graph of a function is not everything, and more precisely, we have:

WARNING 2.8. *The graph is not everything, with for instance the function*

$$f(x) = 2x$$

being best thought of as it comes, elongating all distances by 2.

With this discussed, and getting back now to more concrete mathematics, let us have a look at the simplest functions that we know, namely the degree 2 polynomials:

$$f(x) = ax^2 + bx + c$$

You certainly know that these are best represented by their graphs, which are parabolas. And, in order to draw these, we can use the following formula, from chapter 1:

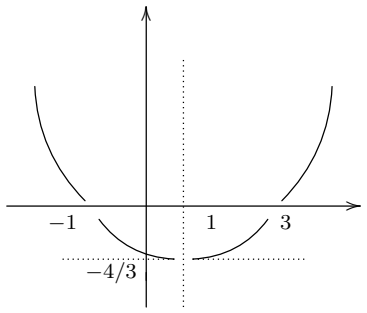
$$ax^2 + bx + c = 0 \iff \left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2}$$

To be more precise, we are led to the following method, for drawing the parabolas:

METHOD 2.9. *In order to draw the graph of $f(x) = ax^2 + bx + c$:*

(1) *We must first compute the discriminant, $\Delta = b^2 - 4ac$.*
(2) *Which leads to 4 cases, depending on whether a, Δ are positive or not.*
(3) *And so to 4 cases, regarding the position and orientation of the parabola.*
(4) *Next, we must compute $x = -b/2a$, where the symmetry axis is.*
(5) *So, we must first draw $(x, f(x))$, and then the parabola, according to (3),*
(6) *With the zeroes $(z, 0)$ with $z = (-b \pm \sqrt{\Delta})/2a$ represented too, when $\Delta \geq 0$.*

As an illustration for this, let us take $f(x) = (x^2 - 2x - 3)/3$. Here the roots are $-1, 3$, the symmetry axis is at $x = 1$, and $f(1) = -4/3$, so the parabola looks as follows:



Observe that I have in fact not bothered with computing Δ, in this example. This is because I have my own tricks for computing the zeroes, as follows:

PROPOSITION 2.10. *The roots r, s of a degree 2 equation, written as*

$$x^2 - ax + b = 0$$

can be computed by using $r + s = a$, $rs = b$.

PROOF. This is something very standard, the idea being as follows:

(1) To start with, given an arbitrary equation $Ax^2 + Bx + C = 0$, we can always divide by A, then switch the sign of B, as to reach to the above form, $x^2 - ax + b = 0$.

(2) Next, let us look for the roots r, s. These must satisfy the following equations:

$$r^2 - ar + b = 0$$

$$s^2 - as + b = 0$$

By making the difference and the sum, these equations are equivalent to:

$$(r - s)(r + s) - a(r - s) = 0$$

$$(r + s)^2 - 2rs - a(r + s) + 2b = 0$$

But, assuming that the roots are distinct, $r \neq s$, the first equation gives $r + s = a$, and with this in hand, the second equation becomes $rs = b$, as desired.

(3) Thus, result proved, modulo a discussion regarding the case $r = s$. But this case appears when $\Delta = a^2 - 4b$ vanishes, and with the common root here being $r = s = a/2$, and this fits with our equations, which are in this case $r + s = a$, $rs = a^2/4$. □

Here is an illustration for this. With the help of the general formula, we find:

$$x^2 - 8x + 15 = 0 \iff x = \frac{8 \pm \sqrt{64 - 60}}{2} = \frac{8 \pm 2}{2} = 3, 5$$

With the above trick, however, the computation is almost instant, as follows:

$$x^2 - 8x + 15 = 0 \iff r + s = 8,\ rs = 15 \iff r, s = 3, 5$$

Which is not bad, hope you agree with me here. Moving on, in order to discuss the analogue of the above trick for the arbitrary polynomials, let us start with:

PROPOSITION 2.11. *For a polynomial $P \in \mathbb{R}[X]$ and a number $r \in \mathbb{R}$, the following conditions are equivalent:*

(1) $P(r) = 0$.
(2) $P(x) = (x - r)Q$, *with* $Q \in \mathbb{R}[X]$.

PROOF. The point here is that we can divide the polynomials, a bit as we divide the integers, with an illustration for the division algorithm being as follows:

$$\begin{aligned}
& x^3 + 1 = (x + 2) * + * \\
\implies\ & x^3 + 1 = (x + 2)(x^2 + *) + * \\
\implies\ & x^3 + 1 = (x + 2)(x^2 - 2x + *) + * \\
\implies\ & x^3 + 1 = (x + 2)(x^2 - 2x + 4) + * \\
\implies\ & x^3 + 1 = (x + 2)(x^2 - 2x + 4) - 7
\end{aligned}$$

Now by dividing P by $x - r$ we are led to a formula as follows, with the quotient being a certain polynomial $Q \in \mathbb{R}[X]$, and the remainder being a constant $c \in \mathbb{R}$:

$$P(x) = (x - r)Q + c$$

But with this in hand, the equivalence in the statement is clear, by taking $x = r$. □

Now by applying this iteratively, we are led to the following key result:

THEOREM 2.12. *Any polynomial $P \in \mathbb{R}[X]$ can be written as*

$$P(x) = (x - r_1)^{n_1} \ldots (x - r_k)^{n_k} Q$$

with $r_1, \ldots, r_k \in \mathbb{R}$ being the roots, $n_1, \ldots, n_k \in \mathbb{N}$, and $Q \in \mathbb{R}[X]$ having no roots.

PROOF. This follows indeed by applying Proposition 2.11 iteratively, with the term $\prod_i (x - r_i)^{n_i}$ growing over the time, until it has to stop, due to the fact that the remainder $Q \in \mathbb{R}[X]$ becomes a constant, or more generally, a polynomial having no roots. □

As a useful complement now to Theorem 2.12, which generalizes and further clarifies what we said in Proposition 2.10 and its proof, in degree 2, we have:

THEOREM 2.13. *Given a polynomial $P \in \mathbb{R}[X]$, with leading coefficient 1,*

$$P(x) = x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$$

assuming that P has the maximum of n roots, when counted with multiplicities, so that

$$P(x) = (x - r_1) \ldots (x - r_n)$$

these roots, taken with multiplicities, satisfy $\sum_i r_i = -a_{n-1}$ and $\prod_i r_i = (-1)^n a_0$.

PROOF. This is clear indeed from the formula $P(x) = (x - r_1) \ldots (x - r_n)$, by expanding the product, and identifying the terms of degree $n - 1$, and of degree 0. □

As an illustration for all this, we can go back now to Proposition 2.10, and we have a more conceptual proof for the equations found there, coming from:

$$x^2 - ax + b = (x - r)(x - s)$$

As a second illustration, consider a degree 3 polynomial, written as follows:

$$P(x) = x^3 - ax^2 + bx - c$$

Assuming that P has its maximum of 3 roots, when counted with multiplicities, these roots r, s, t are then subject to the following formulae, which determine them:

$$r + s + t = a$$

$$rs + rt + st = b$$

$$rst = c$$

Next, as yet another basic thing about roots, that you should know too, we have:

THEOREM 2.14. *Given a polynomial $P \in \mathbb{Z}[X]$, with leading coefficient 1,*

$$P(x) = x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$$

any integer root $r \in \mathbb{Z}$ must satisfy $r | a_0$.

PROOF. This is clear indeed from $P(r) = r^n + a_{n-1} r^{n-1} + \ldots + a_1 r + a_0$, because assuming $P(r) = 0$, this reads $r(r^{n-1} + a_{n-1} r^{n-2} + \ldots + a_1) = -a_0$, which gives $r | a_0$. □

Getting back now to more concrete things, we know how to deal with degree 2. In degree 3 things get more complicated, and as a first observation, we have:

FACT 2.15. *Any degree* 3 *polynomial, say taken with leading coefficient* 1*,*

$$P = x^3 + ax^2 + bx + c$$

must have at least one root, on the grounds that P *must travel as follows:*

$$P(-\infty) = -\infty \qquad \rightsquigarrow \qquad P(\infty) = \infty$$

Moreover, the same argument should apply to any $P \in \mathbb{R}[X]$ *of odd degree.*

However, while intuitive, this is something non-trivial to prove. So, we will prove this in the next section, and then we will get back to polynomials, and their roots.

2b. Continuity basics

In order to say now a number of non-trivial general things about functions, and in particular about polynomials, let us introduce the following key notion:

DEFINITION 2.16. *A function* $f : \mathbb{R} \to \mathbb{R}$*, or more generally* $f : X \to \mathbb{R}$*, with* $X \subset \mathbb{R}$ *being a subset, is called continuous when, for any* $x_n, x \in X$*:*

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

Also, we say that $f : X \to \mathbb{R}$ *is continuous at a given point* $x \in X$ *when the above condition is satisfied, for that point* x*.*

Regarding the basic examples of continuous functions, there are many of them, and we will discuss them in a moment, once we will have some basic tools, in order to prove that this or that function is continuous or not, without much pain. As a matter, however, of having a first illustration for Definition 2.16, let us record the following fact:

PROPOSITION 2.17. *The basic power functions, namely*

$$f(x) = x^k$$

with $k \in \mathbb{N}$*, are all continuous.*

PROOF. According to Definition 2.16, we want to prove that we have:

$$x_n \to x \implies x_n^k \to x^k$$

Which looks quite clear, but you might want a rigorous proof for this. In answer:

(1) A first method is by using the results from chapter 1 regarding the sequences. To be more precise, we know from there that the following formula holds:

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

But with $x_n = y_n$, this leads to the following formula:

$$\lim_{n\to\infty} x_n^2 = \left(\lim_{n\to\infty} x_n\right)^2$$

Obviously, we can iterate this method, and so for any $k \in \mathbb{N}$, we have:

$$\lim_{n\to\infty} x_n^k = \left(\lim_{n\to\infty} x_n\right)^k$$

But now, assuming $x_n \to x$ as above, this formula gives, as desired:

$$\lim_{n\to\infty} x_n^k = x^k$$

(2) Thus, result proved, but let us try as well a second method, which is less conceptual, but is instructive too. Our idea here will be to use no idea at all. Obviously, in order to solve our question, we must estimate quantities of type $(x+t)^k - x^k$, with t small. But we can do this with the binomial formula, which gives, for $|t| \leq 1$:

$$\begin{aligned}
|(x+t)^k - x^k| &= \left|\sum_{s=1}^{k} \binom{k}{s} x^{k-s}t^s\right| \\
&\leq \sum_{s=1}^{k} \binom{k}{s} |x|^{k-s}|t|^s \\
&\leq |t| \sum_{s=1}^{k} \binom{k}{s} |x|^{k-s} \\
&\leq |t|(1+|x|)^k
\end{aligned}$$

Now assume $x_n \to x$. We can then write $x_n = x + t_n$, and by choosing our $n >> 0$ as to have $|t_n| \leq 1$, we can use the above estimate, which gives:

$$|x_n^k - x^k| \leq |t_n|(1+|x|^k)$$

Now since we have $t_n \to 0$, we obtain from this $x_n^k \to x^k$, as desired. □

Still in relation with Definition 2.16, let us record as well a counterexample:

PROPOSITION 2.18. *The basic inverse function, namely*

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

is continuous everywhere, except at 0*. That is, no matter how you pick* $\alpha \in \mathbb{R}$ *and set*

$$f(0) = \alpha$$

the function will be not continuous at 0*.*

PROOF. There are several things going on here, the idea being as follows:

(1) Let us begin with the positive statement, saying that f is continuous at any $x \neq 0$. In order to prove this, the situation is a bit similar to what we had in Proposition 2.17. Indeed, we can use the following formula for sequences, that we know from chapter 1:

$$\lim_{n\to\infty} \frac{1}{x_n} = \frac{1}{\lim_{n\to\infty} x_n}$$

Alternatively, and once again a bit similarly to what we did for Proposition 2.17, we have the 1-neuron proof, based on the following estimate, valid for $|t| < |x|/2$:

$$\left|\frac{1}{x} - \frac{1}{x+t}\right| = \left|\frac{t}{x(x+t)}\right| < \left|\frac{t}{x^2/2}\right|$$

(2) Regarding the second assertion, non-continuity at 0, assume by contradiction that by setting $f(0) = \alpha$, our function f becomes continuous at 0. But this means:

$$x_n \to 0 \implies \frac{1}{x_n} \to \alpha$$

Now with $x_n > 0$ we obtain $\alpha > 0$, and with $x_n < 0$ we obtain $\alpha < 0$. Thus, we have our contradiction, so our assumption that f is continuous was wrong, as desired. □

Getting back now to the general theory, and to Definition 2.16 as stated, many things can be said. To start with, there are many other equivalent formulations of the notion of continuity, with a well-known, useful, and much feared one, being as follows:

THEOREM 2.19. *A function $f : X \to \mathbb{R}$ is continuous when*

$$\forall x \in X, \forall \varepsilon > 0, \exists \delta > 0, |x - y| < \delta \implies |f(x) - f(y)| < \varepsilon$$

holds.

PROOF. Let us prove this, with no fear. According to Definition 2.16, in order for our function f to be continuous, the following must happen, for any $x \in X$:

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

Now when reminding what convergence of a sequence exactly means, for both the convergences $x_n \to x$ and $f(x_n) \to f(x)$, we are led to the conclusion in the statement. □

Nexrt, we have the following useful theoretical result regarding continuity:

THEOREM 2.20. *If f, g are continuous, then so are:*

(1) $f + g$.
(2) fg.
(3) f/g.
(4) $f \circ g$.

PROOF. Before anything, we should mention that the claim is that (1-4) hold indeed, provided that at the level of domains and ranges, the statement makes sense. For instance in (1,2,3) we are talking about functions having the same domain, and with $g(x) \neq 0$ for the needs of (3), and there is a similar discussion regarding (4).

(1) The claim here is that if both f, g are continuous at a point x, then so is the sum $f + g$. But this is clear from the similar result for sequences, namely:

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

(2) Again, the statement here is similar, and the result follows from:

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

(3) Here the claim is that if both f, g are continuous at x, with $g(x) \neq 0$, then f/g is continuous at x. In order to prove this, observe that by continuity, $g(x) \neq 0$ shows that $g(y) \neq 0$ for $|x - y|$ small enough. Thus we can assume $g \neq 0$, and with this assumption made, the result follows from the similar result for sequences, namely:

$$\lim_{n\to\infty} \frac{x_n}{y_n} = \frac{\lim_{n\to\infty} x_n}{\lim_{n\to\infty} y_n}$$

(4) Here the claim is that if g is continuous at x, and f is continuous at $g(x)$, then $f \circ g$ is continuous at x. But this is clear, coming from:

$$\begin{aligned} x_n \to x &\implies g(x_n) \to g(x) \\ &\implies f(g(x_n)) \to f(g(x)) \end{aligned}$$

Alternatively, let us prove this as well by using that scary ε, δ condition from Theorem 2.19. So, let us pick an arbitrary $\varepsilon > 0$. We want in the end to have something of type $|f(g(x)) - f(g(y))| < \varepsilon$, so we must first use that ε, δ condition for the function f. So, let us start in this way. Since f is continuous at $g(x)$, we can find $\delta > 0$ such that:

$$|g(x) - z| < \delta \implies |f(g(x)) - f(z)| < \varepsilon$$

On the other hand, since g is continuous at x, we can find $\gamma > 0$ such that:

$$|x - y| < \gamma \implies |g(x) - g(y)| < \delta$$

Now by combining the above two inequalities, with $z = g(y)$, we obtain:

$$|x - y| < \gamma \implies |f(g(x)) - f(g(y))| < \varepsilon$$

Thus, the composition $f \circ g$ is continuous at x, as desired. □

As a first consequence of the above result, of interest to us, we have:

THEOREM 2.21. *Any polynomial $P \in \mathbb{R}[X]$, regarded as function*

$$P : \mathbb{R} \to \mathbb{R} \quad , \quad x \to P(x)$$

is continuous, over its whole domain.

PROOF. This follows from Theorem 2.20, and from the extra fact, which is something trivial, and that I forgot to mention there, that if a function f is continuous, and $\lambda \in \mathbb{R}$ is a scalar, then the function λf is continuous too. Indeed, since any polynomial $P \in \mathbb{R}[X]$ can be obtained by starting with the function $x \to x$, which is continuous, and making sums, products, and multiplication by scalars, P must be continuous, as stated. □

All this sounds very good, and with polynomials being continuous, we can apply now to them our continuous function technology. This being said, thinking well, what exact continuous function technology? Honestly, we have nothing so far here.

So, we must develop this technology. Let us start with the main result, as follows:

THEOREM 2.22 (Intermediate value property). *Given a continuous function*

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

its image is a closed bounded interval, $Im(f) = [c, d]$.

PROOF. It is convenient to make the convention that all intervals are by definition closed and bounded, and with $[a, b]$ denoting the numbers comprised between a, b, regardless on whether $a \leq b$, or $a > b$. With this convention, the proof goes as follows:

(1) Let us first prove that f takes its intermediate values, in the sense that any $u \in [f(a), f(b)]$ belongs to $Im(f)$. In order to do so, observe that we have:

$$[f(a), f(b)] \subset \left[f(a), f\left(\frac{a+b}{2}\right)\right] \cup \left[f\left(\frac{a+b}{2}\right), f(b)\right]$$

Thus $u \in [f(a), f(b)]$ must belong to one of the intervals on the right. Now by repeating this procedure, indefinitely, we are led to a certain decreasing sequence of closed intervals I_k, with the size of these intervals halving at each step, such that:

$$u \in f(I_k) \quad , \quad \forall k$$

Now consider the limiting point x of the intervals that we found, given by:

$$\bigcap_k I_k = \{x\}$$

And with this, we are done. Indeed, by continuity of our function f at this point x, from the above condition, $u \in f(I_k)$ for any k, we obtain that we have, as desired:

$$f(x) = u$$

(2) Next, we can apply what found to the restriction $f' : [a', b'] \to \mathbb{R}$ of our function to any interval $[a', b'] \subset [a, b]$, and with a bit of thinking here, that we will leave as an exercise, this shows that the image of our function is indeed an interval, as stated. □

In practice now, Theorem 2.22 as stated is something quite compact, and in view of applications, the various findings there are best recalled as follows:

THEOREM 2.23. *The following happen for a continuous function $f : [a, b] \to \mathbb{R}$:*

(1) *f takes all intermediate values between $f(a), f(b)$.*
(2) *f has a minimum and maximum on $[a, b]$.*
(3) *If $f(a), f(b)$ have different signs, $f(x) = 0$ has a solution.*

PROOF. All these statements follow indeed from Theorem 2.22. □

And with this, good news, we have now, eventually, a rigorous proof for Fact 2.15. Let us record this, along with a little bit more, as a theorem, coming as a useful analytic complement to what we already know, of algebraic nature, about polynomials:

THEOREM 2.24. *Any polynomial $P \in \mathbb{R}[X]$ of odd degree has a root. In particular, an arbitrary degree 3 polynomial must decompose as*

$$P(x) = (x - r)Q(x)$$

with $r \in \mathbb{R}$ and with Q being of degree 2, having 0 or 2 roots.

PROOF. Here the first assertion comes from Theorem 2.23 (3), because if P has odd degree, taken with leading coefficient $c > 0$, it must travel as follows:

$$P(-\infty) = -\infty \qquad \rightsquigarrow \qquad P(\infty) = \infty$$

As for the second assertion, this is something self-explanatory, coming from this, and from our previous theory of factorization, and from what we know in degree 2. □

So long for our basic discussion of the intermediate value property. Far more things can be said, along these lines, and we will be back to this, on a more systematic basis, in Part II of this book. In the meantime, what we have in the above will do.

As for the polynomials of degree 3, what we have in Theorem 2.24 is certainly very useful, but there are far more things that can be said, on top of this. More later.

2c. Rational functions

Getting back now to the continuity basics, when thinking a bit, the various operations from Theorem 2.20 allow us to say more about polynomials, as follows:

THEOREM 2.25. *The quotients of real polynomials, called rational functions*

$$f = \frac{P}{Q}$$

are continuous on their domain. To be more precise, with $(P, Q) = 1$, the function

$$f : \mathbb{R} - P_f \to \mathbb{R} \quad , \quad x \to \frac{P(x)}{Q(x)}$$

with $P_f \subset \mathbb{R}$ being the set of zeroes Q, also called poles of f, is continuous.

PROOF. This follows indeed from Theorem 2.20, and more specifically, from the findings (1,2,3) there, and with the above convention $(P,Q)=1$ being there for having the set of poles $P_f\subset\mathbb{R}$ as small as possible. And with the extra comment that, in what regards the term "pole", this does not come from the Poles who invented this, but rather from the fact that, when drawing the graph of f, we are faced with some sort of tent, which is suspended by infinite poles, which lie, guess where, at the poles of f. □

The rational functions are something quite interesting, worth some more discussion. As a first observation, these are stable by all the operations in Theorem 2.20:

THEOREM 2.26. *The rational functions add, multiply and divide according to*

$$\frac{P}{Q}+\frac{R}{S}=\frac{PS+QR}{QS}\quad,\quad\frac{P}{Q}\cdot\frac{R}{S}=\frac{PR}{QS}\quad,\quad\frac{P}{Q}:\frac{R}{S}=\frac{PS}{QR}$$

and they are stable as well by composition, according to the computation

$$\frac{P}{Q}\circ\frac{R}{S}=\frac{P(R/S)}{Q(R/S)}=\frac{P'/S^m}{Q'/S^n}=\frac{P'S^n}{Q'S^m}$$

with m,n being the degrees of P,Q, and with P',Q' being certain polynomials.

PROOF. This is something self-explanatory, with the first three operations being similar to those for the usual numeric fractions, and with the composition computation being self-explanatory too, modulo some thinking, that we will leave as an exercise. □

As an interesting philosophical conclusion of all this, we have:

CONCLUSION 2.27. *With the rational functions being stable by all operations in Theorem 2.20, these are basically the only continuous functions that we have, so far.*

And is this good or not? I would say, good for pure mathematics, we classified all our objects, and can be proud about our classification, but bad for physics and science, because the functions that appear in the real life are, most likely, not rational.

So, let us fix this. Coming as a continuation of our general theorems, we have:

THEOREM 2.28. *A continuous surjective function f is injective, and so invertible, precisely when it is monotone, and in this case, the inverse function f^{-1} must be monotone and continuous too. Moreover, this statement holds both locally, and globally.*

PROOF. The first assertion follows from Theorem 2.22, and the fact that f^{-1} is monotone is clear. Regarding now the continuity of f^{-1}, we want to prove that we have:

$$x_n\to x\implies f^{-1}(x_n)\to f^{-1}(x)$$

But with $x_n=f(y_n)$ and $x=f(y)$, this condition becomes:

$$f(y_n)\to f(y)\implies y_n\to y$$

And this latter condition being true since f is monotone, we are done. □

As an application of this, providing a way out from our impasse, we have:

PROPOSITION 2.29. *The following happen:*

(1) *Given $n \in 2\mathbb{N}+1$, we can extract $\sqrt[n]{x}$, for any $x \in \mathbb{R}$.*
(2) *Given $n \in \mathbb{N}$, we can extract $\sqrt[n]{x}$, for any $x \in [0,\infty)$.*

PROOF. These results come indeed from Theorem 2.28, applied to the power function $f(x) = x^n$, regarded as function $f : \mathbb{R} \to \mathbb{R}$ for n odd, and as function $f : [0,\infty) \to [0,\infty)$ for n even, which is indeed continuous and surjective, as required. □

More generally now, we have the following result, regarding the power functions:

THEOREM 2.30. *The function x^a is defined and continuous on $(0,\infty)$, for any $a \in \mathbb{R}$. Moreover, when trying to extend it to $\mathbb{R}$, we have 4 cases, as follows,*

(1) *For $a \in \mathbb{Q}_{odd}$, $a > 0$, the maximal domain is $\mathbb{R}$.*
(2) *For $a \in \mathbb{Q}_{odd}$, $a \leq 0$, the maximal domain is $\mathbb{R} - \{0\}$.*
(3) *For $a \in \mathbb{R} - \mathbb{Q}$ or $a \in \mathbb{Q}_{even}$, $a > 0$, the maximal domain is $[0,\infty)$.*
(4) *For $a \in \mathbb{R} - \mathbb{Q}$ or $a \in \mathbb{Q}_{even}$, $a \leq 0$, the maximal domain is $(0,\infty)$.*

where $\mathbb{Q}_{odd}$ is the set of rationals $r = p/q$ with q odd, and $\mathbb{Q}_{even} = \mathbb{Q} - \mathbb{Q}_{odd}$.

PROOF. This basically comes from Proposition 2.29, by continuity, as follows:

(1) Assume $a = p/q$, with $p, q \in \mathbb{N}$, $p \neq 0$ and q odd. Given a number $x \in \mathbb{R}$, we can construct the power x^a in the following way, by using Proposition 2.29:

$$x^a = \sqrt[q]{x^p}$$

Then, it is straightforward to prove that x^a is indeed continuous on $\mathbb{R}$.

(2) In the case $a = -p/q$, with $p, q \in \mathbb{N}$ and q odd, the same discussion applies, with the only change coming from the fact that x^a cannot be applied to $x = 0$.

(3) Assume first $a \in \mathbb{Q}_{even}$, $a > 0$. This means $a = p/q$ with $p, q \in \mathbb{N}$, $p \neq 0$ and q even, and as before in (1), we can set $x^a = \sqrt[q]{x^p}$ for $x \geq 0$, by using Proposition 2.29. It is then straightforward to prove that x^a is indeed continuous on $[0,\infty)$, and not extendable either to the negatives. Thus, we are done with the case $a \in \mathbb{Q}_{even}$, $a > 0$, and the case left, namely $a \in \mathbb{R} - \mathbb{Q}$, $a > 0$, follows as well by continuity.

(4) In the cases $a \in \mathbb{Q}_{even}$, $a \leq 0$ and $a \in \mathbb{R} - \mathbb{Q}$, $a \leq 0$, the same discussion applies, with the only change coming from the fact that x^a cannot be applied to $x = 0$. □

Let us record as well a result about the function a^x, as follows:

THEOREM 2.31. *The function a^x is as follows:*

(1) *For $a > 0$, this function is defined and continuous on $\mathbb{R}$.*
(2) *For $a = 0$, this function is defined and continuous on $(0,\infty)$.*
(3) *For $a < 0$, the domain of this function contains no interval.*

PROOF. This is a sort of reformulation of Theorem 2.30, by exchanging the variables, $x \leftrightarrow a$. To be more precise, the situation is as follows:

(1) We know from Theorem 2.30 that things fine with x^a for $x > 0$, no matter what $a \in \mathbb{R}$ is. But this means that things fine with a^x for $a > 0$, no matter what $x \in \mathbb{R}$ is.

(2) This is something trivial, and we have of course $0^x = 0$, for any $x > 0$. As for the powers 0^x with $x \leq 0$, these are impossible to define, for obvious reasons.

(3) Given $a < 0$, we know from Theorem 2.30 that we cannot define a^x for $x \in \mathbb{Q}_{even}$. But since $\mathbb{Q}_{even}$ is dense in $\mathbb{R}$, this gives the result. □

Summarizing, our scare is gone, with Conclusion 2.27 being now obsolete, the point being that, as basic examples of functions, we have the rational functions, their local inverses, and various combinations of these. Which is quite broad, and good to know.

Back now to the rational functions, $f = P/Q$, we know from Theorem 2.25 that it is best to assume that the polynomials P, Q are prime to each other, $(P, Q) = 1$, as for the zeroes of Q to be exactly the poles of f. We can further build on this, as follows:

PROPOSITION 2.32. *Any rational function can be written as*

$$f(x) = \frac{P(x)}{(x - r_1)^{n_1} \ldots (x - r_k)^{n_k} Q(x)}$$

with $r_1, \ldots, r_k \in \mathbb{R}$ being the poles, $P(r_i) \neq 0$, and Q having no roots.

PROOF. This is indeed something self-explanatory, by using first the above-mentioned convention $(P, Q) = 1$, and then factorizing the denominator, using Theorem 2.12. □

In order to exploit the above writing, we will need a standard fact, as follows:

PROPOSITION 2.33. *Given two polynomials which are prime to each other,*

$$(R, Q) = 1$$

we can always find two polynomials A, B such that $AQ + BR = 1$.

PROOF. This is something very standard, exactly as the similar result for numbers, that you surely know. So, consider the following r polynomials, with $r = \deg R$:

$$Q, xQ, x^2Q, \ldots, x^{r-1}Q$$

These polynomials are then, modulo scalars, the various remainders modulo R, so we conclude that, again modulo R and modulo scalars, we have the following equality:

$$\{Q, xQ, x^2Q, \ldots, x^{r-1}Q\} = \{1, x, x^2, \ldots, x^{r-1}\}$$

In particular $AQ = 1(R)$ modulo scalars, for some $A = x^k$, which means $AQ + BR = 1$ modulo scalars, for some A, B, and by dividing by the scalar, $AQ + BR = 1$, as desired. □

Now observe that the formula $AQ+BR=1$ found above can be written as:

$$\frac{1}{RQ}=\frac{A}{R}+\frac{B}{Q}$$

Thus, we can apply this trick in the context of Proposition 2.32, and we obtain:

THEOREM 2.34. *Any rational function can be written as*

$$f(x)=\frac{A_1(x)}{(x-r_1)^{n_1}}+\ldots+\frac{A_k(x)}{(x-r_k)^{n_k}}+\frac{B(x)}{Q(x)}$$

with $r_1,\ldots,r_k\in\mathbb{R}$ being the poles, and with Q having no roots.

PROOF. This follows indeed by writing our function as in Proposition 2.32, and then applying Proposition 2.34, in its fraction form mentioned above, to the various components of the denominator. Thus, we are led to the conclusion in the statement. □

As a continuation of this, we will see soon, in chapter 3, that the last term B/Q actually dissapears when passing to the complex numbers. Thus, we are left with computing the various $A(x)/(x-r)^n$ terms, and here, we can use the following key result:

THEOREM 2.35. *We have the generalized binomial formula*

$$(1+x)^m=\sum_{k=0}^{\infty}\binom{m}{k}x^k$$

with the generalized binomial coefficients being given by

$$\binom{m}{k}=\frac{m(m-1)\ldots(m-k+1)}{k!}$$

valid for any exponent $m\in\mathbb{Z}$, and any $|x|<1$.

PROOF. This is something quite tricky, the idea being as follows:

(1) For exponents $m\in\mathbb{N}$, this is something that we know well from chapter 1, and which is valid for any $x\in\mathbb{R}$, coming from the usual binomial formula.

(2) For the exponent $m=-1$ this is something that we know from chapter 1 too, coming from the following formula, valid for any $|x|<1$:

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

Indeed, this is exactly our generalized binomial formula at $m=-1$, because:

$$\binom{-1}{k}=\frac{(-1)(-2)\ldots(-k)}{k!}=(-1)^k$$

(3) Let us discuss now the general case $m \in -\mathbb{N}$. With $m = -n$, and $n \in \mathbb{N}$, the generalized binomial coefficients are given by the following formula:

$$\begin{aligned}\binom{-n}{k} &= \frac{(-n)(-n-1)\ldots(-n-k+1)}{k!}\\ &= (-1)^k\frac{n(n+1)\ldots(n+k-1)}{k!}\\ &= (-1)^k\frac{(n+k-1)!}{(n-1)!k!}\\ &= (-1)^k\binom{n+k-1}{n-1}\end{aligned}$$

Thus, our generalized binomial formula at $m = -n$, with $n \in \mathbb{N}$, reads:

$$\frac{1}{(1+t)^n} = \sum_{k=0}^{\infty}(-1)^k\binom{n+k-1}{n-1}t^k$$

(4) In order to prove this formula, it is convenient to write it with $-t$ instead of t, in order to get rid of signs. The formula to be proved becomes:

$$\frac{1}{(1-t)^n} = \sum_{k=0}^{\infty}\binom{n+k-1}{n-1}t^k$$

We prove this by recurrence on n. At $n = 1$ this formula definitely holds, as explained in (2) above. So, assume that the formula holds at $n \in \mathbb{N}$. We have then:

$$\begin{aligned}\frac{1}{(1-t)^{n+1}} &= \frac{1}{1-t}\cdot\frac{1}{(1-t)^n}\\ &= \sum_{k=0}^{\infty}t^k\sum_{l=0}^{\infty}\binom{n+l-1}{n-1}t^l\\ &= \sum_{s=0}^{\infty}t^s\sum_{l=0}^{s}\binom{n+l-1}{n-1}\end{aligned}$$

On the other hand, the formula that we want to prove is as follows:

$$\frac{1}{(1-t)^{n+1}} = \sum_{s=0}^{\infty}\binom{n+s}{n}t^k$$

Thus, in order to finish, we must prove the following formula:

$$\sum_{l=0}^{s}\binom{n+l-1}{n-1} = \binom{n+s}{n}$$

(5) In order to prove this latter formula, we proceed by recurrence on $s \in \mathbb{N}$. At $s = 0$ the formula is trivial, $1 = 1$. So, assume that the formula holds at $s \in \mathbb{N}$. In order to prove the formula at $s + 1$, we are in need of the following formula:

$$\binom{n+s}{n} + \binom{n+s}{n-1} = \binom{n+s+1}{n}$$

But this is the Pascal formula, that we know from chapter 1, and we are done. □

In relation now with the rational functions, we have the following result:

THEOREM 2.36. *We have the following formula, for $|x| < r$,*

$$\frac{1}{(r-x)^n} = \frac{1}{r^n} \sum_{k=0}^{\infty} \binom{n+k-1}{n-1} \left(\frac{x}{r}\right)^k$$

which computes the rational functions of type $f(x) = \sum_i A_i(x)/(r_i - x)^{n_i}$.

PROOF. This comes indeed from the formula in Theorem 2.35, or rather from the more digest, equivalent formula found in (4) in its proof, by setting $t = x/r$ there. □

Quite interestingly, the formula in Theorem 2.35 holds in fact at any $m \in \mathbb{R}$, but this is something non-trivial, whose proof will have to wait until Part III below. As for the rational functions, what we have in Theorem 2.34 and Theorem 2.36 certainly provides the key to their study, save for the disappearance of the B/Q factor over the complex numbers, which is something that we will explain soon, in chapter 3. More later.

2d. Complex numbers

Back now to the basics, degree 2 polynomials, what to do when the discriminant is negative? In order to solve $x^2 = -1$, we must trick, in the following way:

DEFINITION 2.37. *The complex numbers are variables of the form*

$$x = a + ib$$

with $a, b \in \mathbb{R}$, which add in the obvious way, and multiply according to the following rule:

$$i^2 = -1$$

Each real number can be regarded as a complex number, $a = a + i \cdot 0$.

In other words, we consider variables as above, without bothering for the moment with their precise meaning. Now consider two such complex numbers:

$$x = a + ib \quad , \quad y = c + id$$

The formula for the sum is then the obvious one, as follows:

$$x + y = (a + c) + i(b + d)$$

As for the formula of the product, by using the rule $i^2 = -1$, we obtain:

$$\begin{aligned} xy &= (a+ib)(c+id) \\ &= ac + iad + ibc + i^2bd \\ &= ac + iad + ibc - bd \\ &= (ac - bd) + i(ad + bc) \end{aligned}$$

The advantage of using the complex numbers comes from the fact that the equation $x^2 = 1$ has now a solution, $x = i$. In fact, this equation has two solutions, namely:

$$x = \pm i$$

This is of course very good news. More generally, we have the following result:

THEOREM 2.38. *The complex solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{R}$ are*

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

with the square root of negative real numbers being defined as

$$\sqrt{-m} = \pm i\sqrt{m}$$

and with the square root of positive real numbers being the usual one.

PROOF. We can write our equation, as usual, in the following way:

$$\begin{aligned} ax^2 + bx + c = 0 &\iff \left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2} \\ &\iff x + \frac{b}{2a} = \pm\frac{\sqrt{b^2 - 4ac}}{2a} \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

We will see later that any degree 2 complex equation has solutions as well, and that more generally, any polynomial equation, real or complex, has solutions. Moving ahead now, we can represent the complex numbers in the plane, in the following way:

PROPOSITION 2.39. *The complex numbers, written as usual*

$$x = a + ib$$

can be represented in the plane, according to the following identification:

$$x = \begin{pmatrix} a \\ b \end{pmatrix}$$

With this convention, the sum of complex numbers is the usual sum of vectors.

PROOF. Consider indeed two arbitrary complex numbers:

$$x = a + ib \quad , \quad y = c + id$$

Their sum is then by definition the following complex number:

$$x + y = (a + c) + i(b + d)$$

Now let us represent x, y in the plane, as in the statement:

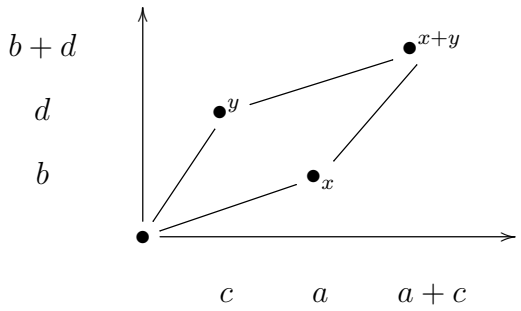
$$x = \begin{pmatrix} a \\ b \end{pmatrix} \quad , \quad y = \begin{pmatrix} c \\ d \end{pmatrix}$$

In this picture, their sum is given by the following formula:

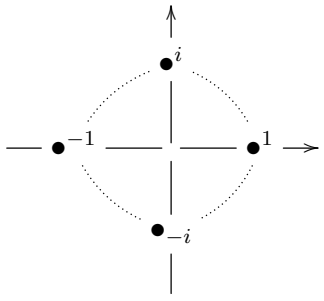
$$x + y = \begin{pmatrix} a + c \\ b + d \end{pmatrix}$$

But this is indeed the vector corresponding to $x + y$, so we are done. □

Here we have assumed that you are a bit familiar with vector calculus. If not, no problem, the idea is simply that vectors add by forming a parallelogram, as follows:



Observe that in our geometric picture from Proposition 2.39, the real numbers correspond to the numbers on the Ox axis. As for the purely imaginary numbers, these lie on the Oy axis. As an illustration for this, we have the following basic picture:



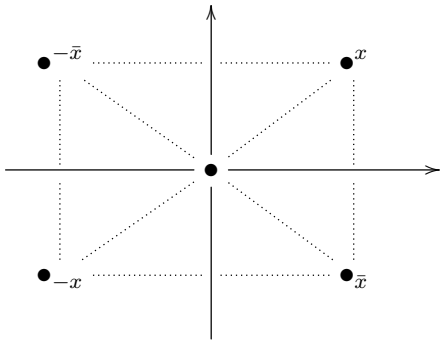
As a last general topic regarding the complex numbers, let us discuss conjugation. This is something quite tricky, complex number specific, as follows:

DEFINITION 2.40. *The complex conjugate of $x = a + ib$ is the following number,*

$$\bar{x} = a - ib$$

obtained by making a reflection with respect to the Ox axis.

As before with other such operations on complex numbers, a quick picture says it all. Here is the picture, with the numbers $x, \bar{x}, -x, -\bar{x}$ being all represented:



Observe that the conjugate of a real number $x \in \mathbb{R}$ is the number itself, $x = \bar{x}$. In fact, the equation $x = \bar{x}$ characterizes the real numbers, among the complex numbers. At the level of non-trivial examples now, we have the following formula:

$$\bar{i} = -i$$

There are many things that can be said about the conjugation of the complex numbers, and here is a summary of basic such things that can be said:

THEOREM 2.41. *The conjugation operation $x \to \bar{x}$ has the following properties:*

(1) *$x = \bar{x}$ precisely when x is real.*
(2) *$x = -\bar{x}$ precisely when x is purely imaginary.*
(3) *$x\bar{x} = |x|^2$, with $|x| = \sqrt{a^2 + b^2}$ for $x = a + ib$.*
(4) *We have the formula $\overline{xy} = \bar{x}\bar{y}$, for any $x, y \in \mathbb{C}$.*
(5) *The solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{R}$ are conjugate.*

PROOF. These results are all elementary, the idea being as follows:

(1) This is something that we already know, coming from definitions.

(2) This is something clear too, because with $x = a + ib$ our equation $x = -\bar{x}$ reads $a + ib = -a + ib$, and so $a = 0$, which amounts in saying that x is purely imaginary.

(3) This is a key formula, which can be proved as follows, with $x = a + ib$:

$$\begin{aligned} x\bar{x} &= (a + ib)(a - ib) \\ &= a^2 + b^2 \\ &= |x|^2 \end{aligned}$$

(4) This is something quite magic, which can be proved as follows:

$$\begin{aligned}\overline{(a+ib)(c+id)} &= \overline{(ac-bd)+i(ad+bc)}\\ &= (ac-bd)-i(ad+bc)\\ &= (a-ib)(c-id)\end{aligned}$$

(5) This comes from the formula of the solutions, that we know from Theorem 2.38, but we can deduce this as well directly, without computations. Indeed, by using our assumption that the coefficients are real, $a,b,c\in\mathbb{R}$, we have:

$$\begin{aligned}ax^2+bx+c=0 &\implies \overline{ax^2+bx+c}=0\\ &\implies \bar{a}\bar{x}^2+\bar{b}\bar{x}+\bar{c}=0\\ &\implies a\bar{x}^2+b\bar{x}+c=0\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Now back to polynomials, as already mentioned before, any degree 2 complex equation has solutions over the complex numbers. In order to discuss this, let us start with:

THEOREM 2.42. *Any complex number $x=a+ib$ has two square roots, given by*

$$\sqrt{x}=\pm\sqrt{\frac{a+\sqrt{a^2+b^2}}{2}}\pm i\sqrt{\frac{-a+\sqrt{a^2+b^2}}{2}}$$

with the signs being identical when $b>0$, and opposite when $b<0$.

PROOF. This is something quite routine, the idea being as follows:

(1) With $x=a+ib$ as in the statement, and $\sqrt{x}=c+id$, our equation is:

$$(c+id)^2=a+ib$$

In terms of the real and imaginary parts, we have two equations, as follows:

$$c^2-d^2=a\quad,\quad 2cd=b$$

(2) Let us first compute the number $u=c^2$. The equation for it is as follows:

$$u-\frac{b^2}{4u}=a$$

Thus, the number $u=c^2$ satisfies the following degree 2 equation:

$$u^2-au-\frac{b^2}{4}=0$$

But this latter equation has a unique positive solution, given by:

$$u=\frac{a+\sqrt{a^2+b^2}}{2}$$

Thus, we are led to the formula of $c=\pm\sqrt{u}$ in the statement.

(3) Similarly, let us compute now $v = d^2$. The equation for it is as follows:

$$\frac{b^2}{4v} - v = a$$

Thus, the number $v = d^2$ satisfies the following degree 2 equation:

$$v^2 + av - \frac{b^2}{4} = 0$$

But this latter equation has a unique positive solution, given by:

$$v = \frac{-a + \sqrt{a^2 + b^2}}{2}$$

Thus, we are led to the formula of $d = \pm\sqrt{v}$ in the statement, and this gives the result, with the last assertion regarding signs being clear, coming from $2cd = b$. □

With this being said, I don't know about you, but personally, for better sleeping at night, I would rather prefer to have this doublechecked. So, given two numbers $a, b \in \mathbb{R}$, consider the following numbers $c, d \in \mathbb{R}$, with the sign on the right being that of b:

$$c = \sqrt{\frac{a + \sqrt{a^2 + b^2}}{2}} \quad , \quad d = \pm\sqrt{\frac{-a + \sqrt{a^2 + b^2}}{2}}$$

We have then $(c + id)^2 = (c^2 - d^2) + 2icd$, whose real part is given by:

$$\begin{aligned} c^2 - d^2 &= \frac{a + \sqrt{a^2 + b^2}}{2} - \frac{-a + \sqrt{a^2 + b^2}}{2} \\ &= \frac{a}{2} + \frac{a}{2} \\ &= a \end{aligned}$$

As for the imaginary part, this can be computed as follows:

$$\begin{aligned} 2cd &= \pm 2\sqrt{\frac{a + \sqrt{a^2 + b^2}}{2} \cdot \frac{-a + \sqrt{a^2 + b^2}}{2}} \\ &= \pm 2\sqrt{\frac{-a^2 + a^2 + b^2}{4}} \\ &= \pm|b| \\ &= b \end{aligned}$$

Thus we have indeed $(c + id)^2 = a + ib$, as desired. Good to know.

Now by getting back to the degree 2 equations, we can formulate a new result regarding them, dealing with the general case, of complex coefficients, as follows:

THEOREM 2.43. *The complex solutions of* $ax^2+bx+c=0$ *with* $a,b,c\in\mathbb{C}$ *are*

$$x_{1,2}=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

with the square root of $b^2-4ac=p+iq$ *being extracted as above, namely*

$$\sqrt{p+iq}=\pm\sqrt{\frac{p+\sqrt{p^2+q^2}}{2}}\pm i\sqrt{\frac{-p+\sqrt{p^2+q^2}}{2}}$$

with the signs being identical when $q>0$*, and opposite when* $q<0$*.*

PROOF. This follows indeed from our old degree 2 computation, from the proof of Theorem 2.38, with the square roots being extracted as in Theorem 2.42. □

As a conclusion to all this, we have learned many interesting things about general functions, and polynomials in particular, and with the degree 2 case fully understood. We will be back to polynomials on numerous occasions, in what follows, and notably in chapter 3 below, with a proof for the key fact that any polynomial equation, of arbitrary degree $N\in\mathbb{N}$, has exactly N complex solutions, counted with multiplicities.

2e. Exercises

Welcome to functions, and to exercises about them. Here are some:

EXERCISE 2.44. *Imagine various mechanical devices, for representing functions.*

EXERCISE 2.45. *What is the best device representing the position of the Sun?*

EXERCISE 2.46. *Rewrite the continuity basics by using the* ε,δ *definition.*

EXERCISE 2.47. *Prove, with bare hands, that the rational functions are continuous.*

EXERCISE 2.48. *Learn about totally discontinuous functions, and other such beasts.*

EXERCISE 2.49. *What happens when representing the complex plane* $\mathbb{C}$ *upside down?*

EXERCISE 2.50. *What about representing* $\mathbb{R}$ *from right to left? Or doing both?*

EXERCISE 2.51. *Practice a bit with extracting complex square roots.*

As bonus exercise, explore a bit more the degree 3 polynomials.

CHAPTER 3

Sin and cos

3a. Angles, triangles

We have seen that we can have some theory going for the polynomials $P \in \mathbb{R}[X]$, and for various related functions, such as the n-th roots $f(x) = \sqrt[n]{x}$, appearing as inverses of the monomials $g(x) = x^n$, or the rational functions $h(x) = P(x)/Q(x)$. Far more things remain to be said here, and more later, but as a more pressing issue, that we would like to discuss now, we have: how to escape from the polynomial function galaxy?

In answer, via geometry, with our plan for what comes next being as follows:

PLAN 3.1. *We can escape from polynomials in two possible ways,*

(1) *Via angles, triangles and plane geometry, leading us into* sin *and* cos*,*
(2) *Via analysis, with* exp *and* log*, but with this being in fact geometry too,*

and we will discuss these two methods in this chapter, and in the next chapter.

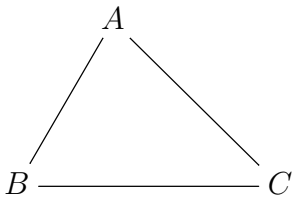
So, this will be our plan for the remainder of the present Part I, talking about angles, triangles and $\sin, \cos$ in this chapter, and then about $\exp, \log$, which belong in fact to plane geometry too, thanks to the Euler formula $e^{it} = \cos t + i \sin t$, in chapter 4.

Getting started now, plane geometry is something very old, with the main results due to the ancient Greeks, found via triangles drawn on sand. We first have:

THEOREM 3.2. *Given a triangle ABC, the following happen:*

(1) *The angle bisectors cross, at a point called incenter.*
(2) *The medians cross, at a point called barycenter.*
(3) *The perpendicular bisectors cross, at a point called circumcenter.*
(4) *The altitudes cross, at a point called orthocenter.*

PROOF. Let us first draw our triangle, with this being always the first thing to be done in geometry, draw a picture, and then thinking and computations afterwards:



(1) Come with a small circle, inside ABC, and then inflate it, as to touch all 3 edges. The center of the circle will be then at equal distance from all 3 edges, so it will lie on all 3 angle bisectors. Thus, we have constructed the incenter, as required.

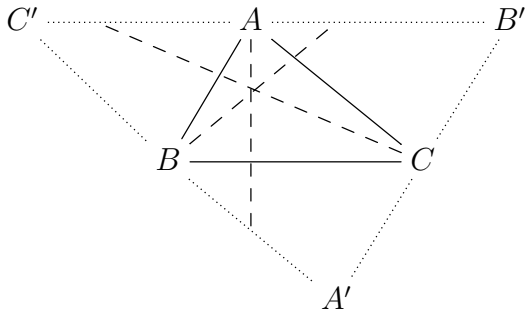
(2) This requires different techniques. Let us call $A, B, C \in \mathbb{C}$ the coordinates of A, B, C, and consider the average $P = (A + B + C)/3$. We have then:

$$P = \frac{1}{3} \cdot A + \frac{2}{3} \cdot \frac{B + C}{2}$$

Thus P lies on the median emanating from A, and a similar argument shows that P lies as well on the medians emanating from B, C. Thus, we have our barycenter.

(3) We can use here the same method as for (1). Indeed, come with a big circle, containing ABC, and then deflate it, as for it to pass through A, B, C. The center of the circle will be then at equal distance from all 3 vertices, so it will lie on all 3 perpendicular bisectors. Thus, we have constructed the circumcenter, as required.

(4) This is tricky. Draw a parallel to BC at A, and similarly, parallels to AB and AC at C and B. We get in this way a bigger triangle, upside-down, $A'B'C'$. But then, the circumcenter of $A'B'C'$, knows to exist from (3), will be the orthocenter of ABC:



Thus, we are led to the conclusions in the statement. □

Getting now to what we wanted to do, angles, we certainly have them, coming in triplets, inside any triangle, and with the name "triangle" witnessing for this. But, how to measure these beasts? In the lack of anything obvious and bright, let us formulate:

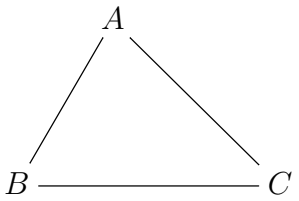
DEFINITION 3.3. *We can talk about the numeric value of angles, as follows:*

(1) *The right angle has value* $90°$.
(2) *We can double angles, in the obvious way.*
(3) *Thus, the half right angle has value* $45°$*, and the flat angle has value* $180°$.
(4) *We can also triple, quadruple and so on, again in the obvious way.*
(5) *Thus, we can talk about arbitrary rational multiples of* $90°$.
(6) *And, with a bit of analysis helping, we can in fact measure any angle.*

So, this will be our starting definition for the numeric values of the angles. Of course, all this might seem a bit improvized, for instance with that 90° figure coming from astronomy, namely 3 times the lunar month, but do not worry, we will come back later to this, with a better, more advanced definition for the numeric values of the angles.

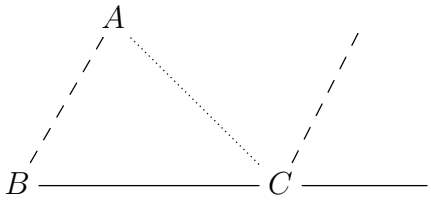
Getting back to work now, theorems and proofs, in relation with the above, here is a key result, which will be our main tool for the study of the angles:

THEOREM 3.4. *In an arbitrary triangle*



the sum of all three angles is 180°.

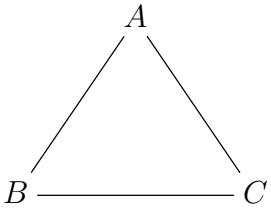
PROOF. This does not seem obvious to prove, with bare hands, but as usual, in such situations, some tricky parallels can come to the rescue. Let us prolong indeed the segment BC a bit, on the C side, and then draw a parallel at C, to the line AB, as follows:



But now, we can see that the three angles around C, summing up to the flat angle 180°, are in fact the 3 angles of our triangle. Thus, theorem proved, just like that. □

As a basic consequence of the above result, making us familiar with 60°, we have:

PROPOSITION 3.5. *In an equilateral triangle, having all sides equal,*

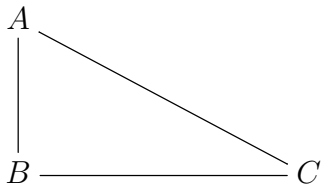


all angles equal 60°.

PROOF. This is clear indeed from the fact that the sum is 180°. □

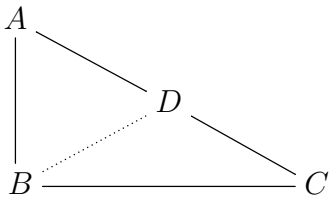
Coming next, we can enlarge our list of familiar angles to $30^\circ, 60^\circ, 90^\circ$, as follows:

PROPOSITION 3.6. *In a right triangle having small angles* $30°, 60°$*,*



we have $AB = AC/2$*.*

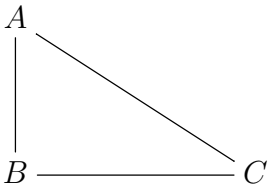
PROOF. This is clear by drawing an equilateral triangle, as follows:



Thus, we are led to the conclusion in the statement. □

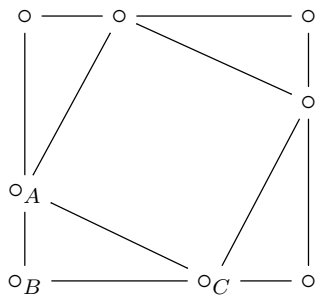
As a key theorem now, dealing with the right triangles, we have:

THEOREM 3.7 (Pythagoras). *In a right triangle* ABC*,*



we have $AB^2 + BC^2 = AC^2$*.*

PROOF. This comes indeed from the following magic configuration, consisting of two squares, and four triangles which are identical to ABC, as indicated:



Indeed, we can compute the area S of the outer square in two ways, as follows:

(1) First, since the side of this square is $AB + BC$, we obtain:

$$\begin{aligned} S &= (AB + BC)^2 \\ &= AB^2 + BC^2 + 2 \times AB \times BC \end{aligned}$$

(2) On the other hand, the outer square is made of the smaller square, having side AC, and of four identical right triangles, having sizes AB, BC. Thus:

$$\begin{aligned} S &= AC^2 + 4 \times \frac{AB \times BC}{2} \\ &= AC^2 + 2 \times AB \times BC \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

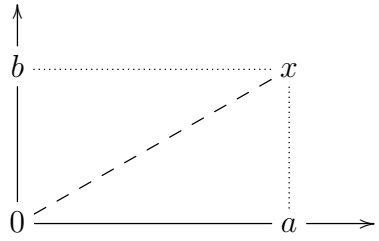
The Pythagoras theorem has many applications. As a first consequence, we have:

THEOREM 3.8. *The distance from a point $x = (a, b) \in \mathbb{R}^2$ to the origin is:*

$$||x|| = \sqrt{a^2 + b^2}$$

Equivalently, the distance from $x = a + ib \in \mathbb{C}$ to the origin is $|x| = \sqrt{a^2 + b^2}$.

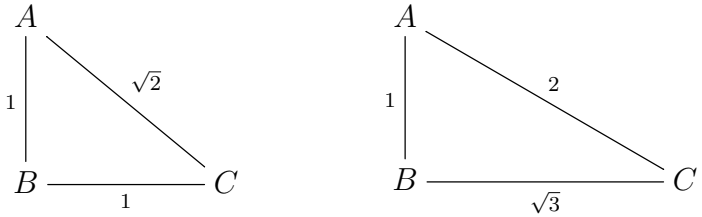
PROOF. This is indeed something self-explanatory, coming from:



Thus, we are led to both the conclusions in the statement. □

As another basic application of the Pythagoras theorem, we have:

PROPOSITION 3.9. *The $45° - 45°$ and $30° - 60°$ right triangles are as follows,*

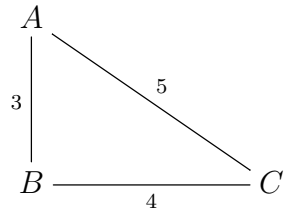


up to a rescaling of the sides.

PROOF. These results come indeed from $1 + 1 = 2$, and from $1 + 3 = 4$. □

As yet another basic application of the Pythagoras theorem, we have:

THEOREM 3.10. *A triangle having sides* $3, 4, 5$ *is a right triangle:*

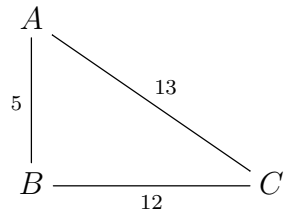


Thus, for drawing right angles, you only need a loop, with 12 *knots on it.*

PROOF. This comes indeed from $9 + 16 = 25$. As for the second assertion, and how can that be used in practice, we will leave this as an engineering exercise. □

Still speaking engineering, having 12 knots equally spaced on a loop is certainly possible, and reliable for most tasks, but if we want to improve our tool, it would be desirable to have more knots on our loop. And here, with a bit of patience, we are led to:

PROPOSITION 3.11. *A triangle having sides* $5, 12, 13$ *is a right triangle:*



Thus, for properly drawing right angles, you need a loop, with 30 *knots on it.*

PROOF. Here the first assertion comes from the following equality, and with the comment that this is the simplest possible one, passed $9 + 16 = 25$:

$$25 + 144 = 169$$

As for the second assertion, we will leave this again as an engineering exercise. □

Along the same lines, at a more advanced level, we have the following result:

THEOREM 3.12. *The Pythagoras equation, namely*

$$a^2 + b^2 = c^2$$

can be fully solved over the integers, the solutions being

$$a = d(m^2 - n^2) \quad , \quad b = 2dmn \quad , \quad c = d(m^2 + n^2)$$

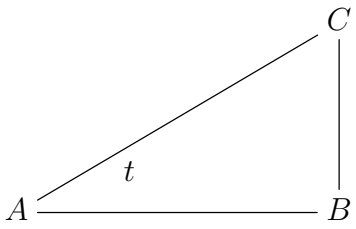
with $(m, n) = 1$*, up to exchanging* a, b*.*

PROOF. This is something quite standard, due to Euclid, that we will actually not really need in what follows. In other words, exercise for you, and enjoy. And as bonus exercise, related to this, try constructing a new right angle device, with 90 knots. □

3b. Sine and cosine

Good news, now that we know about angles and triangles, and about Pythagoras' theorem too, we can start talking about trigonometry. Let us begin with:

DEFINITION 3.13. *Given a right triangle ABC,*

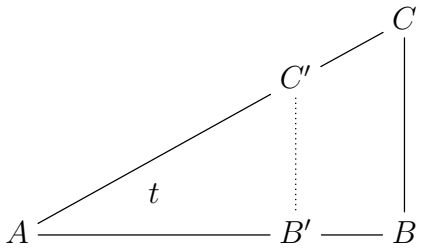


we define the sine and cosine of the angle at A, denoted t, by the following formulae:

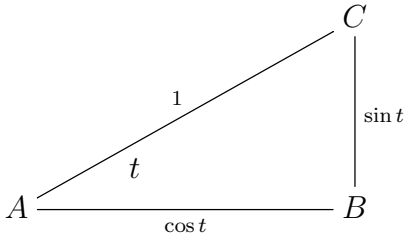
$$\sin t = \frac{BC}{AC} \quad , \quad \cos t = \frac{AB}{AC}$$

We call the sine and cosine basic trigonometric functions.

As a first observation, the sine and cosine do not depend on the choice of the given right triangle ABC having an angle t at A, with this coming according to:



In particular, we can take $AC = 1$ in Definition 3.13, and with this convention, the defining picture for $\sin, \cos$ becomes something very simple, as follows:



Observe also that we have changed the orientation of our right triangles, with our new convention coming from certain geometric considerations, which will appear later.

Question now, what to do with sin and cos? Well, some mathematics I guess, and here are a few basic results regarding them, coming from what we know:

THEOREM 3.14. *The sine and cosine have the following properties:*

(1) $\sin : [0, 90^\circ] \to [0, 1]$ *is bijective, and increasing.*
(2) $\cos : [0, 90^\circ] \to [0, 1]$ *is bijective, and decreasing.*
(3) $\sin(90^\circ - t) = \cos t$.
(4) $\cos(90^\circ - t) = \sin t$.
(5) $\sin(45^\circ + t) = \cos(45^\circ - t)$.
(6) $\cos(45^\circ + t) = \sin(45^\circ - t)$.
(7) *Pythagoras:* $\sin^2 t + \cos^2 t = 1$.

PROOF. Here (1-4) are all clear from definitions, (5-6) follow from (3-4), and finally (7) comes from Pythagoras, applied to our previous $AC = 1$ right triangle. □

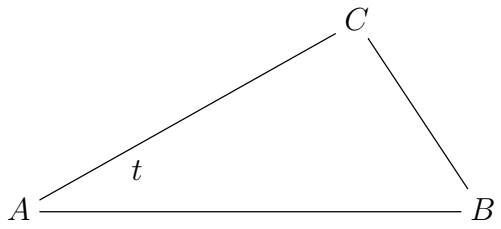
Let us record as well some numerics, coming from things that we know well, about various special right triangles, from the previous section, as follows:

$$\sin 0^\circ = 0 \quad , \quad \sin 30^\circ = \frac{1}{2} \quad , \quad \sin 45^\circ = \frac{1}{\sqrt{2}} \quad , \quad \sin 60^\circ = \frac{\sqrt{3}}{2} \quad , \quad \sin 90^\circ = 1$$

$$\cos 0^\circ = 1 \quad , \quad \cos 30^\circ = \frac{\sqrt{3}}{2} \quad , \quad \cos 45^\circ = \frac{1}{\sqrt{2}} \quad , \quad \cos 60^\circ = \frac{1}{2} \quad , \quad \cos 90^\circ = 0$$

Before getting further with the study of sin and cos, as a question that you might have, what are these good for? In answer, many things, starting with:

THEOREM 3.15 (Law of sines). *The area of an arbitrary triangle, as follows,*

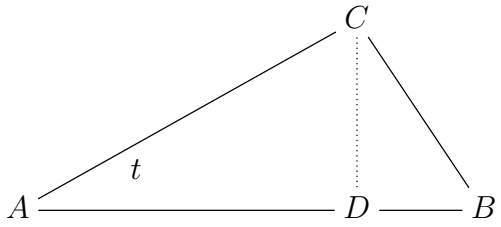


is given by the following formula, making appear the sine,

$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

with the convention $\sin t = \sin(180^\circ - t)$ *for obtuse angles,* $t > 90^\circ$.

PROOF. In order to prove this, we can draw an altitude of our triangle, as follows:



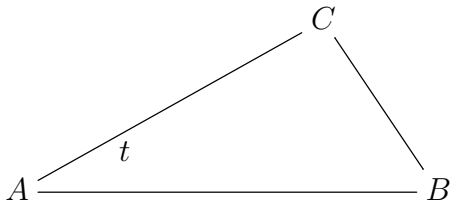
Now with this altitude drawn, we have the following computation:

$$\begin{aligned} area(ABC) &= \frac{basis \times height}{2} \\ &= \frac{AB \times CD}{2} \\ &= \frac{AB \times AC \times \sin t}{2} \end{aligned}$$

Thus, theorem proved, and this working for any $t \in [0°, 180°]$, as stated. □

Regarding now the cosine, things here are a bit more technical, and we have the following result, which is equally useful, generalizing the Pythagoras theorem:

THEOREM 3.16 (Law of cosines). *Given an arbitrary triangle, as follows,*

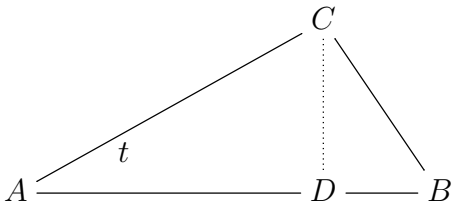


the length of the side which is away from the vertex A is given by the formula

$$BC^2 = AB^2 + AC^2 - 2AB \cdot AC \cdot \cos t$$

with the convention $\cos t = -\cos(180° - t)$ for obtuse angles, $t > 90°$.

PROOF. Let us draw indeed an altitude of our triangle, as follows:



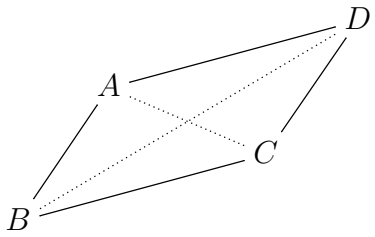
We have then the following computation, coming from Pythagoras, applied twice:

$$\begin{aligned} BC^2 &= CD^2 + BD^2 \\ &= CD^2 + (AB - AD)^2 \\ &= CD^2 + AB^2 + AD^2 - 2AB \cdot AD \\ &= AB^2 + AC^2 - 2AB \cdot AD \\ &= AB^2 + AC^2 - 2AB \cdot AC \cdot \cos t \end{aligned}$$

As for the computation for obtuse triangles, $t > 90°$, this is nearly identical, but with $AB - AD$ replaced by $AB + AD$, leading to $\cos t = -\cos(180° - t)$, as stated. □

Staying with the cosine, as a basic application of Theorem 3.16, we have:

THEOREM 3.17. *Given an arbitrary parallelogram $ABCD$,*



its sides and diagonals are related by the following formula, called parallelogram law,

$$AB^2 + BC^2 + CD^2 + DA^2 = AC^2 + BD^2$$

and this can be used, in the obvious way, in order to compute the triangle medians.

PROOF. There are several things going on here, the idea being as follows:

(1) In the case of a rectangle the parallelogram law is Pythagoras' theorem, and this suggests using the natural generalization of Pythagoras' theorem, which is the law of cosines from Theorem 3.16. Indeed, with O being the middle point of the parallelogram, and with s, t being the angles there of the triangles OAB and OBC, we have:

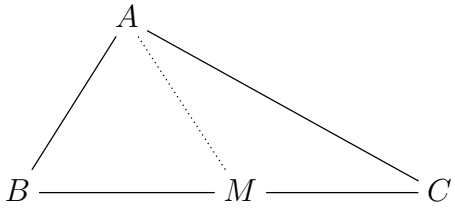
$$AB^2 = OA^2 + OB^2 - 2OA \cdot OB \cdot \cos s$$

$$BC^2 = OB^2 + OC^2 - 2OB \cdot OC \cdot \cos t$$

But $OA = OC$, and $\cos s = -\cos t$, due to $s + t = 180^\circ$, so by summing we get the following formula, which is exactly the parallelogram law, divided by 2:

$$AB^2 + BC^2 = 2OA^2 + 2OB^2$$

(2) Regarding now the medians, consider a triangle ABC, with a median drawn:



By completing to a parallelogram, and using the parallelogram law, we obtain:

$$2AB^2 + 2AC^2 = 4AM^2 + BC^2$$

Thus, we are led to the following formula, for the length of the median:

$$AM = \sqrt{\frac{2AB^2 + 2AC^2 - BC^2}{4}}$$

As for the other two medians of ABC, their formulae are similar. □

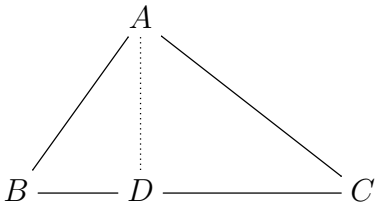
Still in relation with triangles, time now for the sine to strike back, with:

THEOREM 3.18. *Given an arbitrary triangle ABC, we have:*

$$[BC - AC - AB] \sim [\sin A - \sin B - \sin C]$$

That is, the lenghts of the sides are proportional to the sines of the opposite angles.

PROOF. Let us draw indeed an altitude of our triangle, as follows:



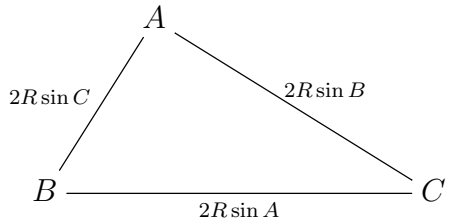
We have then the following computation, for the ratio AB/AC:

$$\frac{AB}{AC} = \frac{AD/\sin B}{AD/\sin C} = \frac{\sin C}{\sin B}$$

As for AB/BC and AC/BC, these are given by similar formulae, again involving quotients of corresponding sines, and this leads to the conclusion in the statement. □

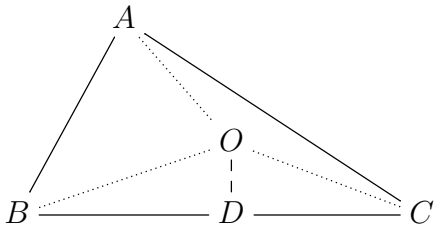
On the same topic, side lengths, we have in fact the following more precise result:

THEOREM 3.19. *The lengths of sides of an arbitrary triangle ABC are given by*



with R being the radius of the circumscribed circle.

PROOF. In order to prove this, let us draw a perpendicular bisector, as follows:



We have then 3 isosceles triangles appearing, say with angles α, β, γ at the point O, satisfying $\alpha + \beta + \gamma = 360°$. The other angles of these isosceles triangles, those coming in

pairs, being $90^\circ-\alpha/2, 90^\circ-\beta/2, 90^\circ-\gamma/2$, we conclude, by looking at what happens at each of the vertices of our triangle ABC, that we have the following formulae:

$$\alpha = 2A \quad , \quad \beta = 2B \quad , \quad \gamma = 2C$$

But with this we can compute the triangle edges. Indeed, we have:

$$BC = 2BD = 2BO \sin\left(\frac{\alpha}{2}\right) = 2R \sin A$$

Similarly, we have $AB = 2R \sin C$ and $AC = 2R \sin B$, as claimed. □

With this discussed, what about the cosine? The competition with the sine was fierce, with sine apparently winning. But the cosine can deliver a fatal blow, as follows:

THEOREM 3.20 (True law of cosines). *The scalar products of vectors,*

$$\left\langle \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}, \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \right\rangle = \sum_i x_i y_i$$

can be computed according to the following formula,

$$< x, y >= ||x|| \cdot ||y|| \cdot \cos t$$

with $t \in [0^\circ, 180^\circ]$ being the angle between the two vectors.

PROOF. This is something a bit off-topic, at this point of this book, just a matter of telling the full story with sin, cos. So, exercise for you, to learn more about this. □

Summarizing, sine dead? Well, never underestimate the sine, and we have:

THEOREM 3.21 (Atomic law of sines). *The vector products in 3 dimensions,*

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} x_2y_3 - x_3y_2 \\ x_3y_1 - x_1y_3 \\ x_1y_2 - x_2y_1 \end{pmatrix}$$

can be computed according to the right-hand rule and the following formula,

$$||x \times y|| = ||x|| \cdot ||y|| \cdot \sin t$$

with $t \in [0^\circ, 180^\circ]$ being the angle between the two vectors.

PROOF. This is something even more specialized than the previous result, although immensely useful in physics, and again exercise for you, to learn more about this. □

Well, as a conclusion to this, you might ask, how wins? In answer, depends on what type of physics you are doing. If you are into very concrete things, in 3D, you will need vector products, and sines. However, in arbitrary N dimensions, including the $N = \infty$ needed for quantum mechanics, the scalar products and the cosines rule.

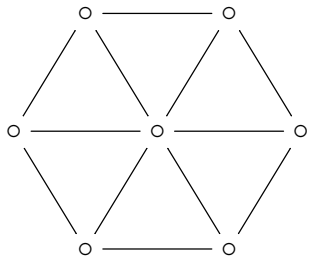
3c. Pi, trigonometry

Let us get now into a more advanced study of the angles. For this purpose, we must first talk about circles, and the number π. And here, we have the following result:

THEOREM 3.22. *The following two definitions of π are equivalent:*

(1) *The length of the unit circle is $L = 2\pi$.*
(2) *The area of the unit disk is $A = \pi$.*

PROOF. In order to prove this theorem let us cut the unit disk as a pizza, into N slices, and forgetting about gastronomy, leave aside the rounded parts:



The area to be eaten can be then computed as follows, where H is the height of the slices, S is the length of their sides, and $P = NS$ is the total length of the sides:

$$A = N \times \frac{HS}{2} = \frac{HP}{2} \simeq \frac{1 \times L}{2}$$

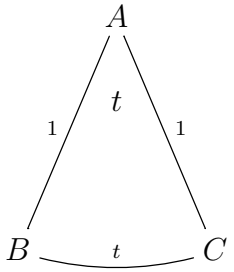
Thus, with $N \to \infty$ we obtain that we have $A = L/2$, as desired. □

In what regards now the precise value of π, the above picture at $N = 6$ shows that we have $\pi > 3$, but not by much. The precise figure is as follows:

$$\pi = 3.14159\ldots$$

Getting now to what we wanted to do, in relation with the angles, we have:

THEOREM 3.23. *We can measure angles by putting them in the middle of a circle of radius 1, and assigning to them the corresponding arc lengths:*



Equivalently, we can use twice the area of the disk slice, which equals the arc length. In this way, the multiples of 90° get converted into corresponding multiples of $\pi/2$.

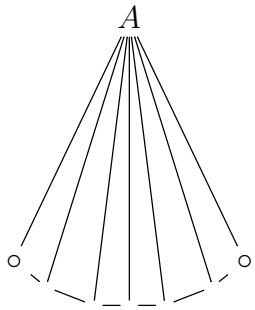
PROOF. We have two things to be proved here, as follows:

(1) First is the fact that our measuring method is indeed good, in the sense that doubling the angles will double their values, tripling the angles will triple their values, and so on. But this is something which is plainly obvious, so done with this.

(2) And then, there is the claim that we have the following formula, with on the left the area of the disk slice ABC, and on the right the arc length BC:

$$2 \times area(ABC) = BC$$

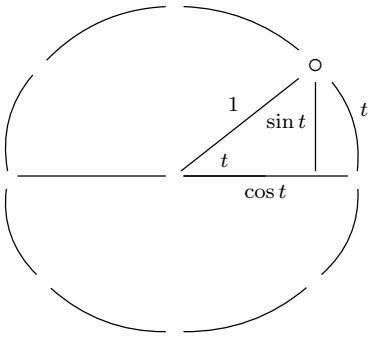
But this is something which is clear for isosceles triangles having altitude 1, and then our disk slice can be approximated by unions of such isosceles triangles, as follows:



Thus, we conclude that our area formula holds indeed, as desired. □

As a first question now that you might have, is doing the above, namely replacing our beloved 90° from astronomy by that crazy $\pi/2$ number, a good thing? In answer, our new convention shines when it comes to trigonometry. We first have:

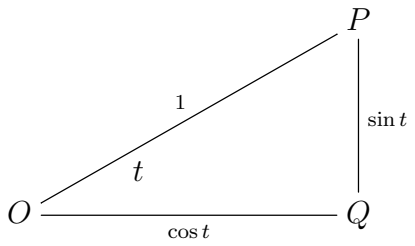
THEOREM 3.24. *The sine and cosine of any $t \in \mathbb{R}$ can be computed according to*



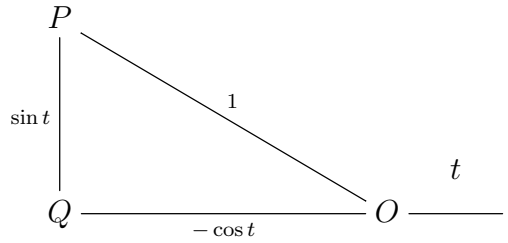
with the convention that inverted segments count as negatives.

PROOF. We have 4 cases to be discussed, which are as as follows:

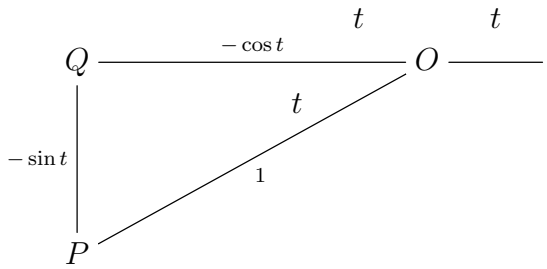
(1) In the simplest case, namely $t \in [0, \pi/2]$, the sine and cosine are indeed computed according to the following picture, which is the one in the statement:



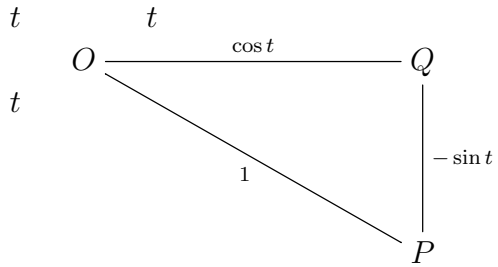
(2) In the case of obtuse angles, $t \in [\pi/2, \pi]$, the picture becomes as follows:



(3) In the next case, namely $t \in [\pi, 3\pi/2]$, the picture becomes as follows:

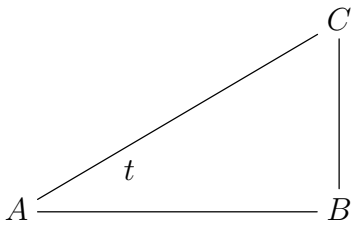


(4) As for the last case, namely $t \in [3\pi/2, 2\pi]$, here our picture is as follows:



Thus, we are led to the conclusions in the statement. □

Before getting into some further study of sin, cos, it is convenient to welcome to our trigonometric family the tangent function. So, consider a basic right triangle:

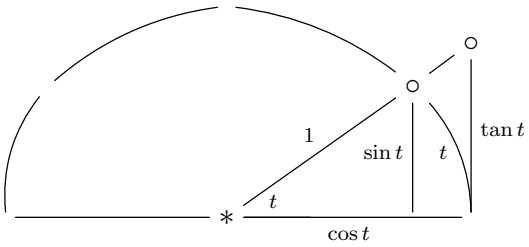


We can define then the tangent of the angle at A by the following formula:

$$\tan t = \frac{BC}{AB} = \frac{\sin t}{\cos t}$$

This tangent function, which is something useful too, and more on this later, naturally fits into the circular picture from Theorem 3.24, in the following way:

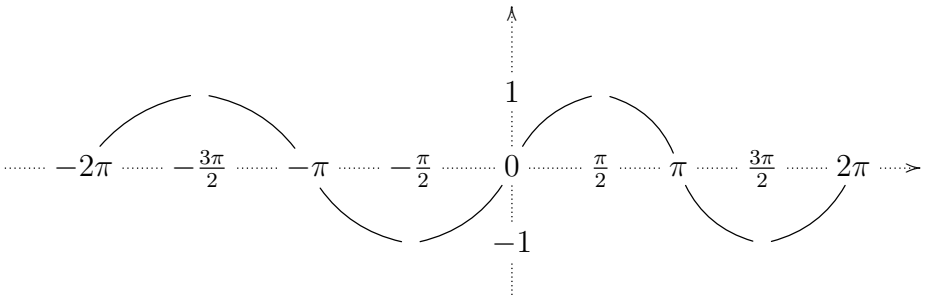
THEOREM 3.25 (addendum). *The tangent can be added as well to the picture,*



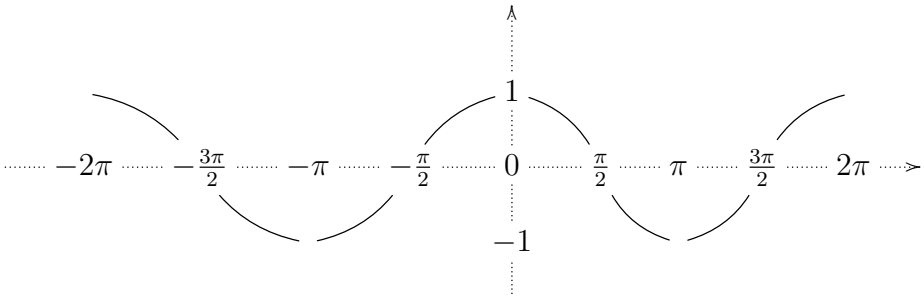
with the convention that this takes the signs $+, -, +, -$, over the four quadrants.

PROOF. This is indeed self-explanatory, and with the comment that there is no simple way of fixing the sign convention, so in a word, better not mess with the tangent. □

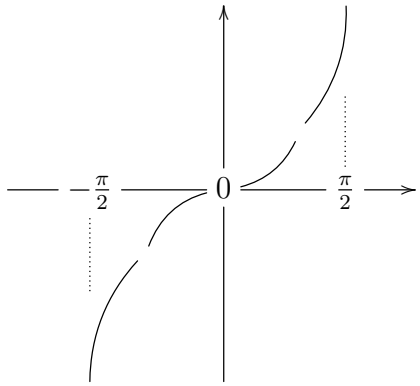
Let us draw as well some graphs. Regarding the sine, the graph is as follows:



For the cosine the graph is similar, translated by $-\pi/2$, as follows:



As for the graph of the tangent, this is as follows, repeated to the left and right:



As a main result now regarding the trigonometric functions, coming from our new convention for the numeric angles, which is something crucial for analysis, we have:

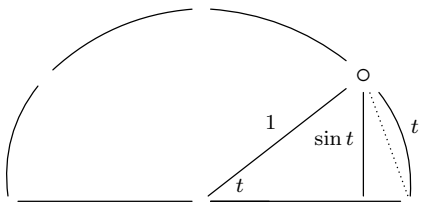
THEOREM 3.26. *We have the following estimates, for small angles*

$$\sin t \leq t \leq \tan t$$

coming from our new convention for numeric angles.

PROOF. Many things can be said here, the idea being as follows:

(1) The general idea is that the estimates are both clear from our circle picture for the angles, and trigonometric functions. Indeed, the picture for the sine is:



Now by using the standard fact that the shortest distance between a point and a line is achieved by constructing the orthogonal projection on that line, we conclude that for any angle $t \in [0, \pi/2]$ we have indeed the following estimate, as claimed:

$$\sin t \leq t$$

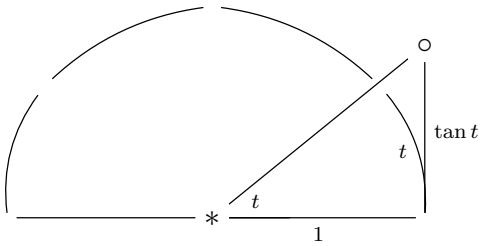
(2) Equivalently, and a bit more rigorously, we can draw the dotted segment above, having length $2\sin(t/2)$, and with Pythagoras on the left, followed by shortest distance between two points being achieved by that dotted segment on the right, we obtain:

$$\sin t \leq 2\sin(t/2) \leq t$$

(3) As yet another proof, we can compare the area of the above isosceles triangle with the area of the disk slice, which gives right away the following estimate, as desired:

$$\frac{\sin t}{2} \leq \frac{t}{2}$$

(4) Regarding now the tangent, again for $t \in [0, \pi/2]$, the picture is as follows:



But here we can argue that the arc t and segment $\tan t$ are related by a projection from $*$, which lands orthogonally on the arc, and obliquely on the segment, and since orthogonal projections notoriously provide the best view, we obtain, as claimed:

$$t \leq \tan t$$

(5) Equivalently, and more rigorously, by comparing areas we get, as desired:

$$\frac{t}{2} \leq \frac{\tan t}{2}$$

(6) Thus, done. Finally, one remaining question concerns the exact range of the above estimates, and we will leave the discussion here as an interesting exercise. □

In fact, by using our circle technology, we are led to the following result:

THEOREM 3.27. *The following happen, for small angles, again coming from our new convention for numeric angles, and best justifying this convention:*

(1) $\sin t \simeq t$.
(2) $\cos t \simeq 1 - t^2/2$.
(3) $\tan t \simeq t$.

PROOF. This can be indeed established as follows:

(1) This is clear indeed on the circle, by arguing like in the previous proof, and we will leave the various details here as an instructive exercise. Equivalently, this follows from $\sin t \leq t \leq \tan t$, by using $\tan t = \sin t / \cos t \simeq \sin t$, coming from $\cos t \simeq 1$.

(2) This comes from (1), and from Pythagoras. Indeed, knowing $\sin t \simeq t$, when looking for a quantity $\cos t$ making the Pythagoras formula $\sin^2 t + \cos^2 t = 1$ hold, we are led, via some quick thinking, to the formula $\cos t \simeq 1 - t^2/2$, according to:

$$t^2 + \left(1 - \frac{t^2}{2}\right)^2 = 1 + \frac{t^4}{4} \simeq 1$$

(3) This is again clear on the circle, or simply follows from (1,2), by dividing. □

Still at the level of the basics, in relation now with continuity issues, we have:

THEOREM 3.28. *The sines and cosines of sums are given by*

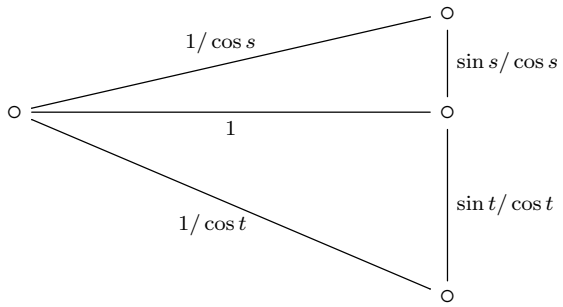
$$\sin(s+t) = \sin s \cos t + \cos s \sin t$$

$$\cos(s+t) = \cos s \cos t - \sin s \sin t$$

and using this, the estimates in Theorem 3.27 can be transported at any $s \in \mathbb{R}$.

PROOF. This is something very standard, the idea being as follows:

(1) Consider the following picture, consisting of a length 1 line segment, with angles s, t drawn on each side, and with the lengths being computed, as indicated:



Now let us compute the area of the big triangle, or rather the double of that area. We can do this in two ways, either directly, with a formula involving $\sin(s+t)$, or by using the two small triangles, involving functions of s, t. We obtain in this way:

$$\frac{1}{\cos s} \cdot \frac{1}{\cos t} \cdot \sin(s+t) = \frac{\sin s}{\cos s} \cdot 1 + \frac{\sin t}{\cos t} \cdot 1$$

But this gives the formula for $\sin(s+t)$ from the statement.

(2) By using $\sin(s+t)$ we can deduce the formula for $\cos(s+t)$, as follows:

$$\begin{aligned}\cos(s+t) &= \sin\left(\frac{\pi}{2}-s-t\right)\\ &= \sin\left[\left(\frac{\pi}{2}-s\right)+(-t)\right]\\ &= \sin\left(\frac{\pi}{2}-s\right)\cos(-t)+\cos\left(\frac{\pi}{2}-s\right)\sin(-t)\\ &= \cos s\cos t-\sin s\sin t\end{aligned}$$

(3) Observe also that by dividing the formulae in (1,2), we obtain as well:

$$\tan(s+t)=\frac{\sin s\cos t+\cos s\sin t}{\cos s\cos t-\sin s\sin t}=\frac{\tan s+\tan t}{1-\tan s\tan t}$$

(4) Finally, in what regards the last assertion, this is something self-explanatory, and we will leave some exploration here as an exercise. We will be back to this in Part II. □

Let us point out now that the formulae in Theorem 3.28 are something quite powerful, which can be useful for many other purposes. For instance, with $s=t$ we obtain:

THEOREM 3.29. *The sines and cosines of the doubles of angles are given by*

$$\sin(2t)=2\sin t\cos t$$

$$\cos(2t)=2\cos^2 t-1$$

and in practice, these formulae can be used as well for passing from t to $t/2$.

PROOF. The formula for sin is clear, and for cos we have 3 useful formulae, namely:

$$\begin{aligned}\cos(2t) &= \cos^2 t-\sin^2 t\\ &= 2\cos^2 t-1\\ &= 1-2\sin^2 t\end{aligned}$$

As for the last assertion, using the above formulae for cos, we have:

$$\cos\left(\frac{t}{2}\right)=\pm\sqrt{\frac{1+\cos t}{2}}\quad,\quad \sin\left(\frac{t}{2}\right)=\pm\sqrt{\frac{1-\cos t}{2}}$$

And we will leave some exploration here, applications of this, as an exercise. For instance, try computing $\sin,\cos,\tan$ for all multiples of $\pi/8$, then $\pi/16$, then $\pi/32$. □

3d. Polar coordinates

Back to the complex numbers, we have since chapter 2 a quite good understanding of their addition. In order to understand now the multiplication operation, we must do something more complicated, namely using polar coordinates. Let us start with:

DEFINITION 3.30. *The complex numbers $x = a+ib$ can be written in polar coordinates,*

$$x = r(\cos t + i \sin t)$$

with the connecting formulae being as follows,

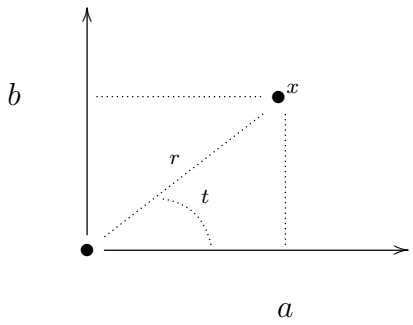
$$a = r \cos t \quad , \quad b = r \sin t$$

and in the other sense being as follows,

$$r = \sqrt{a^2 + b^2} \quad , \quad \tan t = \frac{b}{a}$$

and with r, t being called modulus, and argument.

There is a clear relation here with the vector notation from chapter 2, because r is the length of the vector, and t is the angle made by the vector with the Ox axis. To be more precise, the picture for what is going on in Definition 3.30 is as follows:



The point now is that in polar coordinates, the multiplication formula for the complex numbers, which was so far something quite opaque, takes a very simple form:

THEOREM 3.31. *Two complex numbers written in polar coordinates,*

$$x = r(\cos s + i \sin s) \quad , \quad y = p(\cos t + i \sin t)$$

multiply according to the following formula:

$$xy = rp(\cos(s + t) + i \sin(s + t))$$

In other words, the moduli multiply, and the arguments sum up.

PROOF. We can assume indeed that we have $r = p = 1$, by dividing everything by these numbers. Now with this assumption made, we have the following computation:

$$\begin{aligned} xy &= (\cos s + i \sin s)(\cos t + i \sin t) \\ &= (\cos s \cos t - \sin s \sin t) + i(\cos s \sin t + \sin s \cos t) \\ &= \cos(s + t) + i \sin(s + t) \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

The above result, which is based on some non-trivial trigonometry, is quite powerful. As a basic application of it, we can now compute powers, as follows:

THEOREM 3.32. *The powers of a complex number, written in polar form,*

$$x = r(\cos t + i \sin t)$$

are given by the following formula, valid for any exponent $k \in \mathbb{N}$:

$$x^k = r^k(\cos kt + i \sin kt)$$

Moreover, this formula holds in fact for any $k \in \mathbb{Z}$, and even for any $k \in \mathbb{Q}$.

PROOF. Given a complex number x, written in polar form as above, and an exponent $k \in \mathbb{N}$, we have indeed the following computation, with k terms everywhere:

$$\begin{aligned} x^k &= x \ldots x \\ &= r(\cos t + i \sin t) \ldots r(\cos t + i \sin t) \\ &= r^k([\cos(t + \ldots + t) + i \sin(t + \ldots + t)) \\ &= r^k(\cos kt + i \sin kt) \end{aligned}$$

Thus, we are done with the case $k \in \mathbb{N}$. Regarding now the generalization to the case $k \in \mathbb{Z}$, it is enough here to do the verification for $k = -1$, where the formula is:

$$x^{-1} = r^{-1}(\cos(-t) + i \sin(-t))$$

But this number x^{-1} is indeed the inverse of x, as shown by:

$$\begin{aligned} xx^{-1} &= r(\cos t + i \sin t) \cdot r^{-1}(\cos(-t) + i \sin(-t)) \\ &= \cos(t - t) + i \sin(t - t) \\ &= \cos 0 + i \sin 0 \\ &= 1 \end{aligned}$$

Finally, regarding the generalization to the case $k \in \mathbb{Q}$, it is enough to do the verification for exponents of type $k = 1/n$, with $n \in \mathbb{N}$. The claim here is that:

$$x^{1/n} = r^{1/n} \left[\cos\left(\frac{t}{n}\right) + i \sin\left(\frac{t}{n}\right) \right]$$

In order to prove this, let us compute the n-th power of this number. We can use the power formula for the exponent $n \in \mathbb{N}$, that we already established, and we obtain:

$$\begin{aligned} (x^{1/n})^n &= (r^{1/n})^n \left[\cos\left(n \cdot \frac{t}{n}\right) + i \sin\left(n \cdot \frac{t}{n}\right) \right] \\ &= r(\cos t + i \sin t) \\ &= x \end{aligned}$$

Thus, we have indeed a n-th root of x, and our proof is now complete. □

We kept the best for the end. As a last topic regarding the complex numbers, which is something really beautiful, we have the roots of unity, which are as follows:

THEOREM 3.33. *The equation $x^N = 1$ has N complex solutions, namely*

$$\left\{w^k\Big|k = 0, 1, \ldots, N-1\right\} \quad , \quad w = \cos\left(\frac{2\pi}{N}\right) + i\sin\left(\frac{2\pi}{N}\right)$$

which are called roots of unity of order N.

PROOF. This follows from the general multiplication formula for complex numbers from Theorem 3.31. Indeed, with $x = r(\cos t + i\sin t)$ our equation $x^N = 1$ reads:

$$r^N(\cos Nt + i\sin Nt) = 1$$

Thus $r = 1$, and $t \in [0, 2\pi)$ must be a multiple of $2\pi/N$, as stated. □

The roots of unity are very useful variables, and have many interesting properties. As a first application, we can now solve the ambiguity questions related to the extraction of N-th roots, coming from Theorem 3.32, the statement here being as follows:

THEOREM 3.34. *Any $x = r(\cos t + i\sin t)$ has N roots of order N, namely*

$$y = r^{1/N}\left(\cos\left(\frac{t}{N}\right) + i\sin\left(\frac{t}{N}\right)\right)$$

multiplied by the N roots of unity of order N.

PROOF. We must solve the equation $z^N = x$, over the complex numbers. Since the number y in the statement clearly satisfies $y^N = x$, our equation is equivalent to:

$$z^N = y^N$$

We conclude from this that the solutions z appear by multiplying y by the solutions of $t^N = 1$, which are the N-th roots of unity, as claimed. □

Finally, as a main application of the complex numbers, and therefore of trigonometry too, as previously announced in chapter 2, we have the following key result:

THEOREM 3.35. *Any polynomial $P \in \mathbb{C}[X]$ decomposes as*

$$P = c(X - r_1)\ldots(X - r_N)$$

with $c \in \mathbb{C}$ and with $r_1, \ldots, r_N \in \mathbb{C}$.

PROOF. This is something quite tricky, the idea being as follows:

(1) To start with, the problem is that of proving that our polynomial has at least one root, because afterwards we can proceed by recurrence, in the obvious way.

(2) We prove this fact, that P has at least one root, by contradiction. So, assume that P has no roots, and pick a number $z \in \mathbb{C}$ where $|P|$ attains its minimum:

$$|P(z)| = \min_{x \in \mathbb{C}} |P(x)| > 0$$

(3) Since $Q(t) = P(z + t) - P(z)$ is a polynomial which vanishes at $t = 0$, this polynomial must be of the form ct^k + higher terms, with $c \neq 0$, and with $k \geq 1$ being an integer. We obtain from this that, with $t \in \mathbb{C}$ small, we have the following estimate:

$$P(z + t) \simeq P(z) + ct^k$$

Now let us write $t = rw$, with $r > 0$ small, and with $|w| = 1$. Our estimate becomes:

$$P(z + rw) \simeq P(z) + cr^k w^k$$

(4) But, recall that we assumed $P(z) \neq 0$. We can therefore choose w with $|w| = 1$ such that cw^k points in the opposite direction to that of $P(z)$, and we obtain:

$$\begin{aligned} |P(z + rw)| &\simeq |P(z) + cr^k w^k| \\ &= |P(z)|(1 - |c|r^k) \end{aligned}$$

(5) Now by choosing $r > 0$ small enough, as for the error in the first estimate to be small, and overcame by the negative quantity $-|c|r^k$, we obtain from this:

$$|P(z + rw)| < |P(z)|$$

(6) But this contradicts our definition of $z \in \mathbb{C}$, as a point where $|P|$ attains its minimum. Thus P has a root, and by recurrence it has N roots, as stated. □

3e. Exercises

All good old geometry in this chapter, and as exercises on this, we have:

Exercise 3.36. *What is the barycenter of a triangle, with edges equally weighted?*

Exercise 3.37. *Learn about the Ceva theorem, and further centers of a triangle.*

Exercise 3.38. *Learn also about the Euler line, and the nine-point circle.*

Exercise 3.39. *Further meditate on the practical need for the sine and cosine.*

Exercise 3.40. *Work out trigonometric formulae for the triples of angles.*

Exercise 3.41. *Learn about the Chebycheff polynomials, of first and second kind.*

Exercise 3.42. *Improve our basic estimates for the basic trigonometric functions.*

Exercise 3.43. *Learn more about the roots of unity, and their various properties.*

As bonus exercise, use what we learned here, in order to find estimates for π.

CHAPTER 4

Exp and log

4a. The number e

Time now to get into some truly advanced things, namely exp and log. These are quite basic functions in mathematics and science, but in order to introduce them, we will have to work a bit. The idea will be that the exponential will be $\exp x = e^x$, and the logarithm $\log x$ will be its inverse, but the whole point lies in understanding what is the best number $e \in \mathbb{R}$ that we can use, for good results, and this is something non-trivial.

So, be patient, things here will be non-trivial, and we will have to establish some technical results first, with no obvious goal. But, as analysts, we are supposed to enjoy everything analysis, so take what comes next like this, analysts enjoying analysis.

Let us start with the following remarkable result, which is something having is own interest, namely computing a quite natural 1^∞ type limit, which is non-trivial:

THEOREM 4.1. *We have the following convergence*

$$\left(1+\frac{1}{n}\right)^n \to e$$

where $e = 2.71828\ldots$ *is a certain number.*

PROOF. This is something quite tricky, as follows:

(1) Our first claim is that the following sequence is increasing:

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

In order to prove this claim, we can use the arithmetic-geometric inequality, applied to the number 1, and to n copies of the number $1+1/n$. Indeed, this gives:

$$\frac{1+\sum_{i=1}^n\left(1+\frac{1}{n}\right)}{n+1} \geq \sqrt[n+1]{1\cdot\prod_{i=1}^n\left(1+\frac{1}{n}\right)}$$

In practice, by rearranging a bit, this gives the following inequality:

$$1+\frac{1}{n+1} \geq \left(1+\frac{1}{n}\right)^{n/(n+1)}$$

Now by raising to the power $n+1$ we obtain, as desired:

$$\left(1+\frac{1}{n+1}\right)^{n+1}\geq\left(1+\frac{1}{n}\right)^{n}$$

(2) Normally we are left with proving that x_n is bounded from above, but this is non-trivial, and we have to use a trick. Consider the following sequence:

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

We will prove that this sequence y_n is decreasing, and together with the fact that we have $x_n/y_n\to 1$, this will give the result. So, this will be our plan.

(3) In order to prove now that y_n is decreasing, we use, a bit as before:

$$\frac{1+\sum_{i=1}^{n}\left(1-\frac{1}{n}\right)}{n+1}\geq\sqrt[n+1]{1\cdot\prod_{i=1}^{n}\left(1-\frac{1}{n}\right)}$$

In practice, this gives the following inequality:

$$1-\frac{1}{n+1}\geq\left(1-\frac{1}{n}\right)^{n/(n+1)}$$

Now by raising to the power $n+1$ we obtain from this:

$$\left(1-\frac{1}{n+1}\right)^{n+1}\geq\left(1-\frac{1}{n}\right)^{n}$$

The point now is that we have the following inversion formulae:

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

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

Thus by inverting the inequality that we found, we obtain, as desired:

$$\left(1+\frac{1}{n}\right)^{n+1}\leq\left(1+\frac{1}{n-1}\right)^{n}$$

(4) But with this, we can now finish. Indeed, the sequence x_n is increasing, the sequence y_n is decreasing, and we have $x_n<y_n$, as well as:

$$\frac{y_n}{x_n}=1+\frac{1}{n}\to 1$$

Thus, both sequences x_n,y_n converge to a certain number e, as desired.

(5) Finally, regarding the numerics for our limiting number e, we know from the above that we have $x_n < e < y_n$ for any $n \in \mathbb{N}$, which reads:

$$\left(1+\frac{1}{n}\right)^n < e < \left(1+\frac{1}{n}\right)^{n+1}$$

Thus $e \in [2,3]$, and with a bit of patience, or a computer, we obtain $e = 2.71828\ldots$ We will actually come back to this question in a moment, with better methods. □

More generally now, we have the following result, featuring a variable $x \in \mathbb{R}$:

THEOREM 4.2. *We have the following formula,*

$$\left(1+\frac{x}{n}\right)^n \to e^x$$

valid for any $x \in \mathbb{R}$.

PROOF. We know from Theorem 4.1 that the result holds at $x = 1$, and this because the number e was by definition given by the following formula:

$$\left(1+\frac{1}{n}\right)^n \to e$$

Observe that by taking inverses, we have as well the result at $x = -1$, namely:

$$\left(1-\frac{1}{n}\right)^n \to \frac{1}{e}$$

In general now, when $\in \mathbb{R}$ is arbitrary, the best is to proceed as follows:

$$\left(1+\frac{x}{n}\right)^n = \left[\left(1+\frac{x}{n}\right)^{n/x}\right]^x \to e^x$$

Thus, we are led to the conclusion in the statement. □

Summarizing, we know what the number e, and what the associated exponential function e^x are. However, the story is not over here. As a complement to Theorem 4.1, we have indeed the following result, which is something quite far-reaching:

THEOREM 4.3. *We have the following formula,*

$$e = \sum_{k=0}^{\infty} \frac{1}{k!}$$

which can stand as an alternative definition for e.

PROOF. This can be done in several steps, as follows:

(1) In practice, we want to prove that we have the following equality:

$$\lim_{n\to\infty} \left(1+\frac{1}{n}\right)^n = \sum_{k=0}^{\infty} \frac{1}{k!}$$

(2) For this purpose, the first observation is that we have the following estimate:

$$2 < \sum_{k=0}^{\infty} \frac{1}{k!} < \sum_{k=0}^{\infty} \frac{1}{2^{k-1}} = 3$$

Thus, the series $\sum_{k=0}^{\infty} \frac{1}{k!}$ converges indeed, towards a limit in $(2, 3)$.

(3) In order to prove now that this limit is e, observe that we have:

$$\begin{aligned}
\left(1 + \frac{1}{n}\right)^n &= \sum_{k=0}^{n} \binom{n}{k} \cdot \frac{1}{n^k} \\
&= \sum_{k=0}^{n} \frac{n(n-1)\ldots(n-k+1)}{k!} \cdot \frac{1}{n^k} \\
&\leq \sum_{k=0}^{n} \frac{1}{k!}
\end{aligned}$$

Thus, with $n \to \infty$, we get that the limit of the series $\sum_{k=0}^{\infty} \frac{1}{k!}$ belongs to $[e, 3)$.

(4) For the reverse inequality, we use the following computation:

$$\begin{aligned}
\sum_{k=0}^{n} \frac{1}{k!} - \left(1 + \frac{1}{n}\right)^n &= \sum_{k=0}^{n} \frac{1}{k!} - \sum_{k=0}^{n} \frac{n(n-1)\ldots(n-k+1)}{k!} \cdot \frac{1}{n^k} \\
&= \sum_{k=2}^{n} \frac{1}{k!} - \sum_{k=2}^{n} \frac{n(n-1)\ldots(n-k+1)}{k!} \cdot \frac{1}{n^k} \\
&= \sum_{k=2}^{n} \frac{n^k - n(n-1)\ldots(n-k+1)}{n^k k!} \\
&\leq \sum_{k=2}^{n} \frac{n^k - (n-k)^k}{n^k k!} \\
&= \sum_{k=2}^{n} \frac{1 - \left(1 - \frac{k}{n}\right)^k}{k!}
\end{aligned}$$

(5) In order to estimate the above expression that we found, we can use the following trivial inequality, valid for any number $x \in (0, 1)$:

$$1 - x^k = (1 - x)(1 + x + x^2 + \ldots + x^{k-1}) \leq (1 - x)k$$

Indeed, we can use this with $x = 1 - k/n$, and we obtain in this way:

$$\begin{aligned}
\sum_{k=0}^{n} \frac{1}{k!} - \left(1 + \frac{1}{n}\right)^n &\leq \sum_{k=2}^{n} \frac{\frac{k}{n} \cdot k}{k!} \\
&= \frac{1}{n} \sum_{k=2}^{n} \frac{k}{(k-1)!} \\
&= \frac{1}{n} \sum_{k=2}^{n} \frac{k}{k-1} \cdot \frac{1}{(k-2)!} \\
&\leq \frac{1}{n} \sum_{k=2}^{n} \frac{2}{2^{k-2}} \\
&< \frac{4}{n}
\end{aligned}$$

Now since with $n \to \infty$ this quantity goes to 0, we obtain that the limit of the series $\sum_{k=0}^{\infty} \frac{1}{k!}$ is the same as the limit of the sequence $\left(1 + \frac{1}{n}\right)^n$, namely e, as desired. □

As a fourth and final result in our series, generalizing Theorem 4.3, we have:

THEOREM 4.4. *We have the following formula,*

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

valid for any $x \in \mathbb{R}$.

PROOF. To start with, the above series converges at any $x \in \mathbb{R}$, with this being best seen as an application of our general convergence criteria from chapter 1, coming from:

$$\frac{x^{k+1}}{(k+1)!} : \frac{x^k}{k!} = \frac{x}{k+1} \to 0$$

Alternatively, this comes directly from $x^k/k! < 2^k$, for $k >> 0$. Regarding now the fact that the sum is indeed e^x, we have two possible proofs here, as follows:

(1) As a first idea, which is straightforward, we can adapt the proof of Theorem 4.3, by inserting a variable x there. Indeed, in view of Theorem 4.2, we want to prove that:

$$\lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

As a first observation, exactly as before in the case $x = 1$, we have:

$$\left(1 + \frac{x}{n}\right)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot \frac{x^k}{n^k} \leq \sum_{k=0}^{n} \frac{x^k}{k!}$$

As for the reverse inequality, again by following the proof at $x = 1$, we have:

$$\begin{aligned}\sum_{k=0}^{n}\frac{x^k}{k!}-\left(1+\frac{x}{n}\right)^n &\leq \sum_{k=2}^{n}\frac{1-\left(1-\frac{k}{n}\right)^k}{k!}\cdot x^k\\ &\leq \frac{1}{n}\sum_{k=2}^{n}\frac{2}{2^{k-2}}\cdot x^k\end{aligned}$$

Thus, done with $|x| < 2$, because with $n \to \infty$ we get 0 on the right. As for the general case, $x \in \mathbb{R}$, here we must fine-tune a bit our estimates, say exercise for you.

(2) We have as well a second proof, which is a bit more conceptual, and that we will explain now in detail. Consider the function in the statement, namely:

$$f(x)=\sum_{k=0}^{\infty}\frac{x^k}{k!}$$

By using the binomial formula, we have the following computation:

$$\begin{aligned}f(x+y) &= \sum_{k=0}^{\infty}\frac{(x+y)^k}{k!}\\ &= \sum_{k=0}^{\infty}\sum_{s=0}^{k}\binom{k}{s}\cdot\frac{x^sy^{k-s}}{k!}\\ &= \sum_{k=0}^{\infty}\sum_{s=0}^{k}\frac{x^sy^{k-s}}{s!(k-s)!}\\ &= f(x)f(y)\end{aligned}$$

As a first observation, this shows that f is continuous. Indeed, at $x = 0$ we have:

$$\lim_{t\to 0}f(t)=\lim_{t\to 0}\left(1+t\sum_{k=1}^{\infty}\frac{t^{k-1}}{k!}\right)=1$$

But from this, we get $f(x + t) = f(x)f(t) \to f(x)$ with $t \to 0$, at any x. Thus, as a conclusion, our function f is continuous, and satisfies the following conditions:

$$f(x+y)=f(x)f(y)\quad,\quad f(1)=e$$

But with this, we can finish. Indeed, by iterating, we have $f(nx) = f(x)^n$ for any $n \in \mathbb{N}$. Then, by extracting roots, we have $f(rx) = f(x)^r$ for any $r \in \mathbb{Q}$. Thus $f(r) = e^r$ for any $r \in \mathbb{Q}$, and by continuity we obtain $f(x) = e^x$ for any $x \in \mathbb{R}$, as desired. □

Quite interesting all the above, so let us summarize our findings, as follows:

CONCLUSION 4.5. *We have a number* $e = 2.71828\ldots$ *given by*

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = \sum_{k=0}^{\infty}\frac{1}{k!}$$

with the corresponding power function being given by the following formulae,

$$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = \sum_{k=0}^{\infty}\frac{x^k}{k!}$$

and with all equivalences coming from the basic theory of sequences and series.

Which sounds quite complete, but as a matter of making sure that we have not forgotten anything, time to ask the cats. These days I have two of them constantly roaming around the house, and with Vladimir, who's good at analysis, being currently gone for a hunt, I will have to ask Nicolas, who's more into algebra. And Nicolas answers:

NICOLAS 4.6. *The exponential is the only function mapping* $0 \to 1$*, and equalling its own derivative. And* $e = \exp(1)$*.*

Well, beware I guess of algebraic felines, who can destroy your analytic work in a matter of seconds. This being said, Nicolas makes reference to the notion of derivative, that is scheduled only for Part III in this book. In the meantime, let us record however a number of useful facts, coming as a soft version of what Nicolas is saying:

FACT 4.7. *The number* $e = 2.71828\ldots$ *is the unique one having the property*

$$e^t \simeq 1 + t$$

for $t \simeq 0$*. Geometrically, this means that the slope of* e^x *at* $x = 0$ *must be* 1*.*

To be more precise here, $e^t \simeq 1+t$ comes by truncating $e^t = \sum_k t^k/k!$, and the converse is easily seen to hold too, and more on this in a moment, when talking logarithms. As for the last assertion, this is precisely what $e^t \simeq 1 + t$ tells us, and with the remark here that, more generally, $e^{x+t} \simeq e^x(1 + t)$ tells us that the slope of e^x at any $x \in \mathbb{R}$ must be e^x itself, in agreement with Nicolas 4.6. And more about this later, in Part III.

Observe the similarity with the formula $\sin t \simeq t$ from chapter 3. In fact, we have:

PRINCIPLE 4.8. *Hard mathematics gets axiomatized by functions around* 0*:*

(1) $\sin t \simeq t$ *tells us what the angles* t *are.*
(2) $e^t \simeq 1 + t$ *tells us what the number* e *is.*

Moving on, many other things can be said about e, and we will be back to this on a regular basis, in what follows. As a last basic result about e, let us record:

THEOREM 4.9. *The number e is numerically given by*

$$e = 2.7182818284\ldots$$

and is irrational, $e \notin \mathbb{Q}$.

PROOF. The above assertions are related, both about approximating e, as follows:

(1) Regarding the numerics, the series defining e converges very fast, when compared to the limit, so if you are in a hurry, that series is for you. We have:

$$\begin{aligned}
e &= \sum_{k=0}^{N-1}\frac{1}{k!}+\frac{1}{N!}\left(1+\frac{1}{N+1}+\frac{1}{(N+1)(N+2)}+\ldots\right)\\
&< \sum_{k=0}^{N-1}\frac{1}{k!}+\frac{1}{N!}\left(1+\frac{1}{N+1}+\frac{1}{(N+1)^2}+\ldots\right)\\
&= \sum_{k=0}^{N-1}\frac{1}{k!}+\frac{1}{N!}\left(1+\frac{1}{N}\right)\\
&= \sum_{k=0}^{N}\frac{1}{k!}+\frac{1}{N\cdot N!}
\end{aligned}$$

Thus, the error term in the approximation is really tiny, the estimate being:

$$\sum_{k=0}^{N}\frac{1}{k!} < e < \sum_{k=0}^{N}\frac{1}{k!}+\frac{1}{N\cdot N!}$$

(2) Now by using this, you can easily compute the decimals of e. Actually, you can't call yourself mathematician, or scientist, if you haven't done this by hand, just for the fun, but just in case, here is how the approximation goes, for small values of N:

$$N=2 \implies 2.5 < e < 2.75$$

$$N=3 \implies 2.666\ldots < e < 2.722\ldots$$

$$N=4 \implies 2.70833\ldots < e < 2.71875\ldots$$

$$N=5 \implies 2.71666\ldots < e < 2.71833\ldots$$

$$N=6 \implies 2.71805\ldots < e < 2.71828\ldots$$

$$N=7 \implies 2.71825\ldots < e < 2.71828\ldots$$

Thus, first 4 decimals computed, $e = 2.7182\ldots$, and I would leave the continuation to you. With the remark that, when carefully looking at the above, the estimate on the right works much better than the one on the left, so before getting into more serious numerics, try to find a better lower estimate for e, that can help you in your work.

(3) Getting now to irrationality, a look at $e = 2.7182818284\ldots$ might suggest that the $81, 82, 84\ldots$ values might eventually, after some internal fight, decide for a winner, and so that e might be rational. However, this is wrong, and e is in fact irrational.

(4) So, let us prove now this, that e is irrational. Following Fourier, we will do this by contradiction. So, assume $e = m/n$, and let us look at the following number:

$$x = n!\left(e - \sum_{k=0}^{n} \frac{1}{k!}\right)$$

As a first observation, x is an integer, as shown by the following computation:

$$\begin{aligned} x &= n!\left(\frac{m}{n} - \sum_{k=0}^{n} \frac{1}{k!}\right) \\ &= m(n-1)! - \sum_{k=0}^{n} n(n-1)\ldots(n-k+1) \\ &\in \mathbb{Z} \end{aligned}$$

(5) On the other hand $x > 0$, and we have as well the following estimate:

$$\begin{aligned} x &= n! \sum_{k=n+1}^{\infty} \frac{1}{k!} \\ &= \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \ldots \\ &< \frac{1}{n+1} + \frac{1}{(n+1)^2} + \ldots \\ &= \frac{1}{n} \end{aligned}$$

Thus $x \in (0, 1)$, which contradicts our previous finding $x \in \mathbb{Z}$, as desired. □

Finally, let us mention that there is even worse, with e being transcendental, that is, not being root of any polynomial $P \in \mathbb{Q}[X]$. By the way, the same can be said about π, which is irrational and trascendental too. More on this, hopefully, later in this book.

4b. Euler formula

Now that we know about the number $e = 2.7182\ldots$, we would like to use the well-known formula $x = re^{it}$ for the complex numbers, as everyone does. However, proving this formula is no easy task, normally requiring the use of derivatives, that we will learn only in Part III of this book, and we will be punching here a bit above our weight.

Guess I will have to ask again the cats. And with Nicolas being now gone to a seminar (didn't even know that cats have seminars), and Vladimir being back from his hunt, exactly when the other left, and no surprise here, these two fellows don't quite like each other, I will discuss with Vladimir all this. And here is what he says:

VLADIMIR 4.10. *Mathematics is the part of physics where experiments are cheap.*

Humm, this sounds to my inner philosopher a bit extreme, but after all, isn't this damn true, and efficient. After all, mathematics is just a collection of right and wrong statements, as by the way any other science is, and our job as mathematicians is that of distinguishing between what is right and what is wrong. With all methods allowed.

So, getting now to $x = re^{it}$, as a cheap experiment here, that I just performed, and that you can reproduce too, anytime, we can look up what scientists say about this, on the internet. And with 100% of people agreeing on this, we have our theorem. And with the remark that our social science proof for it, based on 100% agreement, looks far more reliable than our usual formal math proofs, which can, after all, contain mistakes.

This being said, I am pretty much sure that you would like to know more about $x = re^{it}$, before starting using it, like everyone does. So, we first have:

THEOREM 4.11. *We can exponentiate the complex numbers, according to*

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

and the function $x \to e^x$ is continuous, and satisfies $e^{x+y} = e^x e^y$.

PROOF. There are several things going on here, the idea being as follows:

(1) To start with, we are now dealing with complex functions, so we must first argue that our basic knowledge of analysis, namely convergence, sequences, series and continuity, extends to this setting. But this is something which is pretty much obvious, because everything in real analysis comes from the distance function on the reals, namely:

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

To be more precise, it is this distance function which allows us to tell what is small, what is big, and so on, and so to talk about convergence, and anything else analysis.

(2) Now the point is that in the complex setting we have a distance function too, nemely $d(x, y) = |x - y|$, with $|.|$ being this time the modulus of complex numbers. Moreover, as we know well from Pythagoras, this is something very intuitive, namely the usual distance in the plane. And by using this, we can surely talk about convergence, and all our basic results regarding sequences, series and continuity extend to this setting.

(3) We will be talking more in detail about this in chapter 5, and in the meantime, just trust me. Or don't trust me, and have a quick look at what we did in chapter 1, and the first part of chapter 2, in order to prove me wrong, that would be even better.

(4) With this discussed, and getting now to what the statement says, we must first prove that the series converges. But this comes from the following computation:

$$\begin{aligned}
|e^x| &= \left|\sum_{k=0}^{\infty}\frac{x^k}{k!}\right| \\
&\leq \sum_{k=0}^{\infty}\left|\frac{x^k}{k!}\right| \\
&= \sum_{k=0}^{\infty}\frac{|x|^k}{k!} \\
&= e^{|x|} < \infty
\end{aligned}$$

(5) Regarding the formula $e^{x+y} = e^x e^y$, this follows as in the real case, as follows:

$$\begin{aligned}
e^{x+y} &= \sum_{k=0}^{\infty}\frac{(x+y)^k}{k!} \\
&= \sum_{k=0}^{\infty}\sum_{s=0}^{k}\binom{k}{s}\cdot\frac{x^s y^{k-s}}{k!} \\
&= \sum_{k=0}^{\infty}\sum_{s=0}^{k}\frac{x^s y^{k-s}}{s!(k-s)!} \\
&= e^x e^y
\end{aligned}$$

(6) Next, the continuity of $x \to e^x$ comes at $x = 0$ from the following computation:

$$\begin{aligned}
|e^t - 1| &= \left|\sum_{k=1}^{\infty}\frac{t^k}{k!}\right| \\
&\leq \sum_{k=1}^{\infty}\left|\frac{t^k}{k!}\right| \\
&= \sum_{k=1}^{\infty}\frac{|t|^k}{k!} \\
&= e^{|t|} - 1
\end{aligned}$$

(7) As for the continuity of $x \to e^x$ in general, this can be deduced as follows:

$$\begin{aligned}\lim_{t\to 0} e^{x+t} &= \lim_{t\to 0} e^x e^t\\ &= e^x \lim_{t\to 0} e^t\\ &= e^x \cdot 1\\ &= e^x\end{aligned}$$

Thus, we are led to the conclusions in the statement. □

As a consequence of the multiplicativity formula $e^{x+y} = e^x e^y$, we have:

PROPOSITION 4.12. *The exponential of complex numbers is given by*

$$e^{s+it} = e^s e^{it}$$

with e^s being a usual real exponential, and with e^{it}, in need to be computed.

PROOF. This is indeed something self-explanatory, coming from $e^{x+y} = e^x e^y$, and with the somewhat non-standard notation $x = s + it$ being something useful for later. □

Now let us get to the remaining problem, computation of e^{it} with $t \in \mathbb{R}$. Here are a few elementary observations, regarding the operation $t \to e^{it}$:

PROPOSITION 4.13. *For $t \in \mathbb{R}$ the number e^{it} belongs to the unit circle,*

$$e^{it} \in \mathbb{T}$$

and the operation $t \to e^{it}$ is subject to the following formulae,

$$e^{i(s+t)} = e^{is} e^{it} \quad , \quad e^{i0} = 1 \quad , \quad (e^{it})^{-1} = e^{it}$$

telling us $t \to e^{it}$ is a group morphism $\mathbb{R} \to \mathbb{T}$.

PROOF. There are several things going on here, the idea being as follows:

(1) To start with, we have the following formula, valid for any $x \in \mathbb{C}$:

$$e^{\bar{x}} = \sum_{k=0}^{\infty} \frac{\bar{x}^k}{k!} = \overline{\sum_{k=0}^{\infty} \frac{x^k}{k!}} = \overline{e^x}$$

We have as well the following computation, again valid for any $x \in \mathbb{C}$:

$$e^x e^{-x} = e^{x-x} = e^0 = 1 \implies (e^x)^{-1} = e^{-x}$$

(2) But with these two formulae in hand, we can prove the first assertion. Indeed, the first formula, applied with $x = it$, with $t \in \mathbb{R}$, gives the following equality:

$$e^{-it} = \overline{e^{it}}$$

As for the second formula above, again applied with $x = it$, this gives:

$$(e^{it})^{-1} = e^{-it}$$

We conclude that the complex number $z = e^{it}$ has the following property:

$$z^{-1} = \bar{z}$$

But this is exactly the equation of the unit circle $\mathbb{T}$, as desired.

(3) Regarding now the various formulae in the statement, for the operation $t \to e^{it}$, these are all trivial, coming from the multiplicativity formula $e^{x+y} = e^x e^y$.

(4) As for the final conclusion, this is something quite intuitive, telling us that $t \to e^{it}$ transforms the additive structure of $\mathbb{R}$ into the multiplicative structure of $\mathbb{T}$. □

What is next? Well, we will have to improvise a bit, and we are led in this way to the following fundamental result of Euler, regarding the complex exponential:

THEOREM 4.14. *We have the following formula,*

$$e^{it} = \cos t + i \sin t$$

valid for any $t \in \mathbb{R}$.

PROOF. Here is an intuitive proof for this, that will do I hope, and for the formal proof, this will have to wait until chapter 9 below, when discussing derivatives:

(1) We know from Proposition 4.13 that the operation $t \to e^{it}$ is a group morphism $\mathbb{R} \to \mathbb{T}$. But in view of this, barring any pathologies, this operation can only appear by "wrapping". That is, we must have a formula as follows, for a certain $\alpha \in \mathbb{R}$:

$$e^{it} = \cos(\alpha t) + i \sin(\alpha t)$$

(2) In order now to find the parameter $\alpha \in \mathbb{R}$, let us look at what happens around $t = 0$. As a first observation, at $t = 0$ precisely, our formula is as follows, true:

$$e^0 = \cos 0 + i \sin 0$$

The point now is that, around $t = 0$, we have the following elementary estimate, simply obtained by truncating the series defining the exponential:

$$e^{it} \simeq 1 + it$$

On the other hand, according to our basic trigonometry estimates for sin and cos from chapter 3, we have as well the following estimate, again around $t = 0$:

$$\cos(\alpha t) + i \sin(\alpha t) \simeq 1 + i\alpha t$$

(3) Thus, we must have $\alpha = 1$, which gives the Euler formula. With the comment, reiterated, that we will back to this later, in chapter 9, with a fully rigorous proof.

(4) Finally, let us mention that it is possible to prove, with a bit of patience, that the pathologies evoked in (1) cannot appear, that is, that our continuous group morphism $\mathbb{R} \to \mathbb{T}$ must appear indeed via wrapping. And we will leave this, turning what we have into a full, rigorous proof of the Euler formula, as an instructive exercise. □

As a main application now of the Euler formula, we have:

THEOREM 4.15. *The complex numbers $x = a + ib$ can be written exponentially,*

$$x = re^{it}$$

with the connecting formulae being

$$a = r\cos t \quad , \quad b = r\sin t$$

and in the other sense being

$$r = \sqrt{a^2 + b^2} \quad , \quad \tan t = \frac{b}{a}$$

and with r, t being called modulus, and argument.

PROOF. This is a reformulation of our previous polar writing notions, by using the formula $e^{it} = \cos t + i\sin t$ from Theorem 4.14, and multiplying everything by r. □

With this in hand, we can go back now to the basics, namely the addition and multiplication of the complex numbers. We have here the following result:

THEOREM 4.16. *In polar coordinates, the complex numbers multiply as*

$$re^{is} \cdot pe^{it} = rp\, e^{i(s+t)}$$

with the arguments s, t being taken modulo 2π.

PROOF. This is something that we already know, from chapter 3, reformulated by using the notations from Theorem 4.15. Observe that this follows as well directly, from the fact that we have $e^x e^y = e^{x+y}$, that we know from Theorem 4.11. □

As a basic application now of Theorem 4.16, we have the following result:

THEOREM 4.17. *We have the following operations on the complex numbers, written in polar form, as above:*

(1) *Inversion:* $(re^{it})^{-1} = r^{-1}e^{-it}$.
(2) *Square roots:* $\sqrt{re^{it}} = \pm\sqrt{r}e^{it/2}$.
(3) *Powers:* $(re^{it})^a = r^a e^{ita}$.
(4) *Conjugation:* $\overline{re^{it}} = re^{-it}$.

PROOF. This is something that we already know, from chapter 3, but we can discuss now all this, from a more conceptual viewpoint, the idea being as follows:

(1) We have indeed the following computation, using Theorem 4.16:

$$\begin{aligned}(re^{it})(r^{-1}e^{-it}) &= rr^{-1} \cdot e^{i(t-t)}\\ &= 1 \cdot 1\\ &= 1\end{aligned}$$

(2) Once again by using Theorem 4.16, we have:

$$(\pm\sqrt{r}e^{it/2})^2=(\sqrt{r})^2e^{i(t/2+t/2)}=re^{it}$$

(3) Given an arbitrary number $a\in\mathbb{R}$, we can define, as stated:

$$(re^{it})^a=r^ae^{ita}$$

And, due to Theorem 4.16, this operation $x\to x^a$ is indeed the correct one.

(4) This comes from the fact, that we know from chapter 2, that the conjugation operation $x\to\bar{x}$ keeps the modulus, and switches the sign of the argument. □

Let us rewrite as well the theory of roots of unity, in this way. We have here:

THEOREM 4.18. *The equation $x^N=1$ has N complex solutions, namely*

$$\left\{w^k\Big|k=0,1,\ldots,N-1\right\}\quad,\quad w=e^{2\pi i/N}$$

which are called roots of unity of order N.

PROOF. This follows from the general multiplication formula for complex numbers from Theorem 4.16. Indeed, with $x=re^{it}$ our equation reads:

$$r^Ne^{itN}=1$$

Thus $r=1$, and $t\in[0,2\pi)$ must be a multiple of $2\pi/N$, as stated. □

As an illustration here, the roots of unity of small order, along with some of their basic properties, which are very useful for computations, are as follows:

$\underline{N=1}$. Here the unique root of unity is 1.

$\underline{N=2}$. Here we have two roots of unity, namely 1 and -1.

$\underline{N=3}$. Here we have 1, then $w=e^{2\pi i/3}$, and then $w^2=\bar{w}=e^{4\pi i/3}$.

$\underline{N=4}$. Here the roots of unity, read as usual counterclockwise, are $1,i,-1,-i$.

$\underline{N=5}$. Here, with $w=e^{2\pi i/5}$, the roots of unity are $1,w,w^2,w^3,w^4$.

$\underline{N=6}$. Here a useful writing is $\{\pm1,\pm w,\pm w^2\}$, with $w=e^{2\pi i/3}$.

$\underline{N=7}$. Here, with $w=e^{2\pi i/7}$, the roots of unity are $1,w,w^2,w^3,w^4,w^5,w^6$.

$\underline{N=8}$. Here the roots of unity, read as usual counterclockwise, are the numbers $1,w,i,iw,-1,-w,-i,-iw$, with $w=e^{\pi i/4}$, which is also given by $w=(1+i)/\sqrt{2}$.

The roots of unity are very useful variables, and have many interesting properties. We will use them on a regular basis, in what follows, for all sorts of questions.

4c. The logarithm

Getting back now to the real numbers, but this will be temporary, do not worry, we can talk about the logarithm function, which appears as follows:

THEOREM 4.19. *The exponential function, as constructed before,*

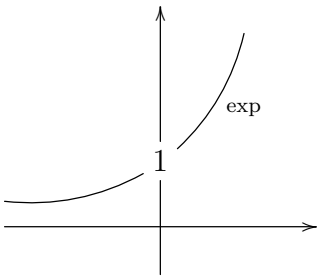
$$\exp : \mathbb{R} \to (0, \infty) \quad , \quad x \to e^x$$

is invertible, with its inverse being the logarithm function, denoted as follows:

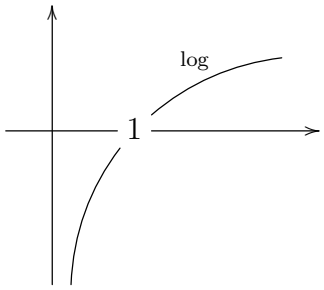
$$\log : (0, \infty) \to \mathbb{R} \quad , \quad \log = \exp^{-1}$$

This logarithm function is continuous, increasing, and bijective.

PROOF. This follows indeed from our general inversion machinery for functions from chapter 2. Let us record as well some pictures. The exponential looks as follows:



As for the logarithm function, obtained by flipping the above graph of the exponential around the main diagonal of the plane, $x = y$, this looks as follows:



Observe the asymptotes of the above graphs, related to each other, coming from $\exp(-\infty) = 0$ and $\log 0 = -\infty$. Also, the slopes of both graphs at the points designated "1" in the above are $\pi/4$, with this coming from $e^t \simeq 1 + t$. More on this later. □

In practice, the logarithm is a very useful function, and more on this later. In the meantime, how to compute it? And here, we have the following collection of results:

THEOREM 4.20. *The logarithm can be computed via the equivalent formulae*

$$e^{\log x} = x \quad , \quad \log(e^x) = x$$

and in practice, we have the following useful rules:

(1) $\log(xy) = \log x + \log y$.
(2) $\log(1/y) = -\log y$.
(3) $\log(x/y) = \log x - \log y$.
(4) $\log(x^p) = p \log x$.

PROOF. The first two formulae come from the fact that the inverse of a function f can be defined either via $f(f^{-1}(x)) = x$, or via $f^{-1}(f(x)) = x$. As for the rest:

(1) This comes indeed from the following computation:

$$e^{\log(xy)} = xy = e^{\log x} e^{\log y} = e^{\log x + \log y}$$

(2) This comes from (1), by setting $x = 1/y$, and using $\log 1 = 0$.

(3) This comes also from (1), by replacing $y \to 1/y$, and using (2).

(4) This formula, generalizing (1) with $x = y$, and also (2), comes as follows:

$$e^{\log(x^p)} = x^p = (e^{\log x})^p = e^{p \log x}$$

Thus, we are led to the conclusions in the statement. □

In what regards now the applications of log, these usually come as follows:

THEOREM 4.21. *We can talk about logarithms in any basis $b > 1$, according to*

$$\log_b x = \frac{\log x}{\log b}$$

with these generalized logarithms solving the following equivalent questions,

$$b^{\log_b x} = x \quad , \quad \log_b(b^x) = x$$

and with $b = e, 2, 10$ being good for hard science, computer science, and social science.

PROOF. We have several assertions here, the idea being as follows:

(1) Given $b > 1$ the power function $[x \to b^x] : \mathbb{R} \to (0, \infty)$ is invertible, and we can call $\log_b x : (0, \infty) \to \mathbb{R}$ is inverse, defined by the two equivalent formulae in the statement, exactly as before in Theorems 4.19 and 4.20, dealing with the case $b = e$.

(2) Next, the good news is that we do not need to compute all these logarithms, and this thanks to the following computation, which gives $\log_b x = \log x / \log b$, as stated:

$$b^{\log x / \log b} = (e^{\log b})^{\log x / \log b} = e^{\log x} = x$$

(3) Finally, the last assertion is something subjective. Regarding computers, everything there is binary, 2^n numbers rule, and the interest in $\log_2$ comes from:

$$\log_2(2^n) = n$$

Pretty much the same can be said about social science, where 10^n numbers, such as millions, billions and trillions rule, and the interest in $\log_{10}$ comes from:

$$\log_{10}(10^n) = n$$

And exercise for you to exploit this technology, and make some money, say with some graphs and advertisements using suitable logarithmic scales, known only to you.

(3) As for $\log_e = \log$ and hard science, more on this later, on several occasions. In the meantime, let us record the following useful formula, related to Fact 4.7:

$$b^t = e^{t\log b} \simeq 1 + t\log b$$

Indeed, this tells us that $e^t \simeq 1 + t$ uniquely determines e, as said in Fact 4.7. □

Time now for some tough mathematics? Here is a main result about log:

THEOREM 4.22. *We have the following formula, valid for $|x| < 1$:*

$$\log(1 + x) = \sum_{k=1}^{\infty}(-1)^{k+1}\frac{x^k}{k}$$

Moreover, this holds as well at $x = 1$, with the formula here being:

$$\log 2 = \sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}$$

As for remaining cases, $|x| > 1$ and $x = -1$, here the series diverges.

PROOF. We have several things going on here, the idea being as follows:

(1) To start with, the series in the statement converges at $|x| < 1$, due to:

$$\frac{x^{k+1}}{k+1} : \frac{x^k}{k} = x \cdot \frac{k+1}{k} \to x$$

The series converges as well at $x = 1$, say as an alternating series. At $x = -1$ however we obtain the Riemann divergent sum. As for the case $|x| > 1$, here the series grossly diverges, in the sense that its terms do not go to 0, as required by convergence.

(2) Getting now to the proof, we have a dillema here, coming from the following two equivalent formulae, that we can both use for our computations:

$$e^{\log x} = x \quad , \quad \log(e^x) = x$$

(3) Let us try the first method. We have the following computation:

$$\begin{aligned}\exp\left(\sum_{k=1}^{\infty}(-1)^{k+1}\frac{x^k}{k}\right) &= \sum_{n=0}^{\infty}\frac{1}{n!}\left(\sum_{k=1}^{\infty}(-1)^{k+1}\frac{x^k}{k}\right)^n\\ &= \sum_{n=0}^{\infty}\frac{1}{n!}\sum_{k_1,\ldots,k_n=1}^{\infty}(-1)^{k_1+\ldots+k_n+n}\frac{x^{k_1+\ldots+k_n}}{k_1\ldots k_n}\\ &= \sum_{s=0}^{\infty}(-x)^s\sum_{n=0}^{\infty}\frac{(-1)^n}{n!}\sum_{k_1+\ldots+k_n=s}\frac{1}{k_1\ldots k_n}\end{aligned}$$

We would like to prove that this equals $1+x$, but the computation of the coefficients on the right looks like a non-trivial business, so we should better stop here.

(4) With the second method, we have the following computation:

$$\begin{aligned}\sum_{k=1}^{\infty}(-1)^{k+1}\frac{(e^x-1)^k}{k} &= \sum_{k=1}^{\infty}(-1)^{k+1}\frac{1}{k}\left(\sum_{r=1}^{\infty}\frac{x^r}{r!}\right)^k\\ &= \sum_{k=1}^{\infty}(-1)^{k+1}\frac{1}{k}\sum_{r_1,\ldots,r_k=1}^{\infty}\frac{x^{r_1+\ldots+r_k}}{r_1!\ldots r_k!}\\ &= \sum_{s=1}^{\infty}x^s\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}\sum_{r_1+\ldots+r_k=s}\frac{1}{r_1!\ldots r_k!}\end{aligned}$$

We would like to prove that this equals x, and although the coefficients on the right look better than those before, finishing does look easy, and we will stop here too.

(5) So, what to do? Third method I guess, by proving that our series has the main required abstract property of $\log(1+x)$, without reference to exp, namely:

$$\log((1+x)(1+y))=\log(1+x)+\log(1+y)$$

So, let us get into this. We have the following computation, to start with:

$$\begin{aligned}\sum_{k=1}^{\infty}(-1)^{k+1}\frac{(x+y+xy)^k}{k} &= \sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}\sum_{r=0}^{k}\binom{k}{r}(x+xy)^r y^{k-r}\\ &= \sum_{r+s\geq 1}\frac{(-1)^{r+s+1}}{r+s}\binom{r+s}{r}x^r(1+y)^r y^s\end{aligned}$$

(6) The contribution of $r=0$ to this sum is the following quantity:

$$C_0=\sum_{s\geq 1}(-1)^{s+1}\frac{y^s}{s}$$

(7) As for the contribution coming from $r \geq 1$, this is the following quantity, with the computation using the inversion formula for $(1+y)^r$, that we know from chapter 2:

$$\begin{aligned}
C_+ &= \sum_{r\geq 1}(-1)^{r+1}x^r(1+y)^r\sum_{s\geq 0}\frac{(-1)^s}{r+s}\binom{r+s}{r}y^s\\
&= \sum_{r\geq 1}(-1)^{r+1}\frac{x^r(1+y)^r}{r}\sum_{s\geq 0}\frac{(-1)^s}{r+s}\binom{r+s-1}{r-1}y^s\\
&= \sum_{r\geq 1}(-1)^{r+1}\frac{x^r(1+y)^r}{r}\cdot\frac{1}{(1+y)^r}\\
&= \sum_{r\geq 1}(-1)^{r+1}\frac{x^r}{r}
\end{aligned}$$

(8) Summarizing, we have proved the desired formula, namely:

$$\sum_{k=1}^{\infty}(-1)^{k+1}\frac{(x+y+xy)^k}{k}=\sum_{r=1}^{\infty}(-1)^{r+1}\frac{x^r}{r}+\sum_{s=1}^{\infty}(-1)^{s+1}\frac{y^s}{s}$$

(9) But with this, we can now finish, a bit as we did before for exp, in the proof of Theorem 4.4. To be more precise, and taking a shortcut here, we can argue that by (8) our candidate series for $\log(1+x)$ must be a certain shifted logarithm, $\log_b(1+x)$. But with the slope at $x=0$ being the correct one, this shifted logarithm must be the usual one, $\log(1+x)$, and we are therefore led to the conclusion in the statement. □

As an interesting observation now about the logarithm, we have:

OBSERVATION 4.23. *We can successfully deal with x^y with $x>0$ by using:*

$$x^y=e^{y\log x}$$

However, this fails at $x<0$, with for instance $(-1)^2=1$ being beyond its reach.

And the problem is, how to fix this? In answer, by using complex numbers, because by setting for instance $\log(-1)=\pi i$, we have the following successful computation:

$$(-1)^2=e^{2\log(-1)}=e^{2\pi i}=1$$

We will be back to such questions in the next chapter, when discussing more on detail the complex functions, and in particular, the complex extensions of the logarithm.

4d. Poisson laws

Getting back now to the basics, I don't know about you, but personally, for some peace of mind, I would like to have as well a combinatorial interpretation of e. In order to reach to this, let us start with the following definition, which is very standard:

DEFINITION 4.24. *A permutation of $\{1,\ldots,N\}$ is a bijection, as follows:*

$$\sigma : \{1,\ldots,N\} \to \{1,\ldots,N\}$$

The set of such permutations is denoted S_N.

Many things can be said about permutations, in particular with the basic fact that there are $N!$ of them. Indeed, in order to construct a permutation $\sigma \in S_N$, we have:

– N choices for the value of $\sigma(N)$.

– $(N-1)$ choices for the value of $\sigma(N-1)$.

– $(N-2)$ choices for the value of $\sigma(N-2)$.

$\vdots$

– and so on, up to 1 choice for the value of $\sigma(1)$.

Thus, we have indeed $N!$ choices, as claimed. Observe also that the set S_N formed by the permutations is a group, being obviously stable by composition, and by inversion.

With this discussed, we have the following remarkable result, which is a bit of group theory flavor, making appear the number e, in a nice combinatorial way:

THEOREM 4.25. *The probability for a random permutation $\sigma \in S_N$ to be a derangement, that is, to have no fixed points, is given by the following formula:*

$$P = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \ldots + (-1)^N \frac{1}{N!}$$

Thus we have the following asymptotic formula, in the $N \to \infty$ limit,

$$P \simeq \frac{1}{e}$$

with $e = 2.7182\ldots$ being the usual constant from analysis.

PROOF. This is something very classical, and beautiful too, which is best viewed by using the inclusion-exclusion principle. Consider indeed the following sets:

$$S_N^k = \left\{\sigma \in S_N \middle| \sigma(k) = k\right\}$$

The set of permutations having no fixed points, called derangements, is then:

$$X_N = \left(\bigcup_k S_N^k\right)^c$$

Now the inclusion-exclusion principle tells us that we have:

$$\begin{aligned}
|X_N| &= \left|\left(\bigcup_k S_N^k\right)^c\right| \\
&= |S_N| - \sum_k |S_N^k| + \sum_{k<l} |S_N^k \cap S_N^l| - \ldots + (-1)^N \sum_{k_1<\ldots<k_N} |S_N^{k_1} \cap \ldots \cap S_N^{k_N}| \\
&= N! - N(N-1)! + \binom{N}{2}(N-2)! - \ldots + (-1)^N \binom{N}{N}(N-N)! \\
&= \sum_{r=0}^{N} (-1)^r \binom{N}{r}(N-r)! \\
&= \sum_{r=0}^{N} (-1)^r \frac{N!}{r!}
\end{aligned}$$

Thus, the probability that we are interested in, for a random permutation $\sigma \in S_N$ to have no fixed points, is given by the following formula:

$$P = \frac{|X_N|}{N!} = \sum_{r=0}^{N} \frac{(-1)^r}{r!}$$

Now since on the right we have the expansion of $1/e$, this gives the result. □

The above is nice, but we can do even better. Let us introduce, indeed:

DEFINITION 4.26. *The Poisson law of parameter* 1 *is the following measure,*

$$p_1 = \frac{1}{e} \sum_{k \in \mathbb{N}} \frac{\delta_k}{k!}$$

and the Poisson law of parameter $t > 0$ *is the following measure,*

$$p_t = e^{-t} \sum_{k \in \mathbb{N}} \frac{t^k}{k!} \delta_k$$

with the letter "p" standing for Poisson.

Generally speaking, these measures appear a bit everywhere, in discrete probability contexts, the reasons for this coming from the Poisson Limit Theorem (PLT). However, this is something quite advanced, and for our purposes here, what we have in Definition 4.26 will do. Observe that our laws have indeed mass 1, as they should, due to:

$$e^t = \sum_{k \in \mathbb{N}} \frac{t^k}{k!}$$

Getting back now to what we wanted to do, generalize Theorem 4.25, we have:

THEOREM 4.27. *The number of fixed points of permutations, $\chi : S_N \to \mathbb{N}$ given by*

$$\chi(\sigma) = \#\left\{i \in \{1, \ldots, N\} \middle| \sigma(i) = i\right\}$$

follows the Poisson law p_1, in the $N \to \infty$ limit. More generally, the variable

$$\chi_t(\sigma) = \#\left\{i \in \{1, \ldots, [tN]\} \middle| \sigma(i) = i\right\}$$

with $t \in (0, 1]$ follows the Poisson law p_t, in the $N \to \infty$ limit.

PROOF. We have two assertions here, the idea being as follows:

(1) In what regards the first assertion, dealing with the case $t = 1$, we must prove here the following formula, for any $r \in \mathbb{N}$, in the $N \to \infty$ limit:

$$P(\chi = r) \simeq \frac{1}{r!e}$$

We already know, from Theorem 4.25, that this formula holds at $r = 0$. In the general case now, we have to count the permutations $\sigma \in S_N$ having exactly r points. Now since having such a permutation amounts in choosing r points among $1, \ldots, N$, and then permuting the $N - r$ points left, without fixed points allowed, we have:

$$\begin{aligned}
\#\left\{\sigma \in S_N \middle| \chi(\sigma) = r\right\} &= \binom{N}{r} \#\left\{\sigma \in S_{N-r} \middle| \chi(\sigma) = 0\right\} \\
&= \frac{N!}{r!(N-r)!} \#\left\{\sigma \in S_{N-r} \middle| \chi(\sigma) = 0\right\} \\
&= N! \times \frac{1}{r!} \times \frac{\#\left\{\sigma \in S_{N-r} \middle| \chi(\sigma) = 0\right\}}{(N-r)!}
\end{aligned}$$

By dividing everything by $N!$, we obtain from this the following formula:

$$\frac{\#\left\{\sigma \in S_N \middle| \chi(\sigma) = r\right\}}{N!} = \frac{1}{r!} \times \frac{\#\left\{\sigma \in S_{N-r} \middle| \chi(\sigma) = 0\right\}}{(N-r)!}$$

Now by using the computation at $r = 0$, that we already have, from Theorem 4.25, it follows that with $N \to \infty$ we have the following estimate:

$$P(\chi = r) \simeq \frac{1}{r!} \cdot P(\chi = 0) \simeq \frac{1}{r!} \cdot \frac{1}{e}$$

Thus, we obtain as limiting measure the Poisson law of parameter 1, as stated.

(2) In the general case now, where $t \in (0,1]$ is arbitrary, we can use the same method. Indeed, as before at $t = 1$, we obtain by inclusion-exclusion that we have:

$$\begin{aligned} P(\chi_t = 0) &= \frac{1}{N!}\sum_{r=0}^{[tN]}(-1)^r \sum_{k_1<\ldots<k_r<[tN]} |S_N^{k_1} \cap \ldots \cap S_N^{k_r}| \\ &= \frac{1}{N!}\sum_{r=0}^{[tN]}(-1)^r \binom{[tN]}{r}(N-r)! \\ &= \sum_{r=0}^{[tN]} \frac{(-1)^r}{r!} \cdot \frac{[tN]!(N-r)!}{N!([tN]-r)!} \end{aligned}$$

Now with $N \to \infty$, we obtain from this the following estimate:

$$P(\chi_t = 0) \simeq \sum_{r=0}^{[tN]} \frac{(-1)^r}{r!} \cdot t^r \simeq e^{-t}$$

More generally, by counting the permutations $\sigma \in S_N$ having exactly r fixed points among $1, \ldots, [tN]$, as in the proof of (2), we obtain:

$$P(\chi_t = r) \simeq \frac{t^r}{r!e^t}$$

Thus, we obtain in the limit a Poisson law of parameter t, as stated. □

Many other things can be said, as a continuation of the above, both of algebraic and analytic nature. We will be back to this on several occassions, in what follows.

4e. Exercises

This was our first tricky analysis chapter, and as exercises on this, we have:

EXERCISE 4.28. *Rewrite the above theory, by starting with* $e = \sum_k 1/k!$.

EXERCISE 4.29. *Rewrite then the theory, by starting with* $e^x = \sum_k x^k/k!$.

EXERCISE 4.30. *Rewrite then again the theory, by starting with* $e^t \simeq 1 + t$.

EXERCISE 4.31. *Then do the same, starting this time with* $e^{x+t} \simeq e^x(1+t)$.

EXERCISE 4.32. *Learn about formal derivatives, and* $\exp' = \exp$, $\exp(0) = 1$.

EXERCISE 4.33. *Learn about analytic derivatives, and* $\exp' = \exp$, $\exp(0) = 1$.

EXERCISE 4.34. *Can you rewrite the theory starting from* $e^{it} = \cos t + i \sin t$*?*

EXERCISE 4.35. *What about rewriting the theory, with* log *coming first?*

As bonus exercise, get a cat or two, their advice might be useful, for such questions.

Part II

Continuity

Here we are
To celebrate a party
In this hot summer night
While the moon is shining bright

CHAPTER 5

Continuity, revised

5a. Continuity, jumps

We have seen so far that there is life in mathematical analysis, with very little continuity involved, the derivatives barely appearing, and the integrals not present at all. This is of course not surprising, because modern analysis as we know it, crucially based on continuity, derivatives and integrals, remains something recent. And there was a long story preceding it, story whose essentials we more or less presented in Part I.

Well, time now to get into modern analysis, with a more detailed discussion of continuity, over the next 100 pages, then a discussion of derivatives, over 100 more pages, and then a discussion of integration, over 100 more pages. This will be our plan.

Getting started, with continuity and related topics, let us make a brief survey of what we know, from Part I. The starting definition, coming in 2 flavors, was as follows:

DEFINITION 5.1. *A function $f : X \to \mathbb{R}$, with $X \subset \mathbb{R}$, is continuous at $x \in X$ when:*

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

Equivalently, the following epsilon-delta condition must be satisfied:

$$\forall \varepsilon > 0 \ \exists \delta > 0 \ , \ |x - y| < \delta \implies |f(x) - f(y)| < \varepsilon$$

We say that f is continuous when it is continuous at all points $x \in X$.

And we refer to chapter 2 for a discussion here. As a main result now regarding continuity, we have the intermediate value theorem, also from chapter 2, as follows:

THEOREM 5.2. *Given a continuous function defined on a closed bounded interval*

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

its image must be a closed bounded interval, $f([a, b]) = f([c, d])$.

PROOF. This is something discussed in chapter 2, the idea being as follows:

(1) Given $u \in [f(a), f(b)]$, we can solve $f(x) = u$ by dividing $[a, b]$ into half, selecting the half whose image contains u, then again dividing in half, and so on, with the limiting point x of these decreasing intervals satisfying, by continuity, $f(x) = u$, as desired.

(2) Thus our function f takes its intermediate values, and with a bit more work, we are led to the conclusion in the statement. We will actually present a second proof of this result, more conceptual, using modern topology methods, later in this chapter. □

Finally, regarding the examples of continuous functions, there are many of them, coming from our work from chapters 2–4, with the summary here being as follows:

THEOREM 5.3. *The following functions are continuous:*

(1) λf, $f + g$, fg, $f \circ g$, *provided that* f, g *are continuous.*
(2) f/g *too, on its domain, meaning outside the zeroes of* g.
(3) *The polynomials* $P \in \mathbb{R}[X]$, *viewed as functions* $P : \mathbb{R} \to \mathbb{R}$.
(4) *The rational functions,* $f = P/Q$ *with* $P, Q \in \mathbb{R}[X]$, *outside their poles.*
(5) *The inverses* $f^{-1} : Y \to X$ *of bijective continuous functions* $f : X \to Y$.
(6) *The power functions* x^a *and* a^x, *over their suitable respective domains.*
(7) *The basic trigonometric functions, namely* sin, cos *and* exp, log.
(8) *The tangent function* tan *too, over its domain,* $\mathbb{R} - (\pi/2 + \pi\mathbb{Z})$.

PROOF. Good result that we have here, based on a lot of work, as follows:

(1) This is something from chapter 2, coming from definitions.

(2) This is again from chapter 2, again coming from definitions.

(3) This comes by applying the operations in (1) to the function $f(x) = x$.

(4) This comes by combining (2) and (3), that is, when using all operations.

(5) This comes from Theorem 5.2, and with the remark that f must be monotone.

(6) This was something more tricky, basically coming from our definition of $\mathbb{R}$.

(7) This is something that we discussed in great detail, in chapters 3 and 4.

(8) This is something that we know too, coming via (2) from $\tan = \sin / \cos$. □

What is next? Many things, and as a first task, inspired by the rational functions, and by the tangent function too, let us get away from continuity, and have a quick look at discontinuity. In relation with this, here is something that we can do, in general:

PROPOSITION 5.4. *Given a function* $f : X \to \mathbb{R}$ *and a point* $x \in X$, *consider*

$$l = \lim_{y \to x} f(y)$$

with this limit being computed over points $y \neq x$. *Then:*

(1) f *is continuous at* x *precisely when* $f(x) = l$.
(2) *If* l *exists and* $f(x) \neq l$, *we can reset* $f(x) = l$, *and* f *becomes continuous at* x.
(3) *If* l *does not exist,* $f(x)$ *cannot be reset, as for* f *to become continuous at* x.

PROOF. This is indeed something self-explanatory, coming from Definition 5.1. □

As a continuation of this, again motivated by rational functions and the tangent, we would like to talk about jumps of discontinuous functions. However, things here are a bit tricky, because given $f : X \to \mathbb{R}$ we would like to talk about both jumps at points $x \in X$ where f is discontinous, and with these being defined or not, and at points $x \notin X$ where jumps can still appear, under certain circumstances, say as for $f = \tan$ at $x = \pi/2$.

In view of this, it is convenient to use the following method, which allows us to get rid to some of the problems, related to the precise geometry of the domain $X \subset \mathbb{R}$:

METHOD 5.5. *Given $f : X \to \mathbb{R}$ with $X \subset \mathbb{R}$, we can regard it as function*

$$f : \mathbb{R} \to \mathbb{R}$$

simply by setting $f(x) = 0$, at any $x \notin X$.

Obviously, this is something quite theoretical, but in order to develop the abstract theory of jumps, without much troubles, this method is what we need. Let us start with a remake of Proposition 5.4 in this setting, that of the functions $f : \mathbb{R} \to \mathbb{R}$:

THEOREM 5.6. *Given a function $f : \mathbb{R} \to \mathbb{R}$ and a point $x \in \mathbb{R}$, consider*

$$f(x_-) = \lim_{y \nearrow x} f(y) \quad , \quad f(x_+) = \lim_{y \searrow x} f(y)$$

with these limits being computed over strictly monotone sequences. Then:

(1) *f is continuous at x precisely when $f(x) = f(x_-) = f(x_+)$.*
(2) *If $f(x_-) = f(x_+) = l$, we can reset $f(x) = l$, and f becomes continuous at x.*
(3) *Otherwise, $f(x)$ cannot be reset, as for f to become continuous at x.*

PROOF. This is indeed something self-explanatory, remake of Proposition 5.4. □

We are now ready to talk about jumps of functions, as follows:

DEFINITION 5.7. *Given a function $f : \mathbb{R} \to \mathbb{R}$ and a point $x \in \mathbb{R}$, set*

$$f(x_-) = \lim_{y \nearrow x} f(y) \quad , \quad f(x_+) = \lim_{y \searrow x} f(y)$$

provided that these two limits exist indeed. We call then the quantity

$$J_f(x) = f(x_+) - f(x_-)$$

which does not depend on $f(x)$, the jump of f at the point $x \in \mathbb{R}$.

As a first observation that you might have, this does not really cover the main examples that we have in mind, namely the rational functions, and tan, which jump by $\pm\infty$, when they do. Good point, and in answer, let us complement the above definition with:

DEFINITION 5.8 (addendum). *We agree to allow infinite limits in the above,*

$$f(x_-), f(x_+) \in [-\infty, \infty]$$

leading to the jumps taking the following values, finite, infinite and formal,

$$J_f(x) \in [-\infty, \infty] \cup \{0^\pm\}$$

with the rules $\infty - \infty = 0^+$ *and* $(-\infty) - (-\infty) = 0^-$*, for their computation.*

Which sounds quite good, with this we can definitely talk about the jumps of all functions that we have in mind. Before that, however, a few theoretical remarks. Generally speaking, in analysis, we have the following formula, never to be forgotten:

$$\infty - \infty = \text{ undefined}$$

To be more precise, this is a theorem, coming from the following computations, which show not only that $\infty - \infty$ is undefined, but that we can say really nothing, about it:

$$n^2 - n \to \infty \quad , \quad n - n^2 \to -\infty \quad , \quad (n + c) - n \to c$$

And the same goes for $(-\infty) - (-\infty)$. In our situation, however, jumps of functions, things are a bit different, because think for instance at $f(x) = 1/x^2$ at $x = 0$, wouldn't you like to say that the jump there is $\infty - \infty = 0$. Similarly, regarding $f(x) = -1/x^2$ at $x = 0$, wouldn't you like to say that the jump there is $(-\infty) - (-\infty) = 0$.

Thus, we are led to the above conventions. At the level of examples, we have:

THEOREM 5.9. *The following happen, under our present conventions:*

(1) *The sign function* $\text{sgn} : \mathbb{R} \to \{-1, 0, 1\}$ *jumps at* 0 *by* $J = 2$.
(2) *The function* $1/x^n$ *with* n *even jumps at* 0 *by* $J = 0^+$.
(3) *The function* $1/x^n$ *with* n *odd jumps at* 0 *by* $J = \infty$.
(4) *The rational functions jump by* $J = 0^\pm$ *or* $J = \pm\infty$*, when they do.*
(5) *The tangent* $\tan : \mathbb{R} - \{\mathbb{Z}\pi + \pi/2\} \to \mathbb{R}$ *jumps by* $J = -\infty$*, at* $x = k\pi + \pi/2$.

PROOF. To start with, as mentioned, all this is to be taken under our various conventions, namely those in Method 5.5, and in Definition 5.8. As for the proof:

(1) This is indeed something trivial, the computation here being as follows:

$$\begin{aligned} J_f(0) &= \lim_{y \searrow 0} f(y) - \lim_{y \nearrow 0} f(y) \\ &= \lim_{y \searrow 0} 1 - \lim_{y \nearrow 0} (-1) \\ &= 1 - (-1) \\ &= 2 \end{aligned}$$

(2) This is again trivial, because for the function $f(x) = 1/x^{2k}$ we have:

$$\begin{aligned} J_f(0) &= \lim_{y\searrow 0} \frac{1}{y^{2k}} - \lim_{y\nearrow 0} \frac{1}{y^{2k}} \\ &= \infty - \infty \\ &= 0^+ \end{aligned}$$

(3) Again, this is trivial, because for the function $f(x) = 1/x^{2k+1}$ we have:

$$\begin{aligned} J_f(0) &= \lim_{y\searrow 0} \frac{1}{y^{2k+1}} - \lim_{y\nearrow 0} \frac{1}{y^{2k+1}} \\ &= \infty - (-\infty) \\ &= \infty \end{aligned}$$

(4) Consider indeed a rational function f, at one of its poles $r \in \mathbb{R}$. As explained in chapter 2, this means that we have a formula as follows, with $n \geq 1$ being a certain exponent, and with $P, Q \in \mathbb{R}[X]$ being polynomials satisfying $P(r) \neq 0$, $Q(r) \neq 0$:

$$f(x) = \frac{P(x)}{(x-r)^n Q(x)}$$

Now when computing the various limits $x \to r$, the formula is as follows:

$$\lim_{x\to r} f(x) = \frac{P(r)}{Q(r)} \times \lim_{x\to r} \frac{1}{(x-r)^n}$$

Thus, we are led to the situations in (2,3), and with the answer being as follows:

$$n \text{ even},\ P(r)/Q(r) > 0 \implies J = 0^+$$

$$n \text{ even},\ P(r)/Q(r) < 0 \implies J = 0^-$$

$$n \text{ odd},\ P(r)/Q(r) > 0 \implies J = \infty$$

$$n \text{ odd},\ P(r)/Q(r) < 0 \implies J = -\infty$$

(5) This is again something trivial, because for $f(x) = \tan x$ we have:

$$\begin{aligned} J_f(\pi/2) &= \lim_{y\searrow \pi/2} \tan y - \lim_{y\nearrow \pi/2} \tan y \\ &= -\infty - \infty \\ &= -\infty \end{aligned}$$

As for the jump at an arbitrary $x = k\pi + \pi/2$, this is also $-\infty$, by periodicity. □

Summarizing, we have examples. At the level of the general theory now, we have:

THEOREM 5.10. *Assuming that a function $f : X \to \mathbb{R}$ does not jump at $x \in X$,*

$$J_f(x) = 0$$

we can modify our function by forgetting the old value $f(x)$, and setting

$$f(x) = f(x_-) = f(x_+)$$

and we obtain in this way a function which is continuous at x.

PROOF. This is indeed something self-explanatory, remake of our previous Theorem 5.6, by using the notion of jump, as axiomatized in Definition 5.7. By the way, observe that we can formally do the same when the jump is $0^\pm$, with the obvious convention for the continuity of the functions taking infinite values, $f : X \to [-\infty, \infty]$. □

The above result is quite interesting, and can be applied to the various points where f is discontinuous, provided that these points are isolated from each other. In the case where f needs a "fix" on a more substantial set of points, such as a whole interval $(a, b) \subset \mathbb{R}$, things are more complicated, requiring advanced technology, such as analytic continuation of the complex functions. More on such questions, hopefully, later in this book.

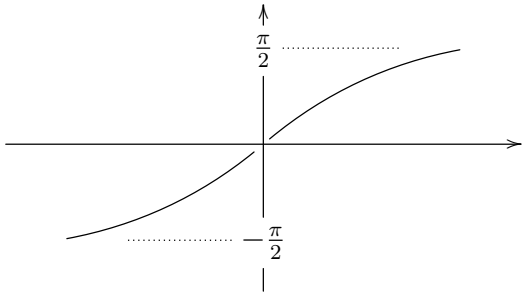
Finally, as a question that you might have, we surely saw in Theorem 5.9 that the sign function has jump $J = 2$ at zero, but this remains a special example, and in what regards the usual functions that we know, the jump always seems to be as follows:

$$J \in \left\{0, 0^+, 0^-, -\infty, \infty\right\}$$

And good question this is, because in order to have a non-trivial jump $J \in \mathbb{R} - \{0\}$, your function f must come from some kind of catastrophic phenomenon, breaking symmetry, and the laws of math and physics, such as a thermonuclear explosion.

In short, I do not have a good answer to your question, and in the lack of the cats around, who know one or two things about jumps, spontaneous symmetry breaking and so on, here is however some sort of modest mathematical answer to this:

FACT 5.11. *Non-trivial jumps appear for instance by horizontally compressing* $\tan^{-1}$*,*



which in the limit leads to $\pi/2$ times the sign function, having jump $J = \pi$ at zero.

And more on such things later, in chapter 6, when talking about limits of continuous functions. Finally, along the same lines, let us mention too, as further examples:

(1) The slope of $f(x) = |x|$ is the sign function, having jump $J = 2$ at zero. And if you find the function $|x|$ to be something quite unnatural, and not fitting with the rest, I can still argue that $f(x) = \sqrt{x^2}$, composition of basic continuous functions.

(2) Given a discrete probability measure $\mu = \sum_i c_i \delta_{x_i}$, such as the Poisson laws that we met in chapter 4, and a variable following it, $\varphi \sim \mu$, the function $f(x) = P(\varphi \leq x)$ makes its way from $f(-\infty) = 0$ to $f(\infty) = 1$, increasingly, by jumping at each x_i.

However, these latter examples remain a bit borderline, due to various reasons, and the main example remains the one in Fact 5.11. We will be back to this, in chapter 6.

5b. Topology methods

Changing topics, we would like to explain now an alternative formulation of the notion of continuity, which is something quite abstract, but useful and powerful, and worth learning. Let us start with the following definition, which is probably new to you:

DEFINITION 5.12. *The open and closed sets are defined as follows:*

(1) *Open means that there is a small interval around each point.*
(2) *Closed means that our set is closed under taking limits.*

As basic illustrations for this, the open intervals are open, and the closed intervals are closed, as you would expect, as shown by the following result:

PROPOSITION 5.13. *The following happen:*

(1) *The open intervals (a, b) are open.*
(2) *The closed intervals $[a, b]$ are closed.*

PROOF. This is something fairly easy, as follows:

(1) Consider indeed an open interval (a, b), and a point inside, $x \in (a, b)$. We have then an inclusion as follows, which does the job, as in Definition 5.12 (1):

$$x \in \left(\frac{a+x}{2}, \frac{x+b}{2} \right) \subset (a, b)$$

Observe that our proof works as well for $a, b = \pm\infty$, with the following rules:

$$x - \infty = -\infty \quad , \quad x + \infty = \infty$$

(2) Consider now a closed interval $[a, b]$, and a sequence inside $\{x_n\} \subset [a, b]$. Assuming that this sequence converges, $x_n \to x$, we have:

$$x_n \geq a \implies x \geq a \quad , \quad x_n \leq b \implies x \leq b$$

Thus we obtain $x \in [a, b]$, as required by Definition 5.12 (2). □

Further basic examples of open and closed sets, or rather results which are easy to establish, producing further examples of open and closed sets, are as follows:

THEOREM 5.14. *The following happen:*

(1) *Union of open sets is open.*
(2) *Intersection of closed sets is closed.*
(3) *Finite intersection of open sets is open.*
(4) *Finite union of closed sets is closed.*

PROOF. This is something elementary, the idea being as follows:

(1) Consider indeed a point, belonging to a union of open sets:

$$x \in \bigcup_i O_i$$

We have then $x \in O_i$ for some i, which tells us that we have a certain interval $x \in (a, b) \subset O_i$. But this gives, as desired, an inclusion as follows:

$$x \in (a, b) \subset O_i \subset \bigcup_i O_i$$

(2) Consider now a converging sequence, inside an intersection of closed sets:

$$\{x_n\} \subset \bigcap_i C_i$$

Since we have $\{x_n\} \subset C_i$ for any i, we deduce that the limit of our sequence $x = \lim_{n\to\infty} x_n$ belongs to all the sets C_i. Thus we obtain, as desired:

$$\lim_{n\to\infty} x_n \in \bigcap_i C_i$$

(3) This is something more subtle, which requires a bit more work. Consider indeed a point, belonging to a finite intersection of open sets:

$$x \in \bigcap_i O_i$$

We have then $x \in O_i$ for any i, which tells us that we have a certain interval $x \in (a_i, b_i) \subset O_i$. Now let us consider the following intersection:

$$(a, b) = \bigcap_i (a_i, b_i)$$

Observe that, since the intersection is finite, we obtain indeed an open interval (a, b) as above, with the precise formulae of the bounds a, b being as follows:

$$a = \max_i a_i \quad , \quad b = \min_i b_i$$

But with this, we are done, because we obtain an inclusion as follows:

$$x \in (a,b) = \bigcap_i (a_i, b_i) \subset \bigcap_i O_i$$

(4) Consider a converging sequence, inside a finite union of closed sets:

$$\{x_n\} \subset \bigcup_i C_i$$

Since the union is finite, we can find a set C_i which contains infinitely many terms of our sequence x_n. That is, we can find a set C_i containing a subsequence of x_n:

$$\{x_{n_k}\} \subset C_i$$

But since C_i was assumed to be closed, this shows that we have:

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

With this done, we can add to the right all the other closed sets, and we get:

$$\lim_{n\to\infty} x_n \in \bigcup_i C_i$$

Thus, we are led to the conclusions in the statement. □

As an important comment, (3,4) above do not hold when removing the finiteness assumption. Indeed, in what regards (3), the simplest counterexample here is:

$$\bigcap_{n\in\mathbb{N}} \left(-\frac{1}{n}\,,\,\frac{1}{n} \right) = \{0\}$$

As for (4), here the simplest counterexample is as follows:

$$\bigcup_{n\in\mathbb{N}} \left[0\,,\,1-\frac{1}{n} \right] = [0,1)$$

We will be back in a moment to all this, with a precise characterization of all the open and closed sets. In the meantime, let us develop some general theory. In order to get truly started, with our study, we first have the following theoretical result:

THEOREM 5.15. *A set $O \subset \mathbb{R}$ is open precisely when its complement $C \subset \mathbb{R}$ is closed, and vice versa.*

PROOF. It is enough to prove the first assertion, since the "vice versa" part will follow from it, by taking complements. But this can be done as follows:

" $\implies$ " Assume that $O \subset \mathbb{R}$ is open, and let $C = \mathbb{R} - O$. In order to prove that C is closed, assume that $\{x_n\}_{n\in\mathbb{N}} \subset C$ converges to $x \in \mathbb{R}$. We must prove that $x \in C$, and we will do this by contradiction. So, assume $x \notin C$. Thus $x \in O$, and since O is open we

can find a small interval $(x-\varepsilon, x+\varepsilon) \subset O$. But since $x_n \to x$ this shows that $x_n \in O$ for n big enough, which contradicts $x_n \in C$ for all n, and we are done.

"$\Longleftarrow$" Assume that $C \subset \mathbb{R}$ is open, and let $O = \mathbb{R} - C$. In order to prove that O is open, let $x \in O$, and consider the intervals $(x-1/n, x+1/n)$, with $n \in \mathbb{N}$. If one of these intervals lies in O, we are done. Otherwise, this would mean that for any $n \in \mathbb{N}$ we have at least one point $x_n \in (x-1/n, x+1/n)$ satisfying $x_n \notin O$, and so $x_n \in C$. But since C is closed and $x_n \to x$, we get $x \in C$, and so $x \notin O$, contradiction, and we are done. □

As a basic illustration for the above result, a disjoint union of two infinite open intervals is open, due to the fact that its complement is closed:

$$(-\infty, a) \cup (b, \infty) = \mathbb{R} - [a, b]$$

As another basic illustration, a disjoint union of two infinite closed intervals is open, due to the fact that its complement is open:

$$(-\infty, a] \cup [b, \infty) = \mathbb{R} - (a, b)$$

Getting now to functions, we have the following key result about them:

THEOREM 5.16. *A function is continuous precisely when $f^{-1}(O)$ is open, for any O open. Equivalently, $f^{-1}(C)$ must be closed, for any C closed.*

PROOF. This is something coming from definitions, the idea being as follows:

(1) The first assertion follows from the ε, δ definition of continuity, namely:

$$\forall x \in X, \forall \varepsilon > 0, \exists \delta > 0, |x-y| < \delta \implies |f(x) - f(y)| < \varepsilon$$

Indeed, if a function f satisfies this condition, it is then clear that if a set O is open, then the set $f^{-1}(O)$ is open too. Moreover, the converse clearly holds too.

(2) As for the second assertion, this can be proved directly, by using the $f(x_n) \to f(x)$ definition of continuity, or can be deduced from what we already know about the open sets, by taking complements. Indeed, assuming that f is continuous, we have:

$$\begin{aligned} C = \text{closed} &\implies C^c = \text{open} \\ &\implies f^{-1}(C^c) = \text{open} \\ &\implies f^{-1}(C) = \text{closed} \end{aligned}$$

As for the converse, this is again clear from (1), by taking complements. □

As a test for the above criterion, let us reprove the fact, that we know well from chapter 2, that if f, g are continuous, so is $f \circ g$. But this is clear, coming from:

$$(f \circ g)^{-1}(O) = g^{-1}(f^{-1}(O))$$

In short, not bad, because our proof using open sets is as simple as the simplest proof, namely the one using $f(x_n) \to f(x)$, and is simpler than the other proof that we know, namely the one with ε, δ. Let us record this as a philosophical conclusion, as follows:

CONCLUSION 5.17. *The open and closed sets are no joke, because with them, we can see right away that a composition of continuous functions is continuous.*

Summarizing, our open and closed set technology is quite interesting, worth developing. As a question now that we still have to solve, we would like to know what the open and closed sets really are. And fortunately, this question can be answered, as follows:

THEOREM 5.18. *The open and closed sets are as follows:*

(1) *The open sets are the disjoint unions of open intervals.*
(2) *The closed sets are the complements of these unions.*

PROOF. We have two assertions to be proved, the idea being as follows:

(1) We know that the open intervals are those of type (a, b) with $a < b$, with the values $a, b = \pm\infty$ allowed, and by Theorem 5.14 a union of such intervals is open.

(2) Conversely, given $O \subset \mathbb{R}$ open, we can cover each point $x \in O$ with an open interval $I_x \subset O$, and we have $O = \cup_x I_x$, so O is a union of open intervals.

(3) In order to finish the proof of the first assertion, it remains to prove that the union $O = \cup_x I_x$ in (2) can be taken to be disjoint. For this purpose, our first observation is that, by approximating points $x \in O$ by rationals $y \in \mathbb{Q} \cap O$, we can make our union to be countable. But once our union is countable, we can start merging intervals, whenever they meet, and we are left in the end with a countable, disjoint union, as desired.

(4) Finally, the second assertion comes from Theorem 5.15. □

Moving towards more concrete things, and applications, let us formulate the following key definition, which is actually one of the most important definitions in analysis:

DEFINITION 5.19. *The compact and connected sets are defined as follows:*

(1) *A subset $K \subset \mathbb{R}$ is called compact when any open cover of it, $K \subset \cup_x O_x$, with all $O_x \subset \mathbb{R}$ being open sets, has a finite subcover.*
(2) *A subset $E \subset \mathbb{R}$ is called connected when it cannot be broken into two parts, in the sense that we cannot have $E \subset A \sqcup B$, with $A, B \neq \emptyset$ open.*

As basic examples for (1), the closed bounded intervals $[a, b]$ are compact, and so are the finite unions of such intervals, as shown by the following result:

THEOREM 5.20. *The closed bounded intervals on the real line,*

$$[a, b] \quad , \quad -\infty < a < b < \infty$$

are compact, and so are the finite unions of such intervals.

PROOF. This is something very standard, the idea being as follows:

(1) We proceed by contradiction. So, assume that $[a, b]$ has no finite subcover, and let us cut this interval in half. Then one of the halves must have no finite subcover either, and we can repeat the procedure, by cutting this smaller interval in half. And so on.

(2) But this leads to a contradiction, because the limiting point $x \in [a,b]$ that we obtain in this way, as the intersection of these smaller and smaller intervals, must be covered by something, and so one of these small intervals leading to it must be covered too, contradiction. Thus, we have proved that our cover has a finite subcover. □

Getting now to connected sets, the basic examples here are the various types of intervals, as follows, with the endpoints being allowed to be finite, or not:

$$(a,b) \quad , \quad (a,b] \quad , \quad [a,b) \quad , \quad [a,b]$$

Thinking a bit at what we have above, as examples for both the compact and the connected sets, it looks impossible to come up with more examples. In fact, we have:

THEOREM 5.21. *The compact and connected sets are as follows:*

(1) *The compact sets are those which are closed and bounded.*
(2) *The connected sets are the various types of intervals.*

PROOF. This is something quite intuitive, the idea being as follows:

(1) Compact implies closed, because assuming the contrary, given $\{x_n\} \subset K$ converging to $x \notin K$, we have the following open cover, having no finite subcover:

$$K \subset \bigcup_{n\in\mathbb{N}} \left[x - \frac{1}{n}, x + \frac{1}{n}\right]^c$$

(2) Similarly, compact implies bounded, and this due to the following open cover:

$$K \subset \bigcup_{n\in\mathbb{N}} (-n, n)$$

(3) As for the converse, stating that closed and bounded implies compact, assume that we have a bounded set, $K \subset [a,b]$, which is closed, and consider an open cover of it:

$$K \subset \bigcup_x O_x$$

By adding to this cover the open set K^c, we obtain an open cover as follows:

$$[a,b] \subset \left(\bigcup_x O_x\right) \cup K^c$$

Now by Theorem 5.20, this cover must have a finite subcover, and by removing from this latter cover the set K^c, we obtain in this way a finite cover of K, as desired.

(4) Finally, regarding the second assertion, this is something quite obvious, because $E \subset \mathbb{R}$ being connected means $a, b \in E \implies [a,b] \subset E$, and this gives the result. □

Now with this discussed, let us go back to the continuous functions. We have here the following result, extending what we already know, from Theorem 5.16:

THEOREM 5.22. *Assuming that f is continuous:*

(1) *If O is open, then $f^{-1}(O)$ is open.*
(2) *If C is closed, then $f^{-1}(C)$ is closed.*
(3) *If K is compact, then $f(K)$ is compact.*
(4) *If E is connected, then $f(E)$ is connected.*

PROOF. This is something elementary, the idea being as follows:

(1,2) These are things from Theorem 5.16, included here for convenience.

(3) Given an open cover $f(K) \subset \cup_x O_x$ we have $K \subset \cup_x f^{-1}(O_x)$, open cover too, and if $K \subset \cup_y f^{-1}(O_y)$ is a finite subcover of this, then $f(K) \subset \cup_y O_y$, as desired.

(4) This is clear too, because assuming $f(E) \subset A \sqcup B$ with $A, B \neq \emptyset$ open, we have $E \subset f^{-1}(A) \sqcup f^{-1}(B)$ with $f^{-1}(A), f^{-1}(B) \neq \emptyset$ open, contradiction. □

Very nice all the above, good mathematical learning that was, but you might perhaps ask at this point, what can all this be good for. In answer, we have:

THEOREM 5.23 (Intermediate value property). *Given a continuous function*

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

its image is a closed bounded interval, $Im(f) = [c, d]$, and this for trivial reasons.

PROOF. This is something that we certainly know, since chapter 2, but the point is that, with our present technology, everything is trivial. Indeed, the result says that:

$$X = \text{compact, connected} \implies f(X) = \text{compact, connected}$$

But this follows from Theorem 5.22 (3) and (4), which themselves were in fact trivial results, so based on this, we can now declare the whole thing trivial. Good. □

There are of course many other applications of our technology. More later.

5c. Uniform continuity

Getting back now to the basics, and going ahead with some more theory, some functions are "obviously" continuous, with a basic result here being as follows:

THEOREM 5.24. *If a function $f : X \to \mathbb{R}$ has the Lipschitz property*

$$|f(x) - f(y)| \leq K|x - y|$$

for some $K > 0$, then it is continuous.

PROOF. This is clear from our ε, δ definition of continuity, namely:

$$\forall x \in X, \forall \varepsilon > 0, \exists \delta > 0, |x - y| < \delta \implies |f(x) - f(y)| < \varepsilon$$

Indeed, by setting $\delta = \varepsilon/K$, this continuity condition is trivially satisfied. □

There are many interesting examples of Lipschitz functions, both concrete and abstract. In what regards the usual functions, the situation is as follows:

THEOREM 5.25. *The following functions are Lipschitz, on any compact interval $[a,b]$ belonging to their domain, with their best constants K there being as indicated:*

(1) x, *with* $K=1$.
(2) x^2, *with* $K=2\max(|a|,|b|)$.
(3) x^n, *with* $K=n\cdot\max(|a|^{n-1},|b|^{n-1})$.
(4) $P(x)$, *polynomial.*
(5) x^{-1}, *with* $K=\max(1/a^2,1/b^2)$.
(6) x^{-n}, *with* $K=n\cdot\max(1/|a|^{n+1},1/|b|^{n+1})$.
(7) $P(x)/Q(x)$, *rational function.*
(8) $\sin x$, *with* $K\leq 1$, *computable.*
(9) $\cos x$, *with* $K\leq 1$, *computable.*
(10) $\tan x$, *with* K *computable.*
(11) $\exp x$, *with* $K=e^b$.
(12) $\log x$, *with* $K=1/a$.

As for the infinite intervals, here the best constants $K\in[0,\infty]$, with $K<\infty$ corresponding to Lipschitz, and $K=\infty$ corresponding to non-Lipschitz, can be computed too.

PROOF. This might sound a bit crazy, because we perfectly know that proving that a function is continuous is not an easy business, and what we say above, with that Lipschitz constants computed, is obviously sharper than what we usually do. In answer, I have my own tricks for computing Lipschitz constants, coming from the following formula:

$$K=\sup_{x\in[a,b]}|f'(x)|$$

To be more precise, here $f'(x)$ stands for the slope, or derivative, at f at the point x, which is something that can be systematically computed, and the above formula is something quite intuitive, because the Lipschitz quotient $(f(x)-f(y))/(x-y)$ from Definition 5.24 should normally correspond to such a slope, at a certain $z\in[x,y]$:

$$\frac{f(x)-f(y)}{x-y}=f'(z)$$

So, these are my tricks, and more on this later, in Part III. In the meantime, here is the proof of the result, with some computations left as exercises, and with my apologies in advance if some of these exercises are in fact a bit too difficult, sorry for this:

(1) The function $f(x)=x$ is certainly Lipschitz everywhere, with constant $K=1$.

(2) In what regards now $f(x)=x^2$, we have here the following equivalence:

$$|x^2-y^2|\leq K|x-y|\iff|x+y|\leq K$$

Thus $f(x) = x^2$ is Lipschitz on any interval $[a, b]$, with constant as follows:

$$K = 2\max(|a|, |b|)$$

Observe also that this shows that x^2 is not Lipschitz on unbounded intervals.

(3) For $f(x) = x^n$ with $n \in \mathbb{N}$ arbitrary, this follows in a similar way, by using:

$$x^n - y^n = (x - y)(x^{n-1} + x^{n-2}y + \ldots + xy^{n-2} + y^{n-1})$$

As an exercise for you, however, try proving this for any $f(x) = x^n$ with $n > 0$.

(4) For polynomials the Lipschitz property on compact intervals comes from (3), by linearity, but in what regards the explicit computation of K, this is something quite tricky. Exercise for you to explore a bit all this, say for polynomials of small degree.

(5) In what regards the function $f(x) = x^{-1}$, we have here the following equivalence:

$$\left|\frac{1}{x} - \frac{1}{y}\right| \leq K|x - y| \iff |xy| \geq \frac{1}{K}$$

Thus x^{-1} is Lipschitz on any interval $[a, b] \subset \mathbb{R} - \{0\}$, with constant as follows:

$$K = \max(1/a^2, 1/b^2)$$

(6) For $f(x) = x^{-n}$ with $n \in \mathbb{N}$ we can use the same trick as in (3), namely:

$$\frac{1}{x^n} - \frac{1}{y^n} = \frac{y^{n-1} + y^{n-2}x + \ldots + yx^{n-2} + x^{n-1}}{x^n y^n}$$

As an exercise for you, however, try proving this for any $f(x) = x^n$ with $n < 0$.

(7) For rational functions the Lipschitz property on compact intervals comes from (4,6) but in what regards the explicit computation of K, this is certainly quite tricky. Exercise for you to explore a bit all this, say for $f = P/Q$ with P, Q of small degree.

(8) In order to deal now with the sine, we can use the following estimate:

$$\begin{aligned} |\sin(x + t) - \sin(x - t)| &= |\sin x \cos t + \cos x \sin t - \sin x \cos t + \cos x \sin t| \\ &= 2|\cos x \sin t| \\ &\leq 2t|\cos x| \end{aligned}$$

We conclude that the sine is indeed Lipschitz, with constant as follows:

$$K \leq \sup_{x \in [a,b]} |\cos x|$$

And then, by arguing that we have $\sin t \simeq t$ for t small, we have in fact equality.

(9) Regarding the cosine, we can get the result from that for the sine, by using:

$$\cos x = \sin(x + \pi/2)$$

Alternatively, we can do a direct computation, as for the sine, as follows:

$$|\cos(x+t)-\cos(x-t)| = 2|\sin x \sin t| \leq 2t|\sin x|$$

We conclude that the cosine is indeed Lipschitz, with constant as follows:

$$K \leq \sup_{x\in[a,b]} |\sin x|$$

And then, by arguing that we have $\sin t \simeq t$ for t small, we have in fact equality.

(10) For the tangent on $[a,b] \subset [-\pi/2, \pi/2]$, we can use the following estimate:

$$|\tan x - \tan y| = \left|\frac{\sin(x-y)}{\cos x \cos y}\right| \leq \frac{|x-y|}{|\cos x \cos y|}$$

We conclude that the tangent is indeed Lipschitz, with constant as follows:

$$K \leq \sup_{x\in[a,b]} \frac{1}{\cos^2 x}$$

And then, by arguing that $\sin(x-y) \simeq x-y$ for $x-y$ small, we have in fact equality.

(11) For exp we have the following estimate, with $t > 0$, coming from $e^{-t} \geq 1-t$:

$$e^x - e^{x-t} = e^x(1-e^{-t}) \leq e^x t$$

Thus we have the Lipschitz property with $K \leq e^b$, and by arguing that $e^{-t} \geq 1-t$ becomes $e^{-t} \simeq 1-t$ for t small, we have in fact $K = e^b$, as claimed.

(12) For log we have the following estimate, with $t > 0$, coming from $\log(1+y) \leq y$:

$$\log(x+t) - \log x = \log\left(1+\frac{t}{x}\right) \leq \frac{t}{x}$$

Thus we have the Lipschitz property with $K \leq 1/a$, and by arguing that $\log(1+y) \leq y$ becomes $\log(1+y) \simeq y$ for y small, we have in fact $K = 1/a$, as claimed. □

Moving on, the story is not over with the Lipschitz property, which is stronger than continuity, because we have as well a third property, in between, as follows:

THEOREM 5.26. *Consider the following properties, regarding $f : X \to \mathbb{R}$ with $X \subset \mathbb{R}$:*

(1) *f has the following property, with $K > 0$, which is the Lipschitz property:*

$$|f(x) - f(y)| \leq K|x-y|$$

(2) *f is uniformly continuous, in the sense that the following happens:*

$$\forall \varepsilon > 0, \exists \delta > 0, |x-y| < \delta \implies |f(x) - f(y)| < \varepsilon$$

(3) *f is continuous in the usual sense, namely:*

$$\forall x \in X, \forall \varepsilon > 0, \exists \delta > 0, |x-y| < \delta \implies |f(x) - f(y)| < \varepsilon$$

We have then (1) $\implies$ (2) $\implies$ (3). *Also, the converse implications do not hold.*

PROOF. This is something quite self-explanatory, and that we partly know:

(1) $\implies$ (2) This is clear, coming by talking $\delta = \varepsilon/K$.

(2) $\implies$ (3) This is something which is plainly trivial.

(3) $\not\implies$ (2) The simplest counterexample here is $f(x) = x^2$. Indeed, this function is continuous, and its uniform continuity property, applied with $\varepsilon = 1$, would lead to the existence of $\delta > 0$ such that the following happens, which is wrong:

$$|x - y| < \delta \implies |x^2 - y^2| < 1$$

(2) $\not\implies$ (1) The simplest counterexample here is $f(x) = \sqrt{x}$. Indeed, observe first that we have the following estimate, valid for any $x > y > 0$:

$$(\sqrt{x} - \sqrt{y})^2 \leq (\sqrt{x} - \sqrt{y})(\sqrt{x} + \sqrt{y}) = x - y$$

Thus our function is indeed uniformly continuous, and this because we have:

$$|x - y| < \varepsilon^2 \implies |\sqrt{x} - \sqrt{y}| < \varepsilon$$

In what regards now the Lipschitz property, observe that we have:

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

Now since this can be arbitrarily big, when x, y are small, Lipschitz fails indeed. □

The interest in the notion of uniform continuity, which remains something a bit abstract, comes from the following remarkable result, due to Heine and Cantor:

THEOREM 5.27. *Any continuous function defined on a compact set*

$$f : X \to \mathbb{R}$$

is automatically uniformly continuous.

PROOF. This is something quite standard, the idea being as follows:

(1) Given $\varepsilon > 0$, for any $x \in X$ we know that we have a $\delta_x > 0$ such that:

$$|x - y| < \delta_x \implies |f(x) - f(y)| < \frac{\varepsilon}{2}$$

So, consider the following open intervals, centered at the various points $x \in X$:

$$I_x = \left(x - \frac{\delta_x}{2}\,,\, x + \frac{\delta_x}{2}\right)$$

These intervals then obviously cover X, in the sense that we have:

$$X \subset \bigcup_{x \in X} I_x$$

(2) But, we know that X is compact. So, consider a finite subcover of this cover:

$$X \subset \bigcup_i I_{x_i}$$

With this done, consider as well the following number, which is strictly positive:

$$\delta = \min_i \frac{\delta_{x_i}}{2}$$

Now assume $|x - y| < \delta$, and pick i such that $x \in I_{x_i}$. By the triangle inequality we have $|x_i - y| < \delta_{x_i}$, which shows that we have $y \in I_{x_i}$ as well. But by applying now f, this gives as desired $|f(x) - f(y)| < \varepsilon$, again via the triangle inequality. □

The Heine-Cantor theorem is something quite useful, in practice. More later.

5d. Complex functions

We would like to end this chapter with a quick discussion of the complex functions, see how the above technology can be extended to them. Normally this is the business of complex analysis, which is something more advanced, coming after real analysis, but in practice, we have seen in chapter 4 that our real variables $x \in \mathbb{R}$ tend to evolve into complex variables, $x \in \mathbb{C}$, and this notably due to the Euler formula, namely:

$$e^{it} = \cos t + i \sin t$$

In short, there is no advanced real analysis without complex analysis, at least a little bit, so our problem mentioned above makes sense, namely see how our technology, developed throughout this chapter, can be adapted to the complex functions.

We will be quite quick. In order to do some complex analysis, let us start with:

DEFINITION 5.28. *The distance in the complex plane is the usual distance, namely:*

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

With this, we can talk about convergence, by saying that $x_n \to x$ when $d(x_n, x) \to 0$.

Observe that in real coordinates, the distance formula is quite complicated, namely:

$$\begin{aligned} d(a + ib, c + id) &= |(a + ib) - (c + id)| \\ &= |(a - c) + i(b - d)| \\ &= \sqrt{(a - c)^2 + (b - d)^2} \end{aligned}$$

However, for most computations, we will not need this, and we can get away with various tricks regarding complex numbers. Talking complex functions now, we have:

DEFINITION 5.29. *A function $f : X \to \mathbb{C}$, with $X \subset \mathbb{C}$, is continuous at $x \in X$ when:*

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

Equivalently, the following epsilon-delta condition must be satisfied:

$$\forall \varepsilon > 0 \ \exists \delta > 0 \ , \ |x - y| < \delta \implies |f(x) - f(y)| < \varepsilon$$

We say that f is continuous when it is continuous at all points $x \in X$.

Observe that, since $x_n \to x$ in the complex sense means that $(a_n, b_n) \to (a, b)$ in the usual, real plane sense, a function $f : \mathbb{C} \to \mathbb{C}$ is continuous precisely when it is continuous when regarded as real function, $f : \mathbb{R}^2 \to \mathbb{R}^2$. Which is something good to know.

Moving on, in what regards the continuity basics, we have:

THEOREM 5.30. *The following complex functions are continuous:*

(1) *λf, $f + g$, fg, $f \circ g$, provided that f, g are continuous.*
(2) *f/g too, on its domain, meaning outside the zeroes of g.*
(3) *The polynomials $P \in \mathbb{C}[X]$, viewed as functions $P : \mathbb{C} \to \mathbb{C}$.*
(4) *The rational functions, $f = P/Q$ with $P, Q \in \mathbb{C}[X]$, outside their poles.*
(5) *The complex exponential function,* $\exp : \mathbb{C} \to \mathbb{C}$.

PROOF. Here (1-4) are things that we know well for the real functions, and the complex extension is straightforward. As for (5), we know this from chapter 4. □

Let us point out now the fact that, contrary to what the above might suggest, everything does not always extend trivally from real to complex. For instance, we have:

PROPOSITION 5.31. *We have the following formula, valid for any $|x| < 1$,*

$$\frac{1}{1 - x} = 1 + x + x^2 + \ldots$$

but, for $x \in \mathbb{C} - \mathbb{R}$, the geometric meaning of this formula is quite unclear.

PROOF. Here the formula in the statement holds indeed, by multiplying and cancelling terms, and with the convergence being justified by the following estimate:

$$\left| \sum_{n=0}^{\infty} x^n \right| \leq \sum_{n=0}^{\infty} |x^n| = \sum_{n=0}^{\infty} |x|^n = \frac{1}{1 - |x|}$$

As for the last assertion, this is something quite informal, the idea being as follows:

(1) To start with, for the simplest possible value of our parameter, $x = 1/2$, our formula is clear, by cutting the interval $[0, 2]$ into half, and so on:

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = 2$$

(2) More generally, for $x \in (-1,1)$ the meaning of the formula in the statement is something quite clear and intuitive, geometrically speaking, using a similar argument.

(3) However, when x is complex, and not real, we are led into a kind of mysterious spiral there, and the only case where the formula is "obvious", geometrically speaking, is that when $x = rw$, with $r \in [0,1)$, and with w being a root of unity.

(4) To be more precise here, assuming $w^N = 1$, we have the following formula:

$$\begin{aligned} 1 + rw + r^2w^2 + \ldots &= (1 + rw + \ldots + r^{N-1}w^{N-1}) \\ &+ (r^N + r^{N+1}w \ldots + r^{2N-1}w^{N-1}) \\ &+ (r^{2N} + r^{2N+1}w \ldots + r^{3N-1}w^{N-1}) \\ &+ \ldots \end{aligned}$$

(5) Thus, by grouping the terms with the same argument, our infinite sum is:

$$\begin{aligned} 1 + rw + r^2w^2 + \ldots &= (1 + r^N + r^{2N} + \ldots) \\ &+ (r + r^{N+1} + r^{2N+1} + \ldots)w \\ &+ \ldots \\ &+ (r^{N-1} + r^{2N-1} + r^{3N-1} + \ldots)w^{N-1} \end{aligned}$$

(6) But the sums of each ray can be computed with the real formula for geometric series, that we know and understand well, and with an extra bit of algebra, we get:

$$\begin{aligned} 1 + rw + r^2w^2 + \ldots &= \frac{1}{1-r^N} + \frac{rw}{1-r^N} + \ldots + \frac{r^{N-1}w^{N-1}}{1-r^N} \\ &= \frac{1}{1-r^N} \cdot \frac{1-(rw)^N}{1-rw} \\ &= \frac{1}{1-rw} \end{aligned}$$

(7) Summarizing, as claimed above, the geometric series formula can be understood, in a purely geometric way, for variables of type $x = rw$, with $r \in [0,1)$, and with w being a root of unity. In general, however, this formula tells us that the numbers on a certain infinite spiral sum up to a certain number, which remains something quite mysterious. □

Getting now to less mysterious mathematics, we have the quite straightforward question of understanding what our topological methods become, in the complex setting. We will be quite brief here. Let us start with the following basic definition:

DEFINITION 5.32. *The open, closed, compact, connected sets in $\mathbb{C}$ are as follows:*

(1) *Open means that there is a small disk around each point.*
(2) *Closed means that our set is closed under taking limits.*
(3) *Compact means that any open cover has a finite subcover.*
(4) *Connected means that it cannot be broken into two parts.*

In what regards the open and closed sets, the basic theory is exactly as in the real case, save for the results involving intervals, with the summary here being as follows:

THEOREM 5.33. *A subset $O \subset \mathbb{C}$ is open precisely when its complement $C \subset \mathbb{C}$ is closed, and vice versa. Also, the following happen:*

(1) *Union of open sets is open.*
(2) *Intersection of closed sets is closed.*
(3) *Finite intersection of open sets is open.*
(4) *Finite union of closed sets is closed.*

Also, the open sets $O \subset \mathbb{C}$ are exactly the unions of open disks, or open rectangles if you prefer, but these unions can no longer be taken disjoint.

PROOF. Here everything is very standard, exactly as in the real case, with the only subtletly being the one at the end. Indeed, you cannot decompose an open disk as a disjoint union of open rectangles, or vice versa, and the problem comes from this. □

Regarding now the compact and connected sets, again the theory is quite similar to the one in the real case, with some twists regarding the connectedness, as follows:

THEOREM 5.34. *The compact and connected sets in $\mathbb{C}$ are as follows:*

(1) *The compact sets are those which are closed and bounded.*
(2) *We have convex $\implies$ path connected $\implies$ connected.*

PROOF. Here (1) is something very standard, exactly as in the real case, by replacing the intervals there by cubes, in the obvious way. As for (2), which is something self-explanatory, this is what can be said, as generalities, about the connected sets in $\mathbb{C}$. And exercise for you to learn more about this, and about the shape of connected sets $E \subset \mathbb{C}$, in general. Many interesting things, for instance regarding holes, can be said here. □

In what regards now the complex functions, we have the following result:

THEOREM 5.35. *Assuming that a function $f : \mathbb{C} \to \mathbb{C}$ is continuous,*

(1) *If O is open, then $f^{-1}(O)$ is open.*
(2) *If C is closed, then $f^{-1}(C)$ is closed.*
(3) *If K is compact, then $f(K)$ is compact.*
(4) *If E is connected, then $f(E)$ is connected.*

and with $(3, 4)$ standing as an intermediate value theorem, for the complex functions.

PROOF. This comes again as a straightforward extension of our previous results regarding the real functions, and with the converses of (1,2) being of course true too. □

Very nice all this, so the theory from the real case basically extends well. Getting now to more concrete things, remember the following challenge, from chapter 4:

$$(-1)^2 = e^{2\log(-1)} = 1$$

So, can we do this, with a suitable definition for the complex logarithm function? For this purpose, let us first study some more the complex exponential function exp, that we would like to invert, over a suitable domain. By using $e^{x+y} = e^x e^y$ we obtain $e^x \neq 0$ for any $x \in \mathbb{C}$, so the complex exponential function is as follows:

$$\exp : \mathbb{C} \to \mathbb{C} - \{0\}$$

Now since we have $e^{x+iy} = e^x e^{iy}$ for $x, y \in \mathbb{R}$, with e^x being surjective onto $(0, \infty)$, and with e^{iy} being surjective onto the unit circle $\mathbb{T}$, we deduce that $\exp : \mathbb{C} \to \mathbb{C} - \{0\}$ is surjective. Also, again by using $e^{x+iy} = e^x e^{iy}$, we deduce that we have:

$$e^x = e^y \iff x - y \in 2\pi i\mathbb{Z}$$

With these ingredients in hand, we can now talk about log, as follows:

THEOREM 5.36. *Given an horizontal strip in the complex plane, having width* 2π*,*

$$S = \left\{ x + iy \middle| x \in \mathbb{R}, y \in [a, a + 2\pi) \right\}$$

$\exp : S \to \mathbb{C} - \{0\}$ *is bijective, so we can define* log *as being the inverse of this map,*

$$\log = \exp^{-1} : \mathbb{C} - \{0\} \to S$$

and with this, we have indeed $(-1)^2 = e^{2\log(-1)} = 1$*, as desired.*

PROOF. This is indeed something self-explanatory, based on the above. Let us also mention that in practice, the best is to choose for instance $a = 0$, or $a = -\pi$, as to have the whole real line included in our strip, $\mathbb{R} \subset S$. In this case on $\mathbb{R}_+$ we recover the usual logarithm, while on $\mathbb{R}_-$ we obtain complex values, as for instance $\log(-1) = \pi i$ in the case $a = 0$, or $\log(-1) = -\pi i$ in the case $a = -\pi$, coming from $e^{\pi i} = -1$. □

Quite exciting all this. More complex analysis, later in this book.

5e. Exercises

For understanding continuity, nothing better than computing Lipschitz constants:

EXERCISE 5.37. *Compute the Lipschitz constant for* x^n*,* $n \in \mathbb{N}$*.*

EXERCISE 5.38. *Compute the Lipschitz constant for* x^{-n}*,* $n \in \mathbb{N}$*.*

EXERCISE 5.39. *Compute the Lipschitz constant for* x^p*,* $p \in \mathbb{R}$*.*

EXERCISE 5.40. *Compute Lipschitz constants for polynomials of small degree.*

EXERCISE 5.41. *Explore the Lipschitz constants, for the rational functions.*

EXERCISE 5.42. *Clarify what we said, about the Lipschitz constants for* $\sin x$*,* $\cos x$*.*

EXERCISE 5.43. *Clarify what we said, about the Lipschitz constant for* $\tan x$*.*

EXERCISE 5.44. *Compute other Lipschitz constants, for functions of your choice.*

As bonus exercise, explore a bit more the complex logarithm constructed above.

CHAPTER 6

Sequences, limits

6a. Fixed points

We discuss in this chapter a number of more specialized questions, in relation with continuity. Let us start with something very basic, and useful in practice, namely:

THEOREM 6.1. *Given a continuous function $f : X \to X$, with $X \subset \mathbb{R}$, if we set*

$$x_0 = x \quad , \quad x_{n+1} = f(x_n)$$

depending on a given $x \in X$, assuming $x_n \to z$ we have $f(z) = z$.

PROOF. This comes from the following computation, using the continuity of f:

$$\begin{aligned} f(z) &= f\left(\lim_{n\to\infty} x_n\right) \\ &= \lim_{n\to\infty} f(x_n) \\ &= \lim_{n\to\infty} x_{n+1} \\ &= z \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

In practice, the above result is a bit too general, for being applied as such, and we have the following version of it, which is its truly useful version:

THEOREM 6.2. *Given a continuous function $f : [a, b] \to [a, b]$ satisfying*

$$f(y) \leq y \quad , \quad \forall y$$

the following sequence, depending on a given $x \in X$, is monotone and converges,

$$x_0 = x \quad , \quad x_{n+1} = f(x_n)$$

and with $x_n \to z$ we have $f(z) = z$. The same happens when assuming $f(y) \geq y$, $\forall y$.

PROOF. This is indeed something self-explanatory, based on Theorem 6.1:

(1) Assuming $f(y) \leq y$, $\forall y$, the sequence $\{x_n\}$ decreases, which gives the result.

(2) Assuming $f(y) \geq y$, $\forall y$, the sequence $\{x_n\}$ increases, which gives the result. □

As a basic application of this method, to the extraction of square roots, we have:

THEOREM 6.3. *We can extract the square root of $a > 0$ by iterating the function*

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

with this being known as the Babylonian method for extracting square roots.

PROOF. This is indeed self-explanatory, based on the above, and on:

$$\frac{x + a/x}{2} = x \iff a/x = x \iff x = \sqrt{a}$$

In practice, however, many things can be said here, as follows:

(1) We can assume $a > 1$ for simplifying, based on the following formula, for $a < 1$:

$$\sqrt{a} = \frac{1}{\sqrt{a^{-1}}}$$

(2) Now let us try to find a suitable interval $[c, d]$ such that $f([c, d]) \subset [c, d]$, as for Theorem 6.1, and hopefully Theorem 6.2 too, to properly apply. Since we are looking for the square root of $a > 1$, satisfying $1 < \sqrt{a} < a$, a reasonable candidate here is:

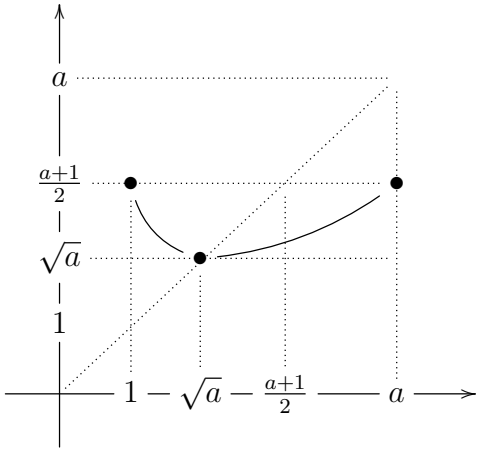
$$[c, d] = [1, a]$$

And, good luck, this works indeed, due to the following two estimates:

$$x > 0 \implies \frac{x + a/x}{2} \geq \sqrt{x \cdot a/x} = 1$$

$$x \in [1, a] \implies \frac{x + a/x}{2} \leq \frac{a + a/1}{2} = a$$

(3) Next, let us examine if the conditions $f(y) \leq y$ or $f(y) \geq y$ from Theorem 6.2 are satisfied. And here, this is not exactly the case, the graph of f being as follows:



To be more precise, the graph of f passes indeed through the 3 points which are indicated, and in what regards the monotony properties, these come from:

$$\frac{x+a/x}{2}\geq x\iff a/x\geq x\iff x\leq\sqrt{a}$$

(4) So, what to do? In what regards $[1,\sqrt{a}]$, better forget about it, due to:

$$x\in[1,\sqrt{a}]\implies f(x)\in[\sqrt{a},a]$$

As for the other interval, namely $[\sqrt{a},a]$, things just fine here, due to:

$$x\in[\sqrt{a},a]\implies f(x)\in[\sqrt{a},a]$$

(5) As a conclusion to this, Theorem 6.2 applies as such on $[\sqrt{a},a]$. However, since it might look a bit unnatural to use $[\sqrt{a},a]$ for our computations, with $\sqrt{a}$ being precisely the number to be computed, the best is to declare the following:

"The iteration in Theorem 6.2 applies on $[1,a]$, after the first step."

(6) With this discussed, let us work out now some numerics, say for $a=2$. As a first observation, when trying to compute $\sqrt{2}$ with bare hands, that is, with computing the decimals one by one, by hand, the algorithm is not very inviting, as follows:

$$1.4^2<2<1.5^2\implies\sqrt{2}=1.4\ldots$$
$$1.41^2<2<1.42^2\implies\sqrt{2}=1.41\ldots$$
$$1.414^2<2<1.415^2\implies\sqrt{2}=1.414\ldots$$
$$1.4142^2<2<1.4143^2\implies\sqrt{2}=1.4142\ldots$$

(7) In order to comment on what happens, with our method, let us first go back to the general case, $a>1$, and see how the error term evolves. So, let us write:

$$x=\sqrt{a}+c$$

The error here is $c>0$, and when applying f, the error evolves according to:

$$\begin{aligned}f(x)&=\frac{\sqrt{a}+c+\frac{a}{\sqrt{a}+c}}{2}\\&=\frac{\sqrt{a}+c}{2}+\frac{\sqrt{a}}{2}\cdot\frac{\sqrt{a}}{\sqrt{a}+c}\\&=\frac{\sqrt{a}+c}{2}+\frac{\sqrt{a}}{2}\left(1-\frac{c}{\sqrt{a}+c}\right)\\&=\sqrt{a}+\frac{c}{2}\left(1-\frac{\sqrt{a}}{\sqrt{a}+c}\right)\end{aligned}$$

Thus, the convergence is exponential, which is nice, and in what regards the numerics at $a=2$, and their comparison with the old method in (6), good exercise for you. □

Observe that Theorem 6.3 used a trick, namely converting our $a \to \sqrt{a}$ problem into a fixed point problem, that we can solve using Theorem 6.2. Many other such tricks are possible, for instance with the computation of zeroes being possible via:

METHOD 6.4. *We can compute zeroes of functions by using*

$$f(x) = 0 \iff f(x) + x = x$$

and iterating $\varphi(x) = f(x) + x$, over suitable intervals.

Obviously, this is something quite vast. We will leave some exploration here, for various interesting functions, say polynomials of small degree, as an exercise.

At a more advanced level now, we have seen in the proof of Theorem 6.3 that upgrading our methods, from Theorem 6.1 to Theorem 6.2, takes some skill, in order to find the correct intervals $[a, b]$ where Theorem 6.2 applies, which can be something non-trivial.

For certain abstract questions, such things are not really possible, and as an alternative to Theorem 6.2, we have the following useful result, again based on Theorem 6.1:

THEOREM 6.5. *A continuous function $f : X \to X$, with X closed, satisfying*

$$|f(x) - f(y)| \leq K|x - y|$$

with $K < 1$, has a unique fixed point $z \in X$, appearing by iterating. Also, we have

$$|f^n(x) - z| \leq \frac{K^n}{1 - K}|f(x) - x|$$

valid for any $x \in X$, regarding the convergence $f^{(n)}(x) \to z$.

PROOF. This is something standard, due to Banach, the idea being as follows:

(1) To start with, observe the similarity with the Lipschitz considerations from chapter 5. In fact, a function as in the statement is called a contraction, and being a contraction means to be Lipschitz, with Lipschitz constant $K < 1$. And more on this in a moment, when discussing examples and illustrations for the present result.

(2) Getting now to the proof of the result, in what regards the uniqueness of the fixed point, our assumption $K < 1$ shows that we have:

$$x \neq y \implies |f(x) - f(y)| < |x - y|$$

But this condition prevents the existence of two fixed points, as desired.

(3) Regarding now the existence part, pick $x = x_0 \in X$, and set $x_n = f^n(x_0)$. In order to prove that the sequence $\{x_n\} \subset X$ is Cauchy, observe first that we have:

$$\begin{aligned} |x_{n+1} - x_n| &\leq K|x_n - x_{n-1}| \\ &\leq K^2|x_{n-1} - x_{n-2}| \\ &\vdots \\ &\leq K^n|x_1 - x_0| \end{aligned}$$

(4) Now by using the triangle inequality, we obtain from this, for $n > m$:

$$\begin{aligned} |x_n - x_m| &\leq \sum_{j=m+1}^{n} |x_j - x_{j-1}| \\ &\leq K^m \sum_{j=0}^{n-m-1} K^j |x_1 - x_0| \\ &\leq \frac{K^m}{1-K}|x_1 - x_0| \end{aligned}$$

(5) Thus the sequence $\{x_n\} \subset X$ is Cauchy, and since our space $X \subset \mathbb{R}$ was assumed to be closed, the limit of this Cauchy sequence must belong to it:

$$x_n \to z \in X$$

But with this in hand, Theorem 6.1 applies, and shows that we have $f(z) = z$.

(6) Finally, in what regards the estimate at the end, in the statement, let us go back to the main estimate obtained before, which was as follows, for any $n > m$:

$$|x_n - x_m| \leq \frac{K^m}{1-K}|x_1 - x_0|$$

But this gives, with $m \to \infty$, the estimate in the statement, as desired. □

Regarding now the applications of these above result, there are many of these, and with the Lipschitz constants computed in chapter 5 being of great help here. We will not get into this in detail, but as a matter of having an illustration for this, let us record:

THEOREM 6.6. *The Babylonian function for extracting the square root of $a > 1$,*

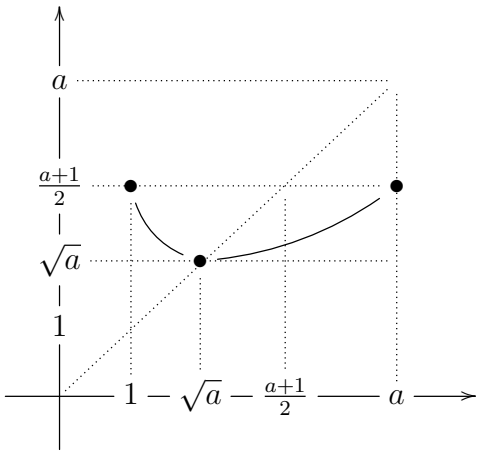
$$f(x) = \frac{x + a/x}{2}$$

has the following properties:

(1) *For $a < 3$ it is a contraction on $[1, a]$, with constant $K = (a-1)/2$.*
(2) *For $a > 3$ it is a contraction on $[\sqrt{a}, a/3]$, with constant $K = (9/a - 1)/2$.*

PROOF. This is something quite interesting, with $a = 3$ producing indeed a mysterious dichotomy, as above. In what regards the proof, however, since all this is not essential to us, we will go quite quick. So, here is how the quick proof goes:

(1) Let us first redraw the graph of $f(x) = (x + a/x)/2$. This was as follows:



(2) Getting now to the computation of the Lipschitz constant, it is pretty much clear that $|(f(x) - f(y))/(x - y)|$ will be maximized at left, for $x, y \simeq 1$. So, let us do the computation there. We have the following estimate, valid for any $t > 0$:

$$\begin{aligned}
\frac{f(1) - f(1+t)}{t} &= \frac{\frac{1+a}{2} - \frac{1+t+a/(1+t)}{2}}{t} \\
&= \frac{1}{2t}\left(a - t - \frac{a}{1+t}\right) \\
&= \frac{1}{2t} \cdot \frac{a + at - t - t^2 - a}{1+t} \\
&= \frac{1}{2} \cdot \frac{a - 1 - t}{1+t} \\
&= \frac{1}{2}\left(\frac{a}{1+t} - 1\right) \\
&< \frac{a-1}{2}
\end{aligned}$$

Now since the last estimate becomes an equality with $t \to 0$, we conclude that the Lipschitz constant, when computed at left, as explained above, is given by:

$$K = \frac{a-1}{2}$$

(3) Of course this is not the end of the story, because all this was based on our graph drawn above, and for being 100% sure that we are not making mistakes here, we still have to compute the Lipschitz constant on the whole $[1,a]$. But the computations here are not too difficult, similar to the above one, say exercise for you, and we obtain indeed:

$$K=\frac{a-1}{2}$$

(4) With this done, let us see if we have a contraction or not. We have:

$$\frac{a-1}{2}<1\iff a<3$$

Thus, done with the first assertion in the statement, dealing with the case $a<3$.

(5) Getting now to the case $a>3$, in view of the above, we certainly must get rid of a certain interval of type $[1,b]$, in order to have a contraction. And here, a natural choice is that of getting rid of the interval $[1,\sqrt{a}]$. Thus, our problem is now on $[\sqrt{a},a]$.

(6) Moreover, since $a>3$ implies $\sqrt{a}<a/3$, it does not make much sense to look for $\sqrt{a}$ inside $[a/3,a]$. So, let us get rid of this latter interval too, leading to $[\sqrt{a},a/3]$.

(7) With these modifications done, the problem is that of computing the Lipschitz constant of f on the interval $[\sqrt{a},a/3]$. But here we can argue, a bit as in (2) before, that $|(f(x)-f(y))/(x-y)|$ will be maximized at right, for $x,y\simeq a/3$. So, let us do the computation there. We have the following estimate, valid for any $t>0$:

$$\begin{aligned}\frac{f(a/3)-f(a/3-t)}{t}&=\frac{(\frac{a}{3}+\frac{3}{a/3})-(\frac{a}{3}-t+\frac{a}{a/3-t})}{2t}\\&=\frac{1}{2t}\left(3+t-\frac{3a}{a-3t}\right)\\&=\frac{1}{2t}\cdot\frac{3a-9t+at-3t^2-3a}{a-3t}\\&=\frac{1}{2}\cdot\frac{a-3t-9}{a-3t}\\&=\frac{1}{2}\left(1-\frac{9}{a-3t}\right)\\&<\frac{1}{2}\left(1-\frac{9}{a}\right)\end{aligned}$$

Now since the last estimate becomes an equality with $t\to 0$, we conclude that the Lipschitz constant, when computed at right, as explained above, is given by:

$$K=\frac{1}{2}\left(1-\frac{9}{a}\right)$$

But then, as before in (3), this is indeed the Lipschitz constant, on $[\sqrt{a},a/3]$.

(8) Now let us see if we have a contraction. And here, some magic strikes, with a remarkable equivalence, which was not expected or really needed, as follows:

$$\frac{1}{2}\left(1-\frac{9}{a}\right)<1 \iff a>3$$

Thus, proof finished, and as a bonus, we have learned something interesting about the number $a=3$, that we can use later, say, for various alchemy purposes.

(9) Finally, for the story to be complete, we have left some exercises in the above, in relation with the full and formal computation of Lipschitz constants, that you are encouraged to solve. But then, as a second exercise, on the same topic, do not hesitate to come back to this after reading Part III, with the formula $f'(x)=(1-a/x^2)/2$, that you will learn there, in hand, and work out an even simpler proof, for all this. □

Getting back now generalities, from Theorems 6.1, 6.2 and 6.5, many other things can be said, as a continuation of that material, regarding the mechanics of the convergence $x_n\to z$. In fact, all this, theory of fixed points and applications, rather belongs to advanced mechanics, and exercise of course for you, to learn a bit about all this.

6b. Uniform convergence

Switching topics, but still in relation with convergence in the context of the real functions, we would like to extend now the material from chapter 1 regarding the numeric sequences and series, to the case of sequences and series of functions.

To start with, with our study, we can talk about the convergence of sequences of functions, $f_n\to f$, in a quite straightforward way, as follows:

DEFINITION 6.7. *We say that f_n converges pointwise to f, and write $f_n\to f$, if*

$$f_n(x)\to f(x)$$

for any x. Equivalently, $\forall x,\forall\varepsilon>0,\exists N\in\mathbb{N},\forall n\geq N,|f_n(x)-f(x)|<\varepsilon$.

The question is now, assuming that f_n are continuous, does it follow that f is continuous? I am pretty much sure that you think that the answer is yes, based on:

$$\begin{aligned}\lim_{y\to x}f(y)&=\lim_{y\to x}\lim_{n\to\infty}f_n(y)\\&=\lim_{n\to\infty}\lim_{y\to x}f_n(y)\\&=\lim_{n\to\infty}f_n(x)\\&=f(x)\end{aligned}$$

However, this proof is wrong, because we know well from chapter 1 that we cannot intervert limits, with this being a common beginner mistake. In fact, the result itself is wrong in general, as shown by the following counterexample:

PROPOSITION 6.8. *The pointwise limit of the functions $f_n : [0,1] \to \mathbb{R}$ given by $f_n(x) = x^n$, which are obviously continuous, is given by*

$$\lim_{n\to\infty} x^n = \begin{cases} 0 & , \quad x \in [0,1) \\ 1 & , \quad x = 1 \end{cases}$$

which is obviously discontinuous.

PROOF. This is indeed something self-explanatory, coming from definitions. □

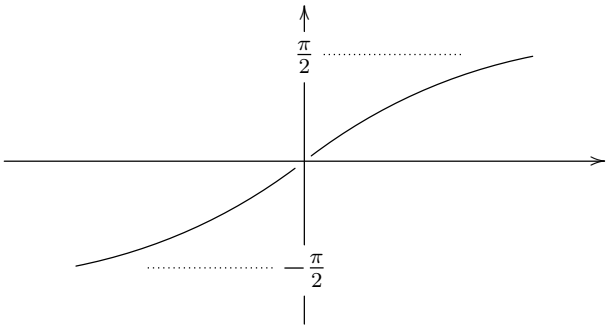
Of course, you might say that allowing $x = 1$ in the above might be a bit unnatural, for whatever reasons, but there is an answer to this too. We can do worse, as follows:

THEOREM 6.9. *The basic step function, namely the sign function*

$$sgn(x) = \begin{cases} -1 & , \quad x < 0 \\ 0 & , \quad x = 0 \\ 1 & , \quad x > 0 \end{cases}$$

can be approximated by suitable rescalings of $\arctan(x) = \tan^{-1}(x)$.

PROOF. The arctangent, obtained by flipping the graph of tan, looks as follows:



Thus $\arctan(x)$ looks a bit like $sgn(x)$, so to say, but one problem comes from the fact that its image is $[-\pi/2, \pi/2]$, instead of the desired $[-1, 1]$. So, let us set:

$$f(x) = \frac{2}{\pi} \arctan(x)$$

Now with this done, we must stretch the variable x, as to get our function closer and closer to $sgn(x)$. This can be done in several ways, a standard one being as follows:

$$g_n(x) = \frac{2}{\pi} \arctan(nx)$$

So, let us see if this works. We have the following computation, for $x > 0$:

$$\lim_{n\to\infty} g_n(x) = \frac{2}{\pi} \arctan(\infty) = \frac{2}{\pi} \cdot \frac{\pi}{2} = 1$$

Similarly, we have the following computation, this time for $x < 0$:

$$\lim_{n\to\infty} g_n(x) = \frac{2}{\pi}\arctan(-\infty) = \frac{2}{\pi}\left(-\frac{\pi}{2}\right) = -1$$

Finally, for $x = 0$ the limit is that of the constant 0 sequence, as follows:

$$\lim_{n\to\infty} g_n(0) = \frac{2}{\pi}\cdot 0 = 0$$

We conclude from this that we have the following pointwise convergence:

$$\lim_{n\to\infty} g_n(x) = \begin{cases} -1 & , \quad x < 0 \\ 0 & , \quad x = 0 \\ 1 & , \quad x > 0 \end{cases}$$

In other words, we have proved that we have the following approximation:

$$\lim_{n\to\infty} \frac{2}{\pi}\arctan(nx) = sgn(x)$$

Thus, we are led to the conclusion in the statement. □

So, this is the situation with pointwise convergence, and not very good all this, hope you agree with me. In fact, even worse, we have truly scary things, as follows:

FACT 6.10. *There are examples of pointwise convergence of functions*

$$f_n \to f$$

with each f_n being continuous, and with f totally discontinuous.

To be more precise, this is something a bit more technical, that we will not really need in what follows, and we will leave some exploration here as an exercise. The idea is that of using suitable linear combinations of functions like in Theorem 6.9, or in Proposition 6.8, shifted as for the discontinuities to appear at various points $x \in \mathbb{R}$.

Time to draw some conclusions? I would say, based on the above, the following:

CONCLUSION 6.11. *Our above notion of convergence of functions is wrong.*

And in the hope that we agree on this, I mean a definition bringing only troubles, and zero theorems, is something that we can safely label as wrong. Or at least this is how things in science go, where your job, following Newton, Maxwell, Einstein and the others, is to constantly fix previous definitions, equations, models and theories, all wrong.

In practice now, what to do? We would like to have in our bag of theorems something saying that $f_n \to f$ with f_n continuous implies f continuous. And fortunately, this can be done indeed, with a suitable refinement of the notion of convergence, as follows:

DEFINITION 6.12. *We say that f_n converges uniformly to f, and write $f_n \to_u f$, if:*

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, |f_n(x) - f(x)| < \varepsilon, \forall x$$

That is, the same condition as for $f_n \to f$ must be satisfied, but with the $\forall x$ at the end.

And it is this "$\forall x$ at the end" which makes the difference, and will make our theory work. As a first observation, when comparing with Definition 6.7, we have:

PROPOSITION 6.13. *Uniform convergence implies pointwise convergence,*

$$f_n \to_u f \implies f_n \to f$$

but the converse is not true, in general.

PROOF. This is something quite obvious, the idea being as follows:

(1) The first assertion is plainly clear from definitions. Indeed, let us compare with the definition of the pointwise convergence, which was as follows:

$$\forall x, \forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, |f_n(x) - f(x)| < \varepsilon$$

Now since the $\forall x$ gets moved to the right, we formally get the result, right away.

(2) As for the second assertion, the simplest counterexamples here are the functions $f_n : [0,1] \to \mathbb{R}$ given by $f_n(x) = x^n$, that we met before in Proposition 6.8. Indeed, the uniform convergence of these functions on $[0,1)$ would mean the following:

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, x^n < \varepsilon, \forall x \in [0,1)$$

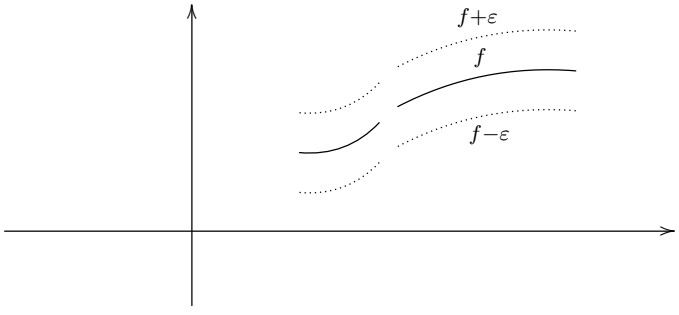
But this is obviously wrong, because no matter how big N is, we have:

$$\lim_{x \to 1} x^N = 1$$

Thus, we can find $x \in [0,1)$ such that $x^N > \varepsilon$, so we have our counterexample. □

In practice now, in order to best understand the uniform convergence, we have:

PROPOSITION 6.14. *The uniform convergence $f_n \to_u f$ means that the $(-\varepsilon, \varepsilon)$ strip drawn along the graph of f, which looks as follows,*



must contain the graphs of all functions f_n, with $n >> 0$.

PROOF. This is again something coming from definitions. Indeed, the uniform convergence condition in Definition 6.12 can be written as follows:

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N \ , \ f(x) - \varepsilon < f_n(x) < f(x) + \varepsilon \ , \forall x$$

But this leads precisely to the conclusion in the statement. □

With this discussed, let us state now our main theorem regarding the uniform convergence, which solves the problems that we were having before, as follows:

THEOREM 6.15. *Assuming that f_n are continuous, and that*

$$f_n \to_u f$$

then f is continuous. That is, uniform limit of continuous functions is continuous.

PROOF. This is something quite fundamental, the idea being as follows:

(1) To start with, philosophically speaking, and as previously advertised, it is the "$\forall x$ at the end" in Definition 6.12 that will make all this work.

(2) Indeed, let us try to prove that the limit f is continuous at some point x. For this purpose, we pick a number $\varepsilon > 0$. Since $f_n \to_u f$, we can find $N \in \mathbb{N}$ such that:

$$|f_N(z) - f(z)| < \frac{\varepsilon}{3} \quad , \quad \forall z$$

(3) On the other hand, since f_N is continuous at x, we can find $\delta > 0$ such that:

$$|x - y| < \delta \implies |f_N(x) - f_N(y)| < \frac{\varepsilon}{3}$$

(4) But with this, we are done. Indeed, for $|x - y| < \delta$ we have:

$$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f_N(x)| + |f_N(x) - f_N(y)| + |f_N(y) - f(y)| \\ &\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \\ &= \varepsilon \end{aligned}$$

Thus, the limit function f is continuous at x, and we are done. □

In view of this, the notion of uniform convergence in Definition 6.12 is something quite interesting, worth some more study. As a next result about it, we have:

THEOREM 6.16. *The following happen, regarding the uniform limits:*

(1) *$f_n \to_u f$, $g_n \to_u g$ imply $f_n + g_n \to_u f + g$.*
(2) *$f_n \to_u f$, $g_n \to_u g$ imply $f_n g_n \to_u fg$.*
(3) *$f_n \to_u f$, $f \neq 0$ imply $1/f_n \to_u 1/f$.*
(4) *$f_n \to_u f$, g continuous imply $f_n \circ g \to_u f \circ g$.*
(5) *$f_n \to_u f$, g continuous imply $g \circ f_n \to_u g \circ f$.*

PROOF. All this is routine, exactly as for the results for numeric sequences from chapter 1, with no difficulties or tricks involved, say exercise for you. □

There is some abstract mathematics to be done as well. Let us start with:

PROPOSITION 6.17. *The uniform convergence condition $f_n \to_u f$ is equivalent to*

$$\sup_x \big|f_n(x) - f(x)\big| \ \longrightarrow_{n\to\infty} \ 0$$

and with the sup and the limit being, as usual, not to be interverted.

PROOF. There are several things going on here, the idea being as follows:

(1) To start with, what we say above is clear from definitions, and is even more clear by using the "strip" interpretation of uniform convergence, from Proposition 6.14.

(2) As for the last assertion, this is our usual word of warning, regarding such things, but out of curiosity, let us see what happens there. Our condition reads:

$$\lim_{n\to\infty} \sup_x \big|f_n(x) - f(x)\big| = 0$$

Now when doing the bad thing, interverting the sup and the limit, we obtain:

$$\sup_x \lim_{n\to\infty} \big|f_n(x) - f(x)\big| = 0$$

(3) So, what does this latter condition mean? Since a supremum of positive numbers vanishes precisely when all the positive numbers vanish, this is equivalent to:

$$\lim_{n\to\infty} \big|f_n(x) - f(x)\big| = 0 \quad , \quad \forall x$$

But, what we have here is the old notion of convergence, the pointwise one. □

As a continuation of this, the following question naturally appears:

QUESTION 6.18. *Can we reach to a better understanding of $f_n \to_u f$, by saying that*

$$d(f_n, f) = \sup_{x\in X} \big|f_n(x) - f(x)\big|$$

is some sort of geometric distance, which must go to zero?

And good question this is. In answer, the spaces of functions $f : X \to \mathbb{R}$ being infinite dimensional, unless $X \subset \mathbb{R}$ is finite, let us first formulate the following definition:

DEFINITION 6.19. *A norm on a real vector space V, which can be finite dimensional or not, is an application $||.|| : V \to [0, \infty)$ having the following properties:*

(1) $||x|| = 0 \iff x = 0$.
(2) $||\lambda x|| = |\lambda| \cdot ||x||$, $\forall \lambda \in \mathbb{R}$.
(3) $||x + y|| \leq ||x|| + ||y||$.

As basic examples, we have $V = \mathbb{R}$, with $||x|| = |x|$, the usual absolute value of real numbers, or $V = \mathbb{C}$, with $||x|| = |x|$, the usual absolute value of complex numbers. More examples in a moment, but before that let us record, as a key result about this:

PROPOSITION 6.20. *Any normed vector space is a metric space, in the sense that*

$$d : V \times V \to [0, \infty) \quad , \quad d(x, y) = ||x - y||$$

has the following properties:

(1) $d(x, y) > 0$ *if* $x \neq y$, *and* $d(x, x) = 0$.
(2) $d(x, y) = d(y, x)$.
(3) $d(x, y) \leq d(x, z) + d(y, z)$.

PROOF. Here (1) and (2) are both clear, and (3) comes from:

$$||x - y|| = ||(x - z) + (z - y)|| \leq ||x - z|| + ||z - y||$$

Thus, we are led to the conclusion in the statement. □

Getting now to the examples of normed vector spaces, there are many of them, both finite and infinite dimensional. In finite dimensions, we have the following result:

THEOREM 6.21. *The following are normed vector spaces:*

(1) $\mathbb{R}^N$, *with* $||x||_2 = \sqrt{\sum_i x_i^2}$ *being the usual vector length.*
(2) $\mathbb{R}^N$ *again, with* $||x||_1 = \sum_i |x_i|$, *or with* $||x||_\infty = \max_i |x_i|$.
(3) *More generally,* $\mathbb{R}^N$ *with* $||x||_p = \sqrt[p]{\sum_i |x_i|^p}$, *for any* $p \in [1, \infty]$.

PROOF. This is something very standard, the idea being as follows:

(1) The first two norm axioms are clear, and in order to check the third one, the best is to note that we have $||x|| = \sqrt{< x, x >}$, with $< x, y >= \sum_i x_i y_i$. Thus, we have:

$$\begin{aligned} ||x + y|| \leq ||x|| + ||y|| &\iff ||x + y||^2 \leq ||x||^2 + ||y||^2 + 2||x|| \cdot ||y|| \\ &\iff ||x||^2 + ||y||^2 + 2 < x, y > \leq ||x||^2 + ||y||^2 + 2||x|| \cdot ||y|| \\ &\iff < x, y > \leq ||x|| \cdot ||y|| \end{aligned}$$

But this latter inequality, called Cauchy-Schwarz inequality, which is something very useful, when doing analysis, can be established by considering the following function:

$$f(t) = ||x + ty||^2 = ||x||^2 + 2t < x, y > +t^2||y||^2$$

Indeed, this function being a degree 2 polynomial in t, that we know to be positive, its discriminant must be negative, $\Delta \leq 0$. But the discriminant is:

$$\Delta = 4 < x, y >^2 -4||x||^2||y||^2$$

Thus we have Cauchy-Schwarz, and then the third norm axiom, as explained above.

(2) The fact that $||.||_1$ and $||.||_\infty$ are indeed norms is clear from definitions.

(3) This is something more technical, that we will not need in what follows next, and that we included only for making it clear that $||.||_1$, $||.||_2$, $||.||_\infty$ are of the same nature,

namely particular cases of $||.||_p$, with $p \in [1,\infty]$, with the case $p = \infty$ coming from:

$$\lim_{p\to\infty} \sqrt[p]{|x_1|^p + \ldots + |x_N|^p} = \max(|x_1|, \ldots, |x_N|)$$

And more on this later, in Part IV, when systematically discussing such things. □

The above Theorem 6.21 was level 1, in normed vector spaces. At level 2, we have:

THEOREM 6.22. *When endowing $\mathbb{R}^\infty$ with the would-be norms*

$$||x||_1 = \sum_i |x_i| \quad , \quad ||x||_2 = \sqrt{\sum_i x_i^2} \quad , \quad ||x||_\infty = \sup_i |x_i|$$

the corresponding subspaces l^1, l^2, l^∞ of vectors of norm $< \infty$ are normed vector spaces.

PROOF. This follows exactly as Theorem 6.21, with of course some care for convergence, as indicated in the statement, because what we have been doing there perfectly works at $N = \infty$ too. And with the remark that we can talk, more generally, about the spaces l^p with $p \in [1,\infty]$, in the same way. And more on this later, in Part IV. □

Getting now to level 3, we have here a unification of Theorems 6.21 and 6.22:

THEOREM 6.23. *Given a set X, finite or not, consider the space of functions on it:*

$$F(X) = \Big\{ f : X \to \mathbb{R} \Big\}$$

Then, when endowing this space with the would-be norms

$$||f||_1 = \sum_{x\in X} |f(x)| \quad , \quad ||f||_2 = \sqrt{\sum_{x\in X} f(x)^2} \quad , \quad ||f||_\infty = \sup_{x\in X} |f(x)|$$

the subspaces $l^1(X), l^2(X), l^\infty(X)$ of functions of norm $< \infty$ are normed vector spaces.

PROOF. This is a unification of Theorems 6.21 and 6.22, coming from the fact that the vectors in $\mathbb{R}^N$, with $N \in \mathbb{N} \cup \{\infty\}$, are the functions $f : X \to \mathbb{R}$, with $|X| = N$. Also, we can talk about $L^p(X)$ too, with $p \in [1,\infty]$, and more on this later, in Part IV. □

The above is quite nice, with $L^\infty(X)$ and its norm $||.||_\infty$ making the link with Question 6.18. So, getting now to level 4 in functional analysis, let us formulate here:

THEOREM 6.24. *Given $X \subset \mathbb{R}$, the space of continuous bounded functions on it,*

$$C_b(X) = \Big\{ f : X \to \mathbb{R} \text{ continuous, bounded} \Big\}$$

is a normed vector space, with the usual sup norm of functions, given by:

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

The convergence in $C_b(X)$ is then the uniform convergence $f_n \to_u f$, and $C_b(X)$ is complete. Also, when X is compact, we have $C_b(X) = C(X)$, usual continuous functions.

PROOF. We have several things going on there, the idea being as follows:

(1) To start with, we had a nice theory developing for $p = 1, 2, \infty$, and more generally for $p \in [1, \infty]$. Unfortunately we will have to ditch now all exponents $p < \infty$, because the correct extension of Theorem 6.23 to the subsets $X \subset \mathbb{R}$ involves replacing the sums there by integrals, that we will only learn in Part IV. But we will be back to this.

(2) Focusing now on $p = \infty$, consider, as in Theorem 6.23, the following space:

$$l^\infty(X) = \left\{ f : X \to \mathbb{R} \,\middle|\, ||f||_\infty < \infty \right\}$$

This is then a normed space, with norm $||.||_\infty$, and by Proposition 6.17, we have:

$$f_n \to_u f \iff ||f_n - f|| \to 0$$

(3) Getting now to the continuous functions, we have an inclusion as follows:

$$C_b(X) \subset l^\infty(X)$$

Thus the convergence in $C_b(X)$ is the uniform convergence $f_n \to_u f$, as stated.

(4) Next, we know from Theorem 6.15 that a uniform limit of continuous functions is continuous. Since it is also clear, say by using Proposition 6.14, that a uniform limit of bounded functions is bounded, we conclude that $C_b(X)$ iscomplete, as stated.

(5) Finally, when $X \subset \mathbb{R}$ is compact any continuous function on it $f : X \to \mathbb{R}$ is automatically bounded, so we have in this case $C_b(X) = C(X)$, as stated. □

Still with me, I hope, after this excursion in functional analysis. As a conclusion, Theorem 6.24 is all you need to know about uniform convergence, encapsulating everything that we said before, on the subject. And as a second conclusion, Theorem 6.24 is just the tip of the iceberg, of functional analysis at level 4, and the continuation remains to be discussed later, in Part IV, once we will know how to integrate functions. More later.

6c. Weierstrass theorem

Getting now to more concrete things, we have the following fundamental finding, due to Weierstrass, regarding the approximation of functions by polynomials:

FACT 6.25 (Weierstrass). *Any continuous function on a closed interval*

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

can be uniformly approximated by polynomials.

In order to prove this, we will need some knowledge in probability. We already talked about Poisson laws in chapter 4, and we will need here in fact something simpler than that, namely the basic theory of the Bernoulli and binomial laws. Let us start with:

DEFINITION 6.26. *The Bernoulli law of parameter $x \in [0,1]$ is the law*

$$\rho_x = (1-x)\delta_0 + x\delta_1$$

appearing when flipping a biased coin, $P(\text{heads}) = x$, $P(\text{tails}) = 1-x$.

To be more precise, when flipping a biased coin as above, and betting heads, your winning law is ρ_x. Next, let us flip the biased coin several times in a row. This leads to:

THEOREM 6.27. *When flipping a x-biased coin n times in a row, the law is*

$$\rho_{xn} = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} \delta_k$$

called binomial law of parameters $x \in [0,1]$ and $n \in \mathbb{N}$.

PROOF. This is something very standard, the idea being as follows:

(1) Observe first that at $n = 1$ we have indeed the Bernoulli law ρ_x.

(2) In general, we can argue that when flipping the coin n times in a row, and betting heads, the probability of winning k times, among our n attempts, is given by:

$$P(k \text{ wins}) = \binom{n}{k} P(\text{heads})^k P(\text{tails})^{n-k} = \binom{n}{k} x^k (1-x)^{n-k}$$

Thus, we are led to the formula of ρ_{xn} in the statement.

(3) Alternatively, and being a bit more formal, since our n coin tosses are independent, and independence corresponds to convolution, at the level of laws, we have:

$$\begin{aligned} \rho_{xn} &= \rho_x^{*n} \\ &= \Big[(1-x)\delta_0 + x\delta_1\Big]^{*n} \\ &= \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} \delta_k \end{aligned}$$

(4) Finally, let us mention that there is a relation here with the Poisson laws from chapter 4 too, coming from the Poisson limit theorem, which states that:

$$\lim_{n\to\infty} \left[\left(1 - \frac{t}{n}\right)\delta_0 + \frac{t}{n}\,\delta_1\right]^{*n} = p_t$$

And exercise of course for you, to learn more about all this. □

Getting now to the study of the binomial laws, we have here:

THEOREM 6.28. *The binomial law ρ_{xn} has the following properties:*

(1) *The mean is $E = nx$.*
(2) *The variance is $V = nx(1-x)$.*

PROOF. In what regards the mean, the computation is as follows:

$$\begin{aligned}
E &= \sum_{k=1}^{n} k\binom{n}{k}x^k(1-x)^{n-k}\\
&= \sum_{k=1}^{n}\frac{n!}{(k-1)!(n-k)!}x^k(1-x)^{n-k}\\
&= nx\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!(n-k)!}x^{k-1}(1-x)^{n-k}\\
&= nx\sum_{t=0}^{n-1}\binom{n-1}{t}x^t(1-x)^{n-t-1}\\
&= nx(x+1-x)^{n-1}\\
&= nx
\end{aligned}$$

With the same trick, we can compute the difference of the first two moments:

$$\begin{aligned}
M_2-M_1 &= \sum_{k=2}^{n}(k^2-k)\binom{n}{k}x^k(1-x)^{n-k}\\
&= \sum_{k=2}^{n}\frac{n!}{(k-2)!(n-k)!}x^k(1-x)^{n-k}\\
&= n(n-1)x^2\sum_{k=2}^{n}\frac{(n-2)!}{(k-2)!(n-k)!}x^{k-2}(1-x)^{n-k}\\
&= n(n-1)x^2\sum_{t=0}^{n-2}\binom{n-2}{t}x^t(1-x)^{n-t-2}\\
&= n(n-1)x^2(x+1-x)^{n-2}\\
&= n(n-1)x^2
\end{aligned}$$

We conclude that the second moment is given by the following formula:

$$M_2 = n(n-1)x^2+nx = nx((n-1)x+1)$$

As for the variance $V = M_2 - M_1^2$, this is given by the following formula:

$$V = nx((n-1)x+1)-(nx)^2 = nx(1-x)$$

Thus, we are led to the conclusions in the statement. □

We can now prove the Weierstrass approximation theorem, as follows:

THEOREM 6.29 (Weierstrass). *Any continuous function on a closed interval*

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

can be uniformly approximated by polynomials.

PROOF. This is something very classical, with a well-known, constructive proof being by using an approximation by suitable Bernstein polynomials, as follows:

(1) We can assume by linearity $[a, b] = [0, 1]$. Consider the following polynomials:

$$b_{kn}(x) = \binom{n}{k} x^k (1 - x)^{n-k}$$

Then, given a continuous function $f : [0, 1] \to \mathbb{R}$, consider the following polynomials:

$$f_n(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) b_{kn}(x)$$

Our claim is that we have $f_n \to_u f$, uniform convergence on $[0, 1]$.

(2) In order to prove this, observe that the polynomials b_{kn} encode the densities of the binomial laws ρ_{xn}. Thus, we have the following formulae, with the first one corresponding to the fact that ρ_{xn} is indeed a probability measure, and with the second and third formulae coming from our mean and variance computations from Theorem 6.28:

$$\sum_{k=0}^{n} b_{kn}(x) = 1$$

$$\sum_{k=0}^{n} \frac{k}{n} \cdot b_{kn}(x) = x$$

$$\sum_{k=0}^{n} \left(x - \frac{k}{n}\right)^2 b_{kn}(x) = \frac{x(1 - x)}{n}$$

(3) In order to estimate now the error $|f_n - f|$, we can use the uniform continuity property of f. So, pick $\varepsilon > 0$, and then $\delta > 0$ such that the following happens:

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

(4) We have then the following estimate, using this, and with $M = \sup|f|$:

$$\begin{aligned}
|f_n(x) - f(x)| &= \left|\sum_{k=0}^{n} f\left(\frac{k}{n}\right) b_{kn}(x) - \sum_{k=0}^{n} f(x) b_{kn}(x)\right| \\
&\leq \sum_{k=0}^{n} \left|f\left(\frac{k}{n}\right) - f(x)\right| b_{kn}(x) \\
&= \sum_{|x-\frac{k}{n}|<\delta} \left|f\left(\frac{k}{n}\right) - f(x)\right| b_{kn}(x) + \sum_{|x-\frac{k}{n}|\geq\delta} \left|f\left(\frac{k}{n}\right) - f(x)\right| b_{kn}(x) \\
&\leq \varepsilon + M \sum_{|x-\frac{k}{n}|\geq\delta} b_{kn}(x)
\end{aligned}$$

(5) In order to deal with the sum on the right, we will need some standard estimates. Let us first recall the Markov inequality, which is something trivial, as follows:

$$P\big(|\varphi| \geq b\big) \leq \frac{E(\varphi)}{b}$$

By using this with $\varphi = (\psi - E)^2$, with $E = E(\psi)$, we obtain the Chebycheff inequality:

$$P\big(|\psi - E| \geq a\big) \leq \frac{E((\psi - E)^2)}{a^2} = \frac{V}{a^2}$$

(6) The point now is that this latter inequality applies to the last sum in (4), with ψ being a variable following the binomial law ρ_{xn}, rescaled to $[0, 1]$, and gives:

$$\begin{aligned}
\sum_{|x-\frac{k}{n}|\geq\delta} b_{kn}(x) &\leq \sum_{k=0}^{n} \delta^{-2} \left(x - \frac{k}{n}\right)^2 b_{kn}(x) \\
&= \delta^{-2} \frac{x(1-x)}{n} \\
&\leq \frac{\delta^{-2}}{4n}
\end{aligned}$$

(7) Now by putting everything together, we obtain the following estimate:

$$|f_n(x) - f(x)| \leq \varepsilon + \frac{\delta^{-2} M}{4n}$$

Thus we have indeed $|f_n - f| \to 0$, uniform convergence, as desired. □

6d. Power series

We would like to end this chapter on the convergence of functions with a discussion of the series of functions. It is convenient here to upgrade right away from real functions to complex functions, where far more things can be said. We first have:

THEOREM 6.30. *Each power series $f(x) = \sum_n c_n x^n$ has a radius of convergence*

$$R \in [0, \infty]$$

which is such that f converges for $|x| < R$, and diverges for $|x| > R$. We have:

$$R = \frac{1}{C} \quad , \quad C = \limsup_{n \to \infty} \sqrt[n]{|c_n|}$$

Also, in the case $|x| = R$ the function f can either converge, or diverge.

PROOF. We have several things going on here, the idea being as follows:

(1) To start with, the result follows from the Cauchy criterion for series, from chapter 1, which says that a series $\sum_n x_n$ converges if $c < 1$, and diverges if $c > 1$, where:

$$c = \limsup_{n \to \infty} \sqrt[n]{|x_n|}$$

Indeed, given $f(x) = \sum_n c_n x^n$ as in the statement, with $x_n = |c_n x^n|$ we obtain that the convergence radius $R \in [0, \infty]$ exists, and is given by the above formula $R = 1/C$.

(2) As a comment here, we also know from chapter 1 the d'Alembert convergence criterion for series, in terms of $|x_{n+1}/x_n|$. In the present context, this gives the following alternative formula for the convergence radius, provided that the limit exists indeed:

$$R = \frac{1}{C} \quad , \quad C = \lim_{n \to \infty} \left| \frac{c_{n+1}}{c_n} \right|$$

(3) Finally, for the examples at the end, when $|x| = R$, the simplest here is to use $f(x) = \sum_n x^n/n$, for which $R = 1$. Indeed, at $x = 1$ we obtain the standard Riemann sum, which diverges, and at $x = -1$ we have an alternating series, which converges. □

At the level of examples of power series, the simplest ones are the geometric series, that we discussed in chapter 5. In relation with these, let us record as well:

THEOREM 6.31. *The complex rational functions can be written as follows,*

$$f(x) = \sum_i \frac{A_i(x)}{(r_i - x)^{n_i}}$$

with $A_i \in \mathbb{C}[X]$, and $r_i \in \mathbb{C}$ being the poles. Also, we have the following formula,

$$\frac{1}{(r - x)^n} = \frac{1}{r^n} \sum_{k=0}^{\infty} \binom{n + k - 1}{n - 1} \left(\frac{x}{r} \right)^k$$

valid for $|x| < r$, which computes these rational functions, in practice.

PROOF. This is indeed something that we know since chapter 2, in the real case, save for an extra factor $B(x)/Q(x)$, with Q being a real polynomial having no roots, which dissapears in the complex case, and the proof in the complex case is similar. □

At a more advanced level, we can talk about $\sin, \cos, \exp, \log$. Let us start with:

THEOREM 6.32. *We can exponentiate the complex numbers, according to*

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

and the function $x \to e^x$ is continuous, and satisfies $e^{x+y} = e^x e^y$.

PROOF. This is something that we know from chapter 4, with the convergence coming from $R = \infty$ obtained via Theorem 6.30, or simply from $|e^x| \leq e^{|x|}$, then with $e^{x+y} = e^x e^y$ coming from a direct computation, and with the continuity coming from this. □

Regarding now the sine and cosine, we have about them the following key result:

THEOREM 6.33. *The sine and cosine are given by the following formulae,*

$$\sin x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l+1}}{(2l+1)!} \quad , \quad \cos x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l}}{(2l)!}$$

which can stand as a definition for sin *and* cos*, on the whole $\mathbb{C}$.*

PROOF. This is something quite tricky, based on the Euler formula $e^{ix} = \cos x + i \sin x$ that we discussed in chapter 4, and with the comment that the formal proof of this Euler formula, using derivatives, is still to come, in chapter 9. In the meantime:

(1) We have the following formula, for any $x \in \mathbb{C}$, coming from Theorem 6.32:

$$\begin{aligned} e^{ix} &= \sum_{k=0}^{\infty} \frac{(ix)^k}{k!} \\ &= \sum_{k=2l} \frac{(ix)^k}{k!} + \sum_{k=2l+1} \frac{(ix)^k}{k!} \\ &= \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l}}{(2l)!} + i \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l+1}}{(2l+1)!} \end{aligned}$$

(2) Now when the variable is real, $x \in \mathbb{R}$, by comparing with the Euler formula $e^{ix} = \cos x + i \sin x$, we are led to the formulae for sin and cos in the statement.

(3) Next, since all series involved above have radius of convergence $R = \infty$, we can formally declare that sin and cos are given by the formulae in the statement, for any $x \in \mathbb{C}$. And with this, the Euler formula $e^{ix} = \cos x + i \sin x$ holds now for any $x \in \mathbb{C}$.

(4) Now let us see what we get for $x = iy$ with $y \in \mathbb{R}$. The sine is given by:

$$\sin(iy) = \sum_{l=0}^{\infty} (-1)^l \frac{(iy)^{2l+1}}{(2l+1)!} = i \sum_{l=0}^{\infty} \frac{y^{2l+1}}{(2l+1)!} = i \cdot \frac{e^y - e^{-y}}{2}$$

As for the cosine, this is given by a quite similar formula, as follows:

$$\cos(iy)=\sum_{l=0}^{\infty}(-1)^l\frac{(iy)^{2l}}{(2l)!}=\sum_{l=0}^{\infty}\frac{y^{2l}}{(2l)!}=\frac{e^y+e^{-y}}{2}$$

(5) Which might sound a bit strange, and as a matter of making sure that we made no mistake here, let us doublecheck with Euler. But this works indeed, as follows:

$$\cos(iy)+i\sin(iy)=\frac{e^y+e^{-y}}{2}-\frac{e^y-e^{-y}}{2}=e^{-y}=e^{i\cdot iy}$$

(6) And we will end our study here. More on such things in chapter 7, and then of course in chapter 9, with our formal proof of the Euler formula coming at that time. □

Finally, regarding the logarithm, the result here is as follows:

THEOREM 6.34. *Given an horizontal strip in the complex plane, having width 2π,*

$$S=\left\{x+iy\middle|x\in\mathbb{R},y\in[a,a+2\pi)\right\}$$

$\exp:S\to\mathbb{C}-\{0\}$ *is bijective, so we can define* log *as being the inverse of this map,*

$$\log=\exp^{-1}:\mathbb{C}-\{0\}\to S$$

and with this, by suitably choosing S, we have the formula

$$\log(1+x)=\sum_{k=1}^{\infty}(-1)^{k+1}\frac{x^k}{k}$$

valid for any $|x|<1$, and at $x=1$ too.

PROOF. The first assertion is something that we know from chapter 5, and the proof of the second assertion is identical to the proof from the real case, from chapter 4. To be more precise, observe first that we have the following formula:

$$\begin{aligned}\sum_{k=1}^{\infty}(-1)^{k+1}\frac{(x+y+xy)^k}{k}&=\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}\sum_{r=0}^{k}\binom{k}{r}(x+xy)^ry^{k-r}\\&=\sum_{r+s\geq1}\frac{(-1)^{r+s+1}}{r+s}\binom{r+s}{r}x^r(1+y)^ry^s\end{aligned}$$

The contribution of $r=0$ to this sum is then the following quantity:

$$C_0=\sum_{s\geq1}(-1)^{s+1}\frac{y^s}{s}$$

As for the contribution coming from $r \geq 1$, this is the following quantity, with the computation using the inversion formula for $(1+y)^r$, from Theorem 6.31:

$$\begin{aligned} C_+ &= \sum_{r\geq 1}(-1)^{r+1}x^r(1+y)^r\sum_{s\geq 0}\frac{(-1)^s}{r+s}\binom{r+s}{r}y^s \\ &= \sum_{r\geq 1}(-1)^{r+1}\frac{x^r(1+y)^r}{r}\cdot\frac{1}{(1+y)^r} \\ &= \sum_{r\geq 1}(-1)^{r+1}\frac{x^r}{r} \end{aligned}$$

Thus, we have proved the following formula, regarding the series in the statement:

$$\sum_{k=1}^{\infty}(-1)^{k+1}\frac{(x+y+xy)^k}{k}=\sum_{r=1}^{\infty}(-1)^{r+1}\frac{x^r}{r}+\sum_{s=1}^{\infty}(-1)^{s+1}\frac{y^s}{s}$$

We conclude that our candidate series for $\log(1+x)$ must be a certain shifted logarithm, $\log_b(1+x)$. But with the slope at $x=0$ being the correct one, this shifted logarithm must be the usual one, $\log(1+x)$, and we are led to the conclusion in the statement. □

6e. Exercises

This was a key analysis chapter, and as exercises on this, we have:

EXERCISE 6.35. *Have some fun with numerics, for the Babylonian method for $\sqrt{a}$.*

EXERCISE 6.36. *Experiment with roots of degree 3 polynomials, found via iteration.*

EXERCISE 6.37. *Learn more about fixed points, iterations, mechanics and dynamics.*

EXERCISE 6.38. *Find examples of pointwise limits which are totally discontinuous.*

EXERCISE 6.39. *Learn various generalizations of the Cauchy-Schwarz inequality.*

EXERCISE 6.40. *Learn about the Arzelà-Ascoli theorem, and its consequences.*

EXERCISE 6.41. *Learn more about the Bernoulli, binomial and Poisson laws.*

EXERCISE 6.42. *Experiment a bit with approximation by Bernstein polynomials.*

As bonus exercise, as usual, learn more complex analysis, related to the above.

CHAPTER 7

Special functions

7a. Fractionary powers

According to our general philosophy, since the beginning of this book, and which is actually the philosophy of mathematics itself, as it was historically developed, over the time, the various explicit functions that we can think of fall in two classes:

(1) First we have the usual power functions x^n, and their versions, such as the polynomials P, or the rational functions P/Q. These count as "basic functions".

(2) And then we have more complicated things, like $\sin, \cos, \exp, \log$, and versions of them, such as $\sin(ix)$ or $\cos(ix)$. These count as "special functions".

Our goal, in this chapter and in the next one, will be to fine-tune our knowledge of explicit functions, along these lines, first with a discussion of the special functions, in this chapter, and then with a discussion of the basic functions, in the next chapter.

As a first comment, you might think that I got it wrong with my plan, why not discussing the basic functions first. In answer, here is a general principle:

PRINCIPLE 7.1. *The simpler your mathematical objects, the more is known about them. That is, at the advanced level, beware of the theory of simple things.*

Getting started now, with special functions as planned, we would first like to talk about arbitrary powers, x^p with $p \in \mathbb{R}$. These are actually a bit in-between basic and special, but in regards with their advanced theory, which needs exp and log, they rather count as special. Here is something remarkable that can be said, about them:

THEOREM 7.2. *We have the generalized binomial formula*

$$(1+x)^p = \sum_{k=0}^{\infty} \binom{p}{k} x^k$$

with the generalized binomial coefficients being given by

$$\binom{p}{k} = \frac{p(p-1)\ldots(p-k+1)}{k!}$$

valid for any exponent $p \in \mathbb{R}$, and any $|x| < 1$.

PROOF. This is something quite tricky, that we already met before at $p \in \mathbb{Z}$, that we will explore in this section at $p \in \mathbb{Z}/2$, and whose proof in general, for $p \in \mathbb{R}$, will be eventually relegated to chapter 9. The idea with all this is as follows:

(1) To start with, at $p = n \in \mathbb{N}$ the generalized binomial coefficients are the usual binomial coefficients at $k \leq n$, and vanish at $k = n + 1$ and higher. Thus, we recover the usual binomial formula, valid for any $x \in \mathbb{R}$, and even $x \in \mathbb{C}$, namely:

$$(1 + x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k$$

(2) Next, as already explained in chapter 2, at $p = -1$ the formula in the statement is the usual formula for the geometric series, valid at $|x| < 1$, namely:

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

(3) More generally, again as explained in chapter 2, at $p = -n$ with $n \in \mathbb{N}$ the formula in the statement takes the following form, and holds indeed, for any $|x| < 1$:

$$\frac{1}{(1 + x)^n} = \sum_{k=0}^{\infty} (-1)^k \binom{n + k - 1}{n - 1} x^k$$

(4) This was for what we already knew. Getting now into the unknown, case $p \in \mathbb{R}$, as a first observation, assuming $p \notin \mathbb{N}$, the radius of convergence is indeed $R = 1$, as shown by the following computation, using the d'Alembert criterion from chapter 6:

$$\begin{aligned} \binom{p}{k + 1} : \binom{p}{k} &= \frac{p(p - 1) \ldots (p - k)}{(k + 1)!} : \frac{p(p - 1) \ldots (p - k + 1)}{k!} \\ &= \frac{p - k}{k + 1} \to -1 \end{aligned}$$

(5) So, the problem is, can we prove now the formula in the statement, at any $p \in \mathbb{R}$, and any $|x| < 1$? In answer, the first thought goes to using the following formula:

$$(1 + x)^p = \exp(p \log(1 + x))$$

Indeed, we know the series of exp and log, and this can potentially lead to a proof.

(6) So, let us see if this works. Using the series of exp and log, we have:

$$\begin{aligned}(1+x)^p &= \exp\left(p\sum_{k=1}^{\infty}(-1)^{k+1}\frac{x^k}{k}\right)\\ &= \sum_{n=0}^{\infty}\frac{p^n}{n!}\left(\sum_{k=1}^{\infty}(-1)^{k+1}\frac{x^k}{k}\right)^n\\ &= \sum_{n=0}^{\infty}\frac{p^n}{n!}\sum_{k_1,\ldots,k_n=1}^{\infty}(-1)^{k_1+\ldots+k_n+n}\frac{x^{k_1+\ldots+k_n}}{k_1\ldots k_n}\\ &= \sum_{s=0}^{\infty}(-x)^s\sum_{n=0}^{\infty}\frac{(-p)^n}{n!}\sum_{k_1+\ldots+k_n=s}\frac{1}{k_1\ldots k_n}\end{aligned}$$

(7) But this gets us into some bad souvenirs, from chapter 4, where we tried precisely to compute this series, at $p=1$, in order to establish the formula for $\log(1+x)$, at that time, and eventually decided to go instead for a more clever method. So, not very good all this, and we will stop here too, again wishing for a more clever method.

(8) And the problem is that, contrary to what was going on in chapter 4 with the series of $\log(1+x)$, where we had 3 possible methods to be tried, and one succeeded, here things are much more rigid, and there does not seem to be any clever alternative method, on the horizon. So, we will stop for good here, and come back to this later, in chapter 9, after learning derivatives, which are the good tool for investigating such things. □

Quite interesting the above discussion, we are learning new things, but in practice we have absolutely nothing, compared to what we already knew from chapter 2. So, as a challenging question, is there any way out of this mud, with the tools that we have?

In answer, here is a more modest question, which is something quite interesting by itself, related to the square roots, which are ubiquitous, that we can try to solve:

QUESTION 7.3. *What are the formulae of the square roots, and their inverses,*

$$\sqrt{1+x}=?\quad,\quad\frac{1}{\sqrt{1+x}}=?$$

and do these confirm the generalized binomial formula, at $p=1/2,-1/2$?

And good question this is. In answer now, browsing through what we did in this book, in relation with square roots, we have nothing at this level of sharpness. So, modesty, and instead of attacking the problem directly, it is probably wiser to cheat a bit, and see first what the generalized binomial formula says, at $p=\pm 1/2$. And here we have:

PROPOSITION 7.4. *The generalized binomial coefficients at $p = \pm 1/2$ are*

$$\binom{1/2}{k} = -2\left(\frac{-1}{4}\right)^k C_{k-1} \quad , \quad \binom{-1/2}{k} = \left(\frac{-1}{4}\right)^k D_k$$

with $D_k = \binom{2k}{k}$ being the central binomial coefficients, and $C_k = \frac{1}{k+1}\binom{2k}{k}$.

PROOF. At $p = 1/2$, the generalized binomial coefficients are as follows:

$$\begin{aligned}
\binom{1/2}{k} &= \frac{1/2(-1/2)\ldots(3/2-k)}{k!}\\
&= (-1)^{k-1}\frac{1\cdot 3\cdot 5\ldots(2k-3)}{2^k k!}\\
&= (-1)^{k-1}\frac{(2k-2)!}{2^{k-1}(k-1)!2^k k!}\\
&= -2\left(\frac{-1}{4}\right)^k C_{k-1}
\end{aligned}$$

As for the case $p = -1/2$, here the binomial coefficients are as follows:

$$\begin{aligned}
\binom{-1/2}{k} &= \frac{-1/2(-3/2)\ldots(1/2-k)}{k!}\\
&= (-1)^k\frac{1\cdot 3\cdot 5\ldots(2k-1)}{2^k k!}\\
&= (-1)^k\frac{(2k)!}{2^k k!2^k k!}\\
&= \left(\frac{-1}{4}\right)^k D_k
\end{aligned}$$

Thus, we are led to the conclusions in the statement. □

The above result is quite interesting, making the link with an old conjecture, that we made back in chapter 1, stating that the central binomial coefficients factorize as $D_k = (k+1)C_k$, with C_k being certain integers. So, time now to prove that conjecture:

THEOREM 7.5. *The Catalan numbers $C_k = \frac{1}{k+1}\binom{2k}{k}$ count:*

(1) *The length $2k$ loops on $\mathbb{N}$, based at 0.*
(2) *The noncrossing pairings of $1, \ldots, 2k$.*
(3) *The noncrossing partitions of $1, \ldots, k$.*
(4) *The length $2k$ Dyck paths in the plane.*

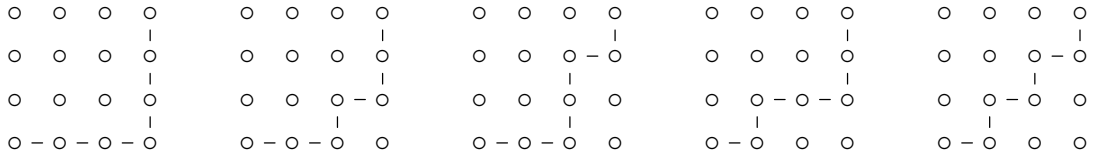
PROOF. All this is standard combinatorics, the idea being as follows:

(1) To start with, in what regards the various objects involved, the length $2k$ loops on $\mathbb{N}$ are the length $2k$ loops on $\mathbb{N}$ that we know, and the same goes for the noncrossing

pairings of $1,\ldots,2k$, and for the noncrossing partitions of $1,\ldots,k$, the idea here being that you must be able to draw the pairing or partition in a noncrossing way.

(2) As for the length $2k$ Dyck paths in the plane, these are by definition the paths from $(0,0)$ to (k,k), marching North-East over the integer lattice $\mathbb{Z}^2\subset\mathbb{R}^2$, by staying inside the square $[0,k]\times[0,k]$, and staying as well under the diagonal of this square.

(3) As an illustration for this, here are the 5 possible Dyck paths at $n=3$:



(4) Thus, we have definitions for all objects involved, and in each case, if you start counting them, you always end up with the same sequence of numbers, namely:

$$1,2,5,14,42,132,429,1430,4862,16796,58786,\ldots$$

(5) In order to prove now that (1-4) produce indeed the same numbers, many things can be said. The idea is that, leaving aside mathematical brevity, and more specifically abstract reasonings of type $a=b,b=c\implies a=c$, what we have to do, in order to fully understand what is going on, is to etablish $\binom{4}{2}=6$ equalities, via bijective proofs.

(6) But this can be done, indeed. As an example here, the noncrossing pairings of $1,\ldots,2k$ from (2) are in bijection with the noncrossing partitions of $1,\ldots,k$ from (3), via fattening the pairings and shrinking the partitions. We will leave the details here as an instructive exercise, and exercise as well, to add (1) and (4) to the picture.

(7) However, as a matter of having our claim formally proved, here is a less elegant argument, which is however quick, and does the job. The point is that, in each of the cases (1-4) under consideration, the numbers C_k that we get are easily seen to satisfy:

$$C_{k+1}=\sum_{a+b=k}C_aC_b$$

Now the initial data being the same, namely $C_1=1$ and $C_2=2$, in each of the cases (1-4) under consideration, we get indeed the same numbers, as desired.

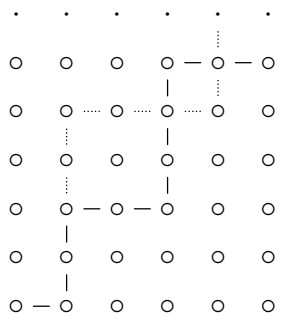
(8) What is next? In view of what we already have, it remains to pick one of the objects (1-4), skilfully do the count, and conclude that we have indeed:

$$C_k=\frac{1}{k+1}\binom{2k}{k}$$

(9) The most convenient is to count the Dyck paths. For this purpose, we can use a trick. Indeed, if we ignore the assumption that our path must stay under the diagonal of

the square, we have $\binom{2k}{k}$ such paths. And among these, we have the "good" ones, those that we want to count, and then the "bad" ones, those that we want to ignore.

(10) So, let us count the bad paths, those crossing the diagonal of the square, and reaching the higher diagonal next to it, the one joining $(0,1)$ and $(k,k+1)$. In order to count these, the trick is to "flip" their bad part over that higher diagonal, as follows:



(11) Now observe that, as it is obvious on the above picture, due to the flipping, the flipped bad path will no longer end in (k,k), but rather in $(k-1,k+1)$. Moreover, more is true, in the sense that, by thinking a bit, we see that the flipped bad paths are precisely those ending in $(k-1,k+1)$. Thus, we can count these flipped bad paths, and so the bad paths, and so the good paths too, and so good news, we are done.

(12) To finish now, by putting everything together, we have:

$$C_k=\binom{2k}{k}-\binom{2k}{k-1}=\binom{2k}{k}-\frac{k}{k+1}\binom{2k}{k}=\frac{1}{k+1}\binom{2k}{k}$$

Thus, we are led to the various conclusions in the statement. □

We can go back now to the generalized binomial formula, and we have:

THEOREM 7.6. *The generalized binomial formula at $p=1/2,-1/2$ reads*

$$\sqrt{1-4t}=1-2\sum_{k=1}^{\infty}C_{k-1}t^k\quad,\quad\frac{1}{\sqrt{1-4t}}=\sum_{k=0}^{\infty}D_kt^k$$

with $C_k=\frac{1}{k+1}\binom{2k}{k}$ and $D_k=\binom{2k}{k}$, and these formulae hold indeed, for $|t|<1/4$.

PROOF. This is quite standard, based on what we have, as follows:

(1) To start with, the formulae in Proposition 7.4 suggest to make the change of variables $x=-4t$ in the generalized binomial formula, and with this change made, that binomial formula at $p=1/2,-1/2$ corresponds precisely to the formulae above.

(2) In order to prove our two formulae, we must establish the following identities:

$$\left(1-2\sum_{k=1}^{\infty}C_{k-1}t^k\right)^2=1-4t\quad,\quad\left(\sum_{k=0}^{\infty}D_kt^k\right)^2=\frac{1}{1-4t}$$

(3) But the first formula is equivalent to the following identity for the Catalan numbers, that we know well to hold, as explained in the proof of Theorem 7.5:

$$\sum_{k+l=n}C_kC_l=C_{n+1}$$

(4) As for the second formula, by using the standard series for $1/(1-4t)$, this is equivalent to the following formula, involving the central binomial coefficients:

$$\sum_{k+l=n}D_kD_l=4^n$$

(5) Now instead of doing again some combinatorics, this time for the numbers D_k, which is certainly possible, and with this being a good exercise for you, let us pull an analysis trick. With $t\to t-\varepsilon$ and $\varepsilon\simeq 0$, our first formula becomes:

$$\begin{aligned}\sqrt{1-4t+4\varepsilon}&=1-2\sum_{k=1}^{\infty}C_{k-1}(t-\varepsilon)^k\\&\simeq 1-2\sum_{k=1}^{\infty}C_{k-1}(t^k-kt^{k-1}\varepsilon)\\&=\sqrt{1-4t}+2\varepsilon\sum_{k=1}^{\infty}D_{k-1}t^{k-1}\end{aligned}$$

(6) On the other hand, again with $\varepsilon\simeq 0$, we have the following estimate:

$$\sqrt{1-4t+4\varepsilon}-\sqrt{1-4t}=\frac{4\varepsilon}{\sqrt{1-4t+4\varepsilon}+\sqrt{1-4t}}\simeq\frac{2\varepsilon}{\sqrt{1-4t}}$$

We conclude from this that we have the following formula, as desired:

$$\frac{1}{\sqrt{1-4t}}=\sum_{k=1}^{\infty}D_{k-1}t^{k-1}$$

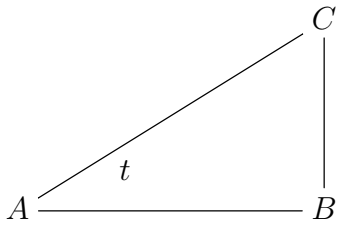
Summarizing, both formulae in the statement proved, one way or another. □

Good work that we did. Along the same lines, it is possible to prove that the generalized binomial formula holds at any $p\in\mathbb{Z}/2$, and we will leave this as an instructive exercise. Also, we will be back to all this in Part III, after learning about derivatives, and about the related notion of Taylor series too, with a general study, at $p\in\mathbb{R}$.

7b. The arcsine family

Switching topics, we know that the bulk of special functions comes from trigonometry, with our collection so far of trigonometric functions consisting of $\sin, \cos, \tan$, and then $\exp, \log$, whose trigonometric nature comes from the Euler formula $e^{it} = \cos t + i \sin t$, and finally hybrid beasts like $\sin(ix)$, $\cos(ix)$. So, time to talk about all this.

To start with, we have some work to do in relation with basic trigonometry, as developed in chapter 3. So, consider a basic right triangle, with an angle t, as follows:



We know from chapter 3 that we have the following formulae:

$$\sin t = \frac{BC}{AC} \quad , \quad \cos t = \frac{AB}{AC} \quad , \quad \tan t = \frac{BC}{AB}$$

However, there are still 3 fractions left, in need of a name, so let us formulate the following definition, completing what we already have, regarding sin, cos, tan:

DEFINITION 7.7. *We can talk about the secant, cosecant and cotangent, as being*

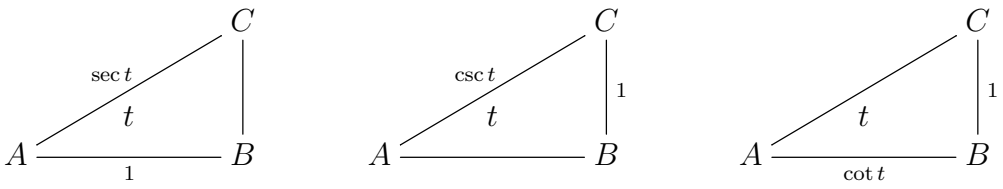
$$\sec t = \frac{AC}{AB} \quad , \quad \csc t = \frac{AC}{BC} \quad , \quad \cot t = \frac{BC}{AB}$$

in the context of a right triangle, as above, or equivalently, as being

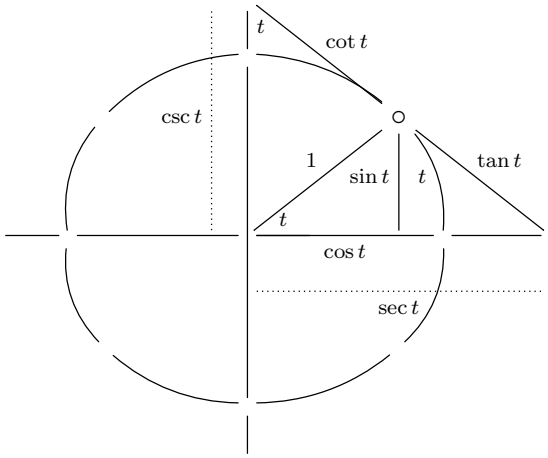
$$\sec t = \frac{1}{\cos t} \quad , \quad \csc t = \frac{1}{\sin t} \quad , \quad \cot t = \frac{1}{\tan t}$$

in terms of the standard trigonometric functions $\sin, \cos, \tan$.

In practice, the secant, cosecant and cotangent can be understood as well geometrically, by using right triangles ABC as above, with a suitable side chosen to be 1:



In relation with this, we have as well the following catch-all picture, featuring a circle too, and justifying the use of the words "secant" and "cosecant" in the above:



But you might probably wonder, what is the usefulness of our new trigonometric functions, sec, csc, cot. In answer, here is a first interesting result, involving csc:

THEOREM 7.8. *The lengths of altitudes in an arbitrary triangle ABC satisfy*

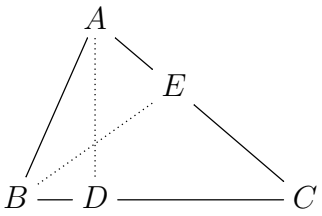
$$[AD - BE - CF] \sim [\csc A - \csc B - \csc C]$$

and in fact we have the following formulae for these altitude lengths,

$$AD = \delta \csc A \quad , \quad BE = \delta \csc B \quad , \quad CF = \delta \csc C$$

with $\delta = S/R$, where S is the area, and R is the radius of the circumscribed circle.

PROOF. Consider a triangle ABC, with two altitudes drawn, as follows:



We have then the following computation for the ratio AD/BE, which along with similar formulae for AD/CF and BE/CF leads to the first assertion:

$$\frac{AD}{BE} = \frac{AB \sin B}{AB \sin A} = \frac{\csc A}{\csc B}$$

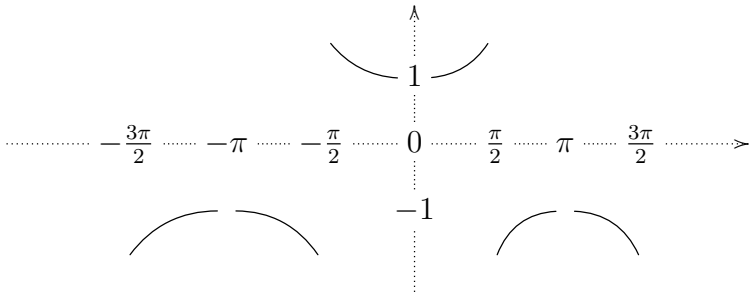
As for the second assertion, recall from chapter 3 that the lengths of sides are:

$$BC = 2R \sin A \quad , \quad AC = 2R \sin B \quad , \quad AB = 2R \sin C$$

But this gives the formulae in the statement, via $S = side \times height/2$. □

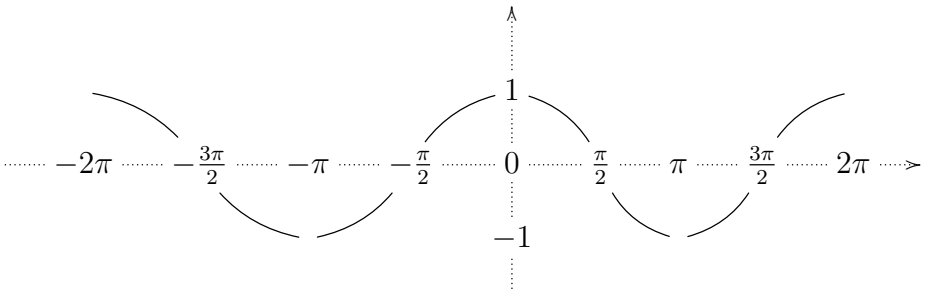
Let us draw now the graphs of our new trigonometric functions. We first have:

PROPOSITION 7.9. *The graph of* $\sec : \mathbb{R} - (\mathbb{Z}\pi + \pi/2) \to \mathbb{R}$ *is as follows,*



with this pattern being repeated indefinitely, to the left and to the right.

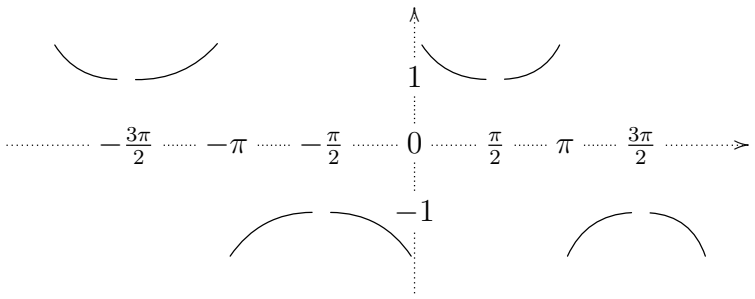
PROOF. This comes indeed from $\sec = 1/\cos$, by applying $x \to 1/x$ to the graph of cos that we found in chapter 3, which was as follows:



Thus, we obtain the graph in the statement, with flattened curves at the multiples of π, and with asymptotes at the multiples of π plus $\pi/2$. □

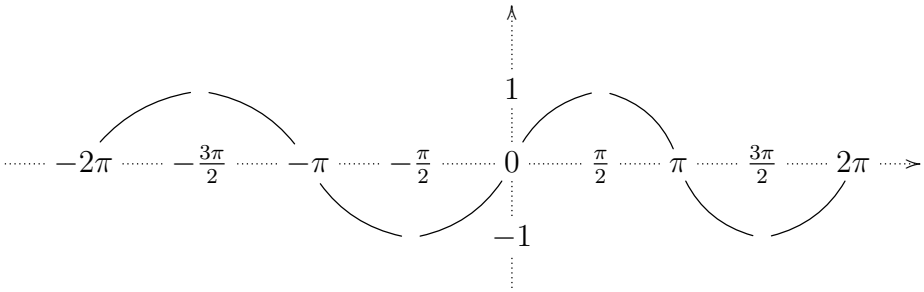
Regarding the cosecant, the graph here is quite similar, as follows:

PROPOSITION 7.10. *The graph of* $\csc : \mathbb{R} - \mathbb{Z}\pi \to \mathbb{R}$ *is as follows,*



with this pattern being repeated indefinitely, to the left and to the right.

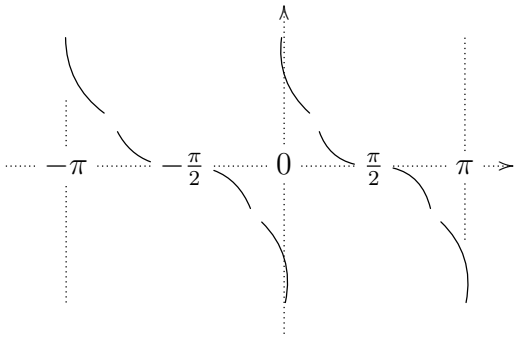
PROOF. This comes indeed from $\csc = 1/\sin$, by applying $x \to 1/x$ to the graph of sin that we found in chapter 3, which was as follows:



Equivalently, we can get this from Proposition 7.9, via $\csc x = \sec(x - \pi/2)$. □

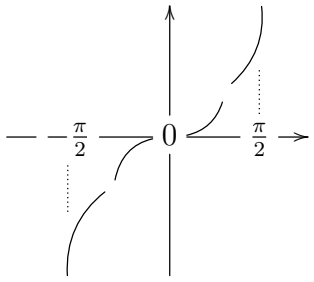
Finally, regarding the cotangent, the graph here is as follows:

PROPOSITION 7.11. *The graph of* $\cot : \mathbb{R} - \mathbb{Z}\pi \to \mathbb{R}$ *is as follows,*



with this pattern being repeated indefinitely, to the left and to the right.

PROOF. This comes indeed from $\cot = 1/\tan$, by applying $x \to 1/x$ to the graph of tan that we found in chapter 3, which was as follows:



In practice, the $x \to 1/x$ procedure amounts in symmetrizing the graph, and then translating it by $\pi/2$, and we will leave some thinking here as an exercise. □

Good work that we did, and the story is not over with this, because we have:

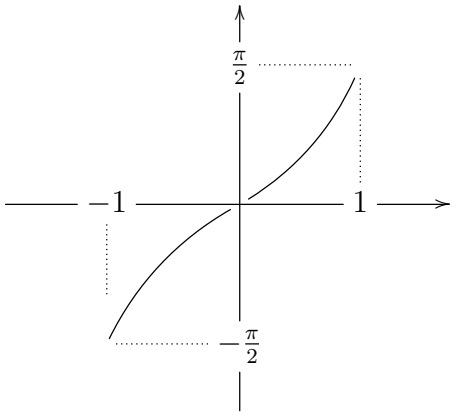
THEOREM 7.12. *We can talk about the inverse trigonometric functions,*

(1) $\arcsin : [-1,1] \to [-\pi/2, \pi/2]$.
(2) $\arccos : [-1,1] \to [0,\pi]$.
(3) $\arctan : \mathbb{R} \to (-\pi/2, \pi/2)$.
(4) $\text{arcsec} : \mathbb{R} - (-1,1) \to (0,\pi) - \{\pi/2\}$.
(5) $\text{arccsc} : \mathbb{R} - (-1,1) \to (-\pi/2, \pi/2) - \{0\}$.
(6) $\text{arccot} : \mathbb{R} \to (0,\pi)$.

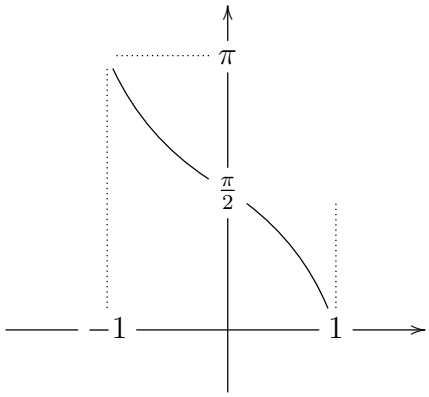
whose graphs can be obtained by flipping those of $\sin, \cos, \tan, \sec, \csc, \cot$.

PROOF. This is something self-explanatory, the idea being as follows:

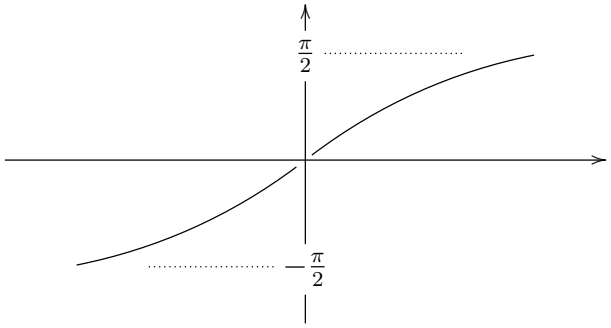
(1) Consider the function $\sin : [-\pi/2, \pi/2] \to [-1,1]$, which is bijective. Its inverse function $\arcsin : [-1,1] \to [-\pi/2, \pi/2]$, obtained by flipping the graph, is as follows:



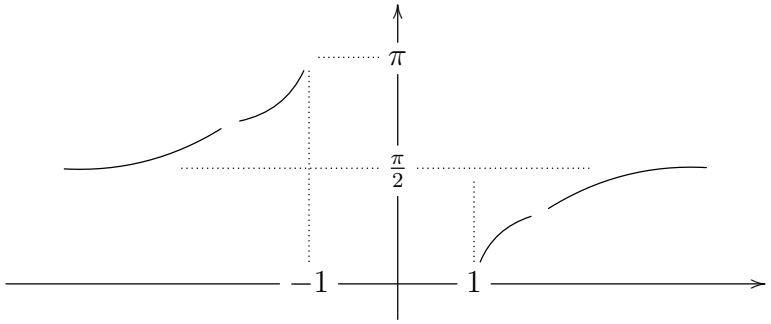
(2) Consider the function $\cos : [0,\pi] \to [-1,1]$, which is bijective. Its inverse function $\arccos : [-1,1] \to [0,\pi]$, obtained by flipping the graph, is then as follows:



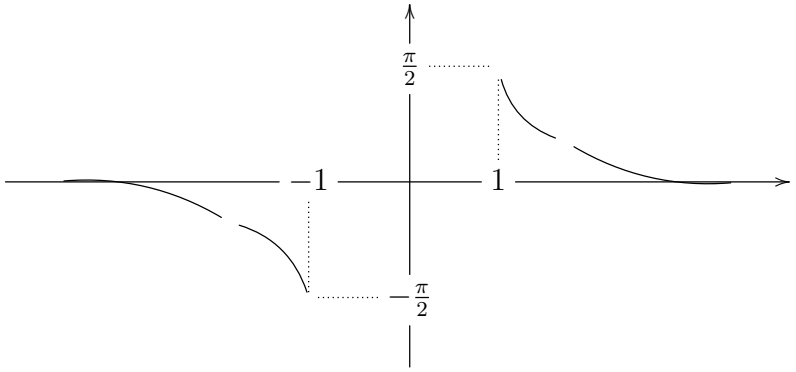
(3) Consider the function $\tan : (-\pi/2, \pi/2) \to \mathbb{R}$, which is bijective. Its inverse function $\arctan : \mathbb{R} \to (-\pi/2, \pi/2)$, obtained by flipping the graph, is as follows:



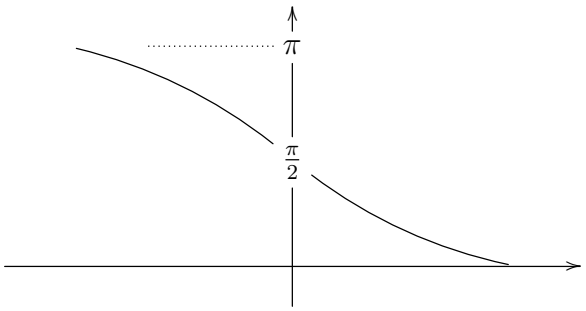
(4) Consider the function $\sec : (0, \pi) - \{\pi/2\} \to \mathbb{R} - (-1, 1)$, which is bijective. Its inverse function $\text{arcsec} : \mathbb{R} - (-1, 1) \to (0, \pi) - \{\pi/2\}$ is then as follows:



(5) Consider the function $\csc : (-\pi/2, \pi/2) - \{0\} \to \mathbb{R} - (-1, 1)$, which is bijective. Its inverse function $\text{arccsc} : \mathbb{R} - (-1, 1) \to (-\pi/2, \pi/2) - \{0\}$ is then as follows:



(6) Consider the function $\cot : (0,\pi) \to \mathbb{R}$, which is bijective. Its inverse function $\text{arccot} : \mathbb{R} \to (0,\pi)$, obtained by flipping the graph, is as follows:



Thus, we are led to the conclusions in the statement. □

Many other things can be said about the inverse trigonometric functions, notably with all sorts of basic formulae for them, coming from the formulae that we know well for the usual trigonometric functions. We will leave some exploration here as an exercise.

7c. Further trigonometry

Back now to the basics, $\sin, \cos, \tan$, we have seen in chapter 3 that some interesting mathematics appears in relation with the sums of angles. This suggests, as a continuation, summing 3 or more angles, and we will explore this here. To start with, we have:

THEOREM 7.13. *The sines of sums of* 3 *angles are given by the formula*

$$\begin{aligned}\sin(x+y+z) &= \sin x\cos y\cos z+\cos x\sin y\cos z\\ &\quad+\cos x\cos y\sin z-\sin x\sin y\sin z\end{aligned}$$

the cosines of sums of 3 *angles are given by the formula*

$$\begin{aligned}\cos(x+y+z) &= \cos x\cos y\cos z-\cos x\sin y\sin z\\ &\quad-\sin x\cos y\sin z-\sin x\sin y\cos z\end{aligned}$$

and we have a formula for the tangent too, namely

$$\tan(x+y+z)=\frac{\tan x+\tan y+\tan z-\tan x\tan y\tan z}{1-\tan x\tan y-\tan x\tan z-\tan y\tan z}$$

provided of course that the denominator is nonzero.

PROOF. We use the addition formulae from chapter 3, namely:

$$\sin(x+y)=\sin x\cos y+\cos x\sin y$$

$$\cos(x+y)=\cos x\cos y-\sin x\sin y$$

In what regards the sine, the computation is as follows:

$$\begin{aligned}
& \sin(x+y+z) \\
= \ & \sin x\cos(y+z)+\cos x\sin(y+z) \\
= \ & \sin x(\cos y\cos z-\sin y\sin z)+\cos x(\sin y\cos z+\cos y\sin z) \\
= \ & \sin x\cos y\cos z+\cos x\sin y\cos z+\cos x\cos y\sin z-\sin x\sin y\sin z
\end{aligned}$$

In what regards the cosine, the computation here is similar, as follows:

$$\begin{aligned}
& \cos(x+y+z) \\
= \ & \cos x\cos(y+z)-\sin x\sin(y+z) \\
= \ & \cos x(\cos y\cos z-\sin y\sin z)-\sin x(\sin y\cos z+\cos y\sin z) \\
= \ & \cos x\cos y\cos z-\cos x\sin y\sin z-\sin x\cos y\sin z-\sin x\sin y\cos z
\end{aligned}$$

Regarding now the tangent, this follows by taking the quotient, or by using:

$$\tan(x+y)=\frac{\tan x+\tan y}{1-\tan x\tan y}$$

Thus, we are led to the conclusions in the statement. □

As a consequence of the above result, obtained with $x=y=z=t$, we have:

THEOREM 7.14. *The sines and cosines of sums of triples of angles are given by*

$$\sin(3t)=3\sin t-4\sin^3 t$$

$$\cos(3t)=4\cos^3 t-3\cos t$$

and we have a formula for the tangent too, namely

$$\tan(3t)=\frac{3\tan t-\tan^3 t}{1-3\tan^2 t}$$

provided of course that the denominator is nonzero.

PROOF. With $x=y=z=t$ in the sine formula from Theorem 7.13, we obtain:

$$\begin{aligned}
\sin(3t) &= 3\sin t\cos^2 t-\sin^3 t \\
&= 3\sin t(1-\sin^2 t)-\sin^3 t \\
&= 3\sin t-4\sin^3 t
\end{aligned}$$

Similarly, with $x=y=z=t$ in the cosine formula from Theorem 7.13, we obtain:

$$\begin{aligned}
\cos(3t) &= \cos^3 t-3\cos t\sin^2 t \\
&= \cos^3 t-3\cos t(1-\cos^2 t) \\
&= 4\cos^3 t-3\cos t
\end{aligned}$$

Finally, with $x=y=z=t$ in the tangent formula from Theorem 7.13 we obtain the formula for the tangent in the statement, without any further manipulation. □

Getting now to potential numeric applications, let us record here:

PROPOSITION 7.15. *The quantities* $a = \sin 10^\circ$, $b = \cos 10^\circ$, $c = \tan 10^\circ$ *satisfy*

$$3a - 4a^3 = \frac{1}{2} \quad , \quad 4b^3 - 3b = \frac{\sqrt{3}}{2} \quad , \quad \frac{3c - c^3}{1 - 3c^2} = \frac{1}{\sqrt{3}}$$

and we have similar equations, for the other multiples of 10°.

PROOF. By taking $t = 10^\circ$ in the formulae from Theorem 7.14, we obtain:

$$\sin(30^\circ) = 3a - 4a^3$$
$$\cos(30^\circ) = 4b^3 - 3b$$
$$\tan(30^\circ) = \frac{3c - c^3}{1 - 3c^2}$$

Thus, we are led indeed to the formulae in the statement. □

Moving on, let us see now what happens for a sum of 4 angles. In view of Theorem 7.13, we do not really want to deal with the sine and cosine, where the formulae will be most likely quite complicated, so we will focus on the tangent instead. We have:

THEOREM 7.16. *The tangents of sums of* 4 *angles are given by*

$$\tan(x+y+z+t) = \frac{\begin{pmatrix} \tan x + \tan y + \tan z + \tan t - \tan x \tan y \tan z \\ - \tan x \tan y \tan t - \tan x \tan z \tan t - \tan y \tan z \tan t \end{pmatrix}}{\begin{pmatrix} 1 - \tan x \tan y - \tan x \tan z - \tan x \tan t - \tan y \tan z \\ - \tan y \tan t - \tan z \tan t + \tan x \tan y \tan z \tan t \end{pmatrix}}$$

provided of course that the denominator is nonzero.

PROOF. We have indeed the following computation:

$$\begin{aligned} & \tan(x+y+z+t) \\ = \ & \frac{\tan(x+y) + \tan(z+t)}{1 - \tan(x+y)\tan(z+t)} \\ = \ & \frac{\frac{\tan x + \tan y}{1 - \tan x \tan y} + \frac{\tan z + \tan t}{1 - \tan z \tan t}}{1 - \frac{\tan x + \tan y}{1 - \tan x \tan y} \cdot \frac{\tan z + \tan t}{1 - \tan z \tan t}} \\ = \ & \frac{\begin{pmatrix} \tan x + \tan y + \tan z + \tan t - \tan x \tan y \tan z \\ - \tan x \tan y \tan t - \tan x \tan z \tan t - \tan y \tan z \tan t \end{pmatrix}}{\begin{pmatrix} 1 - \tan x \tan y - \tan x \tan z - \tan x \tan t - \tan y \tan z \\ - \tan y \tan t - \tan z \tan t + \tan x \tan y \tan z \tan t \end{pmatrix}} \end{aligned}$$

Thus, we are led to the formula in the statement. □

And the problem is now, is what we found in Theorem 7.16 good news, or not? You would probably say no way, but in answer, here is something quite nice:

THEOREM 7.17. *The tangents of sums of angles are given by*

$$\tan(x+y)=\frac{a+b}{1-ab}$$

$$\tan(x+y+z)=\frac{a+b+c-abc}{1-ab-ac-bc}$$

$$\tan(x+y+z+t)=\frac{a+b+c+d-abc-abd-acd-bcd}{1-ab-ac-ad-bc-bd-cd+abcd}$$

$$\vdots$$

where $a=\tan x$, $b=\tan y$, $c=\tan z$, $d=\tan t$, $\ldots$, *with on top odd symmetric functions of* $a,b,c,d,\ldots$, *and on the bottom even symmetric functions of* $a,b,c,d,\ldots$

PROOF. Here the formulae in the statement are those from chapter 3 and from Theorems 7.13 and 7.16, and the conclusion at the end is something quite self-explanatory. We will leave some thinking and further exploration here as an interesting exercise. □

Switching topics now, and back to $\sin,\cos$, we have seen in the above that for small $k\in\mathbb{N}$ we have formulae as follows, with P_k,Q_k being certain polynomials:

$$\cos(kt)=P_k(\cos t)\quad,\quad \sin((k+1)t)=Q_k(\cos t)\sin t$$

To be more precise, in what regards the cosine, we have the following formulae:

$$\cos(2t)=2\cos^2t-1$$

$$\cos(3t)=4\cos^3t-3\cos t$$

$$\vdots$$

As for the sine, the formulae here, coming from what we know, are as follows:

$$\sin(2t)=2\cos t\sin t$$

$$\sin(3t)=(4\cos^2t-1)\sin t$$

$$\vdots$$

To be more precise, in what regards the formula of $\sin(3t)$, we have indeed:

$$\begin{aligned}\sin(3t)&=3\sin t-4\sin^3t\\&=(3-4\sin^2t)\sin t\\&=(3-4+4\cos^2t)\sin t\\&=(4\cos^2t-1)\sin t\end{aligned}$$

In order to see now if our conjecture regarding P_k,Q_k is true, let us compute as well the sine and cosine of $4t$. We have here the following result, confirming our conjecture:

PROPOSITION 7.18. *We have the following formulae,*

$$\cos(4t) = 8\cos^4 t - 8\cos^2 t + 1$$
$$\sin(4t) = (8\cos^3 t - 4\cos t)\sin t$$

confirming our conjectures $\cos(kt) = P_k(\cos t)$ *and* $\sin((k+1)t) = Q_k(\cos t)\sin t$.

PROOF. Regarding the cosine, we have the following computation:

$$\begin{aligned}\cos(4t) &= 2\cos^2(2t) - 1\\ &= 2(2\cos^2 t - 1)^2 - 1\\ &= 2(4\cos^4 t - 4\cos^2 t + 1) - 1\\ &= 8\cos^4 t - 8\cos^2 t + 1\end{aligned}$$

Regarding the sine, we have the following computation:

$$\begin{aligned}\sin(4t) &= 2\sin(2t)\cos(2t)\\ &= 4\sin t\cos t(2\cos^2 t - 1)\\ &= (8\cos^3 t - 4\cos t)\sin t\end{aligned}$$

Thus, we are led to the conclusions in the statement. □

In general now, we can proceed by recurrence, and we obtain:

THEOREM 7.19. *The cosines and sines of multiplied angles are given by*

$$\cos(kt) = P_k(\cos t) \quad , \quad \sin((k+1)t) = Q_k(\cos t)\sin t$$

with P_k, Q_k *being certain polynomials with integer coefficients, given by*

$$P_{k+1}(x) = P_k(x)x - Q_{k-1}(x)(1-x^2)$$
$$Q_k(x) = Q_{k-1}(x)x + P_k(x)$$

called Chebycheff polynomials of the first and second kind.

PROOF. This is indeed something very standard, the idea being as follows:

(1) We use our basic formulae for the sums, which are as follows:

$$\cos(x+y) = \cos x\cos y - \sin x\sin y$$
$$\sin(x+y) = \sin x\cos y + \cos x\sin y$$

We conclude that we have the following formulae, valid for any $k \in \mathbb{N}$:

$$\cos((k+1)t) = \cos(kt)\cos t - \sin(kt)\sin t$$
$$\sin((k+1)t) = \sin(kt)\cos t + \cos(kt)\sin t$$

Now by recurrence, these formulae take the following form:

$$\cos((k+1)t) = P_k(\cos t)\cos t - Q_{k-1}(\cos t)\sin^2 t$$
$$\sin((k+1)t) = Q_{k-1}(\cos t)\sin t\cos t + P_k(\cos t)\sin t$$

We can write these latter formulae in a more convenient way, as follows:

$$\cos((k+1)t) = P_k(\cos t)\cos t - Q_{k-1}(\cos t)(1-\cos^2 t)$$

$$\sin((k+1)t) = (Q_{k-1}(\cos t)\cos t + P_k(\cos t))\sin t$$

Thus, we have the formulae in the statement, with P_k, Q_k being as follows:

$$P_{k+1}(x) = P_k(x)x - Q_{k-1}(x)(1-x^2)$$

$$Q_k(x) = Q_{k-1}(x)x + P_k(x)$$

Observe in particular that both P_k, Q_k much have integer coefficients.

(2) Let us do as well some numerics, as a matter of doublechecking what we found. As input for our computations, we have the following initial values:

$$P_0 = 1 \quad , \quad P_1 = x \quad , \quad Q_0 = 1$$

At the first step of our recurrence we obtain the following formulae:

$$P_2 = 2x^2 - (1-x^2) = 2x^2 - 1$$

$$Q_1 = x + x = 2x$$

At the second step of our recurrence we obtain the following formulae:

$$P_3 = (2x^3 - x) - (2x - 2x^3) = 4x^3 - 3x$$

$$Q_2 = 2x^2 + (2x^2 - 1) = 4x^2 - 1$$

At the third step of our recurrence we obtain the following formulae:

$$P_4 = (4x^4 - 3x^2) - (4x^2 - 1)(1 - x^2) = 8x^4 - 8x^2 + 1$$

$$Q_3 = (4x^3 - x) + (4x^3 - 3x) = 8x^3 - 4x$$

And, good news, this agrees with what we found in Proposition 7.18, and before. □

For future reference, let us record now the above numerics, along with some more:

THEOREM 7.20. *The Chebycheff polynomials of the first kind are*

$$1 \ , \quad x \ , \quad 2x^2 - 1 \ , \quad 4x^3 - 3x \ , \quad 8x^4 - 8x^2 + 1 \ , \quad 16x^5 - 20x^3 + 5x \ , \quad \ldots$$

and the Chebycheff polynomials of the second kind are

$$1 \ , \quad 2x \ , \quad 4x^2 - 1 \ , \quad 8x^3 - 4x \ , \quad 16x^4 - 12x^2 + 1 \ , \quad 32x^5 - 32x^3 + 6x \ , \quad \ldots$$

and this list can be indefinitely enlarged, by recurrence, when needed.

PROOF. Here the formulae of P_0, P_1, P_2, P_3, P_4 and Q_0, Q_1, Q_2, Q_3 are those found above, and those of P_5 and Q_4, Q_5 can be found similarly, by recurrence. □

Many other things can be said, as a continuation of this. We will be back to the Chebycheff polynomials on several occasions, in what follows, with more about them.

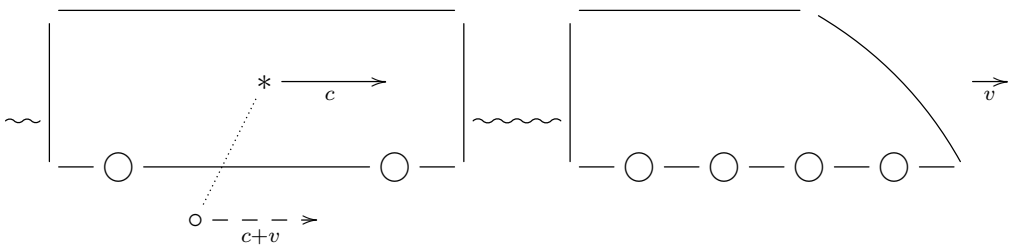
7d. Hyperbolic functions

Ready for some physics? We would like to talk now about the hyperbolic functions, which appear for instance in Einstein's relativity theory. Let us start with:

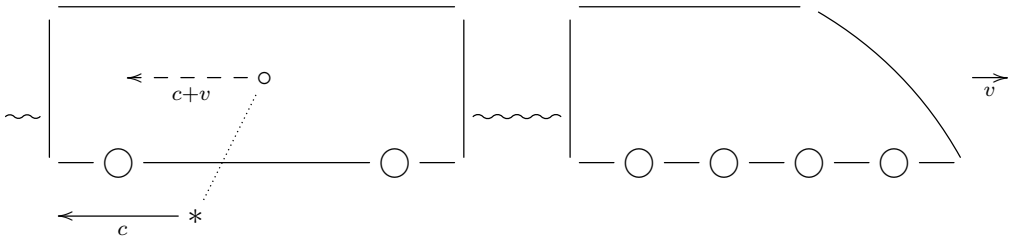
FACT 7.21 (Einstein principles). *The following happen:*

(1) *Light travels in vacuum at a finite speed,* $c < \infty$.
(2) *This speed* c *is the same for all inertial observers.*
(3) *In non-vacuum, the light speed is lower,* $v < c$.
(4) *Nothing can travel faster than light,* $v \not> c$.

The point now is that, obviously, something is wrong here. Indeed, assuming for instance that we have a train, running in vacuum at speed $v > 0$, and someone on board lights a flashlight $*$ towards the locomotive, then an observer $\circ$ on the ground will see the light traveling at speed $c + v > c$, which is a contradiction:



Equivalently, with the same train running, in vacuum at speed $v > 0$, if the observer on the ground lights a flashlight $*$ towards the back of the train, then viewed from the train, that light will travel at speed $c + v > c$, which is a contradiction again:



Summarizing, Fact 7.21 implies $c + v = c$, so contradicts classical mechanics, which therefore needs a fix. By dividing all speeds by c, as to have $c = 1$, and by restricting the attention to the 1D case, to start with, we are led to the following puzzle:

PUZZLE 7.22. *How to define speed addition on the space of* 1D *speeds, which is*

$$I = [-1, 1]$$

with our $c = 1$ *convention, as to have* $1 + c = 1$, *as required by physics?*

In view of our geometric knowledge so far, a natural idea here would be that of wrapping $[-1,1]$ into a circle, and then stereographically projecting on $\mathbb{R}$. Indeed, we can then "import" to $[-,1,1]$ the usual addition on $\mathbb{R}$, via the inverse of this map. So, let us see where all this leads us. First, the formula of our map is as follows:

THEOREM 7.23. *The map wrapping $[-1,1]$ into the unit circle, and then stereographically projecting on $\mathbb{R}$ is given by the formula*

$$\varphi(u)=\tan\left(\frac{\pi u}{2}\right)$$

with the convention that our wrapping is the most straightforward one, making correspond $\pm1\to i$, with negatives on the left, and positives on the right.

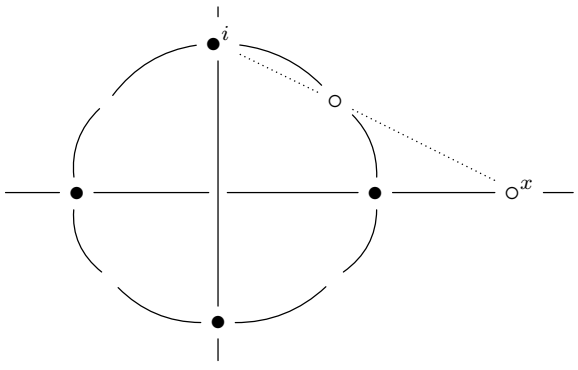
PROOF. Regarding the wrapping, as indicated, this is given by:

$$u\to e^{it}\quad,\quad t=\pi u-\frac{\pi}{2}$$

Indeed, this correspondence wraps $[-1,1]$ as above, the basic instances of our correspondence being as follows, and with everything being fine modulo 2π:

$$-1\to\frac{\pi}{2}\quad,\quad -\frac{1}{2}\to-\pi\quad,\quad 0\to-\frac{\pi}{2}\quad,\quad \frac{1}{2}\to 0\quad,\quad 1\to\frac{\pi}{2}$$

Regarding now the stereographic projection, the picture here is as follows:



Thus, by Thales, the formula of the stereographic projection is as follows:

$$\frac{\cos t}{x}=\frac{1-\sin t}{1}\implies x=\frac{\cos t}{1-\sin t}$$

Now if we compose our wrapping operation above with the stereographic projection, what we get is, via the above Thales formula, and some trigonometry:

$$\begin{aligned}
x &= \frac{\cos t}{1-\sin t}\\
&= \frac{\cos\left(\pi u-\frac{\pi}{2}\right)}{1-\sin\left(\pi u-\frac{\pi}{2}\right)}\\
&= \frac{\cos\left(\frac{\pi}{2}-\pi u\right)}{1+\sin\left(\frac{\pi}{2}-\pi u\right)}\\
&= \frac{\sin(\pi u)}{1+\cos(\pi u)}\\
&= \frac{2\sin\left(\frac{\pi u}{2}\right)\cos\left(\frac{\pi u}{2}\right)}{2\cos^2\left(\frac{\pi u}{2}\right)}\\
&= \tan\left(\frac{\pi u}{2}\right)
\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

The above result is very nice, but when it comes to physics, things do not work, for instance because of the wrong slope of the function $\varphi(u)=\tan\left(\frac{\pi u}{2}\right)$ at the origin, which makes our summing on $[-1,1]$ not compatible with the Galileo addition, at low speeds.

So, what to do? Obviously, trash Theorem 7.23, and start all over again. Getting back now to Puzzle 7.22, this has in fact a simpler solution, based this time on algebra, and which in addition is the good, physically correct solution, as follows:

THEOREM 7.24. *If we sum the speeds according to the Einstein formula*

$$u+_e v=\frac{u+v}{1+uv}$$

then the Galileo formula still holds, approximately, at low speeds

$$u+_e v\simeq u+v$$

and if we have $u=1$ or $v=1$, the resulting sum is $u+_e v=1$.

PROOF. All this is self-explanatory, and clear from definitions, and with the Einstein formula of $u+_e v$ itself being just an obvious solution to Puzzle 7.22, provided that, importantly, we know 0 geometry, and rely on very basic algebra only. □

So, very nice, problem solved, at least in 1D. But, shall we give up with geometry, and the stereographic projection? Certainly not, let us try to recycle that material. In

order to do this, let us recall that the usual trigonometric functions are given by:

$$\sin x = \frac{e^{ix} - e^{-ix}}{2i} \quad , \quad \cos x = \frac{e^{ix} + e^{-ix}}{2} \quad , \quad \tan x = \frac{e^{ix} - e^{-ix}}{i(e^{ix} + e^{-ix})}$$

The point now is that, mathematically speaking, the above functions have some natural "hyperbolic" or "imaginary" analogues, constructed as follows:

$$\sinh x = \frac{e^{x} - e^{-x}}{2} \quad , \quad \cosh x = \frac{e^{x} + e^{-x}}{2} \quad , \quad \tanh x = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

But the function on the right, tanh, starts reminding the formula of Einstein addition, from Theorem 7.24. So, we have our idea, and we are led to the following result:

THEOREM 7.25. *The Einstein speed summation in 1D is given by*

$$\tanh x +_e \tanh y = \tanh(x + y)$$

with $\tanh : [-\infty, \infty] \to [-1, 1]$ *being the hyperbolic tangent function.*

PROOF. This follows by putting together our various formulae above, but it is perhaps better, for clarity, to prove this directly. Our claim is that we have:

$$\tanh(x + y) = \frac{\tanh x + \tanh y}{1 + \tanh x \tanh y}$$

But this can be checked via direct computation, from the definitions, as follows:

$$\begin{aligned}
\frac{\tanh x + \tanh y}{1 + \tanh x \tanh y} &= \left(\frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} + \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}}\right) \Big/ \left(1 + \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \cdot \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}}\right) \\
&= \frac{(e^{x} - e^{-x})(e^{y} + e^{-y}) + (e^{x} + e^{-x})(e^{y} - e^{-y})}{(e^{x} + e^{-x})(e^{y} + e^{-y}) + (e^{x} - e^{-x})(e^{y} + e^{-y})} \\
&= \frac{2(e^{x+y} - e^{-x-y})}{2(e^{x+y} + e^{-x-y})} \\
&= \tanh(x + y)
\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Very nice all this, hope you agree. As a conclusion, passing from the Riemann stereographic projection sum to the Einstein summation basically amounts in replacing:

$$\tan \to \tanh$$

Let us formulate as well this finding more philosophically, as follows:

CONCLUSION 7.26. *The Einstein speed summation in 1D is the imaginary analogue of the summation on* $[-1, 1]$ *obtained via the Riemann stereographic projection.*

As a continuation of this, many other things can be said about relativity, with the next obvious challenge being that of understanding what happens to the Einstein summation formula when passing to 3D. And here, the summation formula is as follows, making appear the vector products $\times$ that we already talked about, in chapter 3:

$$u +_e v = \frac{1}{1+ < u, v >} \left(u + v + \frac{u \times (u \times v)}{1 + \sqrt{1 - ||u||^2}} \right)$$

This being said, let us be reasonable. As mathematicians, we definitely have good reasons for adopting sinh, cosh, tanh, as trigonometric functions. But then we can talk about secondary and inverse functions too, in the obvious way, which leads us to:

CONCLUSION 7.27 (mathematics). *There are in fact* 24 *trigonometric functions,*

sin	cos	tan	sec	csc	cot
arcsin	arccos	arctan	arcsec	arccsc	arccot
sinh	cosh	tanh	sech	csch	coth
arcsinh	arccosh	arctanh	arcsech	arccsch	arccoth

with the hyperbolic ones being useful in relativity, and perhaps in other physics too.

We will be back to these hyperbolic functions, on several occasions. In the meantime, as a good exercise, you can draw the graphs of the $6 + 6$ above hyperbolic functions.

7e. Exercises

This was a special chapter, on special functions, and as exercises here, we have:

EXERCISE 7.28. *Do the missing combinatorics for the Catalan numbers C_k.*

EXERCISE 7.29. *Do as well the combinatorics for the central binomials D_k.*

EXERCISE 7.30. *Prove the generalized binomial formula at any $p \in \mathbb{Z}/2$.*

EXERCISE 7.31. *Meditate at the domains of inverse trigonometric functions.*

EXERCISE 7.32. *Learn more about Chebycheff polynomials, and related topics.*

EXERCISE 7.33. *Study in detail* sinh*,* cosh*,* tanh*, and draw their graphs.*

EXERCISE 7.34. *Study as well* sech*,* csch*,* coth*, and draw their graphs.*

EXERCISE 7.35. *Discuss* arcsinh*,* arccosh*,* arctanh*,* arcsech*,* arccsch*,* arccoth*.*

As bonus exercise, and no surprise here, learn some systematic relativity theory.

CHAPTER 8

Polynomials, again

8a. Multiple roots

Time to end the present first half of this book, which was introductory to functions, before getting into more specialized and modern theory, involving derivatives and integrals, and we will talk here about the most basic functions of them all, the polynomials.

We already know a bit about the polynomials $P \in \mathbb{R}[X]$ from chapter 2. However, remember from chapter 3 that any such polynomial can be regarded as a complex polynomial, $P \in \mathbb{C}[X]$, and with this being a great thing, because any complex polynomial $P \in \mathbb{C}[X]$ has a full collection of $\deg P$ roots, when counted with multiplicities.

So, it is about complex polynomials $P \in \mathbb{C}[X]$ that we will talk about, here. Let us start with a quick remake of the general theory from chapter 2, which was mostly of algebraic nature, in the complex case. At the beginning of everything, we have:

PROPOSITION 8.1. *For a polynomial $P \in \mathbb{C}[X]$ and a number $r \in \mathbb{C}$, the following conditions are equivalent:*

(1) $P(r) = 0$.
(2) $P(x) = (x - r)Q$, *with* $Q \in \mathbb{C}[X]$.

PROOF. The point here is that we can divide the polynomials, a bit as we divide the integers, with an illustration for the division algorithm being as follows:

$$\begin{aligned}
& x^3 + 1 = (x + 2) * + * \\
\implies \quad & x^3 + 1 = (x + 2)(x^2 + *) + * \\
\implies \quad & x^3 + 1 = (x + 2)(x^2 - 2x + *) + * \\
\implies \quad & x^3 + 1 = (x + 2)(x^2 - 2x + 4) + * \\
\implies \quad & x^3 + 1 = (x + 2)(x^2 - 2x + 4) - 7
\end{aligned}$$

Now by dividing P by $x - r$ we are led to a formula as follows, with the quotient being a certain polynomial $Q \in \mathbb{C}[X]$, and the remainder being a constant $c \in \mathbb{C}$:

$$P(x) = (x - r)Q + c$$

But with this in hand, the equivalence in the statement is clear, by taking $x = r$. □

Next, we have the following key result, called fundamental theorem of algebra:

THEOREM 8.2. *Any polynomial $P \in \mathbb{C}[X]$ can be written as*

$$P(x) = c(x - r_1)^{n_1} \dots (x - r_k)^{n_k}$$

with $r_1, \dots, r_k \in \mathbb{R}$ being the roots, $n_1, \dots, n_k \in \mathbb{N}$, and $c \in \mathbb{C}$.

PROOF. This is something quite tricky, coming in two steps, as follows:

(1) To start with, we can apply Proposition 8.1 iteratively, and we obtain a certain factor $\prod_i (x - r_i)^{n_i}$ which grows over the time, until it has to stop, due to the fact that the remainder $Q \in \mathbb{C}[X]$ has no roots. Thus, we have a decomposition as follows:

$$P(x) = (x - r_1)^{n_1} \dots (x - r_k)^{n_k} Q(x)$$

(2) On the other hand, as explained in chapter 3, some basic complex analysis shows that any polynomial having degree ≥ 1 must have a root. Thus the above remainder $Q \in \mathbb{C}[X]$ must be a polynomial of degree zero, $Q(x) = c \in \mathbb{C}$, as desired. □

Let us record as well the following result, useful for dealing with the roots:

THEOREM 8.3. *Given a polynomial $P \in \mathbb{C}[X]$, with leading coefficient 1,*

$$P(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$$

consider its n roots $r_1, \dots, r_n$, counted with multiplicities, so that we have:

$$P(x) = (x - r_1) \dots (x - r_n)$$

These roots satisfy then the formulae $\sum_i r_i = -a_{n-1}$ and $\prod_i r_i = (-1)^n a_0$.

PROOF. This is clear indeed from the formula $P(x) = (x-r_1) \dots (x-r_n)$, by expanding the product, and identifying the terms of degree $n-1$, and of degree 0. □

Finally, we have our beloved result regarding degree 2 polynomials, as follows:

THEOREM 8.4. *The solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{C}$ are*

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

with the square root of complex numbers being defined as $\sqrt{re^{it}} = \sqrt{r}e^{it/2}$.

PROOF. We can indeed write our equation in the following way:

$$\begin{aligned} ax^2 + bx + c = 0 &\iff \left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2} \\ &\iff x + \frac{b}{2a} = \pm\frac{\sqrt{b^2 - 4ac}}{2a} \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

What is next? Many things, presumably in relation with the extension of Theorem 8.4 to higher degree, and here is an interesting question, that we can try to solve:

QUESTION 8.5. *What is the analogue of $\Delta = b^2 - 4ac$ in higher degree? Also, what is the analogue of the condition $\Delta = 0$, in higher degree?*

In answer, you would probably say go with the first question, what is Δ, and then $\Delta = 0$ will come for free. However, with the first question being quite difficult, and you can have some fun here with degree 3 polynomials, in order to understand what I am talking about, we will go instead with the second question, reformulated as follows:

QUESTION 8.6 (update). *Given $P \in \mathbb{C}[X]$, can we decide, via some algebraic trick, if P has or not double roots, without actually computing the roots?*

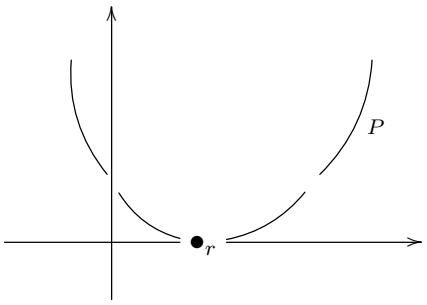
Which does not look trivial at all, but the point is that some magic happens here, no one really understands why, and we have the following surprising answer, to this:

ANSWER 8.7. *Yes, algebraic tricks allow us to decide if $P \in \mathbb{C}[X]$ has multiple roots. And even more, in certain cases, to explicitly compute these multiple roots.*

And it is the second assertion in the above which is the crazy thing, because computing roots of high degree polynomials is reputed to be a hopeless business. Amazing.

Getting to work now, what can be the algebraic tricks mentioned above? And here, as a further twist to the plot, these algebraic tricks actually come from analysis:

PRINCIPLE 8.8. *Given a polynomial $P \in \mathbb{R}[X]$, a multiple root $r \in \mathbb{R}$, corresponding to a formula of type $P(x) = (x - r)^2 Q(x)$, can be identified on the graph of P,*



due to the fact that this graph must be tangent to Ox at this point $r \in \mathbb{R}$, as indicated. Moreover, the same formally happens for complex polynomials, $P \in \mathbb{C}[X]$.

So, getting back to analysis. In order to decide now what is tangent to Ox and what is not, the simplest is to compute the slope of the graph. And here, for the simplest polynomials, $P(x) = x^n$, this is easily done using the binomial formula, which gives:

$$(x + t)^n \simeq x^n + nx^{n-1}t$$

Summarizing, problem solved for $P(x) = x^n$, whose slope at a given $x \in \mathbb{R}$ is the quantity $P'(x) = nx^{n-1}$. But now, with this in hand, we can forget this analysis intermezzo, and start doing the algebra, in an independent and rigorous way, as follows:

THEOREM 8.9. *We can formally differentiate the polynomials, according to*

$$(x^n)' = nx^{n-1}$$

and to the following linearity rules, allowing to pass to linear combinations:

$$(P+Q)' = P' + Q' \quad , \quad (\lambda P)' = \lambda P'$$

This differentiation operation satisfies the following rules,

$$(PQ)' = P'Q + PQ' \quad , \quad (P \circ Q)' = P'(Q)Q'$$

called Leibnitz rule for products, and chain derivative rule.

PROOF. This is indeed something standard, the idea being as follows:

(1) To start with, we can certainly differentiate the polynomials according to the recipe in the statement, with the precise general formula being as follows:

$$\begin{aligned} & P = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 \\ \implies \quad & P' = na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \ldots + a_1 \end{aligned}$$

(2) In what regards the Leibnitz rule, by linearity we can assume that we are dealing with monomials, $P = x^m$ and $Q = x^n$. But here, the Leibnitz rule comes from:

$$\begin{aligned} (x^{m+n})' &= (m+n)x^{m+n-1} \\ &= mx^{m-1}x^n + nx^m x^{n-1} \\ &= (x^m)'x^n + x^m(x^n)' \end{aligned}$$

(3) Regarding now the chain rule, again by linearity we can assume that we have $P = x^m$ and $Q = x^n$. And here, the result comes via the following computation:

$$\begin{aligned} (x^{mn})' &= mnx^{mn-1} \\ &= mx^{mn-n} \cdot nx^{n-1} \\ &= [(mx^{m-1}) \circ x^n] \cdot nx^{n-1} \\ &= [(x^m)' \circ x^n] \cdot (x^n)' \end{aligned}$$

(4) Finally, and we insist here, in case you might already know a bit above derivatives, in general, all the above is just pure algebra, and of quite trivial type. Of course there is an analytic interpretation of all this, that we will discuss later, in chapter 9. For our purposes here, in the present chapter 8, we will not need that analytic interpretation. □

As an application now of our derivatives, as introduced above, we have:

THEOREM 8.10. *Given a polynomial $P \in \mathbb{C}[X]$, the following happen:*

(1) *Any multiple root of P must be a root of P'.*
(2) *In fact, the multiple roots of P are the common roots of P, P'.*
(3) *If $P(r) = 0$ with multiplicity k, then $P'(r) = 0$ with multiplicity $k-1$.*

PROOF. This is something quite magic, the idea being as follows:

(1) We have indeed the following computation, based on the general differentiation rules from Theorem 8.9, and more specifically, on the Leibnitz rule there:

$$\begin{aligned}[(x-r)^2Q]' &= [(x-r)^2]'Q+(x-r)^2Q'\\ &= 2(x-r)Q+(x-r)^2Q'\\ &= (x-r)(2Q+(x-r)Q')\end{aligned}$$

Here we have used the following formula, which is something trivial:

$$[(x-r)^2]'=2(x-r)$$

But with $P=(x-r)^2Q$, this leads to the conclusion in the statement.

(2) We know this in one sense from (1). In the other sense, assume that:

$$P(r)=P'(r)=0$$

Now let us divide P by $(x-r)^2$. This must give a formula as follows:

$$P=(x-r)^2Q+c(x-r)$$

By using now the computation in (1), we can see that $P'(r)=0$ amounts in saying that $(c(x-r))'$ vanishes at r, so that $c=0$. Thus, $P=(x-r)^2Q$, as desired.

(3) We have indeed the following computation, generalizing the one in (1):

$$\begin{aligned}[(x-r)^kQ]' &= [(x-r)^k]'Q+(x-r)^kQ'\\ &= k(x-r)^{k-1}Q+(x-r)^kQ'\\ &= (x-r)^{k-1}(kQ+(x-r)Q')\end{aligned}$$

Here we have used the following formula, coming from the chain rule:

$$[(x-r)^k]'=k(x-r)^{k-1}$$

Thus, with $P=(x-r)^kQ$, we are led to the conclusion in the statement. □

The above result is something quite amazing, raising the possibility of deciding if P has multiple roots, without computing the roots in question. Indeed, for this purpose:

– To start with, we must compute P', with this being something quickly done.

– Then we must successively perform the division algorithm for P,P', a bit like for the usual integers, as to compute the greatest common divisor:

$$D=(P,P')$$

– And then, getting now to conclusions, if this common divisor D has degree ≥ 1, this means that our original polynomial P must have a double root.

Let us summarize this remarkable finding, which answers our original Question 8.6, majestically, along with a bit more, in the following way:

THEOREM 8.11. *Given a polynomial $P \in \mathbb{C}[X]$, compute P', and perform the division algorithm for P, P', as to get to the greatest common divisor $D = (P, P')$.*

(1) *P has multiple roots precisely when* $\deg D \geq 1$.
(2) *In fact, the multiple roots of P are precisely the roots of D.*
(3) *Moreover, via $P \to D$, all root multiplicities get lowered by* 1.

PROOF. This follows indeed as indicated above. To be more precise, assume that P factorizes as follows, with r_i being its multiple roots, with multiplicities $n_i \geq 2$:

$$P(x) = (x - r_1)^{n_1} \ldots (x - r_k)^{n_k} Q$$

According to Theorem 8.10, the polynomial P' is then of the following form:

$$P'(x) = (x - r_1)^{n_1 - 1} \ldots (x - r_k)^{n_k - 1} R$$

Thus, the common divisor $D = (P, P')$ is given by the following formula:

$$D(x) = (x - r_1)^{n_1 - 1} \ldots (x - r_k)^{n_k - 1}$$

But this leads to the various conclusions in the statement. □

Summarizing, Question 8.6 solved. In relation now with our last claim from Answer 8.7, which is perhaps even more amazing, let us record as well:

THEOREM 8.12 (continuation). *In the above context, with $D = (P, P')$ computed, assuming that we can factorize D, which is something that we can do when*

(1) *D has degree* 1 *ot* 2*, as we know well.*
(2) *D has degree* 3 *or* 4*, as we will soon discover.*
(3) *D has multiple roots, and by recursion we can factorize it.*
(4) *D has various other special features, allowing us to factorize it.*

our method computes in practice all the multiple roots of P.

PROOF. This is something self-explanatory, based on Theorem 8.11, and coming with the warning that in (2), things are in fact a bit more complicated, as we will soon discover, and coming as well with the comment that, in relation with (3), there are more things that can be said, and we will leave the continuation there as an interesting exercise. □

8b. The discriminant

Moving on, we would like to understand now what the analogue of the discriminant $\Delta = b^2 - 4ac$ is. But this is something quite tricky, because we would like to have $\Delta = 0$ precisely when $(P, P') \neq 1$, which leads us into the question of deciding, given two polynomials $P, Q \in \mathbb{C}[X]$, if these polynomials have a common root, $(P, Q) \neq 1$, or not.

Fortunately this latter question has a nice answer. We will need:

THEOREM 8.13. *Given a monic polynomial $P \in \mathbb{C}[X]$, factorized as*

$$P = (X - a_1) \ldots (X - a_k)$$

the following happen:

(1) *The coefficients of P are symmetric functions in $a_1, \ldots, a_k$.*
(2) *The symmetric functions in $a_1, \ldots, a_k$ are polynomials in the coefficients of P.*

PROOF. This is something very standard, the idea being as follows:

(1) By expanding our polynomial, we have the following formula:

$$P = \sum_{r=0}^{k} (-1)^r \sum_{i_1 < \ldots < i_r} a_{i_1} \ldots a_{i_r} \cdot X^{k-r}$$

Thus the coefficients of P are, up to some signs, the following functions:

$$f_r = \sum_{i_1 < \ldots < i_r} a_{i_1} \ldots a_{i_r}$$

But these are indeed symmetric functions in $a_1, \ldots, a_k$, as claimed.

(2) Conversely now, let us look at the symmetric functions in the roots $a_1, \ldots, a_k$. These appear as linear combinations of the basic symmetric functions, given by:

$$S_r = \sum_i a_i^r$$

Moreover, when allowing polynomials instead of linear combinations, we need in fact only the first k such sums, namely $S_1, \ldots, S_k$. That is, the symmetric functions $\mathcal{F}$ in our variables $a_1, \ldots, a_k$, with integer coefficients, appear as follows:

$$\mathcal{F} = \mathbb{Z}[S_1, \ldots, S_k]$$

(3) The point now is that, alternatively, the symmetric functions in our variables $a_1, \ldots, a_k$ appear as well as linear combinations of the functions f_r that we found in (1), and that when allowing polynomials instead of linear combinations, we need in fact only the first k functions, namely $f_1, \ldots, f_k$. That is, we have as well:

$$\mathcal{F} = \mathbb{Z}[f_1, \ldots, f_k]$$

But this gives the result, because we can pass from $\{S_r\}$ to $\{f_r\}$, and vice versa.

(4) This was for the idea, and in practice now up to you to clarify all the details. In fact, we will also need in what follows the extension of all this to the case where P is no longer assumed to be monic, and with this being, again, a good exercise for you. □

Getting back now to our original question, namely that of deciding whether two polynomials $P, Q \in \mathbb{C}[X]$ have a common root or not, this has the following nice answer:

THEOREM 8.14. *Given two polynomials $P, Q \in \mathbb{C}[X]$, written as*

$$P = c(X - a_1) \ldots (X - a_k) \quad , \quad Q = d(X - b_1) \ldots (X - b_l)$$

the following quantity, which is called resultant of P, Q,

$$R(P, Q) = c^l d^k \prod_{ij} (a_i - b_j)$$

is a certain polynomial in the coefficients of P, Q, with integer coefficients, and we have $R(P, Q) = 0$ precisely when P, Q have a common root.

PROOF. This is something quite tricky, the idea being as follows:

(1) Given two polynomials $P, Q \in \mathbb{C}[X]$, we can certainly construct the quantity $R(P, Q)$ in the statement, with the role of the normalization factor $c^l d^k$ to become clear later on, and then we have $R(P, Q) = 0$ precisely when P, Q have a common root:

$$R(P, Q) = 0 \iff \exists i, j, a_i = b_j$$

(2) As bad news, however, this quantity $R(P, Q)$, defined in this way, is a priori not very useful in practice, because it depends on the roots a_i, b_j of our polynomials P, Q, that we cannot compute in general. However, and here comes our point, as we will prove below, it turns out that $R(P, Q)$ is in fact a polynomial in the coefficients of P, Q, with integer coefficients, and this is where the power of $R(P, Q)$ comes from.

(3) You might perhaps say, nice, but why not doing things the other way around, that is, formulating our theorem with the explicit formula of $R(P, Q)$, in terms of the coefficients of P, Q, and then proving that we have $R(P, Q) = 0$, via roots and everything. Good point, but this is not exactly obvious, the formula of $R(P, Q)$ in terms of the coefficients of P, Q being something quite complicated. In short, trust me, let us prove our theorem as stated, and for alternative formulae of $R(P, Q)$, we will see later.

(4) Getting started now, let us expand the formula of $R(P, Q)$, by making all the multiplications there, abstractly, in our head. Everything being symmetric in $a_1, \ldots, a_k$, we obtain in this way certain symmetric functions in these variables, which will be therefore certain polynomials in the coefficients of P. Moreover, due to our normalization factor c^l, these polynomials in the coefficients of P will have integer coefficients.

(5) With this done, let us look now what happens with respect to the remaining variables $b_1, \ldots, b_l$, which are the roots of Q. Once again what we have here are certain symmetric functions in these variables $b_1, \ldots, b_l$, and these symmetric functions must be certain polynomials in the coefficients of Q. Moreover, due to our normalization factor d^k, these polynomials in the coefficients of Q will have integer coefficients.

(6) Thus, we are led to the conclusion in the statement, that $R(P, Q)$ is a polynomial in the coefficients of P, Q, with integer coefficients, and with the remark that the $c^l d^k$ factor is there for these latter coefficients to be indeed integers, instead of rationals. □

As an illustration, consider a polynomial of degree 2, and one of degree 1:

$$P = ax^2 + bx + c \quad , \quad Q = dx + e$$

In order to compute the resultant, let us factorize our polynomials:

$$P = a(x - p)(x - q) \quad , \quad Q = d(x - r)$$

The resultant can be then computed as follows, by using the method above:

$$\begin{aligned} R(P, Q) &= ad^2(p - r)(q - r) \\ &= ad^2(pq - (p + q)r + r^2) \\ &= cd^2 + bd^2r + ad^2r^2 \\ &= cd^2 - bde + ae^2 \end{aligned}$$

Finally, observe that $R(P, Q) = 0$ corresponds indeed to the fact that P, Q have a common root. Indeed, the root of Q is $r = -e/d$, and we have:

$$P(r) = \frac{ae^2}{d^2} - \frac{be}{d} + c = \frac{R(P, Q)}{d^2}$$

In higher degree the computations are more complicated. To the rescue comes the following result of Sylvester, based on advanced technology, namely determinants:

THEOREM 8.15. *The resultant of two polynomials, written as*

$$P = p_kX^k + \ldots + p_1X + p_0 \quad , \quad Q = q_lX^l + \ldots + q_1X + q_0$$

appears as the determinant of an associated matrix, as follows,

$$R(P, Q) = \begin{vmatrix} p_k & & & q_l & & \\ \vdots & \ddots & & \vdots & \ddots & \\ p_0 & & p_k & q_0 & & q_l \\ & \ddots & \vdots & & \ddots & \vdots \\ & & p_0 & & & q_0 \end{vmatrix}$$

with the matrix having size $k + l$, and having 0 coefficients at the blank spaces.

PROOF. We have not talked yet about linear algebra and determinants in this book, which correspond to functions of several variables, which are scheduled for later, so here is the proof, for what this is worth, assuming some familiarity with this material:

(1) Consider the vector space $\mathbb{C}_k[X]$ formed by the polynomials of degree $< k$:

$$\mathbb{C}_k[X] = \left\{ P \in \mathbb{C}[X] \middle| \deg P < k \right\}$$

This is a vector space of dimension k, having as basis the monomials $1, X, \ldots, X^{k-1}$. Now given polynomials P, Q as in the statement, consider the following linear map:

$$\Phi : \mathbb{C}_l[X] \times \mathbb{C}_k[X] \to \mathbb{C}_{k+l}[X] \quad , \quad (A, B) \to AP + BQ$$

(2) Our first claim is that with respect to the standard bases for all the vector spaces involved, namely those consisting of the monomials $1, X, X^2, \ldots$, the matrix of Φ is the matrix in the statement. But this is something which is clear from definitions.

(3) Our second claim is that $\det \Phi = 0$ happens precisely when P, Q have a common root. Indeed, our polynomials P, Q having a common root means that we can find A, B such that $AP + BQ = 0$, and so that $(A, B) \in \ker \Phi$, which reads $\det \Phi = 0$.

(4) Finally, our claim is that we have $\det \Phi = R(P, Q)$. But this follows from the uniqueness of the resultant, up to a scalar, and with this uniqueness property being elementary to establish, along the lines of the proofs of Theorems 8.13 and 8.14. □

As an illustration for this, consider our favorite polynomials, namely:

$$P = ax^2 + bx + c \quad , \quad Q = dx + e$$

According to the above result, the resultant should be then, as it should:

$$R(P, Q) = \begin{vmatrix} a & d & \\ b & e & d \\ c & & e \end{vmatrix} = ae^2 - bde + cd^2$$

We can go back now to our original question regarding discriminants, and we have:

THEOREM 8.16. *Given a polynomial $P \in \mathbb{C}[X]$, written as*

$$P(X) = aX^N + bX^{N-1} + cX^{N-2} + \ldots$$

its discriminant, defined as being the following quantity,

$$\Delta(P) = \frac{(-1)^{\binom{N}{2}}}{a} R(P, P')$$

is a polynomial in the coefficients of P, with integer coefficients, and $\Delta(P) = 0$ happens precisely when P has a double root.

PROOF. The fact that the discriminant $\Delta(P)$ is a polynomial in the coefficients of P, with integer coefficients, comes from Theorem 8.14, coupled with the fact that the division by the leading coefficient a is indeed possible, under $\mathbb{Z}$, as being shown by:

$$R(P, P') = \begin{vmatrix} a & & & Na & & \\ \vdots & \ddots & & \vdots & \ddots & \\ z & & a & y & & Na \\ & \ddots & \vdots & & \ddots & \vdots \\ & & z & & & y \end{vmatrix}$$

Also, the fact that we have $\Delta(P) = 0$ precisely when P has a double root is clear from Theorem 8.14. Finally, let us mention that the sign $(-1)^{\binom{N}{2}}$ is there for various reasons, including the compatibility with the formula $\Delta(P) = b^2 - 4ac$ in degree 2. □

As an illustration, let us see what happens in degree 2. Here we have:

$$P = aX^2 + bX + c \quad , \quad P' = 2aX + b$$

Thus, the resultant is given by the following formula:

$$\begin{aligned} R(P, P') &= ab^2 - b(2a)b + c(2a)^2 \\ &= 4a^2c - ab^2 \\ &= -a(b^2 - 4ac) \end{aligned}$$

It follows that the discriminant of our polynomial is, as it should:

$$\Delta(P) = b^2 - 4ac$$

Alternatively, we can use the formula in Theorem 8.15, and we obtain:

$$\begin{aligned} \Delta(P) = &= -\frac{1}{a}\begin{vmatrix} a & 2a & \\ b & b & 2a \\ c & & b \end{vmatrix} \\ &= -\begin{vmatrix} 1 & 2 & \\ b & b & 2a \\ c & & b \end{vmatrix} \\ &= -b^2 + 2(b^2 - 2ac) \\ &= b^2 - 4ac \end{aligned}$$

At the theoretical level now, we have the following result, which is not trivial:

THEOREM 8.17. *The discriminant of a polynomial P is given by the formula*

$$\Delta(P) = a^{2N-2} \prod_{i<j} (r_i - r_j)^2$$

where a is the leading coefficient, and $r_1, \ldots, r_N$ are the roots.

PROOF. This is something quite tricky, the idea being as follows:

(1) The first thought goes to the formula in Theorem 8.14, so let us see what that formula teaches us, in the case $Q = P'$. Let us write P, P' as follows:

$$P = a(x - r_1) \ldots (x - r_N)$$

$$P' = Na(x - p_1) \ldots (x - p_{N-1})$$

According to Theorem 8.14, the resultant of P, P' is then given by:

$$R(P, P') = a^{N-1}(Na)^N \prod_{ij} (r_i - p_j)$$

And bad news, this is not exactly what we wished for, namely the formula in the statement. That is, we are on the good way, but certainly have to work some more.

(2) Obviously, we must get rid of the roots $p_1, \ldots, p_{N-1}$ of the polynomial P'. In order to do this, let us rewrite the formula that we found in (1) in the following way:

$$\begin{aligned} R(P,P') &= N^N a^{2N-1} \prod_i \left(\prod_j (r_i - p_j) \right) \\ &= N^N a^{2N-1} \prod_i \frac{P'(r_i)}{Na} \\ &= a^{N-1} \prod_i P'(r_i) \end{aligned}$$

(3) In order to compute now P', and more specifically the values $P'(r_i)$ that we are interested in, we can use the Leibnitz rule. So, consider our polynomial:

$$P(x) = a(x - r_1) \ldots (x - r_N)$$

The Leibnitz rule for derivatives tells us that $(fg)' = f'g + fg'$, but then also that $(fgh)' = f'gh + fg'h + fgh'$, and so on. Thus, for our polynomial, we obtain:

$$P'(x) = a \sum_i (x - r_1) \ldots \underbrace{(x - r_i)}_{missing} \ldots (x - r_N)$$

Now when applying this formula to one of the roots r_i, we obtain:

$$P'(r_i) = a(r_i - r_1) \ldots \underbrace{(r_i - r_i)}_{missing} \ldots (r_i - r_N)$$

By making now the product over all indices i, this gives the following formula:

$$\prod_i P'(r_i) = a^N \prod_{i \neq j} (r_i - r_j)$$

(4) Time now to put everything together. By taking the formula in (2), making the normalizations in Theorem 8.16, and then using the formula found in (3), we obtain:

$$\begin{aligned} \Delta(P) &= (-1)^{\binom{N}{2}} a^{N-2} \prod_i P'(r_i) \\ &= (-1)^{\binom{N}{2}} a^{2N-2} \prod_{i \neq j} (r_i - r_j) \\ &= a^{2N-2} \prod_{i<j} (r_i - r_j)^2 \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

As applications now, the formula in Theorem 8.17 is quite useful for the real polynomials $P \in \mathbb{R}[X]$ in small degree, because it allows to say when the roots are real, or complex, or at least have some partial information about this. For instance, we have:

THEOREM 8.18. *Consider a polynomial with real coefficients, $P \in \mathbb{R}[X]$, assumed for simplicity to have nonzero discriminant, $\Delta \neq 0$.*

(1) *In degree 2, the roots are real when $\Delta > 0$, and complex when $\Delta < 0$.*
(2) *In degree 3, all roots are real precisely when $\Delta > 0$.*

PROOF. This is very standard, the idea being as follows:

(1) The first assertion is something that we certainly know well, since ages, formally coming from Theorem 8.4, but let us see how this comes via the formula in Theorem 8.17. In degree $N = 2$, this formula looks as follows, with r_1, r_2 being the roots:

$$\Delta(P) = a^2(r_1 - r_2)^2$$

Thus $\Delta > 0$ amounts in saying that we have $(r_1 - r_2)^2 > 0$. Now since r_1, r_2 are conjugate, and with this being something trivial, meaning no need here for the computations in Theorem 8.4, we conclude that $\Delta > 0$ means that r_1, r_2 are real, as stated.

(2) In degree $N = 3$ now, we know from analysis that P has at least one real root, and the problem is whether the remaining 2 roots are real, or complex conjugate. For this purpose, we can use the formula in Theorem 8.17, which in degree 3 reads:

$$\Delta(P) = a^4(r_1 - r_2)^2(r_1 - r_3)^2(r_2 - r_3)^2$$

We can see that in the case $r_1, r_2, r_3 \in \mathbb{R}$, we have $\Delta(P) > 0$. Conversely now, assume that $r_1 = r$ is the real root, coming from analysis, and that the other roots are $r_2 = z$ and $r_3 = \bar{z}$, with z being a complex number, which is not real. We have then:

$$\begin{aligned}
\Delta(P) &= a^4(r - z)^2(r - \bar{z})^2(z - \bar{z})^2 \\
&= a^4|r - z|^4(2iIm(z))^2 \\
&= -4a^4|r - z|^4 Im(z)^2 \\
&< 0
\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

In degree 4 a similar result is available. More on this, later in this chapter.

8c. Degree 3 equations

Let us work out now in detail what happens in degree 3, with the explicit computation of the discriminant, in terms of the coefficients. Here the result is as follows:

THEOREM 8.19. *The discriminant of a degree 3 polynomial,*

$$P = aX^3 + bX^2 + cX + d$$

is given by $\Delta(P) = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd$.

PROOF. We have two methods available, based on Theorem 8.14 and Theorem 8.15, and both being instructive, we will try them both. The computations are as follows:

(1) Let us first go the pedestrian way, based on the definition of the resultant, from Theorem 8.14. Consider two polynomials, of degree 3 and degree 2, written as follows:

$$P = aX^3 + bX^2 + cX + d$$

$$Q = eX^2 + fX + g = e(X - s)(X - t)$$

The resultant of these two polynomials is then given by:

$$\begin{aligned} R(P,Q) &= a^2e^3(p-s)(p-t)(q-s)(q-t)(r-s)(r-t) \\ &= a^2 \cdot e(p-s)(p-t) \cdot e(q-s)(q-t) \cdot e(r-s)(r-t) \\ &= a^2Q(p)Q(q)Q(r) \\ &= a^2(ep^2+fp+g)(eq^2+fq+g)(er^2+fr+g) \end{aligned}$$

By expanding, we obtain the following formula for this resultant:

$$\begin{aligned} \frac{R(P,Q)}{a^2} &= e^3p^2q^2r^2 + e^2f(p^2q^2r + p^2qr^2 + pq^2r^2) \\ &+ e^2g(p^2q^2 + p^2r^2 + q^2r^2) + ef^2(p^2qr + pq^2r + pqr^2) \\ &+ efg(p^2q + pq^2 + p^2r + pr^2 + q^2r + qr^2) + f^3pqr \\ &+ eg^2(p^2 + q^2 + r^2) + f^2g(pq + pr + qr) \\ &+ fg^2(p + q + r) + g^3 \end{aligned}$$

Note in passing that we have 27 terms on the right, as we should, and with this kind of check being mandatory, when doing such computations. Next, we have:

$$p + q + r = -\frac{b}{a} \quad , \quad pq + pr + qr = \frac{c}{a} \quad , \quad pqr = -\frac{d}{a}$$

By using these formulae, we can produce some more, as follows:

$$p^2 + q^2 + r^2 = (p + q + r)^2 - 2(pq + pr + qr) = \frac{b^2}{a^2} - \frac{2c}{a}$$

$$p^2q + pq^2 + p^2r + pr^2 + q^2r + qr^2 = (p + q + r)(pq + pr + qr) - 3pqr = -\frac{bc}{a^2} + \frac{3d}{a}$$

$$p^2q^2 + p^2r^2 + q^2r^2 = (pq + pr + qr)^2 - 2pqr(p + q + r) = \frac{c^2}{a^2} - \frac{2bd}{a^2}$$

By plugging now this data into the formula of $R(P,Q)$, we obtain:

$$\begin{aligned}R(P,Q) &= a^2e^3\cdot\frac{d^2}{a^2}-a^2e^2f\cdot\frac{cd}{a^2}+a^2e^2g\left(\frac{c^2}{a^2}-\frac{2bd}{a^2}\right)+a^2ef^2\cdot\frac{bd}{a^2}\\ &+ a^2efg\left(-\frac{bc}{a^2}+\frac{3d}{a}\right)-a^2f^3\cdot\frac{d}{a}\\ &+ a^2eg^2\left(\frac{b^2}{a^2}-\frac{2c}{a}\right)+a^2f^2g\cdot\frac{c}{a}-a^2fg^2\cdot\frac{b}{a}+a^2g^3\end{aligned}$$

Thus, we have the following formula for the resultant:

$$\begin{aligned}R(P,Q) &= d^2e^3-cde^2f+c^2e^2g-2bde^2g+bdef^2-bcefg+3adefg\\ &- adf^3+b^2eg^2-2aceg^2+acf^2g-abfg^2+a^2g^3\end{aligned}$$

Getting back now to our discriminant problem, with $Q=P'$, which corresponds to $e=3a$, $f=2b$, $g=c$, we obtain the following formula:

$$\begin{aligned}R(P,P') &= 27a^3d^2-18a^2bcd+9a^2c^3-18a^2bcd+12ab^3d-6ab^2c^2+18a^2bcd\\ &- 8ab^3d+3ab^2c^2-6a^2c^3+4ab^2c^2-2ab^2c^2+a^2c^3\end{aligned}$$

By simplifying terms, and dividing by a, we obtain the following formula:

$$-\Delta(P)=27a^2d^2-18abcd+4ac^3+4b^3d-b^2c^2$$

But this gives the formula in the statement, namely:

$$\Delta(P)=b^2c^2-4ac^3-4b^3d-27a^2d^2+18abcd$$

(2) Let us see as well how the computation goes, by using Theorem 8.15, which is our most advanced tool, so far. Consider a polynomial of degree 3, and its derivative:

$$P=aX^3+bX^2+cX+d$$

$$P'=3aX^2+2bX+c$$

By using now Theorem 8.15 and computing the determinant, we obtain:

$$\begin{aligned}
R(P,P') &= \begin{vmatrix} a & & 3a & & \\ b & a & 2b & 3a & \\ c & b & c & 2b & 3a \\ d & c & & c & 2b \\ & d & & & c \end{vmatrix} \\
&= \begin{vmatrix} a & & & & \\ b & a & -b & 3a & \\ c & b & -2c & 2b & 3a \\ d & c & -3d & c & 2b \\ & d & & & c \end{vmatrix} \\
&= a\begin{vmatrix} a & -b & 3a & \\ b & -2c & 2b & 3a \\ c & -3d & c & 2b \\ d & & & c \end{vmatrix} \\
&= -ad\begin{vmatrix} -b & 3a & \\ -2c & 2b & 3a \\ -3d & c & 2b \end{vmatrix} + ac\begin{vmatrix} a & -b & 3a \\ b & -2c & 2b \\ c & -3d & c \end{vmatrix} \\
&= -ad(-4b^3 - 27a^2d + 12abc + 3abc) \\
&\quad +ac(-2ac^2 - 2b^2c - 9abd + 6ac^2 + b^2c + 6abd) \\
&= a(4b^3d + 27a^2d^2 - 15abcd + 4ac^3 - b^2c^2 - 3abcd) \\
&= a(4b^3d + 27a^2d^2 - 18abcd + 4ac^3 - b^2c^2)
\end{aligned}$$

Now according to Theorem 8.16, the discriminant of our polynomial is given by:

$$\begin{aligned}
\Delta(P) &= -\frac{R(P,P')}{a} \\
&= -4b^3d - 27a^2d^2 + 18abcd - 4ac^3 + b^2c^2 \\
&= b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd
\end{aligned}$$

Thus, we have again obtained the formula in the statement. □

Still talking degree 3 equations, let us try now to solve such an equation $P = 0$, with $P = aX^3 + bX^2 + cX + d$ as above. By linear transformations we can assume $a = 1, b = 0$, and then it is convenient to write $c = 3p, d = 2q$. Thus, our equation becomes:

$$x^3 + 3px + 2q = 0$$

Regarding such equations, many things can be said, and to start with, we have the following famous result, dealing with real roots, due to Cardano:

THEOREM 8.20. *For a normalized degree 3 equation, namely*

$$x^3+3px+2q=0$$

the discriminant is $\Delta=-108(p^3+q^2)$. Assuming $p,q\in\mathbb{R}$ and $\Delta<0$, the number

$$x=\sqrt[3]{-q+\sqrt{p^3+q^2}}+\sqrt[3]{-q-\sqrt{p^3+q^2}}$$

is a real solution of our equation.

PROOF. The formula of Δ is clear from definitions, and with $108=4\times 27$. Now with x as in the statement, by using $(a+b)^3=a^3+b^3+3ab(a+b)$, we have:

$$\begin{aligned}x^3&=\left(\sqrt[3]{-q+\sqrt{p^3+q^2}}+\sqrt[3]{-q-\sqrt{p^3+q^2}}\right)^3\\&=-2q+3\sqrt[3]{-q+\sqrt{p^3+q^2}}\cdot\sqrt[3]{-q-\sqrt{p^3+q^2}}\cdot x\\&=-2q+3\sqrt[3]{q^2-p^3-q^2}\cdot x\\&=-2q-3px\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Regarding the other roots, we know from Theorem 8.18 that these are both real when $\Delta<0$, and complex conjugate when $\Delta<0$. Thus, in the context of Theorem 8.20, the other two roots are complex conjugate, the formula for them being as follows:

PROPOSITION 8.21. *For a normalized degree 3 equation, namely*

$$x^3+3px+2q=0$$

with $p,q\in\mathbb{R}$ and discriminant $\Delta=-108(p^3+q^2)$ negative, $\Delta<0$, the numbers

$$z=w\sqrt[3]{-q+\sqrt{p^3+q^2}}+w^2\sqrt[3]{-q-\sqrt{p^3+q^2}}$$

$$\bar{z}=w^2\sqrt[3]{-q+\sqrt{p^3+q^2}}+w\sqrt[3]{-q-\sqrt{p^3+q^2}}$$

with $w=e^{2\pi i/3}$ are the complex conjugate solutions of our equation.

PROOF. As before, by using $(a+b)^3=a^3+b^3+3ab(a+b)$, we have:

$$\begin{aligned}z^3&=\left(w\sqrt[3]{-q+\sqrt{p^3+q^2}}+w^2\sqrt[3]{-q-\sqrt{p^3+q^2}}\right)^3\\&=-2q+3\sqrt[3]{-q+\sqrt{p^3+q^2}}\cdot\sqrt[3]{-q-\sqrt{p^3+q^2}}\cdot z\\&=-2q+3\sqrt[3]{q^2-p^3-q^2}\cdot z\\&=-2q-3pz\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

As a conclusion, we have the following statement, unifying the above:

THEOREM 8.22. *For a normalized degree 3 equation, namely*

$$x^3+3px+2q=0$$

the discriminant is $\Delta=-108(p^3+q^2)$. *Assuming* $p,q\in\mathbb{R}$ *and* $\Delta<0$, *the numbers*

$$x=w\sqrt[3]{-q+\sqrt{p^3+q^2}}+w^2\sqrt[3]{-q-\sqrt{p^3+q^2}}$$

with $w=1,e^{2\pi i/3},e^{4\pi i/3}$ *are the solutions of our equation.*

PROOF. This follows indeed from Theorem 8.20 and Proposition 8.21. Alternatively, we can redo the computation in their proof, which was nearly identical anyway, in the present setting, with x being given by the above formula, by using $w^3=1$. □

As a comment here, the above formula holds in the case $\Delta>0$ too, and also when the coefficients are complex numbers, $p,q\in\mathbb{C}$. However, these extensions are quite often not very useful, because when it comes to extract the above square and cubic roots, for complex numbers, you can end up with the initial question, the one you started with.

8d. Degree 4 equations

In higher degree things become quite complicated. In degree 4, to start with, we first have the following result, dealing with the discriminant and its applications:

THEOREM 8.23. *The discriminant of* $P=ax^4+bx^3+cx^2+dx+e$ *is given by the following formula:*

$$\begin{aligned}\Delta &= 256a^3e^3-192a^2bde^2-128a^2c^2e^2+144a^2cd^2e-27a^2d^4\\ &\quad +144ab^2ce^2-6ab^2d^2e-80abc^2de+18abcd^3+16ac^4e\\ &\quad -4ac^3d^2-27b^4e^2+18b^3cde-4b^3d^3-4b^2c^3e+b^2c^2d^2\end{aligned}$$

In the case $\Delta<0$ *we have 2 real roots and 2 complex conjugate roots, and in the case* $\Delta>0$ *the roots are either all real or all complex.*

PROOF. The formula of Δ follows from the definition of the discriminant, from Theorem 8.16, with the resultant computed via Theorem 8.15, as follows:

$$\Delta=\frac{1}{a}\begin{vmatrix}a & & & 4a & & & \\ b & a & & 3b & 4a & & \\ c & b & a & 2c & 3b & 4a & \\ d & c & b & d & 2c & 3b & 4a\\ e & d & c & & d & 2c & 3b\\ & e & d & & & d & 2c\\ & & e & & & & d\end{vmatrix}$$

As for the last assertion, the study here is routine, a bit as in degree 3. □

In practice, as in degree 3, we can do first some manipulations on our polynomials, as to have them in simpler form, and we have the following version of Theorem 8.23:

PROPOSITION 8.24. *The discriminant of $P = x^4 + cx^2 + dx + e$, normalized degree 4 polynomial, is given by the following formula:*

$$\Delta = 16c^4e - 4c^3d^2 - 128c^2e^2 + 144cd^2e - 27d^4 + 256e^3$$

As before, if $\Delta < 0$ we have 2 real roots and 2 complex conjugate roots, and if $\Delta > 0$ the roots are either all real or all complex.

PROOF. This is a consequence of Theorem 8.23, with $a = 1, b = 0$, but we can deduce this as well directly. Indeed, the formula of Δ follows, quite easily, from:

$$\Delta = \frac{1}{a}\begin{vmatrix} 1 & & & 4 & & & \\ & 1 & & & 4 & & \\ c & & 1 & 2c & & 4 & \\ d & c & & d & 2c & & 4 \\ e & d & c & & d & 2c & \\ & e & d & & & d & 2c \\ & & e & & & & d \end{vmatrix}$$

As for the last assertion, this is something that we know, from Theorem 8.23. □

We still have some work to do. Indeed, looking back at what we did in degree 3, the passage there from Theorem 8.19 to Theorem 8.20 was made of two operations, namely "depressing" the equation, that is, getting rid of the next-to-highest term, and then rescaling the coefficients, as for the formula of Δ to become as simple as possible.

In our present setting now, degree 4, with the depressing done as above, in Proposition 8.24, it remains to rescale the coefficients, as for the formula of Δ to become as simple as possible. And here, a bit of formula hunting, in relation with $2, 3$ powers, leads to:

THEOREM 8.25. *The discriminant of a normalized degree 4 polynomial, written as*

$$P = x^4 + 6px^2 + 4qx + 3r$$

is given by the following formula:

$$\Delta = 256 \times 27 \times \left(9p^4r - 2p^3q^2 - 6p^2r^2 + 6pq^2r - q^4 + r^3\right)$$

In the case $\Delta < 0$ we have 2 real roots and 2 complex conjugate roots, and in the case $\Delta > 0$ the roots are either all real or all complex.

PROOF. This follows from Proposition 8.24, with $c = 6p, d = 4q, e = 3r$, but we can deduce this as well directly. Indeed, the formula of Δ follows, quite easily, from:

$$\Delta = \begin{vmatrix} 1 & & & 4 & & & \\ & 1 & & & 4 & & \\ 6p & & 1 & 12p & & 4 & \\ 4q & 6p & & 4q & 12p & & 4 \\ 3r & 4q & 6p & & 4q & 12p & \\ & 3r & 4q & & & 4q & 12p \\ & & 3r & & & & 4q \end{vmatrix}$$

As for the last assertion, this is something that we know from Theorem 8.23. □

Time now to get to the real thing, solving the equation. We have here:

THEOREM 8.26. *The roots of a normalized degree 4 equation, written as*

$$x^4 + 6px^2 + 4qx + 3r = 0$$

are as follows, with y satisfying the equation $(y^2 - 3r)(y - 3p) = 2q^2$,

$$x_1 = \frac{1}{\sqrt{2}}\left(-\sqrt{y - 3p} + \sqrt{-y - 3p + \frac{4q}{\sqrt{2y - 6p}}}\right)$$

$$x_2 = \frac{1}{\sqrt{2}}\left(-\sqrt{y - 3p} - \sqrt{-y - 3p + \frac{4q}{\sqrt{2y - 6p}}}\right)$$

$$x_3 = \frac{1}{\sqrt{2}}\left(\sqrt{y - 3p} + \sqrt{-y - 3p - \frac{4q}{\sqrt{2y - 6p}}}\right)$$

$$x_4 = \frac{1}{\sqrt{2}}\left(\sqrt{y - 3p} - \sqrt{-y - 3p - \frac{4q}{\sqrt{2y - 6p}}}\right)$$

and with y being computable via the Cardano formula.

PROOF. This is something quite tricky, the idea being as follows:

(1) To start with, let us write our equation in the following form:

$$x^4 = -6px^2 - 4qx - 3r$$

The idea will be that of adding a suitable common term, to both sides, as to make square on both sides, as to eventually end with a sort of double quadratic equation. For this purpose, our claim is that what we need is a number y satisfying:

$$(y^2 - 3r)(y - 3p) = 2q^2$$

Indeed, assuming that we have this number y, our equation becomes:

$$\begin{aligned}(x^2+y)^2 &= x^4+2x^2y+y^2\\ &= -6px^2-4qx-3r+2x^2y+y^2\\ &= (2y-6p)x^2-4qx+y^2-3r\\ &= (2y-6p)x^2-4qx+\frac{2q^2}{y-3p}\\ &= \left(\sqrt{2y-6p}\cdot x-\frac{2q}{\sqrt{2y-6p}}\right)^2\end{aligned}$$

(2) Which looks very good, leading us to the following degree 2 equations:

$$x^2+y+\sqrt{2y-6p}\cdot x-\frac{2q}{\sqrt{2y-6p}}=0$$

$$x^2+y-\sqrt{2y-6p}\cdot x+\frac{2q}{\sqrt{2y-6p}}=0$$

Now let us write these two degree 2 equations in standard form, as follows:

$$x^2+\sqrt{2y-6p}\cdot x+\left(y-\frac{2q}{\sqrt{2y-6p}}\right)=0$$

$$x^2-\sqrt{2y-6p}\cdot x+\left(y+\frac{2q}{\sqrt{2y-6p}}\right)=0$$

(3) Regarding the first equation, the solutions there are as follows:

$$x_1=\frac{1}{2}\left(-\sqrt{2y-6p}+\sqrt{-2y-6p+\frac{8q}{\sqrt{2y-6p}}}\right)$$

$$x_2=\frac{1}{2}\left(-\sqrt{2y-6p}-\sqrt{-2y-6p+\frac{8q}{\sqrt{2y-6p}}}\right)$$

As for the second equation, the solutions there are as follows:

$$x_3=\frac{1}{2}\left(\sqrt{2y-6p}+\sqrt{-2y-6p-\frac{8q}{\sqrt{2y-6p}}}\right)$$

$$x_4=\frac{1}{2}\left(\sqrt{2y-6p}-\sqrt{-2y-6p-\frac{8q}{\sqrt{2y-6p}}}\right)$$

(4) Now by cutting a $\sqrt{2}$ factor from everything, this gives the formulae in the statement. As for the last claim, regarding the nature of y, this comes from Cardano. □

We still have to compute the number y appearing in the above via Cardano, and the result here, adding to what we already have in Theorem 8.26, is as follows:

THEOREM 8.27 (continuation). *The value of y in the previous theorem is*

$$y = t + p + \frac{a}{t}$$

where the number t is given by the formula

$$t = \sqrt[3]{b + \sqrt{b^2 - a^3}}$$

with $a = p^2 + r$ and $b = 2p^2 - 3pr + q^2$.

PROOF. The legend has it that this is what comes from Cardano, but depressing and normalizing and solving $(y^2 - 3r)(y - 3p) = 2q^2$ makes it for too many operations, so the most pragmatic is to simply check this equation. With y as above, we have:

$$\begin{aligned} y^2 - 3r &= t^2 + 2pt + (p^2 + 2a) + \frac{2pa}{t} + \frac{a^2}{t^2} - 3r \\ &= t^2 + 2pt + (3p^2 - r) + \frac{2pa}{t} + \frac{a^2}{t^2} \end{aligned}$$

With this in hand, we have the following computation:

$$\begin{aligned} (y^2 - 3r)(y - 3p) &= \left(t^2 + 2pt + (3p^2 - r) + \frac{2pa}{t} + \frac{a^2}{t^2}\right)\left(t - 2p + \frac{a}{t}\right) \\ &= t^3 + (a - 4p^2 + 3p^2 - r)t + (2pa - 6p^3 + 2pr + 2pa) \\ &\quad + (3p^2a - ra - 4p^2a + a^2)\frac{1}{t} + \frac{a^3}{t^3} \\ &= t^3 + (a - p^2 - r)t + 2p(2a - 3p^2 + r) + a(a - p^2 - r)\frac{1}{t} + \frac{a^3}{t^3} \\ &= t^3 + 2p(-p^2 + 3r) + \frac{a^3}{t^3} \end{aligned}$$

Now by using the formula of t in the statement, this gives:

$$\begin{aligned} (y^2 - 3r)(y - 3p) &= b + \sqrt{b^2 - a^3} - 4p^2 + 6pr + \frac{a^3}{b + \sqrt{b^2 - a^3}} \\ &= b + \sqrt{b^2 - a^3} - 4p^2 + 6pr + b - \sqrt{b^2 - a^3} \\ &= 2b - 4p^2 + 6pr \\ &= 2(2p^2 - 3pr + q^2) - 4p^2 + 6pr \\ &= 2q^2 \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

In degree 5 and more, things become complicated, and the conceptual explanations for what happens here come from the Galois theory of field extensions:

THEOREM 8.28. *Unlike in degree $N \leq 4$, there is no formula for the roots of polynomials of degree $N = 5$ and higher, with the reason for this, coming from Galois theory, being that S_5 is not solvable. The simplest numeric example is $P = X^5 - X - 1$.*

PROOF. This is something quite tricky, normally requiring some good knowledge in abstract algebra, but here is the idea with all this, for what this is worth:

(1) The first assertion, for generic polynomials, is due to Abel-Ruffini, but Galois theory helps in better understanding this, and comes with a number of bonus points too, namely the possibility of formulating a finer result, with Abel-Ruffini's original "generic", which was something algebraic, being now replaced by an analytic "generic", and also with the possibility of dealing with concrete polynomials, such as:

$$P = X^5 - X - 1$$

(2) Regarding now the details of the Galois proof of the Abel-Ruffini theorem, assume that the roots of a polynomial $P \in F[X]$ can be computed by using iterated roots, a bit as for the degree 2 equation, or for the degree 3 and 4 equations, via Cardano. Then, algebrically speaking, this gives rise to a tower of fields as folows, with $F_0 = F$, and each F_{i+1} being obtained from F_i by adding a root, $F_{i+1} = F_i(x_i)$, with $x_i^{n_i} \in F_i$:

$$F_0 \subset F_1 \subset \ldots \subset F_k$$

(3) In order for Galois theory to apply well to this situation, we must make all the extensions normal, which amounts in replacing each $F_{i+1} = F_i(x_i)$ by its extension $K_i(x_i)$, with K_i extending F_i by adding a n_i-th root of unity. Thus, with this replacement, we can assume that the tower in (2) in normal, meaning that all Galois groups are cyclic.

(4) Now by Galois theory, at the level of the corresponding Galois groups we obtain a tower of groups as follows as follows, which is a resolution of the last group G_k, the Galois group of P, in the sense of group theory, in the sense that all quotients are cyclic:

$$G_1 \subset G_2 \subset \ldots \subset G_k$$

As a conclusion, Galois theory tells us that if the roots of a polynomial $P \in F[X]$ can be computed by using iterated roots, then its Galois group $G = G_k$ must be solvable.

(5) In the generic case, the conclusion is that Galois theory tells us that, in order for all polynomials of degree 5 to be solvable, via square roots, the group S_5, which appears there as Galois group, must be solvable, in the sense of group theory. But this is wrong, because the alternating subgroup $A_5 \subset S_5$ is simple, and therefore not solvable.

(6) Finally, regarding the polynomial $P = X^5 - X - 1$, some elementary computations here, based on arithmetic over $\mathbb{F}_2, \mathbb{F}_3$, and involving various cycles of length $2, 3, 5$, show that its Galois group is S_5. Thus, we have our counterexample.

(7) To be more precise, our polynomial factorizes over $\mathbb{F}_2$ as follows:

$$X^5 - X - 1 = (X^2 + X + 1)(X^3 + X^2 + 1)$$

We deduce from this the existence of an element $\tau\sigma \in G \subset S_5$, with $\tau \in S_5$ being a transposition, and with $\sigma \in S_5$ being a 3-cycle, disjoint from it. Thus, we have:

$$\tau = (\tau\sigma)^3 \in G$$

(8) On the other hand since $P = X^5 - X - 1$ is irreducible over $\mathbb{F}_5$, we have as well available a certain 5-cycle $\rho \in G$. Now since $< \tau, \rho >= S_5$, we conclude that the Galois group of P is full, $G = S_5$, and by (4) and (5) we have our counterexample.

(9) Finally, as mentioned in (1), all this shows as well that a random polynomial of degree 5 or higher is not solvable by square roots, and with this being an elementary consequence of the main result from (5), via some standard analysis arguments. □

So long for Galois theory and its main applications, quickly explained. Exercise of course for you, to read more about this, and about polynomials in general.

8e. Exercises

This was a rather advanced chapter, and as exercises on this, we have:

EXERCISE 8.29. *Clarify what has been said above, about symmetric functions.*

EXERCISE 8.30. *Clarify as well all the details in relation with the resultant.*

EXERCISE 8.31. *Learn other formulations of the Cardano formula in degree* 3.

EXERCISE 8.32. *Experiment with Cardano, in relation with trigonometry questions.*

EXERCISE 8.33. *Learn the other formulations of the Cardano formula in degree* 4.

EXERCISE 8.34. *Learn more about field extensions, algebra, and Galois theory.*

EXERCISE 8.35. *Work out all the Galois theory details for* $P = X^5 - X - 1$.

EXERCISE 8.36. *Learn also about the applications of Galois theory to finite fields.*

As bonus exercise, start reading some algebraic geometry. All good polynomials.

Part III

Derivatives

Put me up, put me down
Put my feet back on the ground
Put me up, take my heart
And make me happy

CHAPTER 9

Derivatives, rules

9a. Derivatives

Welcome to analysis I guess, after the long introduction to functions we have been through, which was mostly of algebraic nature. In this second part of this book we intend to explain the basics of the modern theory of functions, based on two key operations, namely differentiation and integration. And with this being something quite recent, due to Newton, Leibnitz and others, going back no more than a few centuries ago.

The basic idea of modern analysis is very simple, coming from the following question:

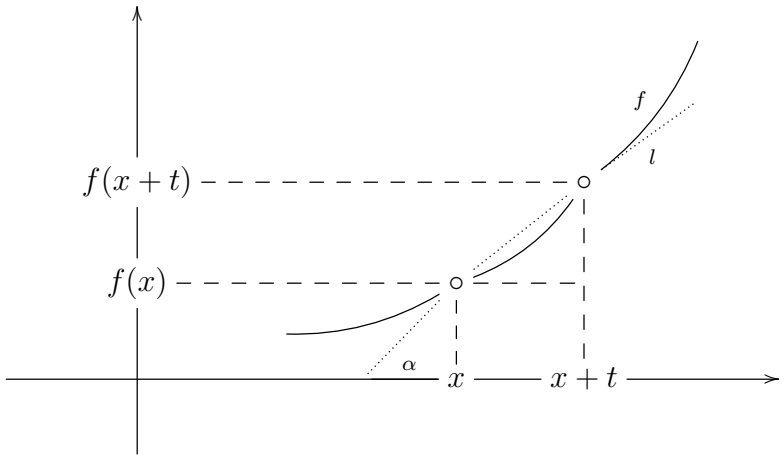
QUESTION 9.1. *When $f : \mathbb{R} \to \mathbb{R}$ is continuous at $x \in \mathbb{R}$, we have, for $t \simeq 0$:*

$$f(x + t) \simeq f(x)$$

How to improve this formula, into something of type $f(x + t) \simeq f(x) + \varepsilon(t)$?

To be more precise, as we know well since chapter 2, the condition $f(x + t) \simeq f(x)$ with $t \simeq 0$ means precisely that f is continuous at x. And we would like to improve this estimate, into $f(x + t) \simeq f(x) + \varepsilon(t)$, with $\varepsilon(t)$ being a certain simple function of t.

In answer, let us draw a picture. With x fixed, we can see appearing some geometry, with a right triangle, a line l, and an angle α, all depending on t, as follows:



But this provides us with an answer to our question. Indeed, trigonometry inside the small right triangle tells us that we have the following formula:

$$\tan\alpha = \frac{f(x+t)-f(x)}{t}$$

Now the point is that we can write this latter formula in the following way:

$$f(x+t) = f(x) + \tan\alpha \cdot t$$

But with $t \to 0$ the line l will become the tangent at $(x, f(x))$ to the graph of f, and so $\tan\alpha$ will become the slope of this tangent, depending only on x. Thus, we have:

ANSWER 9.2. *The basic estimate $f(x+t) \simeq f(x)$ can be improved into*

$$f(x+t) \simeq f(x) + \tan\alpha \cdot t$$

with $\tan\alpha$ being the slope of the tangent at $(x, f(x))$ to the graph of f.

Which is very nice, but in practice, things can be a bit more complicated than this. For instance for the modulus function $f(x) = |x|$, which is certainly continuous, the tangent at $x = 0$ does obviously not exist, and so our method above will not apply.

Summarizing, our method will work for certain functions, and will not work for some other. And with this being not surprising, philosophically, because $f(x+t) \simeq f(x)$ being plainly equivalent to the continuity at x, any improvement of this estimate, by whatever means, should normally require f to be continuous in some stronger sense.

Nevermind. In any case, we have a valuable solution to our initial question, and in view of the above discussion, we must proceed carefully, as follows:

DEFINITION 9.3. *A function $f : \mathbb{R} \to \mathbb{R}$ is called differentiable at x when*

$$f'(x) = \lim_{t\to 0} \frac{f(x+t)-f(x)}{t}$$

called derivative of f at that point x, exists.

Observe that, geometrically, f being differentiable at x means precisely that the graph of f admits a tangent at $(x, (f(x))$, and with $f'(x)$ being the slope of this tangent. However, and here comes the point, the power of Definition 9.3 comes precisely from the fact that it is something very simple, making no reference to this, graphs and geometry.

As another remark, in order for f to be differentiable at x, that is to say, in order for the above limit to converge, the numerator must go to 0, as the denominator t does:

$$\lim_{t\to 0} [f(x+t) - f(x)] = 0$$

Thus, f must be continuous at x. However, the converse is not true, the basic counterexample here being $f(x) = |x|$ at $x = 0$. Let us summarize these findings as follows:

PROPOSITION 9.4. *If f is differentiable at x, then f must be continuous at x. However, the converse is not true, a basic counterexample being the modulus function*

$$f(x) = |x|$$

at the point $x = 0$.

PROOF. The first assertion is something that we already know, from the above. As for the second assertion, regarding $f(x) = |x|$, this is something quite clear on the picture of f, but let us prove this mathematically, based on Definition 9.3. We have:

$$\lim_{t\searrow 0} \frac{|0+t| - |0|}{t} = \lim_{t\searrow 0} \frac{t-0}{t} = 1$$

On the other hand, we have as well the following computation:

$$\lim_{t\nearrow 0} \frac{|0+t| - |0|}{t} = \lim_{t\nearrow 0} \frac{-t-0}{t} = -1$$

Thus, the limit in Definition 9.3 does not converge, as desired. □

Generally speaking, the last assertion in Proposition 9.4 should not bother us much, because most of the basic continuous functions are differentiable, and we will see examples in a moment, with these basically covering all the functions that we know.

Before that, however, let us recall why we are here, namely improving the basic estimate $f(x+t) \simeq f(x)$. We can now do this, using the derivative, as follows:

THEOREM 9.5. *Assuming that f is differentiable at x, we have:*

$$f(x+t) \simeq f(x) + f'(x)t$$

In other words, f is, approximately, locally affine at x.

PROOF. Assume indeed that f is differentiable at x, and let us set, as before:

$$f'(x) = \lim_{t\to 0} \frac{f(x+t) - f(x)}{t}$$

By multiplying by t, we obtain that we have, once again in the $t \to 0$ limit:

$$f(x+t) - f(x) \simeq f'(x)t$$

Thus, we are led to the conclusion in the statement. □

The above result has many practical applications, to all branches of mathematics, physics and other sciences, and many theoretical consequences too. We will explore all this later, after getting more familiar with the derivatives, and their computation.

As a first computation, the derivatives of the power functions are as follows:

THEOREM 9.6. *We have the differentiation formula*

$$(x^p)' = px^{p-1}$$

valid for any exponent $p \in \mathbb{R}$.

PROOF. We can do this in three steps, as follows:

(1) In the case $p \in \mathbb{N}$ we can use the binomial formula, which gives, as desired:

$$\begin{aligned}(x+t)^p &= \sum_{k=0}^{n}\binom{p}{k}x^{p-k}t^k\\ &= x^p + px^{p-1}t + \ldots + t^p\\ &\simeq x^p + px^{p-1}t\end{aligned}$$

(2) Let us discuss now the general case $p \in \mathbb{Q}$. We write $p = m/n$, with $m \in \mathbb{Z}$ and $n \in \mathbb{N}$. In order to do the computation, we use the following formula:

$$a^n - b^n = (a-b)(a^{n-1} + a^{n-2}b + \ldots + b^{n-1})$$

We set in this formula $a = (x+t)^{m/n}$ and $b = x^{m/n}$. We obtain, as desired:

$$\begin{aligned}(x+t)^{m/n} - x^{m/n} &= \frac{(x+t)^m - x^m}{(x+t)^{m(n-1)/n} + \ldots + x^{m(n-1)/n}}\\ &\simeq \frac{(x+t)^m - x^m}{nx^{m(n-1)/n}}\\ &\simeq \frac{mx^{m-1}t}{nx^{m(n-1)/n}}\\ &= \frac{m}{n}\cdot x^{m-1-m+m/n}\cdot t\\ &= \frac{m}{n}\cdot x^{m/n-1}\cdot t\end{aligned}$$

(3) In the general case now, where $p \in \mathbb{R}$ is real, we can use a similar argument. Indeed, given any integer $n \in \mathbb{N}$, we have the following computation:

$$\begin{aligned}(x+t)^p - x^p &= \frac{(x+t)^{pn} - x^{pn}}{(x+t)^{p(n-1)} + \ldots + x^{p(n-1)}}\\ &\simeq \frac{(x+t)^{pn} - x^{pn}}{nx^{p(n-1)}}\end{aligned}$$

Now observe that we have the following estimate, with [.] being the integer part:

$$(x+t)^{[pn]} \leq (x+t)^{pn} \leq (x+t)^{[pn]+1}$$

By using the binomial formula on both sides, for the integer exponents $[pn]$ and $[pn]+1$ there, we deduce that with $n >> 0$ we have the following estimate:

$$(x+t)^{pn} \simeq x^{pn} + pnx^{pn-1}t$$

Thus, we can finish our computation started above as follows:

$$(x+t)^p - x^p \simeq \frac{pnx^{pn-1}t}{nx^{pn-p}} = px^{p-1}t$$

But this gives $(x^p)' = px^{p-1}$, which finishes the proof. □

As a comment now, answering a question that you might have, Theorem 9.6 is of course valid at $x > 0$, and this due to difficulties, and even impossibility in general, of talking about x^p for $x < 0$. In relation with this, let us make the following conventions:

CONVENTIONS 9.7. *When talking derivatives $f'(x)$, we agree on the following:*

(1) *f needs only to be defined on a small interval around x, and this because $f'(x)$ is something local, only depending on f on that small interval around x.*
(2) *When talking about the derivative of an abstract function $f : X \to \mathbb{R}$, with $X \subset \mathbb{R}$, we will therefore assume, by definition, that $X \subset \mathbb{R}$ is open.*
(3) *With an exception, however, for the functions $f : [a, b] \to \mathbb{R}$, where the right/left derivatives at a, b can be defined in an obvious way, as the right/left limits.*

Summarizing, a bit technical all this, and with this book being something quite modest, namely introduction to calculus, as opposed to professional treatise on calculus, we will not hesitate from time to time to be a bit sloppy with all this, if this can be of help, that is, if this can simplify the presentation, and make the ideas of calculus more visible.

Finally, for ending this discussion, shall you ever end up in a key scientific job, say engineer designing things, based on calculus, you will need at some point to upgrade your calculus knowledge, learned from here, into something fully professional and rigorous. And I can recommend here to you, in advance, the books of Rudin [73], [74], which tell the same story as here, and much more, in a fully professional and rigorous way.

Back to work now, we know that the derivatives are, and we have $(x^p)' = px^{p-1}$. Here are some further computations of derivatives, for other basic functions that we know:

THEOREM 9.8. *We have the following results:*

(1) $(\sin x)' = \cos x$.
(2) $(\cos x)' = -\sin x$.
(3) $(e^x)' = e^x$.
(4) $(\log x)' = x^{-1}$.

PROOF. This is quite tricky, as always when computing derivatives, as follows:

(1) Regarding sin, the computation here goes as follows:

$$\begin{aligned}(\sin x)' &= \lim_{t\to 0}\frac{\sin(x+t)-\sin x}{t}\\ &= \lim_{t\to 0}\frac{\sin x\cos t+\cos x\sin t-\sin x}{t}\\ &= \lim_{t\to 0}\sin x\cdot\frac{\cos t-1}{t}+\cos x\cdot\frac{\sin t}{t}\\ &= \cos x\end{aligned}$$

Here we have used the standard fact, that we know well from chapter 3, that we have $\sin t\simeq t$ for $t\simeq 0$, plus the fact, which follows from this and from Pythagoras, $\sin^2+\cos^2=1$, that we have as well $\cos t\simeq 1-t^2/2$, for $t\simeq 0$.

(2) The computation for cos is similar, as follows:

$$\begin{aligned}(\cos x)' &= \lim_{t\to 0}\frac{\cos(x+t)-\cos x}{t}\\ &= \lim_{t\to 0}\frac{\cos x\cos t-\sin x\sin t-\cos x}{t}\\ &= \lim_{t\to 0}\cos x\cdot\frac{\cos t-1}{t}-\sin x\cdot\frac{\sin t}{t}\\ &= -\sin x\end{aligned}$$

(3) For the exponential, the derivative can be computed as follows:

$$\begin{aligned}(e^x)' &= \left(\sum_{k=0}^{\infty}\frac{x^k}{k!}\right)'\\ &= \sum_{k=0}^{\infty}\frac{kx^{k-1}}{k!}\\ &= e^x\end{aligned}$$

(4) As for the logarithm, the computation here is as follows, using $\log(1+y)\simeq y$ for $y\simeq 0$, which follows from $e^y\simeq 1+y$ that we found in (3), by taking the logarithm:

$$\begin{aligned}(\log x)' &= \lim_{t\to 0}\frac{\log(x+t)-\log x}{t}\\ &= \lim_{t\to 0}\frac{\log(1+t/x)}{t}\\ &= \frac{1}{x}\end{aligned}$$

Thus, we are led to the formulae in the statement. □

Speaking exponentials, we can now formulate a nice result about them, as follows:

THEOREM 9.9. *The exponential function, namely*

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

is the unique power series satisfying $f' = f$ and $f(0) = 1$.

PROOF. Consider indeed a power series satisfying $f' = f$ and $f(0) = 1$. Due to $f(0) = 1$, the first term must be 1, and so our function must look as follows:

$$f(x) = 1 + \sum_{k=1}^{\infty} c_k x^k$$

According to our differentiation rules, the derivative of this series is given by:

$$f(x) = \sum_{k=1}^{\infty} k c_k x^{k-1}$$

Thus, the equation $f' = f$ is equivalent to the following equalities:

$$c_1 = 1 \quad , \quad 2c_2 = c_1 \quad , \quad 3c_3 = c_2 \quad , \quad 4c_4 = c_3 \quad , \quad \ldots$$

But this system of equations can be solved by recurrence, as follows:

$$c_1 = 1 \quad , \quad c_2 = \frac{1}{2} \quad , \quad c_3 = \frac{1}{2 \times 3} \quad , \quad c_4 = \frac{1}{2 \times 3 \times 4} \quad , \quad \ldots$$

Thus we have $c_k = 1/k!$, leading to the conclusion in the statement. □

Observe that the above result leads to a more conceptual explanation for the number e itself. To be more precise, $e \in \mathbb{R}$ is the unique number satisfying:

$$(e^x)' = e^x$$

Which is very nice, at least we know one thing. And, more on this later.

9b. Theorems, rules

Let us work out now some general results, for the computation of derivatives. We have here the following statement, summarizing all that you need to know:

THEOREM 9.10. *We have the following formulae:*

(1) $(f + g)' = f' + g'$.
(2) $(fg)' = f'g + fg'$.
(3) $(f \circ g)' = (f' \circ g) \cdot g'$.

PROOF. All these formulae are elementary, the idea being as follows:

(1) This follows indeed from definitions, the computation being as follows:

$$\begin{aligned}(f+g)'(x) &= \lim_{t\to 0}\frac{(f+g)(x+t)-(f+g)(x)}{t}\\ &= \lim_{t\to 0}\left(\frac{f(x+t)-f(x)}{t}+\frac{g(x+t)-g(x)}{t}\right)\\ &= \lim_{t\to 0}\frac{f(x+t)-f(x)}{t}+\lim_{t\to 0}\frac{g(x+t)-g(x)}{t}\\ &= f'(x)+g'(x)\end{aligned}$$

(2) This follows from definitions too, the computation, by using the more convenient formula $f(x+t)\simeq f(x)+f'(x)t$ as a definition for the derivative, being as follows:

$$\begin{aligned}(fg)(x+t) &= f(x+t)g(x+t)\\ &\simeq (f(x)+f'(x)t)(g(x)+g'(x)t)\\ &\simeq f(x)g(x)+(f'(x)g(x)+f(x)g'(x))t\end{aligned}$$

Indeed, we obtain from this that the derivative is the coefficient of t, namely:

$$(fg)'(x)=f'(x)g(x)+f(x)g'(x)$$

(3) Regarding compositions, the computation here is as follows, again by using the more convenient formula $f(x+t)\simeq f(x)+f'(x)t$ as a definition for the derivative:

$$\begin{aligned}(f\circ g)(x+t) &= f(g(x+t))\\ &\simeq f(g(x)+g'(x)t)\\ &\simeq f(g(x))+f'(g(x))g'(x)t\end{aligned}$$

Indeed, we obtain from this that the derivative is the coefficient of t, namely:

$$(f\circ g)'(x)=f'(g(x))g'(x)$$

Thus, we are led to the conclusions in the statement. □

We can of course combine the above formulae, and we obtain for instance:

THEOREM 9.11. *The derivatives of fractions are given by:*

$$\left(\frac{f}{g}\right)'=\frac{f'g-fg'}{g^2}$$

In particular, we have the following formula, for the derivative of inverses:

$$\left(\frac{1}{f}\right)'=-\frac{f'}{f^2}$$

In fact, we have $(f^p)'=pf^{p-1}$, for any exponent $p\in\mathbb{R}$.

PROOF. This statement is in fact written a bit upside down, and for the proof it is technically most convenient to proceed backwards, as follows:

(1) By using the formula $(x^p)' = px^{p-1}$, that we know from Theorem 9.6, and then the clain rule from Theorem 9.10 (3), we obtain the third formula.

(2) Then, with $p = -1$, we obtain from this the second formula.

(3) And finally, by using this second formula and Theorem 9.10 (2), we obtain:

$$\begin{aligned}\left(\frac{f}{g}\right)' &= \left(f \cdot \frac{1}{g}\right)' \\ &= f' \cdot \frac{1}{g} + f\left(\frac{1}{g}\right)' \\ &= \frac{f'}{g} - \frac{fg'}{g^2} \\ &= \frac{f'g - fg'}{g^2}\end{aligned}$$

Thus, we are led to the various conclusions in the statement. □

All the above might seem to start to be a bit too complex, with too many things to be memorized and so on, and as a piece of advice here, we have:

ADVICE 9.12. *Memorize and cherish the formula for fractions*

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$$

along with the usual addition formula, that you know well

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$$

and generally speaking, never mess with fractions.

With this coming from a lifelong calculus teacher and scientist, mathematics can be difficult, and many things can be pardoned, but not messing with fractions. And with this going beyond mathematics too, say if you want to make a living by selling apples or tomatoes at the market, fine, but you'll need to know well fractions, trust me.

Back to work now, with the above formulae in hand, we can do all sorts of computations for other basic functions that we know. Let us start with the tangent:

THEOREM 9.13. *We have the following formula,*

$$(\tan x)' = \frac{1}{\cos^2 x}$$

provided that the denominator does not vanish.

PROOF. This is very standard, with two proofs being possible, as follows:

(1) We have indeed the following computation, using Theorem 9.10:

$$\begin{aligned}(\tan x)' &= \left(\frac{\sin x}{\cos x}\right)' \\ &= \frac{\sin' x\cos x-\sin x\cos' x}{\cos^2 x} \\ &= \frac{\cos^2 x+\sin^2 x}{\cos^2 x} \\ &= \frac{1}{\cos^2 x}\end{aligned}$$

(2) Alternatively, we can get this by using $\tan t\simeq t$ for $t\simeq 0$, as follows:

$$\begin{aligned}(\tan x)' &= \lim_{t\to 0}\frac{\tan(x+t)-\tan x}{t} \\ &= \lim_{t\to 0}\frac{\frac{\tan x+\tan t}{1-\tan x\tan t}-\tan x}{t} \\ &= \lim_{t\to 0}\frac{\tan t+\tan^2 x\tan t}{t(1-\tan x\tan t)} \\ &= \lim_{t\to 0}\frac{\tan t+\tan^2 x\tan t}{t} \\ &= 1+\tan^2 x \\ &= \frac{1}{\cos^2 x}\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

As a comment now, observe that the formula for the tangent can be written as follows, in terms of the secant function, and with this looking like an improvement:

$$(\tan x)'=\sec^2 x$$

However, it is better not to do so, and this for a quite subtle reason, as follows:

FACT 9.14. *We will learn later that the operation inverse to $f\to f'$, called integration, is something interesting too, and in view of this, it is better to express our f' functions in terms of* $\sin,\cos$ *only, for subsequent quick identification and integration, when needed.*

Well, hope you get my point, while the various secondary trigonometric functions are certainly very interesting objects, worth the study, and valuable as input for our present derivative computations, they are not recommended as output, for the above reasons.

Talking now secondary trigonometric functions, we have, regarding them:

THEOREM 9.15. *We have the following formulae,*

$$(\sec x)' = \frac{\sin x}{\cos^2 x} \quad , \quad (\csc x)' = -\frac{\cos x}{\sin^2 x} \quad , \quad (\cot x)' = -\frac{1}{\sin^2 x}$$

provided that the denominators do not vanish.

PROOF. For the secant, we have the following computation:

$$(\sec x)' = \left(\frac{1}{\cos x}\right)' = -\frac{\cos' x}{\cos^2 x} = \frac{\sin x}{\cos^2 x}$$

For the cosecant, we have a similar computation, as follows:

$$(\csc x)' = \left(\frac{1}{\sin x}\right)' = -\frac{\sin' x}{\sin^2 x} = -\frac{\cos x}{\sin^2 x}$$

For the cotangent, we have the following computation, as for the tangent:

$$\begin{aligned}(\cot x)' &= \left(\frac{\cos x}{\sin x}\right)' \\ &= \frac{\cos' x \sin x - \cos x \sin' x}{\sin^2 x} \\ &= -\frac{\sin^2 x + \cos^2 x}{\sin^2 x} \\ &= -\frac{1}{\sin^2 x}\end{aligned}$$

Alternatively, we can use our previous formula for the tangent, and we obtain:

$$(\cot x)' = -\frac{\tan' x}{\tan^2 x} = -\frac{1/\cos^2 x}{\sin^2 x/\cos^2 x} = -\frac{1}{\sin^2 x}$$

Thus, we are led to the conclusions in the statement. □

9c. Basic functions

The computation of derivatives being something quite addictive, let us investigate now the remaining trigonometric functions. We first have the following result:

THEOREM 9.16. *The derivatives of the basic inverse trigonometric functions are*

$$(\arcsin x)' = \frac{1}{\sqrt{1-x^2}} \quad , \quad (\arccos x)' = -\frac{1}{\sqrt{1-x^2}} \quad , \quad (\arctan x)' = \frac{1}{1+x^2}$$

and the derivatives of the secondary inverse trigonometric functions are

$$(\operatorname{arcsec} x)' = \frac{1}{|x|\sqrt{x^2-1}} \ , \ (\operatorname{arccsc} x)' = -\frac{1}{|x|\sqrt{x^2-1}} \ , \ (\operatorname{arccot} x)' = -\frac{1}{1+x^2}$$

provided that the denominators do not vanish.

PROOF. This is something routine, by using what we already have, along with the formula $(f \circ g)' = (f' \circ g) \cdot g'$ from Theorem 9.10 (3), as follows:

(1) For the arcsine, we can use the following computation:

$$\begin{aligned}(\sin \circ \arcsin)'(x) &= \sin'(\arcsin x) \arcsin'(x) \\ &= \cos(\arcsin x)\ \arcsin'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x' = 1$, we obtain from this:

$$\arcsin'(x) = \frac{1}{\cos(\arcsin x)}$$

But with $t = \arcsin x$ we have $\sin t = x$, so we obtain the result, via:

$$\cos(\arcsin x) = \cos t = \sqrt{1 - \sin^2 t} = \sqrt{1 - x^2}$$

(2) For the arcosine, we have a similar computation, as follows:

$$\begin{aligned}(\cos \circ \arccos)'(x) &= \cos'(\arccos x) \arccos'(x) \\ &= -\sin(\arccos x)\ \arccos'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x' = 1$, we obtain from this:

$$\arccos'(x) = -\frac{1}{\sin(\arccos x)}$$

But with $t = \arccos x$ we have $\cos t = x$, so we obtain the result, via:

$$\sin(\arccos x) = \sin t = \sqrt{1 - \cos^2 t} = \sqrt{1 - x^2}$$

(3) For the arctangent, we can use the following computation:

$$\begin{aligned}(\tan \circ \arctan)'(x) &= \tan'(\arctan x) \arctan'(x) \\ &= \frac{1}{\cos^2(\arctan x)} \arctan'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x' = 1$, we obtain from this:

$$\arctan'(x) = \cos^2(\arctan x)$$

But with $t = \arctan x$ we have $\tan t = x$, so we obtain the result, via:

$$\cos^2(\arctan x) = \cos^2 t = \frac{1}{1 + \tan^2 t} = \frac{1}{1 + x^2}$$

(4) For the arcsecant, we can use the following computation:

$$\begin{aligned}(\sec \circ \operatorname{arcsec})'(x) &= \sec'(\operatorname{arcsec} x) \operatorname{arcsec}'(x) \\ &= \frac{\sin(\operatorname{arcsec} x)}{\cos^2(\operatorname{arcsec} x)} \operatorname{arcsec}'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x' = 1$, we obtain from this:

$$\text{arcsec}'(x) = \frac{\cos^2(\text{arcsec}\, x)}{\sin(\text{arcsec}\, x)}$$

On the other hand, with $t = \text{arcsec}\, x$ we have $\sec t = x$, and so:

$$\cos(\text{arcsec}\, x) = \cos t = \frac{1}{x}$$

As for the sine of the arcsecant, we can compute it as well, as follows:

$$\sin(\text{arcsec}\, x) = \sin t = \sqrt{1 - \cos^2 t} = \sqrt{1 - \frac{1}{x^2}} = \frac{\sqrt{x^2 - 1}}{|x|}$$

Thus, we are led to the formula in the statement, namely:

$$(\text{arcsec}\, x)' = \frac{1}{x^2} \cdot \frac{|x|}{\sqrt{x^2 - 1}} = \frac{1}{|x|\sqrt{x^2 - 1}}$$

(5) For the arcosecant, we can use the following computation:

$$\begin{aligned} (\csc \circ \text{arccsc})'(x) &= \csc'(\text{arccsc}\, x)\, \text{arccsc}'(x) \\ &= -\frac{\cos(\text{arcsec}\, x)}{\sin^2(\text{arccsc}\, x)}\, \text{arccsc}'(x) \end{aligned}$$

Indeed, since the term on the left is simply $x' = 1$, we obtain from this:

$$\text{arccsc}'(x) = -\frac{\sin^2(\text{arccsc}\, x)}{\cos(\text{arccsc}\, x)}$$

On the other hand, with $t = \text{arccsc}\, x$ we have $\csc t = x$, and so:

$$\sin(\text{arccsc}\, x) = \sin t = \frac{1}{x}$$

As for the cosine of the arcosecant, we can compute it as well, as follows:

$$\cos(\text{arccsc}\, x) = \cos t = \sqrt{1 - \sin^2 t} = \sqrt{1 - \frac{1}{x^2}} = \frac{\sqrt{x^2 - 1}}{|x|}$$

Thus, we are led to the formula in the statement, namely:

$$(\text{arccsc}\, x)' = -\frac{1}{x^2} \cdot \frac{|x|}{\sqrt{x^2 - 1}} = -\frac{1}{|x|\sqrt{x^2 - 1}}$$

(6) For the arcotangent, we can use the following computation:

$$\begin{aligned} (\cot \circ \text{arccot})'(x) &= \cot'(\text{arccot}\, x)\, \text{arccot}'(x) \\ &= -\frac{1}{\sin^2(\text{arccot}\, x)}\, \text{arccot}'(x) \end{aligned}$$

Indeed, since the term on the left is simply $x' = 1$, we obtain from this:

$$\text{arccot}'(x) = -\sin^2(\text{arccot}\, x)$$

But with $t = \text{arccot}\, x$ we have $\cot t = x$, so we obtain the result, via:

$$\sin^2(\text{arccot}\, x) = \sin^2 t = \frac{1}{1+\cot^2 t} = \frac{1}{1+x^2}$$

Summarizing, theorem proved, we are now experts in computing derivatives. □

And with this, computations over? You must be kidding. Next, we have:

THEOREM 9.17. *The derivatives of basic hyperbolic trigonometric functions are*

$$(\sinh x)' = \cosh x \quad , \quad (\cosh x)' = \sinh x \quad , \quad (\tanh x)' = \frac{1}{\cosh^2 x}$$

and the derivatives of secondary hyperbolic trigonometric functions are

$$(\text{sech}\, x)' = -\frac{\sinh x}{\cosh^2 x} \quad , \quad (\text{csch}\, x)' = -\frac{\cosh x}{\sinh^2 x} \quad , \quad (\coth x)' = -\frac{1}{\sinh^2 x}$$

provided that the denominators do not vanish.

PROOF. This is again something routine, the idea being as follows:

(1) We recall from chapter 7 that we have the following formulae:

$$\sinh x = \frac{e^x - e^{-x}}{2} \quad , \quad \cosh x = \frac{e^x + e^{-x}}{2}$$

But this shows right away that we have the following formulae, as claimed:

$$(\sinh x)' = \cosh x \quad , \quad (\cosh x)' = \sinh x$$

(2) Regarding now the hyperbolic secant, we have the following computation:

$$(\text{sech}\, x)' = \left(\frac{1}{\cosh x}\right)' = -\frac{\cosh' x}{\cosh^2 x} = -\frac{\sinh x}{\cosh^2 x}$$

(3) For the hyperbolic cosecant the computation is similar, as follows:

$$(\text{csch}\, x)' = \left(\frac{1}{\sinh x}\right)' = -\frac{\sinh' x}{\sinh^2 x} = -\frac{\cosh x}{\sinh^2 x}$$

(4)) In what regards now the tangent, we have the following computation:

$$\begin{aligned}(\tanh x)' &= \left(\frac{\sinh x}{\cosh x}\right)' \\ &= \frac{\sinh' x\cosh x-\sinh x\cosh' x}{\cosh^2 x} \\ &= \frac{\cosh^2 x-\sinh^2 x}{\cosh^2 x} \\ &= \frac{1}{\cosh^2 x}\end{aligned}$$

(5) For the cotangent the computation is similar, as follows:

$$\begin{aligned}(\coth x)' &= \left(\frac{\cosh x}{\sinh x}\right)' \\ &= \frac{\cosh' x\sinh x-\cosh x\sinh' x}{\sinh^2 x} \\ &= \frac{\sinh^2 x-\cosh^2 x}{\sinh^2 x} \\ &= -\frac{1}{\sinh^2 x}\end{aligned}$$

Thus, we are led to the conclusions in the statement. □

Finally, regarding the inverse hyperbolic trigonometric functions, we have:

THEOREM 9.18. *The derivatives of basic inverse hyperbolic functions are given by*

$$(\operatorname{arcsinh} x)'=\frac{1}{\sqrt{1+x^2}}\quad,\quad(\operatorname{arccosh} x)'=\frac{1}{\sqrt{x^2-1}}\quad,\quad(\operatorname{arctanh} x)'=\frac{1}{1-x^2}$$

and the derivatives of secondary inverse hyperbolic functions are given by

$$(\operatorname{arcsech} x)'=-\frac{1}{|x|\sqrt{1+x^2}}\ ,\ (\operatorname{arccsch} x)'=-\frac{1}{|x|\sqrt{1-x^2}}\ ,\ (\operatorname{arccoth} x)'=\frac{1}{1-x^2}$$

provided that the denominators do not vanish.

PROOF. This is again routine, by using $(f\circ g)'=(f'\circ g)\cdot g'$, as follows:

(1) For the arcsine, we can use the following computation:

$$\begin{aligned}(\sinh\circ\operatorname{arcsinh})'(x) &= \sinh'(\operatorname{arcsinh} x)\operatorname{arcsinh}'(x) \\ &= \cosh(\operatorname{arcsinh} x)\operatorname{arcsinh}'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x'=1$, we obtain from this:

$$\operatorname{arcsinh}'(x)=\frac{1}{\cosh(\operatorname{arcsinh} x)}$$

But with $t = \operatorname{arcsinh} x$ we have $\sinh t = x$, so we obtain the result, via:

$$\cosh(\operatorname{arcsinh} x) = \cosh t = \sqrt{1 + \sinh^2 t} = \sqrt{1 + x^2}$$

(2) For the arcosine, we have a similar computation, as follows:

$$\begin{aligned}(\cosh \circ \operatorname{arccosh})'(x) &= \cosh'(\operatorname{arccosh} x) \operatorname{arccosh}'(x) \\ &= \sinh(\operatorname{arccosh} x) \operatorname{arccosh}'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x' = 1$, we obtain from this:

$$\operatorname{arccosh}'(x) = \frac{1}{\sinh(\operatorname{arccosh} x)}$$

But with $t = \operatorname{arccosh} x$ we have $\cosh t = x$, so we obtain the result, via:

$$\sinh(\operatorname{arccosh} x) = \sinh t = \sqrt{\cosh^2 t - 1} = \sqrt{x^2 - 1}$$

(3) For the arctangent, we can use the following computation:

$$\begin{aligned}(\tanh \circ \operatorname{arctanh})'(x) &= \tanh'(\operatorname{arctanh} x) \operatorname{arctanh}'(x) \\ &= \frac{1}{\cosh^2(\operatorname{arctanh} x)} \operatorname{arctanh}'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x' = 1$, we obtain from this:

$$\operatorname{arctanh}'(x) = \cosh^2(\operatorname{arctanh} x)$$

But with $t = \operatorname{arctanh} x$ we have $\tanh t = x$, so we obtain the result, via:

$$\cosh^2(\operatorname{arctanh} x) = \cosh^2 t = \frac{1}{1 - \tan^2 t} = \frac{1}{1 - x^2}$$

(4) For the arcsecant, we can use the following computation:

$$\begin{aligned}(\operatorname{sech} \circ \operatorname{arcsech})'(x) &= \operatorname{sech}'(\operatorname{arcsech} x) \operatorname{arcsech}'(x) \\ &= -\frac{\sinh(\operatorname{arcsech} x)}{\cosh^2(\operatorname{arcsech} x)} \operatorname{arcsech}'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x' = 1$, we obtain from this:

$$\operatorname{arcsech}'(x) = -\frac{\cosh^2(\operatorname{arcsech} x)}{\sinh(\operatorname{arcsech} x)}$$

On the other hand, with $t = \operatorname{arcsech} x$ we have $\operatorname{sech} t = x$, and so:

$$\cosh(\operatorname{arcsech} x) = \cosh t = \frac{1}{x}$$

As for the sine of the arcsecant, we can compute it as well, as follows:

$$\sinh(\operatorname{arcsech} x) = \sinh t = \sqrt{1 + \cosh^2 t} = \sqrt{1 + \frac{1}{x^2}} = \frac{\sqrt{1 + x^2}}{|x|}$$

Thus, we are led to the formula in the statement, namely:

$$(\operatorname{arcsech} x)' = -\frac{1}{x^2} \cdot \frac{|x|}{\sqrt{1+x^2}} = -\frac{1}{|x|\sqrt{1+x^2}}$$

(5) For the arcosecant, we can use the following computation:

$$\begin{aligned}(\operatorname{csch} \circ \operatorname{arccsch})'(x) &= \operatorname{csch}'(\operatorname{arccsch} x)\operatorname{arccsch}'(x)\\ &= -\frac{\cosh(\operatorname{arccsch} x)}{\sinh^2(\operatorname{arccsch} x)}\operatorname{arccsch}'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x' = 1$, we obtain from this:

$$\operatorname{arccsch}'(x) = -\frac{\sinh^2(\operatorname{arccsch} x)}{\cosh(\operatorname{arccsch} x)}$$

On the other hand, with $t = \operatorname{arccsch} x$ we have $\operatorname{csch} t = x$, and so:

$$\sinh(\operatorname{arccsch} x) = \sinh t = \frac{1}{x}$$

As for the cosine of the arcsecant, we can compute it as well, as follows:

$$\cosh(\operatorname{arccsch} x) = \cosh t = \sqrt{\sinh^2 t - 1} = \sqrt{\frac{1}{x^2} - 1} = \frac{\sqrt{1-x^2}}{|x|}$$

Thus, we are led to the formula in the statement, namely:

$$(\operatorname{arccsch} x)' = -\frac{1}{x^2} \cdot \frac{|x|}{\sqrt{1-x^2}} = -\frac{1}{|x|\sqrt{1-x^2}}$$

(6) For the arcotangent, we can use the following computation:

$$\begin{aligned}(\coth \circ \operatorname{arccoth})'(x) &= \coth'(\operatorname{arccoth} x)\operatorname{arccoth}'(x)\\ &= -\frac{1}{\sinh^2(\operatorname{arccoth} x)}\operatorname{arccoth}'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x' = 1$, we obtain from this:

$$\operatorname{arccoth}'(x) = -\sinh^2(\operatorname{arccoth} x)$$

But with $t = \operatorname{arccoth} x$ we have $\coth t = x$, so we obtain the result, via:

$$\sinh^2(\operatorname{arccoth} x) = \sinh^2 t = \frac{1}{\coth^2 t - 1} = \frac{1}{x^2 - 1}$$

And so, theorem proved, we are now experts in hyperbolic trigonometry. □

9d. Local extrema

Enough fun I guess with the computations, and time now for some more theory. At the theoretical level, further building on Theorem 9.5, we have:

THEOREM 9.19. *The local minima and maxima of a differentiable function $f : \mathbb{R} \to \mathbb{R}$ appear at the points $x \in \mathbb{R}$ where:*

$$f'(x) = 0$$

However, the converse of this fact is not true in general.

PROOF. The first assertion follows from the formula $f(x+t) \simeq f(x) + f'(x)t$. Indeed, let us rewrite this formula, more conveniently, in the following way:

$$f(x+t) - f(x) \simeq f'(x)t$$

Now saying that our function f has a local maximum at $x \in \mathbb{R}$ means that there exists a number $\varepsilon > 0$ such that the following happens:

$$f(x+t) \geq f(x) \quad , \quad \forall t \in [-\varepsilon, \varepsilon]$$

We conclude that we must have $f'(x)t \geq 0$ for sufficiently small t, and since this small t can be both positive or negative, this gives, as desired:

$$f'(x) = 0$$

Similarly, saying that our function f has a local minimum at $x \in \mathbb{R}$ means that there exists a number $\varepsilon > 0$ such that the following happens:

$$f(x+t) \leq f(x) \quad , \quad \forall t \in [-\varepsilon, \varepsilon]$$

Thus $f'(x)t \leq 0$ for small t, and this gives, as before, $f'(x) = 0$. Finally, in what regards the converse, the simplest counterexample here is the following function:

$$f(x) = x^3$$

Indeed, we have $f'(x) = 3x^2$, and in particular $f'(0) = 0$. But our function being clearly increasing, $x = 0$ is not a local maximum, nor a local minimum. □

In practice, Theorem 9.19 can be used in order to find the minimum and maximum of any differentiable function, and this method is best recalled as follows:

ALGORITHM 9.20. *In order to find the minimum and maximum of $f : [a, b] \to \mathbb{R}$:*

(1) *Compute the derivative f'.*
(2) *Solve the equation $f'(x) = 0$.*
(3) *Add a, b to your set of solutions.*
(4) *Compute $f(x)$, for all your solutions.*
(5) *Compute the min/max of all these $f(x)$ values.*
(6) *Then this is the min/max of your function.*

Needless to say, all this is very interesting, and powerful. The general problem in any type of applied mathematics is that of finding the minimum or maximum of some function, and we have now an algorithm for dealing with such questions. Very nice.

Back to theory, as an important consequence of Theorem 9.19, we have:

THEOREM 9.21. *Assuming that $f : [a, b] \to \mathbb{R}$ is differentiable, we have*

$$\frac{f(b) - f(a)}{b - a} = f'(c)$$

for some $c \in (a, b)$, called mean value property of f.

PROOF. This is something fundamental, coming in two steps, as follows:

(1) In the case $f(a) = f(b)$, the result, called Rolle theorem, states that we must have $f'(c) = 0$ for some $c \in (a, b)$. But this, which by the way is obvious on pictures, follows from Theorem 9.19, because due to $f(a) = f(b)$, our function must have a minimum and maximum on (a, b), and the derivative at either of these must vanish, $f'(c) = 0$.

(2) Now in what regards our theorem as stated, which is a result due to Lagrange, also clear on pictures, this follows from Rolle, applied to the following function:

$$g(x) = f(x) - \frac{f(b) - f(a)}{b - a} \cdot x$$

Indeed, observe first that we have indeed $g(a) = g(b)$, due to:

$$g(b) - g(a) = (f(b) - f(a)) - \frac{f(b) - f(a)}{b - a} \cdot (b - a) = 0$$

Thus Rolle applies and gives $g'(c) = 0$ for some $c \in (a, b)$. But:

$$g'(x) = f'(x) - \frac{f(b) - f(a)}{b - a}$$

Thus $g'(c) = 0$ translates into the formula in the statement. □

As a key consequence of Theorem 9.21, of great practical interest, we have:

THEOREM 9.22. *For a differentiable function we have*

$$f' = 0 \quad \implies \quad f = \text{constant}$$

and with the converse of this being of course true too.

PROOF. This is indeed something self-explanatory, coming from Theorem 9.21. □

As a first comment, this latter result, which might sound a bit philosophical, reminds a bit the main principle in probability, which is something very useful, namely:

$$P(X) > 0 \quad \implies \quad X \text{ happens}$$

To be more precise, you might know from probability that, quite often, the best way of proving that something X happens is by computing the hard way $P(X)$, via lots of formulae and work, then getting something > 0, and applying the above principle.

Well, in what regards the study of functions, pretty much the same can be said, by using Theorem 9.22. That is, if you want to prove $f = g$, an efficient method is that of differentiating $f - g$, or f/g, the hard way, getting 0, and applying Theorem 9.22.

Here is an illustration for this general principle, which is of key importance:

THEOREM 9.23. *We have indeed the Euler formula, namely*

$$e^{it} = \cos t + i \sin t$$

for any $t \in \mathbb{R}$. As a consequence, we have the formulae

$$\cos t = \sum_{k=0}^{\infty} (-1)^k \frac{t^{2k}}{(2k)!} \quad , \quad \sin t = \sum_{k=0}^{\infty} (-1)^k \frac{t^{2k+1}}{(2k+1)!}$$

also valid for any $t \in \mathbb{R}$.

PROOF. We certainly know about the Euler formula and its various consequences, since chapter 4, but our discussion so far on this subject was lacking a bit or rigor. Time to fix this. In order to prove the formula, consider the following function $f : \mathbb{R} \to \mathbb{C}$:

$$f(t) = \frac{\cos t + i \sin t}{e^{it}}$$

The point now is that we can compute the derivative of this function f by using our first derivative formulae for exp, sin, cos, and we obtain in this way:

$$\begin{aligned} f'(t) &= (e^{-it}(\cos t + i \sin t))' \\ &= -ie^{-it}(\cos t + i \sin t) + e^{-it}(-\sin t + i \cos t) \\ &= e^{-it}(-i \cos t + \sin t) + e^{-it}(-\sin t + i \cos t) \\ &= 0 \end{aligned}$$

We conclude that our function $f : \mathbb{R} \to \mathbb{C}$ is constant, and the constant in question can be found by setting $t = 0$, where we obtain:

$$f(0) = \frac{\cos 0 + i \sin 0}{e^{i0}} = \frac{1}{1} = 1$$

Thus we have $f(t) = 1$ for any t, and we have proved the Euler formula. As for the formulae for $\sin, \cos$, these follow from this, as already explained in chapter 6. □

Along the same lines, as another key application, we have as well:

THEOREM 9.24. *We have the generalized binomial formula*

$$(1+x)^p=\sum_{k=0}^{\infty}\binom{p}{k}x^k$$

with the generalized binomial coefficients being given by

$$\binom{p}{k}=\frac{p(p-1)\ldots(p-k+1)}{k!}$$

valid for any exponent $p\in\mathbb{R}$, and any $|x|<1$.

PROOF. As before with Theorem 9.23, this is something which closes a recurrent discussion, throughout this book. If f is the series in the statement, we have:

$$(1+x)f'(x)=pf(x)$$

Now by using this formula, we have the following computation:

$$\big((1+x)^{-p}f(x)\big)'=-p(1+x)^{-p-1}f(x)+(1+x)^{-p}f'(x)=0$$

Thus we have $f(x)=c(1+x)^p$, with $c=f(0)=1$, as desired. □

As yet another application of our derivative techniques, we have:

THEOREM 9.25. *The Lipschitz constant of $f:[a,b]\to\mathbb{R}$ is*

$$K=\sup_{x\in[a,b]}|f'(x)|$$

with the derivatives at a,b being computed using right and left limits.

PROOF. This is indeed something self-explanatory, based on Theorem 9.21. □

The above result is quite powerful, and as an application, we have:

THEOREM 9.26. *The following functions are Lipschitz, on any compact interval $[a,b]$ belonging to their domain, with their best constants K there being as indicated:*

(1) *x^n, with $K=n\cdot\max(|a|^{n-1},|b|^{n-1})$.*
(2) *x^{-n}, with $K=n\cdot\max(1/|a|^{n+1},1/|b|^{n+1})$.*
(3) *$\sin x$, with $K=\sup_{x\in[a,b]}|\cos x|$.*
(4) *$\cos x$, with $K=\sup_{x\in[a,b]}|\sin x|$.*
(5) *$\tan x$, with $K=\sup_{x\in[a,b]}1/\cos^2 x$.*
(6) *$\exp x$, with $K=e^b$.*
(7) *$\log x$, with $K=1/a$.*

As for the infinite intervals, here the best constants $K\in[0,\infty]$, with $K<\infty$ corresponding to Lipschitz, and $K=\infty$ corresponding to non-Lipschitz, can be computed too.

PROOF. This is something that we proved the hard way in chapter 5, and with some details missing, and which is now trivial, by using Theorem 9.25. □

Still talking simplifications of previous Lipschitz computations, we have as well:

THEOREM 9.27. *The Babylonian function for extracting the square root of $a > 1$,*

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

has the following properties:

(1) *For $a < 3$ it is a contraction on $[1, a]$, with constant $K = (a-1)/2$.*
(2) *For $a > 3$ it is a contraction on $[\sqrt{a}, a/3]$, with constant $K = (9/a - 1)/2$.*

PROOF. Again, this is something that we proved the hard way before, in chapter 6, and with some details missing, and which is now trivial, by using Theorem 9.25. □

Finally, as yet another application of Theorem 9.21, we have the following interesting result, due to Darboux, showing that not every function can appear as a derivative:

THEOREM 9.28. *Given a differentiable function on an interval*

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

its derivative f' has the intermediate value property.

PROOF. Consider indeed the following two functions $\varphi, \psi : [a, b] \to \mathbb{R}$, which are continuous, defined at the endpoints by using the differentiability of f:

$$\varphi(x) = \frac{f(x) - f(a)}{x - a} \quad , \quad \psi(x) = \frac{f(x) - f(b)}{x - b}$$

Since $f'(a), f'(b)$ belong to the interval $I = Im(\varphi) \cup Im(\psi)$, any $y \in [f'(a), f'(b)]$ must belong to I too, and by Theorem 9.21 we get $y = f'(c)$ for some $c \in [a, b]$, as desired. □

9e. Exercises

Welcome to calculus, eventually. As exercises on derivatives, we have:

EXERCISE 9.29. *Find other functions, besides $|x|$, which are not differentiable.*

EXERCISE 9.30. *Clarify all the details in our proof of $(x^p)' = px^{p-1}$.*

EXERCISE 9.31. *Rewrite the theory of e, by starting with $f' = f$, $f(0) = 1$.*

EXERCISE 9.32. *Compute $(fg/h)'$, and then $(f/gh)'$ and $(fg/hk)'$ too.*

EXERCISE 9.33. *Compute $(1/f^2)'$, and then $(1/f^k)'$, with $k \in \mathbb{N}$.*

EXERCISE 9.34. *Learn more, via examples and counterexamples, about $f'(x) = 0$.*

EXERCISE 9.35. *Learn more about Rolle, Lagrange and the mean value property.*

EXERCISE 9.36. *Apply our extrema algorithm, to some functions of your choice.*

As bonus exercise, and no surprise here, compute 100 derivatives.

CHAPTER 10

Second derivatives

10a. Second derivatives

The derivative theory that we have is already quite powerful, and can be used in order to solve all sorts of interesting questions, but with a bit more effort, we can do better. Indeed, at a more advanced level, we can come up with the following notion:

DEFINITION 10.1. *We say that $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable if it is differentiable, and its derivative $f' : \mathbb{R} \to \mathbb{R}$ is differentiable too. The derivative of f' is denoted*

$$f'' : \mathbb{R} \to \mathbb{R}$$

and is called second derivative of f.

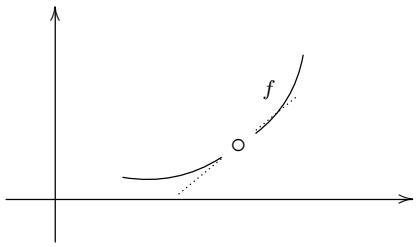
You might probably wonder why coming with this definition, which looks a bit abstract and complicated, instead of further developing the theory of the first derivative, which looks like something very reasonable and useful. Good point, and answer to this coming in a moment. But before that, let us get a bit familiar with f''. We have here:

INTERPRETATION 10.2. *The second derivative $f''(x) \in \mathbb{R}$ is the number which:*

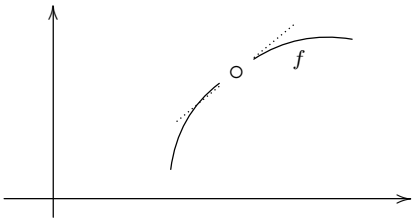
(1) *Expresses the growth rate of the slope $f'(z)$ at the point x.*
(2) *Gives us the acceleration of the function f at the point x.*
(3) *Computes how much different is $f(x)$, compared to $f(z)$ with $z \simeq x$.*
(4) *Tells us how much convex or concave is f, around the point x.*

So, this is the truth about the second derivative, making it clear that what we have here is a very interesting notion. In practice now, the situation is as follows:

(1) This is something very intuitive, which follows from the usual interpretation of the derivative, both as a growth rate, and a slope, according to the following picture:



To be more precise, in the above picture the second derivative is positive. The second derivative can be as well negative, in which case the picture is as follows:



(2) This is some sort of reformulation of (1), using the intuitive meaning of the word "acceleration", with the relevant physics equations, due to Newton, being as follows:

$$v = \dot{x} \quad , \quad a = \dot{v}$$

To be more precise, here x, v, a are the position, speed and acceleration, and the dot denotes the time derivative. Now according to these equations, we have the following formula, saying that the acceleration appears as the second derivative of the position:

$$a = \ddot{x}$$

But with this in hand, and getting now a bit abstract, we can intuitively say that the second derivative of an arbitrary function, $f''(x) \in \mathbb{R}$, is the number which gives us the acceleration of the function f at the point x, as started in Interpretation 10.2 (2).

(3) This is something more subtle. We know from the derivative theory from chapter 9 that, approximately, the differentiable functions are locally linear:

$$f(x + t) \simeq f(x) + at$$

By writing this formula with $t \to -t$ too, and making the average, we obtain:

$$\frac{f(x + t) + f(x - t)}{2} \simeq f(x)$$

Which is of course something not very interesting, but here comes the point. Assuming that the second derivative works a bit like the first one, but at second order, we can expect, approximately, the twice differentiable functions to be locally quadratic:

$$f(x + t) \simeq f(x) + at + bt^2$$

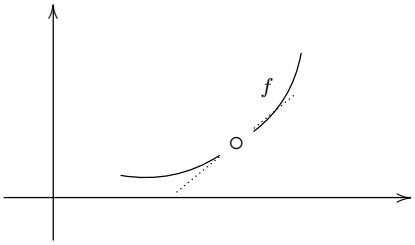
By writing this formula with $t \to -t$ too, and making the average, we obtain:

$$\frac{f(x + t) + f(x - t)}{2} \simeq f(x) + bt^2$$

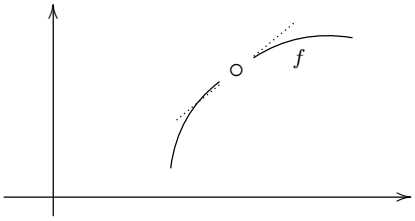
Which suddenly becomes interesting, and save for some clarification about what the above number b is, in relation with the second derivative $f''(x)$, and for a clarification regarding the average of t^2 too, when t is small, around 0, we are led to the conclusion in

Interpretation 10.2 (3), namely that $f''(x) \in \mathbb{R}$ is the number which computes how much different is $f(x)$, compared to the average of $f(z)$ with $z \simeq x$. More on this later.

(4) This is something quite subtle too, and again very useful in practice, that we will again clarify with some mathematics, later on this chapter, and with lots of applications too. In the meantime, let us mention that the convexity of a function is something quite intuitive, here being a typical example of such a convex function:



As for the notion of concavity, this is something a bit similar, corresponding to the opposite situation, where the slope of the function tends to decrease, as follows:



But, we recognize here the pictures that we used in (1), so we are led to the conclusion in Interpretation 10.2 (4), namely that the second derivative $f''(x) \in \mathbb{R}$ is the number which tells us how much convex or concave is f, around the point x.

All in all, what we have above, in Interpretation 10.2, is a mixture of trivial and non-trivial facts, and do not worry, we will get familiar with all this, in the next few pages, as this chapter develops. At a theoretical level now, let us record the following result:

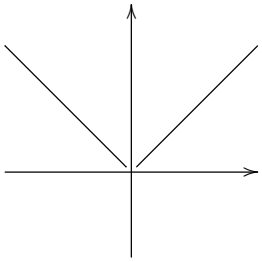
THEOREM 10.3. *There are functions which are differentiable, such as*

$$f(x) = \begin{cases} -x^2 & (x \leq 0) \\ x^2 & (x \geq 0) \end{cases}$$

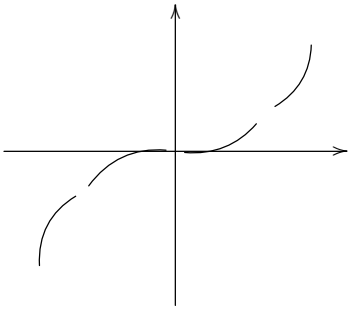
but not twice differentiable.

PROOF. In order to construct a counterexample, recall first that the simplest example of a function which is continuous, but not differentiable, was $f(x) = |x|$, the idea behind

this being to use a "piecewise linear function whose branches do not fit well":



In connection now with our question, piecewise linear will not do, but we can use a similar idea, namely "piecewise quadratic function whose branches do not fit well":



So, following this idea, consider the following function, depending on $a, b \in \mathbb{R}$:

$$f(x) = \begin{cases} ax^2 & (x \leq 0) \\ bx^2 & (x \geq 0) \end{cases}$$

This function is then differentiable, with its derivative being:

$$f'(x) = \begin{cases} 2ax & (x \leq 0) \\ 2bx & (x \geq 0) \end{cases}$$

Now for getting our counterexample, we can set $a = -1, b = 1$, so that f is:

$$f(x) = \begin{cases} -x^2 & (x \leq 0) \\ x^2 & (x \geq 0) \end{cases}$$

Indeed, the derivative is $f'(x) = 2|x|$, which is not differentiable, as desired. □

In practice now, in order to get familiar with the second derivatives, let us first compute the second derivatives of the functions that we are familiar with, and see what we get. The basic result here, which is perhaps not very enlightening at this stage of things, but which looks technically useful, or at least let us hope so, is as follows:

THEOREM 10.4. *The second derivatives of the basic functions are as follows:*

(1) $(x^p)'' = p(p-1)x^{p-2}$.
(2) $\sin'' = -\sin$.
(3) $\cos'' = -\cos$.
(4) $\exp'' = \exp$.
(5) $\log''(x) = -1/x^2$.

PROOF. The various formulae in the statement all follow from the various formulae for the derivatives established before, in chapter 9, as follows:

$$(x^p)'' = (px^{p-1})' = p(p-1)x^{p-2}$$

$$(\sin x)'' = (\cos x)' = -\sin x$$

$$(\cos x)'' = (-\sin x)' = -\cos x$$

$$(e^x)'' = (e^x)' = e^x$$

$$(\log x)'' = (-1/x)' = -1/x^2$$

Thus, we are led to the formulae in the statement. □

The above result might suggest that the second derivative is somehow similar to the first derivative. However, this is wrong, as shown by the following result:

THEOREM 10.5. *We have the following second derivative formula,*

$$(\tan x)'' = \frac{2\sin x}{\cos^3 x}$$

provided that the denominator does not vanish.

PROOF. We have indeed the following computation:

$$\begin{aligned}(\tan x)'' &= \left(\frac{1}{\cos^2 x}\right)' \\ &= -\frac{(\cos^2 x)'}{\cos^4 x} \\ &= \frac{2\cos x \sin x}{\cos^4 x} \\ &= \frac{2\sin x}{\cos^3 x}\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

10b. Basic examples

As before with the first derivatives, the computation of second derivatives is something addictive. So, now that we know about cos, sin, tan, next come sec, csc, cot:

THEOREM 10.6. *We have the following formulae,*

$$(\sec x)'' = \frac{1+\sin^2 x}{\cos^3 x} \quad , \quad (\csc x)'' = \frac{1+\cos^2 x}{\sin^3 x} \quad , \quad (\cot x)'' = \frac{2\cos x}{\sin^3 x}$$

provided that the denominators do not vanish.

PROOF. For the secant, we have the following computation:

$$\begin{aligned}(\sec x)'' &= \left(\frac{\sin x}{\cos^2 x}\right)' \\ &= \frac{\sin' x\cos^2 x - \sin x(\cos^2 x)'}{\cos^4 x} \\ &= \frac{\cos x\cdot\cos^2 x + \sin x\cdot 2\cos x\sin x}{\cos^4 x} \\ &= \frac{\cos^2 x + 2\sin^2 x}{\cos^3 x} \\ &= \frac{1+\sin^2 x}{\cos^3 x}\end{aligned}$$

For the cosecant, we have a similar computation, as follows:

$$\begin{aligned}(\csc x)'' &= \left(-\frac{\cos x}{\sin^2 x}\right)' \\ &= -\frac{\cos' x\sin^2 x - \cos x(\sin^2 x)'}{\sin^4 x} \\ &= \frac{\sin x\cdot\sin^2 x + \cos x\cdot 2\sin x\cos x}{\sin^4 x} \\ &= \frac{\sin^2 x + 2\cos^2 x}{\sin^3 x} \\ &= \frac{1+\cos^2 x}{\sin^3 x}\end{aligned}$$

For the cotangent, we have the following computation, as for the tangent:

$$(\cot x)'' = \left(-\frac{1}{\sin^2 x}\right)' = \frac{(\sin^2 x)'}{\sin^4 x} = \frac{2\sin x\cos x}{\sin^4 x} = \frac{2\cos x}{\sin^3 x}$$

Thus, we are led to the conclusions in the statement. □

Next, time for the inverse trigonometric functions. We have here:

THEOREM 10.7. *The second derivatives of basic inverse trigonometric functions are*

$$(\arcsin x)'' = \frac{x}{(1-x^2)^{3/2}} \ , \ (\arccos x)'' = -\frac{x}{(1-x^2)^{3/2}} \ , \ (\arctan x)'' = -\frac{2x}{(1+x^2)^2}$$

and the second derivatives of secondary inverse trigonometric functions are

$$(\operatorname{arcsec} x)'' = \frac{|x|(1-2x^2)}{x^3(x^2-1)^{3/2}} \ , \ (\operatorname{arccsc} x)'' = -\frac{|x|(1-2x^2)}{x^3(x^2-1)^{3/2}} \ , \ (\operatorname{arccot} x)'' = \frac{2x}{(1+x^2)^2}$$

provided that the denominators do not vanish.

PROOF. This is routine, by using the formulae from chapter 9, as follows:

(1) For the arcsine, the computation is as follows:

$$(\arcsin x)'' = \left(\frac{1}{\sqrt{1-x^2}}\right)' = \frac{x/\sqrt{1-x^2}}{1-x^2} = \frac{x}{(1-x^2)^{3/2}}$$

(2) For the arcosine the computation, using the one above, is as follows:

$$(\arccos x)'' = \left(-\frac{1}{\sqrt{1-x^2}}\right)' = -\frac{x}{(1-x^2)^{3/2}}$$

(3) For the arctangent, the computation is as follows:

$$(\arctan x)'' = \left(\frac{1}{1+x^2}\right)' = -\frac{2x}{(1+x^2)^2}$$

(4) For the arcsecant, the computation is as follows:

$$\begin{aligned}
(\operatorname{arcsec} x)'' &= \left(\frac{1}{|x|\sqrt{x^2-1}}\right)' \\
&= -\frac{(|x|\sqrt{x^2-1})'}{x^2(x^2-1)} \\
&= -\frac{|x|'\sqrt{x^2-1}+|x|\sqrt{x^2-1}'}{x^2(x^2-1)} \\
&= -\frac{sgn(x)\sqrt{x^2-1}+|x|x/\sqrt{x^2-1}}{x^2(x^2-1)} \\
&= -\frac{sgn(x)(x^2-1)+|x|x}{x^2(x^2-1)^{3/2}} \\
&= -\frac{|x|(x^2-1)+|x|x^2}{x^3(x^2-1)^{3/2}} \\
&= \frac{|x|(1-2x^2)}{x^3(x^2-1)^{3/2}}
\end{aligned}$$

(5) For the arcosecant the computation, using the one above, is as follows:

$$(\operatorname{arccsc} x)'' = \left(-\frac{1}{|x|\sqrt{x^2-1}}\right)' = -\frac{|x|(1-2x^2)}{x^3(x^2-1)^{3/2}}$$

(6) For the arcotangent the computation, using the one for the tangent, is:

$$(\operatorname{arccot} x)'' = \left(-\frac{1}{1+x^2}\right)' = \frac{2x}{(1+x^2)^2}$$

Thus, we are led to the formulae in the statement. □

Regarding now the hyperbolic functions, we first have the following result:

THEOREM 10.8. *The second derivatives of basic hyperbolic functions are*

$$(\sinh x)'' = \sinh x \quad , \quad (\cosh x)'' = \cosh x \quad , \quad (\tanh x)'' = -\frac{2\sinh x}{\cosh^3 x}$$

and the second derivatives of secondary hyperbolic functions are

$$(\operatorname{sech} x)'' = \frac{2-3\cosh^2 x}{\cosh^3 x} \quad , \quad (\operatorname{csch} x)'' = -\frac{2+3\sinh^2 x}{\sinh^3 x} \quad , \quad (\coth x)'' = \frac{2\cosh x}{\sinh^3 x}$$

provided that the denominators do not vanish.

PROOF. This is again routine, by using the formulae from chapter 9, as follows:

(1) For the sine the computation is trivial, as follows:

$$(\sinh x)'' = (\cosh x)' = \sinh x$$

(2) For the cosine the computation is trivial too, as follows:

$$(\cosh x)'' = (\sinh x)' = \cosh x$$

(3) For the tangent, the computation is as follows:

$$\begin{aligned}(\tanh x)'' &= \left(\frac{1}{\cosh^2 x}\right)' \\ &= -\frac{(\cosh^2 x)'}{\cosh^4 x} \\ &= -\frac{2\cosh x \sinh x}{\cosh^4 x} \\ &= -\frac{2\sinh x}{\cosh^3 x}\end{aligned}$$

(4) For the secant, the computation is as follows:

$$\begin{aligned}(\operatorname{sech} x)'' &= \left(-\frac{\sinh x}{\cosh^2 x}\right)' \\ &= -\frac{(\sinh x)'\cosh^2 x + \sinh x(\cosh^2 x)'}{\cosh^4 x} \\ &= -\frac{\cosh x\cdot\cosh^2 x + \sinh x\cdot 2\cosh x\sinh x}{\cosh^4 x} \\ &= -\frac{\cosh^2 x + 2\sinh^2 x}{\cosh^3 x} \\ &= \frac{2-3\cosh^2 x}{\cosh^3 x}\end{aligned}$$

(5) For the cosecant, the computation is as follows:

$$\begin{aligned}(\operatorname{csch} x)'' &= \left(-\frac{\cosh x}{\sinh^2 x}\right)' \\ &= -\frac{(\cosh x)'\sinh^2 x + \cosh x(\sinh^2 x)'}{\sinh^4 x} \\ &= -\frac{\sinh x\cdot\sinh^2 x + \cosh x\cdot 2\sinh x\cosh x}{\sinh^4 x} \\ &= -\frac{\sinh^2 x + 2\cosh^2 x}{\sinh^3 x} \\ &= -\frac{2+3\sinh^2 x}{\sinh^3 x}\end{aligned}$$

(6) For the cotangent, the computation is as follows:

$$\begin{aligned}(\coth x)'' &= \left(-\frac{1}{\sinh^2 x}\right)' \\ &= \frac{(\sinh^2 x)'}{\sinh^4 x} \\ &= \frac{2\sinh x\cosh x}{\sinh^4 x} \\ &= \frac{2\cosh x}{\sinh^3 x}\end{aligned}$$

Thus, we are led to the formulae in the statement. □

Finally, regarding the inverse hyperbolic trigonometric functions, we have:

THEOREM 10.9. *The second derivatives of basic inverse hyperbolic functions are*

$$(\operatorname{arcsinh} x)'' = -\frac{x}{(1+x^2)^{3/2}} \ , \ (\operatorname{arccosh} x)'' = -\frac{x}{(x^2-1)^{3/2}} \ , \ (\operatorname{arctanh} x)'' = \frac{2x}{(1-x^2)^2}$$

and the second derivatives of secondary inverse hyperbolic functions are

$$(\operatorname{arcsech} x)'' = -\frac{|x|(1+2x^2)}{x^3(1+x^2)^{3/2}} \ , \ (\operatorname{arccsch} x)'' = -\frac{|x|(1-2x^2)}{x^3(1-x^2)^{3/2}} \ , \ (\operatorname{arccoth} x)'' = \frac{2x}{(1-x^2)^2}$$

provided that the denominators do not vanish.

PROOF. This is again routine, by using the formulae from chapter 9, as follows:

(1) For the arcsine, the computation is as follows:

$$\begin{aligned}(\operatorname{arcsinh} x)'' &= \left(\frac{1}{\sqrt{1+x^2}}\right)' \\ &= -\frac{\sqrt{1+x^2}'}{1+x^2} \\ &= -\frac{x/\sqrt{1+x^2}}{1+x^2} \\ &= -\frac{x}{(1+x^2)^{3/2}}\end{aligned}$$

(2) For the arcosine the computation is as follows:

$$\begin{aligned}(\operatorname{arccosh} x)'' &= \left(\frac{1}{\sqrt{x^2-1}}\right)' \\ &= -\frac{\sqrt{x^2-1}'}{x^2-1} \\ &= -\frac{x/\sqrt{x^2-1}}{x^2-1} \\ &= -\frac{x}{(x^2-1)^{3/2}}\end{aligned}$$

(3) For the arctangent, the computation is as follows:

$$(\operatorname{arctanh} x)'' = \left(\frac{1}{1-x^2}\right)' = \frac{2x}{(1-x^2)^2}$$

(4) For the arcsecant, the computation is as follows:

$$\begin{aligned}
(\operatorname{arcsech} x)'' &= \left(-\frac{1}{|x|\sqrt{1+x^2}}\right)' \\
&= \frac{(|x|\sqrt{1+x^2})'}{x^2(1+x^2)} \\
&= \frac{|x|'\sqrt{1+x^2}+|x|\sqrt{1+x^2}'}{x^2(1+x^2)} \\
&= \frac{sgn(x)\sqrt{1+x^2}+|x|x/\sqrt{1+x^2}}{x^2(1+x^2)} \\
&= \frac{sgn(x)(1+x^2)+|x|x}{x^2(1+x^2)^{3/2}} \\
&= \frac{|x|(1+x^2)+|x|x^2}{x^3(1+x^2)^{3/2}} \\
&= -\frac{|x|(1+2x^2)}{x^3(1+x^2)^{3/2}}
\end{aligned}$$

(5) For the arcosecant the computation is similar, as follows:

$$\begin{aligned}
(\operatorname{arccsch} x)'' &= \left(-\frac{1}{|x|\sqrt{1-x^2}}\right)' \\
&= \frac{(|x|\sqrt{1-x^2})'}{x^2(1-x^2)} \\
&= \frac{|x|'\sqrt{1-x^2}+|x|\sqrt{1-x^2}'}{x^2(1-x^2)} \\
&= \frac{sgn(x)\sqrt{1-x^2}-|x|x/\sqrt{1-x^2}}{x^2(1-x^2)} \\
&= \frac{sgn(x)(1-x^2)-|x|x}{x^2(1-x^2)^{3/2}} \\
&= \frac{|x|(1-x^2)-|x|x^2}{x^3(1-x^2)^{3/2}} \\
&= -\frac{|x|(1-2x^2)}{x^3(1-x^2)^{3/2}}
\end{aligned}$$

(6) Finally, for the arcotangent the computation is the same as for the tangent. □

10c. Taylor formula

Getting back now to theory, our main purpose will be that of improving, with the help of the second derivative, the basic approximation formula for functions, namely:

$$f(x+t) \simeq f(x) + f'(x)t$$

In order to do so, things will be quite tricky, and a bit more geometric, and perhaps less intuitive, than before. We will be in need of the following standard result:

THEOREM 10.10. *The* $0/0$ *type limits can be computed according to the formula*

$$\frac{f(x)}{g(x)} \simeq \frac{f'(x)}{g'(x)}$$

called L'Hôpital's rule.

PROOF. The above formula holds indeed, as an application of the general derivative theory from chapter 9, which gives, in the situation from the statement:

$$\begin{aligned}\frac{f(x+t)}{g(x+t)} &\simeq \frac{f(x)+f'(x)t}{g(x)+g'(x)t}\\ &= \frac{f'(x)t}{g'(x)t}\\ &= \frac{f'(x)}{g'(x)}\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

We can now formulate the following key result:

THEOREM 10.11. *Any twice differentiable function* $f : \mathbb{R} \to \mathbb{R}$ *is locally quadratic,*

$$f(x+t) \simeq f(x) + f'(x)t + \frac{f''(x)}{2}\,t^2$$

with $f''(x)$ *being as usual the derivative of the function* $f' : \mathbb{R} \to \mathbb{R}$ *at the point* x*.*

PROOF. This can be proved by using Theorem 10.10, as follows:

(1) Assume indeed that f is twice differentiable at x, and let us try to construct an approximation of f around x by a quadratic function, as follows:

$$f(x+t) \simeq a + bt + ct^2$$

As a first observation, obtained with $t \to 0$, the order 0 term in the above approximation must be the value of the function at the point under consideration:

$$a = f(x)$$

We also know from chapter 9 that the correct choice for the coefficient of t is the derivative of the function at the point under consideration:

$$b = f'(x)$$

Summarizing, our approximation must be a formula of the following type, with the coefficient c still in need to be computed:

$$f(x+t) \simeq f(x) + f'(x)t + ct^2$$

(2) In order to find now the correct choice for $c \in \mathbb{R}$, consider the following two functions, coming from the original function f, and from its desired approximation:

$$\varphi(t) = f(x+t)$$

$$\psi(t) = f(x) + f'(x)t + ct^2$$

These two functions match in what regards the value at $t = 0$, because we have:

$$\varphi(0) = \psi(0) = f(x)$$

These functions match as well in what regards the derivative at $t = 0$, because:

$$\varphi'(0) = \psi'(0) = f'(x)$$

Thus, the correct choice of $c \in \mathbb{R}$ should be the one making match the second derivatives at $t = 0$. But here, we have the following formulae:

$$\varphi''(0) = f''(x) \quad , \quad \psi''(0) = 2c$$

We conclude that the number $c \in \mathbb{R}$ must satisfy the following equation:

$$f''(x) = 2c$$

We are therefore led to the approximation formula in the statement, namely:

$$f(x+t) \simeq f(x) + f'(x)t + \frac{f''(x)}{2}\, t^2$$

(3) In order to prove now that this formula holds indeed, we can use L'Hôpital's rule. Indeed, by using it, if we denote by $\psi(t) \simeq P(t)$ the formula to be proved, we have:

$$\begin{aligned}
\frac{\psi(t) - P(t)}{t^2} &\simeq \frac{\psi'(t) - P'(t)}{2t} \\
&\simeq \frac{\psi''(t) - P''(t)}{2} \\
&= \frac{f''(x) - f''(x)}{2} \\
&= 0
\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

This was for the story of the Taylor formula at order 2. In case you are still wondering where that 2 coefficient comes from, here is something to always have in mind:

PROPOSITION 10.12. *The Taylor formula at order* 2*, namely*

$$f(x+t) \simeq f(x) + f'(x)t + \frac{f''(x)}{2}\,t^2$$

is an equality for degree 2 *polynomials.*

PROOF. This is something that we already know, coming from the proof of Theorem 10.11, but since we are here for clarifying things, let us do this again, with no reference to that proof. Consider indeed a degree 2 polynomial, written as follows:

$$f(x) = a + bx + cx^2$$

We have then the following formula, for the function to be approximated:

$$\begin{aligned} f(x+t) &= a + b(x+t) + c(x+t)^2 \\ &= a + bx + bt + cx^2 + ct^2 + 2cxt \end{aligned}$$

As for the Taylor approximation at order 2, this is given by the following formula:

$$\begin{aligned} f(x) + f'(x)t + \frac{f''(x)}{2}\,t^2 &= (a + bx + cx^2) + (b + 2cx)t + ct^2 \\ &= a + bx + cx^2 + bt + 2cxt + ct^2 \end{aligned}$$

We conclude that in this case we have, as claimed:

$$f(x+t) = f(x) + f'(x)t + \frac{f''(x)}{2}\,t^2$$

Finally, observe that the converse of this holds too, trivially, because the Taylor approximation at order 2 is by definition a certain degree 2 polynomial. □

Still talking polynomials, let us have a look as well at what happens for the degree 3 polynomials. Here the result, which is quite instructive, is as follows:

PROPOSITION 10.13. *For a degree* 3 *polynomial, written as*

$$f(x) = a + bx + cx^2 + dx^3$$

we have the following formula, valid for any x *and any* t*,*

$$f(x+t) = f(x) + f'(x)t + \frac{f''(x)}{2}\,t^2 + dt^3$$

so the error in the Taylor approximation formula at order 2 *is* $\varepsilon(t) = dt^3$*.*

PROOF. Consider indeed a degree 3 polynomial, as in the statement:

$$f(x) = a + bx + cx^2 + dx^3$$

As a first observation, when looking at the remainder in the order 2 Taylor formula, according to Proposition 10.12 and by linearity we can neglect if we want the low order terms, those of order $0, 1, 2$, which means in practice that we can assume $f(x) = dx^3$.

However, for full transparency, we will do the computation for a general f, as above. We have the following formula, for the function to be approximated:

$$\begin{aligned} & f(x+t) \\ = \ & a+b(x+t)+c(x+t)^2+d(x+t)^3 \\ = \ & a+bx+bt+cx^2+ct^2+2cxt+dx^3+dt^3+3dxt^2+3dx^2t \end{aligned}$$

As for the Taylor approximation at order 2, this is given by the following formula:

$$\begin{aligned} & f(x)+f'(x)t+\frac{f''(x)}{2}\,t^2 \\ = \ & (a+bx+cx^2+dx^3)+(b+2cx+3dx^2)t+(c+3dx)t^2 \\ = \ & a+bx+cx^2+dx^3+bt+2cxt+3dx^2t+ct^2+3dxt^2 \end{aligned}$$

Thus, the error term is given by the following formula:

$$\begin{aligned} \varepsilon(t) &= f(x+t)-f(x)-f'(x)t-\frac{f''(x)}{2}\,t^2 \\ &= a+bx+bt+cx^2+ct^2+2cxt+dx^3+dt^3+3dxt^2+3dx^2t \\ &\quad -a-bx-cx^2-dx^3-bt-2cxt-3dx^2t-ct^2-3dxt^2 \\ &= dt^3 \end{aligned}$$

We are therefore led to the conclusion in the statement. □

We will be back to such things, error term, and further improvements of the Taylor formula, in chapter 10 below, and also later in Part IV, when doing integration.

Getting now to applications of the Taylor formula, we have the following key result:

THEOREM 10.14. *The local minima and local maxima of a twice differentiable function $f:\mathbb{R}\to\mathbb{R}$ appear at the points $x\in\mathbb{R}$ where*

$$f'(x)=0$$

with the local minima coming from the case $f'(x)\geq 0$, and the local maxima coming from the case $f''(x)\leq 0$.

PROOF. The first assertion is something that we already know. As for the second assertion, we can use the formula in Theorem 10.11, which in the case $f'(x)=0$ reads:

$$f(x+t)\simeq f(x)+\frac{f''(x)}{2}\,t^2$$

Indeed, assuming $f''(x)\neq 0$, it is clear that the condition $f''(x)>0$ will produce a local minimum, and that the condition $f''(x)<0$ will produce a local maximum. □

As before with first derivatives, the above result is not the end of the story with the study of the local minima and maxima, because things are undetermined when:

$$f'(x) = f''(x) = 0$$

For instance the functions $\pm x^n$ with $n \in \mathbb{N}$ all satisfy this condition at $x = 0$, which is a minimum for the functions of type x^{2m}, a maximum for the functions of type $-x^{2m}$, and not a local minimum or local maximum for the functions of type $\pm x^{2m+1}$.

We will be back to such questions in the next chapter, with a more advanced discussion about this, by using higher derivatives, which provide the key to the answer.

There are some comments to be made as well in relation with the algorithm discussed in the previous chapter, for finding in practice the extrema of the function. Normally that algorithm stays strong, because Theorem 10.14 can only help in relation with the final steps, and is it worth it to compute the second derivative f'', just for getting rid of roughly $1/2$ of the $f(x)$ values to be compared. However, in certain cases, this method proves to be useful, so Theorem 10.14 is good to know, when applying that algorithm.

As a second application now of the Taylor formula, justifying Interpretation 10.2 (3), we have the following result, which is of course a bit heuristic:

PROPOSITION 10.15. *Intuitively speaking, the second derivative $f''(x) \in \mathbb{R}$ computes how much different is $f(x)$, compared to the average of $f(z)$, with $z \simeq x$.*

PROOF. As already mentioned, this is something a bit heuristic, but which is good to know. Let us write the formula in Theorem 10.11, as such, and with $t \to -t$ too:

$$f(x+t) \simeq f(x) + f'(x)t + \frac{f''(x)}{2}\,t^2$$

$$f(x-t) \simeq f(x) - f'(x)t + \frac{f''(x)}{2}\,t^2$$

By making the average, we obtain the following formula:

$$\frac{f(x+t) + f(x-t)}{2} \simeq f(x) + \frac{f''(x)}{2}\,t^2$$

Now assume that we have found a way of averaging things over $t \in [-\varepsilon, \varepsilon]$, with the corresponding averages being denoted I. We obtain from the above:

$$I(f) \simeq f(x) + f''(x) I\left(\frac{t^2}{2}\right)$$

But this is what our statement says, save for some uncertainties regarding the averaging method, and for the precise value of $I(t^2/2)$. We will leave this for later. □

Ready for some physics? Nothing better than this, in order to understand the second derivatives. To start with, we can talk about free falls in 1 dimension, as follows:

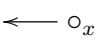
THEOREM 10.16. *The equation of a gravitational free fall, in* 1 *dimension, is*

$$\ddot{x} = -\frac{GM}{x^2}$$

with dot denoting time derivatives, M being the attracting mass, and $G \simeq 6.674 \times 10^{-11}$.

PROOF. Assume indeed that we have a free falling object, in 1 dimension:

In order to reach to calculus as we know it, we must peform a rotation, as to have all this happening on the Ox axis. By doing this, and assuming that M is fixed at 0, our picture becomes as follows, with the attached numbers being now the coordinates:



Now comes the physics. The gravitational force exterted by M, which is fixed in our formalism, on the object m which moves, is subject to the following equations:

$$F = -G \cdot \frac{Mm}{x^2} \quad , \quad F = ma \quad , \quad a = \dot{v} \quad , \quad v = \dot{x}$$

Now observe that, with the above data for F, the equation $F = ma$ reads:

$$-G \cdot \frac{Mm}{x^2} = m\ddot{x}$$

Thus, by simplifying, we are led to the equation in the statement. □

As more phsyics, we can talk as well about waves in 1 dimension, as follows:

THEOREM 10.17. *The wave equation in* 1 *dimension is*

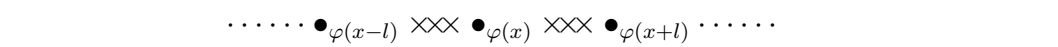
$$\ddot{\varphi} = v^2\varphi''$$

with the dot denoting time derivatives, and $v > 0$ being the propagation speed.

PROOF. This is not exactly a theorem, but rather what comes out of physics experiments, but we can justify that findings mathematically, as follows:

(1) In order to understand the propagation of waves, let us model the space $\mathbb{R}$ as a network of balls, with springs between them, as follows:

Now let us send an impulse, and see how balls will be moving. For this purpose, we zoom on one ball. The situation here is as follows, l being the spring length:



We have two forces acting at x. First is the Newton motion force, mass times acceleration, which is as follows, with m being the mass of each ball:

$$F_n = m \cdot \ddot{\varphi}(x)$$

And second is the Hooke force, displacement of the spring, times spring constant. Since we have two springs at x, this is as follows, k being the spring constant:

$$\begin{aligned} F_h &= F_h^r - F_h^l \\ &= k(\varphi(x+l) - \varphi(x)) - k(\varphi(x) - \varphi(x-l)) \\ &= k(\varphi(x+l) - 2\varphi(x) + \varphi(x-l)) \end{aligned}$$

We conclude that the equation of motion, in our model, is as follows:

$$m \cdot \ddot{\varphi}(x) = k(\varphi(x+l) - 2\varphi(x) + \varphi(x-l))$$

(2) Now let us take the limit of our model, as to reach to continuum. For this purpose we will assume that our system consists of $N >> 0$ balls, having a total mass M, and spanning a total distance L. Thus, our previous infinitesimal parameters are as follows, with K being the spring constant of the total system, which is of course lower than k:

$$m = \frac{M}{N} \quad , \quad k = KN \quad , \quad l = \frac{L}{N}$$

With these changes, our equation of motion found in (1) reads:

$$\ddot{\varphi}(x) = \frac{KN^2}{M}(\varphi(x+l) - 2\varphi(x) + \varphi(x-l))$$

Now observe that this equation can be written, more conveniently, as follows:

$$\ddot{\varphi}(x) = \frac{KL^2}{M} \cdot \frac{\varphi(x+l) - 2\varphi(x) + \varphi(x-l)}{l^2}$$

With $N \to \infty$, and therefore $l \to 0$, we obtain in this way:

$$\ddot{\varphi}(x) = \frac{KL^2}{M} \cdot \frac{d^2\varphi}{dx^2}(x)$$

Thus, we are led to the conclusion in the statement. □

Along the same lines, we can talk as well about heat in 1D, as follows:

THEOREM 10.18. *The heat diffusion equation in* 1 *dimension is*

$$\dot{\varphi} = \alpha\varphi''$$

where $\alpha > 0$ *is the thermal diffusivity of the medium.*

PROOF. As before with the wave equation, this is not exactly a theorem, but rather what comes out of experiments, but we can justify this mathematically, as follows:

(1) As an intuitive explanation for this equation, since the second derivative φ'' computes the average value of a function φ around a point, minus the value of φ at that point, as we know from Proposition 10.15, the heat equation as formulated above tells us that the rate of change $\dot{\varphi}$ of the temperature of the material at any given point must be proportional, with proportionality factor $\alpha > 0$, to the average difference of temperature between that given point and the surrounding material. Which sounds reasonable.

(2) In practice now, we can use, a bit like before for the wave equation, a lattice model as follows, with distance $l > 0$ between the neighbors:

$$\text{—— } \circ_{x-l} \overset{l}{\text{——}} \circ_x \overset{l}{\text{——}} \circ_{x+l} \text{ ——}$$

In order to model now heat diffusion, we have to implement the intuitive mechanism explained above, and in practice, this leads to a condition as follows, expressing the change of the temperature φ, over a small period of time $\delta > 0$:

$$\varphi(x, t+\delta) = \varphi(x,t) + \frac{\alpha\delta}{l^2} \sum_{x \sim y} [\varphi(y,t) - \varphi(x,t)]$$

(3) Now let us do the math. In the context of our 1D model the neighbors of x are the points $x \pm l$, and so the equation that we wrote above takes the following form:

$$\frac{\varphi(x,t+\delta) - \varphi(x,t)}{\delta} = \frac{\alpha}{l^2} \Big[(\varphi(x+l,t) - \varphi(x,t)) + (\varphi(x-l,t) - \varphi(x,t)) \Big]$$

Now observe that we can write this equation as follows:

$$\frac{\varphi(x,t+\delta) - \varphi(x,t)}{\delta} = \alpha \cdot \frac{\varphi(x+l,t) - 2\varphi(x,t) + \varphi(x-l,t)}{l^2}$$

(4) As it was the case with the wave equation before, we recognize on the right the usual approximation of the second derivative, coming from calculus. Thus, when taking the continuous limit of our model, $l \to 0$, we obtain the following equation:

$$\frac{\varphi(x,t+\delta) - \varphi(x,t)}{\delta} = \alpha \cdot \varphi''(x,t)$$

Now with $t \to 0$, we are led in this way to the heat equation in the statement. □

All this is very nice, so with the calculus that we know, we can certainly talk about physics. We will see later in this book how to deal with the above equations.

10d. Convex functions

As a main concrete application now of the second derivative, which is something very useful in practice, and related to Interpretation 10.2 (4), we have the following result:

THEOREM 10.19. *Given a convex function $f : \mathbb{R} \to \mathbb{R}$, we have the following Jensen inequality, for any $x_1, \ldots, x_N \in \mathbb{R}$, and any $\lambda_1, \ldots, \lambda_N > 0$ summing up to 1,*

$$f(\lambda_1 x_1 + \ldots + \lambda_N x_N) \leq \lambda_1 f(x_1) + \ldots + \lambda_N x_N$$

with equality when $x_1 = \ldots = x_N$. In particular, by taking the weights λ_i to be all equal, we obtain the following Jensen inequality, valid for any $x_1, \ldots, x_N \in \mathbb{R}$,

$$f\left(\frac{x_1 + \ldots + x_N}{N}\right) \leq \frac{f(x_1) + \ldots + f(x_N)}{N}$$

and once again with equality when $x_1 = \ldots = x_N$. A similar statement holds for the concave functions, with all the inequalities being reversed.

PROOF. This is indeed something quite routine, the idea being as follows:

(1) First, we can talk about convex functions in a usual, intuitive way, with this meaning by definition that the following inequality must be satisfied:

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x) + f(y)}{2}$$

(2) But this means, via a simple argument, by approximating numbers $t \in [0, 1]$ by sums of powers 2^{-k}, that for any $t \in [0, 1]$ we must have:

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

Alternatively, via yet another simple argument, this time by doing some geometry with triangles, this means that we must have:

$$f\left(\frac{x_1 + \ldots + x_N}{N}\right) \leq \frac{f(x_1) + \ldots + f(x_N)}{N}$$

But then, again alternatively, by combining the above two simple arguments, the following must happen, for any $\lambda_1, \ldots, \lambda_N > 0$ summing up to 1:

$$f(\lambda_1 x_1 + \ldots + \lambda_N x_N) \leq \lambda_1 f(x_1) + \ldots + \lambda_N x_N$$

(3) Summarizing, all our Jensen inequalities, at $N = 2$ and at $N \in \mathbb{N}$ arbitrary, are equivalent. The point now is that, if we look at what the first Jensen inequality, that we took as definition for the convexity, exactly means, this is simply equivalent to:

$$f''(x) \geq 0$$

(4) Thus, we are led to the conclusions in the statement, regarding the convex functions. As for the concave functions, the proof here is similar. Alternatively, we can say that f is concave precisely when $-f$ is convex, and get the results from what we have. □

As a basic application of the Jensen inequality, which is very classical, we have:

THEOREM 10.20. *For any $p \in (1,\infty)$ we have the following inequality,*

$$\left|\frac{x_1+\ldots+x_N}{N}\right|^p \leq \frac{|x_1|^p+\ldots+|x_N|^p}{N}$$

and for any $p \in (0,1)$ we have the following inequality,

$$\left|\frac{x_1+\ldots+x_N}{N}\right|^p \geq \frac{|x_1|^p+\ldots+|x_N|^p}{N}$$

with in both cases equality precisely when $|x_1| = \ldots = |x_N|$.

PROOF. This follows indeed from Theorem 10.19, because we have:

$$(x^p)'' = p(p-1)x^{p-2}$$

Thus x^p is convex for $p > 1$ and concave for $p < 1$, which gives the results. □

Observe that at $p = 2$ we obtain as particular case of the above inequality the Cauchy-Schwarz inequality, or rather something equivalent to it, namely:

$$\left(\frac{x_1+\ldots+x_N}{N}\right)^2 \leq \frac{x_1^2+\ldots+x_N^2}{N}$$

As yet another important application of the Jensen inequality, we have:

THEOREM 10.21. *We have the Young inequality,*

$$ab \leq \frac{a^p}{p}+\frac{b^q}{q}$$

valid for any $a,b \geq 0$, and any exponents $p,q > 1$ satisfying $\frac{1}{p}+\frac{1}{q} = 1$.

PROOF. We use the logarithm function, which is concave on $(0,\infty)$, due to:

$$(\log x)'' = \left(-\frac{1}{x}\right)' = -\frac{1}{x^2}$$

Thus we can apply the Jensen inequality, and we obtain in this way:

$$\begin{aligned}\log\left(\frac{a^p}{p}+\frac{b^q}{q}\right) &\geq \frac{\log(a^p)}{p}+\frac{\log(b^q)}{q}\\ &= \log(a)+\log(b)\\ &= \log(ab)\end{aligned}$$

Now by exponentiating, we obtain the Young inequality. □

Moving forward now, as a consequence of the Young inequality, we have:

THEOREM 10.22 (Hölder). *Assuming that $p, q \geq 1$ are conjugate, in the sense that*

$$\frac{1}{p} + \frac{1}{q} = 1$$

we have the following inequality, valid for any two vectors $x, y \in \mathbb{C}^N$,

$$\sum_i |x_i y_i| \leq \left(\sum_i |x_i|^p \right)^{1/p} \left(\sum_i |y_i|^q \right)^{1/q}$$

with the convention that an ∞ exponent produces a $\max |x_i|$ quantity.

PROOF. This is something very standard, the idea being as follows:

(1) Assume first that we are dealing with finite exponents, $p, q \in (1, \infty)$. By linearity we can assume that x, y are normalized, in the following way:

$$\sum_i |x_i|^p = \sum_i |y_i|^q = 1$$

In this case, we want to prove that the following inequality holds:

$$\sum_i |x_i y_i| \leq 1$$

For this purpose, we use the Young inequality, which gives, for any i:

$$|x_i y_i| \leq \frac{|x_i|^p}{p} + \frac{|y_i|^q}{q}$$

By summing now over $i = 1, \ldots, N$, we obtain from this, as desired:

$$\begin{aligned} \sum_i |x_i y_i| &\leq \sum_i \frac{|x_i|^p}{p} + \sum_i \frac{|y_i|^q}{q} \\ &= \frac{1}{p} + \frac{1}{q} \\ &= 1 \end{aligned}$$

(2) In the case $p = 1$ and $q = \infty$, or vice versa, the inequality holds too, trivially, with the convention that an ∞ exponent produces a max quantity, according to:

$$\lim_{p \to \infty} \left(\sum_i |x_i|^p \right)^{1/p} = \max |x_i|$$

Thus, we are led to the conclusion in the statement. □

As a consequence now of the Hölder inequality, we have:

THEOREM 10.23 (Minkowski). *Assuming $p \in [1,\infty]$, we have the inequality*

$$\left(\sum_i |x_i+y_i|^p\right)^{1/p} \leq \left(\sum_i |x_i|^p\right)^{1/p} + \left(\sum_i |y_i|^p\right)^{1/p}$$

for any two vectors $x, y \in \mathbb{C}^N$, with our usual conventions at $p = \infty$.

PROOF. We have indeed the following estimate, using the Hölder inequality, and the conjugate exponent $q \in [1,\infty]$, given by $1/p + 1/q = 1$:

$$\begin{aligned}
\sum_i |x_i+y_i|^p &= \sum_i |x_i+y_i| \cdot |x_i+y_i|^{p-1} \\
&\leq \sum_i |x_i| \cdot |x_i+y_i|^{p-1} + \sum_i |y_i| \cdot |x_i+y_i|^{p-1} \\
&\leq \left(\sum_i |x_i|^p\right)^{1/p} \left(\sum_i |x_i+y_i|^{(p-1)q}\right)^{1/q} \\
&+ \left(\sum_i |y_i|^p\right)^{1/p} \left(\sum_i |x_i+y_i|^{(p-1)q}\right)^{1/q} \\
&= \left[\left(\sum_i |x_i|^p\right)^{1/p} + \left(\sum_i |y_i|^p\right)^{1/p}\right] \left(\sum_i |x_i+y_i|^p\right)^{1-1/p}
\end{aligned}$$

Here we have used the following fact, at the end:

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

Now by dividing both sides by the last quantity at the end, we obtain:

$$\left(\sum_i |x_i+y_i|^p\right)^{1/p} \leq \left(\sum_i |x_i|^p\right)^{1/p} + \left(\sum_i |y_i|^p\right)^{1/p}$$

Thus, we are led to the conclusion in the statement. □

Good news, done with inequalities, and as a consequence of this, we have:

THEOREM 10.24. *Given an exponent $p \in [1,\infty]$, the formula*

$$||x||_p = \left(\sum_i |x_i|^p\right)^{1/p}$$

with usual conventions at $p = \infty$, defines a norm on $\mathbb{C}^N$.

PROOF. This follows indeed from the Minkowski inequality, which proves the triangle inequality for our norm, and with the other two norm axioms being both clear. □

Very nice all this, but you might wonder at this point, what is the relation of all this with functions. In answer, Theorem 10.24 can be reformulated as follows:

THEOREM 10.25. *Given an exponent $p \in [1, \infty]$, the formula*

$$||f||_p = \left(\sum_x |f(x)|^p \right)^{1/p}$$

defines a norm on the space of functions $f : \{1, \ldots, N\} \to \mathbb{C}$.

PROOF. This is a just fancy reformulation of Theorem 10.24, by using the fact that the space formed by the functions $f : \{1, \ldots, N\} \to \mathbb{C}$ is canonically isomorphic to $\mathbb{C}^N$. □

And with this, end of our excursion into functional analysis, clarifying some things that we previously said, in chapter 6. And more on this, later in this book.

10e. Exercises

Welcome to true calculus, which means second derivatives, which are something quite subtle, and extremely useful. As exercises on them, we have:

EXERCISE 10.26. *Further meditate on the various interpretations of f''.*

EXERCISE 10.27. *Find other functions which are differentiable once, but not twice.*

EXERCISE 10.28. *Fill in all the details in our trigonometric function computations.*

EXERCISE 10.29. *Learn more about L'Hôpital's rule, and its various applications.*

EXERCISE 10.30. *Find a geometric interpretation of the Taylor formula at order 2.*

EXERCISE 10.31. *Learn more, via examples and counterexamples, about $f''(x) = 0$.*

EXERCISE 10.32. *Clarify what we said above, in regards with convex functions.*

EXERCISE 10.33. *Prove the Cauchy-Schwarz inequality over $\mathbb{C}$, directly.*

As bonus exercise, with what we learned here, you are good to read some basic, introductory functional analysis. That would be some very useful learning.

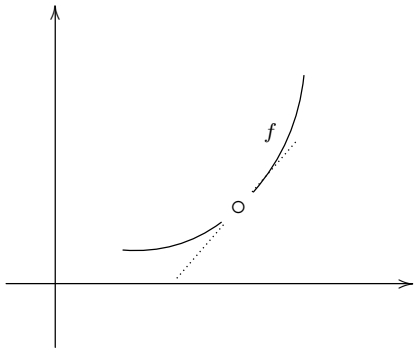
CHAPTER 11

Taylor formula

11a. Higher derivatives

Back now to the general theory of the derivatives, and their applications, we can further develop our basic approximation method, at order 3, at order 4, and so on. Discussing all this, and the general Taylor formula emerging for this, will be our purpose here.

Let us start with some general discussion. We know that $f'(x)$ expresses the growth of $f(x)$, and that $f''(x)$ expresses the growth of $f'(x)$, and this looks quite complete. However, it may happen that the function $f(x)$ sharply grows, in the sense that the growth of the slope $f''(x)$ grows itself, as shown on the following picture:



In this case, which happens for instance for $f(x) = x^3$, or $f(x) = x^4$, or $f(x) = e^x$, our picture using $f'(x)$ and $f''(x)$ is not complete, and we must take into account the third derivative $f'''(x)$, which expresses the growth of the second derivative $f''(x)$.

Moreover, some further thinking along these lines, say with the function $f(x) = e^x$ in mind, suggests that, for certain delicate questions, we might need as well derivatives of order four, $f''''(x)$, or higher. So, let us formulate the following definition:

DEFINITION 11.1. *We say that $f : \mathbb{R} \to \mathbb{R}$ is n times differentiable if*

$$f', f'', f''', \ldots, f^{(n)}$$

exist, with each $f^{(k)}$ being by definition the derivative of $f^{(k-1)}$.

Here the same comments as in chapter 9 apply. To be more precise, differentiability of a function f at a given point $x \in \mathbb{R}$ is local property, only needing f to be defined on a small interval around x. Thus, we can differentiate in fact functions $f : X \to \mathbb{R}$ with $X \subset \mathbb{R}$ open, and our theory below will tacitly use this assumption. With the remark that it is sometimes convenient to make an exception for the functions of type $f : [a, b] \to \mathbb{R}$, whose differentiability at a, b can be dealt with by using right and left limits.

In short, quite complicated all this, and as mentioned in chapter 9, with this being an introduction to calculus, and not a calculus treatise, enjoy this at it comes, and later, if really loving calculus, upgrade to something more rigorous, say Rudin [73], [74].

Getting to work now, it is tempting to take Definition 11.1 as such, involving a new parameter $n \in \mathbb{N}$, and start doing the math, in analogy with what we already know about f' and f'', corresponding to the cases $n = 1, 2$. However, this would most likely lead to some sort of algebra, not very understandable. For doing things slowly and correctly, in the analytic way, we should better relax, and answer first the following question:

QUESTION 11.2. *We know that the second derivative f'' can be interpreted in a variety of interesting ways. Can we have a similar understanding of f''', and of f''''?*

In short, my proposal would be to focus on the cases $n = 3, 4$ first, with some work here, and leave the general theory, and the case where $n \in \mathbb{N}$ is arbitrary, for later.

In answer now, at $n = 3$, the third derivative f''' is something quite intuitive and understandable, and this because we can invoke some familiar physics, as follows:

FACT 11.3. *In analogy with the fact that the second derivative f'' measures the acceleration of the slope f', the third derivative f''' measures the jerk of the slope f'.*

To be more precise, all this comes from real life and classical mechanics, where the jerk, a familiar quantity, is by definition the derivative of the acceleration:

$$j = \dot{a}$$

Now since the Newton laws tell us that the acceleration is the derivative of the speed, $a = \dot{v}$, and the speed is the derivative of the position, $v = \dot{x}$, we conclude that the jerk is the third derivative of the position, according to the following computation:

$$j = \dot{a} = \ddot{v} = \dddot{x}$$

In a word, we have reached here to the conclusions from Fact 11.3. As before with second derivatives, in relation with acceleration, many other things can be said, along these lines. We will be back to this later, when discussing physics and applications.

Getting started now with some mathematical study, let us first record the formulae of the third derivatives of the basic functions, which are as follows:

THEOREM 11.4. *The third derivatives of the basic functions are as follows:*

(1) $(x^p)''' = p(p-1)(p-2)x^{p-3}$.
(2) $\sin''' = -\cos$.
(3) $\cos''' = \sin$.
(4) $\exp''' = \exp$.
(5) $\log'''(x) = 2/x^3$.
(6) $\sinh''' = \cosh$.
(7) $\cosh''' = \sinh$.

PROOF. The various formulae in the statement all follow from the various formulae for the second derivatives established before, in chapter 10, as follows:

$$(x^p)''' = (p(p-1)x^{p-2})' = p(p-1)(p-2)x^{p-3}$$

$$(\sin x)''' = (-\sin x)' = -\cos x$$

$$(\cos x)''' = (-\cos x)' = \sin x$$

$$(e^x)''' = (e^x)' = e^x$$

$$(\log x)''' = (-1/x^2)' = 2/x^3$$

$$(\sinh x)''' = (\sinh x)' = \cosh x$$

$$(\cosh x)''' = (\cosh x)' = \sinh x$$

Thus, we are led to the formulae in the statement. □

It is possible of course to come up as well with formulae for the third derivatives of the other trigonometric functions, and we will leave this as an instructive exercise.

Getting now to the fourth derivatives, things are less intuitive here, in what regards the interpretation, but we can nevertheless formulate, as some sort of conjecture:

FACT 11.5. *The fourth derivative f'''' measures the acceleration of the jerk f''', and with this being best felt in the context of various catastrophic events.*

In short, basic physics cannot really help us, in understanding the fourth derivatives. This being said, math comes to the rescue, and we can do a few computations:

THEOREM 11.6. *The fourth derivatives of the basic functions are as follows:*

(1) $(x^p)'''' = p(p-1)(p-2)(p-3)x^{p-4}$.
(2) $\sin'''' = \sin$.
(3) $\cos'''' = \cos$.
(4) $\exp'''' = \exp$.
(5) $\log''''(x) = -6/x^4$.
(6) $\sinh''' = \sinh$.
(7) $\cosh''' = \cosh$.

PROOF. The various formulae in the statement all follow from the various formulae for the third derivatives established before, as follows:

$$(x^p)'''' = (p(p-1)(p-2)x^{p-3})' = p(p-1)(p-2)(p-3)x^{p-4}$$
$$(\sin x)'''' = (-\cos x)' = \sin x$$
$$(\cos x)'''' = (\sin x)' = \cos x$$
$$(e^x)'''' = (e^x)' = e^x$$
$$(\log x)'''' = (2/x^3)' = -6/x^4$$
$$(\sinh x)'''' = (\cosh x)' = \sinh x$$
$$(\cosh x)'''' = (\sinh x)' = \cosh x$$

Thus, we are led to the formulae in the statement. □

It is possible of course to come up as well with formulae for the fourth derivatives of the other trigonometric functions, and we will leave this as an instructive exercise.

Observe now the magic brought by the fourth derivative at the level of basic trigonometric functions. This is perhaps something worth recording, as follows:

OBSERVATION 11.7. *The fourth derivative brings some periodicity magic at the level of basic trigonometric functions.*

In view of this, which looks interesting, let us see as well what happens for the tangent. However, the result here is as follows, contradicting this observation:

THEOREM 11.8. *The first two derivatives of the tangent function are*

$$(\tan x)' = \frac{1}{\cos^2 x} \quad , \quad (\tan x)'' = \frac{2\sin x}{\cos^3 x}$$

and the third and fourth derivatives are

$$(\tan x)''' = \frac{2 + 4\sin^2 x}{\cos^4 x} \quad , \quad (\tan x)'''' = \frac{16\sin x + 8\sin^3 x}{\cos^5 x}$$

provided that the denominators do not vanish.

PROOF. We already know the first two formulae, from chapters 9 and 10. Regarding now the third formula, the computation here goes as follows:

$$\begin{aligned}
(\tan x)''' &= \left(\frac{2\sin x}{\cos^3 x}\right)' \\
&= \frac{2\cos x \cdot \cos^3 x + 2\sin x \cdot 3\cos^2 x \sin x}{\cos^6 x} \\
&= \frac{2\cos^2 x + 6\sin^2 x}{\cos^4 x} \\
&= \frac{2 + 4\sin^2 x}{\cos^4 x}
\end{aligned}$$

As for the fourth formula, the computation here is as follows:

$$\begin{aligned}(\tan x)''''&=\left(\frac{2+4\sin^2x}{\cos^4x}\right)'\\&=\frac{8\sin x\cos x\cdot\cos^4x+(2+4\sin^2x)\cdot 4\cos^3x\sin x}{\cos^8x}\\&=\frac{8\sin x\cos^2x+(2+4\sin^2x)\cdot 4\sin x}{\cos^5x}\\&=\frac{8\sin x(1-\sin^2x)+(2+4\sin^2x)\cdot 4\sin x}{\cos^5x}\\&=\frac{16\sin x+8\sin^3x}{\cos^5x}\end{aligned}$$

Thus, we are led to the conclusions in the statement. □

As a conclusion to this, Observation 11.7 seems to be something rather superficial. Also, in relation with higher derivatives, beware of the tangent.

So long for our preliminary discussion of third and fourth derivatives, with Question 11.2 in mind. This being said, I don't know about you, but personally I feel quite frustrated of having nothing intuitive regarding the fourth derivative. So, let us think about this, some more. Here is an interesting speculation that can be made, inspired by our main interpretation of the second derivative, as an average, from chapter 10:

SPECULATION 11.9. *Assuming that the fourth differentiable functions satisfy*

$$f(x+t)\simeq f(x)+at+bt^2+ct^3+dt^4$$

with d being a certain multiple of $f''''(x)$, we would have the formula

$$\frac{f(x+t)+f(x+it)+f(x-t)+f(x-it)}{4}\simeq f(x)+dt^4$$

telling us that $f''''(x)$ measures how far is $f(z)$ with $z\simeq x$ from $f(x)$, over $\mathbb{C}$.

To be more precise here, in what regards our assumption from the beginning, this is something quite natural, in view of the Taylor formula, that we already have at order 1 and 2, and we will see in a moment that this is indeed the case. As for the average formula given above, this comes from this, via the following computation:

$$\begin{aligned}f(x+t)+f(x+it)+f(x-t)+f(x-it)&\simeq f(x)+at+bt^2+ct^3+dt^4\\&+f(x)+iat-bt^2-ict^3+dt^4\\&+f(x)-at+bt^2-ct^3+dt^4\\&+f(x)-iat-bt^2+ict^3+dt^4\\&=4f(x)+4dt^4\end{aligned}$$

Which sounds very nice, it looks like, save for the fact that our real functions $f : \mathbb{R} \to \mathbb{R}$ do not necessarily extend into complex functions $f : \mathbb{C} \to \mathbb{C}$, but let us not bother for the moment with this, with most of our functions $f : \mathbb{R} \to \mathbb{R}$ having this property, we have, eventually, something interesting regarding the meaning of the fourth derivative.

However, not so quick. Indeed, thinking a bit, we can in fact do the same at order 3, and we reach here to the following counterspeculation, ruining everything:

COUNTERSPECULATION 11.10. *Assuming that for the third differentiable functions*

$$f(x+t) \simeq f(x) + at + bt^2 + ct^3$$

with c being a certain multiple of $f'''(x)$, we would have, with $w = e^{2\pi i/3}$,

$$\frac{f(x+t) + f(x+wt) + f(x+w^2t)}{3} \simeq f(x) + ct^3$$

telling us that $f'''(x)$ measures too how far is $f(z)$ with $z \simeq x$ from $f(x)$, over $\mathbb{C}$.

And weird thing this is, with our Taylor series assumption being something quite natural, as before, and with the average computation being also as before, namely:

$$\begin{aligned} f(x+t) + f(x+wt) + f(x+w^2t) &\simeq f(x) + at + bt^2 + ct^3 \\ &+ f(x) + wat + w^2bt^2 + ct^3 \\ &+ f(x) + w^2at + wbt^2 + ct^3 \\ &= 3f(x) + 3ct^3 \end{aligned}$$

As a conclusion to all this, the third and fourth derivatives appear to be something quite mysterious, mathematically speaking, and in order to intuitively understand them, we are left with the notion of jerk, and the related notion of acceleration of jerk, from Fact 11.3 and Fact 11.5. Well, nevermind. This is not that bad, and we will see in what follows that we can have some interesting theory going, relying only on this.

11b. Taylor formula

With this discussed, and getting back now to our usual approximation business, the ultimate result on the subject, called Taylor formula, is as follows:

THEOREM 11.11. *Assuming that $f : \mathbb{R} \to \mathbb{R}$ is differentiable n times, we have*

$$f(x+t) \simeq \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k$$

where $f^{(k)}(x)$ are the higher derivatives of f at the point x.

PROOF. Consider indeed the function to be approximated, namely:

$$\varphi(t) = f(x+t)$$

Now, let us try to best approximate this function at order n. We are therefore looking for a certain polynomial in t, of the following type:

$$P(t) = a_0 + a_1 t + \ldots + a_n t^n$$

The natural conditions to be imposed are those stating that P and the function φ should match at $t = 0$, at the level of the actual value, of the derivative, second derivative, and so on up the n-th derivative. But, these conditions are as follows:

$$a_0 = f(x) \quad , \quad a_1 = f'(x) \quad , \quad 2a_2 = f''(x) \quad , \quad \ldots \quad , \quad n!a_n = f^{(n)}(x)$$

We are therefore led to the approximation formula in the statement, namely:

$$f(x+t) \simeq \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k$$

In order to prove now that this approximation holds indeed, we can use L'Hôpital's rule, applied several times, exactly as we did in the proof at order 2. To be more precise, if we denote by $\varphi(t) \simeq P(t)$ the approximation to be proved, we have:

$$\begin{aligned}
\frac{\varphi(t) - P(t)}{t^n} &\simeq \frac{\varphi'(t) - P'(t)}{nt^{n-1}} \\
&\simeq \frac{\varphi''(t) - P''(t)}{n(n-1)t^{n-2}} \\
&\vdots \\
&\simeq \frac{\varphi^{(n)}(t) - P^{(n)}(t)}{n!} \\
&= \frac{f^{(n)}(x) - f^{(n)}(x)}{n!} \\
&= 0
\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Here is a related interesting statement, inspired from the above proof:

PROPOSITION 11.12. *For a polynomial of degree n, the Taylor approximation*

$$f(x+t) \simeq \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k$$

is an equality. The converse of this statement holds too.

PROOF. By linearity, it is enough to check the equality in question for the monomials $f(x) = x^p$, with $p \leq n$. But here, the formula to be proved is as follows:

$$(x+t)^p = \sum_{k=0}^{p} \frac{p(p-1)\ldots(p-k+1)}{k!}\, x^{p-k}t^k$$

We recognize the binomial formula, so our result holds indeed. As for the converse, this is clear, because the Taylor approximation is a polynomial of degree n. □

In order to further comment now on Theorem 11.11, which remains something quite subtle, it is perhaps time to clarify our meaning of $\simeq$. We have been using this sign, since the beginning of this book, for approximation in a general, intuitive sense. However, since in Theorem 11.11 we have all kinds of infinitesimals appearing, namely $t, t^2, \ldots, t^n$, this intuitive sign $\simeq$ is no longer appropriate, and we must invent something better.

In answer, for such things, nothing beats the Landau o and O notations:

DEFINITION 11.13. *We use the Landau o and O notations, as follows:*

(1) $f(t) = g(t) + o(h(t))$ *means* $o(h(t))/h(t) \to 0$, *with* $t \to 0$.
(2) $f(t) = g(t) + O(h(t))$ *means* $O(h(t))/h(t)$ *bounded, with* $t \simeq 0$.

Now with these conventions in hand, we can reformulate Theorem 11.11 in a more rigorous way, and which is more useful too, in practice, as follows:

THEOREM 11.14. *Assuming that $f : \mathbb{R} \to \mathbb{R}$ is differentiable n times we have*

$$f(x+t) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k + o(t^n)$$

and assuming that $f : \mathbb{R} \to \mathbb{R}$ is differentiable $n+1$ times we even have

$$f(x+t) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k + O(t^{n+1})$$

with o and O being the standard Landau symbols.

PROOF. This is a reformulation of Theorem 11.11, the idea being as follows:

(1) In what regards the first assertion, according to our convention for the Landau symbol o, at $n = 0$ this is the very definition of the continuity, according to:

$$f(x+t) = f(x) + o(1) \iff \frac{f(x+t) - f(x)}{1} \to 0$$

Next, at $n = 1$, our formula is the definition of differentiability, according to:

$$f(x+t) = f(x) + f'(x)t + o(t) \iff \frac{f(x+t) - f(x)}{t} - f'(x) \to 0$$

As for the case $n = 2$ and higher, here our formula is precisely what comes out of the proof of Theorem 11.11, after a close inspection.

(2) Regarding now the second assertion, which is finer, assuming that f is differentiable $n + 1$ times, this comes from the Taylor formula at order $n + 1$, which gives:

$$\begin{aligned} f(x+t) &= \sum_{k=0}^{n+1} \frac{f^{(k)}(x)}{k!}\, t^k + o(t^{n+1}) \\ &= \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k + \frac{f^{(n+1)}(x)}{(n+1)!}\, t^{n+1} + o(t^{n+1}) \\ &= \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k + O(t^{n+1}) + o(t^{n+1}) \\ &= \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k + O(t^{n+1}) \end{aligned}$$

Thus, we are led to the conclusions in the statement. □

As a last comment about Theorem 11.11, an interesting situation, which appears quite often, is that when f is infinitely differentiable. Here the result is as follows:

THEOREM 11.15. *Assuming that $f : \mathbb{R} \to \mathbb{R}$ is infinitely differentiable, we have*

$$f(x+t) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k + o(t^n)$$

for any $n \in \mathbb{N}$, according to the Taylor theorem. However, the asymptotic formula

$$f(x+t) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x)}{k!}\, t^k$$

might hold or not, depending on f, and generically, does not hold.

PROOF. This is something quite tricky, the idea being as follows:

(1) To start with, the first assertion is something that we know well.

(2) The second assertion is something more subtle, with the examples there abounding, including polynomials, or $\sin, \cos, \exp, \log$, as we will soon discover, and many more. However, we have as well counterexamples, with a standard counterexample being:

$$f(x) = \begin{cases} e^{-1/x^2} & (x \neq 0) \\ 0 & (x = 0) \end{cases}$$

Indeed, for this function we have the following estimate, valid for any $n \in \mathbb{N}$:

$$f(t) = (e^{1/t^2})^{-1} \leq \left(\frac{1/t^{2n}}{n!} \right)^{-1} = n! t^{2n} = o(t^n)$$

Thus f is infinitely differentiable at 0, with all its derivatives vanishing there, and so its Taylor series at 0 is the null series, which cannot be equal to f itself. That is, f is designed not to take off from 0, but it manages however to take off, very slowly.

(3) In what regards now the very last claim, this is something more technical, an intuitive explanation here being that there should be more functions $f : \mathbb{R} \to \mathbb{R}$, even taken infinitely differentiable, than series $\psi = \sum c_k t^k$. And in practice, up to you to learn here how to count such beasts, as an exercise, and reach to the above conclusion. □

Finally, getting back to Theorem 11.11 as stated, many interesting things can be said, about the remainder, in that approximation. We will be back to this later in this book, once we will have better techniques for investigating such questions.

In relation now with the local extrema, and getting back to our usual, informal $\simeq$ convention, in order to quickly explain what happens, we have the following result:

THEOREM 11.16. *Assuming that $f : \mathbb{R} \to \mathbb{R}$ is n times differentiable, and*

$$f(x+t) \simeq f(x) + \frac{f^{(n)}(x)}{n!} t^n$$

with $f^{(n)}(x) \neq 0$, this tells us if x is a local minimum or maximum of f.

PROOF. This is a quite compact statement, coming from the Taylor formula, the idea in practice being that we have an algorithm here, as follows:

(1) We can start with $n = 1$, and with the following formula, that we know well:

$$f(x+t) \simeq f(x) + f'(x)t$$

Indeed, this formula tells us that when $f'(x) \neq 0$, the point x cannot be a local minimum or maximum, due to the fact that $t \to -t$ will invert the growth.

(2) In the case left, $f'(x) = 0$, we switch to $n = 2$, where the Taylor formula is:

$$f(x+t) \simeq f(x) + \frac{f''(x)}{2} t^2$$

And here, when $f''(x) < 0$ we have a local maximum, and when $f''(x) > 0$ we have a local minimum. As for the remaining case, $f''(x) = 0$, things here remain open.

(3) In the case left, $f''(x) = 0$, we switch to $n = 3$, where the Taylor formula is:

$$f(x+t) \simeq f(x) + \frac{f'''(x)}{6} t^3$$

But this solves the problem in the case $f'''(x) \neq 0$, because here we cannot have a local minimum or maximum, due to $t \to -t$, which switches growth. As for the remaining case, $f'''(x) = 0$, things here remain open, and we have to go at higher order.

(4) Summarizing, we have a recurrence method for solving our problem. In order to comment now on what happens at the n-th step, let us write, as in the statement:

$$f(x+t) \simeq f(x) + \frac{f^{(n)}(x)}{n!}\, t^n$$

Then, when n is even, if $f^{(n)}(x) < 0$ we have a local maximum, and if $f^{(n)}(x) > 0$ we have a local minimum. As for the case where n is odd, here with $f^{(n)}(x) \neq 0$ we cannot have a local minimum or maximum, due to $t \to -t$ which switches growth.

(5) And so on, until the algorithm stops, either due to $f^{(n)}(x) \neq 0$, as it would be ideal, solving our problem, or due to the fact that $f^{(n)}$ is no longer differentiable at x, telling us that we have to use some alternative methods, or due the fact that we are facing a tricky function, resisting our algorithm until the very end, after $n = \infty$ steps, such as the function $f(x) = e^{-1/x^2}$, that we met in the proof of Theorem 11.15. □

Getting now to more concrete things, let us compute the Taylor series of the basic functions that we know. We first have here the following result:

THEOREM 11.17. *We have the following formulae for* sin *and* cos*,*

$$\sin t = \sum_{k=0}^{\infty} (-1)^k \frac{t^{2k+1}}{(2k+1)!} \quad , \quad \cos t = \sum_{k=0}^{\infty} (-1)^k \frac{t^{2k}}{(2k)!}$$

and the following formulae for exp *and* log*,*

$$e^t = \sum_{k=0}^{\infty} \frac{t^k}{k!} \quad , \quad \log(1+t) = \sum_{k=0}^{\infty} (-1)^{k+1} \frac{t^k}{k}$$

and the following formulae for sinh *and* cosh*,*

$$\sinh t = \sum_{k=0}^{\infty} \frac{t^{2k+1}}{(2k+1)!} \quad , \quad \cosh t = \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!}$$

as Taylor series, and in general as well, with $|t| < 1$ *needed for* log*.*

PROOF. There are several statements here, the proofs being as follows:

(1) Regarding the sine and cosine, we can use here the following formulae:

$$(\sin x)' = \cos x \quad , \quad (\cos x)' = -\sin x$$

Thus, we can differentiate $\sin, \cos$ as many times as we want to, and we get:

$$\sin^{(k)}(0) = \begin{cases} 0 & k = 0(4) \\ 1 & k = 1(4) \\ 0 & k = 2(4) \\ -1 & k = 3(4) \end{cases} \qquad \cos^{(k)}(0) = \begin{cases} 1 & k = 0(4) \\ 0 & k = 1(4) \\ -1 & k = 2(4) \\ 0 & k = 3(4) \end{cases}$$

But this leads to the Taylor series at $x = 0$ in the statement.

(2) Regarding now the exponential, nothing much to be proved here, because the exponential is given by definition by the formula in the statement, namely:

$$e^t = \sum_{k=0}^{\infty} \frac{t^k}{k!}$$

Observe that this is indeed a Taylor series, because from the formula $(e^x)' = e^x$ we deduce that we have $(e^x)^{(k)} = e^x$ for any k, which at $x = 0$ takes the value $e^0 = 1$. Thus, the Taylor series of the exponential is indeed the above one.

(3) Regarding now the logarithm, we can use here the following formulae:

$$(\log x)' = \frac{1}{x} \quad , \quad (x^p)' = px^{p-1}$$

Indeed, this shows that the derivatives of log are given by the following formulae:

$$(\log x)' = \frac{1}{x} \ , \ (\log x)'' = -\frac{1}{x^2} \ , \ (\log x)''' = \frac{2}{x^3} \ , \ (\log x)'''' = -\frac{6}{x^4} \ , \ \ldots$$

Thus, we get by recurrence the following formula, for the derivatives:

$$(\log x)^{(k)} = (-1)^{k+1} \frac{(k-1)!}{k^n}$$

Now by replacing the variable, $x \to 1 + x$, we obtain the following formula:

$$(\log(1+x))^{(k)} = (-1)^{k+1} \frac{(k-1)!}{(1+x)^k}$$

And finally, by setting $x = 0$, we obtain the following formula:

$$(\log(1+x))^{(k)}(0) = (-1)^{k+1}(k-1)!$$

But this gives the formula in the statement for the Taylor series.

(4) Regarding now sinh and cosh, we can use here the following formulae:

$$(\sinh x)' = \cosh x \quad , \quad (\cosh x)' = \sinh x$$

Indeed, as before for $\sin, \cos$, we can differentiate $\sinh, \cosh$ as many times as we want to, and we obtain Taylor series in the statement, similar to those for $\sin, \cos$.

(5) Finally, the fact that our various formulae extend beyond the small t setting, coming from Taylor series, as indicated in the statement, is something more subtle. Fortunately, we know all this, from our study from the previous chapters. □

As another basic illustration for the Taylor formula, we have:

THEOREM 11.18. *We have the following generalized binomial formula, with $p \in \mathbb{R}$,*

$$(x+t)^p = \sum_{k=0}^{\infty} \binom{p}{k} x^{p-k} t^k$$

with the generalized binomial coefficients being given by the formula

$$\binom{p}{k} = \frac{p(p-1)\ldots(p-k+1)}{k!}$$

valid for any $|t| < |x|$. With $p \in \mathbb{N}$, we recover the usual binomial formula.

PROOF. For the function $f(x) = x^p$, the derivatives are given by:

$$f^{(k)}(x) = p(p-1)\ldots(p-k+1)x^{p-k}$$

Thus, the Taylor approximation of our function is as follows:

$$f(x+t) = \sum_{k=0}^{\infty} \frac{p(p-1)\ldots(p-k+1)}{k!}\, x^{p-k} t^k$$

Which is precisely the formula in the statement. As for the fact that we have convergence for any $|t| < |x|$, this is something that we already know, from chapter 9. □

As a main application now of our generalized binomial formula, we have:

THEOREM 11.19. *We have the following formula,*

$$\sqrt{1+t} = 1 - 2\sum_{k=1}^{\infty} C_{k-1} \left(\frac{-t}{4}\right)^k$$

with $C_k = \frac{1}{k+1}\binom{2k}{k}$ being the Catalan numbers. Also, we have

$$\frac{1}{\sqrt{1+t}} = \sum_{k=0}^{\infty} D_k \left(\frac{-t}{4}\right)^k$$

with $D_k = \binom{2k}{k}$ being the central binomial coefficients.

PROOF. This is indeed something that we already know from before, and for any $|t| < 1$, coming from Theorem 11.18 applied at $p = 1/2$ and $p = -1/2$. □

11c. Arctangent, Leibnitz

Getting now to more specialized results, let us discuss the computation of the Taylor series of the other basic trigonometric functions. We will be interested, as usual in this book, in the fundamental 24 trigonometric functions, which are as follows:

sin	cos	tan	sec	csc	cot
arcsin	arccos	arctan	arcsec	arccsc	arccot
sinh	cosh	tanh	sech	csch	coth
arcsinh	arccosh	arctanh	arcsech	arccsch	arccoth

We already have 4 Taylor series, that of sin, cos, sinh, cosh, but in what regards the other 20 functions, things can be quite tricky, as shown for instance by Theorem 11.8, dealing with the first few derivatives of tan, which do not look very good.

Thus, expect some tricky mathematics to come. Let us start with a reminder of the results for sin, cos, sinh, cosh, written more conveniently, by using a variable x:

THEOREM 11.20. *We have the following formulae for* sin *and* cos*,*

$$\sin x=\sum_{k=0}^{\infty}(-1)^k\frac{x^{2k+1}}{(2k+1)!}\quad,\quad\cos x=\sum_{k=0}^{\infty}(-1)^k\frac{x^{2k}}{(2k)!}$$

and the following formulae for sinh *and* cosh*,*

$$\sinh x=\sum_{k=0}^{\infty}\frac{x^{2k+1}}{(2k+1)!}\quad,\quad\cosh x=\sum_{k=0}^{\infty}\frac{x^{2k}}{(2k)!}$$

as Taylor series at zero, and in general as well, for any $x\in\mathbb{R}$, and even $x\in\mathbb{C}$.

PROOF. This is something that we know from Theorem 11.17, written by using a more familiar variable x. Let us record as well some numerics. For sin, cos we have:

$$\sin x=x-\frac{x^3}{6}+\frac{x^5}{120}-\frac{x^7}{5040}+\ldots\quad,\quad\cos x=1-\frac{x^2}{2}+\frac{x^4}{24}-\frac{x^6}{720}+\ldots$$

As for sinh, cosh, here the Taylor series are identical, save for the signs:

$$\sinh x=x+\frac{x^3}{6}+\frac{x^5}{120}+\frac{x^7}{5040}+\ldots\quad,\quad\cosh x=1+\frac{x^2}{2}+\frac{x^4}{24}+\frac{x^6}{720}+\ldots$$

Finally, let us mention that, shall you ever need such Taylor series expansions at other points, $y\neq 0$, these can be found by using the standard formulae for sums. □

Next on our list, let us talk now about arcsin, arccos, arctan, arccot. These functions can be investigated by using their first derivatives, computed in chapter 9, and then Theorem 11.19 for extracting the square roots, the result being as follows:

THEOREM 11.21. *The Taylor series of* arcsin, arccos *are given by*

$$\arcsin x=\sum_{k=0}^{\infty}\frac{D_k}{4^k(2k+1)}\,x^{2k+1}\quad,\quad\arccos x=\frac{\pi}{2}-\sum_{k=0}^{\infty}\frac{D_k}{4^k(2k+1)}\,x^{2k+1}$$

and the Taylor series of arctan, arccot *are given by*

$$\arctan x=\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\,x^{2k+1}\quad,\quad\operatorname{arccot}x=\frac{\pi}{2}-\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\,x^{2k+1}$$

with $D_k=\binom{2k}{k}$ *being the central binomial coefficients.*

PROOF. This is something routine, by using the formulae of the first derivatives that we computed before, in chapter 9, which were as follows:

$$(\arcsin x)'=\frac{1}{\sqrt{1-x^2}}\quad,\quad(\arccos x)'=-\frac{1}{\sqrt{1-x^2}}$$

$$(\arctan x)'=\frac{1}{1+x^2}\quad,\quad(\operatorname{arccot}x)'=-\frac{1}{1+x^2}$$

(1) Indeed, let us recall from Theorem 11.19 that we can extract the inverse square roots as follows, with $D_k=\binom{2k}{k}$ being the central binomial coefficients:

$$\frac{1}{\sqrt{1+t}}=\sum_{k=0}^{\infty}D_k\left(\frac{-t}{4}\right)^k$$

With the change of variables $t=-x^2$, this formula becomes:

$$\frac{1}{\sqrt{1-x^2}}=\sum_{k=0}^{\infty}D_k\left(\frac{x^2}{4}\right)^k$$

The question is now, what is the function having this as derivative? Since the arcsine must vanish at $x=0$, we are led to the formula in the statement, namely:

$$\arcsin x=\sum_{k=0}^{\infty}\frac{D_k}{4^k(2k+1)}\,x^{2k+1}$$

(2) A similar study applies to the arcosine, and we obtain here, again as claimed:

$$\arccos x=\frac{\pi}{2}-\sum_{k=0}^{\infty}\frac{D_k}{4^k(2k+1)}\,x^{2k+1}$$

Alternatively, we can simply say that this formula follows from the one of arcsin.

(3) Regarding now the arctangent, we can use here the following formula:

$$\frac{1}{1+x^2}=\sum_{k=0}^{\infty}(-1)^kx^{2k}$$

By arguing like before for the arcsine, we obtained here, as claimed:

$$\arctan x=\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\,x^{2k+1}$$

(4) A similar study applies to the arcotangent, and we obtain, as claimed:

$$\text{arccot}\,x=\frac{\pi}{2}-\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\,x^{2k+1}$$

Alternatively, we can simply say that this formula follows from the one of arctan.

(5) Finally, let us record some numerics. For arcsin, arccos we have:

$$\arcsin x=x+\frac{x^3}{6}+\frac{3x^5}{40}+\frac{5x^7}{112}+\ldots\quad,\quad\arccos x=\frac{\pi}{2}-x-\frac{x^3}{6}-\frac{3x^5}{40}-\ldots$$

As for the arctangent and arcotangent, the Taylor series here are as follows:

$$\arctan x=x-\frac{x^3}{3}+\frac{x^5}{5}-\frac{x^7}{7}+\ldots\quad,\quad\text{arccot}\,x=\frac{\pi}{2}-x+\frac{x^3}{3}-\frac{x^5}{5}+\ldots$$

Thus, we are led to the conclusions in the statement. □

Next, we can talk about arcsinh, arccosh, arctanh, arccoth, as follows:

THEOREM 11.22. *The Taylor series of* arcsinh, arccosh *at* $0,\infty$ *are*

$$\text{arcsinh}\,x=\sum_{k=0}^{\infty}(-1)^k\frac{D_k}{4^k(2k+1)}\,x^{2k+1}\quad,\quad\text{arccosh}\,x=\log(2x)-\frac{1}{2}\sum_{k=1}^{\infty}\frac{D_k}{4^kk}\,x^{-2k}$$

and the Taylor series of arctanh, arccoth *at* $0,\infty$ *are*

$$\text{arctanh}\,x=\sum_{k=0}^{\infty}\frac{x^{2k+1}}{2k+1}\quad,\quad\text{arccoth}\,x=\sum_{k=0}^{\infty}\frac{x^{-(2k+1)}}{2k+1}$$

with $D_k=\binom{2k}{k}$ *being the central binomial coefficients.*

PROOF. This is again routine, by using as before the formulae of the first derivatives that we computed before, in chapter 9, which were as follows:

$$(\text{arcsinh}\,x)'=\frac{1}{\sqrt{1+x^2}}\quad,\quad(\text{arccosh}\,x)'=\frac{1}{\sqrt{x^2-1}}$$

$$(\text{arctanh}\,x)'=\frac{1}{1-x^2}\quad,\quad(\text{arccoth}\,x)'=\frac{1}{1-x^2}$$

(1) For arcsinh the derivative is the same as for arcsin, save for a $-$ sign, which at the level of corresponding Taylor series will produce $(-1)^k$ factors, as above.

(2) For arccosh things are more tricky. The function cosh is as follows, increasing:

$$\cosh x = \frac{e^x + e^{-x}}{2} : [0, \infty) \to [1, \infty)$$

We conclude that the inverse function arccosh is as follows, increasing too:

$$\text{arccosh}\, x : [1, \infty) \to [0, \infty)$$

Now let us try to approximate this function at ∞. As a first observation, we have:

$$\cosh x \simeq \frac{e^x}{2} \implies \text{arccosh}\, x \simeq \log(2x)$$

Thus, what we can try to do is to write $\text{arccosh}\, x - \log(2x)$ as a power series in x^{-1}, which amounts in finding the Taylor series at 0 of the following function:

$$f(y) = \text{arccosh}(y^{-1}) - \log(2y^{-1})$$

Now for this latter purpose, observe that, by using the formula of $\text{arccosh}'$, we have:

$$\begin{aligned} f'(y) &= -\frac{y^{-2}}{\sqrt{y^{-2} - 1}} + \frac{1}{y} \\ &= \frac{1}{y}\left(1 - \frac{1}{\sqrt{1 - y^2}}\right) \\ &= \frac{1}{y}\left(1 - \sum_{k=0}^{\infty} \frac{D_k}{4^k}\, y^{2k}\right) \\ &= -\sum_{k=1}^{\infty} \frac{D_k}{4^k}\, y^{2k-1} \end{aligned}$$

But this gives the following formula, which with $y = x^{-1}$ is the one in the statement:

$$f(y) = -\frac{1}{2} \sum_{k=1}^{\infty} \frac{D_k}{4^k k}\, y^{2k}$$

(3) For arctanh the derivative is the same as for arctan, save for a $-$ sign, which at the level of corresponding Taylor series will produce $(-1)^k$ factors, as above.

(4) Regarding now arccoth, we can use here the following computation:

$$y = \text{arctanh}\,(x^{-1}) \implies \tanh y = x^{-1} \implies \coth y = x$$

Indeed, this shows that arccoth and arctanh are related by the following formula:

$$\text{arccoth}\, x = \text{arctanh}\,(x^{-1})$$

But with this, we are led to the formula in the statement, for arccoth at ∞.

(5) Finally, at the level of the numerics, we first have the following formulae:

$$\text{arcsinh}\, x = x - \frac{x^3}{6} + \frac{3x^5}{40} - \frac{5x^7}{112} + \ldots, \ \text{arccosh}\, x = \log(2x) - \frac{1}{4x^2} - \frac{3}{32x^4} - \frac{5}{96x^6} - \ldots$$

As for the arctangent and arcotangent, the Taylor series here are as follows:

$$\text{arctanh}\, x = x + \frac{x^3}{3} + \frac{x^5}{5} + \frac{x^7}{7} + \ldots \quad , \quad \text{arccoth}\, x = \frac{1}{x} + \frac{1}{3x^3} + \frac{1}{5x^5} + \frac{1}{7x^5} + \ldots$$

Thus, we are led to the conclusions in the statement. □

Next, we can talk about arcsec, arccsc, arcsech, arccsch, as follows:

THEOREM 11.23. *The Taylor series of* arcsec, arccsc *at* ∞ *are given by*

$$\text{arcsec}\, x = \frac{\pi}{2} - \sum_{k=0}^{\infty} \frac{D_k}{4^k(2k+1)}\, x^{-(2k+1)} \quad , \quad \text{arccsc}\, x = \sum_{k=0}^{\infty} \frac{D_k}{4^k(2k+1)}\, x^{-(2k+1)}$$

and the Taylor series of arcsech, arccsch *at* $0, \infty$ *are given by*

$$\text{arcsech}\, x = \log\left(\frac{2}{x}\right) - \frac{1}{2}\sum_{k=1}^{\infty} \frac{D_k}{4^k k}\, x^{2k} \quad , \quad \text{arccsch}\, x = \sum_{k=0}^{\infty} (-1)^k \frac{D_k}{4^k(2k+1)}\, x^{-(2k+1)}$$

with $D_k = \binom{2k}{k}$ *being the central binomial coefficients.*

PROOF. This is again routine, by using the formulae of the first derivatives, along with the binomial formula, as we did in the proof of Theorems 11.21 and 11.22:

$$(\text{arcsec}\, x)' = \frac{1}{|x|\sqrt{x^2-1}} \quad , \quad (\text{arccsc}\, x)' = -\frac{1}{|x|\sqrt{x^2-1}}$$

$$(\text{arcsech}\, x)' = -\frac{1}{|x|\sqrt{1+x^2}} \quad , \quad (\text{arccsch}\, x)' = -\frac{1}{|x|\sqrt{1-x^2}}$$

Alternatively, the formulae for arcsec, arccsc follow from those for arccos, arcsin, and those for arcsech, arccsch follow from those for arccosh, arcsinh. Numerically:

$$\text{arcsec}\, x = \frac{\pi}{2} - \frac{1}{x} - \frac{1}{6x^3} - \frac{3}{40x^5} - \ldots \quad , \quad \text{arccsc}\, x = \frac{1}{x} + \frac{1}{6x^3} + \frac{3}{40x^5} + \frac{5}{112x^7} + \ldots$$

As for the hyperbolic arcsecant and arcosecant, the Taylor series here are as follows:

$$\text{arcsech}\, x = \log\left(\frac{2}{x}\right) - \frac{x^2}{4} - \frac{3x^4}{32} - \frac{5x^6}{96} - \ldots, \ \text{arccsch}\, x = \frac{1}{x} - \frac{1}{6x^3} + \frac{3}{40x^5} - \frac{5}{112x^7} + \ldots$$

Thus, we are led to the conclusions in the statement. □

Good work that we did, and in what regards the remaining functions, which are more complicated, we will leave them for later. As a main application now, we have:

THEOREM 11.24 (Leibnitz). *We have the following formula,*

$$\frac{\pi}{4}=1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\frac{1}{9}-\frac{1}{11}+\ldots$$

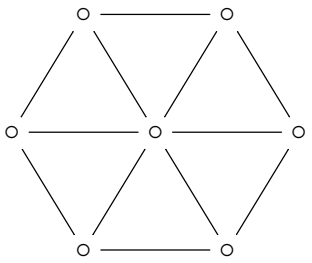
which can be used in order to find the decimals of π.

PROOF. This follows indeed from the Taylor series of arctan, which gives:

$$\frac{\pi}{4}=\arctan(1)=1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\frac{1}{9}-\frac{1}{11}+\ldots$$

However, regarding this, and the approximation of π in general, there is a long story here, involving many mathematicians, and their ideas, worth to be told, as follows:

(1) To start with, forgetting about calculus and everything advanced, that is, fast forward to the ancient times, we know that π appears as the semiperimeter of the circle having radius 1. And we also know, by drawing a hexagon, that $\pi>3$, not by much:



(2) But this suggests approximating π by computing the perimeter of various inscribed and circumscribed regular polygons. In modern terms, what we get from a N-gon is:

$$N\sin\left(\frac{\pi}{N}\right)<\pi<N\tan\left(\frac{\pi}{N}\right)$$

(3) So, let us see how this works. In view of our halving formulae for $\sin,\cos,\tan$, it makes sense to do the computations for $N=2^s$. And for $N=4,8,16$, we get:

$$2\sqrt{2}<\pi<4$$

$$4\sqrt{2-\sqrt{2}}<\pi<8(\sqrt{2}-1)$$

$$8\sqrt{2-\sqrt{2+\sqrt{2}}}<\pi<16\left(\sqrt{4+2\sqrt{2}}-\sqrt{2}-1\right)$$

(4) Numerically, we obtain in this way the following estimates:

$$2.828<\pi<4$$

$$3.061<\pi<3.314$$

$$3.121<\pi<3.183$$

Which does not look great, lots of work for doing all this, especially in extracting the square roots, and this, for not that many decimals, by the end of the day.

(5) This being said, with this being the only method available, let us see what we get at higher $N = 2^s$. And here, forgetting about the tangent, and focusing on the sine, we are led to the following formula for π, coming via the above method, with $k = s - 1$:

$$\pi = \lim_{k\to\infty} 2^k \underbrace{\sqrt{2 - \sqrt{2 + \sqrt{2 + \ldots + \sqrt{2}}}}}_{k \text{ square roots}}$$

(6) So, this was the old method for computing π, but in practice this has not prevented the ancients from doing lots of computations, and approximating π to a fair amount of decimals, a few dozens, which is just perfect for usual engineering purposes:

$$\pi = 3.14159265358979323846\ldots$$

(7) The continuation of the story involves calculus, with the formula in the statement by Leibnitz, and several modifications of this Leibnitz formula. Among others, we have here the following beautiful formula of Euler, that we will discuss later in this book:

$$\frac{\pi^2}{6} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \ldots$$

However, in practice, while these formulae are certainly easier to evaluate than the old ones, their convergence is very slow too, leading to a few hundred decimals for π.

(8) And then, in more modern times, there was the following key formula by Ramanujan, coming from advanced arithmetic, which converges remarkably fast:

$$\frac{1}{\pi} = \frac{2\sqrt{2}}{99^2} \sum_{k=0}^{\infty} \frac{(4k)!}{k!^4} \cdot \frac{26390k + 1103}{396^{4k}}$$

And it is this formula, and its versions, those which are presently used. □

11d. Bernoulli, Euler

We still have some trigonometric functions left, and in order to identify them, the best is to start with a kill list. According to what we have, the situation is as follows:

$\sin^\dagger$	$\cos^\dagger$	$\tan$	$\sec$	$\csc$	$\cot$
$\arcsin^\dagger$	$\arccos^\dagger$	$\arctan^\dagger$	$\text{arcsec}^\dagger$	$\text{arccsc}^\dagger$	$\text{arccot}^\dagger$
$\sinh^\dagger$	$\cosh^\dagger$	$\tanh$	sech	csch	$\coth$
$\text{arcsinh}^\dagger$	$\text{arccosh}^\dagger$	$\text{arctanh}^\dagger$	$\text{arcsech}^\dagger$	$\text{arccsch}^\dagger$	$\text{arccoth}^\dagger$

Thus, we have 16 functions discussed, and 8 functions left. Getting to work now, the idea is that the key to all our computations to follow is the hyperbolic cotangent:

$$\coth x = \frac{e^x + e^{-x}}{e^x - e^{-x}}$$

As a first observation, we can turn coth into something a bit simpler, as follows:

$$\begin{aligned}\coth x=\frac{e^{2x}+1}{e^{2x}-1} &\implies \coth\frac{x}{2}=\frac{e^x+1}{e^x-1}\\ &\implies \coth\frac{x}{2}-1=\frac{2}{e^x-1}\\ &\implies \frac{x}{2}\left(\coth\frac{x}{2}-1\right)=\frac{x}{e^x-1}\end{aligned}$$

Now the point is that, in relation with the function on the right, we have:

THEOREM 11.25. *When defining the Bernoulli numbers B_n according to*

$$\frac{x}{e^x-1}=\sum_{n=0}^{\infty}\frac{B_n}{n!}\,x^n$$

the odd Bernoulli numbers all vanish, except for $B_1=-1/2$, and the even ones are:

$$1\ ,\quad \frac{1}{6}\ ,\quad -\frac{1}{30}\ ,\quad \frac{1}{42}\ ,\quad -\frac{1}{30}\ ,\quad \frac{5}{66}\ ,\quad -\frac{691}{2730}\ ,\quad \frac{7}{6}\ ,\quad -\frac{3617}{510}\ ,\quad \ldots$$

Also, we have the following formula for these Bernoulli numbers,

$$\sum_{k=0}^{m}\binom{m+1}{k}B_k=\delta_{m0}$$

which fully computes them, by recurrence.

PROOF. Many things can be said here, the idea with this being as follows:

(1) Consider the function in the statement, known as hyperbolic cat function:

$$\operatorname{cath} x=\frac{x}{e^x-1}$$

This function expands then into a power series, in the obvious way, as follows:

$$\operatorname{cath} x=\frac{1}{1+\frac{x}{2}+\frac{x^2}{6}+\frac{x^3}{24}+\ldots}=1-\frac{x}{2}+\frac{x^2}{12}-\ldots$$

To be more precise, the first coefficient $-1/2$ is needed for x^2, and the next coefficient, $1/12$, comes according to $1/6-1/4+1/12=0$, which is needed for x^3.

(2) In order to further study the above series, observe that we have:

$$\begin{aligned}\operatorname{cath}(x)-\operatorname{cath}(-x) &= \frac{x}{e^x-1}+\frac{x}{e^{-x}-1}\\ &= \frac{x}{e^x-1}-\frac{xe^x}{e^x-1}\\ &= -x\end{aligned}$$

Thus the odd coefficients all vanish, except for that first odd coefficient, $-1/2$.

(3) In practice now, the coefficients of the hyperbolic cat function can be computed with some patience, using the fraction in (1), and numerically, we obtain:

$$\text{cath}\, x = 1 - \frac{1}{2}\, x + \frac{1}{6} \cdot \frac{x^2}{2!} - \frac{1}{30} \cdot \frac{x^4}{4!} + \frac{1}{42} \cdot \frac{x^6}{6!} - \frac{1}{30} \cdot \frac{x^8}{8!} + \frac{5}{66} \cdot \frac{x^{10}}{10!} - \ldots$$

Thus, with $\text{cath}\, x = \sum_n B_n x^n/n!$, standing as a definition for the Bernoulli numbers B_n, we are led to the numeric values for these numbers B_n in the statement.

(4) At a more advanced level, what we have in the above is a Taylor series, whose coefficients can be obtained by derivating. But this leads, via some computations that we will leave as an exercise, to the following recurrence formula, for these coefficients:

$$\sum_{k=0}^{m} \binom{m+1}{k} B_k = \delta_{m0}$$

(5) So, let us see how this latter formula works. At $m = 0, 1, 2, 3, 4$, we get:

$$B_0 = 1 \implies B_0 = 1$$

$$B_0 + 2B_1 = 0 \implies B_1 = -1/2$$

$$B_0 + 3B_1 + 3B_2 = 0 \implies B_2 = 1/6$$

$$B_0 + 4B_1 + 6B_2 + 4B_3 = 0 \implies B_3 = 0$$

$$B_0 + 5B_1 + 10B_2 + 10B_3 + 5B_4 = 0 \implies B_4 = -1/30$$

And so on, the idea being that the Bernoulli numbers are computable by recurrence, by using the above formula, but unfortunately, there is no explicit formula for them.

(6) So, this was for the story of the hyperbolic cat function, and of the Bernoulli numbers, quickly told. We will be back to all this later, in chapter 14. □

Now with this in hand, we can go back to coth, and to the related functions cot and csc, csch too, and we have the following result, regarding them:

THEOREM 11.26. *The Taylor series of* cot, coth *are given by*

$$\cot x = \sum_{k=0}^{\infty} (-1)^k \frac{4^k B_{2k}}{(2k)!}\, x^{2k-1} \quad , \quad \coth x = \sum_{k=0}^{\infty} \frac{4^k B_{2k}}{(2k)!}\, x^{2k-1}$$

and the Taylor series of csc, csch *are given by*

$$\csc x = \sum_{k=0}^{\infty} (-1)^{k-1} \frac{(4^k - 2) B_{2k}}{(2k)!}\, x^{2k-1} \quad , \quad \text{csch}\, x = -\sum_{k=0}^{\infty} \frac{(4^k - 2) B_{2k}}{(2k)!}\, x^{2k-1}$$

with B_n being the Bernoulli numbers.

PROOF. The formula for coth comes from Theorem 11.25, which gives:

$$\begin{aligned}\frac{x}{2}\left(\coth\frac{x}{2}-1\right) &= \frac{x}{e^x-1}\\ &= \sum_{n=0}^{\infty}\frac{B_n}{n!}\,x^n\\ &= -\frac{x}{2}+\sum_{k=0}^{\infty}\frac{B_{2k}}{(2k)!}\,x^{2k}\end{aligned}$$

Indeed, with $x\to 2x$ we obtain from this the following formula:

$$x(\coth x-1)=-x+\sum_{n=0}^{\infty}\frac{4^kB_{2k}}{(2k)!}\,x^{2k}$$

Thus, we obtain the formula for coth in the statement. As for the other formulae, for cot and csc, csch, these appear as variations of this. At the level of numerics, we have:

$$\cot x=\frac{1}{x}-\frac{x}{3}-\frac{x^3}{45}-\frac{2x^5}{945}-\ldots\quad,\quad \coth x=\frac{1}{x}+\frac{x}{3}-\frac{x^3}{45}+\frac{2x^5}{945}-\ldots$$

As for the Taylor series of the hyperbolic and usual cosecant, these are:

$$\csc x=\frac{1}{x}+\frac{x}{6}+\frac{7x^3}{360}+\frac{31x^5}{15120}+\ldots\quad,\quad \operatorname{csch} x=\frac{1}{x}-\frac{x}{6}+\frac{7x^3}{360}-\frac{31x^5}{15120}+\ldots$$

Thus, we are led to the conclusions in the statement. □

Regarding the functions tan and tanh, we have a similar result here, as follows:

THEOREM 11.27. *The Taylor series of* tan, tanh *are given by*

$$\tan x=\sum_{k=0}^{\infty}(-1)^k\frac{T_{2k+1}}{(2k+1)!}\,x^{2k+1}\quad,\quad \tanh x=\sum_{k=0}^{\infty}\frac{T_{2k+1}}{(2k+1)!}\,x^{2k+1}$$

with T_k being the tangent numbers, given by the following formula,

$$T_{2k+1}=4^{k+1}(4^{k+1}-1)\frac{B_{2k+2}}{2k+2}$$

with B_n being the Bernoulli numbers.

PROOF. These formulae come as variations of the formulae in Theorem 11.26, and we will leave the proof here as an exercise. At the level of numerics, we have:

$$\tan x=x+\frac{x^3}{3}+\frac{2x^5}{15}+\frac{17x^7}{315}+\ldots\quad,\quad \tanh x=x-\frac{x^3}{3}+\frac{2x^5}{15}-\frac{17x^7}{315}+\ldots$$

Thus, we are led to the conclusions in the statement. □

Finally, we can talk as well about sec, sech, the result here being as follows:

THEOREM 11.28. *The Taylor series of* sec, sech *are given by*

$$\sec x = \sum_{k=0}^{\infty}(-1)^k\frac{E_{2k}}{(2k)!}\,x^{2k} \quad , \quad \text{sech}\,x = \sum_{k=0}^{\infty}\frac{E_{2k}}{(2k)!}\,x^{2k}$$

with E_n being Euler numbers, given by the formula

$$\sum_{k=1}^{n}\binom{2n}{2k}E_{2k} = -1$$

which computes them by recurrence.

PROOF. These formulae come as variations of the formulae in Theorems 11.26 and 11.27, and we will leave the proof as an exercise. Let us record as well the first few values of the even Euler numbers, with the odd ones all vanishing, which are as follows:

$$1\ , \quad -1\ , \quad 5\ , \quad -61\ , \quad 1385\ , \quad -50521\ , \quad 2702765\ , \quad \ldots$$

At the level of numerics, the Taylor series in the statement are as follows:

$$\sec x = 1+\frac{x^2}{2}+\frac{5x^4}{24}+\frac{61x^6}{720}+\ldots \quad , \quad \text{sech}\,x = 1-\frac{x^2}{2}+\frac{5x^4}{24}-\frac{61x^6}{720}+\ldots$$

Thus, we are led to the conclusions in the statement. □

11e. Exercises

This was a key calculus chapter, about the key calculus formula, which is the Taylor formula, and its ramifications, and as exercises on all this, we have:

EXERCISE 11.29. *Learn more from physicists about the jerk, and the third derivative.*

EXERCISE 11.30. *Compute the third derivatives of the other basic functions.*

EXERCISE 11.31. *Compute the fourth derivatives of the other basic functions.*

EXERCISE 11.32. *Can we make something out of our $k = 4$ periodicity observation?*

EXERCISE 11.33. *Apply our maximization algorithm to various concrete functions.*

EXERCISE 11.34. *Study how the binomial formula applies, for exponents $p \in \mathbb{Z}/3$.*

EXERCISE 11.35. *Learn more about the Bernoulli numbers, and their properties.*

EXERCISE 11.36. *Learn more about the Euler numbers, and their properties.*

As bonus exercise, with what we learned here, you are good to go for some bonus learning, on special functions. Having a look at that would be a quite good idea.

CHAPTER 12

Differential equations

12a. Differential equations

Eventually. Time to see if our calculus technology is any good, for solving questions appearing from the real life. Let us recall from chapter 10 that we have:

FACT 12.1. *The gravity free fall, wave and heat propagation equations in* $1D$ *are*

$$\ddot{x} = -\frac{k}{x^2} \quad , \quad \ddot{\varphi} = v^2\varphi'' \quad , \quad \dot{\varphi} = \alpha\varphi''$$

with $\varphi = \varphi(x,t)$ *being the height and temperature, and the dots being time derivatives.*

To be more precise, the gravity free fall equation is indeed the one above, where $k = GM$, with M being the attracting mass, and $G \simeq 6.674 \times 10^{-11}$, and with this coming from the Newton formula for the gravitational force. As for the wave and heat propagation equations, where $v > 0$ is the propagation speed, and $\alpha > 0$ is the thermal diffusivity of the medium, these were discussed as well in chapter 10.

So, we can solve right away the above equations? Not really, our comments being:

COMMENTS 12.2. *The above equations are not the easiest ones to start with:*

(1) *They involve second derivatives, instead of the more familiar first derivatives.*
(2) *In addition, the wave and heat equations involve two variables, instead of one.*

In short, modesty, and my feeling is that we should better start with some pure mathematics, solving equations for the sake of solving equations, more and more complicated, and once we get good at this, come back to Fact 12.1, and see what we can do.

Getting started now, with some pure mathematics, here is a first interesting thing that we can say in relation with the differential equations, and with this being a key mathematical fact, already announced in chapter 4, by one of my feline collaborators:

THEOREM 12.3. *The exponential function is the unique solution of*

$$f' = f \quad , \quad f(0) = 1$$

and as a consequence, $e = f(1)$*, with* f *being this unique solution.*

PROOF. Since we have $f(0) = 1$ and $f' = f$ we conclude that we have $f \geq 1$ for $x \geq 0$, and a similar backwards argument shows that we have as well $f > 0$, for $x < 0$. In short, we have $f > 0$ over the whole $\mathbb{R}$, and in particular $f \neq 0$. But with this, we have:

$$\begin{aligned} f' = f &\implies \frac{f'}{f} = 1 \\ &\implies (\log f)' = 1 \\ &\implies \log f = x + c \\ &\implies f = \lambda e^x \end{aligned}$$

Now by using $f(0) = 1$ we conclude that we have $f(x) = e^x$, as desired. □

As a next piece of mathematics, here is a useful generalization of Theorem 12.3:

PROPOSITION 12.4. *The equations $f' = a + bf$ are as follows:*

(1) *The solutions of $f' = f$ are the functions $f(x) = \lambda e^x$.*
(2) *The solutions of $f' = bf$ are the functions $f(x) = \lambda e^{bx}$.*
(3) *The solutions of $f' = a + bf$ with $b \neq 0$ are the functions $f(x) = \lambda e^{bx} - a/b$.*
(4) *And the solutions of $f' = a$ are the functions $f(x) = ax + \mu$.*

PROOF. This comes from Theorem 12.3 and its proof, as follows:

(1) This is something that we already know, from the proof of Theorem 12.3.

(2) This comes from a similar computation, by adding a parameter b, as follows:

$$\begin{aligned} f' = bf &\implies \frac{f'}{f} = b \\ &\implies (\log f)' = b \\ &\implies \log f = bx + c \\ &\implies f = \lambda e^{bx} \end{aligned}$$

(3) Assuming $b \neq 0$ we have indeed the following computation, using (2):

$$\begin{aligned} f' = a + bf &\iff f' = b\left(f + \frac{a}{b}\right) \\ &\iff \left(f + \frac{a}{b}\right)' = b\left(f + \frac{a}{b}\right) \\ &\iff f + \frac{a}{b} = \lambda e^{bx} \\ &\iff f = \lambda e^{bx} - \frac{a}{b} \end{aligned}$$

(4) This is something trivial, coming as a complement to (3). We have indeed:

$$f' = a \iff (f - ax)' = 0 \iff f - ax = \mu \iff f = ax + \mu$$

Thus, we are led to the conclusions in the statement. □

What is next? More mathematics I guess, and looking at what we have, in Proposition 12.4, certainly many interesting things there, but the weak point is the assertion at the end, (4), which, while trivial, suggests looking at the following good question:

QUESTION 12.5. *How to solve the following equations, g being a given function,*

$$f' = g$$

via some kind of "antiderivative" procedure, which must be unique modulo scalars?

To be more precise here, we know from Proposition 12.4 (4) that in the simplest case, where $g(x) = a$, constant function, the solutions are $f(x) = ax + \mu$. Thus, the above question makes sense, and with the assertion at the end coming from:

$$f_1' = f_2' \implies (f_1 - f_2)' = 0 \implies f_1 - f_2 = \text{constant}$$

In answer now, we can easily solve this question for $g \in \mathbb{R}[X]$, as follows:

$$f' = a_n x^n + \ldots + a_1 x + a_0 \iff f = \frac{a_n}{n+1} x^{n+1} + \ldots + \frac{a_1}{2} x^2 + a_0 x + \mu$$

Which makes it quite clear that far more things can be said here, by systematically developing a theory of "antiderivatives", based on what we know about derivatives, from chapter 9. However, we will keep this for later, starting with chapter 13 below, the point being that the antiderivatives, also called integrals, are something extremely vast.

In short, we are learning new things, but the problem remains, namely based on Proposition 12.4, what is next? In answer, and with Proposition 12.4 (3,4) in mind, which do a great job, namely solving $f' = a + bf$ for any $a, b \in \mathbb{R}$, I would formulate:

QUESTION 12.6. *How to solve the equations of the following type,*

$$f'' = a + bf + cf'$$

with $a, b, c \in \mathbb{R}$?

And good question this is, because besides the linear functions and exponentials from Proposition 12.4, we have as well as solutions all sorts of trigonometric beasts, coming from the following double derivation formulae, that we know well from chapter 10:

$$\sin'' = -\sin \quad , \quad \cos'' = -\cos \quad , \quad \sinh'' = \sinh \quad , \quad \cosh'' = \cosh$$

Let us start our study with a basic result, regarding these key examples:

THEOREM 12.7. *The equations $f'' = \pm f$ are as follows:*

(1) *The solutions of $f'' = f$ are the functions $f(x) = \lambda e^x + \mu e^{-x}$.*
(2) *Equivalently, the solutions of $f'' = f$ are $f(x) = \alpha \sinh x + \beta \cosh x$.*
(3) *The solutions of $f'' = -f$ are the functions $f(x) = \alpha \sin x + \beta \cos x$.*
(4) *Equivalently, the solutions of $f'' = -f$ are $f(x) = \lambda e^{ix} + \mu e^{-ix}$, with $\mu = \bar{\lambda}$.*

PROOF. Many things can be said here, the idea being as follows:

(1) To start with, the various functions in the statement verify indeed the equations $f'' = \pm f$, as indicated. As for the uniqueness, since the solution of $f'' = \pm f$ should normally depend on two parameters, say $a = f(0)$ and $b = f'(0)$, and the solutions in the statement do depend on two parameters, as indicated, we intuitively have this too.

(2) This being said, it is instructive to solve our equations for the analytic functions, those appearing as power series, see if we have indeed uniqueness. So, assume that:

$$f = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + \ldots$$

The double derivative of this function is then given by:

$$f'' = 2a_2 + 6a_3x + 12a_4x^2 + 20a_5x^3 + 30a_6x^4 + \ldots$$

(3) Thus, the equation $f'' = f$ is equivalent to the following conditions:

$$a_2 = \frac{a_0}{2} \quad , \quad a_4 = \frac{a_0}{24} \quad , \quad a_6 = \frac{a_0}{720} \quad , \quad \ldots$$

$$a_3 = \frac{a_1}{6} \quad , \quad a_5 = \frac{a_1}{120} \quad , \quad a_7 = \frac{a_1}{5040} \quad , \quad \ldots$$

We are therefore led to the conclusion in the statement, that f must be a linear combination of sinh and cosh. As for the study of $f'' = -f$, this is similar.

(4) In fact, with a bit more work, we can solve if we want $f'' = \pm f$ in a fully rigorous way. Indeed, let us first study the equation $f'' = f$. This equation is equivalent to:

$$(f' + f)' = f' + f$$

Now by using Proposition 12.4 (1), we conclude that we must have:

$$f' + f = \lambda e^x$$

But now, we can use the method in the proof of Proposition 12.4 (3), and we get:

$$\begin{aligned} f' + f = \lambda e^x &\iff f' + f = (\lambda e^x/2)' + \lambda e^x/2 \\ &\iff (f - \lambda e^x/2)' = -(f - \lambda e^x/2) \\ &\iff f - \lambda e^x/2 = \mu e^{-x} \\ &\iff f = \lambda e^x/2 + \mu e^{-x} \end{aligned}$$

As for the study of the equation $f'' = -f$, this is similar, using the same trick.

(5) Finally, there is a discussion to be made, in relation with real vs complex numbers. In (1,2,3) the solutions are as indicated over $\mathbb{R}$, with real parameters, and the solutions over $\mathbb{C}$ are given by the same formulae, this time with complex parameters. As for (4), the solutions there are as indicated over $\mathbb{C}$, with complex parameters $\lambda, \mu \in \mathbb{C}$, and if looking for the solutions which are real, we must impose the condition $\mu = \bar{\lambda}$, as indicated. □

With a bit more work, we can fully solve Question 12.6. Let us start with:

PROPOSITION 12.8. *The equations $f'' = a + bf + cf'$ are as follows,*

(1) *If $b \neq 0$, our equation is equivalent to $g'' = bg + cg'$, with $g = f + a/b$.*
(2) *The solutions of $f'' = a + cf'$ with $c \neq 0$ are $f(x) = \rho e^{cx} - ax/c + \eta$.*
(3) *The solutions of $f'' = a$ are $f(x) = ax^2/2 + \mu x + \nu$.*

and modulo this, it remains to solve the linear case, $f'' = bf + cf'$.

PROOF. As said at the end, our goal here is to get rid of the parameter $a \in \mathbb{R}$, which is something quite natural, mathematically speaking, and is something very reasonable too, physically speaking, because the main order 2 equations coming from physics are linear, $a = 0$, as we will soon discover. In practice, this can be done as follows:

(1) To start with, assuming $b \neq 0$, we can indeed get rid of a, as follows:

$$f'' = a + bf + cf' \iff \left(f + \frac{a}{b}\right)'' = b\left(f + \frac{a}{b}\right) + c\left(f + \frac{a}{b}\right)'$$

(2) In view of this, let us discuss now in detail what happens when $b = 0$. Here the equation is $f'' = a + cf'$, which with $g = f'$ takes the following form:

$$g' = a + cg$$

But, according to Proposition 12.4, the solutions of this latter equation are the functions $g(x) = \lambda e^{cx} - a/c$ when $c \neq 0$, and $g(x) = ax + \mu$ when $c = 0$.

(3) Thus, almost done with the case $b = 0$, we just need to apply the antiderivative procedure from Question 12.5 to the solutions that we found. And here, we get:

$$g(x) = \lambda e^{cx} - \frac{a}{c} \implies f(x) = \frac{\lambda}{c} e^{cx} - \frac{ax}{c} + \eta$$

$$g(x) = ax + \mu \implies f(x) = \frac{ax^2}{2} + \mu x + \nu$$

But this leads to the various conclusions in the statement. □

Getting now to the linear case, the result here is as follows:

THEOREM 12.9. *Given an equation $f'' = bf + cf'$, let r, s be the roots of:*

$$x^2 - cx - b = 0$$

(1) *In the case $r \neq s$, the solutions are $f(x) = \rho e^{rx} + \eta e^{sx}$.*
(2) *In the case $r = s$, the solutions are $f(x) = (\lambda x + \mu)e^{rx}$.*

PROOF. This is a straightforward generalization of Theorem 12.7, with the proof using the same methods as there, the details being as follows:

(1) Our first goal is to put our equation in a simpler form. Observe that we have:

$$(f' - rf)' = f'' - rf' = bf + (c - r)f'$$

Now let us look for numbers r, s such that this equals $s(f' - rf)$. We have:

$$bf + (c - r)f' = s(f' - rf) \iff rs = -b, r + s = c$$

Thus, we certainly have numbers $r, s \in \mathbb{C}$ as desired, appearing as solutions of:

$$x^2 - cx - b = 0$$

(2) As a conclusion to this, with $r, s \in \mathbb{C}$ being as above, our equation reads:

$$(f' - rf)' = s(f' - rf)$$

Now with $g = f' - rf$ this equation reads $g' = sg$, which by Proposition 12.4 has as solutions the functions $g(x) = \lambda e^{sx}$. Thus, we are left with solving:

$$f' = rf + \lambda e^{sx}$$

But for $r \neq s$, the solutions of this latter equation are the following functions:

$$f(x) = \mu e^{rx} + \frac{\lambda}{s - r} e^{sx}$$

As for the case $r = s$, where our equation is $f' = rf + \lambda e^{rx}$, the solutions here are:

$$f(x) = (\lambda x + \mu)e^{rx}$$

(3) Thus, we are led to the conclusions in the statement. There is of course some further discussion to be made here, in relation with real vs complex numbers, the idea being that our result works as stated over $\mathbb{C}$, and that when looking for real solutions, we must impose various conditions on the parameters ρ, η, λ, μ involved. We will leave this as an exercise for now, and come back to it later, when talking physics. □

Along the same lines, at a more advanced level, we can use linear algebra:

THEOREM 12.10. *A differential equation of type $f'' = bf + cf'$ can be written as*

$$\begin{pmatrix} f' \\ f'' \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ b & c \end{pmatrix} \begin{pmatrix} f \\ f' \end{pmatrix}$$

which in terms of $g = \binom{f}{f'}$ and $A = \left(\begin{smallmatrix} 0 & 1 \\ b & c \end{smallmatrix}\right)$ takes the following compact form,

$$g' = Ag$$

and whose solutions are as follows, with v being the initial data vector,

$$g = e^{Ax}v$$

and with e^{Ax} being the exponential of the 2×2 matrix Ax.

PROOF. This is something more advanced, more or less equivalent to Theorem 12.9, but certainly looking more conceptual. However, we will not get in detail into this, the idea being that linear algebra belongs to the "several variables" theory, going beyond what we usually do in this book. So, here are some brief explanations on all this:

(1) To start with, by using the trick in the statement, and with the matrix multiplication being the usual one, "multiply rows by columns", our equation reads:

$$g' = Ag$$

Now the point is that this latter equation reminds the one-variable equation $f' = bf$ from Proposition 12.4, having as solutions the functions $f(x) = \lambda e^{bx}$.

(2) Equivalently, with some changes in the notations, we can say that this reminds the one-variable equation $g' = Ag$ with $A \in \mathbb{R}$, that we know how to solve, having as solutions the functions $g(x) = e^{Ax}v$, with $v \in \mathbb{R}$. But, based on this, we can conjecture that the solutions of our original equation are as follows, as in the statement:

$$g = e^{Ax}v$$

(3) So, can we make some sense of this? To start with, we need to know how to exponentiate the matrices of type $B = Ax$, and in answer, we can declare that:

$$e^B = \sum_{k=0}^{\infty} \frac{B^k}{k!}$$

(4) But, will this work. In order to talk about convergence of the above series, let us endow the space of 2×2 matrices with the following norm:

$$||B|| = \sup_{||x||=1} ||Bx||$$

To be more precise, it is easy to see that this is indeed a norm, which in addition satisfies $||AB|| \leq ||A|| \cdot ||B||$. But with this, we have convergence indeed, coming from:

$$\begin{aligned} ||e^B|| &\leq \sum_{k=0}^{\infty} \frac{||B^k||}{k!} \\ &\leq \sum_{k=0}^{\infty} \frac{||B||^k}{k!} \\ &= e^{||B||} < \infty \end{aligned}$$

(5) Summarizing, our $g = e^{Ax}v$ conjecture above makes sense. Now regarding the proof of this conjecture, in one sense this is clear, coming from the following computation, and I will leave it to you, to check that all the algebra here works just fine:

$$(e^{Ax}v)' = (e^{Ax})'v = (Ae^{ax})v = A(e^{Ax}v)$$

(6) As for the uniqueness of our solutions, this is something a bit more complicated, but we can argue here that since by Theorem 12.9 the solutions of $g' = Ag$ depend on two parameters, these solutions can only be the functions $g = e^{Ax}v$ that we found here, depending on two parameters too, namely the entries of the vector $v \in \mathbb{R}^2$.

(7) Next, for the discussion to be complete, it still makes sense to explicitly compute our solutions $g = e^{Ax}v$, see if we get indeed what we previously found in Theorem 12.9. But this can be done, in two steps. First, we must compute the powers of A:

$$\begin{pmatrix} 0 & 1 \\ b & c \end{pmatrix}^k = ?$$

But this can be done, with some combinatorial pain, and we will leave this as an instructive exercise, and then we can compute e^{Ax}, and the solutions $g = e^{Ax}v$, and these solutions agree of course with what we previously found, in Theorem 12.9.

(8) But then, you might ask, was this excursion into linear algebra, which rather seems to complicate things, worth it? In answer, yes, but provided that we have a more advanced knowledge of linear algebra. Indeed, the main problem that we have, with the linear algebra approach, is that of exponentiating matrices, and at a more advanced level, this problem can be quickly solved by using diagonalization, as follows:

$$\begin{aligned} B = P\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} P^{-1} &\implies B^k = P\begin{pmatrix} \lambda_1^k & 0 \\ 0 & \lambda_2^k \end{pmatrix} P^{-1} \\ &\implies e^B = P\begin{pmatrix} e^{\lambda_1} & 0 \\ 0 & e^{\lambda_2} \end{pmatrix} P^{-1} \end{aligned}$$

(9) So, let us see what these advanced linear algebra methods teach us, in relation with our original problem. The eigenvalues r, s of a 2×2 matrix A are computable by using the well-known formulae $r + s = Tr(A)$ and $rs = \det A$, and in our case, we get:

$$A = \begin{pmatrix} 0 & 1 \\ b & c \end{pmatrix} \implies r + s = c, rs = -b$$

(10) Which is great, because what we have here are the roots of $x^2 - cx - b = 0$, so at least we have now a conceptual explanation for the occurrence of that roots. Good.

(11) As for the continuation of the story, this is a bit more complicated. To start with, in the case $r \neq s$ the matrix is diagonalizable, and we know from the above that the solutions appear as linear combinations of e^{rx}, e^{sx}, as in Theorem 12.9.

(12) As for the remaining case, $r = s$, here is where things get more complicated, because the matrix is no longer diagonalizable. However, by assuming some further linear algebra know-how, namely the Jordan form, we can again do the computations, and we reach to the linear combinations of e^{rx}, xe^{rx}, as in Theorem 12.9. □

Still with me I hope, after all this linear algebra, and I am pretty much sure that you are not impressed, with all these damn things being no match for Theorem 12.9, which was something quite simple. Good point, and in answer, the power of the linear algebra method comes from the fact that this works in higher degree too, as follows:

THEOREM 12.11. *The equation* $f^{(n)} = a_0 f + a_1 f' + \ldots + a_{n-1} f^{(n-1)}$ *reads*

$$\begin{pmatrix} f' \\ f'' \\ \vdots \\ f^{(n)} \end{pmatrix} = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & & 1 \\ a_0 & a_1 & \ldots & a_{n-1} \end{pmatrix} \begin{pmatrix} f \\ f' \\ \vdots \\ f^{(n-1)} \end{pmatrix}$$

which in terms of the matrix A *and vector* g *on the right takes the compact form*

$$g' = Ag$$

and whose solutions are as follows, with v *being the initial data vector,*

$$g = e^{Ax} v$$

and with e^{Ax} *being the exponential of the* $n \times n$ *matrix* Ax*.*

PROOF. This is again something quite self-explanatory, generalizing Theorem 12.10, and we will leave some further learning here as an exercise. Needless to say, there are countless things that can be said here, and the more you know, the better. □

12b. Parabolas, pendulum

We have learned many interesting things about differential equations, but back to physics, the equations in Fact 12.1 still look quite complicated, beyond our methods.

So, let us start with something more modest. In relation with gravity, the simplest question regards a free fall under uniform gravity, and we have here:

PROPOSITION 12.12. *In the context of a* $1D$ *free fall from distance* $x_0 = R >> 0$*, with initial velocity* $v_0 = 0$*, the equation of the trajectory is*

$$x \simeq R - \frac{gt^2}{2}$$

with the constant being $g = GM/R^2$*, called gravity of* M*, at distance* R *from it.*

PROOF. We know that the equation of motion is as follows, with $k = GM$:

$$\ddot{x} = -\frac{k}{x^2} \simeq -\frac{k}{R^2}$$

We conclude that the approximate trajectory is given by the following formula:

$$x \simeq R - \frac{k}{R^2} \cdot \frac{t^2}{2}$$

Thus, we have indeed $x \simeq R - gt^2/2$, with g being the following number:

$$g = \frac{k}{R^2} = \frac{GM}{R^2}$$

We are therefore led to the conclusion in the statement. □

As an illustration for the above result, let us do a numeric terrestrial check, based on it. The gravitational constant, the mass of the Earth, and the average radius of the Earth are as follows, expressed as usual in meters and kilograms:

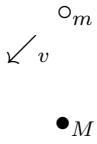
$$G = 6.674 \times 10^{-11} \quad , \quad M = 5.972 \times 10^{24} \quad , \quad R = 6.371 \times 10^6$$

We obtain the following value for the number g computed above:

$$g = \frac{6.674 \times 5.972}{6.371 \times 6.371} \times 10 = 9.819$$

Which is quite decent, when compared to the observed value, $g = 9.806$.

As a generalization of what we have in Proposition 12.12, which is more advanced, lying somewhere between 1D and 2D, let us add an arbitrary initial speed $v_0 = v$ to the above situation, which in addition is allowed to be a vector in $\mathbb{R}^2$, as follows:



We obtain in this way the following generalization of Proposition 12.12:

THEOREM 12.13. *In the context of a free fall from distance $x_0 = R >> 0$, with initial plane velocity vector $v_0 = v$, the equation of the trajectory is*

$$x \simeq R + vt - \frac{gt^2}{2}$$

where $g = GM/R^2$ as usual, and with the quantities R, g in the above being regarded now as vectors, pointing upwards. The approximate trajectory is a parabola.

PROOF. We have several assertions here, the idea being as follows:

(1) Let us first discuss the simpler case where we are still in 1D, as in Proposition 12.12, but with an initial velocity $v_0 = v$ added. In order to find the equation of motion, we can just redo the computations from the proof of Proposition 12.12, with now looking for a general solution of type $x \simeq R + vt + ct^2$, and we get, as stated above:

$$x \simeq R + vt - \frac{gt^2}{2}$$

Alternatively, we can simply argue that, by linearity, what we have to do is to take the solution $x \simeq R - gt^2/2$ found in Proposition 12.12, and add an extra vt term to it.

(2) In the general 2D case now, where the initial velocity $v_0 = v$ is a vector in $\mathbb{R}^2$, the same arguments apply, either by redoing the computations from the proof of Proposition 12.12, or simply by arguing that by linearity we can just take the solution $x \simeq R - gt^2/2$ found there, and add an extra vt term to it. Thus, we have our solution.

(3) Let us study now the solution that we found. In standard (x, y) coordinates, with $v = (p, q)$, and with R, g being now back scalars, our solution looks as follows:

$$x = pt \quad , \quad y \simeq R + qt - \frac{gt^2}{2}$$

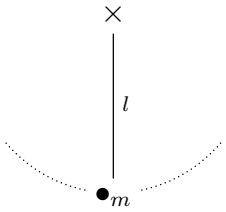
From the first equation we get $t = x/p$, and by substituting into the second:

$$y \simeq R + \frac{qx}{p} - \frac{gx^2}{2p^2}$$

We recognize here the approximate equation of a parabola, and we are done. □

As a continuation of this, still physics of the uniform gravity, and still lying somewhere between 1D and 2D, let us discuss now a fascinating device, the pendulum:

DEFINITION 12.14. *A simple pendulum is a device of type*



consisting of a bob of mass m, attached to a rigid rod of length l.

In order to study the physics of the pendulum, which can easily lead to a lot of complicated computations, when approached with bare hands, the most convenient is to use the notion of energy. For a particle moving under the influence of a force F, the position x, speed v and acceleration a are related by the following formulae:

$$v = \dot{x} \quad , \quad a = \dot{v} = \ddot{x} \quad , \quad F = ma$$

The kinetic energy of our particle is then given by the following formula:

$$T = \frac{mv^2}{2}$$

By differentiating with respect to time t, we obtain the following formula:

$$\dot{T} = mv\dot{v} = mva = Fv$$

But this suggests to define the potential energy V by the following formula, up to a constant, with the derivative being with respect to the space variable x:

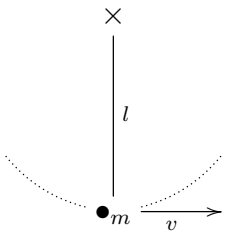
$$V' = -F$$

Indeed, we know from the above that we have $T' = F$, so if we define the total energy to be $E = T + V$, then this total energy is constant, as shown by:

$$E' = T' + V' = 0$$

Very nice all this, and by getting back now to the pendulum from Definition 12.14, we can have this understood with not many computations involved, as follows:

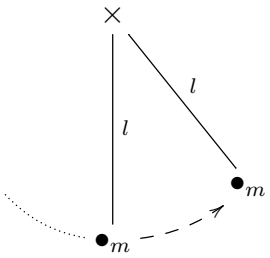
THEOREM 12.15. *For a pendulum starting with speed v from the equilibrium position,*



the motion will be confined if $v^2 < 4gl$, and circular if $v^2 > 4gl$.

PROOF. There are many ways of proving this result, along with working out several other useful related formulae, for which we will refer to the proof below, and with a quite elegant approach to this, using no computations or almost, being as follows:

(1) Let us first examine what happens when the bob has traveled an angular distance $\theta > 0$, with respect to the vertical. The picture here is as follows:



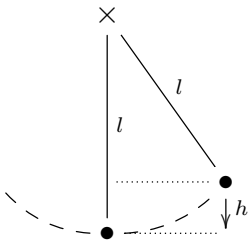
The distance traveled is then $x = l\theta$. As for the force acting, this is $F_{total} = mg$ oriented downwards, with the component alongside x being given by:

$$\begin{aligned} F &= -||F_{total}||\sin\theta \\ &= -mg\sin\theta \\ &= -mg\sin\left(\frac{x}{l}\right) \end{aligned}$$

(2) But with this, we can compute the potential energy. With the convention that this vanishes at the equilibrium position, $V(0) = 0$, we obtain the following formula:

$$\begin{aligned} V' = -F \quad &\implies \quad V' = mg\sin\left(\frac{x}{l}\right) \\ &\implies \quad V = mgl\left(1 - \cos\left(\frac{x}{l}\right)\right) \\ &\implies \quad V = mgl(1 - \cos\theta) \end{aligned}$$

(3) Alternatively, in case this sounds too wizarding, we can compute the potential energy in the old fashion, by letting the bob fall, the picture being as follows:



The height of the fall is then $h = l - l\cos\theta$, and since for this fall the force is constant, $\mathcal{F} = -mg$, we obtain the following formula for the potential energy:

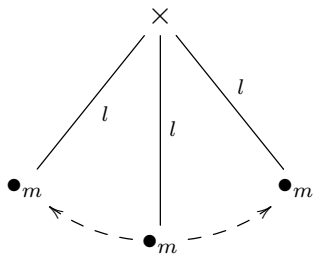
$$\begin{aligned} V' = -\mathcal{F} &\implies V' = mg \\ &\implies V = mgh \\ &\implies V = mgl(1 - \cos\theta) \end{aligned}$$

Summarizing, one way or another we have our formula for the potential energy V.

(4) Now comes the discussion. The motion will be confined when the initial kinetic energy, namely $E = mv^2/2$, satisfies the following condition:

$$\begin{aligned} E < \sup_\theta V = 2mgl &\iff \frac{mv^2}{2} < 2mgl \\ &\iff v^2 < 4gl \end{aligned}$$

In this case, the motion will be confined between two angles $-\theta, \theta$, as follows:



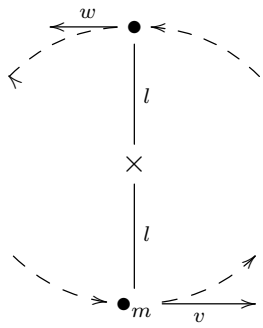
To be more precise here, the two extreme angles $-\theta, \theta \in (-\pi, \pi)$ can be explicitly computed, as being solutions of the following equation:

$$\begin{aligned} V = E &\iff mgl(1 - \cos\theta) = \frac{mv^2}{2} \\ &\iff 1 - \cos\theta = \frac{v^2}{2gl} \end{aligned}$$

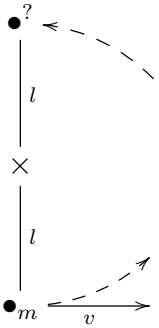
(5) Regarding now the case $v^2 > 4gl$, here the bob will certainly reach the upwards position, with the speed $w > 0$ there being given by the following formula:

$$\begin{aligned}\frac{mw^2}{2} = E - 2mgl \quad &\implies \quad \frac{mw^2}{2} = \frac{mv^2}{2} - 2mgl\\ &\implies \quad w^2 = v^2 - 4gl\\ &\implies \quad w = \sqrt{v^2 - 4gl}\end{aligned}$$

Thus, with the convention in the statement for v, that is, going to the right, the motion of the pendulum will be counterclockwise circular, and perpetual:



(6) Finally, in the case $v^2 = 4gl$, the bob will also reach the upwards position, but with speed $w = 0$ there, and then, at least theoretically, will remain there:



(7) Actually, it is quite interesting in this latter situation, $v^2 = 4gl$, to further speculate on what can happen, when making our problem more realistic. For instance, we can add to our setting the assumption that when the bob is stuck on top, with speed 0, there is a $1/3$ chance for it to keep going, to the left, a $1/3$ chance for it to come back, to the right, and a $1/3$ chance for it to remain stuck. In this case there are infinitely many possible trajectories, which are best investigated by using probability. Welcome to chaos. □

And with this, end of our discussion regarding the pendulum. Never ever shall we be scared about it, with a bit of physics know-how, it all comes down to simple math.

12c. Harmonic oscillators

As a continuation of the above, let us discuss now the motion of a particle near an equilibrium point. We have two examples of such points provided by the pendulum, namely the downwards one, which is stable, and the upwards one, which is unstable.

However, our discussion below will be valid for the movement of any particle, under the influence of any reasonable force. As a first observation, our generalities about motion and energy provide us with some useful information, summarized as follows:

THEOREM 12.16. *For a particle moving near an equilibrium point $x = 0$, the following equivalent conditions must be satisfied, infinitesimally:*

(1) *The potential energy is $V = kx^2/2$, when assuming $V(0) = 0$.*
(2) *The force acting on our particle is $F = -kx$.*
(3) *The equation of motion is $m\ddot{x} + kx = 0$, with m being the mass.*

PROOF. This is something very standard, the idea being as follows:

(1) Let us start with some generalities regarding the potential energy V. Around any given point, that we can choose by translation to be $x = 0$, we can write:

$$V(x) = V(0) + V'(0)x + \frac{V''(0)x^2}{2} + \frac{V'''(0)x^3}{6} + \ldots$$

By definition of V, we can assume $V(0) = 0$. Regarding now the second term, this vanishes too, because our condition of equilibrium reads:

$$V'(0) = -F(0) = 0$$

Thus, with the above normalizations $x = 0$ and $V(0) = 0$ made, our general formula above for V takes at equilibrium the following form, with $k = V''(0)$:

$$V(x) = \frac{kx^2}{2} + \ldots$$

Thus, we are led to the conclusion in the statement, provided that we are indeed in the non-degenerate case, where $k \neq 0$, which amounts in saying that $F'(0) \neq 0$.

(2) This follows indeed from (1), and from $V' = -F$.

(3) This follows indeed from (2), and from $F = ma = m\ddot{x}$. □

The above result suggests formulating the following definition:

DEFINITION 12.17. *A harmonic oscillator is a particle moving as above, following*

$$m\ddot{x} + kx = 0$$

with $k \neq 0$. In the case $k > 0$, we say that we have a simple harmonic oscillator.

Here the last convention comes from the fact that our oscillator is unstable when $k < 0$, and stable $k > 0$, and it is in this latter case that we are mostly interested in. And with this, stability depending on the sign of k, coming either from some abstract reasoning along the lines of Theorem 12.16, or from the explicit formulae below.

Very nice, so let us solve now the equation of motion. We have here:

THEOREM 12.18. *The solutions of the equation of motion $m\ddot{x} + kx = 0$ for the harmonic oscillators are as follows:*

(1) $x = ae^{pt} + be^{-pt}$ *with* $p = \sqrt{-k/m}$, *when* $k < 0$.
(2) $x = c\cos wt + d\sin wt$ *with* $w = \sqrt{k/m}$, *when* $k > 0$.

PROOF. This is standard mathematics, that we already know, as follows:

(1) Assume first that we are in the case $k < 0$. Here, with $p = \sqrt{-k/m}$ as in the statement, the equation of motion takes the following form:

$$\ddot{x} = p^2 x$$

But the functions e^{pt}, e^{-pt} being solutions of this equation, by linearity we obtain that the solutions are exactly the linear combinations of these two functions, as claimed.

(2) Assume now that we are in the case $k > 0$. Here, with $w = \sqrt{k/m}$ as in the statement, the equation of motion takes the following form:

$$\ddot{x} = -w^2 x$$

But the functions $\cos wt, \sin wt$ being solutions, by linearity we obtain that the solutions are exactly the linear combinations of these two functions, as claimed. □

Observe that, as already mentioned above, the formulae that we obtained make it clear that our oscillator is unstable when $k < 0$, and stable when $k > 0$. In fact, we have the following simple consequences of the general formulae obtained above:

PROPOSITION 12.19. *The short and long time behavior of a harmonic oscillator, moving according to $m\ddot{x} + kx = 0$, are as follows:*

(1) *In the case $k < 0$, with $x = ae^{pt} + be^{-pt}$ as above, we have $x \simeq (a+b) + p(a-b)t$ for $t > 0$ small, and $x \simeq ae^{pt}$ for $t >> 0$.*
(2) *In the case $k > 0$, with $x = c\cos wt + d\sin wt$ as above, we have $x \simeq c + dwt$ for $t > 0$ small, and there is no asymptotics for $t >> 0$.*

PROOF. This is indeed standard mathematics based on Theorem 12.18, as follows:

(1) In the case $k<0$, with $x=ae^{pt}+be^{-pt}$ as in Theorem 12.18, in the $t>0$ small regime we have indeed the following estimate, coming from $e^z\simeq 1+z$:

$$\begin{aligned}x&=ae^{pt}+be^{-pt}\\&\simeq a(1+pt)+b(1-pt)\\&=(a+b)+p(a-b)t\end{aligned}$$

As for the other estimate, namely $x\simeq ae^{pt}$ for $t>>0$, this is clear.

(2) In the case $k>0$, with $x=c\cos wt+d\sin wt$ as in Theorem 12.18, in the $t>0$ small regime we have indeed the following estimate, coming from standard calculus:

$$\begin{aligned}x&=c\cos wt+d\sin wt\\&\simeq c(1+o(t))+dwt\\&\simeq c+dwt\end{aligned}$$

As for the last assertion, regarding the lack of asymptotics at $k>0$ in the $t>>0$ regime, this is clear, because neither cos, nor sin have such asymptotics, and the same happens for any linear combination of them, with non-trivial coefficients. □

As more physics, let us study now the damped oscillators, obtained by adding to the picture friction. In terms of the equation of motion, the result is as follows:

THEOREM 12.20. *For a damped oscillator, which is subject by definition to a force of type $F=-kx-\lambda\dot{x}$, the equation of motion is*

$$m\ddot{x}+\lambda\dot{x}+kx=0$$

with m being as before the mass.

PROOF. This is clear indeed from $F=ma=m\ddot{x}$, which gives:

$$\begin{aligned}F=-kx-\lambda\dot{x}&\iff m\ddot{x}=-kx-\lambda\dot{x}\\&\iff m\ddot{x}+\lambda\dot{x}+kx=0\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Now let us try to solve the equation of motion. When looking for solutions of type $x=e^{pt}$, with $p\in\mathbb{C}$ constant, the equation of motion takes the following form:

$$mp^2+\lambda p+k=0$$

But this is a degree 2 equation, that we can solve right away, and we get:

$$p=\frac{-\lambda\pm\sqrt{\lambda^2-4mk}}{2m}$$

It is convenient to write these solutions that we found, and the overall final result about damping, in the following more convenient way:

THEOREM 12.21. *The generic solutions of $m\ddot{x} + \lambda\dot{x} + kx = 0$ are the real linear combinations of the functions e^{pt}, with the parameter $p \in \mathbb{C}$ being given by*

$$p = -\gamma \pm \sqrt{\gamma^2 - w^2} \quad , \quad \gamma = \frac{\lambda}{2m} \quad , \quad w = \sqrt{\frac{k}{m}}$$

with on the right w being the frequency of the usual, undamped oscillator.

PROOF. This follows indeed from the formula of p found above, by dividing everything by $2m$. Observe that w is indeed the frequency of the usual, undamped oscillator. □

Now assume that we are in the case $\lambda > 0$, which is the most usual one, in practice, meaning that our oscillator loses energy. We have then three cases, as follows:

PROPOSITION 12.22. *The oscillator damping with $\lambda > 0$, with this assumption meaning that our oscillator loses energy over the time, can be of three types:*

(1) *Large damping, with $\lambda > 0$ being such that $\gamma > w$. Here the roots found above $p = -\gamma \pm \sqrt{\gamma^2 - w^2}$ are both real, negative, and distinct.*
(2) *Small damping, with $\lambda > 0$ being such that $\gamma < w$. In this case the roots that we found $p = -\gamma \pm \sqrt{\gamma^2 - w^2}$ are complex and conjugate.*
(3) *Critical damping, with $\lambda > 0$ being such that $\gamma = w$. Here we have a double root, which is real and negative, namely $p = -\gamma$.*

PROOF. All this is clear indeed from the formula found in Theorem 12.21. □

Let us study now more in detail the above three types of damping. In what regards the large damping, things here are very simple and intuitive, as follows:

PROPOSITION 12.23. *For large oscillator damping, the trajectory is given by*

$$x = ae^{-\gamma_+ t} + be^{-\gamma_- t}$$

with the parameters $\gamma_+ > \gamma_- > 0$ being given by the formulae

$$\gamma_\pm = \gamma \pm \sqrt{\gamma^2 - w^2}$$

and with $a, b \in \mathbb{R}$. With $t >> 0$ we have $x \simeq be^{-\gamma_- t} \to 0$.

PROOF. All this is indeed self-explanatory, and clear from the formula that we found in Theorem 12.21, under the large damping assumption from Proposition 12.22 (1). □

In what regards the small damping, things here are again quite simple, as follows:

PROPOSITION 12.24. *For small oscillator damping, the trajectory is given by*

$$x = 2re^{-\gamma t}\cos(\rho t - \theta)$$

with $\gamma = \lambda/2m$ as before, with the parameter $\rho > 0$ being given by the formula

$$\rho = \sqrt{w^2 - \gamma^2}$$

and with $r > 0$ and $\theta \in \mathbb{R}$. With $t >> 0$ we have $x \to 0$, exponentially.

PROOF. Assume indeed that we are in the small damping regime, where $\lambda > 0$ is such that $\gamma < w$. The roots that we found are then complex conjugate, as follows:

$$p = -\gamma \pm i\rho \quad , \quad \rho = \sqrt{w^2 - \gamma^2}$$

As for the solution itself, this is given by the following formula, with $c, d \in \mathbb{C}$:

$$\begin{aligned} x &= ce^{p_+ t} + de^{p_- t} \\ &= ce^{-\gamma t + i\rho t} + de^{-\gamma t - i\rho t} \\ &= e^{-\gamma t}(ce^{i\rho t} + de^{-i\rho t}) \end{aligned}$$

Now in order to have $x \in \mathbb{R}$, which is the same as saying $x = \bar{x}$, we must have $c = \bar{d}$. Thus we can write $c = re^{-i\theta}$ and $d = re^{i\theta}$ with $r > 0$ and $\theta \in \mathbb{R}$, and we obtain:

$$\begin{aligned} x &= e^{-\gamma t}(ce^{i\rho t} + \bar{c}e^{-i\rho t}) \\ &= 2e^{-\gamma t} Re(ce^{i\rho t}) \\ &= 2re^{-\gamma t} Re(e^{i(\rho t - \theta)}) \\ &= 2re^{-\gamma t} \cos(\rho t - \theta) \end{aligned}$$

Finally, the fact that we have indeed $x \to 0$, exponentially, is clear. □

As for the critical damping case, the result here is as follows:

THEOREM 12.25. *For critical oscillator damping, the trajectory is given by*

$$x = (a + bt)e^{-\gamma t}$$

with the parameter $\gamma > 0$ being given as usual by the formula

$$\gamma = \frac{\lambda}{2m}$$

and with $a, b \in \mathbb{R}$. With $t >> 0$ we have $x \simeq bte^{-\gamma t} \to 0$.

PROOF. Assume indeed that we are in the critical damping regime, where $\lambda > 0$ is such that $\gamma = w$. In this case the roots that we found in Theorem 12.21 are both given by $p = -\gamma$, and so that result provides us with solutions as follows, with $a \in \mathbb{R}$:

$$x = ae^{-\gamma t}$$

Thus, we must find in this case some further solutions of the equation $m\ddot{x} + \lambda\dot{x} + kx = 0$, and our claim is that the following functions are solutions too:

$$x = bte^{-\gamma t}$$

Indeed, let us verify this, under the above critical damping assumption. By linearity it is enough to do this for $b = 1$, and here the derivatives are computed as follows:

$$\begin{aligned} x = te^{-\gamma t} \quad &\implies \quad \dot{x} = e^{-\gamma t} - \gamma te^{-\gamma t} = (1 - \gamma t)e^{-\gamma t} \\ &\implies \quad \ddot{x} = -\gamma e^{-\gamma t} - \gamma(1 - \gamma t)e^{-\gamma t} = -\gamma(2 - \gamma t)e^{-\gamma t} \end{aligned}$$

We can verify now that the equation is indeed satisfied, as follows:

$$\begin{aligned} m\ddot{x} + \lambda\dot{x} + kx &= -m\gamma(2-\gamma t)e^{-\gamma t} + \lambda(1-\gamma t)e^{-\gamma t} + kte^{-\gamma t} \\ &= (-2m\gamma + m\gamma^2 t + \lambda - \lambda\gamma t + kt)e^{-\gamma t} \\ &= (m\gamma^2 t - \lambda\gamma t + kt)e^{-\gamma t} \\ &= (m\gamma^2 - \lambda\gamma + k)te^{-\gamma t} \\ &= 0 \end{aligned}$$

Here we have used at the end the fact, that we know from Theorem 12.21 and its proof, that the solutions of the equation $mp^2 + \lambda p + k = 0$ are given by $p = -\gamma \pm \sqrt{\gamma^2 - w^2}$. Indeed, in the present critical damping regime these solutions are both given by $p = -\gamma$, and so substituting this particular value in the equation gives zero, as needed:

$$m\gamma^2 - \lambda\gamma + k = 0$$

Thus, we are led to the conclusions in the statement. □

As a conclusion to all this, the precise mathematics of the oscillator damping can be explicitly worked out, in each of the three cases that can appear, large, small or critical, and of particular interest is the mathematics and physics of the critical damping.

12d. Falls, waves, heat

With the above discussed, basically coming from uniform gravity, time now to get back to the equations in Fact 12.1. We first have, regarding free falls:

THEOREM 12.26. *The equation of a gravitational free fall, in* 1 *dimension,*

$$\ddot{x} = -\frac{k}{x^2}$$

can be successfully studied, by computing $t = t(x)$ *instead of* $x = x(t)$*, and we have*

$$t = \sqrt{\frac{x_0^3}{2k}}\left(\sqrt{\frac{x}{x_0}\left(1 - \frac{x}{x_0}\right)} + \arccos\sqrt{\frac{x}{x_0}}\right)$$

with x_0 *being the initial position, at launching.*

PROOF. Many things can be said here, the idea being as follows:

(1) To start with, the equation in the statement, $\ddot{x} = -k/x^2$, is not really solvable. As a first idea here, we can look for a solution as follows, with $T > 0$ and $\lambda > 0$:

$$x = (T - \lambda t)^p$$

In order for our equation to be satisfied, the following quantities must be equal:

$$\ddot{x} = p(p-1)\lambda^2(T-\lambda t)^{p-2} \quad , \quad -\frac{k}{x^2} = -k(T-\lambda t)^{-2p}$$

At the level of the exponent, we must have $p-2=-2p$, and so $p=2/3$. As for the coefficient, the equation here is $(-2/9)\lambda^2=-k$, so $\lambda^2=9k/2$, and $\lambda=3\sqrt{k/2}$. Summarizing, we have our particular solution, which is as follows, with $T>0$:

$$x=\left(T-3\sqrt{\frac{k}{2}}\,t\right)^{2/3}$$

However, this is not the correct solution for our problem, and any attempt of further perturbing this solution, as to get the correct solution, fails. In fact, there is no explicit formula for the solution of $\ddot{x}=-k/x^2$, and we will have to live with this.

(2) In order to say however something on the subject, we can trick as in the statement, by computing $t=t(x)$ instead of $x=x(t)$. Now in order to do the inversion, we will need the following standard computation, coming from the chain rule, applied twice:

$$\begin{aligned}
f(g(x))=x &\implies f'(g(x))g'(x)=1\\
&\implies f'(g(x))=\frac{1}{g'(x)}\\
&\implies f''(g(x))g'(x)=-\frac{g''(x)}{g'(x)^2}\\
&\implies f''(g(x))=-\frac{g''(x)}{g'(x)^3}
\end{aligned}$$

(3) So, consider our equation, written as follows, with $f(t)$ being the position:

$$f''(t)=-\frac{k}{f(t)^2}$$

When setting $t=g(x)$, with $f(g(x))=x$ as above, our equation becomes:

$$\begin{aligned}
-\frac{g''(x)}{g'(x)^3}=-\frac{k}{x^2} &\implies \left(\frac{1}{g'(x)^2}\right)'=\left(\frac{2k}{x}\right)'\\
&\implies \frac{1}{g'(x)^2}=\frac{2k}{x}+c\\
&\implies g'(x)=-\frac{1}{\sqrt{2k/x+c}}
\end{aligned}$$

Here we have chosen the above $-$ sign at the end, when extracting the root, because the function $g=g(x)$, expressing time in terms of position, must be decreasing.

(4) Next, at the initial position x_0 the time must vanish, $g(x_0)=0$, and we can expect its derivative to explode there, $g'(x_0)=-\infty$. Thus the constant c appearing in the above

must be $c = -2k/x_0$, and our equation takes the following form:

$$g'(x) = -\frac{1}{\sqrt{2k}} \cdot \frac{1}{\sqrt{1/x - 1/x_0}}$$

(5) Next, with the change of variables $x = x_0 y$, this equation becomes:

$$g'(x_0 y) = -\frac{x_0}{\sqrt{2k}} \cdot \frac{1}{\sqrt{1/(x_0 y) - 1/x_0}} = -\sqrt{\frac{x_0^3}{2k}} \cdot \frac{1}{\sqrt{1/y - 1}}$$

(6) Now in order to solve this latter equation, observe that we have:

$$\begin{aligned}\left(\sqrt{y - y^2} + \arccos\sqrt{y}\right)' &= \frac{1 - 2y}{2\sqrt{y - y^2}} - \frac{1}{2\sqrt{y}} \cdot \frac{1}{\sqrt{1 - y}} \\ &= -\frac{y}{\sqrt{y - y^2}} \\ &= -\frac{1}{\sqrt{1/y - 1}}\end{aligned}$$

Of course, this might sound a bit rude, but with a bit of patience, and some suitable changes of variables, you can certainly come upon all this by yourself, I mean without knowing the answer in advance. And with this being a good exercise for you.

(7) In any case, problem solved, one way or another, and as a conclusion, the solution of our initial equation from (3), with initial data $g(x_0) = 0$, is as follows:

$$g(x_0 y) = \sqrt{\frac{x_0^3}{2k}} \left(\sqrt{y - y^2} + \arccos\sqrt{y}\right)$$

Now with $x = x_0 y$ we are led to the formula in the statement, namely:

$$g(x) = \sqrt{\frac{x_0^3}{2k}} \left(\sqrt{\frac{x}{x_0}\left(1 - \frac{x}{x_0}\right)} + \arccos\sqrt{\frac{x}{x_0}}\right)$$

(8) What is next? Many possible things, based on this, and as a first application, we can compute the stopping time as function of the initial position x_0, as follows:

$$t_{final} = g(0) = \sqrt{\frac{x_0^3}{2k}} \cdot \arccos(0) = \pi\sqrt{\frac{x_0^3}{8k}}$$

And we will leave some further exploration of this as an exercise. Among others, with the above formula of $g = f^{-1}$ in hand, you can do some numerics for the power series expansion of f, with the conclusion that this leads nowhere, as stated in (1). □

Good to have this basic gravity question solved. As more physics, again with the equations from Fact 12.1 in mind, we can talk as well about waves, as follows:

THEOREM 12.27. *The 1D wave equation with propagation speed $v > 0$, namely*

$$\ddot{\varphi} = v^2\varphi''$$

has as basic solutions the following functions,

$$\varphi(x,t) = A\cos(kx - wt + \delta)$$

with A being called amplitude, $kx-wt+\delta$ being called the phase, k being the wave number, w being the angular frequency, and δ being the phase constant. We have

$$\lambda = \frac{2\pi}{k} \quad , \quad T = \frac{2\pi}{kv} \quad , \quad \nu = \frac{1}{T} \quad , \quad w = 2\pi\nu$$

relating the wavelength λ, period T, frequency ν, and angular frequency w. Moreover, any solution of the wave equation appears as a linear combination of such basic solutions.

PROOF. There are several things going on here, the idea being as follows:

(1) Our first claim is that the function φ in the statement satisfies indeed the wave equation, with speed $v = w/k$. For this purpose, observe that we have:

$$\ddot{\varphi} = -w^2\varphi \quad , \quad \varphi'' = -k^2\varphi$$

Thus, the wave equation is indeed satisfied, with speed $v = w/k$:

$$\ddot{\varphi} = \left(\frac{w}{k}\right)^2\varphi'' = v^2\varphi''$$

(2) Regarding now the other things in the statement, all this is basically terminology, which is very natural, when thinking how $\varphi(x,t) = A\cos(kx - wt + \delta)$ propagates.

(3) Finally, the last assertion is something quite standard, and we will be back to this later, when discussing Fourier analysis, which is the key to such results. □

As a first observation, the above result invites the use of complex numbers. Indeed, we can write the solutions that we found in a more convenient way, as follows:

$$\varphi(x,t) = Re\left[A\,e^{i(kx-wt+\delta)}\right]$$

And we can in fact do even better, by absorbing the quantity $e^{i\delta}$ into the amplitude A, which becomes now a complex number, and writing our formula as:

$$\varphi = Re(\widetilde{\varphi}) \quad , \quad \widetilde{\varphi} = \widetilde{A}e^{i(kx-wt)}$$

Many other things can be said here, and we will be back to waves later in this book, with some further details, directly in the higher dimensional setting.

Along the same lines, we can talk as well about heat in 1D, as follows:

THEOREM 12.28. *The 1D heat equation with thermal diffusivity $\alpha > 0$, namely*

$$\dot{\varphi} = \alpha\varphi''$$

has as basic solution the following function:

$$\varphi(x,t) = \frac{1}{\sqrt{t}}\, e^{-x^2/4\alpha t}$$

Moreover, any solution appears from this, via a standard convolution procedure.

PROOF. As before with waves, this is a mixture of trivial and non-trivial facts:

(1) The time derivative of the function in the statement is given by:

$$\dot{\varphi} = -\frac{1}{2t\sqrt{t}}\, e^{-x^2/4\alpha t} + \frac{x^2}{4\alpha t^2\sqrt{t}}\, e^{-x^2/4\alpha t}$$

Regarding the first space derivative, this is given by the following formula:

$$\varphi' = -\frac{x}{2\alpha t\sqrt{t}}\, e^{-x^2/4\alpha t}$$

As for the second space derivative, this is given by the following formula:

$$\varphi'' = -\frac{1}{2\alpha t\sqrt{t}}\, e^{-x^2/4\alpha t} + \frac{x^2}{4\alpha^2 t^2\sqrt{t}}\, e^{-x^2/4\alpha t}$$

We conclude that the heat equation $\dot{\varphi} = \alpha\varphi''$ is indeed satisfied, as claimed.

(2) As for the second assertion, this is something more advanced, the idea being that the arbitrary solutions can be obtained by convolving the solution in the statement, called heat kernel, with the initial data. And more on this, later in this book. □

12e. Exercises

This was a standard physics chapter, and as exercises on this, we have:

EXERCISE 12.29. *Find out why our figure for g is not exactly the observed value.*

EXERCISE 12.30. *What happens if the bob is attached to a string, instead of a rod?*

EXERCISE 12.31. *Meditate on what happens when the bob is motionless, on top.*

EXERCISE 12.32. *Learn more about potential energy, and about energy in general.*

EXERCISE 12.33. *Learn more about harmonic oscillators, and about damping too.*

EXERCISE 12.34. *Further study the equation of the free fall, in 1 dimension.*

EXERCISE 12.35. *Learn more about the various types of waves, in the real life.*

EXERCISE 12.36. *Learn more about the heat equation, and its finer versions too.*

As bonus exercise, read more basic physics, say from Feynman [31], [32], [33].

Part IV

Integrals

You'll die as you lived, in a flash of the blade
In a corner forgotten by no one
You lived for the touch, for the feel of the steel
One man and his honor

CHAPTER 13

Integration theory

13a. Integration theory

We have seen so far the foundations of calculus, with lots of interesting results regarding the functions $f : \mathbb{R} \to \mathbb{R}$, and their derivatives $f' : \mathbb{R} \to \mathbb{R}$. The general idea was that in order to understand f, we first need to compute its derivative f'. The overall conclusion, coming from the Taylor formula, was that if we are able to compute f', but then also f'', and f''' and so on, we will have a good understanding of f itself.

However, the story is not over here, and there is one more twist to the plot. Which will be a major twist, of similar magnitude to that of the Taylor formula. For reasons which are quite tricky, that will become clear later on, we will be interested in the integration of the functions $f : \mathbb{R} \to \mathbb{R}$. With the claim that this is related to calculus.

There are several possible viewpoints on the integral, which are all useful, and good to know. To start with, we have something very simple, as follows:

DEFINITION 13.1. *The integral of a continuous function $f : [a, b] \to \mathbb{R}$, denoted*

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

is the area below the graph of f, signed $+$ where $f \geq 0$, and signed $-$ where $f \leq 0$.

Here it is of course understood that the area in question can be computed, and with this being something quite subtle, that we will get into later. For the moment, let us just trust our intuition, our function f being continuous, the area in question can "obviously" be computed. More on this later, but for being rigorous, however, let us formulate:

METHOD 13.2. *In practice, the integral of $f \geq 0$ can be computed as follows,*

(1) *Cut the graph of f from 3mm plywood,*
(2) *Plunge that graph into a square container of water,*
(3) *Measure the water displacement, as to have the volume of the graph,*
(4) *Divide by 3×10^{-3} that volume, as to have the area,*

and for general f, we can use this plus $f = f_+ - f_-$, with $f_+, f_- \geq 0$.

So far, so good, we have a rigorous definition, so let us do now some computations. In order to compute areas, and so integrals of functions, without wasting precious water, we can use our geometric knowledge. Here are some basic results of this type:

PROPOSITION 13.3. *We have the following results:*

(1) *When f is linear, we have the following formula:*

$$\int_a^b f(x)dx = (b-a)\cdot\frac{f(a)+f(b)}{2}$$

(2) *In fact, when f is piecewise linear on $[a=a_1,a_2,\ldots,a_n=b]$, we have:*

$$\int_a^b f(x)dx = \sum_{i=1}^{n-1}(a_{i+1}-a_i)\cdot\frac{f(a_i)+f(a_{i+1})}{2}$$

(3) *We have as well the formula $\int_{-1}^1\sqrt{1-x^2}\,dx=\pi/2$.*

PROOF. These results all follow from basic geometry, as follows:

(1) Assuming $f\geq 0$, we must compute the area of a trapezoid having sides $f(a),f(b)$, and height $b-a$. But this is the same as the area of a rectangle having side $(f(a)+f(b))/2$ and height $b-a$, and we obtain $(b-a)(f(a)+f(b))/2$, as claimed.

(2) This is clear indeed from the formula found in (1), by additivity.

(3) The integral in the statement is by definition the area of the upper unit half-disc. But since the area of the whole unit disc is π, this half-disc area is $\pi/2$. □

As an interesting observation, (2) in the above result makes it quite clear that f does not necessarily need to be continuous, in order to talk about its integral. Indeed, assuming that f is piecewise linear on $[a=a_1,a_2,\ldots,a_n=b]$, but not necessarily continuous, we can still talk about its integral, in the obvious way, exactly as in Definition 13.1, and we have an explicit formula for this integral, generalizing the one found in (2), namely:

$$\int_a^b f(x)dx = \sum_{i=1}^{n-1}(a_{i+1}-a_i)\cdot\frac{f(a_i^+)+f(a_{i+1}^-)}{2}$$

Based on this observation, let us upgrade our formalism, as follows:

DEFINITION 13.4. *We say that a function $f:[a,b]\to\mathbb{R}$ is integrable when the area below its graph is computable. In this case we denote by*

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

this area, signed $+$ where $f\geq 0$, and signed $-$ where $f\leq 0$.

As basic examples of integrable functions, we have the continuous ones, provided indeed that our intuition, or that Method 13.2, works indeed for any such function. We will soon see that this is indeed true, coming with mathematical proof. As further examples, we have the functions which are piecewise linear, or piecewise continuous. We will also see, later, as another class of examples, that the piecewise monotone functions are integrable. But more on this later, let us not bother for the moment with all this.

This being said, one more thing regarding theory, that you surely have in mind: is any function integrable? Not clear. I would say that if the Devil comes with some sort of nasty, totally discontinuous function $f : \mathbb{R} \to \mathbb{R}$, then you will have big troubles in cutting its graph from 3mm plywood, as required by Method 13.2. More on this later.

Back to work now, here are some general results regarding the integrals:

PROPOSITION 13.5. *We have the following formulae,*

$$\int_a^b f(x) + g(x)dx = \int_a^b f(x)dx + \int_a^b g(x)dx$$

$$\int_a^b \lambda f(x) = \lambda \int_a^b f(x)$$

valid for any functions f, g and any scalar $\lambda \in \mathbb{R}$.

PROOF. Both these formulae are indeed clear from definitions. □

Moving ahead now, passed the above results, which are of purely algebraic and geometric nature, and perhaps a few more of the same type, which are all quite trivial and that we we will not get into here, we must do some analysis, in order to compute integrals. This is something quite tricky, and we have here the following result:

THEOREM 13.6. *We have the Riemann integration formula,*

$$\int_a^b f(x)dx = (b-a) \times \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f\left(a + \frac{b-a}{N} \cdot k\right)$$

which can serve as a definition for the integral.

PROOF. This is standard, by drawing rectangles. We have indeed the following formula, which can stand as a definition for the signed area below the graph of f:

$$\int_a^b f(x)dx = \lim_{N\to\infty} \sum_{k=1}^{N} \frac{b-a}{N} \cdot f\left(a + \frac{b-a}{N} \cdot k\right)$$

Thus, we are led to the formula in the statement. □

Observe that the above formula suggests that $\int_a^b f(x)dx$ is the length of the interval $[a,b]$, namely $b-a$, times the average of f on the interval $[a,b]$. Thinking a bit, this is indeed something true, with no need for Riemann sums, coming directly from Definition 13.1, because area means side times average height. Thus, we can formulate:

THEOREM 13.7. *The integral of a function $f:[a,b]\to\mathbb{R}$ is given by*

$$\int_a^b f(x)dx=(b-a)\times A(f)$$

where $A(f)$ is the average of f over the interval $[a,b]$.

PROOF. As explained above, this is clear from Definition 13.1, via some geometric thinking. Alternatively, this is something which certainly comes from Theorem 13.6. □

The point of view in Theorem 13.7 is something quite useful, and as an illustration for this, let us review the results that we already have, by using this interpretation. First, we have the formula for linear functions from Proposition 13.3, namely:

$$\int_a^b f(x)dx=(b-a)\cdot\frac{f(a)+f(b)}{2}$$

But this formula is totally obvious with our new viewpoint, from Theorem 13.7. The same goes for the results in Proposition 13.5, which become even more obvious with the viewpoint from Theorem 13.7. Thus, what we have in Theorem 13.7 is definitely useful.

However, not everything trivializes in this way, and the result which is left, from what we have so far, namely the formula $\int_{-1}^1\sqrt{1-x^2}\,dx=\pi/2$ from Proposition 13.3 (3), not only does not trivialize, but becomes quite opaque with our new philosophy.

In short, modesty. Integration is a quite delicate business, and we have several equivalent points of view on what an integral means, and all these points of view are useful, and must be learned, with none of them being clearly better than the others.

Going ahead with more interpretations of the integral, we have:

THEOREM 13.8. *We have the Monte Carlo integration formula,*

$$\int_a^b f(x)dx=(b-a)\times\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^N f(x_i)$$

with $x_1,\ldots,x_N\in[a,b]$ being random.

PROOF. We recall from Theorem 13.7 that the idea is that we have a formula as follows, with the points $x_1, \ldots, x_N \in [a, b]$ being uniformly distributed:

$$\int_a^b f(x)dx = (b-a) \times \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(x_i)$$

But this works as well when the points $x_1, \ldots, x_N \in [a, b]$ are randomly distributed, for somewhat obvious reasons, and this gives the result. □

Observe that Monte Carlo integration works better than Riemann integration, for instance when trying to improve the estimate, via $N \to N + 1$. Indeed, in the context of Riemann integration, assume that we managed to find an estimate as follows, which in practice requires computing N values of our function f, and making their average:

$$\int_a^b f(x)dx \simeq \frac{b-a}{N} \sum_{k=1}^{N} f\left(a + \frac{b-a}{N} \cdot k\right)$$

In order to improve this estimate, any extra computed value of our function $f(y)$ will be unuseful. For improving our formula, what we need are N extra values of our function, $f(y_1), \ldots, f(y_N)$, with the points $y_1, \ldots, y_N$ being the midpoints of the previous division of $[a, b]$, so that we can write an improvement of our formula, as follows:

$$\int_a^b f(x)dx \simeq \frac{b-a}{2N} \sum_{k=1}^{2N} f\left(a + \frac{b-a}{2N} \cdot k\right)$$

With Monte Carlo, things are far more flexible. Assume indeed that we managed to find an estimate as follows, which again requires computing N values of our function:

$$\int_a^b f(x)dx \simeq \frac{b-a}{N} \sum_{k=1}^{N} f(x_i)$$

Now if we want to improve this, any extra computed value of our function $f(y)$ will be helpful, because we can set $x_{n+1} = y$, and improve our estimate as follows:

$$\int_a^b f(x)dx \simeq \frac{b-a}{N+1} \sum_{k=1}^{N+1} f(x_i)$$

And isn't this potentially useful, and powerful, when thinking at practically computing integrals, either by hand, or by using a computer. Let us record this finding as follows:

CONCLUSION 13.9. *Monte Carlo integration works better than Riemann integration, when it comes to computing as usual, by estimating, and refining the estimate.*

As another interesting feature of Monte Carlo integration, this works better than Riemann integration, for functions having various symmetries, because Riemann integration can get "fooled" by these symmetries, while Monte Carlo remains strong.

As an example for this phenomeon, chosen to be quite drastic, let us attempt to integrate, via both Riemann and Monte Carlo, the following function $f : [0, \pi] \to \mathbb{R}$:

$$f(x) = \Big|\sin(120x)\Big|$$

The first few Riemann sums for this function are then as follows:

$$I_2(f) = \frac{\pi}{2}(|\sin 0| + |\sin 60\pi|) = 0$$

$$I_3(f) = \frac{\pi}{3}(|\sin 0| + |\sin 40\pi| + |\sin 80\pi|) = 0$$

$$I_4(f) = \frac{\pi}{4}(|\sin 0| + |\sin 30\pi| + |\sin 60\pi| + |\sin 90\pi|) = 0$$

$$I_5(f) = \frac{\pi}{5}(|\sin 0| + |\sin 24\pi| + |\sin 48\pi| + |\sin 72\pi| + |\sin 96\pi|) = 0$$

$$I_6(f) = \frac{\pi}{6}(|\sin 0| + |\sin 20\pi| + |\sin 40\pi| + |\sin 60\pi| + |\sin 80\pi| + |\sin 100\pi|) = 0$$

$$\vdots$$

Based on this evidence, we will conclude, obviously, that we have:

$$\int_0^\pi f(x)dx = 0$$

With Monte Carlo, however, such things cannot happen. Indeed, since there are finitely many points $x \in [0, \pi]$ having the property $\sin(120x) = 0$, a random point $x \in [0, \pi]$ will have the property $|\sin(120x)| > 0$, so Monte Carlo will give, at any $N \in \mathbb{N}$:

$$\int_0^\pi f(x)dx \simeq \frac{b-a}{N}\sum_{k=1}^{N} f(x_i) > 0$$

Again, this is something interesting, when practically computing integrals, either by hand, or by using a computer. So, let us record, as a complement to Conclusion 13.9:

Conclusion 13.10. *Monte Carlo integration is smarter than Riemann integration, because the symmetries of the function can fool Riemann, but not Monte Carlo.*

All this is good to know, when computing integrals in practice, especially with a computer. Finally, here is one more useful interpretation of the integral:

Theorem 13.11. *The integral of a function $f : [a, b] \to \mathbb{R}$ is given by*

$$\int_a^b f(x)dx = (b-a) \times E(f)$$

where $E(f)$ is the expectation of f, regarded as random variable.

Proof. This is just some sort of fancy reformulation of Theorem 13.7, the idea being that what we can "expect" from a random variable is of course its average. We will be back to this later in this book, when systematically discussing probability theory. □

13b. Riemann sums

Our purpose now will be to understand which functions $f : \mathbb{R} \to \mathbb{R}$ are integrable, and how to compute their integrals. For this purpose, the Riemann formula in Theorem 13.6 will be our favorite tool. Let us begin with some theory. We first have:

THEOREM 13.12. *The following functions are integrable:*

(1) *The piecewise continuous functions.*
(2) *The piecewise monotone functions.*

PROOF. This is indeed something quite standard, as follows:

(1) It is enough to prove the first assertion for a function $f : [a, b] \to \mathbb{R}$ which is continuous, and our claim here is that this follows from the uniform continuity of f. To be more precise, given $\varepsilon > 0$, let us choose $\delta > 0$ such that the following happens:

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

In order to prove the result, let us pick two divisions of $[a, b]$, as follows:

$$I = [a = a_1 < a_2 < \ldots < a_n = b]$$

$$I' = [a = a'_1 < a'_2 < \ldots < a'_m = b]$$

Our claim, which will prove the result, is that if these divisions are sharp enough, of resolution $< \delta/2$, then the associated Riemann sums $\Sigma_I(f), \Sigma_{I'}(f)$ are close within ε:

$$a_{i+1} - a_i < \frac{\delta}{2}\ ,\ a'_{i+1} - a'_i < \delta_2 \implies \Big|\Sigma_I(f) - \Sigma_{I'}(f)\Big| < \varepsilon$$

(2) In order to prove this claim, let us denote by l the length of the intervals on the real line. Our assumption is that the lengths of the divisions I, I' satisfy:

$$l\big([a_i, a_{i+1}]\big) < \frac{\delta}{2} \quad , \quad l\big([a'_i, a'_{i+1}]\big) < \frac{\delta}{2}$$

Now let us intersect the intervals of our divisions I, I', and set:

$$l_{ij} = l\big([a_i, a_{i+1}] \cap [a'_j, a'_{j+1}]\big)$$

The difference of Riemann sums that we are interested in is then given by:

$$\begin{aligned}\Big|\Sigma_I(f) - \Sigma_{I'}(f)\Big| &= \left|\sum_{ij} l_{ij} f(a_i) - \sum_{ij} l_{ij} f(a'_j)\right| \\ &= \left|\sum_{ij} l_{ij}(f(a_i) - f(a'_j))\right|\end{aligned}$$

(3) Now let us estimate $f(a_i) - f(a'_j)$. Since in the case $l_{ij} = 0$ we do not need this estimate, we can assume $l_{ij} > 0$. Now by remembering what the definition of the numbers l_{ij} was, we conclude that we have at least one point $x \in \mathbb{R}$ satisfying:

$$x \in [a_i, a_{i+1}] \cap [a'_j, a'_{j+1}]$$

But then, by using this point x and our assumption on I, I' involving δ, we get:

$$\begin{aligned} |a_i - a'_j| &\leq |a_i - x| + |x - a'_j| \\ &\leq \frac{\delta}{2} + \frac{\delta}{2} \\ &= \delta \end{aligned}$$

Thus, according to our definition of δ from (1), in relation to ε, we get:

$$|f(a_i) - f(a'_j)| < \varepsilon$$

(4) But this is what we need, in order to finish. Indeed, with the estimate that we found, we can finish the computation started in (2), as follows:

$$\begin{aligned} \left|\Sigma_I(f) - \Sigma_{I'}(f)\right| &= \left|\sum_{ij} l_{ij}(f(a_i) - f(a'_j))\right| \\ &\leq \varepsilon \sum_{ij} l_{ij} \\ &= \varepsilon(b - a) \end{aligned}$$

Thus our two Riemann sums are close enough, provided that they are both chosen to be fine enough, and this finishes the proof of the first assertion.

(5) Regarding now the second assertion, this is something more technical, that we will not really need in what follows. We will leave the proof here, which uses similar ideas to those in the proof of (1) above, namely subdivisions and estimates, as an exercise.

(6) Finally, let us mention that the present result is just a beginning, and many other things can be said about the integrable functions, and about the non-integrable functions too. For more on all this, have a look at any specialized measure theory book. □

In what follows we will be mainly interested in the continuous functions. As a useful complement to the various theoretical results from the previous section, we have:

THEOREM 13.13. *Given a continuous function $f : [a, b] \to \mathbb{R}$, we have*

$$\exists c \in [a, b] \quad , \quad \int_a^b f(x)dx = (b - a)f(c)$$

with this being called mean value property.

PROOF. Our claim is that this follows from the following trivial estimate:

$$\min(f) \leq f \leq \max(f)$$

Indeed, by integrating this over $[a, b]$, we obtain the following estimate:

$$(b-a)\min(f) \leq \int_a^b f(x)dx \leq (b-a)\max(f)$$

Now observe that this latter estimate can be written as follows:

$$\min(f) \leq \frac{\int_a^b f(x)dx}{b-a} \leq \max(f)$$

Since f must takes all values on $[\min(f), \max(f)]$, we get a $c \in [a, b]$ such that:

$$\frac{\int_a^b f(x)dx}{b-a} = f(c)$$

Thus, we are led to the conclusion in the statement. □

At the level of examples now, let us first look at the simplest functions that we know, namely the power functions $f(x) = x^p$. However, things here are tricky, as follows:

THEOREM 13.14. *We have the integration formula*

$$\int_a^b x^p dx = \frac{b^{p+1} - a^{p+1}}{p+1}$$

valid for any $p \in \mathbb{N}$.

PROOF. This is something quite tricky, the idea being as follows:

(1) By linearity we can assume that our interval $[a, b]$ is of the form $[0, c]$, and the formula that we want to establish is as follows:

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

(2) We can further assume $c = 1$, and by expressing the left term as a Riemann sum, we are in need of the following estimate, in the $N \to \infty$ limit:

$$1^p + 2^p + \ldots + N^p \simeq \frac{N^{p+1}}{p+1}$$

(3) So, let us try to prove this. At $p = 0$, obviously nothing to do, because we have the following formula, which is exact, and which proves our estimate:

$$1^0 + 2^0 + \ldots + N^0 = N$$

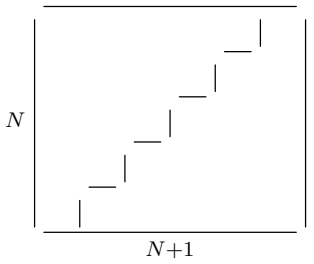
(4) At $p = 1$ now, we are confronted with a well-known question, namely the computation of $1 + 2 + \ldots + N$. But this is simplest done by arguing that the average of the numbers $1, 2, \ldots, N$ being the number in the middle, we have:

$$\frac{1 + 2 + \ldots + N}{N} = \frac{N + 1}{2}$$

Thus, we obtain the following formula, which again solves our question:

$$1 + 2 + \ldots + N = \frac{N(N + 1)}{2} \simeq \frac{N^2}{2}$$

(5) At $p = 2$ now, go compute $1^2 + 2^2 + \ldots + N^2$. This is not obvious, so as a preliminary, let us go back to the case $p = 1$, and try to find a new proof there, which might have some chances to extend at $p = 2$. And here, we have the following trick:



Now the point is that this trick works at $p = 2$ too. Indeed, if we consider the 3D shape P formed by a succession of solid squares, having sizes 1×1, 2×2, 3×3, and so on up to $N \times N$, if we stack 6 copies of P we get a parallelepiped, which gives:

$$1^2 + 2^2 + \ldots + N^2 = \frac{N(N + 1)(2N + 1)}{6} \simeq \frac{N^3}{3}$$

Alternatively, you can get this by computing $1^2 + 2^2 + \ldots + N^2$ for small values of N, then conjecturing the above formula, and proving your conjecture by recurrence.

(6) At $p = 3$ now, the legend has it that by deeply thinking in 4D we are led to the following formula, a bit as in the cases $p = 1, 2$, explained above:

$$1^3 + 2^3 + \ldots + N^3 = \frac{N^2(N + 1)^2}{4} \simeq \frac{N^4}{4}$$

Alternatively, assuming that the gods of combinatorics are with us, we can see right away the following formula, which coupled with (2) gives the result:

$$1^3 + 2^3 + \ldots + N^3 = (1 + 2 + \ldots + N)^2$$

In any case, in practice, the above formula holds indeed, and you can of course come upon it via some numerics too, and then check it by recurrence.

(7) All this is very nice, but coming now as bad news, at $p = 4, 5, 6, \ldots$ the situation is more complicated, with the formulae, provable by recurrence, being as follows:

$$1^4 + 2^4 + \ldots + N^4 = \frac{N(N+1)(2N+1)(3N^2+3N-1)}{30} \simeq \frac{N^5}{5}$$

$$1^5 + 2^5 + \ldots + N^5 = \frac{N^2(N+1)^2(2N^2+2N+1)}{12} \simeq \frac{N^6}{6}$$

$$1^6 + 2^6 + \ldots + N^6 = \frac{N(N+1)(2N+1)(3N^4+6N^3-3N+1)}{42} \simeq \frac{N^7}{7}$$

$$\vdots$$

(8) In fact, we have the following formula, valid at any $p \in \mathbb{N}$, which is provable by recurrence, making appear the Bernoulli numbers B_k, that we met in chapter 11:

$$1^p + 2^p + \ldots + N^p = \frac{1}{p+1} \sum_{k=0}^{p} (-1)^k \binom{p+1}{k} B_k N^{p+1-k}$$

To be more precise, this formula is compatible with those above at $p = 0, \ldots, 6$, with the data for the first Bernoulli numbers, from chapter 11, being as follows:

$$B_0 = 1 \ , \quad B_1 = -\frac{1}{2} \ , \quad B_2 = \frac{1}{6} \ , \quad B_3 = 0 \ , \quad B_4 = -\frac{1}{30} \ , \quad B_5 = 0 \ , \quad B_6 = \frac{1}{42}$$

As for the proof, when trying to establish this by recurrence, we are led into the recurrence formula for the Bernoulli numbers, from chapter 11, which was as follows:

$$\sum_{k=0}^{m} \binom{m+1}{k} B_k = \delta_{m0}$$

And we will leave clarifying this, and some further learning here, as an exercise.

(9) Now the point is that the above formula does the job for our integration purposes, because we obtain right away the following estimate, exactly as needed:

$$1^p + 2^p + \ldots + N^p \simeq \frac{N^{p+1}}{p+1}$$

Summarizing, theorem proved, with the cases $p = 0, 1, 2, 3$ being elementary, then $p = 4, 5, 6, \ldots$ being doable too, and $p \in \mathbb{N}$ general requiring Bernoulli numbers. □

As a continuation of this, and following somehow the story with the other occurrences of the power functions x^p, from this book, in the context of the generalized binomial formula, and of the computation of derivatives too, we can safely conjecture that the formula in Theorem 13.14 should hold as well for negative exponents, $p \in -\mathbb{N}$, via some combinatorial work, then in fact for any $p \in \mathbb{Q}$, via some further combinatorial work, and then in fact at any $p \in \mathbb{R}$, say by invoking a suitable continuity argument.

Ready for this? In the hope that it is so, because you are young and enthusiastic, and loving complicated computations. However, in what regards myself, well, I'm now quite old, and I'd rather leave this as a conjecture, in waiting for better times and ideas:

CONJECTURE 13.15. *We have the following estimate,*

$$1^p + 2^p + \ldots + N^p \simeq \frac{N^{p+1}}{p+1}$$

and so, by Riemann sums, we have the following integration formula,

$$\int_a^b x^p dx = \frac{b^{p+1} - a^{p+1}}{p+1}$$

valid for any exponent $p \in \mathbb{R}$.

Now, instead of struggling with the above conjecture, let us look at some other functions, which are not polynomial. And here, as good news, we have:

THEOREM 13.16. *We have the following integration formula,*

$$\int_a^b e^x dx = e^b - e^a$$

valid for any two real numbers $a < b$.

PROOF. This follows indeed from the Riemann integration formula, because:

$$\begin{aligned}
\int_a^b e^x dx &= \lim_{N\to\infty} \frac{e^a + e^{a+(b-a)/N} + e^{a+2(b-a)/N} + \ldots + e^{a+(N-1)(b-a)/N}}{N} \\
&= \lim_{N\to\infty} \frac{e^a}{N} \cdot \left(1 + e^{(b-a)/N} + e^{2(b-a)/N} + \ldots + e^{(N-1)(b-a)/N}\right) \\
&= \lim_{N\to\infty} \frac{e^a}{N} \cdot \frac{e^{b-a} - 1}{e^{(b-a)/N} - 1} \\
&= (e^b - e^a) \lim_{N\to\infty} \frac{1}{N(e^{(b-a)/N} - 1)} \\
&= e^b - e^a
\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Summarizing, we have some Riemann sum knowledge for the power functions, and the exponentials. For other functions, such as the trigonometric ones, such computations can be quite complicated, and so, with due apologies, we will have to stop our study here.

13c. Primitives, rules

The problem is now, what to do with what we have, from the above. Not obvious, so stuck, and as always in such situations, time to ask the cats. And the cats say:

CATS 13.17. *Summing the infinitesimals of the rate of change of the function should give you the global change of the function. Obvious.*

Which is quite puzzling, usually my cats are quite helpful. It is either that Nicolas learned some bad analytic tricks from Vladimir, or perhaps vice versa, that Vladimir leaned some bad algebraic methods from Nicolas, or most likely, both.

This being said, wait. There is suggestion to connect integrals and derivatives, and this is in fact what we have, coming from what we have from before, due to:

$$\left(\frac{x^{p+1}}{p+1}\right)' = x^p \quad , \quad (e^x)' = e^x$$

So, eureka, we have our idea, thanks cats. Moving ahead now, following this idea, we first have the following result, called fundamental theorem of calculus:

THEOREM 13.18. *Given a continuous function $f : [a, b] \to \mathbb{R}$, if we set*

$$F(x) = \int_a^x f(s)ds$$

then $F' = f$. That is, the derivative of the integral is the function itself.

PROOF. This follows from the Riemann integration picture, and more specifically, from the mean value property from Theorem 13.13. Indeed, we have:

$$\frac{F(x+t) - F(x)}{t} = \frac{1}{t}\int_x^{x+t} f(x)dx$$

On the other hand, our function f being continuous, by using the mean value property from Theorem 13.13, we can find a number $c \in [x, x+t]$ such that:

$$\frac{1}{t}\int_x^{x+t} f(x)dx = f(x)$$

Thus, putting our formulae together, we conclude that we have:

$$\frac{F(x+t) - F(x)}{t} = f(c)$$

Now with $t \to 0$, no matter how the number $c \in [x, x+t]$ varies, one thing that we can be sure about is that we have $c \to x$. Thus, by continuity of f, we obtain:

$$\lim_{t\to 0} \frac{F(x+t) - F(x)}{t} = f(x)$$

But this means exactly that we have $F' = f$, and we are done. □

We have as well the following result, also called fundamental theorem of calculus:

THEOREM 13.19. *Given a function $F : \mathbb{R} \to \mathbb{R}$, we have*

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

for any interval $[a, b]$.

PROOF. As already mentioned, this is something which follows from Theorem 13.18, and is in fact equivalent to it. Indeed, consider the following function:

$$G(s) = \int_a^s F'(x)dx$$

By using Theorem 13.18 we have $G' = F'$, and so our functions F, G differ by a constant. But with $s = a$ we have $G(a) = 0$, and so the constant is $F(a)$, and we get:

$$F(s) = G(s) + F(a)$$

Now with $s = b$ this gives $F(b) = G(b) + F(a)$, which reads:

$$F(b) = \int_a^b F'(x)dx + F(a)$$

Thus, we are led to the conclusion in the statement. □

As a first illustration for all this, solving our previous problems, we have:

THEOREM 13.20. *We have the following integration formulae,*

$$\int_a^b x^p dx = \frac{b^{p+1} - a^{p+1}}{p+1} \quad , \quad \int_a^b \frac{1}{x}\, dx = \log\left(\frac{b}{a}\right)$$

$$\int_a^b \sin x\, dx = \cos a - \cos b \quad , \quad \int_a^b \cos x\, dx = \sin b - \sin a$$

$$\int_a^b e^x dx = e^b - e^a \quad , \quad \int_a^b \log x\, dx = b\log b - a\log a - b + a$$

all obtained, in case you ever forget them, via the fundamental theorem of calculus.

PROOF. We already know some of these formulae, but the best is to do everything, using the fundamental theorem of calculus. The computations go as follows:

(1) With $F(x) = x^{p+1}$ we have $F'(x) = px^p$, and we get, as desired:

$$\int_a^b px^p\, dx = b^{p+1} - a^{p+1}$$

(2) Observe first that the formula (1) does not work at $p = -1$. However, here we can use $F(x) = \log x$, having as derivative $F'(x) = 1/x$, which gives, as desired:

$$\int_a^b \frac{1}{x}\,dx = \log b - \log a = \log\left(\frac{b}{a}\right)$$

(3) With $F(x) = \cos x$ we have $F'(x) = -\sin x$, and we get, as desired:

$$\int_a^b -\sin x\,dx = \cos b - \cos a$$

(4) With $F(x) = \sin x$ we have $F'(x) = \cos x$, and we get, as desired:

$$\int_a^b \cos x\,dx = \sin b - \sin a$$

(5) With $F(x) = e^x$ we have $F'(x) = e^x$, and we get, as desired:

$$\int_a^b e^x\,dx = e^b - e^a$$

(6) This is something more tricky. We are looking for a function satisfying:

$$F'(x) = \log x$$

This does not look doable, but fortunately the answer to such things can be found on the internet. But, what if the internet connection is down? So, let us think a bit, and try to solve our problem. Speaking logarithm and derivatives, what we know is:

$$(\log x)' = \frac{1}{x}$$

But then, in order to make appear log on the right, the idea is quite clear, namely multiplying on the left by x. We obtain in this way the following formula:

$$(x\log x)' = 1\cdot\log x + x\cdot\frac{1}{x} = \log x + 1$$

We are almost there, all we have to do now is to substract x from the left, as to get:

$$(x\log x - x)' = \log x$$

But this this formula in hand, we can go back to our problem, and we get the result. □

Getting back now to theory, inspired by the above, let us formulate:

DEFINITION 13.21. *Given f, we call primitive of f any function F satisfying:*

$$F' = f$$

We denote such primitives by $\int f$, and also call them indefinite integrals.

Observe that the primitives are unique up to an additive constant, in the sense that if F is a primitive, then so is $F+c$, for any $c\in\mathbb{R}$, and conversely, if F,G are two primitives, then we must have $G=F+c$, for some $c\in\mathbb{R}$, with this latter fact coming from a result from chapter 9, saying that the derivative vanishes when the function is constant.

As for the convention at the end, $F=\int f$, this comes from the fundamental theorem of calculus, which can be written as follows, by using this convention:

$$\int_a^b f(x)dx=\left(\int f\right)(b)-\left(\int f\right)(a)$$

By the way, observe that there is no contradiction here, coming from the indeterminacy of $\int f$. Indeed, when adding a constant $c\in\mathbb{R}$ to the chosen primitive $\int f$, when computing the above difference the c quantities will cancel, and we obtain the same result.

We can now reformulate Theorem 13.20 in a more digest form, as follows:

THEOREM 13.22. *We have the following formulae for primitives,*

$$\int x^p=\frac{x^{p+1}}{p+1}\quad,\quad\int\frac{1}{x}=\log x$$

$$\int\sin x=-\cos x\quad,\quad\int\cos x=\sin x$$

$$\int e^x=e^x\quad,\quad\int\log x=x\log x-x$$

allowing us to compute the corresponding definite integrals too.

PROOF. Here the various formulae in the statement follow from Theorem 13.20, or rather from the proof of Theorem 13.20, or even from chapter 9, for most of them, and the last assertion comes from the integration formula given after Definition 13.21. □

Getting back now to theory, we have the following key result:

THEOREM 13.23. *We have the formula*

$$\int f'g+\int fg'=fg$$

called integration by parts.

PROOF. This follows by integrating the Leibnitz formula, namely:

$$(fg)'=f'g+fg'$$

Indeed, with our convention for primitives, this gives the formula in the statement. □

In terms of usual integrals, Theorem 13.23 reformulates as follows, with this being called integration by parts too, with the convention $[\varphi]_a^b = \varphi(b) - \varphi(a)$:

$$\int_a^b f'g + \int_a^b fg' = \Big[fg\Big]_a^b$$

In practice, the most interesting case is that when fg vanishes on the boundary $\{a, b\}$ of our interval, leading to the following formula, in this case:

$$\int_a^b f'g = -\int_a^b fg'$$

Now still at the theoretical level, completing our series of theorems, we have:

THEOREM 13.24. *We have the change of variable formula*

$$\int_a^b f(x)dx = \int_c^d f(\varphi(t))\varphi'(t)dt$$

where $c = \varphi^{-1}(a)$ *and* $d = \varphi^{-1}(b)$.

PROOF. This follows with $f = F'$, from the following differentiation rule, that we know from chapter 9, and whose proof is something elementary:

$$(F\varphi)'(t) = F'(\varphi(t))\varphi'(t)$$

Indeed, by integrating between c and d, we obtain the result. □

As a first application now of our theory, in relation with advanced calculus, and more specifically with the Taylor formula from chapter 11, we have:

THEOREM 13.25. *Given a function* $f : \mathbb{R} \to \mathbb{R}$, *we have the formula*

$$f(x+t) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k + \int_x^{x+t} \frac{f^{(n+1)}(s)}{n!}(x+t-s)^n\, ds$$

called Taylor formula with integral formula for the remainder.

PROOF. This is something which looks a bit complicated, so we will first do some verifications, and then go for the proof in general:

(1) At $n = 0$ the formula in the statement is as follows, and certainly holds, due to the fundamental theorem of calculus, which gives $\int_x^{x+t} f'(s)ds = f(x+t) - f(x)$:

$$f(x+t) = f(x) + \int_x^{x+t} f'(s)ds$$

(2) At $n = 1$, the formula in the statement becomes more complicated, as follows:

$$f(x+t) = f(x) + f'(x)t + \int_x^{x+t} f''(s)(x+t-s)ds$$

As a first observation, this formula holds indeed for the linear functions, where we have $f(x+t)=f(x)+f'(x)t$, and $f''=0$. So, let us try $f(x)=x^2$. Here we have:

$$f(x+t)-f(x)-f'(x)t=(x+t)^2-x^2-2xt=t^2$$

On the other hand, the integral remainder is given by the same formula, namely:

$$\begin{aligned}\int_x^{x+t}f''(s)(x+t-s)ds&=2\int_x^{x+t}(x+t-s)ds\\&=2t(x+t)-2\int_x^{x+t}sds\\&=2t(x+t)-((x+t)^2-x^2)\\&=2tx+2t^2-2tx-t^2\\&=t^2\end{aligned}$$

(3) Still at $n=1$, let us try now to prove the formula in the statement, in general. Since what we have to prove is an equality, this cannot be that hard, and the first thought goes towards differentiating. But this method works indeed, and we obtain the result.

(4) In general, the proof is similar, by differentiating, the computations being similar to those at $n=1$, and we will leave this as an instructive exercise. □

Many other things can be said, as a continuation of the above. We will be back to this in chapter 14, with a number of more specialized results, on approximation.

13d. Areas, volumes

Getting now towards more concrete applications of our technology, we can compute all sorts of areas and volumes. Normally such things are rather the business of multivariable calculus, and more on this later in this book, in the final chapter, dedicated to several variables, but with the techniques that we have so far, we can do a number of things.

Let us first talk about ellipses. These are very familiar objects, generalizing the circles, and with our integration methods we can now compute their areas, as follows:

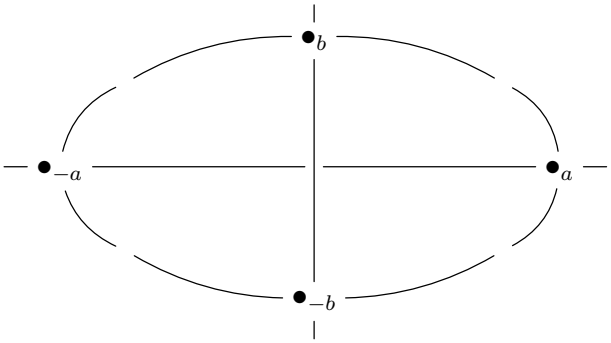
THEOREM 13.26. *The area of an ellipse, given by the equation*

$$\left(\frac{x}{a}\right)^2+\left(\frac{y}{b}\right)^2=1$$

with $a,b>0$ being half the size of a box containing the ellipse, is $A=\pi ab$.

PROOF. This is something very standard, the idea being as follows:

(1) To start with, let us draw a picture. Here that is, making it clear what the parameters $a, b > 0$ stand for, with $2a \times 2b$ being the gift box size for our ellipse:



(2) The idea will be that of cutting the ellipse into vertical slices. First observe that, according to our equation $(x/a)^2 + (y/b)^2 = 1$, the x coordinate can range as follows:

$$x \in [-a, a]$$

For any such x, the other coordinate y, satisfying $(x/a)^2 + (y/b)^2 = 1$, is given by:

$$y = \pm b\sqrt{1 - \frac{x^2}{a^2}}$$

Thus the length of the vertical ellipse slice at x is given by the following formula:

$$l(x) = 2b\sqrt{1 - \frac{x^2}{a^2}}$$

(3) We conclude from this discussion that the area of the ellipse is given by:

$$\begin{aligned} A &= 2b\int_{-a}^{a}\sqrt{1 - \frac{x^2}{a^2}}\,dx \\ &= \frac{4b}{a}\int_{0}^{a}\sqrt{a^2 - x^2}\,dx \\ &= 4ab\int_{0}^{1}\sqrt{1 - y^2}\,dy \\ &= 4ab \cdot \frac{\pi}{4} \\ &= \pi ab \end{aligned}$$

(4) Finally, as a verification, for $a = b = 1$ we get $A = \pi$, as we should. □

Still talking ellipses, in what regards the length things are quite tricky, as follows:

THEOREM 13.27. *The length of an ellipse, given by* $(x/a)^2 + (y/b)^2 = 1$*, is*

$$L = 4\int_0^{\pi/2} \sqrt{a^2 \sin^2 t + b^2 \cos^2 t}\, dt$$

and with this integral being generically not computable.

PROOF. This is something quite surprising, the idea being as follows:

(1) To start with, in the case where our ellipse is a circle, say of radius R, the area and length are related by the formula $A = LR/2$, as we know well from chapter 3, and so we get $L = 2\pi R$. The problem, however, is that the "pizza" argument from chapter 3 obviously does not work for general ellipses, so we must find something else.

(2) So, what is the length of a curve $\gamma : [a, b] \to \mathbb{R}^2$? Good question, and in answer, a physicist would say that this is the quantity obtained by integrating the magnitude of the velocity vector over the curve, with respect to time. But this velocity vector is $\gamma'(t)$, having magnitude $||\gamma'(t)||$, so we are led to the following formula:

$$L(\gamma) = \int_a^b ||\gamma'(t)||dt$$

(3) Regarding now mathematicians, these would say that the length of a curve is the following quantity, with $(t_1 = a, t_2, \ldots, t_{n-1}, t_n = b)$ being a uniform division of (a, b):

$$L(\gamma) = \lim_{n\to\infty} \sum_{i=1}^{n} ||\gamma(t_i) - \gamma(t_{i-1})||$$

But, by using the fundamental theorem of calculus, we can write this as follows:

$$L(\gamma) = \lim_{n\to\infty} \sum_{i=1}^{n} \left|\left| \int_{t_{i-1}}^{t_i} \gamma'(t)dt \right|\right|$$

And the point now is that, by doing some standard analysis, that we will leave here as an instructive exercise, we are led to the formula in (2).

(4) Getting back now to the ellipses, we can compute their length, as follows:

$$\begin{aligned} L &= 4\int_0^{\pi/2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt \\ &= 4\int_0^{\pi/2} \sqrt{\left(\frac{da\cos t}{dt}\right)^2 + \left(\frac{db\sin t}{dt}\right)^2}\, dt \\ &= 4\int_0^{\pi/2} \sqrt{a^2 \sin^2 t + b^2 \cos^2 t}\, dt \end{aligned}$$

(5) As for the last assertion, when $a = b = R$ we get of course $L = 2\pi R$, as we should, but in general, when $a \neq b$, there is no trick for computing the above integral. □

Moving now to 3D, as an obvious challenge here, we can try to compute the area and volume of the sphere, and more generally of the ellipsoids. We have here:

THEOREM 13.28. *The volume of the unit sphere in $\mathbb{R}^3$ is given by:*

$$V = \frac{4\pi}{3}$$

More generally, the volume of an ellipsoid, $(x/a)^2 + (y/b)^2 + (z/c)^2 = 1$, is:

$$V = \frac{4\pi abc}{3}$$

The area of the unit sphere is $A = 4\pi$. For ellipsoids, this is generically not computable.

PROOF. There are several things going on here, as follows:

(1) Let us first compute the volume of the ellipsoid, which at $a = b = c = 1$ will give the volume of the unit sphere. The range of the first coordinate x is as follows:

$$x \in [-a, a]$$

Now observe that when the first coordinate x is fixed, in this range, the other coordinates y, z vary on an ellipse, given by the following equation:

$$\left(\frac{y}{b}\right)^2 + \left(\frac{z}{c}\right)^2 = 1 - \left(\frac{x}{a}\right)^2$$

Next, observe that this latter equation can be written as follows:

$$\left(\frac{y}{\beta}\right)^2 + \left(\frac{z}{\gamma}\right)^2 = 1 \quad : \quad \beta = b\sqrt{1 - \left(\frac{x}{a}\right)^2}\ ,\ \gamma = c\sqrt{1 - \left(\frac{x}{a}\right)^2}$$

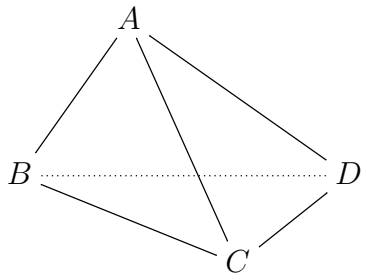
Thus, the vertical slice of our ellipsoid at x has area as follows:

$$A(x) = \pi\beta\gamma = \pi bc\left[1 - \left(\frac{x}{a}\right)^2\right]$$

(2) We conclude that the volume of the ellipsoid is given, as claimed, by:

$$\begin{aligned} V &= \pi bc\int_{-a}^{a} 1 - \left(\frac{x}{a}\right)^2 dx \\ &= \pi bc\left[x - \frac{x^3}{3a^2}\right]_{-a}^{a} \\ &= \pi bc\left(\frac{2a}{3} + \frac{2a}{3}\right) \\ &= \frac{4\pi abc}{3} \end{aligned}$$

(3) Getting to the unit sphere, its volume follows to be $V = 4\pi/3$. In order to convert this into an area formula, we can use the same "pizza" argument as in 2 dimensions, but with a 3 factor appearing. Indeed, consider a tetrahedron, in 3D space:



The volume of this tetrahedron is then given by the following formula, coming for instance by constructing a triangular prism, out of 3 copies of this tetrahedron:

$$\text{volume} = \frac{1}{3} \times \text{basis area} \times \text{height}$$

(4) Now recall the "pizza" argument from 2 dimensions, from chapter 3. By using (3) the same will apply in 3D, giving the following formula, for the area of the sphere:

$$A = 3 \times V = 3 \times \frac{4\pi}{3} = 4\pi$$

(5) Finally, the last assertion, regarding the area of ellipsoids, is something quite informal, coming from the last assertion in Theorem 13.27, which was informal too. □

13e. Exercises

Welcome to integration, and as theoretical exercises on this, we have:

EXERCISE 13.29. *Learn more about the Monte Carlo formula, and its applications.*

EXERCISE 13.30. *Find, then sell, a good algorithm for producing random numbers.*

EXERCISE 13.31. *Learn more, from probabilists, about variables and their expectations.*

EXERCISE 13.32. *Clarify the integration property of piecewise monotone functions.*

EXERCISE 13.33. *Learn a bit about Lebesgue measure, and measurable functions.*

EXERCISE 13.34. *Learn also about the convergence theorems of Lebesgue and Fatou.*

EXERCISE 13.35. *Have a look as well at several variables, and the Fubini theorem.*

EXERCISE 13.36. *And do not hesitate to learn further things, such as Radon-Nikodym.*

As bonus exercise, and no surprise here, explicitly compute 100 integrals.

CHAPTER 14

Heavy analysis

14a. Gauss, Fresnel

Good news, with the differentiation theory discussed in Part III, complemented by the integration theory from chapter 13, we have now all the basic tools of 1-variable calculus, in our bag. And so, starting from this point of this book, there will be no excuses, we will have to do heavy analysis and computations, as heavy as these can get. Hang on.

As a first goal, we have unfinished business with the differential equations. We discussed these in chapter 12, but time and again we said there that the proper tool for dealing with them are the "antiderivatives". But now that we know that the antiderivatives are integrals, time to see how our integration theory applies to them.

Getting to work, we would like to review the 1D wave and heat equations, discussed in chapter 12. In order to deal with the waves, following d'Alembert, we will need the following useful technical result, which is something having its own interest:

PROPOSITION 14.1. *The derivative of a function of type*

$$\varphi(x) = \int_{g(x)}^{h(x)} f(s)ds$$

is given by the formula $\varphi'(x) = f(h(x))h'(x) - f(g(x))g'(x)$.

PROOF. Consider a primitive of the function that we integrate, $F' = f$. We have:

$$\begin{aligned}\varphi(x) &= \int_{g(x)}^{h(x)} f(s)ds\\ &= \int_{g(x)}^{h(x)} F'(s)ds\\ &= F(h(x)) - F(g(x))\end{aligned}$$

By using now the chain rule for derivatives, we obtain from this:

$$\begin{aligned}\varphi'(x) &= F'(h(x))h'(x) - F'(g(x))g'(x)\\ &= f(h(x))h'(x) - f(g(x))g'(x)\end{aligned}$$

Thus, we are led to the formula in the statement. □

Now back to the 1D waves, the general result here, due to d'Alembert, along with a little more, in relation with our lattice models from chapter 10, is as follows:

THEOREM 14.2. *The solution of the 1D wave equation $\ddot{\varphi} = v^2\varphi''$ with initial value conditions $\varphi(x,0) = f(x)$ and $\dot{\varphi}(x,0) = g(x)$ is given by the d'Alembert formula:*

$$\varphi(x,t) = \frac{f(x-vt) + f(x+vt)}{2} + \frac{1}{2v}\int_{x-vt}^{x+vt} g(s)ds$$

Also, in the context of our previous lattice model discretizations, what happens is more or less that the above d'Alembert integral gets computed via Riemann sums.

PROOF. There are several things going on here, the idea being as follows:

(1) Let us first check that the d'Alembert solution is indeed a solution of the wave equation $\ddot{\varphi} = v^2\varphi''$. The first time derivative is computed as follows:

$$\dot{\varphi}(x,t) = \frac{-vf'(x-vt) + vf'(x+vt)}{2} + \frac{1}{2v}(vg(x+vt) + vg(x-vt))$$

The second time derivative is computed as follows:

$$\ddot{\varphi}(x,t) = \frac{v^2f''(x-vt) + v^2f(x+vt)}{2} + \frac{vg'(x+vt) - vg'(x-vt)}{2}$$

Regarding now space derivatives, the first one is computed as follows:

$$\varphi'(x,t) = \frac{f'(x-vt) + f'(x+vt)}{2} + \frac{1}{2v}(g'(x+vt) - g'(x-vt))$$

As for the second space derivative, this is computed as follows:

$$\varphi''(x,t) = \frac{f''(x-vt) + f''(x+vt)}{2} + \frac{g''(x+vt) - g''(x-vt)}{2v}$$

Thus we have indeed $\ddot{\varphi} = v^2\varphi''$. As for the initial conditions, $\varphi(x,0) = f(x)$ is clear from our definition of φ, and $\dot{\varphi}(x,0) = g(x)$ is clear from our above formula of $\dot{\varphi}$.

(2) As a comment here, consider the basic solutions of the wave equation $\ddot{\varphi} = v^2\varphi''$ that we found in chapter 12, which were as follows, with $w/k = v$:

$$\varphi(x,t) = A\cos(kx - wt + \delta)$$

The initial data for these particular solutions is then as follows:

$$f(x) = \varphi(x,0) = A\cos(kx + \delta)$$

$$g(x) = \dot{\varphi}(x,0) = wA\sin(kx + \delta)$$

Now by plugging this into the d'Alembert formula, and doing the computation, with this being an easy exercise for you, we obtain indeed $\varphi(x,t) = A\cos(kx - wt + \delta)$.

(3) Next, we must show that our solution is unique. And here, instead of going into abstract arguments, we will simply solve our equation, which among others will doublecheck the computations in (1). Let us make the following change of variables:

$$y = x - vt \quad , \quad z = x + vt$$

The point now is that, with this change of variables, which is something quite tricky, mixing space and time variables, and by assuming some multivariable calculus know-how, our wave equation $\ddot{\varphi} = v^2\varphi''$ reformulates in a very simple way, as follows:

$$\frac{d^2\varphi}{dydz} = 0$$

But this latter equation tells us that our new y, z variables get separated, and we conclude from this that the solution must be of the following special form:

$$\varphi(x,t) = F(y) + G(z) = F(x - vt) + G(x + vt)$$

Now by taking into account the intial conditions $\varphi(x,0) = f(x)$ and $\dot{\varphi}(x,0) = g(x)$, and then integrating, we are led to the d'Alembert formula in the statement.

(4) Summarizing, uniqueness proved, modulo some multivariable calculus trickery, that we will actually learn later in this book, in chapter 16. We will be back to this.

(5) In regards now with our discretization questions, by using a 1D lattice model with balls and springs as before, what happens to all the above is more or less that the d'Alembert integral gets computed via Riemann sums, in our model, as stated.

(6) Finally, as one more thing that can be said, and which is something quite intuitive, when thinking at waves, the solutions appear in fact as superpositions of the basic solutions from (2). And more on this in chapter 15, when discussing Fourier analysis.

(7) And we will end our discussion here. To make a summary, we talked about waves in chapter 10, then in chapter 12 with more tools, then here with even more tools, but the story is certainly not over here. And more on this, later in this book. □

Very nice all this, at least we know one thing. Getting now to the heat equation, $\dot{\varphi} = \alpha\varphi''$, we know from chapter 12 that the basic solutions are as follows:

$$\varphi(x,t) = \frac{1}{\sqrt{t}}\, e^{-x^2/4\alpha t}$$

Which brings us first, mathematically, into the question of integrating the exponentials on the right. And here, we have the following famous result, due to Gauss:

THEOREM 14.3. *We have the following formula,*

$$\int_{\mathbb{R}} e^{-x^2} dx = \sqrt{\pi}$$

called Gauss integral formula.

PROOF. This is something truly magic, the idea being as follows:

(1) To start with, we can certainly integrate e^{-x^2} by using the formula of the exponential series, and the primitive which is worth 0 at $x = 0$ is given by:

$$\int e^{-x^2} = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)k!}$$

However, this series is not computable, in terms of the known, familiar series.

(2) Thus, no primitive, but we can still ask for the computation of $\int_{\mathbb{R}} e^{-x^2} dx$, who knows. And here, another surprise awaits us, this is simply undoable, with bare hands, I mean all the formulae and tricks that we learned in chapter 13 fail, for this integral.

(3) Which seems to send our problem to the trash can. However, and here comes the magic, the Gauss integral can be computed by using two dimensions, as follows:

$$\begin{aligned}
\int_{\mathbb{R}} \int_{\mathbb{R}} e^{-x^2-y^2} dxdy &= 4 \int_0^{\infty} \int_0^{\infty} e^{-x^2-y^2} dxdy \\
&= 4 \int_0^{\infty} \int_0^{\infty} e^{-t^2y^2-y^2} ydtdy \\
&= 4 \int_0^{\infty} \int_0^{\infty} ye^{-y^2(1+t^2)} dydt \\
&= 2 \int_0^{\infty} \int_0^{\infty} \left(-\frac{e^{-y^2(1+t^2)}}{1+t^2} \right)' dydt \\
&= 2 \int_0^{\infty} \frac{dt}{1+t^2} \\
&= 2 \int_0^{\infty} (\arctan t)' dt \\
&= \pi
\end{aligned}$$

(4) Amazing all this, isn't it. Let us mention too that there are some other known proof of the Gauss formula, but all using two dimensions. Among these, we have for instance the following quick computation, assuming some polar coordinate know-how:

$$\begin{aligned}
\int_{\mathbb{R}} \int_{\mathbb{R}} e^{-x^2-y^2} dxdy &= \int_0^{2\pi} \int_0^{\infty} e^{-r^2} rdrdt \\
&= 2\pi \int_0^{\infty} \left(-\frac{e^{-r^2}}{2} \right)' dr \\
&= \pi
\end{aligned}$$

And more on this, and more specifically on $dxdy = rdrdt$, in chapter 16. □

Getting back now to the heat equation, we have the following result, about it:

THEOREM 14.4. *The heat equation, $\dot{\varphi} = \alpha\varphi''$ with $\alpha > 0$, with initial condition $\varphi(x, 0) = f(x)$, has as solution the function*

$$\varphi(x, t) = \int_{\mathbb{R}} K_t(x - y) f(y) dy$$

where the function $K_t : \mathbb{R} \to \mathbb{R}$, called heat kernel, given by

$$K_t(x) = \frac{1}{\sqrt{4\pi\alpha t}}\, e^{-x^2/4\alpha t}$$

is the standard solution, coming from $f = \delta_0$, normalized as to have mass 1.

PROOF. There are several things going on here, the idea being as follows:

(1) To start with, the function K_t in the statement is indeed the standard solution that we found in chapter 12, normalized as to have mass 1, with this coming from:

$$\begin{aligned}
\int_{\mathbb{R}} K_t(x) dx &= \frac{1}{\sqrt{4\pi\alpha t}} \int_{\mathbb{R}} e^{-x^2/4\alpha t} dx \\
&= \frac{1}{\sqrt{4\pi\alpha t}} \int_{\mathbb{R}} e^{-y^2} \sqrt{4\alpha t}\, dy \\
&= \frac{1}{\sqrt{4\pi\alpha t}} \cdot \sqrt{4\alpha t} \cdot \sqrt{\pi} \\
&= 1
\end{aligned}$$

(2) Next, we can see that by plugging $f = \delta_0$ in the statement, we obtain precisely this standard solution K_t. And with this being something which is not surprising, K_t coming from the simplest situation, that of a radiating point body placed at 0.

(3) Getting now to computations, let us first recall from chapter 12 the verifications for the basic solution K_t itself. Taking into account our normalization factor $1/\sqrt{4\pi\alpha}$, introduced for having mass 1, the time derivative of this function was given by:

$$\dot{K}_t = -\frac{1}{2t\sqrt{4\pi\alpha t}}\, e^{-x^2/4\alpha t} + \frac{x^2}{4\alpha t^2\sqrt{4\pi\alpha t}}\, e^{-x^2/4\alpha t}$$

Regarding the first space derivative, this was given by the following formula:

$$K_t' = -\frac{x}{2\alpha t\sqrt{4\pi\alpha t}}\, e^{-x^2/4\alpha t}$$

As for the second space derivative, this was given by the following formula:

$$K_t'' = -\frac{1}{2\alpha t\sqrt{4\pi\alpha t}}\, e^{-x^2/4\alpha t} + \frac{x^2}{4\alpha^2 t^2\sqrt{4\pi\alpha t}}\, e^{-x^2/4\alpha t}$$

And we can see that the heat equation $\dot{\varphi} = \alpha\varphi''$ is indeed satisfied.

(4) Now the point is that when perturbing the heat kernel K_t by an arbitrary function f, according to the procedure in the statement, which is called convolution, all these computations will perturb well to, according to the following two formulae:

$$\dot{\varphi}(x,t)=\int_{\mathbb{R}}\dot{K}_t(x-y)f(y)dy$$

$$\varphi''(x,t)=\int_{\mathbb{R}}K_t''(x-y)f(y)dy$$

Thus, we can see that the heat equation $\dot{\varphi}=\alpha\varphi''$ is indeed satisfied.

(5) So, this was for the story of the heat equation, first studied in chapter 10, then in chapter 12 with more tools, and then here with even more tools, and as before with the waves, things are certainly not over here. More about this, later in this book. □

We would like to end this section with a quick discussion of the Fresnel integrals. These integrals, coming as well from physics, and more specifically from the work on Fresnel on optics, are quite similar to the Gauss integral, the result being as follows:

THEOREM 14.5. *We have the following formulae,*

$$\int_0^\infty\sin(t^2)dt=\int_0^\infty\cos(t^2)dt=\sqrt{\frac{\pi}{8}}$$

with these being called Fresnel integrals.

PROOF. This is something quite tricky, a bit as before for the Gauss integral, that we discussed in Theorem 14.3, the idea with this being as follows:

(1) As before with Gauss, we can certainly integrate $\sin(t^2)$ and $\cos(t^2)$ by using the series of sin and cos, and the primitives which are worth 0 at $x=0$ are given by:

$$\int\sin(t^2)=\sum_{k=0}^\infty(-1)^k\frac{x^{4k+3}}{(4k+3)(2k+1)!}$$

$$\int\cos(t^2)=\sum_{k=0}^\infty(-1)^k\frac{x^{4k+1}}{(4k+1)(2k)!}$$

However, these series are not computable, in terms of the known, familiar series.

(2) Next, and again as before with Gauss, we can try however to compute the integrals in the statement, and here the basic tools of one-variable calculus fail. But then, again as with Gauss, the two dimensions come to the rescue. In order to discuss this, observe first that, due to $e^{it^2}=\cos(t^2)+i\sin(t^2)$, the Fresnel formulae are equivalent to:

$$\int_0^\infty e^{it^2}dt=\sqrt{\frac{\pi}{2}}\cdot\frac{1+i}{2}$$

(3) So, let us prove this. For this purpose, consider the following function:

$$f(t) = \int_0^\infty \frac{e^{(i-u^2)t^2}}{i-u^2}\,du$$

The derivative of this function is then given by the following formula:

$$\begin{aligned} f'(t) &= 2t\int_0^\infty e^{(i-u^2)t^2}du \\ &= 2te^{it^2}\int_0^\infty e^{-u^2t^2}du \\ &= 2e^{it^2}\int_0^\infty e^{-v^2}dv \\ &= 2e^{it^2}\times\frac{\sqrt{\pi}}{2} \\ &= \sqrt{\pi}e^{it^2} \end{aligned}$$

(4) Now let us integrate this derivative, from 0 to ∞. We obtain in this way:

$$\begin{aligned} \sqrt{\pi}\int_0^\infty e^{it^2}dt &= \int_0^\infty f'(t)dt \\ &= f(\infty)-f(0) \\ &= 0-\int_0^\infty \frac{1}{i-u^2}\,du \\ &= \int_0^\infty \frac{1}{u^2-i}\,du \end{aligned}$$

Summarizing, we have obtained the following formula, for the Fresnel integral:

$$\int_0^\infty e^{it^2}dt = \frac{1}{\sqrt{\pi}}\int_0^\infty \frac{1}{u^2-i}\,du$$

(5) In order to compute the latter integral, set $w = e^{\pi i/4}$. Then $w^2 = i$, and so:

$$\begin{aligned} \int_0^\infty \frac{1}{u^2-i}\,du &= \int_0^\infty \frac{1}{u^2-w^2}\,du \\ &= \int_0^\infty \frac{1}{2w}\left(\frac{1}{u-w}-\frac{1}{u+w}\right)du \\ &= \frac{1}{2w}\left[\log\left(\frac{u-w}{u+w}\right)\right]_0^\infty \\ &= -\frac{1}{2w}\log(-1) \end{aligned}$$

Which brings us to the question, what is $\log(-1)$, in relation with this. In answer, and trust me here, $\log(-1) = -\pi i$, and we can finish our computation as follows:

$$\int_0^\infty \frac{1}{u^2 - i}\, du = -\frac{1}{2w} \times (-\pi i) = \frac{\pi i}{2w} = \frac{\pi w}{2} = \frac{\pi}{2} \cdot \frac{1+i}{\sqrt{2}}$$

(6) Alternatively, for avoiding this complex number mess, observe first that:

$$\int_0^\infty \frac{1}{u^2 - i}\, du = \int_0^\infty \frac{u^2}{u^4 + 1}\, du + i \int_0^\infty \frac{1}{u^4 + 1}\, du$$

Next, the two real integrals are equal, because with $u \to u^{-1}$ we obtain:

$$\int_0^\infty \frac{u^2}{u^4 + 1}\, du = \int_0^\infty \frac{u^{-2}}{u^{-4} + 1}\, u^{-2} du = \int_0^\infty \frac{1}{u^4 + 1}\, du$$

Also, we can compute the sum of these integrals by using $t = u - u^{-1}$, as follows:

$$\int_0^\infty \frac{u^2 + 1}{u^4 + 1}\, du = \int_0^\infty \frac{1 + u^{-2}}{u^2 + u^{-2}}\, du = \int_0^\infty \frac{dt}{t^2 + 2} = \frac{\pi}{\sqrt{2}}$$

Thus, we are led to the following conclusion, exactly as we found in (5):

$$\int_0^\infty \frac{1}{u^2 - i}\, du = \frac{\pi}{2\sqrt{2}} + i\, \frac{\pi}{2\sqrt{2}} = \frac{\pi}{2} \cdot \frac{1+i}{\sqrt{2}}$$

(7) Summarizing, computation done, one way or another, and this gives:

$$\int_0^\infty e^{it^2} dt = \frac{1}{\sqrt{\pi}} \times \frac{\pi}{2} \cdot \frac{1+i}{\sqrt{2}} = \sqrt{\frac{\pi}{2}} \cdot \frac{1+i}{2}$$

But this is exactly what we wanted, and this ends the proof of our result. □

We will be back to all this later in this chapter, when discussing probability and normal variables, which are intimately related to the Gauss integral formula.

14b. Wallis, Stirling

Switching topics, but still in relation with questions that we have unsolved, we would like to discuss now the computation of the volume of the unit sphere in N dimensions. For this purpose, we will need the very useful Wallis formula, which is as follows:

THEOREM 14.6 (Wallis). *We have the following formulae,*

$$\int_0^{\pi/2} \cos^n t\, dt = \int_0^{\pi/2} \sin^n t\, dt = \left(\frac{\pi}{2}\right)^{\varepsilon(n)} \frac{n!!}{(n+1)!!}$$

where $\varepsilon(n) = 1$ if n is even, and $\varepsilon(n) = 0$ if n is odd, and where

$$m!! = (m-1)(m-3)(m-5)\ldots$$

with the product ending at 2 if m is odd, and ending at 1 if m is even.

PROOF. Let us first compute the integral on the left in the statement:

$$I_n = \int_0^{\pi/2} \cos^n t\, dt$$

We do this by partial integration. We have the following formula:

$$\begin{aligned}(\cos^n t \sin t)' &= n\cos^{n-1} t(-\sin t)\sin t + \cos^n t\cos t\\ &= n\cos^{n+1} t - n\cos^{n-1} t + \cos^{n+1} t\\ &= (n+1)\cos^{n+1} t - n\cos^{n-1} t\end{aligned}$$

By integrating between 0 and $\pi/2$, we obtain the following formula:

$$(n+1)I_{n+1} = nI_{n-1}$$

Thus we can compute I_n by recurrence, and we obtain:

$$\begin{aligned}I_n &= \frac{n-1}{n} I_{n-2}\\ &= \frac{n-1}{n}\cdot\frac{n-3}{n-2} I_{n-4}\\ &= \frac{n-1}{n}\cdot\frac{n-3}{n-2}\cdot\frac{n-5}{n-4} I_{n-6}\\ &\vdots\\ &= \frac{n!!}{(n+1)!!} I_{1-\varepsilon(n)}\end{aligned}$$

But $I_0 = \frac{\pi}{2}$ and $I_1 = 1$, so we get the result. As for the second formula, this follows from the first one, with $t = \frac{\pi}{2} - s$. Thus, we have proved both formulae in the statement. □

Before going further, let us record a useful generalization of Theorem 14.6:

THEOREM 14.7 (Wallis 2). *We have the following formula,*

$$\int_0^{\pi/2} \cos^p t \sin^q t\, dt = \left(\frac{\pi}{2}\right)^{\varepsilon(p)\varepsilon(q)} \frac{p!!q!!}{(p+q+1)!!}$$

where $\varepsilon(p) = 1$ if p is even, and $\varepsilon(p) = 0$ if p is odd, as before.

PROOF. We use the same idea as for Theorem 14.6. Let I_{pq} be the integral in the statement. In order to do the partial integration, observe that we have:

$$\begin{aligned}(\cos^p t \sin^q t)' &= p\cos^{p-1} t(-\sin t)\sin^q t\\ &+ \cos^p t\cdot q\sin^{q-1} t\cos t\\ &= -p\cos^{p-1} t\sin^{q+1} t + q\cos^{p+1} t\sin^{q-1} t\end{aligned}$$

By integrating between 0 and $\pi/2$, we obtain, for $p, q > 0$:

$$pI_{p-1,q+1} = qI_{p+1,q-1}$$

Thus, we can compute I_{pq} by recurrence. When q is even we have:

$$\begin{aligned} I_{pq} &= \frac{q-1}{p+1}\, I_{p+2,q-2} \\ &= \frac{q-1}{p+1} \cdot \frac{q-3}{p+3}\, I_{p+4,q-4} \\ &= \frac{q-1}{p+1} \cdot \frac{q-3}{p+3} \cdot \frac{q-5}{p+5}\, I_{p+6,q-6} \\ &= \vdots \\ &= \frac{p!!q!!}{(p+q)!!}\, I_{p+q} \end{aligned}$$

But the last term comes from the formula in Theorem 14.6, and we obtain the result:

$$\begin{aligned} I_{pq} &= \frac{p!!q!!}{(p+q)!!}\, I_{p+q} \\ &= \frac{p!!q!!}{(p+q)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(p+q)} \frac{(p+q)!!}{(p+q+1)!!} \\ &= \left(\frac{\pi}{2}\right)^{\varepsilon(p)\varepsilon(q)} \frac{p!!q!!}{(p+q+1)!!} \end{aligned}$$

Observe that this gives the result for p even as well, by symmetry. Indeed, we have $I_{pq} = I_{qp}$, by using the following change of variables:

$$t = \frac{\pi}{2} - s$$

In the remaining case now, where both p, q are odd, we can use once again the formula $pI_{p-1,q+1} = qI_{p+1,q-1}$ established above, and the recurrence goes as follows:

$$\begin{aligned} I_{pq} &= \frac{q-1}{p+1}\, I_{p+2,q-2} \\ &= \frac{q-1}{p+1} \cdot \frac{q-3}{p+3}\, I_{p+4,q-4} \\ &= \frac{q-1}{p+1} \cdot \frac{q-3}{p+3} \cdot \frac{q-5}{p+5}\, I_{p+6,q-6} \\ &= \vdots \\ &= \frac{p!!q!!}{(p+q-1)!!}\, I_{p+q-1,1} \end{aligned}$$

Thus, we are led to the formula in the statement. □

We can now compute the volumes of the N-dimensional spheres, as follows:

THEOREM 14.8. *The volume of the unit sphere in $\mathbb{R}^N$ is given by*

$$V=\left(\frac{\pi}{2}\right)^{[N/2]}\frac{2^N}{(N+1)!!}$$

with our usual convention $N!!=(N-1)(N-3)(N-5)\ldots$

PROOF. If we denote by V_N the volume of the unit sphere in $\mathbb{R}^N$, we have:

$$\begin{aligned}V_N&=\int_{-1}^1(1-x^2)^{(N-1)/2}dx\cdot V_{N-1}\\&=2V_{N-1}\int_0^1(1-x^2)^{(N-1)/2}dx\\&=2V_{N-1}\int_0^{\pi/2}(1-\sin^2t)^{(N-1)/2}\cos tdt\\&=2V_{N-1}\int_0^{\pi/2}\cos^{N-1}t\cos tdt\\&=2V_{N-1}\int_0^{\pi/2}\cos^N tdt\end{aligned}$$

Now by recurrence, and using the formula in Theorem 14.6, we obtain:

$$\begin{aligned}V_N&=2^N\int_0^{\pi/2}\cos^N tdt\int_0^{\pi/2}\cos^{N-1}tdt\ldots\ldots\int_0^{\pi/2}\cos tdt\\&=2^N\left(\frac{\pi}{2}\right)^{\varepsilon(N)+\varepsilon(N-1)+\ldots+\varepsilon(1)}\frac{N!!}{(N+1)!!}\cdot\frac{(N-1)!!}{N!!}\ldots\frac{1!!}{2!!}\\&=\left(\frac{\pi}{2}\right)^{\varepsilon(N)+\varepsilon(N-1)+\ldots+\varepsilon(1)}\frac{2^N}{(N+1)!!}\\&=\left(\frac{\pi}{2}\right)^{[N/2]}\frac{2^N}{(N+1)!!}\end{aligned}$$

Thus, we are led to the formula in the statement. □

As main particular cases of the above formula, we have:

PROPOSITION 14.9. *The volumes of the low-dimensional spheres are as follows:*

(1) *At $N=1$, the length of the unit interval is $V=2$.*
(2) *At $N=2$, the area of the unit disk is $V=\pi$.*
(3) *At $N=3$, the volume of the unit sphere is $V=\frac{4\pi}{3}$*
(4) *At $N=4$, the volume of the corresponding unit sphere is $V=\frac{\pi^2}{2}$.*

PROOF. Most of these results are well-known, but we can obtain all of them as particular cases of the general formula in Theorem 14.8, in the obvious way. □

Let us record as well the formula of the area of the sphere, as follows:

THEOREM 14.10. *The area of the unit sphere in $\mathbb{R}^N$ is given by:*

$$A = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N-1)!!}$$

In particular, at $N = 2, 3, 4$ we obtain respectively $A = 2\pi, 4\pi, 2\pi^2$.

PROOF. As shown by the pizza argument from chapters 3 and 13, which extends to N dimensions, the area and volume of the sphere in $\mathbb{R}^N$ are related by:

$$A = N \cdot V$$

Together with the formula in Theorem 14.8 for V, this gives the result. □

Moving on, in order to estimate the volumes of the spheres, we will need:

THEOREM 14.11. *We have the Stirling formula*

$$N! \simeq \left(\frac{N}{e}\right)^N \sqrt{2\pi N}$$

valid in the $N \to \infty$ limit.

PROOF. This is something quite tricky, the idea being as follows:

(1) Let us first see what we can get with Riemann sums. We have:

$$\log(N!) = \sum_{k=1}^{N} \log k \approx \int_1^N \log x \, dx = N \log N - N + 1$$

By exponentiating, this gives the following estimate, which is not bad:

$$N! \approx \left(\frac{N}{e}\right)^N \cdot e$$

(2) We can improve our estimate by replacing the rectangles from the Riemann sum approach to the integrals by trapezoids. In practice, this gives the following estimate:

$$\log(N!) \approx \int_1^N \log x \, dx + \frac{\log 1 + \log N}{2} = N \log N - N + 1 + \frac{\log N}{2}$$

By exponentiating, this gives the following estimate, which gets us closer:

$$N! \approx \left(\frac{N}{e}\right)^N \cdot e \cdot \sqrt{N}$$

(3) In order to conclude, we must take some kind of mathematical magnifier, and carefully estimate the error made in (2). Fortunately, this mathematical magnifier exists, called Euler-Maclaurin formula, and after some tough computations, we get to:

$$N! \simeq \left(\frac{N}{e}\right)^N \sqrt{2\pi N}$$

(4) Alternatively, here is another approach to (3), which better does the job, explaining where the $\sqrt{2\pi}$ factor comes from. First, by partial integration we have:

$$N! = \int_0^\infty x^N e^{-x} dx$$

(5) Since the integrand is sharply peaked at $x = N$, as you can see by computing the derivative of $\log(x^N e^{-x})$, this suggests writing $x = N + y$, and we obtain:

$$\begin{aligned}
\log(x^N e^{-x}) &= N \log x - x \\
&= N \log(N + y) - (N + y) \\
&= N \log N + N \log\left(1 + \frac{y}{N}\right) - (N + y) \\
&\simeq N \log N + N\left(\frac{y}{N} - \frac{y^2}{2N^2}\right) - (N + y) \\
&= N \log N - N - \frac{y^2}{2N}
\end{aligned}$$

(6) By exponentiating, we obtain from this the following estimate:

$$x^N e^{-x} \simeq \left(\frac{N}{e}\right)^N e^{-y^2/2N}$$

(7) Now by integrating, we obtain from this the following estimate, as desired:

$$\begin{aligned}
N! &= \int_0^\infty x^N e^{-x} dx \\
&\simeq \int_{-N}^N \left(\frac{N}{e}\right)^N e^{-y^2/2N}\, dy \\
&\simeq \left(\frac{N}{e}\right)^N \int_{\mathbb{R}} e^{-y^2/2N}\, dy \\
&= \left(\frac{N}{e}\right)^N \sqrt{2N} \int_{\mathbb{R}} e^{-z^2}\, dz \\
&= \left(\frac{N}{e}\right)^N \sqrt{2\pi N}
\end{aligned}$$

(8) And exercise of course for you to learn more about this, as much as you can. □

We can now estimate the volumes of the spheres, as follows:

THEOREM 14.12. *The volume of the unit sphere in $\mathbb{R}^N$ is given by*

$$V \simeq \left(\frac{2\pi e}{N}\right)^{N/2} \frac{1}{\sqrt{\pi N}}$$

in the $N \to \infty$ limit.

PROOF. This is something very standard, the idea being as follows:

(1) We know that the volume of the unit sphere in $\mathbb{R}^N$ is given by:

$$V = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!}$$

(2) But the double factorials can be estimated by using the Stirling formula. Indeed, in the case where $N = 2K$ is even, we have the following computation:

$$\begin{aligned}(N+1)!! &= 2^K K! \\ &\simeq \left(\frac{2K}{e}\right)^K \sqrt{2\pi K} \\ &= \left(\frac{N}{e}\right)^{N/2} \sqrt{\pi N}\end{aligned}$$

As for the case where $N = 2K-1$ is odd, here the estimate goes as follows:

$$\begin{aligned}(N+1)!! &= \frac{(2K)!}{2^K K!} \\ &\simeq \frac{1}{2^K}\left(\frac{2K}{e}\right)^{2K} \sqrt{4\pi K}\left(\frac{e}{K}\right)^K \frac{1}{\sqrt{2\pi K}} \\ &= \left(\frac{2K}{e}\right)^K \sqrt{2} \\ &= \left(\frac{N+1}{e}\right)^{(N+1)/2} \sqrt{2} \\ &= \left(\frac{N}{e}\right)^{N/2} \left(\frac{N+1}{N}\right)^{N/2} \sqrt{\frac{N+1}{e}} \cdot \sqrt{2} \\ &\simeq \left(\frac{N}{e}\right)^{N/2} \sqrt{e} \cdot \sqrt{\frac{N}{e}} \cdot \sqrt{2} \\ &= \left(\frac{N}{e}\right)^{N/2} \sqrt{2N}\end{aligned}$$

(3) Now back to the spheres, when N is even, the estimate goes as follows:

$$\begin{aligned} V &= \left(\frac{\pi}{2}\right)^{N/2}\frac{2^N}{(N+1)!!}\\ &\simeq \left(\frac{\pi}{2}\right)^{N/2}2^N\left(\frac{e}{N}\right)^{N/2}\frac{1}{\sqrt{\pi N}}\\ &= \left(\frac{2\pi e}{N}\right)^{N/2}\frac{1}{\sqrt{\pi N}} \end{aligned}$$

As for the case where N is odd, here the estimate goes as follows:

$$\begin{aligned} V &= \left(\frac{\pi}{2}\right)^{(N-1)/2}\frac{2^N}{(N+1)!!}\\ &\simeq \left(\frac{\pi}{2}\right)^{(N-1)/2}2^N\left(\frac{e}{N}\right)^{N/2}\frac{1}{\sqrt{2N}}\\ &= \sqrt{\frac{2}{\pi}}\left(\frac{2\pi e}{N}\right)^{N/2}\frac{1}{\sqrt{2N}}\\ &= \left(\frac{2\pi e}{N}\right)^{N/2}\frac{1}{\sqrt{\pi N}} \end{aligned}$$

Thus, we are led to the uniform formula in the statement. □

So long for high dimensional spheres and their volumes. We will be back to the above formulae, which are key to modern mathematics, on several occasions, in what follows.

14c. Normal variables

Switching topics, but still in relation with questions that we can now solve, using integration theory, we can talk, in a more systematic way, about probability.

We already met the Poisson laws in chapter 4, and the binomial laws in chapter 6, in relation with our considerations there. As a starting point, in general, we have:

DEFINITION 14.13. *Let X be a probability space, that is, a space with a probability measure, and with the corresponding integration denoted E, and called expectation.*

(1) *The random variables are the integrable functions $f : X \to \mathbb{R}$.*
(2) *The moments of such a variable are the numbers $M_k(f) = E(f^k)$.*
(3) *The law of such a variable is the measure given by $M_k(f) = \int_{\mathbb{R}} x^k d\mu_f(x)$.*

This is of course something quite compact, and the fact that μ_f as above exists indeed is not exactly trivial. But we can do this by looking at formulae of the following type:

$$E(\varphi(f)) = \int_{\mathbb{R}} \varphi(x) d\mu_f(x)$$

Indeed, having this for monomials $\varphi(x) = x^n$, as above, is the same as having it for polynomials $\varphi \in \mathbb{R}[X]$, which in turn is the same as having it for the characteristic functions $\varphi = \chi_I$ of measurable sets $I \subset \mathbb{R}$. Thus, in the end, what we need is:

$$P(f \in I) = \mu_f(I)$$

But this formula can serve as a definition for μ_f, and we are done. Regarding now independence, which is the key notion in probability, we can formulate here:

DEFINITION 14.14. *Two variables $f, g : X \to \mathbb{R}$ are called independent when*

$$E(f^k g^l) = E(f^k)\, E(g^l)$$

happens, for any $k, l \in \mathbb{N}$.

Again, this definition hides some non-trivial things, the idea being a bit as before, namely that of looking at formulae of the following type:

$$E[\varphi(f)\psi(g)] = E[\varphi(f)]\, E[\psi(g)]$$

To be more precise, passing as before from monomials to polynomials, then to characteristic functions, we are led to the usual definition of independence, namely:

$$P(f \in I, g \in J) = P(f \in I)\, P(g \in J)$$

As a first result now, dealing with the mechanism of independence, we have:

THEOREM 14.15. *Assuming that $f, g : X \to \mathbb{R}$ are independent, we have*

$$\mu_{f+g} = \mu_f * \mu_g$$

where $*$ is the convolution operation for real probability measures, given by

$$\int_{\mathbb{R}} \varphi(x) d(\mu * \nu)(x) = \int_{\mathbb{R}\times\mathbb{R}} \varphi(x+y) d\mu(x) d\nu(y)$$

for any function φ. The converse of this holds too.

PROOF. We have the following computation, using the independence of f, g:

$$\int_{\mathbb{R}} x^k d\mu_{f+g}(x) = E((f+g)^k) = \sum_r \binom{k}{r} M_r(f) M_{k-r}(g)$$

On the other hand, we have as well the following computation:

$$\begin{aligned} \int_{\mathbb{R}} x^k d(\mu_f * \mu_g)(x) &= \int_{\mathbb{R}\times\mathbb{R}} (x+y)^k d\mu_f(x) d\mu_g(y) \\ &= \sum_r \binom{k}{r} M_r(f) M_{k-r}(g) \end{aligned}$$

Thus μ_{f+g} and $\mu_f * \mu_g$ have the same moments, so they coincide, as claimed. As for the converse, this is clear too, coming as well from the above computation. □

As a second result now on independence, which is more advanced, we have:

THEOREM 14.16. *Assuming that $f, g : X \to \mathbb{R}$ are independent, we have*

$$F_{f+g} = F_f F_g$$

where $F_f(x) = E(e^{ixf})$ is the Fourier transform.

PROOF. We have indeed the following computation, using Theorem 14.15:

$$\begin{aligned} F_{f+g}(x) &= \int_{\mathbb{R}} e^{ixz} d(\mu_f * \mu_g)(z) \\ &= \int_{\mathbb{R}\times\mathbb{R}} e^{ix(z+t)} d\mu_f(z) d\mu_g(t) \\ &= \int_{\mathbb{R}} e^{ixz} d\mu_f(z) \int_{\mathbb{R}} e^{ixt} d\mu_g(t) \\ &= F_f(x) F_g(x) \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

As a first application, we can now prove the Poisson Limit Theorem, as follows:

THEOREM 14.17 (PLT). *We have the following convergence, in moments,*

$$\left(\left(1 - \frac{t}{n} \right) \delta_0 + \frac{t}{n} \delta_1 \right)^{*n} \to p_t$$

where $p_t = e^{-t} \sum_{k\in\mathbb{N}} \frac{t^k}{k!} \delta_k$ is the Poisson law of parameter $t > 0$.

PROOF. If we denote by ν_n the measure under the convolution sign, we have the following computation, for the Fourier transform of the limit:

$$\begin{aligned} F_{\delta_r}(y) = e^{iry} &\implies F_{\nu_n}(y) = \left(1 - \frac{t}{n} \right) + \frac{t}{n} e^{iy} \\ &\implies F_{\nu_n^{*n}}(y) = \left(\left(1 - \frac{t}{n} \right) + \frac{t}{n} e^{iy} \right)^n \\ &\implies F_{\nu_n^{*n}}(y) = \left(1 + \frac{(e^{iy} - 1)t}{n} \right)^n \\ &\implies F(y) = \exp\left((e^{iy} - 1)t \right) \end{aligned}$$

On the other hand, the Fourier transform of the Poisson law p_t is given by:

$$F_{p_t}(y) = e^{-t} \sum_k \frac{t^k}{k!} e^{iky} = e^{-t} \sum_k \frac{(e^{iy} t)^k}{k!} = \exp\left((e^{iy} - 1)t \right)$$

Thus, we are led to the conclusion in the statement. □

Moving on, to the context of continuous probability, let us formulate:

DEFINITION 14.18. *The normal law of parameter* 1 *is the following measure:*

$$g_1 = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

More generally, the normal law of parameter $t > 0$ *is the following measure:*

$$g_t = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t} dx$$

These are also called Gaussian distributions, with "g" standing for Gauss.

Observe that the above laws have indeed mass 1, as they should. This follows indeed from the Gauss formula, which gives, with the change of variables $x = \sqrt{2t}\, y$:

$$\int_{\mathbb{R}} e^{-x^2/2t} dx = \sqrt{2t} \int_{\mathbb{R}} e^{-y^2} dy = \sqrt{2\pi t}$$

Generally speaking, the normal laws appear as bit everywhere, in real life. The reasons behind this phenomenon come from the Central Limit Theorem (CLT), that we will explain in a moment, after developing some general theory. As a first result, we have:

PROPOSITION 14.19. *We have the variance formula*

$$V(g_t) = t$$

valid for any $t > 0$.

PROOF. The first moment is 0, because our normal law g_t is centered. As for the second moment, this can be computed as follows:

$$\begin{aligned} M_2 &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} x^2 e^{-x^2/2t} dx \\ &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} (tx) \left(-e^{-x^2/2t} \right)' dx \\ &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} t e^{-x^2/2t} dx \\ &= t \end{aligned}$$

We conclude from this that the variance is $V = M_2 = t$. □

More generally now, we have the following useful result, regarding the normal laws:

PROPOSITION 14.20. *The even moments of the normal law are the numbers*

$$M_k(g_t) = t^{k/2} \times k!!$$

where $k!! = (k-1)(k-3)(k-5)\ldots$, *and the odd moments vanish.*

PROOF. We have the following computation, valid for any integer $k \in \mathbb{N}$:

$$\begin{aligned}
M_k &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} y^k e^{-y^2/2t} dy \\
&= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} (ty^{k-1}) \left(-e^{-y^2/2t}\right)' dy \\
&= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} t(k-1) y^{k-2} e^{-y^2/2t} dy \\
&= t(k-1) \times \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} y^{k-2} e^{-y^2/2t} dy \\
&= t(k-1) M_{k-2}
\end{aligned}$$

Thus by recurrence, we are led to the formula in the statement. □

Here is another result, which is the key one for the study of the normal laws:

THEOREM 14.21. *We have the following formula, valid for any $t > 0$:*

$$F_{g_t}(x) = e^{-tx^2/2}$$

In particular, the normal laws satisfy $g_s * g_t = g_{s+t}$, for any $s, t > 0$.

PROOF. The Fourier transform formula can be established as follows:

$$\begin{aligned}
F_{g_t}(x) &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-y^2/2t + ixy} dy \\
&= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-(y/\sqrt{2t} - \sqrt{t/2}ix)^2 - tx^2/2} dy \\
&= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-z^2 - tx^2/2} \sqrt{2t} dz \\
&= \frac{1}{\sqrt{\pi}} e^{-tx^2/2} \int_{\mathbb{R}} e^{-z^2} dz \\
&= e^{-tx^2/2}
\end{aligned}$$

As for the last assertion, this follows from the fact that $\log F_{g_t}$ is linear in t. □

We are now ready to state and prove the CLT, as follows:

THEOREM 14.22 (CLT). *Given random variables $f_1, f_2, f_3, \ldots$ which are i.i.d., centered, and with variance $t > 0$, we have, with $n \to \infty$, in moments,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} f_i \sim g_t$$

where g_t is the Gaussian law of parameter t.

PROOF. In terms of moments, the Fourier transform is given by:

$$F_f(x) = E(e^{ixf}) = E\left(\sum_{k=0}^{\infty}\frac{(ixf)^k}{k!}\right) = \sum_{k=0}^{\infty}\frac{i^kM_k(f)}{k!}\,x^k$$

Thus, the Fourier transform of the variable in the statement is:

$$\begin{aligned} F(x) &= \left[F_f\left(\frac{x}{\sqrt{n}}\right)\right]^n \\ &= \left[1-\frac{tx^2}{2n}+O(n^{-2})\right]^n \\ &\simeq \left[1-\frac{tx^2}{2n}\right]^n \\ &\simeq e^{-tx^2/2} \end{aligned}$$

But this latter function being the Fourier transform of g_t, we obtain the result. □

14d. Gamma, zeta, eta

We have kept the best for the end, arithmetic, featuring the gamma, zeta and eta functions, Bernoulli numbers, and many more. Following Riemann, let us start with:

PROPOSITION 14.23. *We can talk about the Riemann zeta function*

$$\zeta(s) = \sum_{n=1}^{\infty}\frac{1}{n^s}$$

with the exponent being any $s \in \mathbb{C}$ satisfying $Re(s) > 1$.

PROOF. We have indeed the following computation, with $s = r + it$ with $r > 1$:

$$\begin{aligned} |\zeta(s)| &\leq \sum_{n=1}^{\infty}\frac{1}{|n^s|} \\ &= \sum_{n=1}^{\infty}\frac{1}{n^r} \\ &< 1+\int_1^{\infty}\frac{1}{x^r}\,dx \\ &= 1+\frac{1}{r-1} \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

But, how to study zeta? The answer comes via the gamma function:

THEOREM 14.24. *We can talk about the gamma function*

$$\Gamma(s) = \int_0^\infty x^{s-1}e^{-x}\,dx$$

extending the usual factorial of integers, $\Gamma(s) = (s-1)!$.

PROOF. The integral converges indeed, and by partial integration we have:

$$\begin{aligned}\Gamma(s+1) &= \int_0^\infty x^s e^{-x}\,dx\\ &= \int_0^\infty sx^{s-1}e^{-x}\,dx\\ &= s\Gamma(s)\end{aligned}$$

Regarding now the case $s \in \mathbb{N}$, for the initial value $s = 1$ we have:

$$\Gamma(1) = \int_0^\infty e^{-x}dx = 1$$

Thus, for $s \in \mathbb{N}$ we have indeed $\Gamma(s) = (s-1)!$, as claimed. □

Many interesting things can be said about the gamma function, notably with:

THEOREM 14.25. *The gamma function is given at half-integers by*

$$\Gamma(n) = (n-1)! \quad , \quad \Gamma\left(n + \frac{1}{2}\right) = \frac{(2n)!!}{2^n}\sqrt{\pi}$$

and we have the following formula for it, with $c = \sqrt{2}, \sqrt{\pi}$ *for* N *even, odd:*

$$\Gamma\left(\frac{N}{2}\right) = \frac{(N-1)!!}{2^{(N-1)/2}}\,c$$

Moreover, the volume of the unit sphere in $\mathbb{R}^N$*, which is given by*

$$V = \left(\frac{\pi}{2}\right)^{[N/2]}\frac{2^N}{(N+1)!!}$$

can be expressed in terms of these values of the gamma function.

PROOF. There are several things going on here, the idea being as follows:

(1) We already know, from Theorem 14.24, that for $n \in \mathbb{N}$ we have:

$$\Gamma(n) = (n-1)!$$

(2) Regarding now the half-integers, we first have the following computation:

$$\begin{aligned}\Gamma\left(\frac{1}{2}\right) &= \int_0^\infty x^{-1/2}e^{-x}\,dx\\ &= \int_0^\infty y^{-1}e^{-y^2}2y\,dy\\ &= \sqrt{\pi}\end{aligned}$$

Thus, by using $\Gamma(s+1)=s\Gamma(s)$, we obtain the formula in the statement:

$$\Gamma\left(n+\frac{1}{2}\right)=\frac{1.3.5\ldots(2n-1)}{2^n}\sqrt{\pi}=\frac{(2n)!!}{2^n}\sqrt{\pi}$$

(3) Regarding the unification, this comes by writing the formula in (1) as follows:

$$\Gamma(n)=(n-1)!=\frac{(2n-1)!!}{2^{n-1}}=\frac{(2n-1)!!}{2^{n-1/2}}\sqrt{2}$$

(4) Finally, the assertion regarding the spheres is something self-explanatory. □

Getting back now to zeta, we can formulate a key result about it, as follows:

THEOREM 14.26. *We have the following formula,*

$$\zeta(s)=\frac{1}{\Gamma(s)}\int_0^\infty\frac{x^{s-1}}{e^x-1}\,dx$$

valid for any $s\in\mathbb{C}$ with $Re(s)>1$.

PROOF. We have indeed the following computation:

$$\begin{aligned}\int_0^\infty\frac{x^{s-1}}{e^x-1}\,dx &= \int_0^\infty\frac{x^{s-1}}{e^x}\cdot\frac{1}{1-e^{-x}}\,dx\\ &= \int_0^\infty x^{s-1}(e^{-x}+e^{-2x}+e^{-3x}+\ldots)\\ &= \sum_{n=1}^\infty\int_0^\infty x^{s-1}e^{-nx}\,dx\\ &= \sum_{n=1}^\infty\int_0^\infty\left(\frac{y}{n}\right)^{s-1}e^{-y}\,\frac{dy}{n}\\ &= \sum_{n=1}^\infty\frac{1}{n^s}\int_0^\infty y^{s-1}e^{-y}\,dy\\ &= \zeta(s)\Gamma(s)\end{aligned}$$

Thus, we are led to the formula in the statement. □

At a more advanced level now, following Euler and others, we have:

THEOREM 14.27. *The values of zeta at even integers are given by*

$$\zeta(2k) = (-1)^{k+1}\frac{(2\pi)^{2k}B_{2k}}{2\cdot(2k)!}$$

with B_n being the Bernoulli numbers, which in practice gives the formulae

$$\zeta(2) = \frac{\pi^2}{6} \quad , \quad \zeta(4) = \frac{\pi^4}{90} \quad , \quad \zeta(6) = \frac{\pi^6}{945} \quad , \quad \zeta(8) = \frac{\pi^8}{9450} \quad , \quad \ldots$$

generalizing the formula $\zeta(2) = \pi^2/6$ of Euler, solving the Basel problem.

PROOF. This is something quite tricky, the idea being as follows:

(1) To start with, at $s = 2$ we have the following heuristics, following Euler, obtained by factorizing the function $\sin x$, having zeroes at $\mathbb{Z}/2$, a bit like a polynomial:

$$\begin{aligned}
\frac{\sin x}{x} &= 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \ldots \\
&= \left(1 - \frac{x}{\pi}\right)\left(1 + \frac{x}{\pi}\right)\left(1 - \frac{x}{2\pi}\right)\left(1 + \frac{x}{2\pi}\right)\ldots \\
&= \left(1 - \frac{x^2}{\pi^2}\right)\left(1 - \frac{x^2}{4\pi^2}\right)\left(1 - \frac{x^2}{9\pi^2}\right)\ldots \\
&= 1 - \frac{1}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\,x^2 + \ldots
\end{aligned}$$

Thus we get $\zeta(2) = \pi^2/6$, as claimed. Of course this is far from being rigorous, but there are several ways of fixing this. We will be back to this in chapter 15.

(2) Next, at $s = 2k$, we have the following computation, using Theorem 14.26:

$$\begin{aligned}
\zeta(2k) &= \frac{1}{\Gamma(2k)}\int_0^{\infty}\frac{x^{2k-1}}{e^x - 1}\,dx \\
&= \frac{1}{(2k-1)!}\int_0^{\infty}\frac{(2\pi t)^{2k-1}}{e^{2\pi t} - 1}\,2\pi dt \\
&= \frac{(2\pi)^{2k}}{(2k-1)!}\int_0^{\infty}\frac{t^{2k-1}}{e^{2\pi t} - 1}\,dt
\end{aligned}$$

(3) Now the point is that the integrals on the right can be computed, and with all this being related to the Euler formula from (1), and we are led to the following formula:

$$\int_0^{\infty}\frac{t^{2k-1}}{e^{2\pi t} - 1}\,dt = (-1)^{k+1}\frac{B_{2k}}{4k}$$

Thus, we are led to the various conclusions in the statement. So, this was for the idea, and we will be back to all this in chapter 15, with fixes and details. □

As yet another key result about zeta, this extends to $Re(s) > 0$, as follows:

THEOREM 14.28. *We have the following formula,*

$$\zeta(s) = \frac{1}{1-2^{1-s}} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s}$$

which can stand as definition for ζ, in the strip $0 < Re(s) < 1$.

PROOF. To start with, we can define the Dirichlet eta function η as being the signed version of the Riemann zeta function ζ, according to the following formula:

$$\eta(s) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s}$$

We must now connect ζ and η, at $Re(s) > 1$, and this can be done as follows:

$$\zeta(s) + \eta(s) = 2 \sum_{k=1}^{\infty} \frac{1}{(2k)^s} = 2^{1-s} \sum_{k=1}^{\infty} \frac{1}{k^s} = 2^{1-s}\zeta(s)$$

Thus, we have the formula in the statement. Now since η is defined at $Re(s) > 0$, we can say that our formula can stand as a definition for ζ, at $0 < Re(s) < 1$, as stated. □

And with this, we can now talk about the Riemann hypothesis, as follows:

CONJECTURE 14.29 (Riemann hypothesis). *The zeroes of zeta in the critical strip*

$$0 < Re(s) < 1$$

can only appear on the critical line, $Re(s) = 1/2$.

And good question this is. Exercise for you, to learn more about all this.

14e. Exercises

This was a quite tricky chapter, and as exercises on this, we have:

EXERCISE 14.30. *Learn more about 1D waves, d'Alembert, and related topics.*

EXERCISE 14.31. *Experiment some more with the Gauss integral.*

EXERCISE 14.32. *Learn more about the heat equation and kernel.*

EXERCISE 14.33. *Experiment some more with the Fresnel integrals.*

EXERCISE 14.34. *Learn more about the Stirling formula, and the error term.*

EXERCISE 14.35. *Learn also the Euler-Maclaurin formula, and its applications.*

EXERCISE 14.36. *Learn more about the gamma function, and its properties.*

EXERCISE 14.37. *Learn more about the zeta function, and the Riemann hypothesis.*

As bonus exercise, with what we learned here, you're good for physics and arithmetic.

CHAPTER 15

Function spaces

15a. Function spaces

We learned a lot of interesting things in the previous chapter, but despite our study there being certainly sharp and professional, we still have many questions left, in relation with the differential equations, and with the Basel problem and zeta function too.

So, question for the two of us, what is the way out? In answer, geometry:

PRINCIPLE 15.1. *Difficult analysis questions can be solved via geometry:*

(1) $||f|| = \sqrt{\int f(x)^2 dx}$ *can be thought of as being the length of* f.
(2) *And* $< f, g >= \int f(x)g(x)dx$, *as being the scalar product of* f, g.

So, let us get into this, function spaces and their geometry. It is technically convenient to aim for the complex functions, $f : X \to \mathbb{C}$, and with this in mind, let us formulate:

DEFINITION 15.2. *A Banach space is a complex vector space* V, *with a map*

$$||.|| : V \to \mathbb{R}_+$$

called norm, subject to the following conditions:

(1) $||x|| = 0$ *implies* $x = 0$.
(2) $||\lambda x|| = |\lambda| \cdot ||x||$, *for any* $x \in V$, *and* $\lambda \in \mathbb{C}$.
(3) $||x + y|| \leq ||x|| + ||y||$, *for any* $x, y \in V$.
(4) V *is complete with respect to the distance* $d(x, y) = ||x - y||$.

This is of course something quite compact, but we already have some familiarity with such things, from chapters 6 and 10. In fact, what we did in chapters 6 and 10, extended to the complex setting, and along with a bit more, can be summarized as follows:

THEOREM 15.3. *Given a set* X, *finite or not, and* $p \in [1, \infty]$, *the space*

$$l^p(X) = \left\{ f : X \to \mathbb{C} \,\middle|\, ||f||_p < \infty \right\} \quad , \quad ||f||_p = \left(\sum_{x \in X} |f(x)|^p \right)^{1/p}$$

is a Banach space. For $X = \{1, \ldots, N\}$ *and* $p = 2$, *this is the usual* $\mathbb{C}^N$.

PROOF. We basically know all this, from chapters 6 and 10, as follows:

(1) Let us first discuss the case where X is finite, $X = \{1, \ldots, N\}$. Here the functions $f : X \to \mathbb{C}$ can be identified with the vectors $v \in \mathbb{C}^N$, and the norm is:

$$||v||_p = \left(\sum_i |v_i|^p \right)^{1/p}$$

But, is this a norm? In order to discuss this, let us first examine the exponents $p = 1, 2, \infty$. Here the corresponding norms are as follows, with the one on the right coming by definition, or as the limit of the p-norm above, with $p \to \infty$:

$$||v||_1 = \sum_i |v_i| \quad , \quad ||v||_2 = \sqrt{\sum_i |v_i|^2} \quad , \quad ||v||_\infty = \max_i |v_i|$$

And these are norms indeed, with this being clear for $||.||_1$, being well-known for $||.||_2$, this being the usual norm on $\mathbb{C}^N$, and with the norm property formally coming from Cauchy-Schwarz, and finally with the norm property of $||.||_\infty$ being clear too.

(2) In the general case now, where $X = \{1, \ldots, N\}$ is still finite, but $p \in [1, \infty]$ is arbitrary, the norm property follows from our inequality know-how from chapter 10. Indeed, as explained there, Jensen for $\log x$ gives the Young inequality, namely:

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \quad , \quad \frac{1}{p} + \frac{1}{q} = 1$$

But from this Young inequality we obtain the Hölder inequality, namely:

$$\sum_i |u_i v_i| \leq \left(\sum_i |u_i|^p \right)^{1/p} \left(\sum_i |v_i|^q \right)^{1/q}$$

And then, from this Hölder inequality we get the Minkowski inequality, namely:

$$\left(\sum_i |u_i + v_i|^p \right)^{1/p} \leq \left(\sum_i |u_i|^p \right)^{1/p} + \left(\sum_i |v_i|^p \right)^{1/p}$$

And this is exactly what we need, proving that $||.||_p$ is a norm on $l^p(X) = \mathbb{C}^N$.

(3) Next, let us discuss the case where both X and $p \in [1, \infty]$ are arbitrary. Here the functions $f : X \to \mathbb{C}$ can be identified with the sequences $(v_x)_{x \in X}$, the norm is given by the same formula as in (1), and all inequalities extend well. However, there is a subtlety here, because we have to get rid of the sequences of infinite norm, by setting:

$$l^p(X) = \left\{ (v_x)_{x \in X} \;\middle|\; ||v||_p < \infty \right\}$$

Thus, we have our normed space, and the fact that this space is indeed complete, and so is a Banach space, is something quite clear, say easy exercise for you. □

Getting now to what we wanted to do, spaces of functions, we have here:

THEOREM 15.4. *Given a measured space X, and $p \in [1, \infty]$, the following space, with the convention that functions are identified up to equality almost everywhere,*

$$L^p(X) = \left\{ f : X \to \mathbb{C} \,\middle|\, \int_X |f(x)|^p dx < \infty \right\}$$

is a vector space, and the following quantity is a norm on it,

$$||f||_p = \left(\int_X |f(x)|^p \right)^{1/p}$$

making it a Banach space. In the discrete case, that is, when X is set, endowed with its counting measure, we obtain in this way the previous spaces $l^p(X)$.

PROOF. This is a straightforward generalization of Theorem 15.3, but since this will be our final saying on the subject, here are the technical details:

(1) Let us first prove Hölder, stating that with $1/p + 1/q = 1$ we have:

$$\int_X |fg| \leq \left(\int_X |f|^p \right)^{1/p} \left(\int_X |g|^q \right)^{1/q}$$

By linearity we can assume that f, g are normalized, in the following way:

$$\int_X |f|^p = \int_X |g|^q = 1$$

But with this assumption made, we can get Hölder from Young, as follows:

$$\int_X |fg| \leq \int_X \frac{|f|^p}{p} + \int_X \frac{|g|^q}{q} = \frac{1}{p} + \frac{1}{q} = 1$$

(2) Next, let us prove Minkowski. By using the Hölder inequality, we have:

$$\begin{aligned}
\int_X |f+g|^p &= \int_X |f+g| \cdot |f+g|^{p-1} \\
&\leq \int_X |f| \cdot |f+g|^{p-1} + \int_X |g| \cdot |f+g|^{p-1} \\
&\leq \left(\int_X |f|^p \right)^{1/p} \left(\int_X |f+g|^{(p-1)q} \right)^{1/q} \\
&+ \left(\int_X |g|^p \right)^{1/p} \left(\int_X |f+g|^{(p-1)q} \right)^{1/q} \\
&= \left[\left(\int |f|^p \right)^{1/p} + \left(\int_X |g|^p \right)^{1/p} \right] \left(\int_X |f+g|^p \right)^{1-1/p}
\end{aligned}$$

Thus, we are led in this way to the Minkowski inequality, namely:

$$\left(\int_X |f+g|^p\right)^{1/p} \leq \left(\int_X |f|^p\right)^{1/p} + \left(\int_X |g|^p\right)^{1/p}$$

(3) Before getting further, let us mention that in the above it was understood that the conjugate exponents p, q were finite, $p, q \in (1, \infty)$. However, for infinite exponents we have similar results, which are trivial this time, with the following convention:

$$||f||_\infty = \lim_{p\to\infty} \left(\int_X |f|^p\right)^{1/p} = \text{ess sup}\,|f|$$

(4) Summarizing, Minkowski inequality proved at any $p \in [1, \infty]$, which basically tells us that $||.||_p$ is a norm. However, as before in Theorem 15.3, there is a subtlety here, because we must get rid of the functions having infinite norm, by imposing the condition $||f||_p < \infty$ in the statement. And there is a second subtlety too, because in order to have the first norm axiom working, namely $||f||_p = 0 \implies f = 0$, we must proceed as in the statement, by identifying the functions which are equal almost everywhere.

(5) Thus, we have our normed space, modulo some thinking about the above two subtleties, that we will leave as an exercise. As for the fact that our space is indeed complete, so is a Banach space, we will leave this again as an easy exercise for you. □

Very nice all this, and getting back to Principle 15.1, we have now a good understanding of (1) there, corresponding to the case $p = 2$, plus generalization to the case of the arbitrary exponents $p \in [1, \infty]$. So, time I guess to do some geometry, based on this.

Well, in theory at least. In practice, geometry rather needs scalar products as in Principle 15.1 (2), and more on this in a moment. In the meantime, let us record a few interesting things that can be done in the general Banach space context, as follows:

THEOREM 15.5. *Given a Banach space V, the following happen:*

(1) *The continuous linear maps $f : V \to \mathbb{C}$ form a Banach space V^*.*
(2) *We have an isometric embedding $V \subset V^{**}$, given by $v \to (f \to f(v))$.*
(3) *For $V = L^p(X)$ we have $V^* = L^q(X)$, with $1/p + 1/q = 1$, and $V = V^{**}$.*

PROOF. This is a mixture of trivial and non-trivial facts, as follows:

(1) This is something very standard, with the norm on V^* being given by:

$$||f|| = \sup_{||v||=1} |f(v)|$$

To be more precise, it is easy to see that a linear map $f : V \to \mathbb{C}$ is continuous precisely when $||f|| < \infty$, and then that $||.||$ is indeed a norm on V^*, and finally that V^* is indeed complete. We will leave all this, as an easy exercise for you.

(2) We certainly have a continuous linear map $V \to V^{**}$ given by $v \to (f \to f(v))$, but when it comes to prove that this map is indeed injective, and even isometric, surprise, this is not obvious. But this follows from the Hahn-Banach theorem, stating that given a nonzero vector $v \in V$, we can always find $f \in V^*$ with $f(v) \neq 0$. And, exercise for you to learn more about all this, have a look here at any functional analysis book.

(3) This result is more or less what the Hölder inequality says, because the coupling $(f, g) \to \int fg$ from Hölder can be interpreted as corresponding to a continuous linear form on V, and vice versa. As before, exercise for you to clarify all this, or look it up.

(4) Finally, as an interesting bonus exercise, related to all this, think as well as examples of Banach spaces V which are not reflexive, $V \neq V^{**}$. Enjoy. □

So long for Principle 15.1 (1), and Banach space theory. At our present level in functional analysis, which is beginner, all this was rather a failure, because we were unable to derive anything concrete, out of this. Nevermind. Please be sure that the above was good learning, and that one day, you might need Theorems 15.4 and 15.5.

Upgrading now to scalar products, according to Principle 15.1 (2), let us start with the following key definition, meant to replace Definition 15.2:

Definition 15.6. *A Hilbert space is a complex vector space H, with a map*

$$< , >: H \times H \to \mathbb{C}$$

called scalar product, subject to the following conditions:

(1) $< x, y >$ *is linear in* x*, and antilinear in* y*.*
(2) $\overline{< x, y >} =< y, x >$*, for any* x, y*.*
(3) $< x, x >> 0$*, for any* $x \neq 0$*.*
(4) H *is complete with respect to the norm* $||x|| = \sqrt{< x, x >}$*.*

Here, as before in the context of Definition 15.2, we are going quite quick, because we are already a bit familiar with such things, from chapters 6 and 10. We will see examples in a moment, but before that, some explanations in regards with (4). Given two vectors $x, y \in H$, consider the following degree 2 function of $t \in \mathbb{R}$, depending on $w \in \mathbb{T}$:

$$f(t) = ||wx + ty||^2 = ||x||^2 + 2tRe(w < x, y >) + t^2||y||^2$$

This function being positive, its discriminant must be negative, which gives:

$$| < x, y > | \leq ||x|| \cdot ||y||$$

But this gives in turn $||x + y|| \leq ||x| + ||y||$, and so we have indeed a norm, as said in (4), so we can talk about the completness of H with respect to this norm. Observe the relation with Definition 15.2, the Hilbert spaces being certain Banach spaces.

At the level of examples, our result, meant to replace Theorem 15.4, is as follows:

THEOREM 15.7. *Given a measured space X, the following space, with the convention that functions are identified up to equality almost everywhere,*

$$L^2(X) = \left\{ f : X \to \mathbb{C} \,\middle|\, \int_X |f(x)|^2 dx < \infty \right\}$$

is a vector space, and the following is a scalar product on it, making it a Hilbert space:

$$< f, g >= \int_X f(x)\overline{g(x)}\, dx$$

In the discrete case, where X is set, with its counting measure, we obtain the previous spaces $l^2(X)$, which generalize $\mathbb{C}^N$ with its usual scalar product, $< u, v >= \sum_i u_i \bar{v}_i$.

PROOF. This is indeed something self-explanatory, based on Theorem 15.4. Observe that we do not need in fact Theorem 15.5, for talking about all this, because we already have Cauchy-Schwarz, as explained above, so everything is in fact a triviality. □

Getting now to the linear forms, as a replacement for Theorem 15.5, we have:

THEOREM 15.8. *Given a Hilbert space H, the following happen:*

(1) *The continuous linear maps $f : H \to \mathbb{C}$ are $f(x) =< x, y >$, with $y \in H$.*
(2) *Thus we have $H^* = \bar{H}$, and as a consequence, we have $H^{**} = H$.*

PROOF. Again, this is something self-explanatory, based this time on Theorem 15.5, and with the remark that Theorem 15.5 is in fact not needed, because $f(x) =< x, y >$ is elementary to establish, say exercise for you, and everything else comes from this. □

Summarizing, the Hilbert spaces appear to be something far more powerful than the Banach spaces, with all our Banach space theory corresponding to mere trivialities, in the Hilbert space setting. Getting now to the real thing, geometry, we first have:

PROPOSITION 15.9. *Given a Hilbert space H, the following happen:*

(1) *Norm formula: $||x|| = \sqrt{< x, x >}$.*
(2) *Cauchy-Schwarz: $| < x, y > | \leq ||x|| \cdot ||y||$.*
(3) *Parallelogram identity: $||x + y||^2 + ||x - y||^2 = 2(||x||^2 + ||y||^2)$.*
(4) *Polarization: $4 < x, y >= ||x + y||^2 - ||x - y||^2 + i||x + iy||^2 - i||x - iy||^2$.*

PROOF. Here (1) was part of the axioms, (2) is something that we already know, which was explained above, and (3) comes from the following computation:

$$\begin{aligned} & ||x + y||^2 + ||x - y||^2 \\ = \ & < x + y, x + y > + < x - y, x - y > \\ = \ & ||x||^2 + ||y||^2 + < x, y > + < y, x > + ||x||^2 + ||y||^2 - < x, y > - < y, x > \\ = \ & 2(||x||^2 + ||y||^2) \end{aligned}$$

As for (4), this comes from a similar computation, as follows:

$$\begin{aligned}
& ||x+y||^2-||x-y||^2+i||x+iy||^2-i||x-iy||^2\\
= \;& ||x||^2+||y||^2-||x||^2-||y||^2+i||x||^2+i||y||^2-i||x||^2-i||y||^2\\
& +2Re(<x,y>)+2Re(<x,y>)+2iIm(<x,y>)+2iIm(<x,y>)\\
= \;& 4<x,y>
\end{aligned}$$

Finally, observe that (1,4) show that the norm is uniquely determined by the scalar product, and vice versa. This is something that we will often use, in what follows. □

As a piece of true geometry now, in the Hilbert space setting, we have:

THEOREM 15.10. *Given a Hilbert space H, we can talk about:*

(1) *Bounded operators $T:H\to H$, for which $||T||=\sup_{||x||=1}||Tx||$ is bounded.*
(2) *Adjoints of such bounded operators, given by $<T^*x,y>=<x,Ty>$.*
(3) *Unitaries $U:H\to H$, which must satisfy the condition $U^*=U^{-1}$.*
(4) *Projections $P:H\to H$, which must satisfy the condition $P^2=P^*=P$.*

PROOF. This is obviously something quite advanced, generalizing some key facts from linear algebra, which linear algebra is scheduled for chapter 16 in this book. So, as before with Theorem 15.5, take this as something which is good to know, making the point with geometry, and whose details you can get into later, when really needing such things. □

Moving on, as a key feature of the Hilbert spaces, we have bases:

THEOREM 15.11. *Any Hilbert space H has an orthonormal basis $\{e_i\}_{i\in I}$, which is by definition a set of vectors whose span is dense in H, and which satisfy*

$$<e_i,e_j>=\delta_{ij}$$

with δ being a Kronecker symbol. The cardinality $|I|$ of the index set, which can be finite, countable, or uncountable, depends only on H, and is called dimension of H. We have

$$H\simeq l^2(I)$$

in the obvious way, mapping $\sum\lambda_ie_i\to(\lambda_i)$. The Hilbert spaces with $\dim H=|I|$ being countable, such as $l^2(\mathbb{N})$, are all isomorphic, and are called separable.

PROOF. We have many assertions here, the idea being as follows:

(1) In finite dimensions an orthonormal basis $\{e_i\}_{i\in I}$ can be constructed by starting with any vector space basis $\{f_i\}_{i\in I}$, and using the well-known Gram-Schmidt procedure. As for the other assertions, these are all clear, from basic linear algebra.

(2) In general, the same method works, namely Gram-Schmidt, with a subtlety coming from the fact that the basis $\{e_i\}_{i\in I}$ will not span in general the whole H, but just a dense subspace of it, as it is in fact obvious by looking at the standard basis of $l^2(\mathbb{N})$.

(3) And there is a second subtlety as well, coming from the fact that the recurrence procedure needed for Gram-Schmidt must be replaced by some sort of "transfinite recurrence", using standard tools from logic, and more specifically the Zorn lemma.

(4) Finally, everything at the end, regarding our notion of separability for the Hilbert spaces, is clear from definitions, and from our various results above. □

As a continuation of this, we have the following result, dealing with separability:

THEOREM 15.12. *The following happen, in relation with separability:*

(1) *The Hilbert space $H = L^2[-1, 1]$ is separable, with orthonormal basis coming by applying Gram-Schmidt to the basis $\{x^k\}_{k\in\mathbb{N}}$, coming from Weierstrass.*
(2) *In fact, $H = L^2(X)$ with $X \subset \mathbb{R}$ and $d\mu(x) = f(x)dx$ is separable, with standard basis $\{P_k\}_{k\in\mathbb{N}}$ formed by the orthogonal polynomials with respect to μ.*
(3) *More generally, given a separable abstract measured space X, the associated Hilbert space of square-summable functions $H = L^2(X)$ is separable.*

PROOF. Many things can be said here, the idea being as follows:

(1) The fact that $H = L^2[-1, 1]$ is separable is clear indeed from the Weierstrass density theorem, which provides us with the algebraic basis $f_k = x^k$, which can be orthogonalized by using the Gram-Schmidt procedure, as explained in Theorem 15.11.

(2) This is a straightforward generalization of (1), with the polynomials $\{P_k\}_{k\in\mathbb{N}}$ coming from the Weierstrass basis $\{x^k\}_{k\in\mathbb{N}}$, via Gram-Schmidt with respect to the measure $d\mu(x) = f(x)dx$, being called the orthogonal polynomials with respect to μ.

(3) As for the last assertion, regarding the general spaces of type $H = L^2(X)$, which generalizes what we have in (1,2), this comes as a consequence of general measure theory, and we will leave learning the details here as a long, instructive exercise. □

As a concrete illustration for all this, making the link with the Chebycheff polynomials that we studied back in chapter 7, we have the following key result:

THEOREM 15.13. *The orthogonal polynomials for $L^2[-1, 1]$, with measure*

$$d\mu(x) = (1 - x)^a(1 + x)^b dx$$

called Jacobi polynomials, satisfy a degree 2 equation, namely

$$(1 - x^2)J_k''(x) + (b - a - (a + b + 2)x)J_k'(x) + k(k + a + b + 1)J_k(x) = 0$$

and are given by the following formula, featuring derivatives:

$$J_k(x) = \frac{(-1)^k}{2^k k!}(1 - x)^{-a}(1 + x)^{-b}\frac{d^k}{dx^k}\left[(1 - x)^a(1 + x)^b(1 - x^2)^k\right]$$

At $a = b = 0$ we recover the Legendre polynomials from physics, and at $a = b = \pm\frac{1}{2}$ we recover the Chebycheff polynomials of the first and second kind.

PROOF. This is obviously something quite advanced, the idea being as follows:

(1) Generally speaking, the statement appears as a generalization of the well-known result for Legendre polynomials, which corresponds to the particular case $a = b = 0$, and the proof is quite similar. We will leave learning more about all this as an exercise.

(2) Regarding now the main particular cases of the Jacobi polynomials, these are the Gegenbauer polynomials, appearing at $a = b$. However, there is not that much of a simplification when passing from general parameters a, b to equal parameters, $a = b$, so in practice, the main particular cases are those indicated in the statement, namely:

– The Legendre polynomials, which naturally appear in questions from quantum mechanics, coming at the simplest values of the parameters, namely $a = b = 0$.

– The Chebycheff polynomials of the first kind P_k, which are given by the formula $P_k(\cos t) = \cos(kt)$ from trigonometry, appearing at $a = b = -\frac{1}{2}$.

– The Chebycheff polynomials of the second kind Q_k, which are given by the formula $Q_k(\cos t)\sin t = \sin((k+1)t)$, appearing at $a = b = \frac{1}{2}$. □

So long for Hilbert spaces, abstract or concrete, and their basic theory. The above material was quite advanced, yes I know, but functional analysis is a delicate business, requiring a lot of learning, and all the above was an introduction to this. So, take that as it came, and come back regularly to this, later, until fully understanding all this.

15b. Fourier, Parseval

Getting now to more concrete things, which will eventually lead to applications, we have the following key fact, which is the starting point for Fourier analysis:

CLAIM 15.14. *The space of square-summable functions on the unit circle,*

$$L^2(\mathbb{T}) = \left\{ f : \mathbb{T} \to \mathbb{C} \;\middle|\; \int_{\mathbb{T}} |f(z)|^2 dz < \infty \right\}$$

with respect to the usual mass 1 *measure, has* $\{z^n\}_{n\in\mathbb{Z}}$ *as orthonormal basis.*

As a first observation, this reminds a bit what we said above regarding the Weierstrass basis $\{x^n\}_{n\in\mathbb{N}}$ for the various spaces of functions $f : X \to \mathbb{C}$, with $X \subset \mathbb{R}$, and the related notion of orthogonal polynomials. That is, what Claim 15.14 says is that we have some sort of Weierstrass approximation theorem on the circle, but with respect to the 2-norm, and with the standard basis which is used, $\{z^n\}_{n\in\mathbb{Z}}$, being orthonormal for free.

In practice now, Claim 15.14 remains something which, while being certainly simple, beautiful and understandable, is a bit compact and abstract. For most purposes, it is better to replace it by the following equivalent formulation, in terms of real functions:

THEOREM 15.15. *The space of 2π-periodic square-summable functions on $\mathbb{R}$,*

$$L^2(\mathbb{R})_{per} = \left\{ f : \mathbb{R} \to \mathbb{C} \;\middle|\; f(t) = f(t+2\pi),\ \int_{-\pi}^{\pi} |f(t)|^2 dt < \infty \right\}$$

has $\{e^{int}\}_{n\in\mathbb{Z}}$ as orthonormal basis, with respect to the normalized mass 1 measure.

PROOF. This is something quite tricky, which came as a big surprise at the time of its discovery by Fourier, the idea with all this being as follows:

(1) As a first observation, which is philosophical, as already mentioned above, this is a real function reformulation of Claim 15.14, and for various reasons, in relation with both the proof and the future applications, this is the version that we will prefer.

(2) Next, still talking generalities, according to what the statement says at the end, the scalar product on $L^2(\mathbb{R})_{per}$ is by definition given by the following formula:

$$< f, g >= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)\overline{g(t)}\, dt$$

As for the corresponding norm on $L^2(\mathbb{R})_{per}$, this is given by the following formula:

$$||f|| = \frac{1}{2\pi} \int_{-\pi}^{\pi} |f(t)|^2 dt$$

Observe that we could have used $[0, 2\pi]$ for integrating, or more generally, any interval $I \subset \mathbb{R}$ having length 2π. For certain technical reasons, we prefer to use $I = [-\pi, \pi]$.

(3) Getting now to the proof, as a first basic computation that we can do, coming from the 2π-periodicity of $e^{it} = \cos t + i \sin t$, we have the following formula:

$$< e^{int}, e^{imt} >= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i(n-m)t} dt = \delta_{nm}$$

Thus $\{e^{int}\}_{n\in\mathbb{Z}}$ is as orthonormal system, and our theorem says that this is a basis.

(4) But, how to prove this? The first thought goes to the Weierstrass approximation theorem, that we know since chapter 6, but after thinking a bit, that is not of much use. Indeed, Weierstrass approximates the functions $f : [-\pi, \pi] \to \mathbb{C}$ by polynomials $\sum_n c_n t^n$, and what Fourier says is that the same functions $f : [-\pi, \pi] \to \mathbb{C}$ can be approximated by trigonometric polynomials $\sum_n c_n e^{int}$, which is a different story.

(5) In fact, the difference between Fourier and Weierstrass is even more visible with Theorem 15.13 in mind. Indeed, up to a rescaling, Theorem 15.13 tells us that what we get from Weierstrass is the basis of $L^2(\mathbb{R})_{per}$ formed by Legendre polynomials $\{L_k\}_{k\in\mathbb{N}}$, and the Fourier basis $\{e^{int}\}_{n\in\mathbb{Z}}$ is obviously something of different nature.

(6) Time now for the proof? For reasons that will become clear in a moment, consider the following trigonometric polynomials, with $c_k \in \mathbb{R}$ being chosen for having mass 1:

$$E_k(t) = c_k \left(\frac{1+\cos t}{2}\right)^k$$

Observe that c_k can be computed explicitly, by using $1+\cos t = 2\cos^2(t/2)$ and the Wallis formulae from chapter 14, but in what follows, we will not need this. Our claim, which is the reason for introducing these functions E_k, is that for any $\delta > 0$ we have:

$$\lim_{k\to\infty} \sup_{\delta<|t|<\pi} E_k(t) = 0$$

In other words, our claim is that $E_k \to 0$ uniformly on any $[-\pi,-\delta] \cup [\delta,\pi]$.

(7) So, let us prove this claim. As mentioned, c_k can be computed explicitly, but in what follows we will only need the following elementary estimate:

$$\begin{aligned}
1 &= \frac{c_k}{\pi}\int_0^\pi \left(\frac{1+\cos t}{2}\right)^k dt\\
&> \frac{c_k}{\pi}\int_0^\pi \left(\frac{1+\cos t}{2}\right)^k \sin t dt\\
&= \frac{c_k}{\pi}\left[-\frac{2}{k+1}\left(\frac{1+\cos t}{2}\right)^{k+1}\right]_0^\pi\\
&= \frac{2c_k}{\pi(k+1)}
\end{aligned}$$

Now since E_k is decreasing on $[0,\pi]$, we obtain from this, for $\delta < |t| < \pi$:

$$E_k(t) < E_k(\delta) < \frac{\pi(k+1)}{2}\left(\frac{1+\cos\delta}{2}\right)^k$$

But this proves our claim, because for $\delta > 0$ we have $(1+\cos\delta)/2 < 1$, as needed.

(8) Getting now to what we wanted to do, we must prove that $\{e^{int}\}_{n\in\mathbb{Z}}$ spans a dense subset of $L^2(\mathbb{R})_{per}$. Since $C(\mathbb{R})_{per} \subset L^2(\mathbb{R})_{per}$ is dense, it is enough to prove that any $f \in C(\mathbb{R})_{per}$ can be approximated by trigonometric polynomials $\sum_n c_n e^{int}$. Moreover, since $||.||_2 \leq ||.||_\infty$, it is enough to prove our approximation with respect to $||.||_\infty$.

(9) All in all, it remains to prove that given a function $f \in C(\mathbb{R})_{per}$ and a number $\varepsilon > 0$, we can always come with a trigonometric polynomial $\sum_n c_n e^{int}$, such that:

$$\left|f(t) - \sum_n c_n e^{int}\right| < \varepsilon \quad , \quad \forall t \in [-\pi,\pi]$$

(10) But for this, we can use the polynomials E_k from (6). Let us set indeed:

$$Q_k(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t-s)E_k(s)\,ds$$

As a first observation, with the change of variables $s \to t-s$, we have the following alternative formula, which shows that $Q_k(t)$ are indeed trigonometric polynomials:

$$Q_k(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(s)E_k(t-s)\,ds$$

(11) Now given $\varepsilon > 0$, let us prove that the estimate in (9) holds indeed, with the trigonometric polynomial there being $Q_k(t)$, for $k >> 0$ large enough. For this purpose, we use the uniform continuity of f, which tells us that we can find $\delta > 0$ such that:

$$|s-t| < \delta \implies |f(s) - f(t)| < \varepsilon$$

Indeed, by using this, we have the following estimate, for the error in (9):

$$\begin{aligned} |Q_k(t) - f(t)| &= \frac{1}{2\pi}\left|\int_{-\pi}^{\pi}(f(t-s) - f(t))E_k(s)\,ds\right| \\ &\leq \frac{1}{2\pi}\int_{-\pi}^{\pi}|f(t-s) - f(t)|E_k(s)\,ds \end{aligned}$$

(12) Now let us split the last integral into three parts, according to:

$$[-\pi,\pi] = [-\pi,-\delta] \cup [-\delta,\delta] \cup [\delta,\pi]$$

On the middle part the integrand is $< \varepsilon$, so the middle integral is $< \varepsilon$. As for the other two integrals, on $[-\pi,-\delta] \cup [\delta,\pi]$, we can use here (6), telling us that $E_k(t) \to 0$ uniformly, on that domain. Indeed, with $k >> 0$ big enough the other two integrals are $< \varepsilon$ too, so we have obtained (9) as desired, with $\varepsilon \to 3\varepsilon$, which finishes the proof. □

Still with me I hope, after all these computations. In practice now, a bit as Theorem 15.15 itself was a useful version of Claim 15.14, there are several other useful versions of Theorem 15.15, which can be of great use, in practice. We notably have here:

THEOREM 15.16. *We have an isomorphism $L^2(\mathbb{R})_{per} \simeq l^2(\mathbb{Z})$, as follows:*

(1) *Associated to $f \in L^2(\mathbb{R})_{per}$ are its Fourier coefficients, given by:*

$$\widehat{f}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)e^{int}dt$$

(2) *Associated to $g \in l^2(\mathbb{Z})$ is the series $S_g(t) = \sum_{n\in\mathbb{Z}} g(n)e^{-int}$.*

PROOF. This is something self-explanatory, based on the general orthonormal basis theory from Theorem 15.11, with the Fourier coefficients of $f \in L^2(\mathbb{R})_{per}$ being its coefficients $\widehat{f}(n) =< f, e^{-int} >$ with respect to the basis $\{e^{int}\}$ from Theorem 15.15. □

In practice, it is possible to talk as well about real Fourier coefficients, given by $\widehat{f}(n) = a_n + ib_n$, and with the formula of these, for $f : \mathbb{R} \to \mathbb{R}$, being as follows:

$$a_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\cos(nt)dt \quad , \quad b_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\sin(nt)dt$$

And exercise of course for you, to learn more about all this. Finally, as yet another reformulation, or rather consequence of what we have, we have the Parseval formula:

THEOREM 15.17. *The Fourier coefficients $\widehat{f}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)e^{int}dt$ satisfy*

$$\sum_{n\in\mathbb{Z}} \widehat{f}(n)\overline{\widehat{g}(n)} = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\overline{g(t)}\,dt$$

for any $f, g \in L^2(\mathbb{R})_{per}$, and in particular satisfy the formula

$$\sum_{n\in\mathbb{Z}} |\widehat{f}(n)|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |f(t)|^2 dt$$

for any $f \in L^2(\mathbb{R})_{per}$, with this being called Parseval formula.

PROOF. This is indeed yet another reformulation of what we have, coming from the fact that the scalar products and norms are invariant under $L^2(\mathbb{R})_{per} \simeq l^2(\mathbb{Z})$. □

With this discussed, time perhaps for some applications? And here, there are countless of them, because the above technology can be used in order to decompose various signals, such as mechanical, electromagnetic, seismic or acoustic waves, or even solutions of more complicated differential equations, somewhat of wave type, into sinusoids.

Thus, many things to be learned here, and for more, have a look at any advanced calculus book. In what concerns us, let us just present an application to arithmetic:

THEOREM 15.18. *We have the following formula of Euler,*

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \ldots = \frac{\pi^2}{6}$$

computing $\zeta(2)$, and solving the Basel problem.

PROOF. To start with, as already mentioned in chapter 14, we have the following computation of Euler, obtained by factorizing the function $\sin x$, having zeroes at $\mathbb{Z}/2$, a

bit like a polynomial, which gives the result, by looking at the coefficient of x^2:

$$\begin{aligned}\frac{\sin x}{x} &= 1-\frac{x^2}{3!}+\frac{x^4}{5!}-\frac{x^6}{7!}+\ldots\\ &= \left(1-\frac{x}{\pi}\right)\left(1+\frac{x}{\pi}\right)\left(1-\frac{x}{2\pi}\right)\left(1+\frac{x}{2\pi}\right)\ldots\\ &= \left(1-\frac{x^2}{\pi^2}\right)\left(1-\frac{x^2}{4\pi^2}\right)\left(1-\frac{x^2}{9\pi^2}\right)\ldots\\ &= 1-\frac{1}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\,x^2+\ldots\end{aligned}$$

Of course this is far from being rigorous, but following Weierstrass, it is possible to fix this, with the above formula for $\sin x/x$ being indeed true. And, exercise of course for you, to learn more about all this, with this being first-class mathematics, for sure.

(2) As an alternative approach, much quicker, we can use our Fourier series knowledge. Indeed, the nonzero Fourier coefficients of the function $f(t)=t$ on $[-\pi,\pi]$ are:

$$\widehat{f}(n)=\frac{1}{2\pi}\int_{-\pi}^{\pi}te^{int}\,dt=\frac{1}{2\pi}\left[\frac{1-int}{n^2}\,e^{int}\right]_{-\pi}^{\pi}=\frac{(-1)^{n+1}}{n}\,i$$

Thus, the Parseval formula for the function $f(t)=t$ gives:

$$\sum_{n\in\mathbb{Z}^*}\frac{1}{n^2}=\frac{1}{2\pi}\int_{-\pi}^{\pi}t^2dt=\frac{1}{2\pi}\cdot\frac{2\pi^3}{3}=\frac{\pi^2}{3}$$

And we therefore solved the Basel problem, just like that. Amazing. □

Along the same lines, we have as well the following more general result:

THEOREM 15.19. *The generating function of the numbers $\zeta(2k)$ with $k\in\mathbb{N}$ is*

$$\sum_{k=1}^{\infty}\zeta(2k)\left(\frac{x}{\pi}\right)^{2k}=\frac{1-x\cot x}{2}$$

and the numbers $\zeta(2k)$ with $k\in\mathbb{N}$ are themselves given by the formula

$$\zeta(2k)=(-1)^{k+1}\frac{(2\pi)^{2k}B_{2k}}{2\cdot(2k)!}$$

with B_n being the Bernoulli numbers, which in practice gives the formulae

$$\zeta(2)=\frac{\pi^2}{6}\quad,\quad\zeta(4)=\frac{\pi^4}{90}\quad,\quad\zeta(6)=\frac{\pi^6}{945}\quad,\quad\zeta(8)=\frac{\pi^8}{9450}\quad,\quad\ldots$$

generalizing the formula $\zeta(2)=\pi^2/6$ of Euler, solving the Basel problem.

PROOF. This comes as continuation of Theorem 15.18, following a discussion from chapter 14, and we will leave all this, as an interesting exercise for you. □

15c. Fourier transform

With the above discussed, end of Fourier analysis? You must be kidding. As we will soon discover, the Fourier series that we learned are just one type of Fourier analysis, and there is more. Let us start our discussion with something of general interest, namely the convolution operation, that we already met in chapter 14, when talking about heat:

DEFINITION 15.20. *The convolution of two functions $f, g : \mathbb{R} \to \mathbb{C}$ is the function*

$$(f * g)(x) = \int_{\mathbb{R}} f(x-y)g(y)dy$$

provided that the function $y \to f(x-y)g(y)$ is indeed integrable, for any x.

There are many reasons for introducing this operation, that we will gradually discover, in what follows. As a basic example, let us take $g = \chi_{[0,1]}$. We have then:

$$(f * g)(x) = \int_0^1 f(x-y)dy$$

Thus, with this choice of g, the operation $f \to f * g$ has some sort of "regularizing effect", that can be useful for many purposes. We will be back to this, later.

Goinh ahead with more theory, let us try to understand now when the convolution operation is well-defined. We have here the following basic result:

PROPOSITION 15.21. *The convolution operation is well-defined on the space*

$$C_c(\mathbb{R}) = \left\{ f \in C(\mathbb{R}) \middle| supp(f) = \text{ compact} \right\}$$

of continuous functions $f : \mathbb{R} \to \mathbb{C}$ having compact support.

PROOF. We have several things to be proved, the idea being as follows:

(1) First we must show that given two functions $f, g \in C_c(\mathbb{R})$, their convolution $f * g$ is well-defined, as a function $f * g : \mathbb{R} \to \mathbb{C}$. But this follows from the following estimate, where l denotes the length of the compact subsets of $\mathbb{R}$:

$$\begin{aligned}
\int_{\mathbb{R}} |f(x-y)g(y)|dy &= \int_{supp(g)} |f(x-y)g(y)|dy \\
&\leq \max(g) \int_{supp(g)} |f(x-y)|dy \\
&\leq \max(g) \cdot l(supp(g)) \cdot \max(f) \\
&< \infty
\end{aligned}$$

(2) Next, we must show that the function $f * g : \mathbb{R} \to \mathbb{C}$ that we constructed is indeed continuous. But this follows from the following estimate, where K_f is the constant of

uniform continuity for the function $f \in C_c(\mathbb{R})$:

$$\begin{aligned}|(f*g)(x+\varepsilon)-(f*g)(x)| &= \left|\int_{\mathbb{R}} f(x+\varepsilon-y)g(y)dy-\int_{\mathbb{R}} f(x-y)g(y)dy\right| \\ &= \left|\int_{\mathbb{R}} \left(f(x+\varepsilon-y)-f(x-y)\right)g(y)dy\right| \\ &\leq \int_{\mathbb{R}} |f(x+\varepsilon-y)-f(x-y)|\cdot|g(y)|dy \\ &\leq K_f\cdot\varepsilon\cdot\int_{\mathbb{R}}|g|\end{aligned}$$

(3) Finally, we must show that the function $f*g\in C(\mathbb{R})$ that we constructed has indeed compact support. For this purpose, our claim is that we have:

$$supp(f+g)\subset supp(f)+supp(g)$$

In order to prove this claim, observe that we have, by definition of $f*g$:

$$(f*g)(x)=\int_{\mathbb{R}} f(x-y)g(y)dy=\int_{supp(g)} f(x-y)g(y)dy$$

But this latter quantity being 0 for $x\notin supp(f)+supp(g)$, this gives the result. □

In relation with derivatives, and with the "regularizing effect" of the convolution operation mentioned after Definition 15.20, we have the following result:

THEOREM 15.22. *Given two functions $f,g\in C_c(\mathbb{R})$, assuming that g is differentiable, then so is $f*g$, with derivative given by the following formula:*

$$(f*g)'=f*g'$$

More generally, given $f,g\in C_c(\mathbb{R})$, and assuming that g is k times differentiable, then so is $f*g$, with k-th derivative given by $(f*g)^{(k)}=f*g^{(k)}$.

PROOF. In what regards the first assertion, with $y=x-t$, then $t=x-y$, we get:

$$\begin{aligned}(f*g)'(x) &= \frac{d}{dx}\int_{\mathbb{R}} f(x-y)g(y)dy \\ &= \frac{d}{dx}\int_{\mathbb{R}} f(t)g(x-t)dt \\ &= \int_{\mathbb{R}} f(t)g'(x-t)dt \\ &= \int_{\mathbb{R}} f(x-y)g'(y)dy \\ &= (f*g')(x)\end{aligned}$$

As for the second assertion, this follows form the first one, by recurrence. □

Finally, getting beyond the compactly supported continuous functions, we have the following result, which is of particular theoretical importance:

THEOREM 15.23. *The convolution operation is well-defined on $L^1(\mathbb{R})$, and we have:*

$$||f*g||_1 \leq ||f||_1||g||_1$$

Thus, if $f \in L^1(\mathbb{R})$ and $g \in C_c^k(\mathbb{R})$, then $f*g$ is well-defined, and $f*g \in C_c^k(\mathbb{R})$.

PROOF. In what regards the first assertion, this follows from:

$$\begin{aligned}\int_{\mathbb{R}}|(f*g)(x)|dx &\leq \int_{\mathbb{R}}\int_{\mathbb{R}}|f(x-y)g(y)|dydx\\ &= \int_{\mathbb{R}}\int_{\mathbb{R}}|f(x-y)g(y)|dxdy\\ &= \int_{\mathbb{R}}|f(x)|dx\int_{\mathbb{R}}|g(y)|dy\end{aligned}$$

As for the second assertion, this follows from this, and from Theorem 15.22. □

Let us discuss now the construction and main properties of the Fourier transform, which is the main tool in analysis, and mathematics in general. We first have:

DEFINITION 15.24. *Given $f \in L^1(\mathbb{R})$, we define a function $\widehat{f}:\mathbb{R}\to\mathbb{C}$ by*

$$\widehat{f}(x) = \int_{\mathbb{R}} f(t)e^{ixt}dt$$

and call it Fourier transform of f.

As a first observation, even if f is a real function, $\widehat{f}$ is a complex function, which is not necessarily real. Also, $\widehat{f}$ is obviously well-defined, because $f \in L^1(\mathbb{R})$ and $|e^{ixt}| = 1$. Also, the condition $f \in L^1(\mathbb{R})$ is basically needed for construcing $\widehat{f}$, because:

$$\widehat{f}(0) = \int_{\mathbb{R}} f(t)dt$$

Generally speaking, the Fourier transform is there for helping with various computations, with the above formula $\widehat{f}(0) = \int f$ being something quite illustrating. Here are some basic properties of the Fourier transform, all providing some good motivations:

PROPOSITION 15.25. *The Fourier transform has the following properties:*

(1) *Linearity: $\widehat{f+g} = \widehat{f}+\widehat{g}$, $\widehat{\lambda f} = \lambda\widehat{f}$.*
(2) *Regularity: $\widehat{f}$ is continuous and bounded.*
(3) *If f is even then $\widehat{f}$ is even.*
(4) *If f is odd then $\widehat{f}$ is odd.*

PROOF. All this is very standard, and we will leave the proof as an exercise. □

Here are as well some basic computations of Fourier transforms:

PROPOSITION 15.26. *We have the following Fourier transform formulae,*

$$f = \chi_{[-a,a]} \implies \widehat{f}(x) = \frac{2\sin(ax)}{x}$$

$$f = e^{-at}\chi_{[0,\infty]}(t) \implies \widehat{f}(x) = \frac{1}{a - ix}$$

$$f = e^{at}\chi_{[-\infty,0]}(t) \implies \widehat{f}(x) = \frac{1}{a + ix}$$

$$f = e^{-a|t|} \implies \widehat{f}(x) = \frac{2a}{a^2 + x^2}$$

valid for any number $a > 0$.

PROOF. All this is again standard, and we will leave the proof as an exercise. □

Back now to theory, we have the following result, adding to the various properties in Proposition 15.25, and providing more motivations for the Fourier transform:

PROPOSITION 15.27. *Given $f, g \in L^1(\mathbb{R})$ we have $\widehat{f}g, f\widehat{g} \in L^1(\mathbb{R})$ and*

$$\int_{\mathbb{R}} f(x)\widehat{g}(x)dx = \int_{\mathbb{R}} \widehat{f}(x)g(x)dx$$

called "exchange of hat" formula.

PROOF. Regarding the fact that we have indeed $\widehat{f}g, f\widehat{g} \in L^1(\mathbb{R})$, this is actually a bit non-trivial, but we will be back to this later. Assuming this, we have:

$$\int_{\mathbb{R}} f(x)\widehat{g}(x)dx = \int_{\mathbb{R}}\int_{\mathbb{R}} f(x)g(y)e^{ixy}dxdy$$

On the other hand, we have as well the following formula:

$$\int_{\mathbb{R}} \widehat{f}(x)g(x)dx = \int_{\mathbb{R}}\int_{\mathbb{R}} f(y)e^{iyx}g(x)dydx$$

Thus, with $x \leftrightarrow y$, we are led to the formula in the statement. □

As a key result now, showing the power of the Fourier transform, we have:

THEOREM 15.28. *Given $f : \mathbb{R} \to \mathbb{C}$ such that $f, f' \in L^1(\mathbb{R})$, we have:*

$$\widehat{f'}(x) = -ix\widehat{f}(x)$$

More generally, assuming $f, f', f'', \ldots, f^{(n)} \in L^1(\mathbb{R})$, we have

$$\widehat{f^{(k)}}(x) = (-ix)^k\widehat{f}(x)$$

for any $k = 1, 2, \ldots, n$.

PROOF. Assuming that $f : \mathbb{R} \to \mathbb{C}$ has compact support, we have indeed:

$$\begin{aligned}\widehat{f'}(x) &= \int_{\mathbb{R}} f'(t)e^{ixt}dt\\ &= -\int_{\mathbb{R}} f(t)\cdot ixe^{ixt}dt\\ &= -ix\int_{\mathbb{R}} f(t)e^{ixt}dt\\ &= -ix\widehat{f}(x)\end{aligned}$$

As for the higher derivatives, the formula here follows by recurrence. □

Importantly, we have a converse statement as well, as follows:

THEOREM 15.29. *Assuming that $f \in L^1(\mathbb{R})$ is such that $F(t) = tf(t)$ belongs to $L^1(\mathbb{R})$ too, the function $\widehat{f}$ is differentiable, with derivative given by:*

$$(\widehat{f})'(x) = i\widehat{F}(x)$$

More generally, if $F_k(t) = t^k f(t)$ belongs to $L^1(\mathbb{R})$, for $k = 0, 1, \ldots, n$, we have

$$(\widehat{f})^{(k)}(x) = i^k\widehat{F_k}(x)$$

for any $k = 1, 2, \ldots, n$.

PROOF. Regarding the first assertion, the computation here is as follows:

$$\begin{aligned}(\widehat{f})'(x) &= \frac{d}{dx}\int_{\mathbb{R}} f(t)e^{ixt}dt\\ &= \int_{\mathbb{R}} f(t)\cdot ite^{ixt}dt\\ &= i\int_{\mathbb{R}} tf(t)e^{ixt}dt\\ &= i\widehat{F}(x)\end{aligned}$$

As for the second assertion, this follows from the first one, by recurrence. □

Here is another useful result, of the same type, this time regarding convolutions:

THEOREM 15.30. *Assuming $f, g \in L^1(\mathbb{R})$, the following happens:*

$$\widehat{f * g} = \widehat{f}\cdot\widehat{g}$$

Conversely, we have $\widehat{fg} = \frac{1}{2\pi}\,\widehat{f} * \widehat{g}$, which holds almost everywhere.

PROOF. The first assertion is something elementary, coming as follows:

$$\begin{aligned}
\widehat{f*g}(x) &= \int_{\mathbb{R}}(f*g)(t)e^{ixt}dt\\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}f(t-s)g(s)e^{ixt}dsdt\\
&= \int_{\mathbb{R}}\left(\int_{\mathbb{R}}f(t-s)e^{ix(t-s)}dt\right)g(s)e^{ixs}ds\\
&= \int_{\mathbb{R}}\left(\int_{\mathbb{R}}f(r)e^{ixr}dr\right)g(s)e^{ixs}ds\\
&= \int_{\mathbb{R}}\widehat{f}(x)g(s)e^{ixs}ds\\
&= \widehat{f}(x)\widehat{g}(x)
\end{aligned}$$

As for the second assertion, this is something more tricky, which follows from the first one by using the Fourier inversion formula, that we will soon learn. □

Let us develop now more theory for the Fourier transform. We first have:

THEOREM 15.31. *Given $f \in L^1(\mathbb{R})$, its Fourier transform satisfies*

$$\lim_{x\to\pm\infty}\widehat{f}(x) = 0$$

called Riemann-Lebesgue property of $\widehat{f}$.

PROOF. This is something quite technical, as follows:

(1) Given a function $f : \mathbb{R} \to \mathbb{C}$ and a number $r \in \mathbb{R}$, let us set:

$$f_r(t) = f(t-r)$$

Our claim is then is that if $f \in L^p(\mathbb{R})$, then the following function is uniformly continuous, with respect to the usual p-norm on the right:

$$\mathbb{R} \to L^p(\mathbb{R}) \quad , \quad r \to f_r$$

(2) In order to prove this, fix $\varepsilon > 0$. Since $f \in L^p(\mathbb{R})$, we can find a function of type $g : [-K, K] \to \mathbb{C}$ which is continuous, such that:

$$||f - g||_p < \varepsilon$$

Now since g is uniformly continuous, we can find $\delta \in (0, K)$ such that:

$$|u - v| < \delta \implies |g(u) - g(v)| < (3K)^{-1/p}\varepsilon$$

But this shows that we have the following estimate:

$$\begin{aligned}||g_r-g_s||_p &= \left(\int_{\mathbb{R}}\left|g(t-r)-g(t-s)\right|^p dt\right)^{1/p}\\ &< \left[(3K)^{-1}\varepsilon^p(2k+\delta)\right]^{1/p}\\ &< \varepsilon\end{aligned}$$

By using now the formula $||f||_p=||f_r||_p$, which is clear, we obtain:

$$\begin{aligned}||f_r-f_s||_p &\leq ||f_r-g_r||_p+||g_r-g_s||_p+||g_s-f_s||_p\\ &< \varepsilon+\varepsilon+\varepsilon\\ &= 3\varepsilon\end{aligned}$$

But this being true for any $|r-s|<\delta$, we have proved our claim.

(3) Let us prove now the Riemann-Lebesgue property of $\widehat{f}$, as formulated in the statement. By using $e^{\pi i}=-1$, and the change of variables $t\to t-\pi/x$, we have:

$$\begin{aligned}\widehat{f}(x) &= \int_{\mathbb{R}} f(t)e^{ixt}dt\\ &= -\int_{\mathbb{R}} e^{\pi i}f(t)e^{ixt}dt\\ &= -\int_{\mathbb{R}} f(t)e^{ix(t+\pi/x)}dt\\ &= -\int_{\mathbb{R}} f\left(t-\frac{\pi}{x}\right)e^{ixt}dt\end{aligned}$$

On the other hand, we have as well the following formula:

$$\widehat{f}(x)=\int_{\mathbb{R}} f(t)e^{ixt}dt$$

Thus by summing, we obtain the following formula:

$$2\widehat{f}(x)=\int_{\mathbb{R}}\left(f(t)-f\left(t-\frac{\pi}{x}\right)\right)e^{ixt}dt$$

But this gives the following estimate, with notations from (1):

$$2|\widehat{f}(x)|\leq ||f-f_{\pi/x}||_1$$

Since by (1) this goes to 0 with $x\to\pm\infty$, this gives the result. □

Quite remarkably, and as a main result now regarding Fourier transforms, a function $f:\mathbb{R}\to\mathbb{C}$ can be recovered from its Fourier transform $\widehat{f}:\mathbb{R}\to\mathbb{C}$, as follows:

THEOREM 15.32. *Assuming $f, \widehat{f} \in L^1(\mathbb{R})$, we have*

$$f(t) = \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(x)e^{-itx}dx$$

almost everywhere, called Fourier inversion formula.

PROOF. Consider the following function, depending on a parameter $\lambda > 0$:

$$\varphi_\lambda(s) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-isx-\lambda|x|}dx$$

We have then the following convolution computation:

$$\begin{aligned}(f * \varphi_\lambda)(t) &= \int_{\mathbb{R}} f(t-s)\varphi_\lambda(s)ds\\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} f(t-s)e^{-isx-\lambda|x|}dxds\\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\left(\int_{\mathbb{R}} f(t-s)e^{-isx}ds\right)e^{-\lambda|x|}dx\\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(x)e^{-itx}e^{-\lambda|x|}dx\end{aligned}$$

By letting now $\lambda \to 0$, we obtain from this the following formula:

$$\lim_{\lambda\to 0}(f * \varphi_\lambda)(t) = \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(x)e^{-itx}dx$$

On the other hand, by using Theorem 15.31 we obtain that, almost everywhere:

$$\lim_{\lambda\to 0}(f * \varphi_\lambda)(t) = f(t)$$

We are therefore led to the conclusion in the statement. □

As a first application, we have now the fix for Theorem 15.30. In general, the Fourier transform can be used a bit like the Fourier series, for dealing with all sorts of differential equations, and exercise of course for you, to learn more about all this.

15d. Distributions

We kept the best for the end, the mathematical distributions. These are something quite smart, and as an advertisement for what we will be doing, we have:

ADVERTISEMENT 15.33. *With a suitable theory of distributions, covering both the functions and the Dirac masses, the basic step function, namely*

$$H(x) = \begin{cases} 0 & (x \leq 0)\\ 1 & (x > 0)\end{cases}$$

is differentiable when viewed as distribution, with derivative $H' = \delta_0$.

And isn't this crazy, hope you agree with me. Getting started now, there is a price to pay for doing such things, namely formulating a technical definition, as follows:

DEFINITION 15.34. *A distribution on an open interval $I \subset \mathbb{R}$ is a functional*

$$\varphi : C_c^\infty(I) \to \mathbb{C}$$

such that for any $K \subset I$ compact, there exist $n \in \mathbb{N}$ and $c > 0$ such that

$$|\varphi(f)| \leq c||f||_{C^n(K)}$$

for any $f \in C_c^\infty(I)$ having support in K, where $||f||_{C^n(K)} = \sup_{x \in K} \sum_{i=0}^n |f^{(i)}(x)|$.

At the level of main examples of distributions, we have the integration functionals associated to the measures, and in particular to the measures having a density. In view of this, we can consider any function $f \in L^1(I)$, viewed as density, as being a distribution. Other basic examples include the Dirac masses δ_x at the points $x \in I$.

Regarding the general theory of distributions, that is quite similar to the theory of functions. Algebrically, the distributions form a vector space, and are subject to a number of supplementary operations, such as dilations, translations and so on, and multiplication by functions too. Analytically, we can talk about convergence of distributions, $\varphi_n \to \varphi$, and about their support too, $supp(\varphi) \subset I$, in a quite straightforward way.

Getting now to what we wanted to do, derivatives, we have here:

THEOREM 15.35. *We can talk about the derivatives of distributions, given by*

$$\varphi'(f) = -\varphi(f')$$

and with this notion in hand, the following happen:

(1) *When φ is a usual differentiable function, φ' is the usual derivative.*
(2) *For the basic step function we have $H' = \delta_0$, as previously advertised.*
(3) *In fact, for a function $\varphi = g$ with jumps at $\{x_i\}$, we have $\varphi' = g' + \sum_i J_g(x_i)\delta_{x_i}$.*

PROOF. The first assertion, which by the way explains the need for the $-$ sign, follows from the Leibnitz rule for derivatives. Regarding the second assertion, this follows from:

$$\begin{aligned} H'(f) &= -H(f') \\ &= -\int_0^\infty f'(x)dx \\ &= -f(\infty) + f(0) \\ &= f(0) \\ &= \delta_0(f) \end{aligned}$$

As for the third assertion, which generalizes (1,2), we will leave this as an exercise. □

Summarizing, job done. Many other things can be said about distributions, and we would like to end with a physics trick. We have the following computation, for $a > 0$:

$$\begin{aligned}
\int_{\mathbb{R}} f(x)\delta(x^2-a^2)dx &= \int_{-\infty}^{0} f(x)\delta(x^2-a^2)dx + \int_{0}^{\infty} f(x)\delta(x^2-a^2)dx \\
&= \int_{-\infty}^{a} f(y-a)\delta(y^2-2ay)dy + \int_{-a}^{\infty} f(y+a)\delta(y^2+2ay)dy \\
&\simeq \int_{-\infty}^{a} f(y-a)\delta(-2ay)dy + \int_{-a}^{\infty} f(y+a)\delta(2ay)dy \\
&= \int_{-\infty}^{2a^2} f\left(\frac{z}{2a}-a\right)\delta(-z)\frac{dz}{2a} + \int_{-2a^2}^{\infty} f\left(\frac{z}{2a}+a\right)\delta(z)\frac{dz}{2a} \\
&= \frac{f(-a)}{2a} + \frac{f(a)}{2a} \\
&= \int_{R} f(x)\,\frac{\delta(x-a)+\delta(x+a)}{2a}\,dx
\end{aligned}$$

Sounds like physics, you would say, and in answer, yes physics that is, but in any case, we have in this way a definition for the quadratic Dirac masses, as follows:

$$\delta(x^2-a^2) = \frac{\delta(x-a)+\delta(x+a)}{2a}$$

And we will end our discussion here. Good learning this was, and exercise for you to learn more about distributions, and their applications to mathematics and physics.

15e. Exercises

This was a quite exciting chapter, and as exercises on this, we have:

EXERCISE 15.36. *Learn some other proofs of Young, Hölder, Minkowski.*

EXERCISE 15.37. *Learn about other Banach spaces of sequences, such as c_0.*

EXERCISE 15.38. *Learn about the Hahn-Banach theorem, and related topics.*

EXERCISE 15.39. *Learn more about Hilbert spaces, and linear operators on them.*

EXERCISE 15.40. *Learn also about the various families of orthogonal polynomials.*

EXERCISE 15.41. *Learn more about Fourier series, and their applications.*

EXERCISE 15.42. *Learn more about the Fourier transform, and its applications.*

EXERCISE 15.43. *Learn more about distributions, and their various applications.*

As bonus exercise, find a good functional analysis book, and start reading it.

CHAPTER 16

Several variables

16a. Linear algebra

We would like to end this book with an introduction to multivariable calculus. Which is reputed to be a complicated business, but the basics are fairly understandable, and in any case, far simpler than what we have been doing recently, in chapters 14-15.

Thinking a bit, things look quite complicated in several dimensions, say in the familiar 3D that we live in, and which is of main interest, for instance because the graph of a function $f : \mathbb{R}^3 \to \mathbb{R}$ is a certain 4-dimensional beast, and go understand that.

Fortunately, the main ideas of calculus, as we learned them so far, come to the rescue. Based on what we know well in 1D, let us formulate things as follows:

QUESTION 16.1. *The main idea of calculus was that the functions $f : \mathbb{R} \to \mathbb{R}$ are locally approximately linear. In view of this, when looking for generalizations:*

(1) *What can we say about the linear maps $f : \mathbb{R}^N \to \mathbb{R}^M$?*
(2) *Then, what can we say about the arbitrary functions $f : \mathbb{R}^N \to \mathbb{R}^M$?*

Which sounds good, now we have a serious plan, and time to develop it. Regarding the first question, about the linear maps $f : \mathbb{R}^N \to \mathbb{R}^M$, let us start with some basic linear algebra. At the beginning, we have the following result, that you surely know:

THEOREM 16.2. *The linear maps $f : \mathbb{R}^N \to \mathbb{R}^M$ are in correspondence with the matrices $A \in M_{M\times N}(\mathbb{R})$, with the linear map associated to such a matrix being*

$$f(x) = Ax$$

and with the matrix associated to a linear map being $A_{ij} =< f(e_j), e_i >$.

PROOF. The first assertion is clear, because a linear map $f : \mathbb{R}^N \to \mathbb{R}^M$ must send a vector $x \in \mathbb{R}^N$ to a certain vector $f(x) \in \mathbb{R}^M$, all whose components are linear combinations of the components of x. Thus, we can write, for certain real numbers $A_{ij} \in \mathbb{R}$:

$$f\begin{pmatrix} x_1 \\ \vdots \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} A_{11}x_1 + \ldots + A_{1N}x_N \\ \vdots \\ \vdots \\ A_{M1}x_1 + \ldots + A_{MN}x_N \end{pmatrix}$$

Now the parameters $A_{ij} \in \mathbb{R}$ can be regarded as being the entries of a certain matrix $A \in M_{M\times N}(\mathbb{R})$, and with the usual convention for matrix multiplication, we have:

$$f(x) = Ax$$

Regarding the second assertion, with $f(x) = Ax$ as above, if we denote by $e_1, \ldots, e_N$ the standard basis of $\mathbb{R}^N$, then we have the following formula:

$$f(e_j) = \begin{pmatrix} A_{1j} \\ \vdots \\ \vdots \\ A_{Mj} \end{pmatrix}$$

But this gives the second formula in the statement, $< f(e_j), e_i >= A_{ij}$, as desired. □

In order to reach now to sharper results, we will restrict the attention to the linear maps $f : \mathbb{R}^N \to \mathbb{R}^N$, which amounts in looking at the square matrices $A \in M_N(\mathbb{R})$. And in what regards these latter matrices, we first have the following result:

PROPOSITION 16.3. *Given a matrix $A \in M_N(\mathbb{R})$, let us call $v \in \mathbb{R}^N$ an eigenvector, with corresponding eigenvalue λ, when A multiplies by λ in the direction of v:*

$$Av = \lambda v$$

Then, in case where $\mathbb{R}^N$ has a basis $v_1, \ldots, v_N$ formed by eigenvectors of A, with corresponding eigenvalues $\lambda_1, \ldots, \lambda_N$, in this new basis A becomes diagonal, as follows:

$$A \sim \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix}$$

Equivalently, if we denote by $D = diag(\lambda_1, \ldots, \lambda_N)$ the above diagonal matrix, and by $P = [v_1 \ldots v_N]$ the square matrix formed by the eigenvectors of A, we have:

$$A = PDP^{-1}$$

In this case we say that the matrix A is diagonalizable.

PROOF. This is something which is clear, the idea being as follows:

(1) The first assertion is clear, because the matrix which multiplies each basis element v_i by a number λ_i is precisely the diagonal matrix $D = diag(\lambda_1, \ldots, \lambda_N)$.

(2) The second assertion follows from the first one, by changing the basis. We can prove this by a direct computation as well, because we have $Pe_i = v_i$, and so:

$$PDP^{-1}v_i = PDe_i = P\lambda_i e_i = \lambda_i Pe_i = \lambda_i v_i$$

Thus, the matrices A and PDP^{-1} coincide, as stated. □

In general, in order to study the diagonalization problem, the idea is that the eigenvectors can be grouped into linear spaces, called eigenspaces, as follows:

PROPOSITION 16.4. *Let $A \in M_N(\mathbb{R})$, and for any eigenvalue $\lambda \in \mathbb{R}$ define the corresponding eigenspace as being the vector space formed by the corresponding eigenvectors:*

$$E_\lambda = \left\{ v \in \mathbb{R}^N \middle| Av = \lambda v \right\}$$

These eigenspaces E_λ are then in a direct sum position, in the sense that given vectors $v_1 \in E_{\lambda_1}, \ldots, v_k \in E_{\lambda_k}$ corresponding to different eigenvalues $\lambda_1, \ldots, \lambda_k$, we have:

$$\sum_i c_i v_i = 0 \implies c_i = 0$$

In particular we have the following estimate, with sum over all the eigenvalues,

$$\sum_\lambda \dim(E_\lambda) \leq N$$

and our matrix is diagonalizable precisely when we have equality.

PROOF. We prove the first assertion by recurrence on $k \in \mathbb{N}$. Assume by contradiction that we have a formula as follows, with the scalars $c_1, \ldots, c_k$ being not all zero:

$$c_1 v_1 + \ldots + c_k v_k = 0$$

By dividing by one of these scalars, we can assume that our formula is:

$$v_k = c_1 v_1 + \ldots + c_{k-1} v_{k-1}$$

Now let us apply A to this vector. We obtain in this way the following equality:

$$\lambda_k \left(c_1 v_1 + \ldots + c_{k-1} v_{k-1} \right) = c_1 \lambda_1 v_1 + \ldots + c_{k-1} \lambda_{k-1} v_{k-1}$$

On the other hand, we know by recurrence that the vectors $v_1, \ldots, v_{k-1}$ must be linearly independent. Thus, the coefficients must be equal, at left and at right:

$$\lambda_k c_1 = c_1 \lambda_1 \quad , \quad \ldots \quad , \quad \lambda_k c_{k-1} = c_{k-1} \lambda_{k-1}$$

Now since at least one of the numbers c_i must be nonzero, we obtain $\lambda_k = \lambda_i$, which is a contradiction. Thus, first assertion proved, and the second assertion follows it. □

In order to reach now to more advanced results, based on the above, we can use the characteristic polynomial, which appears in the following way:

PROPOSITION 16.5. *Given $A \in M_N(\mathbb{R})$, consider its characteristic polynomial:*

$$P(x) = \det(A - x 1_N)$$

The eigenvalues of A are then the roots of P. Also, we have the inequality

$$\dim(E_\lambda) \leq m_\lambda$$

where m_λ is the multiplicity of λ, as root of P.

PROOF. The first assertion follows from the following computation, using the fact that a linear map is bijective when the determinant of the associated matrix is nonzero:

$$\begin{aligned}\exists v, Av = \lambda v &\iff \exists v, (A - \lambda 1_N)v = 0\\ &\iff \det(A - \lambda 1_N) = 0\end{aligned}$$

Regarding now the second assertion, given an eigenvalue λ of our matrix A, consider the dimension $d_\lambda = \dim(E_\lambda)$ of the corresponding eigenspace. By changing the basis of $\mathbb{R}^N$, as for the eigenspace E_λ to be spanned by the first d_λ basis elements, our matrix becomes as follows, with B being a certain smaller matrix:

$$A \sim \begin{pmatrix} \lambda 1_{d_\lambda} & 0\\ 0 & B\end{pmatrix}$$

We conclude that the characteristic polynomial of A is of the following form:

$$P_A = P_{\lambda 1_{d_\lambda}} P_B = (\lambda - x)^{d_\lambda} P_B$$

Thus the multiplicity m_λ of our eigenvalue λ, as a root of P, satisfies $m_\lambda \geq d_\lambda$, and this leads to the conclusion in the statement. □

It is convenient now to regard $A \in M_N(\mathbb{R})$ as a complex matrix, $A \in M_N(\mathbb{C})$, as for its characteristic polynomial P to have roots. We are led in this way to:

THEOREM 16.6. *Given a matrix $A \in M_N(\mathbb{C})$, consider its characteristic polynomial*

$$P(X) = \det(A - X1_N)$$

then factorize this polynomial, by computing the complex roots, with multiplicities,

$$P(X) = (-1)^N (X - \lambda_1)^{n_1} \ldots (X - \lambda_k)^{n_k}$$

and finally compute the corresponding eigenspaces, for each eigenvalue found:

$$E_i = \left\{ v \in \mathbb{C}^N \middle| Av = \lambda_i v \right\}$$

The dimensions of these eigenspaces satisfy then the following inequalities,

$$\dim(E_i) \leq n_i$$

and A is diagonalizable precisely when we have equality for any i.

PROOF. This follows by combining Propositions 16.4 and 16.5, or rather their complex versions, whose proofs are identical to those in the real case. Indeed, by summing the inequalities $\dim(E_\lambda) \leq m_\lambda$ from Proposition 16.5, we obtain an inequality as follows:

$$\sum_\lambda \dim(E_\lambda) \leq \sum_\lambda m_\lambda \leq N$$

On the other hand, we know from Proposition 16.4 that our matrix is diagonalizable when we have global equality. Thus, we are led to the conclusion in the statement. □

In practice, diagonalizing a matrix remains something quite complicated. Let us record as well a useful algorithmic version of the above result, as follows:

THEOREM 16.7. *The square matrices $A \in M_N(\mathbb{C})$ can be diagonalized as follows:*

(1) *Compute the characteristic polynomial.*
(2) *Factorize the characteristic polynomial.*
(3) *Compute the eigenvectors, for each eigenvalue found.*
(4) *If there are no N eigenvectors, A is not diagonalizable.*
(5) *Otherwise, A is diagonalizable, $A = PDP^{-1}$.*

PROOF. This is an informal reformulation of Theorem 16.6, with (4) referring to the total number of linearly independent eigenvectors found in (3), and with $A = PDP^{-1}$ in (5) being the usual diagonalization formula, with P, D being as before. □

As an illustration for this, which is a must-know computation, we have:

PROPOSITION 16.8. *The rotation of angle $t \in \mathbb{R}$ in the plane diagonalizes as:*

$$\begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} \begin{pmatrix} e^{-it} & 0 \\ 0 & e^{it} \end{pmatrix} \begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix}$$

Over the reals this is impossible, unless $t = 0, \pi$, where the rotation is diagonal.

PROOF. Observe first that, as stated, unlike we are in the case $t = 0, \pi$, where our rotation is $\pm 1_2$, our rotation is a "true" rotation, having no eigenvectors in the plane. Fortunately the complex numbers come to the rescue, via the following computation:

$$\begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} 1 \\ i \end{pmatrix} = \begin{pmatrix} \cos t - i \sin t \\ i \cos t + \sin t \end{pmatrix} = e^{-it} \begin{pmatrix} 1 \\ i \end{pmatrix}$$

We have as well a second complex eigenvector, coming from:

$$\begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} 1 \\ -i \end{pmatrix} = \begin{pmatrix} \cos t + i \sin t \\ -i \cos t + \sin t \end{pmatrix} = e^{it} \begin{pmatrix} 1 \\ -i \end{pmatrix}$$

Thus, we are led to the conclusion in the statement. □

As another basic illustration, in N dimensions, we have the following result:

PROPOSITION 16.9. *The all-one matrix diagonalizes as follows,*

$$\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix} = \frac{1}{N} F_N \begin{pmatrix} N & & \\ & 0 & \\ & & \ddots \\ & & & 0 \end{pmatrix} F_N^*$$

with $F_N = (w^{ij})_{ij}$ with $w = e^{2\pi i/N}$ being the Fourier matrix.

PROOF. The all-one matrix being N times the projection on the all-one vector, the diagonal form is the one in the statement. In order to find now the explicit diagonalization formula, with passage matrix and its inverse, we must solve the following equation:

$$x_1 + \ldots + x_N = 0$$

And this is not an easy task, if we want a nice basis for the space of solutions. Fortunately, the complex numbers come to the rescue, via the following formula:

$$\sum_{k=0}^{N-1} w^{ks} = N\delta_{N|s}$$

But this leads, after some thinking, to the conclusion in the statement. □

As a key result now regarding diagonalization, we have:

THEOREM 16.10. *Any matrix $A \in M_N(\mathbb{R})$ which is symmetric, $A = A^t$, is diagonalizable over the reals, with the diagonalization being of the following type,*

$$A = UDU^t$$

with U orthogonal, meaning $U^t = U^{-1}$, and D diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because if we take a matrix of the form $A = UDU^t$, with U orthogonal and D diagonal, we have:

$$A^t = (UDU^t)^t = UDU^t = A$$

In the other sense now, assume that A is symmetric, $A = A^t$. Our first claim is that the eigenvalues are real. Indeed, assuming $Av = \lambda v$, we have:

$$\begin{aligned} \lambda < v, v > &= < \lambda v, v > \\ &= < Av, v > \\ &= < v, Av > \\ &= < v, \lambda v > \\ &= \bar{\lambda} < v, v > \end{aligned}$$

Thus we obtain $\lambda \in \mathbb{R}$, as claimed. Our next claim is that the eigenspaces corresponding to different eigenvalues are pairwise orthogonal. Assume indeed that:

$$Av = \lambda v \quad , \quad Aw = \mu w$$

We have then the following computation, using the fact that we have $\lambda, \mu \in \mathbb{R}$:

$$\begin{aligned}
\lambda < v, w > &= < \lambda v, w > \\
&= < Av, w > \\
&= < v, Aw > \\
&= < v, \mu w > \\
&= \mu < v, w >
\end{aligned}$$

Thus $\lambda \neq \mu$ implies $v \perp w$, as claimed. In order now to finish the proof, it remains to prove that the eigenspaces of A span the whole space $\mathbb{R}^N$. For this purpose, we will use a recurrence method. Let us pick an eigenvector of our matrix:

$$Av = \lambda v$$

Assuming now that we have a vector w orthogonal to it, $v \perp w$, we have:

$$\begin{aligned}
< Aw, v > &= < w, Av > \\
&= < w, \lambda v > \\
&= \lambda < w, v > \\
&= 0
\end{aligned}$$

Thus, if v is an eigenvector, then the vector space $v^\perp$ is invariant under A. We can therefore proceed by recurrence, and we obtain the result. □

And with this, good news, done with the linear algebra that we should know. We have now a good understanding of the linear maps $f : \mathbb{R}^N \to \mathbb{R}^M$, and we will make a heavy use of this material, when investigating the arbitrary functions $f : \mathbb{R}^N \to \mathbb{R}^M$.

16b. Partial derivatives

Getting now to the case of the arbitrary functions $f : \mathbb{R}^N \to \mathbb{R}^M$, let us first discuss differentiability. At order 1, the situation is quite simple, as follows:

THEOREM 16.11. *The derivative of a function $f : \mathbb{R}^N \to \mathbb{R}^M$, making the formula*

$$f(x + t) \simeq f(x) + f'(x)t$$

work, must be the matrix of partial derivatives at x, namely

$$f'(x) = \left(\frac{df_i}{dx_j}(x) \right)_{ij} \in M_{M \times N}(\mathbb{R})$$

acting on the vectors $t \in \mathbb{R}^N$ by usual multiplication.

PROOF. As a first observation, the formula in the statement makes sense indeed, as an equality, or rather approximation, of vectors in $\mathbb{R}^M$, as follows:

$$f\begin{pmatrix}x_1+t_1\\ \vdots\\ x_N+t_N\end{pmatrix}\simeq f\begin{pmatrix}x_1\\ \vdots\\ x_N\end{pmatrix}+\begin{pmatrix}\frac{df_1}{dx_1}(x) & \ldots & \frac{df_1}{dx_N}(x)\\ \vdots & & \vdots\\ \frac{df_M}{dx_1}(x) & \ldots & \frac{df_M}{dx_N}(x)\end{pmatrix}\begin{pmatrix}t_1\\ \vdots\\ t_N\end{pmatrix}$$

In order to prove now this formula, we can proceed by recurrence, as follows:

(1) First of all, at $N=M=1$ what we have is a usual 1-variable function $f:\mathbb{R}\to\mathbb{R}$, and the formula in the statement is something that we know well, namely:

$$f(x+t)\simeq f(x)+f'(x)t$$

(2) Let us discuss now the case $N=2, M=1$. Here what we have is a function $f:\mathbb{R}^2\to\mathbb{R}$, and by using twice the basic approximation result from (1), we obtain:

$$\begin{aligned}f\begin{pmatrix}x_1+t_1\\ x_2+t_2\end{pmatrix}&\simeq f\begin{pmatrix}x_1+t_1\\ x_2\end{pmatrix}+\frac{df}{dx_2}(x)t_2\\ &\simeq f\begin{pmatrix}x_1\\ x_2\end{pmatrix}+\frac{df}{dx_1}(x)t_1+\frac{df}{dx_2}(x)t_2\\ &= f\begin{pmatrix}x_1\\ x_2\end{pmatrix}+\begin{pmatrix}\frac{df}{dx_1}(x) & \frac{df}{dx_2}(x)\end{pmatrix}\begin{pmatrix}t_1\\ t_2\end{pmatrix}\end{aligned}$$

(3) More generally, we can deal in this way with the case $N\in\mathbb{N}, M=1$, by recurrence. But this gives the result in the general case $N,M\in\mathbb{N}$ too. Indeed, let us write:

$$f=\begin{pmatrix}f_1\\ \vdots\\ f_M\end{pmatrix}$$

We can apply our result to each of the components $f_i:\mathbb{R}^N\to\mathbb{R}$, and we get:

$$f_i\begin{pmatrix}x_1+t_1\\ \vdots\\ x_N+t_N\end{pmatrix}\simeq f_i\begin{pmatrix}x_1\\ \vdots\\ x_N\end{pmatrix}+\begin{pmatrix}\frac{df_i}{dx_1}(x) & \ldots & \frac{df_i}{dx_N}(x)\end{pmatrix}\begin{pmatrix}t_1\\ \vdots\\ t_N\end{pmatrix}$$

But this is precisely what we want, at the level of the global map $f:\mathbb{R}^N\to\mathbb{R}^M$. □

As a technical complement to the above result, we have:

THEOREM 16.12. *For a function $f:X\to\mathbb{R}^M$, with $X\subset\mathbb{R}^N$, the following conditions are equivalent, and in this case we say that f is continuously differentiable:*

(1) *f is differentiable, and the map $x\to f'(x)$ is continuous.*
(2) *f has partial derivatives, which are continuous with respect to $x\in X$.*

If these conditions are satisfied, $f'(x)$ is the matrix fomed by the partial derivatives at x.

PROOF. We already know, from Theorem 16.11, that the last assertion holds. Regarding now the proof of the equivalence, this goes as follows:

(1) $\implies$ (2) Assuming that f is differentiable, we know from Theorem 16.11 that $f'(x)$ is the matrix fomed by the partial derivatives at x. Thus, for any $x, y \in X$:

$$\frac{df_i}{dx_j}(x) - \frac{df_i}{dx_j}(y) = f'(x)_{ij} - f'(y)_{ij}$$

By applying now the absolute value, we obtain from this the following estimate:

$$\begin{aligned}\left|\frac{df_i}{dx_j}(x) - \frac{df_i}{dx_j}(y)\right| &= |f'(x)_{ij} - f'(y)_{ij}| \\ &= |(f'(x) - f'(y))_{ij}| \\ &\leq ||f'(x) - f'(y)||\end{aligned}$$

But this gives the result, because if the map $x \to f'(x)$ is assumed to be continuous, then the partial derivatives follow to be continuous with respect to $x \in X$.

(2) $\implies$ (1) This is something more technical. For simplicity, let us assume $M = 1$, the proof in general being similar. Given $x \in X$ and $\varepsilon > 0$, let us pick $r > 0$ such that the ball $B = B_x(r)$ belongs to X, and such that the following happens, over B:

$$\left|\frac{df}{dx_j}(x) - \frac{df}{dx_j}(y)\right| < \frac{\varepsilon}{N}$$

Our claim is that, with this choice made, we have the following estimate, for any $t \in \mathbb{R}^N$ satisfying $||t|| < r$, with A being the vector of partial derivatives at x:

$$|f(x+t) - f(x) - At| \leq \varepsilon ||t||$$

In order to prove this claim, the idea will be that of suitably applying the mean value theorem, over the N directions of $\mathbb{R}^N$. Indeed, consider the following vectors:

$$t^{(k)} = \begin{pmatrix} t_1 \\ \vdots \\ t_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

In terms of these vectors, we have the following formula:

$$f(x+t) - f(x) = \sum_{j=1}^{N} f(x + t^{(j)}) - f(x + t^{(j-1)})$$

Also, the mean value theorem gives a formula as follows, with $s_j \in [0,1]$:

$$f(x+t^{(j)})-f(x+t^{(j-1)})=\frac{df}{dx_j}(x+s_jt^{(j)}+(1-s_j)t^{(j-1)})\cdot t_j$$

But, according to our assumption on $r>0$ from the beginning, the derivative on the right differs from $\frac{df}{dx_j}(x)$ by something which is smaller than ε/N:

$$\left|\frac{df}{dx_j}(x+s_jt^{(j)}+(1-s_j)t^{(j-1)})-\frac{df}{dx_j}(x)\right|<\frac{\varepsilon}{N}$$

Now by putting everything together, we obtain the following estimate:

$$\begin{aligned}
|f(x+t)-f(x)-At| &= \left|\sum_{j=1}^{N}f(x+t^{(j)})-f(x+t^{(j-1)})-\frac{df}{dx_j}(x)\cdot t_j\right| \\
&\leq \sum_{j=1}^{N}\left|f(x+t^{(j)})-f(x+t^{(j-1)})-\frac{df}{dx_j}(x)\cdot t_j\right| \\
&= \sum_{j=1}^{N}\left|\frac{df}{dx_j}(x+s_jt^{(j)}+(1-s_j)t^{(j-1)})\cdot t_j-\frac{df}{dx_j}(x)\cdot t_j\right| \\
&= \sum_{j=1}^{N}\left|\frac{df}{dx_j}(x+s_jt^{(j)}+(1-s_j)t^{(j-1)})-\frac{df}{dx_j}(x)\right|\cdot|t_j| \\
&\leq \sum_{j=1}^{N}\frac{\varepsilon}{N}\cdot|t_j| \\
&\leq \varepsilon||t||
\end{aligned}$$

Thus we have proved our claim, and this gives the result. □

Moving on, with this done, our next task will be that of extending to several variables our basic results from one-variable calculus. As a standard result here, we have:

THEOREM 16.13. *We have the chain derivative formula*

$$(f\circ g)'(x)=f'(g(x))\cdot g'(x)$$

as an equality of matrices.

PROOF. This is something standard in one variable, and in several variables the proof is similar, by using the abstract notion of derivative coming from Theorem 16.11. To be more precise, consider a composition of functions, as follows:

$$f:\mathbb{R}^N\to\mathbb{R}^M\quad,\quad g:\mathbb{R}^K\to\mathbb{R}^N\quad,\quad f\circ g:\mathbb{R}^K\to\mathbb{R}^M$$

According to Theorem 16.11, the derivatives of these functions are certain linear maps, corresponding to certain rectangular matrices, as follows:

$$f'(g(x)) \in M_{M\times N}(\mathbb{R}) \quad , \quad g'(x) \in M_{N\times K}(\mathbb{R}) \qquad (f\circ g)'(x) \in M_{M\times K}(\mathbb{R})$$

Thus, our formula makes sense indeed. As for proof, this comes from:

$$\begin{aligned}(f\circ g)(x+t) &= f(g(x+t))\\ &\simeq f(g(x)+g'(x)t)\\ &\simeq f(g(x))+f'(g(x))g'(x)t\end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Moving on, we can talk as well about higher derivatives, simply by performing the operation of taking derivatives recursively. As a key result here, we have:

THEOREM 16.14. *The double derivatives of a function $f:\mathbb{R}^2\to\mathbb{R}$ satisfy*

$$\frac{d^2f}{dxdy}=\frac{d^2f}{dydx}$$

called Clairaut formula.

PROOF. Given a point in the plane, $z=(a,b)$, consider the following functions, depending on $h,k\in\mathbb{R}$ small:

$$u(h,k)=f(a+h,b+k)-f(a+h,b)$$

$$v(h,k)=f(a+h,b+k)-f(a,b+k)$$

$$w(h,k)=f(a+h,b+k)-f(a+h,b)-f(a,b+k)+f(a,b)$$

By the mean value theorem, for $h,k\neq 0$ we can find $\alpha,\beta\in\mathbb{R}$ such that:

$$\begin{aligned}w(h,k) &= u(h,k)-u(0,k)\\ &= h\cdot\frac{d}{dx}\,u(\alpha h,k)\\ &= h\left(\frac{d}{dx}\,f(a+\alpha h,b+k)-\frac{d}{dx}\,f(a+\alpha h,b)\right)\\ &= hk\cdot\frac{d}{dy}\cdot\frac{d}{dx}\,f(a+\alpha h,b+\beta k)\end{aligned}$$

Similarly, again for $h, k \neq 0$, we can find $\gamma, \delta \in \mathbb{R}$ such that:

$$\begin{aligned} w(h,k) &= v(h,k) - v(h,0) \\ &= k \cdot \frac{d}{dy} v(h, \delta k) \\ &= k \left(\frac{d}{dy} f(a+h, b+\delta k) - \frac{d}{dy} f(a, b+\delta k) \right) \\ &= hk \cdot \frac{d}{dx} \cdot \frac{d}{dy} f(a+\gamma h, b+\delta k) \end{aligned}$$

Now by dividing everything by $hk \neq 0$, we conclude from this that the following equality holds, with the numbers $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ being found as above:

$$\frac{d}{dy} \cdot \frac{d}{dx} f(a+\alpha h, b+\beta k) = \frac{d}{dx} \cdot \frac{d}{dy} f(a+\gamma h, b+\delta k)$$

But with $h, k \to 0$ we get from this the Clairaut formula, at $z = (a, b)$, as desired. □

In arbitrary dimensions now, we have the following result:

THEOREM 16.15. *Given $f : \mathbb{R}^N \to \mathbb{R}$, we can talk about its higher derivatives,*

$$\frac{d^k f}{dx_{i_1} \ldots dx_{i_k}} = \frac{d}{dx_{i_1}} \cdots \frac{d}{dx_{i_k}} (f)$$

provided that these derivatives exist indeed. Moreover, due to the Clairaut formula,

$$\frac{d^2 f}{dx_i dx_j} = \frac{d^2 f}{dx_j dx_i}$$

the order in which these higher derivatives are computed is irrelevant.

PROOF. This is indeed something self-explanatory, based on Theorem 16.14, applied to the various two-variable slices $f_{ij} : \mathbb{R}^2 \to \mathbb{R}$ of our function $f : \mathbb{R}^N \to \mathbb{R}$. □

All this is very nice, and as an illustration, let us work out in detail the case $k = 2$. Here things are quite special, and we can formulate the following definition:

DEFINITION 16.16. *Given a twice differentiable function $f : \mathbb{R}^N \to \mathbb{R}$, we set*

$$f''(x) = \left(\frac{d^2 f}{dx_i dx_j} \right)_{ij}$$

which is a symmetric matrix, called Hessian matrix of f at the point $x \in \mathbb{R}^N$.

To be more precise, we know that when $f : \mathbb{R}^N \to \mathbb{R}$ is twice differentiable, its order $k = 2$ partial derivatives are the numbers in the statement. Now since these numbers

naturally form a $N \times N$ matrix, the temptation is high to call this matrix $f''(x)$, and so we will do. And finally, we know from Clairaut that this matrix is symmetric:

$$f''(x)_{ij} = f''(x)_{ji}$$

Observe that at $N = 1$ this is compatible with the usual definition of the second derivative f'', because in this case, the 1×1 matrix from Definition 16.16 is:

$$f''(x) = (f''(x)) \in M_{1\times 1}(\mathbb{R})$$

As a word of warning, however, never use Definition 16.16 for functions $f : \mathbb{R}^N \to \mathbb{R}^M$, where the second derivative can only be something more complicated. Also, never attempt either to do something similar at $k = 3$ or higher, for functions $f : \mathbb{R}^N \to \mathbb{R}$ with $N > 1$, because again, that beast has too many indices, for being a true, honest matrix.

Back now to our usual business, approximation, we have the following result:

THEOREM 16.17. *Given a twice differentiable function $f : \mathbb{R}^N \to \mathbb{R}$, we have*

$$f(x + t) \simeq f(x) + f'(x)t + \frac{< f''(x)t, t >}{2}$$

where $f''(x) \in M_N(\mathbb{R})$ stands as usual for the Hessian matrix.

PROOF. This is something very standard, the idea being as follows:

(1) As a first observation, at $N = 1$ the Hessian matrix as constructed in Definition 16.16 is the 1×1 matrix having as entry the second derivative $f''(x)$, and the formula in the statement is something that we know well from basic calculus, namely:

$$f(x + t) \simeq f(x) + f'(x)t + \frac{f''(x)t^2}{2}$$

(2) In general now, this is in fact something which does not need a new proof, because it follows from the one-variable formula above, applied to the restriction of f to the following segment in $\mathbb{R}^N$, which can be regarded as being a one-variable interval:

$$I = [x, x + t]$$

To be more precise, let $y \in \mathbb{R}^N$, and consider the following function, with $r \in \mathbb{R}$:

$$g(r) = f(x + ry)$$

We know from (1) that the Taylor formula for g, at the point $r = 0$, reads:

$$g(r) \simeq g(0) + g'(0)r + \frac{g''(0)r^2}{2}$$

And our claim is that, with $t = ry$, this is precisely the formula in the statement.

(3) So, let us see if our claim is correct. By using the chain rule, we have the following formula, with on the right, as usual, a row vector multiplied by a column vector:

$$g'(r) = f'(x + ry) \cdot y$$

By using again the chain rule, we can compute the second derivative as well:

$$\begin{aligned}
g''(r) &= (f'(x+ry)\cdot y)' \\
&= \left(\sum_i \frac{df}{dx_i}(x+ry)\cdot y_i\right)' \\
&= \sum_i\sum_j \frac{d^2f}{dx_idx_j}(x+ry)\cdot\frac{d(x+ry)_j}{dr}\cdot y_i \\
&= \sum_i\sum_j \frac{d^2f}{dx_idx_j}(x+ry)\cdot y_iy_j \\
&= < f''(x+ry)y, y >
\end{aligned}$$

(4) Time now to conclude. We know that we have $g(r) = f(x+ry)$, and according to our various computations above, we have the following formulae:

$$g(0) = f(x) \quad , \quad g'(0) = f'(x) \quad , \quad g''(0) =< f''(x)y, y >$$

Buit with this data in hand, the usual Taylor formula for our one variable function g, at order 2, at the point $r = 0$, takes the following form, with $t = ry$:

$$\begin{aligned}
f(x+ry) &\simeq f(x) + f'(x)ry + \frac{< f''(x)y, y > r^2}{2} \\
&= f(x) + f'(x)t + \frac{< f''(x)t, t >}{2}
\end{aligned}$$

Thus, we have obtained the formula in the statement. □

As in the one variable case, the Taylor formula is useful for computing the local extrema of the function. Indeed, let us first look at the order 1 formula, namely:

$$f(x+t) \simeq f(x) + f'(x)t$$

It is clear then, exactly as in the one-variable case, that in order to have a local extremum, we must have $f'(x) = 0$. Next, assuming that this holds, let us look at the order 2 Taylor formula, which in the case $f'(x) = 0$ takes the following form:

$$f(x+t) \simeq f(x) + \frac{< f''(x)t, t >}{2}$$

We conclude from this, again as in the one-variable case, that when $f''(x) > 0$, with this meaning that the symmetric matrix $f''(x) \in M_N(\mathbb{R})$ has eigenvalues $\lambda_1, \ldots, \lambda_N > 0$, we have a local minimum, and that when $f''(x) < 0$, we have a local maximum.

At higher order now, things become more complicated, as follows:

THEOREM 16.18. *Given an order k differentiable function $f:\mathbb{R}^N\to\mathbb{R}$, we have*

$$f(x+t)\simeq f(x)+f'(x)t+\frac{<f''(x)t,t>}{2}+\ldots$$

and this helps in identifying the local extrema, when $f'(x)=0$ and $f''(x)=0$.

PROOF. The study here is very similar to that at $k=2$, from the proof of Theorem 16.17, with everything coming from the usual Taylor formula, applied on:

$$I=[x,x+t]$$

Thus, it is pretty much clear that we are led to the conclusion in the statement. We will leave some study here as an instructive exercise. □

And with this, end of our discussion regarding the foundations of multivariable calculus. Which was actually far simpler than originally expected, hope we agree on this.

16c. Kepler and Newton

Time now for some 3D physics and applications, you would say, but before that, we have some unfinished business in 1D, in relation with the waves. And here, we have:

THEOREM 16.19. *The solution of the 1D wave equation $\ddot{\varphi}=v^2\varphi''$ with initial value conditions $\varphi(x,0)=f(x)$ and $\dot{\varphi}(x,0)=g(x)$ is given by the d'Alembert formula:*

$$\varphi(x,t)=\frac{f(x-vt)+f(x+vt)}{2}+\frac{1}{2v}\int_{x-vt}^{x+vt}g(s)ds$$

Also, in the context of our previous lattice model discretizations, what happens is more or less that the above d'Alembert integral gets computed via Riemann sums.

PROOF. We already talked about waves in this book, in chapters 10, 12, 14, 15, but the above formula is still to be proved. Let us make the following change of variables:

$$y=x-vt\quad,\quad z=x+vt$$

We have the following computation, using the chain rule from Theorem 16.13:

$$\frac{d\varphi}{dt}=\frac{d\varphi}{dy}\cdot\frac{dy}{dt}+\frac{d\varphi}{dz}\cdot\frac{dz}{dt}=-v\frac{d\varphi}{dy}+v\frac{d\varphi}{dz}$$

By using the chain rule again, the second time derivative is given by:

$$\begin{aligned}\frac{d^2\varphi}{dt^2}&=-v\left(\frac{d^2\varphi}{dy^2}\cdot\frac{dy}{dt}+\frac{d^2\varphi}{dydz}\cdot\frac{dz}{dt}\right)+v\left(\frac{d^2\varphi}{dzdy}\cdot\frac{dy}{dt}+\frac{d^2\varphi}{dz^2}\cdot\frac{dz}{dt}\right)\\&=-v\left(-v\frac{d^2\varphi}{dy^2}+v\frac{d^2\varphi}{dydz}\right)+v\left(-v\frac{d^2\varphi}{dzdy}+v\frac{d^2\varphi}{dz^2}\right)\\&=v^2\left(\frac{d^2\varphi}{dy^2}+\frac{d^2\varphi}{dz^2}-2\frac{d^2\varphi}{dydz}\right)\end{aligned}$$

Regarding now the first space derivative, this is given by the following formula:

$$\frac{d\varphi}{dx}=\frac{d\varphi}{dy}\cdot\frac{dy}{dx}+\frac{d\varphi}{dz}\cdot\frac{dz}{dx}=\frac{d\varphi}{dy}+\frac{d\varphi}{dz}$$

By using the chain rule again, the second space derivative is given by:

$$\begin{aligned}\frac{d^2\varphi}{dt^2}&=\left(\frac{d^2\varphi}{dy^2}\cdot\frac{dy}{dx}+\frac{d^2\varphi}{dydz}\cdot\frac{dz}{dx}\right)+\left(\frac{d^2\varphi}{dzd\xi}\cdot\frac{dy}{dx}+\frac{d^2\varphi}{dz^2}\cdot\frac{dz}{dx}\right)\\&=\left(\frac{d^2\varphi}{dy^2}+\frac{d^2\varphi}{dydz}\right)+\left(\frac{d^2\varphi}{dzdy}+\frac{d^2\varphi}{dz^2}\right)\\&=\frac{d^2\varphi}{dy^2}+\frac{d^2\varphi}{dz^2}+2\frac{d^2\varphi}{dydz}\end{aligned}$$

Thus, our wave equation $\ddot{\varphi}=v^2\varphi''$ reformulates in a very simple way, as follows:

$$\frac{d^2\varphi}{dydz}=0$$

But this latter equation tells us that our new y,z variables get separated, and we conclude from this that the solution must be of the following special form:

$$\varphi(x,t)=F(y)+G(z)=F(x-vt)+G(x+vt)$$

Now by taking into account the intial conditions $\varphi(x,0)=f(x)$ and $\dot{\varphi}(x,0)=g(x)$, and then integrating, we are led to the d'Alembert formula in the statement. □

Getting now to the real thing, gravity and celestial mechanics, we have here:

THEOREM 16.20 (Kepler, Newton). *Planets and other celestial bodies move around the Sun on conics, that is, on curves of type*

$$C=\left\{(x,y)\in\mathbb{R}^2\Big|P(x,y)=0\right\}$$

with $P\in\mathbb{R}[x,y]$ being of degree 2. The same is true for any body moving around another body, provided that we are not in the situation of a free fall.

PROOF. This is something very standard, the idea being as follows:

(1) According to observations and calculations performed over the centuries, since the ancient times, and first formalized by Newton, following some groundbreaking work of Kepler, the force of attraction between two bodies of masses M,m is given by:

$$||F||=G\cdot\frac{Mm}{d^2}$$

Here d is the distance between the two bodies, and $G\simeq 6.674\times 10^{-11}$ is a constant. Now assuming that M is fixed at $0\in\mathbb{R}^3$, the force exterted on m positioned at $x\in\mathbb{R}^3$,

regarded as a vector $F \in \mathbb{R}^3$, is given by the following formula:

$$F = -||F|| \cdot \frac{x}{||x||} = -\frac{GMm}{||x||^2} \cdot \frac{x}{||x||} = -\frac{GMmx}{||x||^3}$$

But $F = ma = m\ddot{x}$, with $a = \ddot{x}$ being the acceleration, second derivative of the position, so the equation of motion of m, assuming that M is fixed at 0, is:

$$\ddot{x} = -\frac{GMx}{||x||^3}$$

(2) Obviously, the problem happens in 2 dimensions, and here the most convenient is to use standard x, y coordinates, and denote our point as $z = (x, y)$. With this change made, and by setting $K = GM$, the equation of motion becomes:

$$\ddot{z} = -\frac{Kz}{||z||^3}$$

In other words, in terms of the coordinates x, y, the equations are:

$$\ddot{x} = -\frac{Kx}{(x^2 + y^2)^{3/2}} \quad , \quad \ddot{y} = -\frac{Ky}{(x^2 + y^2)^{3/2}}$$

(3) Let us begin with a simple particular case, that of the circular solutions. To be more precise, we are interested in solutions of the following type:

$$x = r\cos\alpha t \quad , \quad y = r\sin\alpha t$$

In this case we have $||z|| = r$, so our equation of motion becomes:

$$\ddot{z} = -\frac{Kz}{r^3}$$

On the other hand, differentiating x, y leads to the following formula:

$$\ddot{z} = (\ddot{x}, \ddot{y}) = -\alpha^2(x, y) = -\alpha^2 z$$

Thus, we have a circular solution when the parameters r, α satisfy:

$$r^3\alpha^2 = K$$

(4) In the general case now, the problem can be solved via some calculus. Let us write indeed our vector $z = (x, y)$ in polar coordinates, as follows:

$$x = r\cos\theta \quad , \quad y = r\sin\theta$$

We have then $||z|| = r$, and our equation of motion becomes, as in (3):

$$\ddot{z} = -\frac{Kz}{r^3}$$

Let us differentiate now x, y. By using the standard calculus rules, we have:

$$\dot{x} = \dot{r}\cos\theta - r\sin\theta \cdot \dot{\theta}$$

$$\dot{y} = \dot{r}\sin\theta + r\cos\theta \cdot \dot{\theta}$$

Differentiating one more time gives the following formulae:

$$\ddot{x} = \ddot{r}\cos\theta - 2\dot{r}\sin\theta\cdot\dot{\theta} - r\cos\theta\cdot\dot{\theta}^2 - r\sin\theta\cdot\ddot{\theta}$$

$$\ddot{y} = \ddot{r}\sin\theta + 2\dot{r}\cos\theta\cdot\dot{\theta} - r\sin\theta\cdot\dot{\theta}^2 + r\cos\theta\cdot\ddot{\theta}$$

Consider now the following two quantities, appearing as coefficients in the above:

$$a = \ddot{r} - r\dot{\theta}^2 \quad , \quad b = 2\dot{r}\dot{\theta} + r\ddot{\theta}$$

In terms of these quantities, our second derivative formulae read:

$$\ddot{x} = a\cos\theta - b\sin\theta$$

$$\ddot{y} = a\sin\theta + b\cos\theta$$

(5) We can now solve the equation of motion from (4). Indeed, with the formulae that we found for $\ddot{x}, \ddot{y}$, our equation of motion takes the following form:

$$a\cos\theta - b\sin\theta = -\frac{K}{r^2}\cos\theta$$

$$a\sin\theta + b\cos\theta = -\frac{K}{r^2}\sin\theta$$

But these two formulae can be written in the following way:

$$\left(a + \frac{K}{r^2}\right)\cos\theta = b\sin\theta$$

$$\left(a + \frac{K}{r^2}\right)\sin\theta = -b\cos\theta$$

By making now the product, and assuming that we are in a non-degenerate case, where the angle θ varies indeed, we obtain by positivity that we must have:

$$a + \frac{K}{r^2} = b = 0$$

(6) We are almost there. Let us first examine the second equation, $b = 0$. Remembering who b is, from (4), this equation can be solved as follows:

$$\begin{aligned}
b = 0 &\iff 2\dot{r}\dot{\theta} + r\ddot{\theta} = 0\\
&\iff \frac{\ddot{\theta}}{\dot{\theta}} = -2\frac{\dot{r}}{r}\\
&\iff (\log\dot{\theta})' = (-2\log r)'\\
&\iff \log\dot{\theta} = -2\log r + c\\
&\iff \dot{\theta} = \frac{\lambda}{r^2}
\end{aligned}$$

As for the first equation the we found, namely $a + K/r^2 = 0$, remembering from (4) that a was by definition given by $a = \ddot{r} - r\dot{\theta}^2$, this equation now becomes:

$$\ddot{r} - \frac{\lambda^2}{r^3} + \frac{K}{r^2} = 0$$

(7) As a conclusion to all this, in polar coordinates, $x = r\cos\theta$, $y = r\sin\theta$, our equations of motion are as follows, with λ being a constant, not depending on t:

$$\ddot{r} = \frac{\lambda^2}{r^3} - \frac{K}{r^2} \quad , \quad \dot{\theta} = \frac{\lambda}{r^2}$$

Even better now, by writing $K = \lambda^2/c$, these equations read:

$$\ddot{r} = \frac{\lambda^2}{r^2}\left(\frac{1}{r} - \frac{1}{c}\right) \quad , \quad \dot{\theta} = \frac{\lambda}{r^2}$$

(8) As an illustration, let us quickly work out the case of a circular motion, where r is constant. Here $\ddot{r} = 0$, so the first equation gives $c = r$. Also we have $\dot{\theta} = \alpha$, with:

$$\alpha = \frac{\lambda}{r^2}$$

Assuming $\theta = 0$ at $t = 0$, from $\dot{\theta} = \alpha$ we obtain $\theta = \alpha t$, and so, as in (3) above:

$$x = r\cos\alpha t \quad , \quad y = r\sin\alpha t$$

Observe also that the condition found in (3) is indeed satisfied:

$$r^3\alpha^2 = \frac{\lambda^2}{r} = \frac{\lambda^2}{c} = K$$

(9) Back to the general case now, our claim is that we have the following formula, for the distance $r = r(t)$ as function of the angle $\theta = \theta(t)$, for some $\varepsilon, \delta \in \mathbb{R}$:

$$r = \frac{c}{1 + \varepsilon\cos\theta + \delta\sin\theta}$$

Let us first check that this formula works indeed. With r being as above, and by using our second equation found before, $\dot{\theta} = \lambda/r^2$, we have the following computation:

$$\begin{aligned}
\dot{r} &= \frac{c(\varepsilon\sin\theta - \delta\cos\theta)\dot{\theta}}{(1 + \varepsilon\cos\theta + \delta\sin\theta)^2} \\
&= \frac{\lambda c(\varepsilon\sin\theta - \delta\cos\theta)}{r^2(1 + \varepsilon\cos\theta + \delta\sin\theta)^2} \\
&= \frac{\lambda(\varepsilon\sin\theta - \delta\cos\theta)}{c}
\end{aligned}$$

Thus, the second derivative of the above function r is given, as desired, by:

$$\begin{aligned}\ddot{r} &= \frac{\lambda(\varepsilon\cos\theta+\delta\sin\theta)\dot{\theta}}{c}\\ &= \frac{\lambda^2(\varepsilon\cos\theta+\delta\sin\theta)}{r^2c}\\ &= \frac{\lambda^2}{r^2}\left(\frac{1}{r}-\frac{1}{c}\right)\end{aligned}$$

(10) The above check was something quite informal, and now we must prove that our formula is indeed the correct one. For this purpose, we use a trick. Let us write:

$$r(t)=\frac{1}{f(\theta(t))}$$

Abbreviated, and by always reminding that f takes $\theta=\theta(t)$ as variable, this reads:

$$r=\frac{1}{f}$$

With the convention that dots mean as usual derivatives with respect to t, and that the primes will denote derivatives with respect to $\theta=\theta(t)$, we have:

$$\dot{r}=-\frac{f'\dot{\theta}}{f^2}=-\frac{f'}{f^2}\cdot\frac{\lambda}{r^2}=-\lambda f'$$

By differentiating one more time with respect to t, we obtain:

$$\ddot{r}=-\lambda f''\dot{\theta}=-\lambda f''\cdot\frac{\lambda}{r^2}=-\frac{\lambda^2}{r^2}f''$$

On the other hand, our equation for $\ddot{r}$ found in (7) reads:

$$\ddot{r}=\frac{\lambda^2}{r^2}\left(\frac{1}{r}-\frac{1}{c}\right)=\frac{\lambda^2}{r^2}\left(f-\frac{1}{c}\right)$$

Thus, in terms of $f=1/r$ as above, our equation for $\ddot{r}$ simply reads:

$$f''+f=\frac{1}{c}$$

But this latter equation is elementary to solve. Indeed, both functions $\cos t,\sin t$ satisfy $g"+g=0$, so any linear combination of them satisfies as well this equation. But the solutions of $f''+f=1/c$ being those of $g''+g=0$ shifted by $1/c$, we obtain:

$$f=\frac{1+\varepsilon\cos\theta+\delta\sin\theta}{c}$$

Now by inverting, we obtain the formula announced in (9), namely:

$$r=\frac{c}{1+\varepsilon\cos\theta+\delta\sin\theta}$$

(11) But this leads to the conclusion that the trajectory is a conic. Indeed, in terms of the parameter θ, the formulae of the coordinates are:

$$x = \frac{c\cos\theta}{1+\varepsilon\cos\theta+\delta\sin\theta}$$

$$y = \frac{c\sin\theta}{1+\varepsilon\cos\theta+\delta\sin\theta}$$

But these are precisely the equations of conics in polar coordinates.

(12) To be more precise, in order to find the precise equation of the conic, observe that the two functions x, y that we found above satisfy the following formula:

$$\begin{aligned} x^2+y^2 &= \frac{c^2(\cos^2\theta+\sin^2\theta)}{(1+\varepsilon\cos\theta+\delta\sin\theta)^2} \\ &= \frac{c^2}{(1+\varepsilon\cos\theta+\delta\sin\theta)^2} \end{aligned}$$

On the other hand, these two functions satisfy as well the following formula:

$$\begin{aligned} (\varepsilon x+\delta y-c)^2 &= \frac{c^2\big(\varepsilon\cos\theta+\delta\sin\theta-(1+\varepsilon\cos\theta+\delta\sin\theta)\big)^2}{(1+\varepsilon\cos\theta+\delta\sin\theta)^2} \\ &= \frac{c^2}{(1+\varepsilon\cos\theta+\delta\sin\theta)^2} \end{aligned}$$

We conclude that our coordinates x, y satisfy the following equation:

$$x^2+y^2=(\varepsilon x+\delta y-c)^2$$

But what we have here is an equation of a conic, as claimed. □

16d. Spherical integrals

Back to mathematics, we can talk about multiple integrals, in the obvious way. Getting now to the general theory and rules, for computing such integrals, the key result here is the change of variable formula, which is something quite tricky, as follows:

THEOREM 16.21. *Given a transformation $\varphi=(\varphi_1,\ldots,\varphi_N)$, we have*

$$\int_E f(x)dx=\int_{\varphi^{-1}(E)} f(\varphi(t))|J_\varphi(t)|dt$$

with the J_φ quantity, called Jacobian, being given by

$$J_\varphi(t)=\det\left[\left(\frac{d\varphi_i}{dx_j}(x)\right)_{ij}\right]$$

and with this generalizing the usual formula from one-variable calculus.

PROOF. This is something quite tricky, the idea being as follows:

(1) Observe first that this generalizes indeed the change of variable formula in 1 dimension, from chapter 13, the point here being that the absolute value on the derivative appears as to compensate for the lack of explicit bounds for the integral.

(2) In general now, we can first argue that, the formula in the statement being linear in f, we can assume $f = 1$. Thus we want to prove $vol(E) = \int_{\varphi^{-1}(E)} |J_\varphi(t)|dt$, and with $D = \varphi^{-1}(E)$, this amounts in proving $vol(\varphi(D)) = \int_D |J_\varphi(t)|dt$.

(3) Now since this latter formula is additive with respect to D, it is enough to prove that $vol(\varphi(D)) = \int_D J_\varphi(t)dt$, for small cubes D, and assuming $J_\varphi > 0$. But this follows by using the usual definition of the determinant, as a volume.

(4) The details and computations however are quite non-trivial, and can be found for instance in Rudin [73]. So, please read that. With this, reading the complete proof of the present theorem from Rudin, being part of the standard math experience. □

As an application, we can compute the arbitrary spherical integrals, as follows:

THEOREM 16.22. *The polynomial integrals over the unit sphere $S_{\mathbb{R}}^{N-1} \subset \mathbb{R}^N$, with respect to the normalized, mass 1 measure, are given by the following formula,*

$$\int_{S_{\mathbb{R}}^{N-1}} x_1^{k_1} \ldots x_N^{k_N}\, dx = \frac{(N-1)!!k_1!! \ldots k_N!!}{(N+\Sigma k_i - 1)!!}$$

valid when all exponents k_i are even. If an exponent k_i is odd, the integral vanishes.

PROOF. Assume first that one of the exponents k_i is odd. We can make then the following change of variables, which shows that the integral in the statement vanishes:

$$x_i \to -x_i$$

Assume now that all exponents k_i are even. As a first observation, the result holds at $N = 2$, due to the Wallis formula from chapter 14. In the general case now, we have:

$$I = \frac{2^N}{A} \int_0^{\pi/2} \ldots \int_0^{\pi/2} x_1^{k_1} \ldots x_N^{k_N} J\, dt_1 \ldots dt_{N-1}$$

Here A is the area of the sphere, computed in chapter 14, and given by:

$$\frac{2^N}{A} = \left(\frac{2}{\pi}\right)^{[N/2]} (N-1)!!$$

As for the unnormalized integral, this is given by the following formula:

$$\begin{aligned} I' = \int_0^{\pi/2} \ldots \int_0^{\pi/2} \quad & (\cos t_1)^{k_1} (\sin t_1 \cos t_2)^{k_2} \\ & \vdots \\ & (\sin t_1 \sin t_2 \ldots \sin t_{N-2} \cos t_{N-1})^{k_{N-1}} \\ & (\sin t_1 \sin t_2 \ldots \sin t_{N-2} \sin t_{N-1})^{k_N} \\ & \sin^{N-2} t_1 \sin^{N-3} t_2 \ldots \sin^2 t_{N-3} \sin t_{N-2} \\ & dt_1 \ldots dt_{N-1} \end{aligned}$$

By rearranging the terms, we obtain:

$$\begin{aligned} I' &= \int_0^{\pi/2} \cos^{k_1} t_1 \sin^{k_2+\ldots+k_N+N-2} t_1 \, dt_1 \\ &\quad \int_0^{\pi/2} \cos^{k_2} t_2 \sin^{k_3+\ldots+k_N+N-3} t_2 \, dt_2 \\ &\quad \vdots \\ &\quad \int_0^{\pi/2} \cos^{k_{N-2}} t_{N-2} \sin^{k_{N-1}+k_N+1} t_{N-2} \, dt_{N-2} \\ &\quad \int_0^{\pi/2} \cos^{k_{N-1}} t_{N-1} \sin^{k_N} t_{N-1} \, dt_{N-1} \end{aligned}$$

Now by using the Wallis formula at $N = 2$, this gives:

$$\begin{aligned} I' &= \frac{k_1!!(k_2 + \ldots + k_N + N - 2)!!}{(k_1 + \ldots + k_N + N - 1)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(N-2)} \\ &\quad \frac{k_2!!(k_3 + \ldots + k_N + N - 3)!!}{(k_2 + \ldots + k_N + N - 2)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(N-3)} \\ &\quad \vdots \\ &\quad \frac{k_{N-2}!!(k_{N-1} + k_N + 1)!!}{(k_{N-2} + k_{N-1} + l_N + 2)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(1)} \\ &\quad \frac{k_{N-1}!!k_N!!}{(k_{N-1} + k_N + 1)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(0)} \end{aligned}$$

But this gives, after simplifying, the formula in the statement. □

As an interesting application of the above formula, we have:

THEOREM 16.23. *The moments of the hyperspherical variables are*

$$\int_{S_{\mathbb{R}}^{N-1}} x_i^p dx = \frac{(N-1)!!p!!}{(N+p-1)!!}$$

and the rescaled variables $y_i = \sqrt{N}x_i$ become normal and independent with $N \to \infty$.

PROOF. We have two assertions here, the idea being as follows:

(1) The moment formula in the statement follows from the general formula from Theorem 16.22. As a consequence, with $N \to \infty$ we have the following estimate:

$$\int_{S_{\mathbb{R}}^{N-1}} x_i^p dx \simeq N^{-p/2} \times p!!$$

By comparing with the moment formula for normal variables, from chapter 14, we conclude that the rescaled variables $\sqrt{N}x_i$ become normal with $N \to \infty$, as claimed.

(2) As for the proof of the asymptotic independence, this is standard too, once again by using Theorem 16.22. Indeed, the joint moments of $x_1, \ldots, x_N$ are given by:

$$\begin{aligned}\int_{S_{\mathbb{R}}^{N-1}} x_1^{k_1} \ldots x_N^{k_N}\, dx &= \frac{(N-1)!!k_1!! \ldots k_N!!}{(N+\Sigma k_i-1)!!}\\ &\simeq N^{-\Sigma k_i} \times k_1!! \ldots k_N!!\end{aligned}$$

By rescaling, the joint moments of the variables $y_i = \sqrt{N}x_i$ are given by:

$$\int_{S_{\mathbb{R}}^{N-1}} y_1^{k_1} \ldots y_N^{k_N}\, dx \simeq k_1!! \ldots k_N!!$$

Thus, we have multiplicativity, and so independence with $N \to \infty$, as claimed. □

Many other things can be said, as a continuation of this, and for more, you can have a look at any advanced book mixing geometry, probability and physics.

16e. Exercises

Congratulations for having read this book, and no exercises for this final chapter. However, if interested to learn more about functions, and especially about the multivariable ones, which are quite tricky, there are plently of more advanced calculus books, waiting for you. So, have a look at the various books referenced below, pick one, and enjoy.

Bibliography

[1] V.I. Arnold, Ordinary differential equations, Springer (1973).

[2] V.I. Arnold, Mathematical methods of classical mechanics, Springer (1974).

[3] V.I. Arnold, Catastrophe theory, Springer (1984).

[4] V.I. Arnold, Lectures on partial differential equations, Springer (1997).

[5] V.I. Arnold and B.A. Khesin, Topological methods in hydrodynamics, Springer (1998).

[6] M.F. Atiyah, K-theory, CRC Press (1964).

[7] M.F. Atiyah, The geometry and physics of knots, Cambridge Univ. Press (1990).

[8] M.F. Atiyah and I.G. MacDonald, Introduction to commutative algebra, Addison-Wesley (1969).

[9] T. Banica, Calculus and applications (2025).

[10] T. Banica, Measure and integration (2025).

[11] T. Banica, Introduction to modern physics (2025).

[12] R.J. Baxter, Exactly solved models in statistical mechanics, Academic Press (1982).

[13] S.J. Blundell and K.M. Blundell, Concepts in thermal physics, Oxford Univ. Press (2006).

[14] B. Bollobás, Modern graph theory, Springer (1998).

[15] S.M. Carroll, Spacetime and geometry, Cambridge Univ. Press (2004).

[16] A.R. Choudhuri, Astrophysics for physicists, Cambridge Univ. Press (2012).

[17] D.D. Clayton, Principles of stellar evolution and nucleosynthesis, Univ. of Chicago Press (1968).

[18] A. Connes, Noncommutative geometry, Academic Press (1994).

[19] J.B. Conway, A course in functional analysis, Springer (1985).

[20] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).

[21] P.A. Davidson, Introduction to magnetohydrodynamics, Cambridge Univ. Press (2001).

[22] P.A.M. Dirac, Principles of quantum mechanics, Oxford Univ. Press (1930).

[23] M.P. do Carmo, Differential geometry of curves and surfaces, Dover (1976).

[24] M.P. do Carmo, Riemannian geometry, Birkhäuser (1992).

[25] S. Dodelson, Modern cosmology, Academic Press (2003).

[26] R. Durrett, Probability: theory and examples, Cambridge Univ. Press (1990).

[27] A. Einstein, Relativity: the special and the general theory, Dover (1916).

[28] L.C. Evans, Partial differential equations, AMS (1998).

[29] W. Feller, An introduction to probability theory and its applications, Wiley (1950).

[30] E. Fermi, Thermodynamics, Dover (1937).

[31] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics I: mainly mechanics, radiation and heat, Caltech (1963).

[32] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics II: mainly electromagnetism and matter, Caltech (1964).

[33] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics III: quantum mechanics, Caltech (1966).

[34] R.P. Feynman and A.R. Hibbs, Quantum mechanics and path integrals, Dover (1965).

[35] P. Flajolet and R. Sedgewick, Analytic combinatorics, Cambridge Univ. Press (2009).

[36] A.P. French, Special relativity, Taylor and Francis (1968).

[37] W. Fulton, Algebraic topology, Springer (1995).

[38] W. Fulton and J. Harris, Representation theory, Springer (1991).

[39] C. Godsil and G. Royle, Algebraic graph theory, Springer (2001).

[40] H. Goldstein, C. Safko and J. Poole, Classical mechanics, Addison-Wesley (1980).

[41] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).

[42] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).

[43] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).

[44] D.J. Griffiths, Revolutions in twentieth-century physics, Cambridge Univ. Press (2012).

[45] J. Harris, Algebraic geometry, Springer (1992).

[46] R.A. Horn and C.R. Johnson, Matrix analysis, Cambridge Univ. Press (1985).

[47] K. Huang, Introduction to statistical physics, CRC Press (2001).

[48] K. Huang, Quantum field theory, Wiley (1998).

[49] K. Huang, Quarks, leptons and gauge fields, World Scientific (1982).

[50] K. Huang, Fundamental forces of nature, World Scientific (2007).

[51] J.E. Humphreys, Introduction to Lie algebras and representation theory, Springer (1972).

[52] V.F.R. Jones, Subfactors and knots, AMS (1991).

[53] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).

[54] C. Kittel, Introduction to solid state physics, Wiley (1953).

[55] M. Kumar, Quantum: Einstein, Bohr, and the great debate about the nature of reality, Norton (2009).

[56] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).

[57] L.D. Landau and E.M. Lifshitz, Mechanics, Pergamon Press (1960).

[58] L.D. Landau and E.M. Lifshitz, The classical theory of fields, Addison-Wesley (1951).

[59] L.D. Landau and E.M. Lifshitz, Quantum mechanics: non-relativistic theory, Pergamon Press (1959).

[60] S. Lang, Algebra, Addison-Wesley (1993).

[61] P. Lax, Linear algebra and its applications, Wiley (2007).

[62] P. Lax, Functional analysis, Wiley (2002).

[63] P. Lax and M.S. Terrell, Calculus with applications, Springer (2013).

[64] P. Lax and M.S. Terrell, Multivariable calculus with applications, Springer (2018).

[65] J.M. Lee, Introduction to topological manifolds, Springer (2011).

[66] J.M. Lee, Introduction to smooth manifolds, Springer (2012).

[67] J.M. Lee, Introduction to Riemannian manifolds, Springer (2018).

[68] M.L. Mehta, Random matrices, Elsevier (2004).

[69] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge Univ. Press (2000).

[70] R.K. Pathria and and P.D. Beale, Statistical mechanics, Elsevier (1972).

[71] P. Petersen, Linear algebra, Springer (2012).

[72] P. Petersen, Riemannian geometry, Springer (1998).

[73] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[74] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[75] W. Rudin, Functional analysis, McGraw-Hill (1973).

[76] W. Rudin, Fourier analysis on groups, Dover (1974).

[77] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[78] B. Ryden and B.M. Peterson, Foundations of astrophysics, Cambridge Univ. Press (2010).

[79] B. Ryden and R.W. Pogge, Interstellar and intergalactic medium, Cambridge Univ. Press (2021).

[80] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).

[81] J.P. Serre, A course in arithmetic, Springer (1973).

[82] J.P. Serre, Linear representations of finite groups, Springer (1977).

[83] I.R. Shafarevich, Basic algebraic geometry, Springer (1974).

[84] R. Shankar, Fundamentals of physics I: mechanics, relativity, and thermodynamics, Yale Univ. Press (2014).

[85] R. Shankar, Fundamentals of physics II: electromagnetism, optics, and quantum mechanics, Yale Univ. Press (2016).

[86] R. Shankar, Principles of quantum mechanics, Springer (1980).

[87] R. Shankar, Quantum field theory and condensed matter: an introduction, Cambridge Univ. Press (2017).

[88] A.M. Steane, Thermodynamics, Oxford Univ. Press (2016).

[89] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[90] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[91] J. von Neumann and O. Morgenstern, Theory of games and economic behavior, Princeton Univ. Press (1944).

[92] J. Watrous, The theory of quantum information, Cambridge Univ. Press (2018).

[93] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[94] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[95] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).

[96] S. Weinberg, Cosmology, Oxford Univ. Press (2008).

[97] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).

[98] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).

[99] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).

[100] B. Zwiebach, A first course in string theory, Cambridge Univ. Press (2004).

Index

acceleration, 225
all-one matrix, 373
alternating series, 30, 32
altitudes, 57
angle bisectors, 57
approximation, 74
approximation by polynomials, 146
arctan, 213
area, 299
area below graph, 299
area of circle, 69
area of ellipsoid, 319
area of sphere, 319, 332
argument, 94
argument of complex number, 76
attracting mass, 292
average of function, 302

Babylonian method, 129
Banach theorem, 132
barycenter, 57
Bernoulli law, 144
Bernoulli laws, 337
Bernstein polynomial, 146
biased coin, 145
binomial coefficient, 12
binomial coefficients, 14
binomial formula, 13, 153, 261
binomial law, 145

Cardano formula, 192, 194, 196
Catalan numbers, 156, 158
Cauchy criterion, 148
Cauchy sequence, 26
Cauchy-Schwarz, 142, 245
central binomial coefficients, 158
chain rule, 209, 315, 378
change of variable, 315, 389
character, 102
characteristic polynomial, 372
Chebycheff inequality, 146
Chebycheff polynomials, 170
circumcenter, 57
Clairaut formula, 379
closed and bounded, 118
closed set, 113, 114
common roots, 183
compact set, 117, 123
complement, 115
complete space, 26
complex conjugate, 52
complex cosine, 150
complex eigenvalues, 373
complex exponential, 90
complex logarithm, 128
complex number, 50, 51
complex plane, 124
complex roots, 79, 188
complex sine, 150
concave, 225
concave function, 244
conjugation, 53
connected set, 117
continuous and bounded, 143
continuous function, 39, 116
continuously differentiable, 376
convergence radius, 148
convergent sequence, 23, 124
convergent series, 28
convex, 225
convex function, 244

convex set, 127
convolution, 145, 325, 336
cos, 63, 207, 259
cosecant, 160
cosine, 63
cosine of sum, 75
cotangent, 160
cover, 117, 123
critical damping, 290, 291
cut, 17

d'Alembert criterion, 148
d'Alembert formula, 322, 383
damped oscillator, 289
decimal form, 18
Dedekind cut, 17
degree 2 equation, 18, 51, 178
degree 2 polynomial, 188
degree 3 equation, 192, 194
degree 3 polynomial, 188, 192
degree 4 equation, 194, 196
degree 4 polynomial, 194
derangement, 101
derivative, 204, 205, 375
derivative of arctan, 213
derivative of derivative, 225
derivative of fraction, 210
derivative of inverse, 210
derivative of tan, 211
diagonalization, 372
differentiable function, 204
differential equation, 240, 292
discretization, 322, 383
discriminant, 18, 186
distance, 124
distribution, 335
double factorial, 330
double root, 186, 290
Dyck paths, 156

e, 81, 83, 85
eigenvalue, 372
eigenvector, 372
Einstein formula, 174
ellipsoid, 319
equal almost everywhere, 347
Euclid theorem, 62
Euler formula, 150
exchange of hat, 362
exp, 207, 259
expectation, 304
exponential, 83, 85, 90, 208
exponential series, 83
extremum, 220

factorials, 12
Fermat theorem, 16
fixed point, 129, 132
fixed points, 101
flat matrix, 373
Fourier inversion, 365
Fourier matrix, 373
Fourier transform, 337, 361
fraction, 11, 210
free fall, 292
friction, 289
function, 39
function space, 143, 347
fundamental theorem of calculus, 311, 312

Gegenbauer polynomials, 352
generalized binomial formula, 153, 261
generalized binomial numbers, 153
geometric series, 28, 125, 149
gravity, 240, 292
growth of slope, 225

Hölder inequality, 245, 347
heat equation, 242
heat kernel, 325
Heine-Cantor theorem, 123
Hessian, 381
Hessian matrix, 380
higher derivative, 315, 380
Hooke law, 241
hyperspherical law, 391

i, 50
image of compact set, 118
image of connected set, 118
incenter, 57
inclusion-exclusion, 101
indefinite integral, 313
independence, 145, 336, 337
infinitesimal, 311
integrable function, 300, 305

integral, 299
integration by parts, 314
intermediate value, 43, 119
intersection of closed sets, 114
inverse image, 116
irrational, 87
irrationality of e, 87
iteration, 129

Jacobi polynomials, 352
Jacobian, 389
Jacobian matrix, 375
Jensen inequality, 244
jump, 109

large damping, 290
lattice model, 241, 242
law, 335
law of cosines, 65
left limit, 109
Legendre polynomials, 352
Leibnitz formula, 209
length of circle, 69
length of ellipse, 317
lim inf, 25
lim sup, 25
limit of continuous functions, 140
limit of sequence, 23
limit of series, 28
Lipschitz constant, 119
Lipschitz function, 119
Lipschitz property, 119
local extremum, 220, 239, 382
local maximum, 220, 239, 381
local minimum, 220, 239, 381
locally affine, 205
locally quadratic, 236
log, 207, 259
logarithm, 128, 151
loops on graph, 156

main character, 102
Markov inequality, 146
maximum, 43, 119, 220, 381
mean, 145
mean value property, 221, 306
medians, 57
metric, 141
metric space, 141
minimum, 43, 119, 220, 381
Minkowski inequality, 246, 347
modulus, 53, 94, 204
modulus of complex number, 76
moments, 145, 335
monotone function, 45
Monte Carlo integration, 302
multiplication of complex numbers, 94

Newton law, 241
noncrossing pairings, 156
noncrossing partitions, 156
norm, 141
normed space, 141, 247, 347

open cover, 117
open intervals, 117
open set, 113, 114
orthocenter, 57
orthogonal polynomials, 352
oscillator damping, 290

p-norm, 142, 247, 347
parallelogram rule, 51
partial derivatives, 375
Pascal triangle, 14
path connected, 127
periodic decimal form, 20
permutation, 101
perpendicular bisectors, 57
pi, 69
piecewise continuous, 305
piecewise linear, 300
piecewise monotone, 305
plane rotation, 373
PLT, 337
pointwise convergence, 136, 139
Poisson law, 102, 337
Poisson limit, 102
Poisson Limit Theorem, 337
polar coordinates, 76, 94
polar writing, 93
polynomial approximation, 146
positive matrix, 381
power function, 46, 205, 307
power series, 148, 208, 254
powers of complex number, 78

primitive, 313
probability 0, 20
probability space, 335
purely imaginary, 53
Pythagoras equation, 62
Pythagoras theorem, 60

quotient, 11
quotient of polynomials, 25

random number, 302
random permutation, 101
random variable, 304, 335
rational function, 149
rational number, 11
real numbers, 17
real roots, 188
reflection, 52
remainder, 315
resultant, 183, 185
Riemann integration, 301
Riemann series, 28
Riemann sum, 301, 307, 322, 383
Riemann zeta function, 340
Riemann-Lebesgue property, 364
right angle, 60
right limit, 109
right triangle, 60, 63
root of unity, 193
roots of polynomial, 46, 79
roots of unity, 79, 95
rotation, 373

scalar product, 142
secant, 160
second derivative, 225, 380
second derivatives, 379
second order derivative, 381
sequence, 23
sequence of functions, 136
series, 28
sin, 63, 207, 259
sine, 63
sine of sum, 75
single roots, 186
slope, 204
small damping, 290
space of functions, 143
spectral theorem, 374
speed addition, 172
spherical integral, 390
spiral, 125
square root, 17, 18, 46, 51, 129, 158
step function, 137
subcover, 117
subsequence, 25
sum of cubes, 307
sum of squares, 307
sum of vectors, 51
sun of three angles, 166
symmetric functions, 182
symmetric matrix, 374, 380

tan, 211
tangent of sum, 75
tangent of sums, 168
Taylor formula, 254, 259, 315, 381, 382
totally discontinuous, 137, 138
triangle, 57
triangle inequality, 141
triple angle, 167
truncated character, 102
twice differentiable, 225

undamped frequency, 289
undamped oscillator, 289
uniform continuity, 122
uniform convergence, 138
union of intervals, 117
union of open sets, 114
unit sphere, 319

variance, 145
vector, 51
volume, 299
volume of sphere, 319, 330, 331

Wallis formula, 329, 390
wave equation, 241, 294
Weierstrass basis, 352
Weierstrass theorem, 146

Young inequality, 245

zeta function, 340