Calculus and applications

Teo Banica

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

2010 *Mathematics Subject Classification.* 26A06

Key words and phrases. Calculus, Multivariable calculus

Abstract. This is an introduction to calculus, and its applications to basic questions from physics. We first discuss the theory of functions $f : \mathbb{R} \to \mathbb{R}$, with the notion of continuity, and the construction of the derivative $f'(x)$ and of the integral $\int_a^b f(x)dx$. Then we investigate the case of the complex functions $f : \mathbb{C} \to \mathbb{C}$, and notably the holomorphic functions, and harmonic functions. Then, we discuss the multivariable functions, $f : \mathbb{R}^N \to \mathbb{R}^M$ or $f : \mathbb{R}^N \to \mathbb{C}^M$ or $f : \mathbb{C}^N \to \mathbb{C}^M$, with general theory, integration results, maximization questions, and basic applications to physics.

Preface

Understanding what happens in the real life surrounding us, in phenomena involving physics, chemistry, biology and so on, is not an easy task. What we can do as humans is to come up with some machinery, and perform measurements, recording quantities such as length, volume, temperature, pressure and so on, and then see how these quantities, called "variables", and denoted $x, y, z, \ldots$ depend on each other, and change in time.

Calculus is the study of the correspondences $x \to y$ between such variables. Such correspondences are called "functions", and are denoted $y = f(x)$, with f standing for the abstract machinery, or mathematical formula, producing y out of x.

The basics of calculus were developed by Newton, Leibnitz and others, a long time ago. The idea is very simple. The simplest functions $f : \mathbb{R} \to \mathbb{R}$ are the linear ones, $f(x) = a + bx$ with $a, b \in \mathbb{R}$, but of course not any function is linear. Miraculously, however, most functions $f : \mathbb{R} \to \mathbb{R}$ are "locally linear", in the sense that around any given point $c \in \mathbb{R}$, we have a formula of type $f(c + x) \simeq a + bx$, for x small. Why? Obviously, $a \in \mathbb{R}$ can only be the value of our function at that point, $a = f(c)$. As for the number $b \in \mathbb{R}$, this can be taken to be the rate of change of f around that point, called derivative of the function at that point, and denoted $b = f'(c)$.

So, this was the main idea of calculus, "functions are locally linear". This idea applies as well to more complicated functions, such as the "multivariable" ones $f : \mathbb{R}^N \to \mathbb{R}^M$, relating vector variables $x \in \mathbb{R}^N$ to vector variables $y \in \mathbb{R}^M$, with the linear approximation formula $f(c + x) \simeq a + bx$ needing this time as parameters a vector $a = f(c) \in \mathbb{R}^M$, and a linear map, or beast called rectangular matrix, $b = f'(c) \in M_{M \times N}(\mathbb{R})$.

Further ideas of calculus, which are more advanced, include the facts that: (1) the remainder $\varepsilon(x)$ given by $f(c + x) = a + bx + \varepsilon(x)$ can studied by using again derivatives, (2) in several variables, the geometric understanding of the derivatives $f'(c) \in M_{M \times N}(\mathbb{R})$ is best done by using complex numbers, (3) in fact, the use of complex numbers is useful even for one-variable functions $f : \mathbb{R} \to \mathbb{R}$, and (4) in one variable at least, there is a magic relation between derivatives and weighted averages, called integrals and denoted $\int_a^b f(x)dx$, the idea being that "the derivative of the integral is the function itself".

Calculus can be learned from many places, with this being mostly a matter of taste. Personally as a student I used to be quite interested in science and mathematics, and so I skipped classes, that I found quite boring, and read Rudin [74], [75] instead. Later on, however, I had to attend some of these calculus classes, ironically, as the professor teaching them. For preparing my classes I usually still rely on Rudin, with quite often a look into internet, Wikipedia or similar websites, and into other good books, or at least books that I personally like, including those of Lax and Terrell [63], [64].

The present book is an introduction to calculus, based on lecture notes from various classes that I taught at Cergy, and previously at Toulouse. The material inside claims of course no originality, basically going back to Newton, Leibnitz and others. But in what regards the presentation, there are a few ideas behind it, none of these claiming of course originality either, but their combination being something original, I hope:

(1) One complex variable comes before several real variables.

(2) Applications to probability everywhere, scattered throughout the book.

(3) Combinatorics, binomials and factorials all over the place, with joy.

(4) Applications to physics too, including the hydrogen atom, at the end.

In the hope that you will like this book. High-school or undergraduate students, wishing to learn full calculus in a quick way, graduate students in math and science, wishing to fine-tune their calculus knowledge, or just math professionals like me, wishing to have a compact analysis book, so that they can grab the appropriate chapter, before going to class, no matter what the class is about. Hope you will all find this useful.

As already mentioned, the present book is based on lecture notes from classes at Toulouse and Cergy, and I would like to thank my students. Many thanks go as well to my cats, for useful pieces of advice, often complementary to the pieces of advice of my colleagues. And of course, with an homage to Newton's cat, who the legend goes, used to hang out in an apple tree, and was at the beginning of everything.

Contents

Preface 3

Part I. Basic calculus 9

Chapter 1. Sequences, series 11
- 1a. Binomials, factorials 11
- 1b. Real numbers, analysis 16
- 1c. Sequences, convergence 22
- 1d. Series, the number e 25
- 1e. Exercises 32

Chapter 2. Functions, continuity 33
- 2a. Continuous functions 33
- 2b. Intermediate values 39
- 2c. Sequences and series 45
- 2d. Basic functions 48
- 2e. Exercises 56

Chapter 3. Derivatives 57
- 3a. Derivatives, rules 57
- 3b. Second derivatives 66
- 3c. The Taylor formula 72
- 3d. Differential equations 77
- 3e. Exercises 80

Chapter 4. Integration 81
- 4a. Integration theory 81
- 4b. Riemann sums 87
- 4c. Basic results 93
- 4d. Some probability 101
- 4e. Exercises 104

Part II. Complex functions 105

Chapter 5. Complex numbers 107
5a. Complex numbers 107
5b. Exponential writing 115
5c. Equations, roots 121
5d. Roots of unity 126
5e. Exercises 128

Chapter 6. Complex functions 129
6a. Functions, continuity 129
6b. Holomorphic functions 137
6c. Cauchy formula 141
6d. Stieltjes inversion 147
6e. Exercises 152

Chapter 7. Fourier analysis 153
7a. Function spaces 153
7b. Fourier transform 158
7c. Groups, extensions 166
7d. Independence, limits 172
7e. Exercises 176

Chapter 8. Harmonic functions 177
8a. Laplace operator 177
8b. Harmonic functions 183
8c. Light, spectroscopy 191
8d. Atomic spectrum 196
8e. Exercises 200

Part III. Several variables 201

Chapter 9. Linear maps 203
9a. Linear maps 203
9b. Matrix inversion 206
9c. The determinant 213
9d. Diagonalization 220
9e. Exercises 224

Chapter 10. Partial derivatives 225
10a. Functions, continuity 225
10b. Compact sets 231
10c. Partial derivatives 238
10d. The chain rule 244
10e. Exercises 248

Chapter 11. Some geometry 249
11a. Equations, conics 249
11b. Kepler and Newton 255
11c. Algebraic manifolds 261
11d. Differential manifolds 268
11e. Exercises 272

Chapter 12. Higher derivatives 273
12a. Higher derivatives 273
12b. Matrices, positivity 281
12c. Harmonic functions 290
12d. Lagrange multipliers 292
12e. Exercises 296

Part IV. Integration theory 297

Chapter 13. Multiple integrals 299
13a. Multiple integrals 299
13b. Spherical coordinates 302
13c. Stirling estimates 306
13d. Spherical integrals 313
13e. Exercises 320

Chapter 14. Normal variables 321
14a. Probability basics 321
14b. Normal variables 329
14c. Hyperspherical laws 337
14d. Rotations, reflections 340
14e. Exercises 344

Chapter 15. Partial integration 345

15a. Vector products 345
15b. Einstein addition 351
15c. Charges and flux 359
15d. Gauss, Green, Stokes 365
15e. Exercises 368

Chapter 16. Infinite dimensions 369
16a. Operators, matrices 369
16b. Schrödinger equation 375
16c. Spherical coordinates 377
16d. The hydrogen atom 386
16e. Exercises 392

Bibliography 393

Index 397

Part I

Basic calculus

I've got to stand and fight
In this creation
Vanity I know
Can't guide I alone

CHAPTER 1

Sequences, series

1a. Binomials, factorials

We denote by $\mathbb{N}$ the set of positive integers, $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$, with $\mathbb{N}$ standing for "natural". Quite often in computations we will need negative numbers too, and we denote by $\mathbb{Z}$ the set of all integers, $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$, with $\mathbb{Z}$ standing from "zahlen", which is German for "numbers". Finally, there are many questions in mathematics involving fractions, or quotients, which are called rational numbers:

DEFINITION 1.1. *The rational numbers are the quotients of type*

$$r = \frac{a}{b}$$

with $a, b \in \mathbb{Z}$, and $b \neq 0$, identified according to the usual rule for quotients, namely:

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

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

Observe that we have inclusions $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q}$. The integers add and multiply according to the rules that you know well. As for the rational numbers, these add according to the usual rule for quotients, which is as follows, and death penalty for forgetting it:

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

Also, the rational numbers multiply according to the usual rule for quotients, namely:

$$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}$$

Beyond rationals, we have the real numbers, whose set is denoted $\mathbb{R}$, and which include beasts such as $\sqrt{3} = 1.73205\ldots$ or $\pi = 3.14159\ldots$ But more on these later. For the moment, let us see what can be done with integers, and their quotients. As a first theorem, solving a problem which often appears in real life, we have:

THEOREM 1.2. *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. Imagine a set consisting of n objects. We have 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:

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

But is this correct. Normally a mathematical theorem coming with mathematical proof is guaranteed to be 100% correct, and if in addition the proof is truly clever, like the above proof was, with that fraction trick, the confidence rate jumps up to 200%.

This being said, never knows, so let us doublecheck, by taking for instance $n=3, k=2$. Here we have to choose 2 objects among 3 objects, and this is something easily done, because what we have to do is to dismiss one of the objects, and $N=3$ choices here, and keep the 2 objects left. Thus, we have $N=3$ choices. On the other hand our genius math computation gives $N=3!/1!=6$, which is obviously the wrong answer.

So, where is the mistake? Thinking a bit, the 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:

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

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 we have $M=k(k-1)\ldots 2\cdot 1=k!$, and so:

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

And this is the correct answer, because, well, that is how things are. In case you doubt, at $n=3, k=2$ for instance we obtain $3!/2!1!=3$, which is correct. □

All this is quite interesting, and in addition to having some exciting mathematics going on, and more on this in a moment, we have as well some philosophical conclusions. Formulae can be right or wrong, and as the above shows, good-looking, formal mathematical proofs can be right or wrong too. So, what to do? Here is my advice:

ADVICE 1.3. *Always doublecheck what you're doing, regularly, and definitely at the end, either with an alternative proof, or with some numerics.*

This is something very serious. Unless you're doing something very familiar, that you're used to for at least 5-10 years or so, like doing additions and multiplications for you, or some easy calculus for me, formulae and proofs that you can come upon are by default wrong. In order to make them correct, and ready to use, you must check and doublecheck and correct them, helped by alternative methods, or numerics.

Which brings us into the question on whether mathematics is an exact science or not. Not clear. Chemistry for instance is an exact science, because findings of type "a mixture of water and salt cannot explode" look rock-solid. Same for biology, with findings of type "crocodiles eat fish" being rock-solid too. In what regards mathematics however, and theoretical physics too, things are always prone to human mistake.

And for ending this discussion, you might ask then, what about engineering? After all, this is mathematics and physics, which is usually 100% correct, because most of the bridges, buildings and other things built by engineers don't collapse. Well, this is because engineers follow, and in a truly maniac way, the above Advice 1.3. You won't declare a project for a bridge, building, engine and so on final and correct, ready for production, until you checked and doublechecked it with 10 different methods or so, won't you.

Back to work now, as an important adding to Theorem 1.2, we have:

CONVENTION 1.4. *By definition,* $0! = 1$.

This convention comes, and no surprise here, from Advice 1.3. Indeed, we obviously have $\binom{n}{n} = 1$, but if we want to recover this formula via Theorem 1.2 we are a bit in trouble, and so we must declare that $0! = 1$, as for the following computation to work:

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

Going ahead now with more mathematics and less philosophy, with Theorem 1.2 complemented by Convention 1.4 being now in final form (trust me), we have:

THEOREM 1.5. *We have the binomial formula*

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

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 $a^k b^{n-k}$. Thus, we have a formula as follows:

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

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 1.2, 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:

$$a+b = a+b$$

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

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

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

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

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* $\binom{n}{k}$,

$$1 \ , \ 1$$

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

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

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

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

$$\vdots$$

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

PROOF. As a first observation, for having a full triangle we should normally add a $\binom{0}{0} = 1$ entry on top, corresponding to the formula $(a+b)^0 = 1$, but let us not bother with that. In practice, the theorem states that the following formula must hold:

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

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 now the binomial formula, this formula becomes:

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

Now let us perform the multiplication on the right. We obtain a certain sum of terms of type $a^k b^{n-k}$, and to be more precise, each such $a^k b^{n-k}$ 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 $a^k b^{n-1-k}$ multiplied by b. Thus, the coefficient of $a^k b^{n-k}$ 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. □

There are many more things that can be said about binomial coefficients, with all sorts of interesting formulae, but the idea is always the same, namely that in order to find such formulae you have a choice between algebra and combinatorics, and that when it comes to proofs, the brute-force computation method is useful too. In practice, the best

is to master all 3 techniques. Among others, because of Advice 1.3. You will have in this way 3 different methods, for making sure that your formulae are correct indeed.

1b. Real numbers, analysis

All the above was very nice, but remember that we are here for doing science and physics, and more specifically for mathematically understanding the numeric variables $x, y, z, \ldots$ coming from real life. Such variables can be lengths, volumes, pressures and so on, which vary continuously with time, and common sense dictates that there is little to no chance for our variables to be rational, $x, y, z, \ldots \notin \mathbb{Q}$. In fact, we will even see soon a theorem, stating that the probability for such a variable to be rational is exactly 0. Or, to put it in a dramatic way, "rational numbers don't exist in real life".

You are certainly familiar with the real numbers, but let us review now their definition, which is something quite tricky. 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 1.7. *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 1.8. *The number $\sqrt{2}$ is the abstract beast which is bigger than all rationals satisfying $p^2 < 2$, and smaller than all positive rationals satisfying $q^2 > 2$.*

This does not look very good, but you know what, instead of looking for more clever solutions to our problem, what about relaxing, or being lazy, or coward, or you name it, and taking Conclusion 1.8 as a definition for $\sqrt{2}$. This is actually something not that bad, and leads to the following "lazy" definition for the real numbers:

DEFINITION 1.9. *The real numbers $x \in \mathbb{R}$ are formal cuts in the set of rationals,*

$$\mathbb{Q} = \mathbb{Q}_{\leq x} \sqcup \mathbb{Q}_{>x}$$

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

$$p \in \mathbb{Q}_{\leq x}\ ,\ q \in \mathbb{Q}_{>x} \implies p < q$$

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:

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

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:

$$\mathbb{Q}_{\leq x+y} = \mathbb{Q}_{\leq x} + \mathbb{Q}_{\leq y}$$

As for the multiplication, the formula here is similar, namely $\mathbb{Q}_{\leq xy} = \mathbb{Q}_{\leq x}\mathbb{Q}_{\leq y}$, 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:

$$x \leq y \iff \mathbb{Q}_{\leq x} \subset \mathbb{Q}_{\leq y}$$

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 theorem:

THEOREM 1.10. *The equation $x^2 = 2$ has two solutions over the real numbers, namely the positive solution, denoted $\sqrt{2}$, and its negative counterpart, which is $-\sqrt{2}$.*

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:

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

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 1.11. *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 definition for the real numbers, that we can certainly do some math with. However, for anything more advanced we are in need of the decimal writing for the real numbers. The result here is as follows:

THEOREM 1.12. *The real numbers $x\in\mathbb{R}$ can be written in decimal form,*

$$x=\pm a_1\ldots a_n.b_1b_2b_3\ldots\ldots$$

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 quite non-trivial, assuming that you already have some familiarity with such things, for the rational numbers. The idea is 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 $x=\pm a_1\ldots a_n.b_1b_2b_3\ldots\ldots$ 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 $\mathbb{Q}=\mathbb{Q}_{\leq x}\sqcup\mathbb{Q}_{>x}$. Since the set $\mathbb{Q}_{\leq x}\cap\mathbb{Z}$ consists of integers, and is bounded from above by any element $q\in\mathbb{Q}_{>x}$ of your choice, this set has a maximal element, that we can denote $[x]$:

$$[x]=\max\left(\mathbb{Q}_{\leq x}\cap\mathbb{Z}\right)$$

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 $0.b_10, 0.b_11, \ldots, 0.b_19$. 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 $x = 0.b_1b_2b_3 \ldots$ defines a real number $x \in [0, 1]$, simply by declaring that the corresponding cut $\mathbb{Q} = \mathbb{Q}_{\leq x} \sqcup \mathbb{Q}_{>x}$ comes from the following set, and its complement:

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

(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:

$$\mathbb{Q}_{\leq \ldots b999 \ldots} = \mathbb{Q}_{\leq \ldots (b+1)000 \ldots}$$

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. You might perhaps say why bothering with cuts, and not taking $x = \pm a_1 \ldots a_n.b_1b_2b_3 \ldots \ldots$ as definition for the real numbers. Well, this is certainly possible, but when it comes to summing such numbers, or making products, or proving basic things such as the existence of $\sqrt{2}$, things become fairly complicated with the decimal writing picture. So, all the above is not as stupid as it seems. And we will come back anyway to all this later on, with a 3rd picture for the real numbers, involving scary things like ε and δ, and it will be up to you to decide, at that time, which picture is the one that you prefer.

Moving on, we made the claim in the beginning of this chapter that "in real life, real numbers are never rational". Here is a theorem, justifying this claim:

THEOREM 1.13. *The probability for a real number $x \in \mathbb{R}$ to be rational is* 0.

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

(1) Before starting, let us point out the fact that probability theory is something quite tricky, with probability 0 not necessarily meaning that the event cannot happen, but rather meaning that "better not count on that". For instance according to my computations the probability of you winning 1 billion at the lottery is 0, but you are of course free to disagree, and prove me wrong, by playing every day at the lottery.

(2) With this discussion made, and extrapolating now from finance and lottery to our question regarding real numbers, your possible argument of type "yes, but if I pick $x \in \mathbb{R}$ to be $x = 3/2$, I have proof that the probability for $x \in \mathbb{Q}$ is nonzero" is therefore dismissed. Thus, our claim as stated makes sense, so let us try now to prove it.

(3) By translation, it is enough to prove that the probability for a real number $x \in [0, 1]$ to be rational is 0. For this purpose, let us write the rational numbers $r \in [0, 1]$ in the form of a sequence $r_1, r_2, r_3 \ldots$, with this being possible say by ordering our rationals $r = a/b$ according to the lexicographic order on the pairs (a, b):

$$\mathbb{Q} \cap [0, 1] = \left\{ r_1, r_2, r_3, \ldots \right\}$$

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:

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

Here we have used the well-known formula $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = 1$, which comes by dividing $[0, 1]$ into half, and then one of the halves into half again, and so on, and then saying in the end that the pieces that we have must sum up to 1. Thus, we have indeed $P \leq c$, and since the number $c > 0$ was arbitrary, we obtain $P = 0$, as desired. □

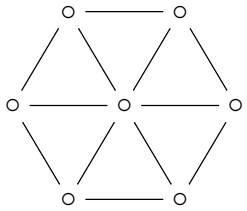
As a comment here, all the above is of course quite tricky, and a bit bordeline in respect to what can be called "rigorous mathematics". But we will be back to this, namely general probability theory, and in particular meaning of the mysterious formula $P = 0$, countable sets, infinite sums and so on, on several occasions, throughout this book.

Moving ahead now, let us construct now some more real numbers. We already know about $\sqrt{2}$ and other numbers of the same type, namely roots of polynomials, and our knowledge here being quite decent, no hurry with this, we will be back to it later. So, let us get now into π and trigonometry. To start with, we have the following result:

THEOREM 1.14. *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:

$$\begin{aligned} A &= N \times \frac{HS}{2} \\ &= \frac{HP}{2} \\ &\simeq \frac{1 \times L}{2} \end{aligned}$$

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 $\pi = 3.14159\ldots$, but we will come back to this later, once we will have appropriate tools for dealing with such questions. It is also possible to prove that π is irrational, $\pi \notin \mathbb{Q}$, but this is not trivial either.

Let us end this discussion about real numbers with some trigonometry. There are many things that can be said, that you certainly know, the basics being as follows:

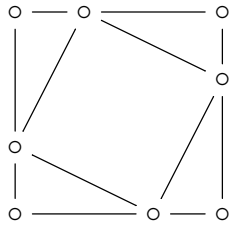
THEOREM 1.15. *The following happen:*

(1) *We can talk about angles $x \in \mathbb{R}$, by using the unit circle, in the usual way, and in this correspondence, the right angle has a value of $\pi/2$.*
(2) *Associated to any $x \in \mathbb{R}$ are numbers $\sin x, \cos x \in \mathbb{R}$, constructed in the usual way, by using a triangle. These numbers satisfy $\sin^2 x + \cos^2 x = 1$.*

PROOF. There are certainly things that you know, the idea being as follows:

(1) The formula $L = 2\pi$ from Theorem 1.14 shows that the length of a quarter of the unit circle is $l = \pi/2$, and so the right angle has indeed this value, $\pi/2$.

(2) As for $\sin^2 x+\cos^2 x=1$, called Pythagoras' theorem, this comes from the following picture, with the edges of the outer and inner square being $\sin x+\cos x$ and 1:



Indeed, when computing the area of the outer square, we obtain:

$$(\sin x+\cos x)^2=1+4\times\frac{\sin x\cos x}{2}$$

Now when expanding we obtain $\sin^2 x+\cos^2 x=1$, as claimed. □

It is possible to say many more things about angles and $\sin x$, $\cos x$, and also talk about some supplementary quantities, such as $\tan x=\sin x/\cos x$. But more on this later, once we will have some appropriate tools, beyond basic geometry, in order to discuss this.

1c. Sequences, convergence

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 1.16. *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 looks quite scary, but isn't. Let us think a bit, how shall we translate $x_n\to x$ into mathematical language. The condition $x_n\to x$ tells us that "when n is big, x_n is close to x", and to be more precise, it tells us that "when n is big enough, x_n gets arbitrarily close to x". But n big enough 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 led into the above definition.

As a basic example for all this, we have:

PROPOSITION 1.17. *We have $1/n\to 0$.*

PROOF. This is obvious, but let us prove it by using Definition 1.16. We have:

$$\left|\frac{1}{n}-0\right|<\varepsilon\iff\frac{1}{n}<\varepsilon\iff\frac{1}{\varepsilon}<n$$

Thus we can take $N=[1/\varepsilon]+1$ in Definition 1.16, and we are done. □

There are many other examples, and more on this in a moment. Going ahead with more theory, let us complement Definition 1.16 with:

DEFINITION 1.18. *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. As a basic illustration, we have:

PROPOSITION 1.19. *We have $n^2 \to \infty$.*

PROOF. As before, this is obvious, but let us prove it using Definition 1.18. We have:

$$n^2 > K \iff n > \sqrt{K}$$

Thus we can take $N = [\sqrt{K}] + 1$ in Definition 1.18, and we are done. □

We can generalize Proposition 1.17 and Proposition 1.19, as follows:

PROPOSITION 1.20. *We have the following convergence, with $n \to \infty$:*

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

PROOF. This follows indeed by using the same method as in the proof of Proposition 1.17 and Proposition 1.19, first for a rational, and then for a real as well. □

We have some general results about limits, summarized as follows:

THEOREM 1.21. *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 $k \in \mathbb{N}$:

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

(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 from definitions. □

Here are as well some general rules for computing limits:

THEOREM 1.22. *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* $x_n y_n \to xy$.
(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:

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

(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:

PROPOSITION 1.23. *The $n \to \infty$ limits of quotients of polynomials are given by*

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

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:

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

As for the second assertion, this comes from Proposition 1.20. □

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 1.24. *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 1.25. *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 1.26. *$\mathbb{R}$ 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. Let us denote the completion operation by $X \to \bar{X} = C_X/\sim$, where C_X is the space of Cauchy sequences over X, and $\sim$ is the above equivalence relation. Since by Theorem 1.25 any Cauchy sequence $(x_n) \in C_{\mathbb{Q}}$ has a limit $x \in \mathbb{R}$, we obtain $\bar{\mathbb{Q}} = \mathbb{R}$. As for the equality $\bar{\mathbb{R}} = \mathbb{R}$, this is clear again by using Theorem 1.25. □

1d. Series, the number e

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

DEFINITION 1.27. *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 1.28. *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 1.29. *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 1.30. *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 1.29, as follows:

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

$$\begin{aligned}1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\ldots&=\left(1-\frac{1}{2}\right)+\left(\frac{1}{3}-\frac{1}{4}\right)+\ldots\\&=\frac{1}{2}+\frac{1}{12}+\frac{1}{30}+\ldots\\&<\frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}+\ldots\\&<\infty\end{aligned}$$

(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 1.31. *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 $x_k=S_k-S_{k-1}\to 0$, and this gives the result. As for the simplest counterexample for the converse, this is $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots=\infty$, coming from Theorem 1.29.

(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 is $1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\ldots<\infty$, coming from Theorem 1.30, coupled with $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots=\infty$ from (1).

(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 1.32. *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 1.30, we have:

THEOREM 1.33. *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. □

All this was a bit theoretical, and as something more concrete now, we have:

THEOREM 1.34. *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, we use the following arithmetic-geometric inequality:

$$\frac{1+\sum_{i=1}^n\left(1+\frac{1}{n}\right)}{n+1}\geq\sqrt[n+1]{1\cdot\prod_{i=1}^n\left(1+\frac{1}{n}\right)}$$

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, 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 later, with better methods. □

We should mention that there are many other ways of getting into e. For instance it is possible to prove that we have the following formula, which is a bit more conceptual than the formula in Theorem 1.34, and also with the convergence being very quick:

$$\sum_{n=0}^{\infty}\frac{1}{n!}=e$$

Importantly, all this not the end of the story with e. For instance, in relation with the first formula that we found, from Theorem 1.34, we have, more generally:

$$\left(1+\frac{x}{n}\right)^{n}\to e^{x}$$

Also, in relation with the second formula, from above, we have, more generally:

$$\sum_{n=0}^{\infty}\frac{x^n}{n!}=e^x$$

To be more precise, these latter two formulae are something that we know at $x=1$. The case $x=0$ is trivial, the case $x=-1$ follows from the case $x=1$, via some simple manipulations, and with a bit more work, we can get these formulae for any $x\in\mathbb{N}$, and then for any $x\in\mathbb{Z}$. However, the general case $x\in\mathbb{R}$ is quite tricky, requiring a good knowledge of the theory of real functions. And, good news, real functions will be what we will be doing in the remainder of this first part, in chapters 2-4 below.

1e. Exercises

This opening chapter was a bit special, containing a lot of material in need to be known, and compacted to the maximum. As exercises, again compacted, we have:

EXERCISE 1.35. *Prove that the rational numbers $r\in\mathbb{Q}$ are exactly the real numbers whose decimal expansion is periodic.*

EXERCISE 1.36. *Find geometric proofs, using triangles in the plane, for the well-known formulae for* $\sin(x+y)$ *and* $\cos(x+y)$.

EXERCISE 1.37. *Develop some convergence theory for $x_n=a^n$ with $a>0$, notably by proving that $a^n/n^k\to\infty$ for any $a>1$, and any $k\in\mathbb{N}$.*

EXERCISE 1.38. *Prove that $\sum_{n=0}^{\infty}\frac{1}{n!}=e$. Also, prove that $\left(1+\frac{x}{n}\right)^n\to e^x$, and that $\sum_{n=0}^{\infty}\frac{x^n}{n!}=e^x$, for $x=-1$, then for $x\in\mathbb{Z}$, then for $x\in\mathbb{R}$.*

These exercises are probably quite difficult, unless you are already a bit familiar with all this. If this is not the case, a good idea at this point is to pick a random entry-level calculus book, and work out a few dozen exercises from there, as a warm-up.

CHAPTER 2

Functions, continuity

2a. Continuous functions

We are now ready to talk about functions, which are the main topic of this book. A function $f : \mathbb{R} \to \mathbb{R}$ is a correspondence $x \to f(x)$, which to each real number $x \in \mathbb{R}$ associates a real number $f(x) \in \mathbb{R}$. As examples, we have $f(x) = x^2$, $f(x) = 2^x$ and so on. This suggests that any function $f : \mathbb{R} \to \mathbb{R}$ should be given by some kind of "mathematical formula", but unfortunately this is not correct, because, with suitable definitions of course, there are more functions than mathematical formulae.

This being said, we will see that under suitable regularity assumptions on $f : \mathbb{R} \to \mathbb{R}$, we have indeed a mathematical formula for $f(x)$ in terms of x, at least locally. And with this being actually the main idea of calculus, that will take some time to be developed. But more on this later, once we will know more about functions.

Getting started now, let us keep from the above discussion the idea that we should focus our study on the functions $f : \mathbb{R} \to \mathbb{R}$ having suitable regularity properties. In what regards these regularity properties, the most basic of them is continuity:

DEFINITION 2.1. *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.

Observe that a function $f : X \to \mathbb{R}$ is continuous precisely when it is continuous at any point $x \in X$. We will see examples in a moment. Still speaking theory, there are many equivalent formulations of the notion of continuity, with a well-known one, coming by reminding in the above definition what convergence of a sequence means, twice, for both the convergences $x_n \to x$ and $f(x_n) \to f(x)$, being as follows:

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

At the level of examples, basically all the functions that you know, including powers x^a, exponentials a^x, and more advanced functions like $\sin, \cos, \exp, \log$, are continuous. However, proving this will take some time. Let us start with:

THEOREM 2.2. *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 indeed. 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:

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

(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 given after Definition 2.1. So, let us pick $\varepsilon > 0$. 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 comment, (3) shows in particular that $1/f$ is continuous, and we will use this many times, in what follows. As a second comment, more philosophical, the proof of (4) shows that the ε, δ formulation of continuity can be sometimes more complicated than the usual formulation, with sequences, which leads us into the question of why bothering at all with this ε, δ condition. Good question, and in answer:

(1) It is usually said that "for doing advanced math, you must use the ε, δ condition", but this is not exactly true, because sometimes what happens is that "for doing advanced math, you must use open and closed sets". With these sets, and the formulation of continuity in terms of them, being something that we will discuss a bit later.

(2) This being said, the point is that the use of open and closed sets, technology that we will discuss in a moment, requires some prior knowledge of the ε, δ condition. So, you cannot really run away from this ε, δ condition, and want it or not, in order to do later some more advanced mathematics, you'll have to get used to that.

(3) But this should be fine, because you're here since you love math and science, aren't you, and good math and science, including this ε, δ condition, will be what you will learn from here. So, everything fine, more on this later, and in the meantime, no matter what we do, always take a few seconds to think at what that means, in ε, δ terms.

Back to work now, at the level of examples, we have:

THEOREM 2.3. *The following functions are continuous:*

(1) x^n, *with* $n \in \mathbb{Z}$.
(2) P/Q, *with* $P, Q \in \mathbb{R}[X]$.
(3) $\sin x$, $\cos x$, $\tan x$, $\cot x$.

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

(1) Since $f(x) = x$ is continuous, by using Theorem 2.2 we obtain the result for exponents $n \in \mathbb{N}$, and then for general exponents $n \in \mathbb{Z}$ too.

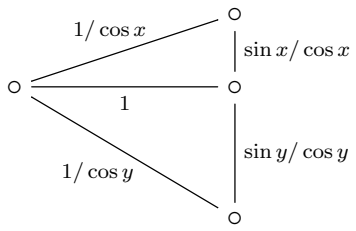
(2) The statement here, which generalizes (1), follows exactly as (1), by using the various findings from Theorem 2.2.

(3) We must first prove here that $x_n \to x$ implies $\sin x_n \to \sin x$, which in practice amounts in proving that $\sin(x + y) \simeq \sin x$ for y small. But this follows from:

$$\sin(x + y) = \sin x \cos y + \cos x \sin y$$

To be more precise, let us first establish this formula. In order to do so, consider the following picture, consisting of a length 1 line segment, with angles x, y drawn on each

side, and with everything being completed, and lengths 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 $\sin(x+y)$, or by using the two small triangles, involving functions of x, y. We obtain in this way:

$$\frac{1}{\cos x}\cdot\frac{1}{\cos y}\cdot\sin(x+y)=\frac{\sin x}{\cos x}\cdot 1+\frac{\sin y}{\cos y}\cdot 1$$

But this gives the formula for $\sin(x+y)$ claimed above.

(4) Now with this formula in hand, we can establish the continuity of $\sin x$, as follows, with the limits at 0 which are used being both clear on pictures:

$$\begin{aligned}\lim_{y\to 0}\sin(x+y) &= \lim_{y\to 0}(\sin x\cos y+\cos x\sin y)\\ &= \sin x\lim_{y\to 0}\cos y+\cos x\lim_{y\to 0}\sin y\\ &= \sin x\cdot 1+\cos x\cdot 0\\ &= \sin x\end{aligned}$$

(5) Moving ahead now with $\cos x$, here the continuity follows from the continuity of $\sin x$, by using the following formula, which is obvious from definitions:

$$\cos x=\sin\left(\frac{\pi}{2}-x\right)$$

(6) Alternatively, and let us do this because we will need later the formula, by using the formula for $\sin(x+y)$ we can deduce a formula for $\cos(x+y)$, as follows:

$$\begin{aligned}\cos(x+y) &= \sin\left(\frac{\pi}{2}-x-y\right)\\ &= \sin\left[\left(\frac{\pi}{2}-x\right)+(-y)\right]\\ &= \sin\left(\frac{\pi}{2}-x\right)\cos(-y)+\cos\left(\frac{\pi}{2}-x\right)\sin(-y)\\ &= \cos x\cos y-\sin x\sin y\end{aligned}$$

But with this, we can use the same method as in (4), and we get, as desired:

$$\begin{aligned}\lim_{y\to 0}\cos(x+y) &= \lim_{y\to 0}(\cos x\cos y-\sin x\sin y)\\ &= \cos x\lim_{y\to 0}\cos y-\sin x\lim_{y\to 0}\sin y\\ &= \cos x\cdot 1-\sin x\cdot 0\\ &= \cos x\end{aligned}$$

(7) Finally, the fact that $\tan x$, $\cot x$ are continuous is clear from the fact that $\sin x$, $\cos x$ are continuous, by using the result regarding quotients from Theorem 2.2. □

We will be back to more examples later, and in particular to functions of type x^a and a^x with $a\in\mathbb{R}$, which are more tricky to define. Also, we will talk as well about inverse functions f^{-1}, with as particular cases the basic inverse trigonometric functions, namely arcsin, arccos, arctan, arccot, once we will have more tools for dealing with them.

Going ahead with more theory, some functions are "obviously" continuous:

PROPOSITION 2.4. *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 indeed clear from our definition of continuity. □

Along the same lines, we can also argue, based on our intuition, that "some functions are more continuous than other". For instance, we have the following definition:

DEFINITION 2.5. *A function $f:X\to\mathbb{R}$ is called uniformly continuous when:*

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

That is, f must be continuous at any $x\in X$, with the continuity being "uniform".

As basic examples of uniformly continuous functions, we have the Lipschitz ones. Also, as a basic counterexample, we have the following function:

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

Indeed, it is clear by looking at the graph of f that, the further our point $x\in\mathbb{R}$ is from 0, the smaller our $\delta>0$ must be, compared to $\varepsilon>0$, in our ε,δ definition of continuity. Thus, given some $\varepsilon>0$, we have no $\delta>0$ doing the $|x-y|<\delta\implies|f(x)-f(y)|<\varepsilon$ job at any $x\in\mathbb{R}$, and so our function is indeed not uniformly continuous.

Quite remarkably, we have the following theorem, due to Heine and Cantor:

THEOREM 2.6. *Any continuous function defined on a closed, bounded interval*

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

is automatically uniformly continuous.

PROOF. This is something quite subtle, and we are punching here a bit above our weight, but here is the proof, with everything or almost included:

(1) Given $\varepsilon > 0$, for any $x \in [a, b]$ 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 $x \in [a, b]$:

$$U_x = \left(x - \frac{\delta_x}{2} \, , \, x + \frac{\delta_x}{2} \right)$$

These intervals then obviously cover $[a, b]$, in the sense that we have:

$$[a, b] \subset \bigcup_{x \in [a,b]} U_x$$

Now assume that we managed to prove that this cover has a finite subcover. Then we can most likely choose our $\delta > 0$ to be the smallest of the $\delta_x > 0$ involved, or perhaps half of that, and then get our uniform continuity condition, via the triangle inequality.

(2) So, let us prove first that the cover in (1) has a finite subcover. For this purpose, 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. 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.

(3) With this done, we are ready to finish, as announced in (1). Indeed, let us denote by $[a, b] \subset \bigcup_i U_{x_i}$ the finite subcover found in (2), and let us set:

$$\delta = \min_i \frac{\delta_{x_i}}{2}$$

Now assume $|x - y| < \delta$, and pick i such that $x \in U_{x_i}$. By the triangle inequality we have then $|x_i - y| < \delta_{x_i}$, which shows that we have $y \in U_{x_i}$ as well. But by applying now f, this gives as desired $|f(x) - f(y)| < \varepsilon$, again via the triangle inequality. □

2b. Intermediate values

Moving ahead with more theory, we would like to explain now an alternative formulation of the notion of continuity, which is quite abstract, and a bit difficult to understand and master when you are a beginner, but which is definitely worth learning, because it is quite powerful, solving some of the questions that we have left. Let us start with:

DEFINITION 2.7. *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 examples, the open intervals (a, b) are open, and the closed intervals $[a, b]$ are closed. Observe also that $\mathbb{R}$ itself is open and closed at the same time. Further examples, or rather results which are easy to establish, include the fact that the finite unions or intersections of open or closed sets are open or closed. We will be back to all this later, with some precise results in this sense. For the moment, we will only need:

PROPOSITION 2.8. *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 if this happens indeed, we are done. Otherwise, we have $x \notin C$, and so $x \in O$, 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.

"$\impliedby$" 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 example for this, $\mathbb{R} - (a, b) = (-\infty, a] \cup [b, \infty)$ is closed, as a union of two closed sets, and $\mathbb{R} - [a, b] = (-\infty, a) \cup (b, \infty)$ is open, as a union of two open sets. More on this in a moment, and getting now back to functions, we have:

THEOREM 2.9. *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. Here the first assertion follows from definitions, and more specifically from the ε, δ definition of continuity, which was as follows:

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

Indeed, if f satisfies this condition, it is clear that if O is open, then $f^{-1}(O)$ is open, and the converse holds too. As for the second assertion, this can be proved either directly, by using the $f(x_n) \to f(x)$ definition of continuity, or by taking complements. □

As a test for the above criterion, let us reprove the fact, that we know from Theorem 2.2, 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 at least in relation with this specific problem, 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.

In order to reach to true applications of Theorem 2.9, we will need to know more about the open and closed sets. Let us begin with a useful result, as follows:

PROPOSITION 2.10. *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. Here (1) is clear from definitions, (3) is clear from definitions too, and (2,4) follow from (1,3) by taking complements $E \to E^c$, using the following formulae:

$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c \quad , \quad \left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c$$

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:

$$\bigcap_{n\in\mathbb{N}} (-1/n, 1/n) = \{0\}$$

As for (4), here the simplest counterexample is as follows:

$$\bigcup_{n\in\mathbb{N}} [0, 1 - 1/n] = [0, 1)$$

All this is quite interesting, and leads us to the question about what the open and closed sets really are. And fortunately, this question can be answered, as follows:

THEOREM 2.11. *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 Proposition 2.10 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 Proposition 2.8. □

The above result is quite interesting, philosophically speaking, because contrary to what we have been doing so far, it makes the open sets appear quite different from the closed sets. Indeed, there is no way of having a simple description of the closed sets $C \subset \mathbb{R}$, similar to the above simples description of the open sets $O \subset \mathbb{R}$.

Moving towards more concrete things, and applications, let us formulate:

DEFINITION 2.12. *The compact and connected sets are defined as follows:*

(1) *Compact means that any open cover has a finite subcover.*
(2) *Connected means that it cannot be broken into two parts.*

As basic examples, the closed bounded intervals $[a, b]$ are compact, as we know from the proof of Theorem 2.6, and so are the finite unions of such intervals. As for connected sets, the basic examples here are the various types of intervals, namely (a, b), $(a, b]$, $[a, b)$, $[a, b]$, and it looks impossible to come up with more examples. In fact, we have:

THEOREM 2.13. *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) The fact that compact implies both closed and bounded is clear from our definition of compactness, because assuming non-closedness or non-boundedness leads to an open cover having no finite subcover. As for the converse, we know from the proof of Theorem 2.6 that any closed bounded interval $[a, b]$ is compact, and it follows that any $K \subset \mathbb{R}$ closed and bounded is a closed subset of a compact set, which follows to be compact.

(2) This is something which is obvious, and this regardless of what "cannot be broken into parts" in Definition 2.12 exactly means, mathematically speaking, with several possible definitions being possible here, all being equivalent. Indeed, $E \subset \mathbb{R}$ having this property is equivalent to $a, b \in E \implies [a, b] \subset E$, and this gives the result. □

We will be back to all this later in this book, when looking at open, closed, compact and connected sets in $\mathbb{R}^N$, or more general spaces, where things are more complicated than in $\mathbb{R}$. Now with this discussed, let us go back to continuous functions. We have:

THEOREM 2.14. *Assuming that f is continuous:*

(1) *If K is compact, then $f(K)$ is compact.*
(2) *If E is connected, then $f(E)$ is connected.*

PROOF. These assertions both follow from our definition of compactness and connectedness, as formulated in Definition 2.12. To be more precise:

(1) This comes from the fact that if a function f is continuous, then the inverse function f^{-1} returns an open cover into an open cover.

(2) This is something clear as well, because if $f(E)$ can be split into two parts, then by applying f^{-1} we can split as well E into two parts. □

Let us record as well the following useful generalization of Theorem 2.6:

THEOREM 2.15. *Any continuous function defined on a compact set*

$$f : X \to \mathbb{R}$$

is automatically uniformly continuous.

PROOF. We can prove this exactly as Theorem 2.6, by using the compactness of X. □

You might perhaps ask at this point, were Theorems 2.14 and 2.15 worth all this excursion into open and closed sets. Good point, and here is our answer, a beautiful and powerful theorem based on the above, which can be used for a wide range of purposes:

THEOREM 2.16. *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 are related, and are called altogether "intermediate value theorem". Regarding now the proof, one way of viewing things is that since $[a, b]$ is compact and connected, the set $f([a, b])$ is compact and connected too, and so it is a certain closed bounded interval $[c, d]$, and this gives all the results. However, this is based on rather advanced technology, and it is possible to prove (1-3) directly as well. □

Along the same lines, we have as well the following result:

THEOREM 2.17. *Assuming that a function f is continuous and invertible, this function must be monotone, and its inverse function f^{-1} must be monotone and continuous too. Moreover, this statement holds both locally, and globally.*

PROOF. The fact that both f and f^{-1} are monotone follows from Theorem 2.16. 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 a basic application of Theorem 2.17, we have:

PROPOSITION 2.18. *The various usual inverse functions, such as the inverse trigonometric functions* arcsin*,* arccos*,* arctan*,* arccot*, are all continuous.*

PROOF. This follows indeed from Theorem 2.17, with a course the full discussion needing some explanations on bijectivity and domains. But you surely know all that, and in what concerns us, our claim is simply that these beasts are all continuous, proved. □

As another basic application of this, we have:

PROPOSITION 2.19. *The following happen:*

(1) *Any polynomial $P \in \mathbb{R}[X]$ of odd degree has a root.*
(2) *Given $n \in 2\mathbb{N}+1$, we can extract $\sqrt[n]{x}$, for any $x \in \mathbb{R}$.*
(3) *Given $n \in \mathbb{N}$, we can extract $\sqrt[n]{x}$, for any $x \in [0,\infty)$.*

PROOF. All these results come as applications of Theorem 2.16, as follows:

(1) This is clear from Theorem 2.16 (3), applied on $[-\infty,\infty]$.

(2) This follows from (1), by using the polynomial $P(z) = z^n - x$.

(3) This follows as well by applying Theorem 2.16 (3) to the polynomial $P(z) = z^n - x$, but this time on $[0,\infty)$. □

There are many other things that can be said about roots of polyomials, and solutions of other equations of type $f(x) = 0$, by using Theorem 2.16. We will be back to this.

As a concrete application, in relation with powers, we have the following result, completing our series of results regarding the basic mathematical functions:

THEOREM 2.20. *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. The idea is that we know how to extract roots by using Proposition 2.19, and all the rest follows by continuity. To be more precise:

(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 2.19:

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

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 $q = p/q$ 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 2.19. 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 2.21. *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 2.20, by exchanging the variables, $x \leftrightarrow a$. To be more precise, the situation is as follows:

(1) We know from Theorem 2.20 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 2.20 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, we have been quite successful with our theory of continuous functions, having how full results, regarding the definition and continuity property, for all basic functions from mathematics. All this is of course just a beginning, and we will be back to these functions on regular occasions, in what follows. In particular, we will discuss the function a^x at the special value $a = e$, and its inverse $\log x$, at the end of this chapter.

2c. Sequences and series

Our goal now is to extend the material from chapter 1 regarding the numeric sequences and series, to the case of the sequences and series of functions. To start with, we can talk about the convergence of sequences of functions, $f_n \to f$, as follows:

DEFINITION 2.22. *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, and with a death penalty coming for that. In fact, the result itself is wrong in general, because if we consider the functions $f_n : [0,1] \to \mathbb{R}$ given by $f_n(x) = x^n$, which are obviously continuous, their limit is discontinuous, given by:

$$\lim_{n\to\infty} x^n = \begin{cases} 0 & , \quad x \in [0,1)\\ 1 & , \quad x = 1\end{cases}$$

Of course, you might say here that allowing $x = 1$ in all this might be a bit unnatural, for whatever reasons, but there is an answer to this too. We can do worse, as follows:

PROPOSITION 2.23. *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 modifications of $\arctan(x)$*. Even worse, there are examples of $f_n \to f$ with each f_n continuous, and with f totally discontinuous.*

PROOF. To start with, $\arctan(x)$ 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 $[-1, 1]$. Thus, we must first rescale $\arctan(x)$ by $\pi/2$. Now with this done, we can further stretch the variable x, as to get our function closer and closer to $sgn(x)$, as desired. This proves the first assertion, and the second assertion, which is a bit more technical, and that we will not really need in what follows, is left as an exercise for you, reader. □

Sumarizing, we are a bit in trouble, because we would like to have in our bag of theorems something saying that $f_n \to f$ with f_n continuous implies f continuous. Fortunately, this can be done, with a suitable refinement of the notion of convergence, as follows:

DEFINITION 2.24. *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. In order to understand this, which is something quite subtle, let us compare Definition 2.22 and Definition 2.24. As a first observation, we have:

PROPOSITION 2.25. *Uniform convergence implies pointwise convergence,*

$$f_n \to_u f \implies f_n \to f$$

but the converse is not true, in general.

PROOF. Here the first assertion is clear from definitions, just by thinking at what is going on, with no computations needed. 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 Proposition 2.23. Indeed, uniform convergence on $[0, 1)$ would mean:

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, x^n < \varepsilon, \forall x \in [0, 1)$$

But this is wrong, because no matter how big N is, we have $\lim_{x\to 1} x^N = 1$, and so we can find $x \in [0, 1)$ such that $x^N > \varepsilon$. Thus, we have our counterexample. □

Moving ahead now, let us state our main theorem on uniform convergence, as follows:

THEOREM 2.26. *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. As previously advertised, it is the "$\forall x$ at the end" in Definition 2.24 that will make this work. Indeed, let us try to prove that the limit f is continuous at some point x. For this, we pick a number $\varepsilon > 0$. Since $f_n \to_u f$, we can find $N \in \mathbb{N}$ such that:

$$|f_N(z) - f(z)| < \frac{\varepsilon}{3} \quad , \quad \forall z$$

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

$$|x - y| < \delta \implies |f_N(x) - f_N(y)| < \frac{\varepsilon}{3}$$

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. □

Obviously, the notion of uniform convergence in Definition 2.24 is something quite interesting, worth some more study. As a first result, we have:

PROPOSITION 2.27. *The following happen, regarding 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* $f_ng_n\to_u fg$.
(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, that we know well, with no difficulties or tricks involved. □

Finally, there is some abstract mathematics to be done as well. Indeed, observe that the notion of uniform convergence, as formulated in Definition 2.24, means that:

$$\sup_x\big|f_n(x)-f(x)\big|\ \longrightarrow_{n\to\infty}\ 0$$

This suggests measuring the distance between functions via a supremum as above, and in relation with this, we have the following result:

THEOREM 2.28. *The uniform convergence, $f_n\to_u f$, means that we have $f_n\to f$ with respect to the following distance,*

$$d(f,g)=\sup_x\big|f(x)-g(x)\big|$$

which is indeed a distance function.

PROOF. Here the fact that d is indeed a distance follows from definitions, and the fact that the uniform convergence can be interpreted as above is clear as well. □

Finally, regarding the series, some general theory can be developed here as well, in connection with the notion of uniform convergence, and in connection with the notion of convergence radius. We will see applications of all this, in a moment.

2d. Basic functions

With the above theory in hand, let us get now to interesting things, namely computations. Among others, because this is what a mathematician's job is, doing all sorts of weird computations. We will be mainly interested in the functions x^a and a^x, which remain something quite mysterious. Regarding x^a, we first have the following result:

THEOREM 2.29. *We have the generalized binomial formula*

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

with the generalized binomial coefficients being given by

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

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

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

(1) For exponents $a\in\mathbb{N}$, 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, namely:

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

(2) For the exponent $a=-1$ 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 $a=-1$, because:

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

(3) Let us discuss now the general case $a\in-\mathbb{N}$. With $a=-n$, 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 $a = -n$, and $n \in \mathbb{N}$, 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:

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

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. □

Quite interestingly, the formula in Theorem 2.29 holds in fact at any $a \in \mathbb{R}$, but this is something non-trivial, whose proof will have to wait until chapter 3 below. However, in the meantime, let us investigate the case $a \in \mathbb{Z}/2$. Indeed, not only the results here are interesting, and very useful in practice, but also they can be proved with elementary methods. At $a = \pm 1/2$, to start with, we have the following result:

THEOREM 2.30. *The generalized binomial formula, namely*

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

holds as well at $a=\pm 1/2$. In practice, at $a=1/2$ we obtain the 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, and at $a=-1/2$ we obtain

$$\frac{1}{\sqrt{1+t}}=\sum_{k=0}^{\infty}D_k\left(\frac{-t}{4}\right)^k$$

with $D_k=\binom{2k}{k}$ being the central binomial coefficients.

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

(1) At $a=1/2$, the generalized binomial coefficients are as follows:

$$\begin{aligned}\binom{1/2}{k}&=\frac{1/2(-1/2)\ldots(3/2-k)}{k!}\\&=(-1)^{k-1}\frac{1\cdot 3\cdot 5\ldots(2k-3)}{2^k k!}\\&=(-1)^{k-1}\frac{(2k-2)!}{2^{k-1}(k-1)!2^k k!}\\&=-2\left(\frac{-1}{4}\right)^k C_{k-1}\end{aligned}$$

(2) At $a=-1/2$, the generalized binomial coefficients are as follows:

$$\begin{aligned}\binom{-1/2}{k}&=\frac{-1/2(-3/2)\ldots(1/2-k)}{k!}\\&=(-1)^k\frac{1\cdot 3\cdot 5\ldots(2k-1)}{2^k k!}\\&=(-1)^k\frac{(2k)!}{2^k k!2^k k!}\\&=\left(\frac{-1}{4}\right)^k D_k\end{aligned}$$

(3) Summarizing, we have proved so far that the binomial formula at $a=\pm 1/2$ is equivalent to the explicit formulae in the statement, involving the Catalan numbers C_k,

and the central binomial coefficients D_k. It remains now to prove that these two explicit formulae hold indeed. For this purpose, let us write these formulae as follows:

$$\sqrt{1-4t}=1-2\sum_{k=1}^{\infty}C_{k-1}t^k\quad,\quad\frac{1}{\sqrt{1-4t}}=\sum_{k=0}^{\infty}D_kt^k$$

In order to check these latter formulae, we must prove the following identities:

$$\left(1-2\sum_{k=1}^{\infty}C_{k-1}t^k\right)^2=1-4t\quad,\quad\left(\sum_{k=0}^{\infty}D_kt^k\right)^2=\frac{1}{1-4t}$$

(4) As a first observation, the formula on the left is equivalent to:

$$\sum_{k+l=n}C_kC_l=C_{n+1}$$

By using the series for $1/(1-4t)$, the formula on the right is equivalent to:

$$\sum_{k+l=n}D_kD_l=4^n$$

Finally, observe that if our formulae hold indeed, by multiplying we must have:

$$\sum_{k+l=n}C_kD_l=\frac{D_{n+1}}{2}$$

(5) Summarizing, we have to understand 3 formulae, which look quite similar. Let us first attempt to prove $\sum_{k+l=n}D_kD_l=4^n$, by recurrence. We have:

$$D_{k+1}=\binom{2k+2}{k+1}=\frac{4k+2}{k+1}\binom{2k}{k}=\left(4-\frac{2}{k+1}\right)D_k$$

Thus, assuming that we have $\sum_{k+l=n}D_kD_l=4^n$, we obtain:

$$\begin{aligned}\sum_{k+l=n+1}D_kD_l&=D_0D_{n+1}+\sum_{k+l=n}\left(4-\frac{2}{k+1}\right)D_kD_l\\&=D_{n+1}+4\sum_{k+l=n}D_kD_l-2\sum_{k+l=n}\frac{D_kD_l}{k+1}\\&=D_{n+1}+4^{n+1}-2\sum_{k+l=n}C_kD_l\end{aligned}$$

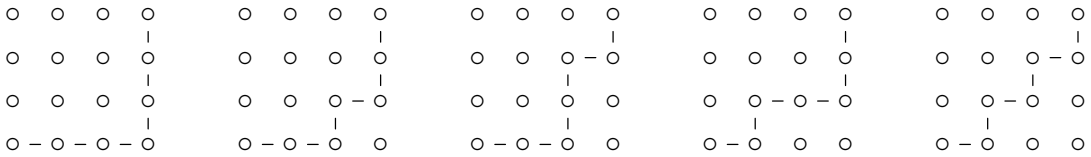
Thus, this leads to a sort of half-failure, the conclusion being that for proving by recurrence the second formula in (4), we need the third formula in (4).

(6) All this suggests a systematic look at the three formulae in (4). According to our various observations above, these three formulae are equivalent, and so it is enough to

prove one of them. We will chose here to prove the first one, namely:

$$\sum_{k+l=n} C_k C_l = C_{n+1}$$

(7) For this purpose, we must trick. Let us count the Dyck paths in the plane, which are by definition the paths from $(0,0)$ to (n,n), marching North-East over the integer lattice $\mathbb{Z}^2 \subset \mathbb{R}^2$, by staying inside the square $[0,n] \times [0,n]$, and staying as well under the diagonal of this square. As an example, here are the 5 possible Dyck paths at $n = 3$:



In fact, the number C'_n of these paths is as follows, coinciding with C_n:

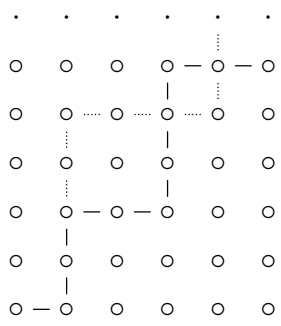
$$1, 1, 2, 5, 14, 42, 132, 429, \ldots$$

(8) So, here is our trick. We will prove on one hand that the numbers C'_n satisfy the recurrence for the numbers C_n that we want to prove, from (6), and on the other hand we will prove that we have $C'_n = C_n$. Which is smart, isn't it. Getting to work now, in what regards our first task, this is easy, because when looking when our path last intersects the diagonal of the square, we obtain the recurrence relation that we want, namely:

$$C'_n = \sum_{k+l=n-1} C'_k C'_l$$

(9) In what regards now our second task, proving that we have $C'_n = C_n$, this is more tricky. If we ignore the assumption that our path must stay under the diagonal of the square, we have $\binom{2n}{n}$ 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 $(n,n+1)$. 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 (n,n), but rather in $(n-1,n+1)$. Moreover, more is true, in the sense that, by thinking a bit, we see that the flipped bad paths are precisely those ending in $(n-1,n+1)$. 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:

$$\begin{aligned} C_n' &= \binom{2n}{n} - \binom{2n}{n-1} \\ &= \binom{2n}{n} - \frac{n}{n+1}\binom{2n}{n} \\ &= \frac{1}{n+1}\binom{2n}{n} \end{aligned}$$

Thus we have indeed $C_n' = C_n$, and this finishes the proof. □

As already mentioned, the binomial formula holds in fact for any exponent $a \in \mathbb{Z}/2$, after some combinatorial pain, and even for any $a \in \mathbb{R}$, but this is non-trivial, and the elementary study stops with Theorem 2.30. However, we will see in chapter 3 below that the problem can be solved, and in a very elegant way, by guess whom: calculus.

As another application of our methods, let us get now into the other version of the exponential function, namely a^x. The idea is that some very interesting results appear with $a = e$, the number that we know from chapter 1. We first have:

PROPOSITION 2.31. *We have the following formula,*

$$\left(1+\frac{x}{n}\right)^n \to e^x$$

valid for any $x \in \mathbb{R}$.

PROOF. We already know from chapter 1 that the result holds at $x = 1$, and this because the number e was by definition given by the following formula:

$$\left(1+\frac{1}{n}\right)^n \to e$$

By taking inverses, we obtain 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. □

We have the following result, which is something quite far-reaching:

THEOREM 2.32. *We have the formula*

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

valid for any $x \in \mathbb{R}$.

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

(1) At $x = 1$, which is the key step, we want to prove that we have the following equality, between the sum of a series, and a limit of a sequence:

$$\sum_{k=0}^{\infty} \frac{1}{k!} = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n$$

(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:

$$\begin{aligned}
\sum_{k=0}^{n}\frac{1}{k!}-\left(1+\frac{1}{n}\right)^n &= \sum_{k=0}^{n}\frac{1}{k!}-\sum_{k=0}^{n}\frac{n(n-1)\ldots(n-k+1)}{k!}\cdot\frac{1}{n^k}\\
&= \sum_{k=2}^{n}\frac{1}{k!}-\sum_{k=2}^{n}\frac{n(n-1)\ldots(n-k+1)}{k!}\cdot\frac{1}{n^k}\\
&= \sum_{k=2}^{n}\frac{n^k-n(n-1)\ldots(n-k+1)}{n^kk!}\\
&\leq \sum_{k=2}^{n}\frac{n^k-(n-k)^k}{n^kk!}\\
&= \sum_{k=2}^{n}\frac{1-\left(1-\frac{k}{n}\right)^k}{k!}
\end{aligned}$$

(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 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 $\left(1+\frac{1}{n}\right)^n$, manely e. Thus, getting back now to what we wanted to prove, our theorem, we are done in this way with the case $x=1$.

(6) In order to deal now with the general case, consider the following function:

$$f(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Observe that, by using our various results above, this function is indeed well-defined. Moreover, again by using our various results above, f is continuous.

(7) Our next claim, which is the key one, is that we have:

$$f(x+y) = f(x)f(y)$$

Indeed, by using the binomial formula, we have the following computation:

$$\begin{aligned} f(x+y) &= \sum_{k=0}^{\infty} \frac{(x+y)^k}{k!} \\ &= \sum_{k=0}^{\infty} \sum_{s=0}^{k} \binom{k}{s} \cdot \frac{x^s y^{k-s}}{k!} \\ &= \sum_{k=0}^{\infty} \sum_{s=0}^{k} \frac{x^s y^{k-s}}{s!(k-s)!} \\ &= f(x)f(y) \end{aligned}$$

(8) In order to finish now, we know that our function f is continuous, that it satisfies $f(x+y) = f(x)f(y)$, and that we have:

$$f(0) = 1 \quad , \quad f(1) = e$$

But it is easy to prove that such a function is necessarily unique, and since e^x obviously has all these properties too, we must have $f(x) = e^x$, as desired. □

We will be back to all this, and to the logarithm and trigonometric functions as well, in chapter 3 below, when talking about derivatives and the Taylor formula.

2e. Exercises

There are many possible exercises on the above, and we have here:

Exercise 2.33. *Find some interesting examples of Lipschitz functions.*

Exercise 2.34. *Find a simple proof for the intermediate value theorem.*

Exercise 2.35. *Prove the generalized binomial formula, for exponents $a \in \mathbb{Z}/2$.*

Exercise 2.36. *Rewrite the theory of e, with $e = \sum_k 1/k!$ as definition.*

For the rest, business as usual, more exercises are easy to find. Find and solve them.

CHAPTER 3

Derivatives

3a. Derivatives, rules

Welcome to calculus. In this chapter we go for the real thing, namely developement of modern calculus, following some amazing ideas of Newton, Leibnitz and others. The material will be quite difficult, mixing geometry and intuition with formal mathematics and computations, and needing some time to be understood. But we will survive.

The basic idea of calculus is very simple. We are interested in functions $f : \mathbb{R} \to \mathbb{R}$, and we already know that when f is continuous at a point x, we can write an approximation as follows, for the values of our function f around that point x:

$$f(x+t) \simeq f(x)$$

The problem is now, how to improve this? And a bit of thinking at all this suggests to look at the slope of f at the point x. Which leads us into the following notion:

DEFINITION 3.1. *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.

As a first 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, a basic counterexample being $f(x) = |x|$ at $x = 0$. Let us summarize these findings as follows:

PROPOSITION 3.2. *If f is differentiable at x, then f must be continuous at x. However, the converse is not true, a basic counterexample being $f(x) = |x|$, at $x = 0$.*

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 3.1. 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 3.1 does not converge, so we have our counterexample. □

Generally speaking, the last assertion in Proposition 3.2 should not bother us much, because most of the basic continuous functions are differentiable, and we will see examples in a moment. 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 3.3. *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. □

All this is very nice, and before developing more theory, let us work out some examples. As a first illustration, the derivatives of the power functions are as follows:

PROPOSITION 3.4. *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 $m\in\mathbb{N}$ and $n\in\mathbb{Z}$. 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 in this way:

$$\begin{aligned}(x+t)^{m/n} - x^{m/n} &= \frac{(x+t)^m - x^m}{(x+t)^{m(n-1)/n} + \ldots + x^{m(n-1)/n}}\\ &\simeq \frac{(x+t)^m - x^m}{nx^{m(n-1)/n}}\\ &\simeq \frac{mx^{m-1}t}{nx^{m(n-1)/n}}\\ &= \frac{m}{n} \cdot x^{m-1-m+m/n} \cdot t\\ &= \frac{m}{n} \cdot x^{m/n-1} \cdot t\end{aligned}$$

(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. □

Here are some further computations, for other basic functions that we know:

PROPOSITION 3.5. *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 fact, which is clear on pictures, by drawing the trigonometric circle, 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 $\cos t$ 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:

THEOREM 3.6. *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:

$$1c_1 = 1 \quad , \quad 2c_2 = c_1 \quad , \quad 3c_3 = c_2 \quad , \quad 4c_4 = c_3 \quad , \quad \ldots$$

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:

$$(e^x)' = e^x$$

Let us work out now some general results. We have here the following statement:

THEOREM 3.7. *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:

PROPOSITION 3.8. *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 written a bit upside down, and for the proof it is better to proceed backwards. To be more precise, by using $(x^p)' = px^{p-1}$ and Theorem 3.7 (3), we obtain the third formula. Then, with $p = -1$, we obtain from this the second formula. And finally, by using this second formula and Theorem 3.7 (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 formulae 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 3.9. *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, 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, including $\tan x$, or $\arctan x$:

PROPOSITION 3.10. *We have the following formulae,*

$$(\tan x)' = \frac{1}{\cos^2 x} \quad , \quad (\arctan x)' = \frac{1}{1 + x^2}$$

and the derivatives of the remaining trigonometric functions can be computed as well.

PROOF. For tan, we have the following computation:

$$\begin{aligned}(\tan x)' &= \left(\frac{\sin x}{\cos x}\right)' \\ &= \frac{\sin' x\cos x-\sin x\cos' x}{\cos^2 x} \\ &= \frac{\cos^2 x+\sin^2 x}{\cos^2 x} \\ &= \frac{1}{\cos^2 x}\end{aligned}$$

As for arctan, we can use here 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:

$$\arctan'(x)=\cos^2(\arctan x)$$

On the other hand, with $t=\arctan x$ we know that we have $\tan t=x$, and so:

$$\cos^2(\arctan x)=\cos^2 t=\frac{1}{1+\tan^2 t}=\frac{1}{1+x^2}$$

Thus, we are led to the formula in the statement, namely:

$$(\arctan x)'=\frac{1}{1+x^2}$$

As for the last assertion, we will leave this as an exercise. □

At the theoretical level now, further building on Theorem 3.3, we have:

THEOREM 3.11. *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 in Theorem 3.3, namely:

$$f(x+t)\simeq f(x)+f'(x)t$$

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. □

As an important consequence of Theorem 3.11, we have:

THEOREM 3.12. *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. In the case $f(a) = f(b)$, the result, called Rolle theorem, states that we have $f'(c) = 0$ for some $c \in (a, b)$, and follows from Theorem 3.11. Now in what regards our statement, due to Lagrange, this follows from Rolle, applied to the following function:

$$g(x) = f(x) - \frac{f(b) - f(a)}{b - a} \cdot x$$

Indeed, we have $g(a) = g(b)$ due to our choice of the constant on the right, so we get $g'(c) = 0$ for some $c \in (a, b)$, which translates into the formula in the statement. □

In practice, Theorem 3.11 can be used in order to find the maximum and minimum of any differentiable function, and the method is best recalled as follows:

ALGORITHM 3.13. *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 these $f(x)$ values.*
(6) *Then this is the min/max of your function.*

To be more precise, we are using here Theorem 3.11, or rather the obvious extension of this result to the case of the functions $f : [a, b] \to \mathbb{R}$. This tells us that the local minima and maxima of our function f, and in particular the global minima and maxima, can be found among the zeroes of the first derivative f', with the endpoints a, b added. Thus, what we have to do is to compute these "candidates", as explained in steps (1-2-3), and then see what each candidate is exactly worth, as explained in steps (4-5-6).

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.

3b. 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 3.14. *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 first have:

INTERPRETATION 3.15. *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, (1) follows from the usual interpretation of the derivative, as both a growth rate, and a slope. Regarding (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:

$$a = \dot{v} \quad , \quad v = \dot{x}$$

To be more precise, here a, v, x are the acceleration, speed and position, and the dot denotes the time derivative, and according to these equations, we have $a = \ddot{x}$, second derivative. We will be back to these equations at the end of the present chapter.

Regarding now (3) in the above, this is something more subtle, of statistical nature, that we will clarify with some mathematics, in a moment. As for (4), this is something quite subtle too, that we will again clarify with some mathematics, in a moment.

All in all, what we have above 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.

In practice now, let us first compute the second derivatives of the functions that we are familiar with, see what we get. The result here, which is perhaps not very enlightening at this stage of things, but which certainly looks technically useful, is as follows:

PROPOSITION 3.16. *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) $\log'(x) = -1/x^2$.

Also, there are functions which are differentiable, but not twice differentiable.

PROOF. We have two assertions here, the idea being as follows:

(1) Regarding the various formulae in the statement, these all follow from the various formulae for the derivatives established before, as follows:

$$(x^p)'' = (px^{p-1})' = p(p-1)x^{p-2}$$

$$(\sin x)'' = (\cos x)' = -\sin x$$

$$(\cos x)'' = (-\sin x)' = -\cos x$$

$$(e^x)'' = (e^x)' = e^x$$

$$(\log x)'' = (-1/x)' = -1/x^2$$

Of course, this is not the end of the story, because these formulae remain quite opaque, and must be examined in view of Interpretation 3.15, in order to see what exactly is going on. Also, we have $\tan(x)$ and the inverse trigonometric functions too. In short, plenty of interesting exercises here, and the more you solve, the better your calculus will be.

(2) Regarding the 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, let us set:

$$f(x) = \begin{cases} ax^2 & (x \leq 0) \\ bx^2 & (x \geq 0) \end{cases}$$

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. □

Getting now to theory, we first have the following key result:

THEOREM 3.17. *Any function of one variable $f:\mathbb{R}\to\mathbb{R}$ is locally quadratic,*

$$f(x+t)\simeq f(x)+f'(x)t+\frac{f''(x)}{2}\,t^2$$

where $f''(x)$ is the derivative of the function $f':\mathbb{R}\to\mathbb{R}$ at the point x.

PROOF. 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:

$$f(x+t)\simeq a+bt+ct^2$$

We must have $a=f(x)$, and we also know from Theorem 3.3 that $b=f'(x)$ is the correct choice for the coefficient of t. Thus, our approximation must be as follows:

$$f(x+t)\simeq f(x)+f'(x)t+ct^2$$

In order to find the correct choice for $c\in\mathbb{R}$, observe that the function $t\to f(x+t)$ matches with $t\to f(x)+f'(x)t+ct^2$ in what regards the value at $t=0$, and also in what regards the value of the derivative at $t=0$. Thus, the correct choice of $c\in\mathbb{R}$ should be the one making match the second derivatives at $t=0$, and this gives:

$$f''(x)=2c$$

We are therefore led to the formula in the statement, namely:

$$f(x+t)\simeq f(x)+f'(x)t+\frac{f''(x)}{2}\,t^2$$

In order to prove now that this formula holds indeed, we will use L'Hôpital's rule, which states that the $0/0$ type limits can be computed as follows:

$$\frac{f(x)}{g(x)}\simeq\frac{f'(x)}{g'(x)}$$

Observe that this formula holds indeed, as an application of Theorem 3.3. Now by using this, if we denote by $\varphi(t) \simeq P(t)$ the formula to be proved, we have:

$$\begin{aligned}\frac{\varphi(t)-P(t)}{t^2} &\simeq \frac{\varphi'(t)-P'(t)}{2t}\\ &\simeq \frac{\varphi''(t)-P''(t)}{2}\\ &= \frac{f''(x)-f''(x)}{2}\\ &= 0\end{aligned}$$

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

The above result substantially improves Theorem 3.3, and there are many applications of itt. As a first such application, justifying Interpretation 3.15 (3), we have the following statement, which is a bit heuristic, but we will call it however Proposition:

PROPOSITION 3.18. *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 3.17, 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:

$$\frac{f(x+t)+f(x-t)}{2} = f(x) + \frac{f''(x)}{2}\,t^2$$

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:

$$I(f) = f(x) + f''(x)I\left(\frac{t^2}{2}\right)$$

But this is what our statement says, save for some uncertainties regarding the averaging method, and the precise value of $I(t^2/2)$. We will leave this for later. □

Back to rigorous mathematics now, and of course with apologies for the physics intermezzo, but Proposition 3.18 is really cool isn't it, and we will be back later to this with full mathematical details, after developing more theory, that is promised, as a second application of Theorem 3.17, we can improve as well Theorem 3.11, as follows:

THEOREM 3.19. *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 corresponding to the case $f'(x) \geq 0$, and with the local maxima corresponding to 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 3.17, 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 Theorem 3.11, the above result is not the end of the story with the mathematics 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 $\pm x^{2m+1}$.

There are some comments to be made in relation with Algorithm 3.13 as well. Normally that algorithm stays strong, because Theorem 3.19 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 3.19 is good to know, when applying that algorithm.

As a main concrete application now of the second derivative, which is something very useful in practice, and related to Interpretation 3.15 (4), we have the following result:

THEOREM 3.20. *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 $x_1 = \ldots = x_N$. We have 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:

$$f(\lambda_1x_1+\ldots+\lambda_Nx_N)\leq\lambda_1f(x_1)+\ldots+\lambda_Nx_N$$

(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, 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 second result on the subject, which is very classical as well, we have:

THEOREM 3.21. *For $p\in(1,\infty)$ we have the following Hölder inequality,*

$$\left|\frac{x_1+\ldots+x_N}{N}\right|^p\leq\frac{|x_1|^p+\ldots+|x_N|^p}{N}$$

and for $p\in(0,1)$ we have the following reverse Hölder 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 3.20, 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 Hölder inequality the Cauchy-Schwarz inequality, or rather something equivalent to it, namely:

$$\left(\frac{x_1 + \ldots + x_N}{N}\right)^2 \leq \frac{x_1^2 + \ldots + x_N^2}{N}$$

We will be back to this later on in this book, when talking scalars products and Hilbert spaces, with some more conceptual proofs for such inequalities.

Finally, as yet another important application of the Jensen inequality, we have:

THEOREM 3.22. *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. □

Observe that for the simplest exponents, namely $p = q = 2$, the Young inequality gives something which is trivial, but is very useful and basic, namely:

$$ab \leq \frac{a^2 + b^2}{2}$$

In general, the Young inequality is something non-trivial, and the idea with it is that "when stuck with a problem, and with $ab \leq \frac{a^2+b^2}{2}$ not working, try Young". We will be back to this general principle, later in this book, with some illustrations.

3c. The Taylor formula

We can further develop our approximation method, at order 3, at order 4, and so on, the ultimate result on the subject, called Taylor formula, being as follows:

THEOREM 3.23. *Any function $f : \mathbb{R} \to \mathbb{R}$ can be locally approximated as*

$$f(x+t) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x)}{k!}\, t^k$$

where $f^{(k)}(x)$ are the higher derivatives of f at the point x.

PROOF. Consider the function to be approximated, namely:

$$\varphi(t) = f(x+t)$$

Let us try to best approximate this function at a given order $n \in \mathbb{N}$. 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 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. Thus, we are led to the approximation in the statement:

$$f(x+t) \simeq \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k$$

In order to prove now that this approximation holds indeed, we use L'Hôpital's rule, applied several times, as in the proof of Theorem 3.17. 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 3.24. *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:

$$(x+t)^p \simeq \sum_{k=0}^{p} \frac{p(p-1)\ldots(p-k+1)}{k!}\, x^{p-k}t^k$$

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. □

There are many other things that can be said about the Taylor formula, at the theoretical level, and we will be back to this later, in chapter 4 below.

As an application of the Taylor formula, we can now improve the binomial formula, which was actually our main tool so far, in the following way:

THEOREM 3.25. *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. It is customary to divide everything by x, which is the same as assuming $x = 1$. The formula to be proved is then as follows, under the assumption $|t| < 1$:

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

Let us discuss now the validity of this formula, depending on $p \in \mathbb{R}$:

(1) Case $p \in \mathbb{N}$. According to our definition of the generalized binomial coefficients, we have $\binom{p}{k} = 0$ for $k > p$, so the series is stationary, and the formula to be proved is:

$$(1+t)^p = \sum_{k=0}^{p} \binom{p}{k} t^k$$

But this is the usual binomial formula, which holds for any $t \in \mathbb{R}$.

(2) Case $p = -1$. Here we can use the following formula, valid for $|t| < 1$:

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

But this is exactly our generalized binomial formula at $p = -1$, because:

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

(3) Case $p \in -\mathbb{N}$. This is a continuation of our study at $p = -1$, which will finish the study at $p \in \mathbb{Z}$. With $p = -m$, the generalized binomial coefficients are:

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

Thus, our generalized binomial formula at $p = -m$ reads:

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

But this is something which holds indeed, as we know from chapter 2.

(4) General case, $p \in \mathbb{R}$. As we can see, things escalate quickly, so we will skip the next step, $p \in \mathbb{Q}$, and discuss directly the case $p \in \mathbb{R}$. Consider the following function:

$$f(x) = x^p$$

The derivatives at $x = 1$ are then given by the following formula:

$$f^{(k)}(1) = p(p-1)\ldots(p-k+1)$$

Thus, the Taylor approximation at $x = 1$ is as follows:

$$f(1+t) = \sum_{k=0}^{\infty}\frac{p(p-1)\ldots(p-k+1)}{k!}\,t^k$$

But this is exactly our generalized binomial formula, so we are done with the case where t is small. With a bit more care, we obtain that this holds for any $|t| < 1$. □

We can see from the above the power of the Taylor formula, saving us from quite complicated combinatorics. Remember indeed the mess from chapter 2.

As a main application now of our generalized binomial formula, which is something very useful in practice, we can extract square roots, as follows:

PROPOSITION 3.26. *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 something that we already know from chapter 2, but time to review all this. At $p = 1/2$, the generalized binomial coefficients are:

$$\begin{aligned}\binom{1/2}{k} &= \frac{1/2(-1/2)\ldots(3/2-k)}{k!}\\ &= (-1)^{k-1}\frac{(2k-2)!}{2^{k-1}(k-1)!2^k k!}\\ &= -2\left(\frac{-1}{4}\right)^k C_{k-1}\end{aligned}$$

Also, at $p = -1/2$, the generalized binomial coefficients are:

$$\begin{aligned}\binom{-1/2}{k} &= \frac{-1/2(-3/2)\ldots(1/2-k)}{k!}\\ &= (-1)^k\frac{(2k)!}{2^k k!2^k k!}\\ &= \left(\frac{-1}{4}\right)^k D_k\end{aligned}$$

Thus, Theorem 3.25 at $p = \pm 1/2$ gives the formulae in the statement. □

As another basic application of the Taylor series, we have:

THEOREM 3.27. *We have the following formulae,*

$$\sin x = \sum_{l=0}^{\infty}(-1)^l\frac{x^{2l+1}}{(2l+1)!} \quad , \quad \cos x = \sum_{l=0}^{\infty}(-1)^l\frac{x^{2l}}{(2l)!}$$

as well as the following formulae,

$$e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!} \quad , \quad \log(1+x) = \sum_{k=0}^{\infty}(-1)^{k+1}\frac{x^k}{k}$$

as Taylor series, and in general as well, with $|x| < 1$ *needed for* $\log$.

PROOF. There are several statements here, the proofs being as follows:

(1) Regarding sin and cos, we can use here the following formulae:

$$(\sin x)' = \cos x \quad , \quad (\cos x)' = -\sin x$$

Thus, we can differentiate sin and cos as many times as we want to, so we can compute the corresponding Taylor series, and we obtain the formulae in the statement.

(2) Regarding exp and log, here the needed formulae, which lead to the formulae in the statement for the corresponding Taylor series, are as follows:

$$(e^x)' = e^x$$

$$(\log x)' = x^{-1}$$

$$(x^p)' = px^{p-1}$$

(3) Finally, the fact that the formulae in the statement extend beyond the small t setting, coming from Taylor series, is something standard too. □

3d. Differential equations

Good news, with the calculus that we know we can do some physics, in 1 dimension. Let us start with something immensely important, in the history of science:

FACT 3.28. *Newton invented calculus for formulating the laws of motion as*

$$a = \dot{v} \quad , \quad v = \dot{x}$$

where a, v, x are the acceleration, speed and position, and the dots are time derivatives.

To be more precise, the variable in Newton's physics is time $t \in \mathbb{R}$, playing the role of the variable $x \in \mathbb{R}$ that we have used in the above. And we are looking at a particle whose position is described by a function $x = x(t)$. Then, it is quite clear that the speed of this particle should be described by the first derivative $v = x'(t)$, and that the acceleration of the particle should be described by the second derivative $a = v'(t) = x''(t)$.

So, this was for the story, and there are some further related stories, involving an apple falling from a tree, on the ground, or even on Newton's head, according to certain accounts. According however to some other accounts, it all started with Newton's cat hunting a bird in that same apple tree, and whether the cat made fall the apple, or even fell himself, on the ground, or on Newton's head, remains a matter of debate.

Leaving this to science historians, and getting back to serious matters, let us mention too that higher derivatives can be considered. For instance the following quantity is called in physics the jerk, in analogy with a jerk driving an open-exhaust Lamborghini:

$$j = \dot{a} = \ddot{v} = \dddot{x}$$

Summarizing, with Newton's theory of derivatives, as we learned it in this chapter, we can certainly do some mathematics for the motion of bodies. But, for these bodies to move, we need them to be acted upon by some forces, right? The simplest such force is gravity, and in our present, modest 1 dimensional setting, we have:

THEOREM 3.29. *The equation of a gravitational free fall, in* 1 *dimension, is*

$$\ddot{x} = -\frac{GM}{x^2}$$

with M *being the attracting mass, and* $G \simeq 6.674 \times 10^{-11}$ *being a constant.*

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:

$$F = -G \cdot \frac{Mm}{x^2} \quad , \quad F = ma \quad , \quad a = \dot{v} \quad , \quad v = \dot{x}$$

To be more precise, in the first equation $G \simeq 6.674 \times 10^{-11}$ is the gravitational constant, in usual SI units, and the sign is $-$ because F is attractive. The second equation is something standard and very intuitive, and the last two equations are those from Fact 3.28. 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 3.30. *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. In order to understand the propagation of the waves, let us model the space, which is $\mathbb{R}$ for us, as a network of balls, with springs between them, as follows:

$$\cdots \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \cdots$$

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:

$$\cdots\cdots \bullet_{\varphi(x-l)} \times\!\!\times\!\!\times \bullet_{\varphi(x)} \times\!\!\times\!\!\times \bullet_{\varphi(x+l)} \cdots\cdots$$

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:

$$m \cdot \ddot{\varphi}(x) = k(\varphi(x+l) - 2\varphi(x) + \varphi(x-l))$$

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 the heat equation in 1D, as follows:

THEOREM 3.31. *The heat 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 3.18, 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 $\delta > 0$:

$$\varphi(x, t+\delta) = \varphi(x,t) + \frac{\alpha\delta}{l^2} \sum_{x \sim y} [\varphi(y,t) - \varphi(x,t)]$$

But this leads, via manipulations as before, to $\dot{\varphi}(x,t) = \alpha \cdot \varphi''(x,t)$, as claimed. □

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.

3e. Exercises

Here are some basic exercises, in relation with the above:

EXERCISE 3.32. *Work out, geometrically, estimates for* $\sin, \cos, \tan$ *at* $x = 0$.

EXERCISE 3.33. *Work out all the details in the proof of the Jensen inequality.*

EXERCISE 3.34. *Clarify the use of l'Hôpital's rule, in the proof of the Taylor formula.*

EXERCISE 3.35. *Complete the proof of the generalized binomial formula, at* $|t| < 1$.

As a bonus exercise, try solving the gravity, wave and heat equations in 1D.

CHAPTER 4

Integration

4a. 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 4.1. *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 4.2. *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 4.3. *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 4.1, 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 4.4. *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 4.2, 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 more generally 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 4.2. More on this later.

Back to work now, here are some general results regarding the integrals:

PROPOSITION 4.5. *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 4.6. *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:

$$\int_a^b f(x)dx = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \frac{b-a}{N} \cdot f\left(a + \frac{b-a}{N} \cdot k\right)$$

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 4.1, because area means side times average height. Thus, we can formulate:

THEOREM 4.7. *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 4.1, via some geometric thinking. Alternatively, this is something which certainly comes from Theorem 4.6. □

The point of view in Theorem 4.7 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 4.3, 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 4.7. The same goes for the results in Proposition 4.5, which become even more obvious with the viewpoint from Theorem 4.7. However, not everything trivializes in this way, and the result which is left, namely the formula $\int_{-1}^1\sqrt{1-x^2}\,dx=\pi/2$ from Proposition 4.3 (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 4.8. *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 4.7 that the idea is to use 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 the 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 precisely 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 4.9. *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 much 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 4.9:

CONCLUSION 4.10. *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, and we will be back to this later, when knowing more about random numbers, whose production and manipulation is in fact something quite tricky. To be more precise, the problem usually comes from the fact that the various algorithms producing random numbers have, as any piece of mathematics has, some "symmetries", that you don't want to see on your output numbers. But more on this later.

Hang on, we are not done yet. Here is one more interpretation of the integral:

THEOREM 4.11. *The integral of a function $f : [a, b] \to \mathbb{R}$ is given by*

$$\int_a^b f(x)dx = (b-a) \times \mathbb{E}(f)$$

where $\mathbb{E}(f)$ is the expectation of f, regarded as random variable.

PROOF. This is just some sort of fancy reformulation of Theorem 4.8, the idea being that what we can "expect" from a random variable is of course its average. □

Summarizing, we have so far a deep knowledge of what the integral is, philosophically speaking, but unfortunately, not many concrete results about it. Fixing this, with several explicit computations of integrals, will be our main purpose, in what follows.

4b. 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 4.6 will be our favorite tool. So, let us recall this formula, namely:

$$\int_a^b f(x)dx = (b-a) \times \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f\left(a + \frac{b-a}{N} \cdot k\right)$$

Let us begin with some theory. We first have the following result:

THEOREM 4.12. *The following functions are integrable:*

(1) *The piecewise continuous functions.*
(2) *The piecewise monotone functions.*

PROOF. This is indeed something quite standard, coming from the definition of the integral as a limit of Riemann sums, via some routine analysis:

(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 arbitrary divisions of the interval $[a, b]$, not necessarily uniform, denoted 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:

$$a_{i+1} - a_i < \frac{\delta}{2}\ ,\ a'_{i+1} - a'_i < \delta_2 \implies \left|\Sigma_I(f) - \Sigma_{I'}(f)\right| < \varepsilon$$

(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:

$$\begin{aligned}\left|\Sigma_I(f)-\Sigma_{I'}(f)\right| &= \left|\sum_{ij} l_{ij}f(a_i)-\sum_{ij} l_{ij}f(a'_j)\right| \\ &= \left|\sum_{ij} l_{ij}(f(a_i)-f(a'_j))\right|\end{aligned}$$

(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. □

Going ahead with more theory, let us establish now some abstract properties of the integration operation. We already know from Proposition 4.5 that the integrals behave

well with respect to sums and multiplication by scalars, the formulae being as follows:

$$\int_a^b f(x)+g(x)dx=\int_a^b f(x)dx+\int_a^b g(x)dx$$

$$\int_a^b \lambda f(x)=\lambda\int_a^b f(x)$$

Along the same lines, but at a more advanced level, we have the following result, which is equally useful, in practice, for the concrete computation of integrals:

THEOREM 4.13. *The integrals behave well with respect to taking limits,*

$$\int_a^b\left(\lim_{n\to\infty}f_n(x)\right)dx=\lim_{n\to\infty}\int_a^b f_n(x)dx$$

and with respect to taking infinite sums as well,

$$\int_a^b\left(\sum_{n=0}^{\infty}f_n(x)\right)dx=\sum_{n=0}^{\infty}\int_a^b f_n(x)dx$$

with both these formulae being valid, undwer mild assumptions.

PROOF. This is something quite standard, by using the general theory developed in chapter 3 for the sequences and series of functions. To be more precise, (1) follows by using the material there, via Riemann sums, and then (2) follows as a particular case of (1). We will leave the clarification of all this as an instructive exercise. □

Finally, still at the general level, let us record as well the following result:

THEOREM 4.14. *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. □

We will be back to more theory later. At the level of examples now, which is what matters the most, let us first look at the simplest functions that we know, namely the power functions $f(x) = x^p$. However, things here are tricky, and we only have:

THEOREM 4.15. *We have the integration formula*

$$\int_a^b x^p dx = \frac{b^{p+1} - a^{p+1}}{p+1}$$

valid at $p = 0, 1, 2, 3$.

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:

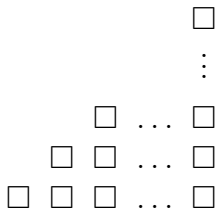
$$\frac{1 + 2 + \ldots + N}{N} = \frac{N+1}{2}$$

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 at all, so as a preliminary here, 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 $p = 2$. The trick is to use 2D geometry. Indeed, consider the following picture, with stacks going from 1 to N:



Now if we take two copies of this, and put them one on the top of the other, with a twist, in the obvious way, we obtain a rectangle having size $N \times (N + 1)$. Thus:

$$2(1 + 2 + \ldots + N) = N(N + 1)$$

But this gives the same formula as before, solving our question, namely:

$$1 + 2 + \ldots + N = \frac{N(N + 1)}{2} \simeq \frac{N^2}{2}$$

(6) Armed with this new method, let us attack now the case $p = 2$. Here we obviously need to do some 3D geometry, namely taking the picture P formed by a succession of solid squares, having sizes 1×1, 2×2, 3×3, and so on up to $N \times N$. Some quick thinking suggests that stacking 3 copies of P, with some obvious twists, will lead us to a parallelepiped. But this is not exactly true, and some further thinking shows that what we have to do is to add 3 more copies of P, leading to the following formula:

$$1^2 + 2^2 + \ldots + N^2 = \frac{N(N + 1)(2N + 1)}{6}$$

Or at least, that's how the legend goes. In practice, the above formula holds indeed, and you can check it for instance by recurrence, and this solves our problem:

$$1^2 + 2^2 + \ldots + N^2 \simeq \frac{2N^3}{6} = \frac{N^3}{3}$$

(7) At $p = 3$ now, the legend goes that by deeply thinking in 4D we are led to the following formula, a bit as in the cases $p = 1, 2$, explained above:

$$1^3 + 2^3 + \ldots + N^3 = \left(\frac{N(N + 1)}{2}\right)^2$$

Alternatively, assuming that the gods of combinatorics are with us, we can see right away the following formula, which coupled with (4) 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 check it for instance by recurrence, and this solves our problem:

$$1^3+2^3+\ldots+N^3\simeq\frac{N^4}{4}$$

(8) Thus, we proved our theorem. Of course, I can hear you screaming what about $p=4$ and higher. But the thing is that, by a strange twist of fate, there is no exact formula for $1^p+2^p+\ldots+N^p$, at $p=4$ and higher. Thus, game over. □

What happened above, with us unable to integrate x^p at $p=4$ and higher, not to mention the exponents $p\in\mathbb{R}-\mathbb{N}$ that we have not even dared to talk about, is quite annoying. As a conclusion to all this, however, let us formulate:

Conjecture 4.16. *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 $p\in\mathbb{N}$, and perhaps for some other $p\in\mathbb{R}$.

We will see later that this conjecture is indeed true, and with the exact details regarding the exponents $p\in\mathbb{R}-\mathbb{N}$ too. Now, instead of struggling with this, let us look at some other functions, which are not polynomial. And here, as good news, we have:

Theorem 4.17. *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. □

4c. Basic results

The problem is now, what to do with what we have, namely Conjecture 4.16 and Theorem 4.17. Not obvious, so stuck, and time to ask the cat. And cat says:

CAT 4.18. *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 cat is quite helpful. Guess he must be either a reincarnation of Newton or Leibnitz, these gentlemen used to talk like that, or that I should take care at some point of my garden, remove catnip and other weeds.

This being said, wait. There is suggestion to connect integrals and derivatives, and this is in fact what we have, coming from Conjecture 4.16 and Theorem 4.17, 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 cat. Moving ahead now, following this idea, we first have the following result, called fundamental theorem of calculus:

THEOREM 4.19. *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 4.14. 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 4.14, 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, which is something equivalent, and a hair more beautiful, also called fundamental theorem of calculus:

THEOREM 4.20. *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 4.19, and is in fact equivalent to it. Indeed, consider the following function:

$$G(s) = \int_a^s F'(x)dx$$

By using Theorem 4.19 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. □

There are many other equivalent formulations of the fundamental theorem of calculus, and countless applications as well. As an illustration for all this, we have:

THEOREM 4.21. *We have the following integration formulae,*

$$\int_a^b x^p dx = \frac{b^{p+1} - a^{p+1}}{p+1}$$

$$\int_a^b \frac{1}{x}\, dx = \log\left(\frac{b}{a}\right)$$

$$\int_a^b \sin x\, dx = \cos a - \cos b$$

$$\int_a^b \cos x\, dx = \sin b - \sin a$$

$$\int_a^b e^x dx = e^b - e^a$$

$$\int_a^b \log x\, dx = b\log b - a\log a - b + a$$

all obtained, in case you ever forget them, via the fundamental theorem of calculus.

PROOF. We already know two of these formulae, namely the one for powers from Theorem 4.15, and the one for exponentials Theorem 4.17, 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:

$$\begin{aligned}\int_a^b \log x\,dx &= (b\log b-b)-(a\log a-a)\\ &= b\log b-a\log a-b+a\end{aligned}$$

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

Getting back now to theory, inspired by the above, let us formulate:

DEFINITION 4.22. *Given a function 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 3, 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 conputing the above difference the c quantities will cancel, and we will obtain the same result.

As an application, we can reformulate Theorem 4.21 in a more digest form, as follows:

THEOREM 4.23. *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 4.21, or rather from the proof of Theorem 4.21, or even from chapter 3, for most of them, and the last assertion comes from the integration formula given after Definition 4.22. □

Getting back now to theory, we have the following key result:

THEOREM 4.24. *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. □

It is then possible to pass to usual integrals, and we obtain a formula here as well, as follows, also called integration by parts, 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:

$$\int_a^b f'g = -\int_a^b fg'$$

Examples of this usually come with $[a, b] = [-\infty, \infty]$, and more on this soon. Now still at the theoretical level, we have as well the following result:

THEOREM 4.25. *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 3, 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. □

All the above formulae are very useful, and altogether, and with some skill that can be acquired over the time, by computing a few dozen integrals, they can be basically used for computing any integral, provided that the integral is indeed computable. Of course, there are a few exceptions to this, such as the Gauss integral, which is as follows:

$$\int_{\mathbb{R}} e^{-x^2}dx = \sqrt{\pi}$$

So, this would be my advice, compute as many integrals as you can, as to get familiar with the above rules, and in your spare time, have a look at the Gauss integral too. And in case you don't find for Gauss, don't worry, we will get back to it, later in this book.

As a main application now of our theory, in relation with advanced calculus, and more specifically with the Taylor formula from chapter 3, we have:

THEOREM 4.26. *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 we will 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 $n = 1$. Thus, we are led to the formula in the statement. □

So long for basic integration theory. As a first concrete application now, we can compute all sorts of areas and volumes. Normally such computations are the business of multivariable calculus, and we will be back to this later, but with the technology that we have so far, we can do a number of things. As a first such computation, we have:

PROPOSITION 4.27. *The area of an ellipsis, given by the equation*

$$ax^2 + by^2 = 1$$

is $A = \pi/\sqrt{ab}$.

PROOF. The idea is that of cutting the ellipsis into vertical slices. First observe that, according to our equation $ax^2 + by^2 = 1$, the x coordinate can range as follows:

$$x \in \left[-\frac{1}{\sqrt{a}}, \frac{1}{\sqrt{a}}\right]$$

Now for any such x, the other coordinate y, satisfying $ax^2 + by^2 = 1$, is given by:

$$y = \pm\sqrt{\frac{1 - ax^2}{b}}$$

Thus the length of the vertical ellipsis slice at x is given by the following formula:

$$l(x) = 2\sqrt{\frac{1 - ax^2}{b}}$$

We conclude from this discussion that the area of the ellipsis is given by:

$$A = 2\int_{-1/\sqrt{a}}^{1/\sqrt{a}} \sqrt{\frac{1 - ax^2}{b}}\, dx$$

The problem is now, can we compute this integral, with the technology that we have. As a first step, by moving the $1/\sqrt{b}$ factor to the left, and then using the fact that we integrate an even function, we obtain the following formula:

$$A = \frac{4}{\sqrt{b}}\int_0^{1/\sqrt{a}} \sqrt{1 - ax^2}\, dx$$

But this suggests making the change of variables $x = y/\sqrt{a}$, which gives:

$$A = \frac{4}{\sqrt{ab}}\int_0^1 \sqrt{1 - y^2}\, dy$$

Regarding now the integral on the right, this can be computed for instance with $x = \cos t$, but in fact no need for this, because that integral computes by definition the area of a quarter of the unit circle, which is $\pi/4$. Thus, we obtain, as claimed:

$$A = \frac{4}{\sqrt{ab}} \cdot \frac{\pi}{4} = \frac{\pi}{\sqrt{ab}}$$

Finally, as a verification, for $a = b = 1$ we get $A = \pi$, as we should. □

Moving now to 3D, as an obvious challenge here, we can try to compute the volume of the sphere. This can be done a bit as for the ellipsis, the answer being as follows:

THEOREM 4.28. *The volume of the unit sphere is:*

$$V = \frac{4\pi}{3}$$

More generally, the volume of the sphere of radius R is $V = 4\pi R^3/3$.

PROOF. We proceed a bit as for the ellipsis. The equation of the sphere is:

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

Thus, the range of the first coordinate x is as follows:

$$x \in [-1, 1]$$

Now when this first coordinate x is fixed, the other coordinates y, z vary on a circle, given by the equation $y^2 + z^2 = 1 - x^2$, and so having radius as follows:

$$r(x) = \sqrt{1 - x^2}$$

Thus, the vertical slice of our sphere at x has area as follows:

$$a(x) = \pi r(x)^2 = \pi(1 - x^2)$$

We conclude from this discussion that the area of the sphere is given by:

$$\begin{aligned}
A &= \int_{-1}^{1} a(x)\,dx \\
&= \pi \int_{-1}^{1} 1 - x^2\,dx \\
&= \pi \int_{-1}^{1} \left(x - \frac{x^3}{3}\right)' dx \\
&= \pi \left[\left(1 - \frac{1}{3}\right) - \left(-1 + \frac{1}{3}\right)\right] \\
&= \pi \left(\frac{2}{3} + \frac{2}{3}\right) \\
&= \frac{4\pi}{3}
\end{aligned}$$

Finally, the last assertion is clear too, by multiplying everything by R, which amounts in multiplying the final result of our volume computation by R^3. □

4d. Some probability

As another application of the integration theory developed above, let us develop now some theoretical probability theory. You probably know, from real life, what probability is. But in practice, when trying to axiomatize this, in mathematical terms, things can be quite tricky. So, here comes our point, the definition saving us is as follows:

DEFINITION 4.29. *A probability density is a function $f : \mathbb{R} \to \mathbb{R}$ satisfying:*

(1) $f \geq 0$.
(2) $\int f(x)dx = 1$.

Observe the obvious relation with intuitive probability theory, where the probability for something to happen is always positive, $P \geq 0$, and where the overall probability for something to happen, with this meaning for one of the possible events to happen, is of course $\Sigma P = 1$, and this because life goes on, and something must happen, right.

In short, what we are proposing with Definition 4.29 is some sort of continuous analogue of basic probability theory, coming from coins, dice and cards, that you surely know. Moving now ahead, let us formulate, as a continuation of Definition 4.29:

DEFINITION 4.30. *Given a probability density function $f : \mathbb{R} \to \mathbb{R}$, we set*

$$P(X \in [a, b]) = \int_a^b f(x)dx$$

and call this probability for our variable to belong to $[a, b]$.

With this, we are now one step closer to what we know from coins, dice, cards and so on. Indeed, we have now a random variable X, that we can try to study. The first questions regard the mean and variance, which are constructed as follows:

DEFINITION 4.31. *Given a variable X, its mean is the following quantity:*

$$M = \int_{\mathbb{R}} xf(x)\,dx$$

More generally, we can say that the k-th moment of X is the following quantity:

$$M_k = \int_{\mathbb{R}} x^k f(x)\,dx$$

With this in hand, the variance of X is the quantity $V = M_2 - M_1^2$.

And with this, we have now a full theory, which is perfectly in tune with what we know from coins, dice, cards and so on. But you might perhaps say, wait, all this looks like some kind of reformulation of integration theory, and weren't we supposed to talk about probability theory instead? Good point, and the answer to this comes from:

FACT 4.32. *Probability theory is the same as integration theory.*

To be more precise, with just a bit more effort, basically amounting in including the Dirac delta functions δ_x into the class of density functions $f : \mathbb{R} \to \mathbb{R}$ that we consider, as to cover discrete probability too, we can turn Definition 4.29, along with the subsequent Definition 4.30 and Definition 4.31, into something foundational for probability at large, covering all types of probability that we know. We will be back to this.

In practice now, let us look for some illustrations for the above. Unfortunately the simplest laws in probability are the normal, or Gaussian laws, whose introduction requires the computation of the Gauss integral, which is a open problem that we have, namely:

$$\int_{\mathbb{R}} e^{-x^2} dx = ?$$

On top of this, as further bad news, even when avoiding the normal laws, and looking at some other laws instead, we need technology such as Fourier transform, convolution and so on, which are the business of complex analysis and multivariable calculus, that we have not learned yet. So, we are stuck, and time to ask the cat. And cat says:

CAT 4.33. *You must be kidding, probability theory comes after full calculus. Or at least, that's how we teach it to our kittens.*

Thanks cat, so we'll keep this for later. This being said, for not leaving things like this, let us talk however about something that we can talk about, namely the Poisson laws. These laws, which are quite important, can be introduced as follows:

DEFINITION 4.34. *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 more generally, 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.

Observe that our laws have indeed mass 1, as they should, and this due to:

$$e^t = \sum_{k \in \mathbb{N}} \frac{t^k}{k!}$$

In general, the idea with the Poisson laws is that these appear a bit everywhere, in discrete contexts, the reasons for this coming from the Poisson Limit Theorem (PLT). However, the proof of this theorem, and even its statement, uses advanced calculus, including convolution and the Fourier transform, and we will leave this for later.

We can, however, have some fun with moments, the result here being as follows:

THEOREM 4.35. *The moments of p_1 are the Bell numbers,*

$$M_k(p_1) = |P(k)|$$

where $P(k)$ is the set of partitions of $\{1, \ldots, k\}$. More generally, we have

$$M_k(p_t) = \sum_{\pi \in P(k)} t^{|\pi|}$$

for any $t > 0$, where $|.|$ is the number of blocks.

PROOF. The moments of p_1 satisfy the following recurrence formula:

$$\begin{aligned} M_{k+1} &= \frac{1}{e} \sum_r \frac{(r+1)^{k+1}}{(r+1)!} \\ &= \frac{1}{e} \sum_r \frac{r^k}{r!} \left(1 + \frac{1}{r}\right)^k \\ &= \frac{1}{e} \sum_r \frac{r^k}{r!} \sum_s \binom{k}{s} r^{-s} \\ &= \sum_s \binom{k}{s} \cdot \frac{1}{e} \sum_r \frac{r^{k-s}}{r!} \\ &= \sum_s \binom{k}{s} M_{k-s} \end{aligned}$$

With this done, let us try now to find a recurrence for the Bell numbers, $B_k = |P(k)|$. A partition of $\{1, \ldots, k+1\}$ appears by choosing s neighbors for 1, among the k numbers available, and then partitioning the $k - s$ elements left. Thus, we have:

$$B_{k+1} = \sum_s \binom{k}{s} B_{k-s}$$

Thus, our moments M_k satisfy the same recurrence as the numbers B_k. Regarding now the initial values, in what concerns the first moment of p_1, we have:

$$M_1 = \frac{1}{e} \sum_r \frac{r}{r!} = 1$$

Also, by using the above recurrence for the numbers M_k, we obtain from this:

$$M_2 = \sum_s \binom{1}{s} M_{k-s} = 1 + 1 = 2$$

On the other hand, we have $B_1 = 1, B_2 = 2$, so we get by recurrence, as claimed:

$$M_k = B_k$$

Regarding now the law p_t with $t > 0$, we have here a similar formula, namely:

$$\begin{aligned}
M_{k+1} &= e^{-t}\sum_r \frac{t^{r+1}(r+1)^{k+1}}{(r+1)!}\\
&= e^{-t}\sum_r \frac{t^{r+1}r^k}{r!}\left(1+\frac{1}{r}\right)^k\\
&= e^{-t}\sum_r \frac{t^{r+1}r^k}{r!}\sum_s \binom{k}{s} r^{-s}\\
&= \sum_s \binom{k}{s}\cdot e^{-t}\sum_r \frac{t^{r+1}r^{k-s}}{r!}\\
&= t\sum_s \binom{k}{s} M_{k-s}
\end{aligned}$$

Regarding now the initial values, the first moment of p_t is given by:

$$M_1 = e^{-t}\sum_r \frac{t^r r}{r!} = e^{-t}\sum_r \frac{t^r}{(r-1)!} = t$$

Now by using the above recurrence we obtain from this:

$$M_2 = t\sum_s \binom{1}{s} M_{k-s} = t(1+t) = t+t^2$$

On the other hand, some standard combinatorics, a bit as before at $t = 1$, shows that the numbers in the statement $S_k = \sum_{\pi\in P(k)} t^{|\pi|}$ satisfy the same recurrence relation, and with the same initial values. Thus we have $M_k = S_k$, as claimed. □

The above result looks quite exciting, making a link between subtle probability, and subtle combinatorics. We will be back to it, after learning more calculus.

4e. Exercises

We had a lot of interesting theory in this chapter, and as exercises, we have:

EXERCISE 4.36. *Clarify the integrability property of piecewise monotone functions.*

EXERCISE 4.37. *Find a geometric proof for* $1^3 + \ldots + N^3 = (1 + \ldots + N)^2$.

EXERCISE 4.38. *Find estimates for the remainder, in the Taylor formula.*

EXERCISE 4.39. *Work out with full details the formula of* $M_k(p_t)$, *at* $t > 0$.

As a bonus exercise, compute areas and volumes. As many as you can.

Part II

Complex functions

How can I change the world
If I can't even change myself
How can I change the way I am
I don't know, I don't know

CHAPTER 5

Complex numbers

5a. Complex numbers

In this second part of the present book we discuss the functions of one complex variable $f : \mathbb{C} \to \mathbb{C}$. This is certainly something to be done before going to several variables, be them real or complex, because we have some unfinished business with $\sin, \cos, \exp, \log$, which remain quite mysterious objects. We will see here that these functions are much better understood as complex functions $f : \mathbb{C} \to \mathbb{C}$. In fact, even a dumb polynomial $P \in \mathbb{R}[X]$ is better understood as complex polynomial, $P \in \mathbb{C}[X]$, because its roots, that might not always exist as real numbers, always exist as complex numbers.

On top of this, we have physics. You might know this or not, but physics is all about waves, which waves require complex numbers for their understanding. And there is even better, the $\mathbb{R}$ that we are so used to, from classical mechanics, dealing with questions at our usual, human scale, naturally becomes, via some sort of complicated procedure, $\mathbb{C}$ in quantum mechanics, that is, when zooming down, where the new science and technologies are. But more on this later, towards the end of the present book.

In short, many interesting things to be discussed, not necessarily for the sake of doing complex numbers, but rather with the goal of better understanding the real numbers themselves. Let us begin with the complex numbers. There is a lot of magic here, and we will carefully explain this material. Their definition is as follows:

DEFINITION 5.1. *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 is regarded as 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:

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

Thus, the complex numbers as introduced above are well-defined. The multiplication formula is of course quite tricky, and hard to memorize, but we will see later some alternative ways, which are more conceptual, for performing the multiplication.

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, regarding the arbitrary degree 2 equations, with real coefficients:

THEOREM 5.2. *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 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. □

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 5.3. *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:

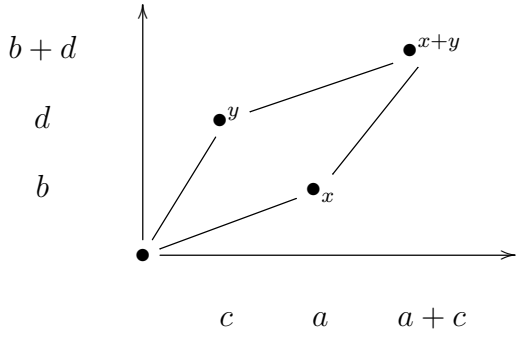
$$x = \begin{pmatrix} a \\ b \end{pmatrix} \quad , \quad y = \begin{pmatrix} c \\ d \end{pmatrix}$$

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. □

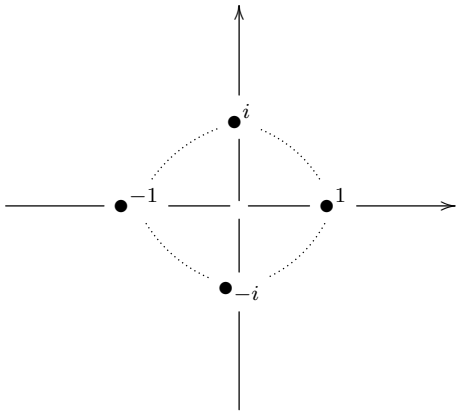
Here we have assumed that you are a bit familiar with vector calculus. If not, no problem, the idea is simply that vectors add by forming a parallelogram, as follows:



Observe that in our geometric picture from Proposition 5.3, the real numbers correspond to the numbers on the Ox axis. As for the purely imaginary numbers, these lie on the Oy axis, with the number i itself being given by the following formula:

$$i = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

As an illustration for this, let us record now a basic picture, with some key complex numbers, namely $1, i, -1, -i$, represented according to our conventions:



You might perhaps wonder why I chose to draw that circle, connecting the numbers $1, i, -1, -i$, which does not look very useful. More on this in a moment, the idea being that that circle can be immensely useful, and coming in advance, some advice:

ADVICE 5.4. *When drawing complex numbers, always begin with the coordinate axes Ox, Oy, and with a copy of the unit circle.*

We have so far a quite good understanding of their complex numbers, and 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 5.5. *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,

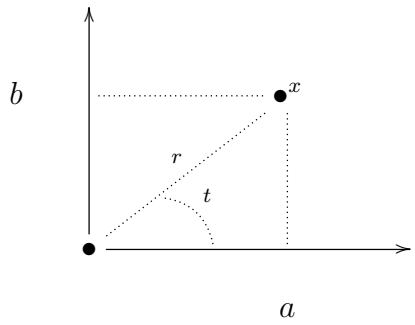
$$a = r \cos t \quad , \quad b = r \sin t$$

and in the other sense being as follows,

$$r = \sqrt{a^2 + b^2} \quad , \quad \tan t = b/a$$

and with r, t being called modulus, and argument.

There is a clear relation here with the vector notation from Proposition 5.3, 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 5.5 is as follows:



As a basic example here, the number i takes the following form:

$$i = \cos\left(\frac{\pi}{2}\right) + i\sin\left(\frac{\pi}{2}\right)$$

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 5.6. *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. This follows from the following formulae, that we know well:

$$\cos(s+t) = \cos s\cos t - \sin s\sin t$$

$$\sin(s+t) = \cos s\sin t + \sin s\cos t$$

Indeed, we can assume 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 5.7. *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:

$$\begin{aligned} x^k &= x \ldots x \\ &= r(\cos t + i \sin t) \ldots r(\cos t + i \sin t) \\ &= r \ldots r([\cos(t + \ldots + t) + i \sin(t + \ldots + t)) \\ &= r^k(\cos kt + i \sin kt) \end{aligned}$$

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, because:

$$\begin{aligned} xx^{-1} &= r(\cos t + i \sin t) \cdot r^{-1}(\cos(-t) + i \sin(-t)) \\ &= \cos(t - t) + i \sin(t - t) \\ &= \cos 0 + i \sin 0 \\ &= 1 \end{aligned}$$

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 should mention that there is a bit of ambiguity in the above, in the case of the exponents $k \in \mathbb{Q}$, due to the fact that the square roots, and the higher roots as well, can take multiple values, in the complex number setting. We will be back to this.

As a basic application of Theorem 5.7, we have the following result:

PROPOSITION 5.8. *Each complex number, written in polar form,*

$$x = r(\cos t + i \sin t)$$

has two square roots, given by the following formula:

$$\sqrt{x} = \pm\sqrt{r}\left[\cos\left(\frac{t}{2}\right) + i\sin\left(\frac{t}{2}\right)\right]$$

When $x > 0$, these roots are $\pm\sqrt{x}$. When $x < 0$, these roots are $\pm i\sqrt{-x}$.

PROOF. The first assertion is clear indeed from the general formula in Theorem 5.7, at $k = 1/2$. As for its particular cases with $x \in \mathbb{R}$, these are clear from it. □

As a comment here, for $x > 0$ we are very used to call the usual $\sqrt{x}$ square root of x. However, for $x < 0$, or more generally for $x \in \mathbb{C} - \mathbb{R}_+$, there is less interest in choosing one of the possible $\sqrt{x}$ and calling it "the" square root of x, because all this is based on our convention that i comes up, instead of down, which is something rather arbitrary. Actually, clocks turning clockwise, i should be rather coming down. All this is a matter of taste, but in any case, for our math, the best is to keep some ambiguity, as above.

With the above results in hand, and notably with the square root formula from Proposition 5.8, we can now go back to the degree 2 equations, and we have:

THEOREM 5.9. *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 complex numbers being defined as above.

PROOF. This is clear, the computations being the same as in the real case. To be more precise, our degree 2 equation can be written as follows:

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2}$$

Now since we know from Proposition 5.8 that any complex number has a square root, we are led to the conclusion in the statement. □

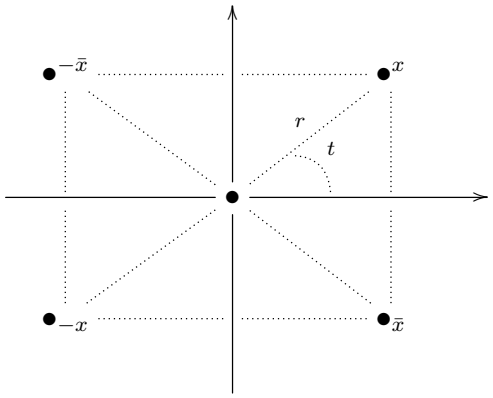
As a last general topic regarding the complex numbers, let us discuss conjugation. This is something quite tricky, complex number specific, as follows:

DEFINITION 5.10. *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 5.11. *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 $|x| = r$ being as usual the modulus.*
(4) *With $x = r(\cos t + i \sin t)$, we have $\bar{x} = r(\cos(-t) + r \sin(-t))$.*
(5) *We have the formula $\overline{xy} = \bar{x}\bar{y}$, for any $x, y \in \mathbb{C}$.*
(6) *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 clear indeed from the picture following Definition 5.10.

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

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

However, what we have been doing here is not very clear, geometrically speaking, and our formula is worth an alternative proof. Here is that proof, which after inspection contains no computations at all, making it clear that the polar writing is the best:

$$\begin{aligned} & \overline{r(\cos s + i\sin s)\cdot p(\cos t + i\sin t)} \\ = \ & \overline{rp(\cos(s+t) + i\sin(s+t))} \\ = \ & rp(\cos(-s-t) + i\sin(-s-t)) \\ = \ & r(\cos(-s) + i\sin(-s))\cdot p(\cos(-t) + i\sin(-t)) \\ = \ & \overline{r(\cos s + i\sin s)}\cdot\overline{p(\cos t + i\sin t)} \end{aligned}$$

(6) This comes from the formula of the solutions, that we know from Theorem 5.2, 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. □

5b. Exponential writing

We discuss now the theory of complex functions $f : \mathbb{C} \to \mathbb{C}$, in analogy with the theory of the real functions $f : \mathbb{R} \to \mathbb{R}$. We will see that many results that we know from the real setting extend to the complex setting. Before starting, two remarks on this:

(1) Most of the real functions $f : \mathbb{R} \to \mathbb{R}$ that we know, such as $\sin, \cos, \exp, \log$, extend into complex functions $f : \mathbb{C} \to \mathbb{C}$, and the study of these latter extensions brings some new light on the original real functions. Thus, what we will be doing here will be, in a certain sense, a refinement of the theory developed in chapters 1-4.

(2) On the other hand, since we have $\mathbb{C} \simeq \mathbb{R}^2$, the complex functions $f : \mathbb{C} \to \mathbb{C}$ that we will study here can be regarded as functions $f : \mathbb{R}^2 \to \mathbb{R}^2$. This is something quite subtle, but in any case, what we will be doing here will stand as well as an introduction to the functions of type $f : \mathbb{R}^N \to \mathbb{R}^M$, that we will study in chapters 9-16 below.

In short, one complex variable is something in between one real variable, and two or more real variables, and we can only expect to end up with a mysterious mixture of surprising and unsurprising results. Welcome to complex analysis. Let us start with:

DEFINITION 5.12. *A complex function $f : \mathbb{C} \to \mathbb{C}$, or more generally $f : X \to \mathbb{C}$, with $X \subset \mathbb{C}$ being a subset, is called continuous when, for any $x_n, x \in X$:*

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

where the convergence of the sequences of complex numbers, $x_n \to x$, means by definition that for n big enough, the quantity $|x_n - x|$ becomes arbitrarily small.

Observe that in real coordinates, $x = (a, b)$, the distances appearing in the definition of the convergence $x_n \to x$ are given by the following formula:

$$|x_n - x| = \sqrt{(a_n - a)^2 + (b_n - b)^2}$$

Thus $x_n \to x$ in the complex sense means that $(a_n, b_n) \to (a, b)$ in the usual, intuitive sense, with respect to the usual distance in the plane $\mathbb{R}^2$, and as a consequence, a function $f : \mathbb{C} \to \mathbb{C}$ is continuous precisely when it is continuous, in an intuitive sense, when regarded as function $f : \mathbb{R}^2 \to \mathbb{R}^2$. But more on this later, in chapters 9-10 below.

At the level of examples, we have the following result:

THEOREM 5.13. *We can exponentiate the complex numbers, according to the formula*

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

and the function $x \to e^x$ is continuous, and satisfies $e^{x+y} = e^x e^y$.

PROOF. We must first prove that the series converges. But this follows from:

$$\begin{aligned} |e^x| &= \left| \sum_{k=0}^{\infty} \frac{x^k}{k!} \right| \\ &\leq \sum_{k=0}^{\infty} \left| \frac{x^k}{k!} \right| \\ &= \sum_{k=0}^{\infty} \frac{|x|^k}{k!} \\ &= e^{|x|} < \infty \end{aligned}$$

Regarding the formula $e^{x+y}=e^xe^y$, this follows too as in the real case, as follows:

$$\begin{aligned}
e^{x+y} &= \sum_{k=0}^{\infty}\frac{(x+y)^k}{k!}\\
&= \sum_{k=0}^{\infty}\sum_{s=0}^{k}\binom{k}{s}\cdot\frac{x^sy^{k-s}}{k!}\\
&= \sum_{k=0}^{\infty}\sum_{s=0}^{k}\frac{x^sy^{k-s}}{s!(k-s)!}\\
&= e^xe^y
\end{aligned}$$

Finally, the continuity of $x\to e^x$ comes at $x=0$ from the following computation:

$$\begin{aligned}
|e^t-1| &= \left|\sum_{k=1}^{\infty}\frac{t^k}{k!}\right|\\
&\leq \sum_{k=1}^{\infty}\left|\frac{t^k}{k!}\right|\\
&= \sum_{k=1}^{\infty}\frac{|t|^k}{k!}\\
&= e^{|t|}-1
\end{aligned}$$

As for the continuity of $x\to e^x$ in general, this can be deduced now as follows:

$$\lim_{t\to0}e^{x+t}=\lim_{t\to0}e^xe^t=e^x\lim_{t\to0}e^t=e^x\cdot1=e^x$$

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

We will be back to more functions later. As an important fact, however, let us point out that, contrary to what the above might suggest, everything does not always extend trivially from the real to the complex case. For instance, we have:

PROPOSITION 5.14. *We have the following formula, valid for any $|x|<1$,*

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

but, unlike in the real case, 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:

$$\left|\sum_{n=0}^{\infty}x^n\right|\leq\sum_{n=0}^{\infty}|x|^n=\frac{1}{1-|x|}$$

As for the last assertion, this is something quite informal. To be more precise, for $x = 1/2$ our formula is clear, by cutting the interval $[0, 2]$ into half, and so on:

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = 2$$

More generally, for $x \in (-1, 1)$ the meaning of the formula in the statement is something quite clear and intuitive, geometrically speaking, by using a similar argument. 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. To be more precise here, by anticipating a bit, assume that we have a number $w \in \mathbb{C}$ satisfying $w^N = 1$, for some $N \in \mathbb{N}$. We have then the following formula, for our infinite sum:

$$\begin{aligned} 1 + rw + r^2w^2 + \ldots &= (1 + rw + \ldots + r^{N-1}w^{N-1}) \\ &+ (r^N + r^{N+1}w \ldots + r^{2N-1}w^{N-1}) \\ &+ (r^{2N} + r^{2N+1}w \ldots + r^{3N-1}w^{N-1}) \\ &+ \ldots \end{aligned}$$

Thus, by grouping the terms with the same argument, our infinite sum is:

$$\begin{aligned} 1 + rw + r^2w^2 + \ldots &= (1 + r^N + r^{2N} + \ldots) \\ &+ (r + r^{N+1} + r^{2N+1} + \ldots)w \\ &+ \ldots \\ &+ (r^{N-1} + r^{2N-1} + r^{3N-1} + \ldots)w^{N-1} \end{aligned}$$

But the sums of each ray can be computed with the real formula for geometric series, that we know and understand well, and together with an extra bit of algebra, we get:

$$\begin{aligned} 1 + rw + r^2w^2 + \ldots &= \frac{1}{1 - r^N} + \frac{rw}{1 - r^N} + \ldots + \frac{r^{N-1}w^{N-1}}{1 - r^N} \\ &= \frac{1}{1 - r^N}\left(1 + rw + \ldots + r^{N-1}w^{N-1}\right) \\ &= \frac{1}{1 - r^N} \cdot \frac{1 - r^N}{1 - rw} \\ &= \frac{1}{1 - rw} \end{aligned}$$

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 back now to less mysterious mathematics, which in fact will turn to be quite mysterious as well, as is often the case with things involving complex numbers, as an

application of all this, let us discuss the final and most convenient writing of the complex numbers, which is a variation on the polar writing, as follows:

$$x = re^{it}$$

The point with this formula comes from the following deep result:

THEOREM 5.15. *We have the following formula,*

$$e^{it} = \cos t + i \sin t$$

valid for any $t \in \mathbb{R}$.

PROOF. Our claim is that this follows from the formula of the complex exponential, and for the following formulae for the Taylor series of cos and sin, that we know well:

$$\cos t = \sum_{l=0}^{\infty} (-1)^l \frac{t^{2l}}{(2l)!} \quad , \quad \sin t = \sum_{l=0}^{\infty} (-1)^l \frac{t^{2l+1}}{(2l+1)!}$$

Indeed, let us first recall from Theorem 5.13 that we have the following formula, for the exponential of an arbitrary complex number $x \in \mathbb{C}$:

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Now let us plug $x = it$ in this formula. We obtain the following formula:

$$\begin{aligned} e^{it} &= \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \\ &= \sum_{k=2l} \frac{(it)^k}{k!} + \sum_{k=2l+1} \frac{(it)^k}{k!} \\ &= \sum_{l=0}^{\infty} (-1)^l \frac{t^{2l}}{(2l)!} + i \sum_{l=0}^{\infty} (-1)^l \frac{t^{2l+1}}{(2l+1)!} \\ &= \cos t + i \sin t \end{aligned}$$

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

As a main application of the above formula, we have:

THEOREM 5.16. *We have the following formula,*

$$e^{\pi i} = -1$$

and we have $E = mc^2$ *as well.*

PROOF. We have two assertions here, the idea being as follows:

(1) The first formula, $e^{\pi i}=-1$, which is actually the main formula in mathematics, comes from Theorem 5.15, by setting $t=\pi$. Indeed, we obtain:

$$\begin{aligned}e^{\pi i}&=\cos\pi+i\sin\pi\\&=-1+i\cdot 0\\&=-1\end{aligned}$$

(2) As for $E=mc^2$, which is the main formula in physics, this is something deep as well. Although we will not really need it here, we recommend learning it too, for symmetry reasons between math and physics, say from Feynman [31], [32], [33]. □

Now back to our $x=re^{it}$ objectives, with the above theory in hand we can indeed use from now on this notation, the complete statement being as follows:

THEOREM 5.17. *The complex numbers $x=a+ib$ can be written in polar coordinates,*

$$x=re^{it}$$

with the connecting formulae being

$$a=r\cos t\quad,\quad b=r\sin t$$

and in the other sense being

$$r=\sqrt{a^2+b^2}\quad,\quad\tan t=b/a$$

and with r,t being called modulus, and argument.

PROOF. This is a reformulation of our previous Definition 5.5, by using the formula $e^{it}=\cos t+i\sin t$ from Theorem 5.15, and multiplying everything by r. □

With this in hand, we can now go back to the basics, namely the addition and multiplication of the complex numbers. We have the following result:

THEOREM 5.18. *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 Theorem 5.6, reformulated by using the notations from Theorem 5.17. Observe that this follows as well directly, from the fact that we have $e^{a+b}=e^ae^b$, that we know from analysis. □

The above formula is obviously very powerful. However, in polar coordinates we do not have a simple formula for the sum. Thus, this formalism has its limitations.

We can investigate as well more complicated operations, as follows:

THEOREM 5.19. *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 Theorem 5.7, but we can now discuss all this, from a more conceptual viewpoint, the idea being as follows:

(1) We have indeed the following computation, using Theorem 5.18:

$$\begin{aligned}(re^{it})(r^{-1}e^{-it}) &= rr^{-1}\cdot e^{i(t-t)}\\ &= 1\cdot 1\\ &= 1\end{aligned}$$

(2) Once again by using Theorem 5.18, we have:

$$(\pm\sqrt{r}e^{it/2})^2 = (\sqrt{r})^2 e^{i(t/2+t/2)} = re^{it}$$

(3) Given an arbitrary number $a \in \mathbb{R}$, we can define, as stated:

$$(re^{it})^a = r^a e^{ita}$$

Due to Theorem 5.18, this operation $x \to x^a$ is indeed the correct one.

(4) This is just a reformulation of the fact, that we know from Theorem 5.11, that the conjugation operation $x \to \bar{x}$ keeps the modulus, and switches the argument. □

5c. Equations, roots

Getting back to algebra, recall from Theorem 5.9 that any degree 2 equation has 2 complex roots. We can in fact prove that any polynomial equation, of arbitrary degree $N \in \mathbb{N}$, has exactly N complex solutions, counted with multiplicities:

THEOREM 5.20. *Any polynomial $P \in \mathbb{C}[X]$ decomposes as*

$$P = c(X - a_1)\ldots(X - a_N)$$

with $c \in \mathbb{C}$ and with $a_1, \ldots, a_N \in \mathbb{C}$.

PROOF. The problem is that of proving that our polynomial has at least one root, because afterwards we can proceed by recurrence. We prove this by contradiction. So, assume that P has no roots, and pick a number $z \in \mathbb{C}$ where $|P|$ attains its minimum:

$$|P(z)| = \min_{x\in\mathbb{C}} |P(x)| > 0$$

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:

$$P(z+rw) \simeq P(z) + cr^k w^k$$

Now recall that we have assumed $P(z) \neq 0$. We can therefore choose $w \in \mathbb{T}$ such that cw^k points in the opposite direction to that of $P(z)$, and we obtain in this way:

$$\begin{aligned} |P(z+rw)| &\simeq |P(z) + cr^k w^k| \\ &= |P(z)|(1 - |c|r^k) \end{aligned}$$

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:

$$|P(z+rw)| < |P(z)|$$

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. □

All this is very nice, and we will see applications in a moment. As a word of warning, however, we should say that the above result remains something quite theoretical. Indeed, the proof is by contradiction, and there is no way of recycling the material there into something explicit, which can be used for effectively computing the roots.

Still talking polynomials and their roots, let us try however to understand what the analogue of $\Delta = b^2 - 4ac$ is, for an arbitrary polynomial $P \in \mathbb{C}[X]$. We will need:

THEOREM 5.21. *Given two polynomials $P, Q \in \mathbb{C}[X]$, written as follows,*

$$P = c(X - a_1) \ldots (X - a_k) \quad , \quad Q = d(X - b_1) \ldots (X - b_l)$$

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 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. Given $P, Q \in \mathbb{C}[X]$, we can certainly construct the quantity $R(P, Q)$ in the statement, and we have then $R(P, Q) = 0$ precisely when P, Q have a common root. The whole point is that of proving that $R(P, Q)$ is a polynomial in the coefficients of P, Q, with integer coefficients. But this can be checked as follows:

(1) We can expand the formula of $R(P,Q)$, and in what regards $a_1,\ldots,a_k$, which are the roots of P, we obtain in this way certain symmetric functions in these variables, which will be therefore polynomials in the coefficients of P, with integer coefficients.

(2) We can then look 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, and so polynomials in the coefficients of Q, with integer coefficients.

(3) 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^ld^k factor is there for these latter coefficients to be indeed integers, instead of rationals. □

All this might seem a bit complicated, so as an illustration, let us work out an example. Consider the case of a polynomial of degree 2, and a polynomial 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:

$$P(r)=\frac{ae^2}{d^2}-\frac{be}{d}+c=\frac{R(P,Q)}{d^2}$$

Thus $P(r)=0$ precisely when $R(P,Q)=0$, as predicted by Theorem 5.21.

With this, we can now talk about the discriminant of any polynomial, as follows:

THEOREM 5.22. *Given a polynomial $P\in\mathbb{C}[X]$, written as*

$$P(X)=cX^N+dX^{N-1}+\ldots$$

its discriminant, defined as being the following quantity,

$$\Delta(P)=\frac{(-1)^{\binom{N}{2}}}{c}R(P,P')$$

is a polynomial in the coefficients of P, with integer coefficients, and

$$\Delta(P)=0$$

happens precisely when P has a double root.

PROOF. This follows from Theorem 5.21, applied with $P = Q$, with the division by c being indeed possible, under $\mathbb{Z}$, and with the sign being there for various reasons, including the compatibility with some well-known formulae, at small values of $N \in \mathbb{N}$. □

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:

$$\begin{aligned} R(P, P') &= ab^2 - b(2a)b + c(2a)^2 \\ &= 4a^2c - ab^2 \\ &= -a(b^2 - 4ac) \end{aligned}$$

With the normalizations in Theorem 5.22 made, we obtain, as we should:

$$\Delta(P) = b^2 - 4ac$$

As another illustration, let us work out what happens in degree 3. Here the result, which is useful and interesting, and is probably new to you, is as follows:

THEOREM 5.23. *The discriminant of a degree* 3 *polynomial,*

$$P = aX^3 + bX^2 + cX + d$$

is the number $\Delta(P) = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd$.

PROOF. We need to do some tough computations here. Let us first compute resultants. Consider two polynomials, of degree 3 and degree 2, written as follows:

$$P = aX^3 + bX^2 + cX + d = a(X - p)(X - q)(X - r)$$

$$Q = eX^2 + fX + g = e(X - s)(X - t)$$

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, and we are done. □

Still talking degree 3 equations, let us try to solve $P=0$, 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 $x^3+3px+2q=0$, and regarding such equations, we have the following famous result, due to Cardano:

THEOREM 5.24. *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 $108=4\times27$. 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. □

There are many more things that can be said about degree 3 equations, along these lines, and we will certainly have an exercise about this, at the end of this chapter.

5d. Roots of unity

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. Let us start with:

THEOREM 5.25. *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 5.16. 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 alternative writing is $\{\pm 1, \pm w, \pm w^2\}$, 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. As a first application, we can now solve the ambiguity questions related to the extraction of N-th roots, from Theorem 5.7 and Theorem 5.19, the statement being as follows:

THEOREM 5.26. *Any nonzero complex number, written as*

$$x = re^{it}$$

has exactly N roots of order N, which appear as

$$y = r^{1/N} e^{it/N}$$

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:

$$z^N = y^N$$

Now observe that we can write this equation as follows:

$$\left(\frac{z}{y}\right)^N = 1$$

We conclude 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. □

The roots of unity appear in connection with many other interesting questions, and there are many useful formulae relating them, which are good to know. Here is a basic such formula, very beautiful, to be used many times in what follows:

THEOREM 5.27. *The roots of unity, $\{w^k\}$ with $w = e^{2\pi i/N}$, have the property*

$$\sum_{k=0}^{N-1} (w^k)^s = N\delta_{N|s}$$

for any exponent $s \in \mathbb{N}$, where on the right we have a Kronecker symbol.

PROOF. The numbers in the statement, when written more conveniently as $(w^s)^k$ with $k = 0, \ldots, N-1$, form a certain regular polygon in the plane P_s. Thus, if we denote by C_s the barycenter of this polygon, we have the following formula:

$$\frac{1}{N}\sum_{k=0}^{N-1} w^{ks} = C_s$$

Now observe that in the case $N\not|\, s$ our polygon P_s is non-degenerate, circling around the unit circle, and having center $C_s = 0$. As for the case $N|s$, here the polygon is degenerate, lying at 1, and having center $C_s = 1$. Thus, we have the following formula:

$$C_s = \delta_{N|s}$$

Thus, we obtain the formula in the statement. □

As an interesting philosophical fact, regarding the roots of unity, and the complex numbers in general, we can now solve the following equation, in a "uniform" way:

$$x_1 + \ldots + x_N = 0$$

With this being not a joke. Frankly, can you find some nice-looking family of real numbers $x_1, \ldots, x_N$ satisfying $x_1 + \ldots + x_N = 0$? Certainly not. But with complex numbers we have now our answer, the sum of the N-th roots of unity being zero.

This was for our basic presentation of the complex numbers. We will be back to more theory regarding them, and the roots of unity, later on. Among others, we will see later some non-trivial applications of our above solution to $x_1 + \ldots + x_N = 0$.

5e. Exercises

This was all beautiful material, and our exercises will be beautiful as well:

EXERCISE 5.28. *Solve easy plane geometry problems, using complex numbers.*

EXERCISE 5.29. *Solve more plane geometry problems, using complex numbers, this time of scary type, involving circles, square angles, and equilateral triangles.*

EXERCISE 5.30. *Further build on the Cardano formula presented above.*

EXERCISE 5.31. *Diagonalize, beautifully, the all-one matrix.*

As bonus exercise, learn some physics. This is where complex numbers really shine.

CHAPTER 6

Complex functions

6a. Functions, continuity

We discuss in this chapter and in the next two ones the theory of complex functions $f : \mathbb{C} \to \mathbb{C}$, in analogy with the theory of the real functions $f : \mathbb{R} \to \mathbb{R}$. We will see that many results that we know from the real setting extend to the complex setting, but there will be quite a number of new phenomena too. We will need, in order to get started:

DEFINITION 6.1. *We define the distance between two complex numbers to be the usual distance in the plane between them, namely:*

$$d(x, y) = |x - y|$$

With this, we can talk about convergence, by saying that $x_n \to x$ when $d(x_n, x) \to 0$.

Here the fact that $d(x, y) = |x-y|$ is indeed the usual distance in the plane is clear for $y = 0$, because we have $d(x, 0) = |x|$, by definition of the modulus $|x|$. As for the general case, $y \in \mathbb{C}$, this comes from the fact that the distance in the plane is given by:

$$d(x, y) = d(x - y, 0) = |x - y|$$

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 formula, and we can get away with all sorts of tricks regarding complex numbers that we know. As a first result now, regarding $\mathbb{C}$ and its distance, that we will need in what follows, we have:

PROPOSITION 6.2. *The complex plane $\mathbb{C}$ is complete, in the sense that any Cauchy sequence converges.*

PROOF. Consider indeed a Cauchy sequence $\{x_n\}_{n\in\mathbb{N}} \subset \mathbb{C}$. If we write $x_n = a_n + ib_n$ for any $n \in \mathbb{N}$, then we have the following estimates:

$$|a_n - a_m| \leq \sqrt{(a_n - a_m)^2 + (b_n - b_m)^2} = |x_n - x_m|$$

$$|b_n - b_m| \leq \sqrt{(a_n - a_m)^2 + (b_n - b_m)^2} = |x_n - x_m|$$

Thus both the sequences $\{a_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ and $\{b_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ are Cauchy, and since we know that $\mathbb{R}$ itself is complete, we can consider the limits of these sequences:

$$a_n \to a \quad , \quad b_n \to b$$

With $x = a + ib$, our claim is that $x_n \to x$. Indeed, we have:

$$\begin{aligned} |x_n - x| &= \sqrt{(a_n - a)^2 + (b_n - b)^2} \\ &\leq |a_n - a| + |b_n - b| \end{aligned}$$

It follows that we have $x_n \to x$, as claimed, and this gives the result. □

Talking complex functions now, we have the following definition:

DEFINITION 6.3. *A complex function $f : \mathbb{C} \to \mathbb{C}$, or more generally $f : X \to \mathbb{C}$, with $X \subset \mathbb{C}$ 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 can talk about pointwise convergence of functions, $f_n \to f$, and about uniform convergence too, $f_n \to_u f$, exactly as for the real functions.

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, $f : \mathbb{R}^2 \to \mathbb{R}^2$. But more on this later in this book. At the level of examples now, we first have the polynomials, $P \in \mathbb{C}[X]$. We have already met such polynomials in chapter 5, so let us recall from there that we have:

THEOREM 6.4. *Each polynomial $P \in \mathbb{C}[X]$ can be regarded as a continuous function $P : \mathbb{C} \to \mathbb{C}$. Moreover, we have the formula*

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

with $a \in \mathbb{C}$, and with the numbers $r_1, \ldots, r_n \in \mathbb{C}$ being the roots of P.

PROOF. This is something that we know well from chapter 5, the idea being that one root can be always constructed, by reasoning by contradiction, and doing some analysis around the minimum of $|P|$, and then a recurrence on the degree $n \in \mathbb{N}$ does the rest. □

Next in line, we have the rational functions, which are defined as follows:

THEOREM 6.5. *The quotients of complex polynomials, called rational funtions, when written in reduced form, as follows, with P, Q prime to each other,*

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

are well-defined and continuous outside the zeroes $P_f \subset \mathbb{C}$ of Q, called poles of f:

$$f : \mathbb{C} - P_f \to \mathbb{C}$$

In addition, the rational functions, regarded as algebraic expressions, are stable under summing, making products and taking inverses.

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

(1) First of all, we can surely talk about quotients of polynomials, $f = P/Q$, regarded as abstract algebraic expressions, and the last assertion is clear, because we can indeed perform sums, products, and take inverses, by using the following formulae:

$$\frac{P}{Q} + \frac{R}{S} = \frac{PS + QR}{QS} \quad , \quad \frac{P}{Q} \cdot \frac{R}{S} = \frac{PR}{QS} \quad , \quad \left(\frac{P}{Q}\right)^{-1} = \frac{Q}{P}$$

(2) The question is now, given a rational function $f \in \mathbb{C}(X)$, can we regard it as a complex function? In general, we cannot say that we have $f : \mathbb{C} \to \mathbb{C}$, for instance because $f(x) = x^{-1}$ is not defined at $x = 0$. More generally, assuming $f = P/Q$ with $P, Q \in \mathbb{C}$, we cannot talk about $f(x)$ when x is a root of Q, unless of course we are in the special situation where x is a root of P too, and we can simplify the fraction.

(3) In view of this discussion, in order to solve our question, we must avoid the situation where P, Q have common roots. But this can be done by writing our rational function $f \in \mathbb{C}(X)$ in reduced form, as follows, with $P, Q \in \mathbb{C}[X]$ prime to each other:

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

(4) Now with this convention made, it is clear that f is well-defined, and continuous too, outside of the zeroes of f. Now since these zeroes can be obviously recovered from the knowledge of f itself, as being the points where "f explodes", we can call them poles of f, and so we have a function $f : \mathbb{C} - P_f \to \mathbb{C}$, as in the statement. □

As a comment here, the term "pole" does not come from the Poles who invented this, but rather from the fact that, when trying to draw the graph of f, or rather imagine that graph, which takes place in $2+2 = 4$ real dimensions, we are faced with some sort of tent, which is suspended by infinite poles, which lie, guess where, at the poles of f.

Getting back now to Theorem 6.5, as stated, that is obviously a mixture of algebra and analysis. So, let us first further clarify the algebra part. We know that the rational functions are stable under summing, making products and taking inverses, and this makes the link with the following notion, from number theory and abstract algebra:

DEFINITION 6.6. *A field is a set F with a sum operation $+$ and a product operation $\times$, subject to the following conditions:*

(1) *$a + b = b + a$, $a + (b + c) = (a + b) + c$, there exists $0 \in F$ such that $a + 0 = 0$, and any $a \in F$ has an inverse $-a \in F$, satisfying $a + (-1) = 0$.*
(2) *$ab = ba$, $a(bc) = (ab)c$, there exists $1 \in F$ such that $a1 = a$, and any $a \neq 0$ has an inverse $a^{-1} \in F$, satisfying $aa^{-1} = 1$.*
(3) *The sum and product are compatible via $a(b + c) = ab + ac$.*

As basic examples of fields, we have the rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, and the complex numbers $\mathbb{C}$. Some further examples of fields of numbers, which are more specialized, and useful in number theory, can be constructed as well. In view of this, it is useful to think of any field F as being a "field of numbers", and this because the elements $a, b, c, \ldots \in F$ behave under the operations $+$ and $\times$ exactly as the usual numbers do.

In what regards the various spaces of functions, such as the polynomials $\mathbb{C}[X]$, or the continuous functions $C(\mathbb{R})$, these certainly have sum and product operations $+$ and $\times$, but are in general not fields, because they do not satisfy the following field axiom:

$$f \neq 0 \implies \exists f^{-1}$$

However, and here comes our point, Theorem 6.5 tells us that the rational functions form a field. This is quite interesting, and opposite to the general spirit of analysis and function spaces, which are in general not fields. Let us record this finding, as follows:

DEFINITION 6.7. *We denote by $\mathbb{C}(X)$ the field of rational functions*

$$f = \frac{P}{Q} \quad , \quad P, Q \in \mathbb{C}[X]$$

with the usual sum and product operations $+$ and $\times$ for the rational functions.

To be more precise, this is some sort of reformulation of Theorem 6.5, or rather of the algebraic content of Theorem 6.5, telling us that the rational functions form indeed a field. And to the question, how can a theorem suddenly become a definition, the answer is that this is rather commonplace in mathematics, and especially in algebra.

Back now to analysis, let us point out that, contrary to what the above might suggest, everyhing does not always extend trivally from the real to the complex case. For instance, we have the following result, that we already talked about a bit in chapter 5:

PROPOSITION 6.8. *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, exactly as in the real case, with the convergence being justified by:

$$\left| \sum_{n=0}^{\infty} x^n \right| \leq \sum_{n=0}^{\infty} |x|^n = \frac{1}{1-|x|}$$

As for the last assertion, this is something informal, the idea being as follows:

(1) To start with, for $x = 1/2$ our formula is clear, by cutting the interval $[0, 2]$ into half, and then half again, and again, and so on, which leads to:

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = 2$$

More generally, for $x \in (-1, 1)$ the meaning of the formula in the statement is something quite intuitive, geometrically speaking, by using a similar argument.

(2) However, for $x \in \mathbb{C} - \mathbb{R}$ we are led into a mysterious spiral, whose geometric understanding is a challenging problem. In fact, the only case where the formula is "obvious" is when $x = rw$, with $r \in [0, 1)$, and with w being a root of unity, $w^N = 1$.

(3) To be more precise, in this case the terms of $1 + x + x^2 + \ldots$ organize on the N rays $0 - 1, 0 - w, 0 - w^2, \ldots, 0 - w^{N-1}$ emanating from the origin, the sum can be performed on each ray, and the result eventually follows from the following computation:

$$1 + rw + (rw)^2 + \ldots = \frac{1}{1 - r^N} \cdot \frac{1 - r^N}{1 - rw} = \frac{1}{1 - rw}$$

(4) Summarizing, we are led to the conclusions in the statement, and of course with a challenge for you, dear reader, to further meditate on all this. □

Getting now to more complicated functions, such as sin, cos, exp, log, again many things extend well from real to complex, the basic theory here being as follows:

THEOREM 6.9. *The functions* $\sin, \cos, \exp, \log$ *have complex extensions, given by*

$$\sin x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l+1}}{(2l+1)!} \quad , \quad \cos x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l}}{(2l)!}$$

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \quad , \quad \log(1 + x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k}$$

with $|x| < 1$ *needed for* log*, which are continuous over their domain, and satisfy the formulae* $e^{x+y} = e^x e^y$ *and* $e^{ix} = \cos x + i \sin x$*.*

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

(1) We already know about e^x from chapter 5, the idea being that the convergence of the series, and then the continuity of e^x, come from the following estimate:

$$|e^x| \leq \sum_{k=0}^{\infty} \frac{|x|^k}{k!} = e^{|x|} < \infty$$

(2) Regarding $\sin x$, the same method works, with the following estimate:

$$|\sin x| \leq \sum_{l=0}^{\infty} \frac{|x|^{2l+1}}{(2l+1)!} \leq \sum_{k=0}^{\infty} \frac{|x|^k}{k!} = e^{|x|}$$

(3) The same goes for $\cos x$, the estimate here being as follows:

$$|\cos x| \leq \sum_{l=0}^{\infty} \frac{|x|^{2l}}{(2l)!} \leq \sum_{k=0}^{\infty} \frac{|x|^k}{k!} = e^{|x|}$$

(4) Regarding now the formulae satisfied by $\sin, \cos, \exp$, we already know from chapter 5 that the exponential has the following property, exactly as in the real case:

$$e^{x+y} = e^x e^y$$

We also have the following formula, connecting $\sin, \cos, \exp$, again as before:

$$\begin{aligned} e^{ix} &= \sum_{k=0}^{\infty} \frac{(ix)^k}{k!} \\ &= \sum_{k=2l} \frac{(ix)^k}{k!} + \sum_{k=2l+1} \frac{(ix)^k}{k!} \\ &= \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l}}{(2l)!} + i \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l+1}}{(2l+1)!} \\ &= \cos x + i \sin x \end{aligned}$$

(5) In order to discuss now the complex logarithm function log, let us first study some more the complex exponential function exp. 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:

$$e^x = e^y \iff x - y \in 2\pi i \mathbb{Z}$$

(6) With these ingredients in hand, we can now talk about log. Indeed, let us fix a horizontal strip in the complex plane, having width 2π:

$$S = \left\{ x + iy \middle| x \in \mathbb{R}, y \in [a, a + 2\pi) \right\}$$

We know from the above that the restriction map $\exp : S \to \mathbb{C} - \{0\}$ is bijective, so we can define log as to be the inverse of this map:

$$\log = \exp^{-1} : \mathbb{C} - \{0\} \to S$$

(7) In practice now, 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 $e^{\pi i} = -1$.

(8) Finally, assuming $|x| < 1$, we can consider the following series, which converges:

$$f(x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k}$$

We have then $e^{f(x)} = 1 + x$, and so $f(x) = \log(1 + x)$, when $1 + x \in S$. □

As an interesting consequence of the above result, which is of great practical interest, we have the following useful method, for remembering the basic math formulae:

METHOD 6.10. *Knowing* $e^x = \sum_k x^k/k!$ *and* $e^{ix} = \cos x + i \sin x$ *gives you*

$$\cos(x + y) = \cos x \cos y - \sin x \sin y$$
$$\sin(x + y) = \sin x \cos y + \cos x \sin y$$

right away, in case you forgot these formulae, as well as

$$\cos x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l}}{(2l)!} \quad , \quad \sin x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l+1}}{(2l+1)!}$$

again, right away, in case you forgot these formulae.

To be more precise, assume that we forgot everything trigonometry, which is something that can happen to everyone, in the real life, but know the formulae $e^x = \sum_k x^k/k!$ and $e^{ix} = \cos x + i \sin x$. Then, we can recover the formulae for sums, as follows:

$$e^{i(x+y)} = e^{ix} e^{iy} \implies \cos(x + y) + i \sin(x + y) = (\cos x + i \sin x)(\cos y + i \sin y)$$
$$\implies \begin{cases} \cos(x + y) = \cos x \cos y - \sin x \sin y \\ \sin(x + y) = \sin x \cos y + \cos x \sin y \end{cases}$$

And isn't this smart. Also, and even more impressively, we can recover the Taylor formulae for $\sin, \cos$, which are certainly difficult to memorize, as follows:

$$e^{ix} = \sum_k \frac{(ix)^k}{k!} \implies \cos x + i \sin x = \sum_k \frac{(ix)^k}{k!}$$
$$\implies \begin{cases} \cos x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l}}{(2l)!} \\ \sin x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l+1}}{(2l+1)!} \end{cases}$$

Finally, in what regards log, there is a trick here too, which is partial, namely:

$$\log(\exp(x)) = x \implies \log\left(1 + x + \frac{x^2}{2} + \ldots\right) = x$$
$$\implies \log(1 + y) = y - \frac{y^2}{2} + \ldots$$

To be more precise, $\log(1 + y) \simeq y$ is clear, and with a bit more work, that we will leave here as an instructive exercise, you can recover $\log(1 + y) = y - y^2/2$ too.

Moving ahead, Theorem 6.9 leads us into the question on whether the other formulae that we know about $\sin, \cos$, such as the values of these functions on sums $x + y$, or on doubles $2x$, extend to the complex setting. Things are quite tricky here, and in relation with this, we have the following result, which is something of general interest:

PROPOSITION 6.11. *The following functions, called hyperbolic cosine and sine,*

$$\cosh x = \frac{e^x + e^{-x}}{2} \quad , \quad \sinh x = \frac{e^x - e^{-x}}{2}$$

are subject to the following formulae:

(1) $e^x = \cosh x + \sinh x$.
(2) $\cosh(ix) = \cos x$, $\sinh(ix) = i \sin x$, *for* $x \in \mathbb{R}$.
(3) $\cosh(x + y) = \cosh x \cosh y + \sinh x \sinh y$.
(4) $\sinh(x + y) = \sinh x \cosh y + \cosh x \sinh y$.
(5) $\cosh x = \sum_l \frac{x^{2l}}{(2l)!}$, $\sinh x = \sum_l \frac{x^{2l+1}}{(2l+1)!}$.

PROOF. The formula (1) follows from definitions. As for (2), this follows from:

$$\cosh(ix) = \frac{e^{ix} + e^{-ix}}{2} = \frac{\cos x + i \sin x}{2} + \frac{\cos x - i \sin x}{2} = \cos x$$

$$\sinh(ix) = \frac{e^{ix} - e^{-ix}}{2} = \frac{\cos x + i \sin x}{2} - \frac{\cos x - i \sin x}{2} = i \sin x$$

Regarding now (3,4), observe first that the formula $e^{x+y} = e^x + e^y$ reads:

$$\cosh(x + y) + \sinh(x + y) = (\cosh x + \sinh x)(\cosh y + \sinh y)$$

Thus, we have some good explanation for (3,4), and in practice, these formulae can be checked by direct computation, as follows:

$$\frac{e^{x+y} + e^{-x-y}}{2} = \frac{e^x + e^{-x}}{2} \cdot \frac{e^y + e^{-y}}{2} + \frac{e^x - e^{-x}}{2} \cdot \frac{e^y - e^{-y}}{2}$$

$$\frac{e^{x+y} - e^{-x-y}}{2} = \frac{e^x - e^{-x}}{2} \cdot \frac{e^y + e^{-y}}{2} + \frac{e^x + e^{-x}}{2} \cdot \frac{e^y - e^{-y}}{2}$$

Finally, (5) is clear from the definition of cosh, sinh, and from $e^x = \sum_k \frac{x^k}{k!}$. □

Finally, we can talk as well about powers, in the following way:

FACT 6.12. *Under suitable assumptions, we can talk about x^y with $x, y \in \mathbb{C}$, and in particular about the complex functions a^x and x^a, with $a \in \mathbb{C}$.*

To be more precise, in what regards x^y, we already know from chapters 1-2 that things are quite tricky, even in the real case. In the complex case the same problems appear, along with some more, but these questions can be solved by using the above theory of $\exp, \log$. To be more precise, in order to solve the first question, we can set:

$$x^y = e^{y \log x}$$

But with this, after some discussion, we obtain what we want, namely definition and results. We will be back to these functions later, when we will have more tools for studying them. In fact, all of a sudden, we are now into quite complicated mathematics, and we cannot really deal with the problems left open above, with bare hands. More later.

At the level of the general theory now, the main tool for dealing with the continuous functions $f : \mathbb{R} \to \mathbb{R}$ was the intermediate value theorem. In the complex setting, that of the functions $f : \mathbb{C} \to \mathbb{C}$, we do not have such a theorem, at least in its basic formulation, because there is no order relation for the complex numbers, or things like complex intervals. However, the intermediate value theorem in its advanced formulation, that with connected sets, extends of course, and we have the following result:

THEOREM 6.13. *Assuming that $f : X \to \mathbb{C}$ with $X \subset \mathbb{C}$ is continuous, if the domain set X is connected, then so is its image $f(X)$.*

PROOF. This follows exactly as in the real case, with just a bit of discussion being needed, in relation with open and closed sets, and then connected sets, inside $\mathbb{C}$. □

6b. Holomorphic functions

Let us study now the differentiability of the complex functions $f : \mathbb{C} \to \mathbb{C}$. Things here are quite tricky, but let us start with a straightforward definition, exactly as the one in the real case, supposedly not leading to any trouble, as follows:

DEFINITION 6.14. *We say that a function $f : X \to \mathbb{C}$ is differentiable in the complex sense when the following limit is defined for any $x \in X$:*

$$f'(x) = \lim_{t \to 0} \frac{f(x + t) - f(x)}{t}$$

In this case, we also say that f is holomorphic, and we write $f \in H(X)$.

As basic examples, we have the power functions $f(x) = x^n$. Indeed, the derivative of such a power function can be computed exactly as in the real case, and we get:

$$\begin{aligned}
(x^n)' &= \lim_{t \to 0} \frac{(x + t)^n - x^n}{t} \\
&= \lim_{t \to 0} \frac{nx^{n-1}t + \binom{n}{2} x^{n-2}t^2 + \ldots + t^n}{t} \\
&= \lim_{t \to 0} \frac{nx^{n-1}t}{t} \\
&= nx^{n-1}
\end{aligned}$$

We will see later more computations of this type, similar to those from the real case. To summarize, our definition of differentiability seems to work nicely, so let us start developing some theory. The general results from the real case extend well, as follows:

PROPOSITION 6.15. *We have the following results:*

(1) $(f+g)' = f' + g'$.
(2) $(\lambda f)' = \lambda f'$.
(3) $(fg)' = f'g + fg'$.
(4) $(f \circ g)' = f'(g)g'$.

PROOF. These formulae are all clear from definitions, following exactly as in the real case. Thus, we are led to the conclusions in the statement. □

As an obvious consequence of (1,2) above, any poynomial $P \in \mathbb{C}[X]$ is differentiable, with its derivative being given by the same formula as in the real case, namely:

$$P(x) = \sum_{k=0}^{n} c_k x^k \implies P'(x) = \sum_{k=1}^{n} k c_k x^{k-1}$$

More generally, any rational function $f \in \mathbb{C}(X)$ is differentiable on its domain, that is, outsides its poles, because if we write $f = P/Q$ with $P, Q \in \mathbb{C}[X]$, we have:

$$f' = \left(\frac{P}{Q}\right)' = \frac{P'Q - PQ'}{Q^2}$$

Let us record our conclusions in a statement, as follows:

PROPOSITION 6.16. *The following happen:*

(1) *Any polynomial $P \in \mathbb{C}[X]$ is holomorphic, and in fact infinitely differentiable in the complex sense, with all its derivatives being polynomials.*
(2) *Any rational function $f \in \mathbb{C}(X)$ is holomorphic, and in fact infinitely differentiable, with all its derivatives being rational functions.*

PROOF. This follows indeed from the above discussion. □

Let us look now into more complicated complex functions that we know. And here, surprise, things are quite tricky, the result being as follows:

THEOREM 6.17. *The following happen:*

(1) $\sin, \cos, \exp, \log$ *are holomorphic, and in fact are infinitely differentiable, with their derivatives being given by the same formulae as in the real case.*
(2) *However, functions like $\bar{x}$ or $|x|$ are not holomorphic, and this because the limit defining $f'(x)$ depends on the way we choose $t \to 0$.*

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

(1) Here the first assertion is standard, because our functions $\sin, \cos, \exp, \log$ have Taylor series that we know, and the derivative can be therefore computed by using the

same rule as in the real case, similar to the one for polynomials, namely:

$$f(x)=\sum_{k=0}^{\infty}c_kx^k\implies f'(x)=\sum_{k=1}^{\infty}kc_kx^{k-1}$$

(2) Regarding now the function $f(x)=\bar{x}$, the point here is that we have:

$$\frac{f(x+t)-f(x)}{t}=\frac{\bar{x}+\bar{t}-\bar{x}}{t}=\frac{\bar{t}}{t}$$

But this limit does not converge with $t\to 0$, for instance because with $t\in\mathbb{R}$ we obtain 1 as limit, while with $t\in i\mathbb{R}$ we obtain -1 as limit. In fact, with $t=rw$ with $|w|=1$ fixed and $r\in\mathbb{R}$, $r\to 0$, we can obtain as limit any number on the unit circle:

$$\lim_{r\to 0}\frac{f(x+rw)-f(x)}{rw}=\lim_{r\to 0}\frac{r\bar{w}}{rw}=\bar{w}^2$$

(3) The situation for the function $f(x)=|x|$ is similar. To be more precise, we have:

$$\frac{f(x+rw)-f(x)}{rw}=\frac{|x+rw|-|x|}{r}\cdot\bar{w}$$

Thus with $|w|=1$ fixed and $r\to 0$ we obtain a certain multiple of $\bar{w}$, with the multiplication factor being computed as follows:

$$\begin{aligned}\frac{|x+rw|-|x|}{r}&=\frac{|x+rw|^2-|x|^2}{(|x+rw|+|x|)r}\\&\simeq\frac{xr\bar{w}+\bar{x}rw}{2|x|r}\\&=Re\left(\frac{x\bar{w}}{|x|}\right)\end{aligned}$$

Now by making w vary on the unit circle, as in (2) above, we can obtain in this way limits pointing in all possible directions, so our limit does not converge, as stated. □

The above result might seem quite surprising, because we are so used, from the real case, with the notion of differentiability to correspond to some form of "smoothness" of the function, and to be more precisely, to "smoothness at first order". Or, if you prefer, to the "non-bumpiness" of the function. So, we are led to the following dilemma:

DILEMMA 6.18. *It's either that $\bar{x}$ and $|x|$ are smooth, as the intuition suggests, and we are wrong with our definition of differentiability. Or that $\bar{x}$ and $|x|$ are bumpy, while this being not very intuitive, and we are right with our definition of differentiability.*

And we won't get discouraged by this. After all, this is just some empty talking, and if there is something to rely upon, mathematics and computations, these are the computations from the proof of Theorem 6.17. So, moving ahead now, based on that computations, let us formulate the following definition, coming as a complement to Definition 6.14:

DEFINITION 6.19. *A function $f : X \to \mathbb{C}$ is called differentiable:*

(1) *In the real sense, if the following two limits converge, for any $x \in X$:*

$$f_1'(x) = \lim_{t \in \mathbb{R} \to 0} \frac{f(x+t) - f(x)}{t} \quad , \quad f_i'(x) = \lim_{t \in i\mathbb{R} \to 0} \frac{f(x+t) - f(x)}{t}$$

(2) *In a radial sense, if the following limit converges, for any $x \in X$, and $w \in \mathbb{T}$:*

$$f_w'(x) = \lim_{t \in w\mathbb{R} \to 0} \frac{f(x+t) - f(x)}{t}$$

(3) *In the complex sense, if the following limit converges, for any $x \in X$:*

$$f'(x) = \lim_{t \to 0} \frac{f(x+t) - f(x)}{t}$$

If f is differentiable in the complex sense, we also say that f is holomorphic.

We can see now more clearly what is going on. We have (3) $\implies$ (2) $\implies$ (1) in general, and most of the functions that we know, namely the polynomials, the rational functions, and $\sin, \cos, \exp, \log$, satisfy (3). As for the functions $\bar{x}, |x|$, these do not satisfy (3), and do not satisfy (2) either, but they satisfy however (1). It is possible to say more about all this, and we will certainly come back to this topic, later in this book.

Back to business now, all the examples of holomorphic functions that we have are infinitely differentiable, and this raises the question of finding a function such that f' exists, while f'' does not exist. Quite surprisingly, we will see that such functions do not exist. In order to get into this latter phenomenon, let us start with:

THEOREM 6.20. *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. This 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, with $x_n = |c_n x^n|$ we obtain that the convergence radius $R \in [0, \infty]$ exists, and is given by the formula in the statement. Finally, for the examples and counterexamples at the end, when $|x| = R$, the simplest here is to use $f(x) = \sum_n x^n$, where $R = 1$. □

Back now to our questions regarding derivatives, we have:

THEOREM 6.21. *Assuming that a function $f : X \to \mathbb{C}$ is analytic, in the sense that it is a series, around each point $x \in X$,*

$$f(x+t) = \sum_{n=0}^{\infty} c_n t^n$$

it follows that f is infinitely differentiable, in the complex sense. In particular, f' exists, and so f is holomorphic in our sense.

PROOF. Assuming that f is analytic, as in the statement, we have:

$$f'(x+t) = \sum_{n=1}^{\infty} n c_n t^{n-1}$$

Moreover, the radius of convergence is the same, as shown by the following computation, using the Cauchy formula for the convergence radius, and $\sqrt[n]{n} \to 1$:

$$\frac{1}{R'} = \limsup_{n\to\infty} \sqrt[n]{|nc_n|} = \limsup_{n\to\infty} \sqrt[n]{|c_n|} = \frac{1}{R}$$

Thus f' exists and is analytic, on the same domain, and this gives the result. □

6c. Cauchy formula

Our goal in what follows will be that of proving that any holomorphic function is analytic. This is something quite subtle, which cannot be proved with bare hands, and requires lots of preliminaries. Getting to these preliminaries now, our claim is that a lot of useful knowledge, in order to deal with the holomorphic functions, can be gained by further studying the analytic functions, and even the usual polynomials $P \in \mathbb{C}[X]$.

So, let us further study the polynomials $P \in \mathbb{C}[X]$, and other analytic functions. We already know from chapter 5 that in the polynomial case, $P \in \mathbb{C}[X]$, some interesting things happen, because any such polynomial has a root, and even $\deg(P)$ roots, after a recurrence. Kepping looking at polynomials, with the same methods, we are led to:

THEOREM 6.22. *Any polynomial $P \in \mathbb{C}[X]$ satisfies the maximum principle, in the sense that given a domain D, with boundary γ, we have:*

$$\exists x \in \gamma \quad , \quad |P(x)| = \max_{y\in D} |P(y)|$$

That is, the maximum of $|P|$ over a domain is attained on its boundary.

PROOF. In order to prove this, we can split D into connected components, and then assume that D is connected. Moreover, we can assume that D has no holes, and so is homeomorphic to a disk, and even assume that D itself is a disk. But with this assumption made, the result follows from by contradiction, by using the same arguments as in the

proof of the existence of a root, from chapter 5. To be more precise, assume $\deg P \geq 1$, and that the maximum of $|P|$ is attained at the center of a disk $D = D(z, r)$:

$$|P(z)| = \max_{x \in D} |P(x)|$$

We can write then $P(z + t) \simeq P(z) + ct^k$ with $c \neq 0$, for t small, and by suitably choosing the argument of t on the unit circle we conclude, exactly as in chapter 5, that the function $|P|$ cannot have a local maximum at z, as stated. □

A good explanation for the fact that the maximum principle holds for polynomials $P \in \mathbb{C}[X]$ could be that the values of such a polynomial inside a disk can be recovered from its values on the boundary. And fortunately, this is indeed the case, and we have:

THEOREM 6.23. *Given a polynomial $P \in \mathbb{C}[X]$, and a disk D, with boundary γ:*

(1) *P satisfies the plain mean value formula $P(x) = \int_D P(y)dy$.*
(2) *P satisfies the boundary mean value formula $P(x) = \int_\gamma P(y)dy$.*

PROOF. As a first observation, the two mean value formulae in the statement are equivalent, by restricting the attention to disks D, having as boundaries circles γ, and using annuli and polar coordinates for the proof of the equivalence. As for the formulae themselves, these can be checked by direct computation for a disk D, with the formulation in (2) being the most convenient. Indeed, for a monomial $P(x) = x^n$ we have:

$$\begin{aligned}
\int_\gamma y^n dy &= \frac{1}{2\pi}\int_0^{2\pi}(x + re^{it})^n dt \\
&= \frac{1}{2\pi}\int_0^{2\pi}\sum_{k=0}^{n}\binom{n}{k}x^k(re^{it})^{n-k}dt \\
&= \sum_{k=0}^{n}\binom{n}{k}x^k r^{n-k}\frac{1}{2\pi}\int_0^{2\pi}e^{i(n-k)t}dt \\
&= \sum_{k=0}^{n}\binom{n}{k}x^k r^{n-k}\delta_{kn} \\
&= x^n
\end{aligned}$$

Here we have used the following key identity, valid for any exponent $m \in \mathbb{Z}$:

$$\begin{aligned}
\frac{1}{2\pi}\int_0^{2\pi}e^{imt}dt &= \frac{1}{2\pi}\int_0^{2\pi}\cos(mt) + i\sin(mt)dt \\
&= \delta_{m0} + i \cdot 0 \\
&= \delta_{m0}
\end{aligned}$$

Thus, we have the result for monomials, and the general case follows by linearity. □

All the above is very nice, but we can in fact do even better, with a more powerful integration formula. Let us start with some preliminaries. We first have:

PROPOSITION 6.24. *We can integrate functions f over curves γ by setting*

$$\int_\gamma f(x)dx = \int_a^b f(\gamma(t))\gamma'(t)dt$$

with this quantity being independent on the parametrization $\gamma : [a, b] \to \mathbb{C}$.

PROOF. We must prove that the quantity in the statement is independent on the parametrization. In other words, we must prove that if we pick two different parametrizations $\gamma, \eta : [a, b] \to \mathbb{C}$ of our curve, then we have the following formula:

$$\int_a^b f(\gamma(t))\gamma'(t)dt = \int_a^b f(\eta(t))\eta'(t)dt$$

But for this purpose, let us write $\gamma = \eta\phi$, with $\phi : [a, b] \to [a, b]$ being a certain function, that we can assume to be bijective, via an elementary cut-and-paste argument. By using the chain rule for derivatives, and the change of variable formula, we have:

$$\begin{aligned}\int_a^b f(\gamma(t))\gamma'(t)dt &= \int_a^b f(\eta\phi(t))(\eta\phi)'(t)dt\\ &= \int_a^b f(\eta\phi(t))\eta'(\phi(t))\phi'(t)dt\\ &= \int_a^b f(\eta(t))\eta'(t)dt\end{aligned}$$

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

The main properties of the above integration method are as follows:

PROPOSITION 6.25. *We have the following formula, for a union of paths:*

$$\int_{\gamma\cup\eta} f(x)dx = \int_\gamma f(x)dx + \int_\eta f(x)dx$$

Also, when reversing the path, the integral changes its sign.

PROOF. Here the first assertion is clear from definitions, and the second assertion comes from the change of variable formula, by using Proposition 6.24. □

Now by getting back to polynomials, we have the following result:

THEOREM 6.26. *Any polynomial $P \in \mathbb{C}[X]$ satisfies the Cauchy formula*

$$P(x) = \frac{1}{2\pi i}\int_\gamma \frac{P(y)}{y - x}\,dy$$

with the integration over γ being constructed as above.

PROOF. This follows by using abstract arguments and computations similar to those in the proof of Theorem 6.23. Indeed, by linearity we can assume $P(x) = x^n$. Also, by using a cut-and-paste argument, we can assume that we are on a circle:

$$\gamma : [0, 2\pi] \to \mathbb{C} \quad , \quad \gamma(t) = x + re^{it}$$

By using now the computation from the proof of Theorem 6.23, we obtain:

$$\begin{aligned} \int_\gamma \frac{y^n}{y-x}\,dy &= \int_0^{2\pi} \frac{(x+re^{it})^n}{re^{it}}\,rie^{it}dt \\ &= i\int_0^{2\pi} (x+re^{it})^n dt \\ &= i \cdot 2\pi x^n \end{aligned}$$

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

All this is quite interesting, and we are now into some serious mathematics here. Importantly, Theorem 6.22, Theorem 6.23 and Theorem 6.26 provide us with a path for proving the converse of Theorem 6.21. Indeed, if we manage to prove the Cauchy formula for any holomorphic function $f : X \to \mathbb{C}$, then it will follow that our function is in fact analytic, and so infinitely differentiable. So, let us start with the following result:

THEOREM 6.27. *The Cauchy formula, namely*

$$f(x) = \frac{1}{2\pi i}\int_\gamma \frac{f(y)}{y-x}\,dy$$

holds for any holomorphic function $f : X \to \mathbb{C}$.

PROOF. This is something standard, which can be proved as follows:

(1) Our first claim is that given $f \in H(X)$, with $f' \in C(X)$, the integral of f' vanishes on any path. Indeed, by using the change of variable formula, we have:

$$\begin{aligned} \int_\gamma f'(x)dx &= \int_a^b f'(\gamma(t))\gamma'(t)dt \\ &= f(\gamma(b)) - f(\gamma(a)) \\ &= 0 \end{aligned}$$

(2) Our second claim is that given $f \in H(X)$ and a triangle $\Delta \subset X$, we have:

$$\int_\Delta f(x)dx = 0$$

Indeed, let us call $\Delta = ABC$ our triangle. Now consider the midpoints A', B', C' of the edges BC, CA, AB, and then consider the following smaller triangles:

$$\Delta_1 = AC'B' \quad , \quad \Delta_2 = BA'C' \quad , \quad \Delta_3 = CB'A' \quad , \quad \Delta_4 = A'B'C'$$

These smaller triangles partition then Δ, and due to our above conventions for the vertex ordering, which produce cancellations when integrating over them, we have:

$$\int_\Delta f(x)dx = \sum_{i=1}^4 \int_{\Delta_i} f(x)dx$$

Thus we can pick, among the triangles Δ_i, a triangle $\Delta^{(1)}$ such that:

$$\left|\int_\Delta f(x)dx\right| \leq 4\left|\int_{\Delta^{(1)}} f(x)dx\right|$$

In fact, by repeating the procedure, we obtain triangles $\Delta^{(n)}$ such that:

$$\left|\int_\Delta f(x)dx\right| \leq 4^n\left|\int_{\Delta^{(n)}} f(x)dx\right|$$

(3) Now let z be the limiting point of these triangles $\Delta^{(n)}$, and fix $\varepsilon > 0$. By using the fact that the functions $1, x$ integrate over paths up to 0, coming from (1), we obtain the following estimate, with $n \in \mathbb{N}$ being big enough, and L being the perimeter of Δ:

$$\begin{aligned}
\left|\int_{\Delta^{(n)}} f(x)dx\right| &= \left|\int_{\Delta^{(n)}} f(x) - f(z) - f'(z)(x-z)dx\right| \\
&\leq \int_{\Delta^{(n)}} |f(x) - f(z) - f'(z)(x-z)|\,dx \\
&\leq \int_{\Delta^{(n)}} \varepsilon|x-z|dx \\
&\leq \varepsilon(2^{-n}L)^2
\end{aligned}$$

Now by combining this with the estimate in (2), this proves our claim.

(4) The rest is quite routine. First, we can pass from triangles to boundaries of convex sets, in a straightforward way, with the same conclusion as in (2), namely:

$$\int_\gamma f(x)dx = 0$$

Now getting back to what we want to prove, namely Cauchy formula for an arbitrary holomorphic function $f \in H(X)$, let $x \in X$, and consider the following function:

$$g(y) = \begin{cases} \frac{f(y)-f(x)}{y-x} & (y \neq x) \\ f'(x) & (y = x) \end{cases}$$

Thus, assuming that γ encloses a convex set, we can apply what we found, namely vanishing of the integral, to this function g, and we obtain the Cauchy formula for f.

(5) Finally, the extension to general curves is standard, and standard as well is the discussion of what exactly happens at x, in the above proof. See Rudin [75]. □

As a main application of the Cauchy formula, we have:

THEOREM 6.28. *The following conditions are equivalent, for a function $f : X \to \mathbb{C}$:*

(1) *f is holomorphic.*
(2) *f is infinitely differentiable.*
(3) *f is analytic.*
(4) *The Cauchy formula holds for f.*

PROOF. This is quite routine from what we have, the idea being as follows:

(1) $\implies$ (4) is non-trivial, but we know this from Theorem 6.27.

(4) $\implies$ (3) is something trivial, because we can expand the series in the Cauchy formula, and we conclude that our function is indeed analytic.

(3) $\implies$ (2) is something elementary, that we know from Theorem 6.22.

(2) $\implies$ (1) is again something elementary, that we know from Theorem 6.22. □

As another application of the Cauchy formula, we have:

THEOREM 6.29. *Any holomorphic function $f : X \to \mathbb{C}$ satisfies the maximum principle, in the sense that given a domain D, with boundary γ, we have:*

$$\exists x \in \gamma \quad , \quad |f(x)| = \max_{y \in D} |f(y)|$$

That is, the maximum of $|f|$ over a domain is attained on its boundary.

PROOF. This follows indeed from the Cauchy formula. Observe that the converse is not true, for instance because functions like $\bar{x}$ satisfy too the maximum principle. We will be back to this later, when talking about harmonic functions. □

As before with polynomials, a good explanation for the fact that the maximum principle holds could be that the values of our function inside a disk can be recovered from its values on the boundary. And fortunately, this is indeed the case, and we have:

THEOREM 6.30. *Given an holomorphic function $f : X \to \mathbb{C}$, and a disk D, with boundary γ, the following happen:*

(1) *f satisfies the plain mean value formula $f(x) = \int_D f(y)dy$.*
(2) *f satisfies the boundary mean value formula $f(x) = \int_\gamma f(y)dy$.*

PROOF. As usual, this follows from the Cauchy formula, with of course some care in passing from integrals constructed as in Proposition 6.24 to integrals viewed as averages, which are those that we refer to, in the present statement. □

Finally, as yet another application of the Cauchy formula, which is something nice-looking and conceptual, we have the following statement, called Liouville theorem:

THEOREM 6.31. *An entire, bounded holomorphic function*

$$f : \mathbb{C} \to \mathbb{C} \quad , \quad |f| \leq M$$

must be constant. In particular, knowing $f \to 0$ with $z \to \infty$ gives $f = 0$.

PROOF. This follows as usual from the Cauchy formula, namely:

$$f(x) = \frac{1}{2\pi i} \int_\gamma \frac{f(y)}{y - x}\, dy$$

Alternatively, we can view this as a consequence of Theorem 6.30, because given two points $x \neq y$, we can view the values of f at these points as averages over big disks centered at these points, say $D = D_x(R)$ and $E = D_y(R)$, with $R >> 0$:

$$f(x) = \int_D f(z)dz \quad , \quad f(y) = \int_E f(z)dz$$

Indeed, the point is that when the radius goes to ∞, these averages tend to be equal, and so we have $f(x) \simeq f(y)$, which gives $f(x) = f(y)$ in the limit. □

6d. Stieltjes inversion

We would like to end this chapter with an interesting application of the complex numbers to probability theory. Let us recall from chapter 4 that we have:

DEFINITION 6.32. *Let X be a probability space, that is, a space with a probability measure, and with the corresponding integration denoted $\mathbb{E}$, and called expectation.*

(1) *The random variables are the real functions $f \in L^\infty(X)$.*
(2) *The moments of such a variable are the numbers $M_k(f) = \mathbb{E}(f^k)$.*
(3) *The law of such a variable is the measure given by $M_k(f) = \int_{\mathbb{R}} x^k d\mu_f(x)$.*

Here the fact that μ_f exists indeed is not trivial. By linearity, we would like to have a real probability measure making hold the following formula, for any $P \in \mathbb{R}[X]$:

$$\mathbb{E}(P(f)) = \int_{\mathbb{R}} P(x) d\mu_f(x)$$

By using a standard continuity argument, it is enough to have this formula for the characteristic functions χ_I of the arbitrary measurable sets of real numbers $I \subset \mathbb{R}$:

$$\mathbb{E}(\chi_I(f)) = \int_{\mathbb{R}} \chi_I(x) d\mu_f(x)$$

But this latter formula, which reads $\mathbb{P}(f \in I) = \mu_f(I)$, can serve as a definition for μ_f, and we are done. Alternatively, assuming some familiarity with measure theory, μ_f is the push-forward of the probability measure on X, via the function $f : X \to \mathbb{R}$.

The problem is now, how to recover a probability measure out of its moments. And the answer here, which is something non-trivial, is as follows:

THEOREM 6.33. *The density of a real probability measure μ can be recaptured from the sequence of moments $\{M_k\}_{k\geq 0}$ via the Stieltjes inversion formula*

$$d\mu(x) = \lim_{t\searrow 0} -\frac{1}{\pi}\, Im\left(G(x+it)\right)\cdot dx$$

where the function on the right, given in terms of moments by

$$G(\xi) = \xi^{-1} + M_1\xi^{-2} + M_2\xi^{-3} + \dots$$

is the Cauchy transform of the measure μ.

PROOF. The Cauchy transform of our measure μ is given by:

$$\begin{aligned} G(\xi) &= \xi^{-1}\sum_{k=0}^{\infty} M_k\xi^{-k} \\ &= \int_{\mathbb{R}} \frac{\xi^{-1}}{1-\xi^{-1}y}\, d\mu(y) \\ &= \int_{\mathbb{R}} \frac{1}{\xi - y}\, d\mu(y) \end{aligned}$$

Now with $\xi = x + it$, we obtain the following formula:

$$\begin{aligned} Im(G(x+it)) &= \int_{\mathbb{R}} Im\left(\frac{1}{x-y+it}\right) d\mu(y) \\ &= \int_{\mathbb{R}} \frac{1}{2i}\left(\frac{1}{x-y+it} - \frac{1}{x-y-it}\right) d\mu(y) \\ &= -\int_{\mathbb{R}} \frac{t}{(x-y)^2+t^2}\, d\mu(y) \end{aligned}$$

By integrating over $[a,b]$ we obtain, with the change of variables $x = y + tz$:

$$\begin{aligned} \int_a^b Im(G(x+it))dx &= -\int_{\mathbb{R}}\int_a^b \frac{t}{(x-y)^2+t^2}\, dx\, d\mu(y) \\ &= -\int_{\mathbb{R}}\int_{(a-y)/t}^{(b-y)/t} \frac{t}{(tz)^2+t^2}\, t\, dz\, d\mu(y) \\ &= -\int_{\mathbb{R}}\int_{(a-y)/t}^{(b-y)/t} \frac{1}{1+z^2}\, dz\, d\mu(y) \\ &= -\int_{\mathbb{R}} \left(\arctan\frac{b-y}{t} - \arctan\frac{a-y}{t}\right) d\mu(y) \end{aligned}$$

Now observe that with $t \searrow 0$ we have:

$$\lim_{t\searrow 0}\left(\arctan\frac{b-y}{t}-\arctan\frac{a-y}{t}\right)=\begin{cases}\frac{\pi}{2}-\frac{\pi}{2}=0 & (y<a)\\ \frac{\pi}{2}-0=\frac{\pi}{2} & (y=a)\\ \frac{\pi}{2}-(-\frac{\pi}{2})=\pi & (a<y<b)\\ 0-(-\frac{\pi}{2})=\frac{\pi}{2} & (y=b)\\ -\frac{\pi}{2}-(-\frac{\pi}{2})=0 & (y>b)\end{cases}$$

We therefore obtain the following formula:

$$\lim_{t\searrow 0}\int_a^b Im(G(x+it))dx=-\pi\left(\mu(a,b)+\frac{\mu(a)+\mu(b)}{2}\right)$$

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

Before getting further, let us mention that the above result does not fully solve the moment problem, because we still have the question of understanding when a sequence of numbers $M_1, M_2, M_3, \ldots$ can be the moments of a measure μ. We have here:

THEOREM 6.34. *A sequence of numbers $M_0, M_1, M_2, M_3, \ldots \in \mathbb{R}$, with $M_0 = 1$, is the series of moments of a real probability measure μ precisely when:*

$$\left|M_0\right| \geq 0 \quad , \quad \begin{vmatrix} M_0 & M_1\\ M_1 & M_2\end{vmatrix} \geq 0 \quad , \quad \begin{vmatrix} M_0 & M_1 & M_2\\ M_1 & M_2 & M_3\\ M_2 & M_3 & M_4\end{vmatrix} \geq 0 \quad , \quad \ldots$$

That is, the associated Hankel determinants must be all positive.

PROOF. This is something a bit more advanced, the idea being as follows:

(1) As a first observation, the positivity conditions in the statement tell us that the following associated linear forms must be positive:

$$\sum_{i,j=1}^{n} c_i\bar{c}_j M_{i+j} \geq 0$$

(2) But this is something very classical, in one sense the result being elementary, coming from the following computation, which shows that we have positivity indeed:

$$\int_{\mathbb{R}}\left|\sum_{i=1}^{n} c_i x^i\right|^2 d\mu(x)=\int_{\mathbb{R}}\sum_{i,j=1}^{n} c_i\bar{c}_j x^{i+j}d\mu(x)=\sum_{i,j=1}^{n} c_i\bar{c}_j M_{i+j}$$

(3) As for the other sense, here the result comes once again from the above formula, this time via some standard functional analysis. □

As a basic application of the Stieltjes formula, let us solve the moment problem for the Catalan numbers C_k, and for the central binomial coefficients D_k. We first have:

THEOREM 6.35. *The real measure having as even moments the Catalan numbers, $C_k = \frac{1}{k+1}\binom{2k}{k}$, and having all odd moments 0 is the measure*

$$\gamma_1 = \frac{1}{2\pi}\sqrt{4-x^2}dx$$

called Wigner semicircle law on $[-2,2]$.

PROOF. In order to apply the inversion formula, our starting point will be the formula from chapter 3 for the generating series of the Catalan numbers, namely:

$$\sum_{k=0}^{\infty} C_k z^k = \frac{1-\sqrt{1-4z}}{2z}$$

By using this formula with $z = \xi^{-2}$, we obtain the following formula:

$$\begin{aligned} G(\xi) &= \xi^{-1}\sum_{k=0}^{\infty} C_k \xi^{-2k} \\ &= \xi^{-1}\cdot\frac{1-\sqrt{1-4\xi^{-2}}}{2\xi^{-2}} \\ &= \frac{\xi}{2}\left(1-\sqrt{1-4\xi^{-2}}\right) \\ &= \frac{\xi}{2}-\frac{1}{2}\sqrt{\xi^2-4} \end{aligned}$$

Now let us apply Theorem 6.33. The study here goes as follows:

(1) According to the general philosophy of the Stieltjes formula, the first term, namely $\xi/2$, which is "trivial", will not contribute to the density.

(2) As for the second term, which is something non-trivial, this will contribute to the density, the rule here being that the square root $\sqrt{\xi^2-4}$ will be replaced by the "dual" square root $\sqrt{4-x^2}\,dx$, and that we have to multiply everything by $-1/\pi$.

(3) As a conclusion, by Stieltjes inversion we obtain the following density:

$$d\mu(x) = -\frac{1}{\pi}\cdot -\frac{1}{2}\sqrt{4-x^2}\,dx = \frac{1}{2\pi}\sqrt{4-x^2}dx$$

Thus, we have obtained the mesure in the statement, and we are done. □

We have the following version of the above result:

THEOREM 6.36. *The real measure having as sequence of moments the Catalan numbers, $C_k = \frac{1}{k+1}\binom{2k}{k}$, is the measure*

$$\pi_1 = \frac{1}{2\pi}\sqrt{4x^{-1}-1}\,dx$$

called Marchenko-Pastur law on $[0,4]$.

PROOF. As before, we use the standard formula for the generating series of the Catalan numbers. With $z = \xi^{-1}$ in that formula, we obtain the following formula:

$$\begin{aligned} G(\xi) &= \xi^{-1}\sum_{k=0}^{\infty} C_k\xi^{-k} \\ &= \xi^{-1}\cdot\frac{1-\sqrt{1-4\xi^{-1}}}{2\xi^{-1}} \\ &= \frac{1}{2}\left(1-\sqrt{1-4\xi^{-1}}\right) \\ &= \frac{1}{2}-\frac{1}{2}\sqrt{1-4\xi^{-1}} \end{aligned}$$

With this in hand, let us apply now the Stieltjes inversion formula, from Theorem 6.33. We obtain, a bit as before in Theorem 6.35, the following density:

$$d\mu(x) = -\frac{1}{\pi}\cdot-\frac{1}{2}\sqrt{4x^{-1}-1}\,dx = \frac{1}{2\pi}\sqrt{4x^{-1}-1}\,dx$$

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

Regarding now the central binomial coefficients, we have here:

THEOREM 6.37. *The real probability measure having as moments the central binomial coefficients, $D_k = \binom{2k}{k}$, is the measure*

$$\alpha_1 = \frac{1}{\pi\sqrt{x(4-x)}}\,dx$$

called arcsine law on $[0,4]$.

PROOF. We have the following computation, using formulae from chapter 3:

$$\begin{aligned} G(\xi) &= \xi^{-1}\sum_{k=0}^{\infty} D_k\xi^{-k} \\ &= \frac{1}{\xi}\sum_{k=0}^{\infty} D_k\left(-\frac{t}{4}\right)^k \\ &= \frac{1}{\xi}\cdot\frac{1}{\sqrt{1-4/\xi}} \\ &= \frac{1}{\sqrt{\xi(\xi-4)}} \end{aligned}$$

But this gives the density in the statement, via Theorem 6.33. □

Finally, we have the following version of the above result:

THEOREM 6.38. *The real probability measure having as moments the middle binomial coefficients, $E_k = \binom{k}{[k/2]}$, is the following law on $[-2,2]$,*

$$\sigma_1 = \frac{1}{2\pi}\sqrt{\frac{2+x}{2-x}}\,dx$$

called "square root" of the arcsine law on $[0,4]$.

PROOF. In terms of the central binomial coefficients D_k, we have:

$$E_{2k} = \binom{2k}{k} = \frac{(2k)!}{k!k!} = D_k$$

$$E_{2k-1} = \binom{2k-1}{k} = \frac{(2k-1)!}{k!(k-1)!} = \frac{D_k}{2}$$

Standard calculus based on the Taylor formula for $(1+t)^{-1/2}$ gives:

$$\frac{1}{2x}\left(\sqrt{\frac{1+2x}{1-2x}} - 1\right) = \sum_{k=0}^{\infty} E_k x^k$$

With $x = \xi^{-1}$ we obtain the following formula for the Cauchy transform:

$$\begin{aligned} G(\xi) &= \xi^{-1}\sum_{k=0}^{\infty} E_k \xi^{-k} \\ &= \frac{1}{\xi}\left(\sqrt{\frac{1+2/\xi}{1-2/\xi}} - 1\right) \\ &= \frac{1}{\xi}\left(\sqrt{\frac{\xi+2}{\xi-2}} - 1\right) \end{aligned}$$

By Stieltjes inversion we obtain the density in the statement. □

All this is very nice, and we are obviously building here, as this book goes by, some good knowledge in probability theory. We will be back to all this later.

6e. Exercises

Here are a few exercises, in relation with the material from this chapter:

EXERCISE 6.39. *Find a geometric proof for $1 + x + x^2 + \ldots = 1/(1-x)$.*

EXERCISE 6.40. *Clarify the definition and properties of x^y, with $x, y \in \mathbb{C}$.*

EXERCISE 6.41. *Complete our proof of the Cauchy formula, in the general case.*

EXERCISE 6.42. *Learn more complex analysis, including the residue formula.*

As a bonus exercise, learn some more probability, and random matrices.

CHAPTER 7

Fourier analysis

7a. Function spaces

In this chapter we go back to the functions of one real variable, $f : \mathbb{R} \to \mathbb{R}$ or $f : \mathbb{R} \to \mathbb{C}$, with some applications, by using our complex number technology. We will be mainly interested in constructing the Fourier transform, which is an operation $f \to \widehat{f}$ on such functions, which can solve various concrete questions. Some advertisement:

ADVERTISEMENT 7.1. *There is only one tool in mathematics, and that is the Fourier transform.*

Or at least, that's what some people say. And as fans of calculus, we should side with them. We will see for instance how, via Fourier transform, light can be decomposed into color components, distant galaxies can be observed, atoms can be distinguished from each other, the past and future of the universe can be predicted, and so on.

Before doing that, however, let us study the spaces that the functions $f : \mathbb{R} \to \mathbb{C}$ can form. These functions can be continuous, differentiable, infinitely differentiable, and so on, but there are many more properties that these functions can have, that we will investigate now. This will lead to various spaces of functions $f : \mathbb{R} \to \mathbb{C}$, that can be used, among others, in order to well-define the Fourier transform operation $f \to \widehat{f}$.

Let us start with some well-known and useful inequalities, as follows:

THEOREM 7.2. *Given two functions $f, g : \mathbb{R} \to \mathbb{C}$ and an exponent $p \geq 1$, we have*

$$\left(\int_{\mathbb{R}} |f+g|^p\right)^{1/p} \leq \left(\int_{\mathbb{R}} |f|^p\right)^{1/p} + \left(\int_{\mathbb{R}} |g|^p\right)^{1/p}$$

called Minkowski inequality. Also, assuming that $p, q \geq 1$ satisfy $1/p + 1/q = 1$, we have

$$\int_{\mathbb{R}} |fg| \leq \left(\int_{\mathbb{R}} |f|^p\right)^{1/p} \left(\int_{\mathbb{R}} |g|^q\right)^{1/q}$$

called Hölder inequality. These inequalities hold as well for ∞ values of the exponents.

PROOF. All this is very standard, the idea being as follows:

(1) As a first observation, at $p=2$, which is a special exponent, we have $q=2$ as well, and the Minkowski and Hölder inequalities are as follows:

$$\left(\int_{\mathbb{R}}|f+g|^2\right)^{1/2}\leq\left(\int_{\mathbb{R}}|f|^2\right)^{1/2}+\left(\int_{\mathbb{R}}|g|^2\right)^{1/2}$$

$$\int_{\mathbb{R}}|fg|\leq\left(\int_{\mathbb{R}}|f|^2\right)^{1/2}\left(\int_{\mathbb{R}}|g|^2\right)^{1/2}$$

But the proof of the Hölder inequality, called Cauchy-Schwarz inequality in this case, is something elementary, coming from the fact that $I(t)=\int|f+twg|^2$ with $|w|=1$ is a positive degree 2 polynomial in $t\in\mathbb{R}$, and so its discriminant must be negative. As for the Minkowski inequality, in the present $p=2$ case this follows from the Hölder, or Cauchy-Schwarz inequality, by taking squares and simplifying.

(2) In the general case now, where $p\geq 1$ is still finite, but arbitrary, the proofs are a bit more complicated, based on the Jensen inequality that we met in chapter 3. In fact, we already know from there the Minkowski and Hölder inequalities for the sequences of numbers, and the extension to functions is straightforward.

(3) Finally, with the convention that $(\int|f|^p)^{1/p}$ takes as value at $p=\infty$ the essential supremum of f, the Minkowski inequality holds as well at $p=\infty$, trivially:

$$\sup\big|f+g\big|\leq\sup\big|f\big|+\sup\big|g\big|$$

The same goes for the Hölder inequality at $p=\infty,q=1$, which is simply:

$$\int_{\mathbb{R}}|fg|\leq\sup\big|f\big|\times\int_{\mathbb{R}}|g|$$

And finally, the same goes for the Hölder inequality at $p=1,q=\infty$. □

As a consequence of the above results, we can formulate:

THEOREM 7.3. *Given an interval $I\subset\mathbb{R}$ and an exponent $p\geq 1$, the following space, with the convention that functions are identified up to equality almost everywhere,*

$$L^p(I)=\left\{f:I\to\mathbb{C}\Big|\int_I|f(x)|^p dx<\infty\right\}$$

is a vector space, and the following quantity

$$||f||_p=\left(\int_I|f(x)|^p\right)^{1/p}$$

is a norm on it, in the sense that it satisfies the usual conditions for a vector space norm. Moreover, $L^p(I)$ is complete with respect to the distance $d(f,g)=||f-g||_p$.

PROOF. This basically follows from Theorem 7.2, the idea being as follows:

(1) Again, let us first see what happens at $p = 2$. Here everything is standard from what we have in Theorem 7.2, and with the remark that the space $L^2(I)$ that we obtain is more than just a normed vector space, because we have as well a scalar product, related to the norm by the formula $||f||_2 = \sqrt{< f, f >}$, constructed as follows:

$$< f, g >= \int_I f(x)\,\overline{g(x)}\,dx$$

(2) In the general case now, where $p \geq 1$ is still finite, but arbitrary, the proof is similar, basically coming from the Minkowski inequality from Theorem 7.2.

(3) Finally, the extension at $p = \infty$ is clear too, coming from definitions, and with the various conventions made at the end of the proof of Theorem 7.2. □

There are many more things that can be said about the above spaces $L^2(I)$. Let us mention, as a key result, that we will not really need in what follows, that the Hölder inequality shows that any continuous linear form $\varphi : L^p(I) \to \mathbb{C}$ must be of the form $\varphi(f) = \int fg$, with $g \in L^q(I)$, where $q \geq 1$ is as usual given by $1/p + 1/q = 1$. Thus, if we denote by $L^p(I)^*$ the space of continuous linear forms $\varphi : L^p(I) \to \mathbb{C}$, we have:

$$L^p(I)^* = L^q(I)$$

Observe in particular that by applying twice this formula, we obtain the following remarkable property, called self-duality, or reflexivity of the Banach space $L^2(I)$:

$$L^p(I)^{**} = L^p(I)$$

For more on all this, we refer to any functional analysis book. Going ahead now with our study of functions $f : \mathbb{R} \to \mathbb{C}$, let us define an interesting operation on such functions, called convolution, which is useful for many purposes. Let us start with:

DEFINITION 7.4. *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 get back to this, later.

Goinh ahead with more theory, let us try to understand when the convolution operation is well-defined. We have here the following basic result:

THEOREM 7.5. *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 $f * g$:

$$\begin{aligned}
(f*g)(x) &= \int_{\mathbb{R}} f(x-y)g(y)dy \\
&= \int_{supp(g)} f(x-y)g(y)dy
\end{aligned}$$

But this latter quantity being 0 for $x \notin supp(f) + supp(g)$, this gives the result. □

Here are now a few remarkable properties of the convolution operation:

PROPOSITION 7.6. *The following hold, for the functions in $C_c(\mathbb{R})$:*

(1) $f * g = g * f$.
(2) $f * (g * h) = (f * g) * h$.
(3) $f * (\lambda g + \mu h) = \lambda f * g + \mu f * h$.

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

(1) This follows from the following computation, with $y = x - t$:

$$\begin{aligned}(f * g)(x) &= \int_{\mathbb{R}} f(x-y)g(y)dy \\ &= \int_{\mathbb{R}} f(t)g(x-t)dt \\ &= \int_{\mathbb{R}} g(x-t)f(t)dt \\ &= (g * f)(x)\end{aligned}$$

(2) This is clear from definitions.

(3) Once again, this is clear from definitions. □

In relation with derivatives, and with the "regularizing effect" of the convolution operation mentioned after Definition 7.4, we have the following result:

THEOREM 7.7. *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 continous functions, we have the following result, which is of particular theoretical importance:

THEOREM 7.8. *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 the following computation, involving an intuitive manipulation on the double integrals, called Fubini theorem, that we will use as such here, and that we will fully clarify later on, when talking more in detail about functions of several real variables, and their integrals:

$$\begin{aligned}\int_{\mathbb{R}}|(f*g)(x)|dx &\leq \int_{\mathbb{R}}\int_{\mathbb{R}}|f(x-y)g(y)|dydx\\ &= \int_{\mathbb{R}}\int_{\mathbb{R}}|f(x-y)g(y)|dxdy\\ &= \int_{\mathbb{R}}|f(x)|dx\int_{\mathbb{R}}|g(y)|dy\end{aligned}$$

As for the second assertion, this follows from the first one, and from Theorem 7.7. □

We should mention that it is possible to go beyond the above result, with a suitable extension of the convolution to the space $L^2(\mathbb{R})$ as well. We will be back to this.

Summarizing, we have now some good knowledge of the various spaces that the functions $f:\mathbb{R}\to\mathbb{C}$ can form, and we have as well an interesting regularization operation $f\to f*g$ on such functions, that can be used for various purposes.

7b. Fourier transform

We discuss here the construction and main properties of the Fourier transform, which is the main tool in analysis, and even in mathematics in general. We first have:

DEFINITION 7.9. *Given $f \in L^1(\mathbb{R})$, we define a function $\widehat{f}:\mathbb{R}\to\mathbb{C}$ by*

$$\widehat{f}(\xi)=\int_{\mathbb{R}}e^{ix\xi}f(x)dx$$

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 $|e^{ix\xi}|=1$. Also, the condition $f\in L^1(\mathbb{R})$ is basically needed for construcing $\widehat{f}$, because:

$$\widehat{f}(0)=\int_{\mathbb{R}}f(x)dx$$

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 7.10. *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. These results are all elementary, as follows:

(1) The additivity formula is clear from definitions, as follows:

$$\begin{aligned}\widehat{f+g}(\xi) &= \int_{\mathbb{R}} e^{ix\xi}(f+g)(x)dx\\ &= \int_{\mathbb{R}} e^{ix\xi}f(x)dx + \int_{\mathbb{R}} e^{ix\xi}f(x)dx\\ &= \widehat{f}(\xi) + \widehat{g}(\xi)\end{aligned}$$

As for the formula $\widehat{\lambda f} = \lambda \widehat{f}$, this is clear as well.

(2) The continuity of $\widehat{f}$ follows indeed from:

$$\begin{aligned}|\widehat{f}(\xi+\varepsilon) - \widehat{f}(\xi)| &\leq \int_{\mathbb{R}} \left|(e^{ix(\xi+\varepsilon)} - e^{ix\xi})f(x)\right| dx\\ &= \int_{\mathbb{R}} \left|e^{ix\xi}(e^{ix\varepsilon} - 1)f(x)\right| dx\\ &\leq |e^{ix\varepsilon} - 1| \int_{\mathbb{R}} |f|\end{aligned}$$

As for the boundedness of $\widehat{f}$, this is clear as well.

(3) This follows from the following computation, assuming that f is even:

$$\begin{aligned}\widehat{f}(-\xi) &= \int_{\mathbb{R}} e^{-ix\xi}f(x)dx\\ &= \int_{\mathbb{R}} e^{ix\xi}f(-x)dx\\ &= \int_{\mathbb{R}} e^{ix\xi}f(x)dx\\ &= \widehat{f}(\xi)\end{aligned}$$

(4) The proof here is similar to the proof of (3), by changing some signs. □

We will be back to more theory in a moment, but let us explore now the examples. Here are some basic computations of Fourier transforms:

PROPOSITION 7.11. *We have the following Fourier transform formulae,*

$$f = \chi_{[-a,a]} \implies \widehat{f}(\xi) = \frac{2\sin(a\xi)}{\xi}$$

$$f = e^{-ax}\chi_{[0,\infty]}(x) \implies \widehat{f}(\xi) = \frac{1}{a - i\xi}$$

$$f = e^{ax}\chi_{[-\infty,0]}(x) \implies \widehat{f}(\xi) = \frac{1}{a + i\xi}$$

$$f = e^{-a|x|} \implies \widehat{f}(\xi) = \frac{2a}{a^2 + \xi^2}$$

$$f = sgn(x)e^{-a|x|} \implies \widehat{f}(\xi) = \frac{2i\xi}{a^2 + \xi^2}$$

valid for any number $a > 0$.

PROOF. All this follows from some calculus, as follows:

(1) In what regards first formula, assuming $f = \chi_{[-a,a]}$, we have, by using the fact that $\sin(x\xi)$ is an odd function, whose integral vanishes on centered intervals:

$$\begin{aligned}
\widehat{f}(\xi) &= \int_{-a}^{a} e^{ix\xi}dx \\
&= \int_{-a}^{a} \cos(x\xi)dx + i\int_{-a}^{a} \sin(x\xi)dx \\
&= \int_{-a}^{a} \cos(x\xi)dx \\
&= \left[\frac{\sin(x\xi)}{\xi}\right]_{-a}^{a} \\
&= \frac{2\sin(a\xi)}{\xi}
\end{aligned}$$

(2) With $f(x)=e^{-ax}\chi_{[0,\infty]}(x)$, the computation goes as follows:

$$\begin{aligned}\widehat{f}(\xi) &= \int_0^\infty e^{ix\xi-ax}dx\\ &= \int_0^\infty e^{(i\xi-a)x}dx\\ &= \left[\frac{e^{(i\xi-a)x}}{i\xi-a}\right]_0^\infty\\ &= \frac{1}{a-i\xi}\end{aligned}$$

(3) Regarding the third formula, this follows from the second one, by using the following fact, generalizing the parity computations from Proposition 7.10:

$$F(x)=f(-x)\implies\widehat{F}(\xi)=\widehat{f}(-\xi)$$

(4) The last 2 formulae follow from what we have, by making sums and differences, and the linearity properties of the Fourier transform, from Proposition 7.10. □

We will see many other examples, in what follows. Getting back now to theory, we have the following result, adding to the various general properties in Proposition 7.10, and providing more motivations for the Fourier transform:

PROPOSITION 7.12. *Given $f,g\in L^1(\mathbb{R})$ we have $\widehat{f}g,f\widehat{g}\in L^1(\mathbb{R})$ and*

$$\int_{\mathbb{R}}f(\xi)\widehat{g}(\xi)d\xi=\int_{\mathbb{R}}\widehat{f}(x)g(x)dx$$

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:

$$\int_{\mathbb{R}}f(\xi)\widehat{g}(\xi)d\xi=\int_{\mathbb{R}}\int_{\mathbb{R}}f(\xi)e^{ix\xi}g(x)dxd\xi$$

On the other hand, we have as well the following formula:

$$\int_{\mathbb{R}}\widehat{f}(x)g(x)dx=\int_{\mathbb{R}}\int_{\mathbb{R}}e^{ix\xi}f(x)g(\xi)dxd\xi$$

Thus, with $x\leftrightarrow\xi$, we are led to the formula in the statement. □

As an important result now, showing the power of the Fourier transform, this transforms the derivative, which can be a quite complicated operation, into something very simple, namely the multiplication by the variable, the precise result being as follows:

THEOREM 7.13. *Given $f : \mathbb{R} \to \mathbb{C}$ such that $f, f' \in L^1(\mathbb{R})$, we have:*

$$\widehat{f'}(\xi) = -i\xi\widehat{f}(\xi)$$

More generally, assuming $f, f', f'', \ldots, f^{(n)} \in L^1(\mathbb{R})$, we have

$$\widehat{f^{(k)}}(\xi) = (-i\xi)^k\widehat{f}(\xi)$$

for any $k = 1, 2, \ldots, n$.

PROOF. These results follow by doing a partial integration, as follows:

(1) Assuming that $f : \mathbb{R} \to \mathbb{C}$ has compact support, we have indeed:

$$\begin{aligned}\widehat{f'}(\xi) &= \int_{\mathbb{R}} e^{ix\xi}f'(x)dx\\ &= -\int_{\mathbb{R}} i\xi e^{ix\xi}f(x)dx\\ &= -i\xi\int_{\mathbb{R}} e^{ix\xi}f(x)dx\\ &= -i\xi\widehat{f}(\xi)\end{aligned}$$

(2) Regarding the higher derivatives, the formula here follows by recurrence. □

Importantly, we have a converse statement as well, as follows:

THEOREM 7.14. *Assuming that $f \in L^1(\mathbb{R})$ is such that $F(x) = xf(x)$ belongs to $L^1(\mathbb{R})$ too, the function $\widehat{f}$ is differentiable, with derivative given by:*

$$(\widehat{f})'(\xi) = i\widehat{F}(\xi)$$

More generally, if $F_k(x) = x^k f(x)$ belongs to $L^1(\mathbb{R})$, for $k = 0, 1, \ldots, n$, we have

$$(\widehat{f})^{(k)}(\xi) = i^k\widehat{F_k}(\xi)$$

for any $k = 1, 2, \ldots, n$.

PROOF. These results are both elementary, as follows:

(1) Regarding the first assertion, the computation here is as follows:

$$\begin{aligned}(\widehat{f})'(\xi) &= \frac{d}{d\xi}\int_{\mathbb{R}} e^{ix\xi}f(x)dx\\ &= \int_{\mathbb{R}} ixe^{ix\xi}f(x)dx\\ &= i\int_{\mathbb{R}} e^{ix\xi}xf(x)dx\\ &= i\widehat{F}(\xi)\end{aligned}$$

(2) As for the second assertion, this follows from the first one, by recurrence. □

As a conclusion to all this, we are on a good way with our theory, and we have:

CONCLUSION 7.15. *Modulo normalization factors, the Fourier transform converts the derivatives into multiplications by the variable, and vice versa.*

And isn't this interesting, because isn't computing derivatives a difficult task. Here is now another useful result, of the same type, this time regarding convolutions:

THEOREM 7.16. *Assuming $f, g \in L^1(\mathbb{R})$, the following happens:*

$$\widehat{f * g} = \widehat{f} \cdot \widehat{g}$$

Moreover, under suitable assumptions, the formula $\widehat{fg} = \widehat{f} * \widehat{g}$ holds too.

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

(1) Regarding the first assertion, this is something elementary, as follows:

$$\begin{aligned}
\widehat{f * g}(\xi) &= \int_{\mathbb{R}} e^{ix\xi}(f * g)(x)dx \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} e^{ix\xi} f(x-y)g(y)dxdy \\
&= \int_{\mathbb{R}} e^{iy\xi}\left(\int e^{i(x-y)\xi} f(x-y)dx\right) g(y)dy \\
&= \int_{\mathbb{R}} e^{iy\xi}\left(\int e^{it\xi} f(t)dt\right) g(y)dy \\
&= \int_{\mathbb{R}} e^{iy\xi}\widehat{f}(\xi)g(y)dy \\
&= \widehat{f}(\xi)\widehat{g}(\xi)
\end{aligned}$$

(2) As for the second assertion, this is something more tricky, and we will be back to this later. In the meantime, here is however some sort of proof, not very honest:

$$\begin{aligned}
(\widehat{f} * \widehat{g})(\xi) &= \int_{\mathbb{R}} \widehat{f}(\xi-\eta)\widehat{g}(\eta)d\eta \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{ix(\xi-\eta)} f(x)e^{iy\eta}g(y)dxdyd\eta \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{ix\eta}e^{i(y-x)\eta} f(x)g(y)dxdyd\eta \\
&= \int_{\mathbb{R}} e^{ix\eta} f(x)g(x)dx \\
&= \widehat{fg}(\eta)
\end{aligned}$$

To be more precise, the point here is that we can pass from the triple to the single integral by arguing that "we must have $x = y$". We will be back to this later. □

As an updated conclusion to all this, we have, modulo a few bugs, to be fixed:

CONCLUSION 7.17. *The Fourier transform converts the derivatives into multiplications by the variable, and convolutions into products, and vice versa.*

We will see applications of this later, after developing some more general theory. So, let us develop now more theory for the Fourier transform. We first have:

THEOREM 7.18. *Given $f \in L^1(\mathbb{R})$, its Fourier transform satisfies*

$$\lim_{\xi\to\pm\infty} \widehat{f}(\xi) = 0$$

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 $y \in \mathbb{R}$, let us set:

$$f_y(x) = f(x - y)$$

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:

$$\mathbb{R} \to L^p(\mathbb{R}) \quad , \quad y \to f_y$$

(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:

$$|s - t| < \delta \implies |g(s) - g(t)| < (3K)^{-1/p}\varepsilon$$

But this shows that we have the following estimate:

$$\begin{aligned} ||g_s - g_t||_p &= \left(\int_{\mathbb{R}} \left|g(x - s) - g(x - t)\right|^p dx\right)^{1/p} \\ &< \left[(3K)^{-1}\varepsilon^p(2k + \delta)\right]^{1/p} \\ &< \varepsilon \end{aligned}$$

By using now the formula $||f||_p = ||f_s||_p$, which is clear, we obtain:

$$\begin{aligned} ||f_s - f_t||_p &\leq ||f_s - g_s||_p + ||g_s - g_t||_p + ||g_t - f_t||_p \\ &< \varepsilon + \varepsilon + \varepsilon \\ &= 3\varepsilon \end{aligned}$$

But this being true for any $|s - t| < \delta$, 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 $x\to x-\pi/\xi$, we have:

$$\begin{aligned}\widehat{f}(\xi)&=\int_{\mathbb{R}}e^{ix\xi}f(x)dx\\&=-\int_{\mathbb{R}}e^{ix\xi}e^{\pi i}f(x)dx\\&=-\int_{\mathbb{R}}e^{i\xi(x+\pi/\xi)}f(x)dx\\&=-\int_{\mathbb{R}}e^{ix\xi}f\left(x-\frac{\pi}{\xi}\right)dx\end{aligned}$$

On the other hand, we have as well the following formula:

$$\widehat{f}(\xi)=\int_{\mathbb{R}}e^{ix\xi}f(x)dx$$

Thus by summing, we obtain the following formula:

$$2\widehat{f}(\xi)=\int_{\mathbb{R}}e^{ix\xi}\left(f(x)-f\left(x-\frac{\pi}{\xi}\right)\right)dx$$

But this gives the following estimate, with notations from (1):

$$2|\widehat{f}(\xi)|\leq||f-f_{\pi/\xi}||_1$$

Since by (1) this goes to 0 with $\xi\to\pm\infty$, 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 7.19. *Assuming $f,\widehat{f}\in L^1(\mathbb{R})$, we have*

$$f(x)=\int_{\mathbb{R}}e^{-ix\xi}\widehat{f}(\xi)d\xi$$

almost everywhere, called Fourier inversion formula.

PROOF. This is something quite tricky, due to the fact that a direct attempt by double integration fails. Consider the following function, depending on a parameter $\lambda>0$:

$$\varphi_\lambda(x)=\int_{\mathbb{R}}e^{-ix\xi-\lambda|\xi|}d\xi$$

We have then the following computation:

$$\begin{aligned}(f*\varphi_\lambda)(x) &= \int_{\mathbb{R}} f(x-y)\varphi_\lambda(y)dy\\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} f(x-y)e^{-iy\xi-\lambda|\xi|}d\xi dy\\ &= \int_{\mathbb{R}} e^{-\lambda|\xi|}\left(\int_{\mathbb{R}} f(x-y)e^{-iy\xi}dy\right)d\xi\\ &= \int_{\mathbb{R}} e^{-\lambda|\xi|}e^{-ix\xi}\widehat{f}(\xi)d\xi\end{aligned}$$

By letting now $\lambda\to 0$, we obtain from this the following formula:

$$\lim_{\lambda\to 0}(f*\varphi_\lambda)(x)=\int_{\mathbb{R}} e^{-ix\xi}\widehat{f}(\xi)d\xi$$

On the other hand, by using Theorem 7.18 we obtain that, almost everywhere:

$$\lim_{\lambda\to 0}(f*\varphi_\lambda)(x)=f(x)$$

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

There are many more things that can be said about Fourier transforms, a key result being the Plancherel formula, allowing us to talk about the Fourier transform over the space $L^2(\mathbb{R})$. Also, we can talk about the Fourier transform over the space $\mathcal{S}$ of functions all whose derivatives are rapidly decreasing, called Schwartz space. It is beyond our scope here to discuss all this, and we recommend instead any good book on the subject.

7c. Groups, extensions

All the above theory concerns the Fourier transform over $\mathbb{R}$, but there are other types of Fourier transforms too, the idea being that we have such a transform over any locally compact abelian group G. In order to understand this, let us start with:

DEFINITION 7.20. *An abelian group is a set G with a multiplication operation*

$$(g,h)\to gh$$

which must satisfy the following conditions:

(1) *Commutativity: we have $gh=hg$, for any $g,h\in G$.*
(2) *Associativity: we have $(gh)k=g(hk)$, for any $g,h,k\in G$.*
(3) *Unit: there is an element $1\in G$ such that $g1=g$, for any $g\in G$.*
(4) *Inverses: for any $g\in G$ there is $g^{-1}\in G$ such that $gg^{-1}=1$.*

There are many examples of abelian groups, and in order to understand what is going on, and to develop our Fourier transform theory, let us first restrict the attention to the

finite group case. Here we have as basic examples the cyclic group $\mathbb{Z}_N$, or more generally the products of such cyclic groups, which are clearly both finite and abelian:

$$G = \mathbb{Z}_{N_1} \times \ldots \times \mathbb{Z}_{N_k}$$

Our goal in what follows will be that of proving that these are in fact all the finite abelian groups, and also that a Fourier transform theory, called "discrete Fourier transform", can be developed over such groups. These two questions are in fact related, and in order to investigate them, we will need the following notion:

PROPOSITION 7.21. *Given a finite abelian group G, the group morphisms*

$$\chi : G \to \mathbb{T}$$

with $\mathbb{T}$ being the unit circle, called characters of G, form a finite abelian group $\widehat{G}$.

PROOF. There are several things to be proved here, the idea being as follows:

(1) Our first claim is that $\widehat{G}$ is a group, with the pointwise multiplication, namely:

$$(\chi\rho)(g) = \chi(g)\rho(g)$$

Indeed, if χ, ρ are characters, so is $\chi\rho$, and so the multiplication is well-defined on $\widehat{G}$. Regarding the unit, this is the trivial character, constructed as follows:

$$1 : G \to \mathbb{T} \quad , \quad g \to 1$$

Finally, we have inverses, with the inverse of $\chi : G \to \mathbb{T}$ being its conjugate:

$$\bar{\chi} : G \to \mathbb{T} \quad , \quad g \to \overline{\chi(g)}$$

(2) Our next claim is that $\widehat{G}$ is finite. Indeed, given a group element $g \in G$, we can talk about its order, which is smallest integer $k \in \mathbb{N}$ such that $g^k = 1$. Now assuming that we have a character $\chi : G \to \mathbb{T}$, we have the following formula:

$$\chi(g)^k = 1$$

Thus $\chi(g)$ must be one of the k-th roots of unity, and in particular there are finitely many choices for $\chi(g)$. Thus, there are finitely many choices for χ, as desired.

(3) Finally, the fact that $\widehat{G}$ is abelian follows from definitions, because the pointwise multiplication of functions, and in particular of characters, is commutative. □

The above construction is quite interesting, and we have:

THEOREM 7.22. *The character group operation $G \to \widehat{G}$ for the finite abelian groups, called Pontrjagin duality, has the following properties:*

(1) *The dual of a cyclic group is the group itself, $\widehat{\mathbb{Z}}_N = \mathbb{Z}_N$.*
(2) *The dual of a product is the product of duals, $\widehat{G \times H} = \widehat{G} \times \widehat{H}$.*
(3) *Any product of cyclic groups $G = \mathbb{Z}_{N_1} \times \ldots \times \mathbb{Z}_{N_k}$ is self-dual, $G = \widehat{G}$.*

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

(1) A character $\chi : \mathbb{Z}_N \to \mathbb{T}$ is uniquely determined by its value $z = \chi(g)$ on the standard generator $g \in \mathbb{Z}_N$. But this value must satisfy:

$$z^N = 1$$

Thus we must have $z \in \mathbb{Z}_N$, with the cyclic group $\mathbb{Z}_N$ being regarded this time as being the group of N-th roots of unity. Now conversely, any N-th root of unity $z \in \mathbb{Z}_N$ defines a character $\chi : \mathbb{Z}_N \to \mathbb{T}$, by setting, for any $r \in \mathbb{N}$:

$$\chi(g^r) = z^r$$

Thus we have an identification $\widehat{\mathbb{Z}}_N = \mathbb{Z}_N$, as claimed.

(2) A character of a product of groups $\chi : G \times H \to \mathbb{T}$ must satisfy:

$$\chi(g, h) = \chi\left[(g, 1)(1, h)\right] = \chi(g, 1)\chi(1, h)$$

Thus χ must appear as the product of its restrictions $\chi_{|G}, \chi_{|H}$, which must be both characters, and this gives the identification in the statement.

(3) This follows from (1) and (2). Alternatively, any character $\chi : G \to \mathbb{T}$ is uniquely determined by its values $\chi(g_1), \ldots, \chi(g_k)$ on the standard generators of $\mathbb{Z}_{N_1}, \ldots, \mathbb{Z}_{N_k}$, which must belong to $\mathbb{Z}_{N_1}, \ldots, \mathbb{Z}_{N_k} \subset \mathbb{T}$, and this gives $\widehat{G} = G$, as claimed. □

At a more advanced level now, we have the following result:

THEOREM 7.23. *The finite abelian groups are the following groups,*

$$G = \mathbb{Z}_{N_1} \times \ldots \times \mathbb{Z}_{N_k}$$

and these groups are all self-dual, $G = \widehat{G}$.

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

(1) In order to prove our result, assume that G is finite and abelian. For any prime number $p \in \mathbb{N}$, let us define $G_p \subset G$ to be the subset of elements having as order a power of p. Equivalently, this subset $G_p \subset G$ can be defined as follows:

$$G_p = \left\{ g \in G \middle| \exists k \in \mathbb{N}, g^{p^k} = 1 \right\}$$

(2) It is then routine to check, based on definitions, that each G_p is a subgroup. Our claim now is that we have a direct product decomposition as follows:

$$G = \prod_p G_p$$

(3) Indeed, by using the fact that our group G is abelian, we have a morphism as follows, with the order of the factors when computing $\prod_p g_p$ being irrelevant:

$$\prod_p G_p \to G \quad , \quad (g_p) \to \prod_p g_p$$

Moreover, it is routine to check that this morphism is both injective and surjective, via some simple manipulations, so we have our group decomposition, as in (2).

(4) Thus, we are left with proving that each component G_p decomposes as a product of cyclic groups, having as orders powers of p, as follows:

$$G_p = \mathbb{Z}_{p^{r_1}} \times \ldots \times \mathbb{Z}_{p^{r_s}}$$

But this is something that can be checked by recurrence on $|G_p|$, via some routine computations, and we are led to the conclusion in the statement.

(5) Finally, the fact that the finite abelian groups are self-dual, $G = \widehat{G}$, follows from the structure result that we just proved, and from Theorem 7.22 (3). □

In relation now with Fourier analysis, the result is as follows:

THEOREM 7.24. *Given a finite abelian group G, we have an isomorphism as follows, obtained by linearizing/delinearizing the characters,*

$$C^*(G) \simeq C(\widehat{G})$$

where $C^*(G)$ is the algebra of functions $\varphi : G \to \mathbb{C}$, with convolution product, and $C(\widehat{G})$ is the algebra of functions $\varphi : \widehat{G} \to \mathbb{C}$, with usual product.

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

(1) Given a finite abelian group G, we can talk about the complex algebra $C(G)$ formed by the complex functions $\varphi : G \to \mathbb{C}$, with usual product, namely:

$$(\varphi\psi)(g) = \varphi(g)\psi(g)$$

Observe that we have $C(G) \simeq \mathbb{C}^N$ as an algebra, where $N = |G|$, with this being best seen via the basis of $C(G)$ formed by the Dirac masses at the points of G:

$$C(G) = \left\{ \sum_{g \in G} \lambda_g \delta_g \Big| \lambda_g \in \mathbb{C} \right\}$$

(2) On the other hand, we can talk as well about the algebra $C^*(G)$ formed by the same functions $\varphi : G \to \mathbb{C}$, but this time with the convolution product, namely:

$$(\varphi * \psi)(g) = \sum_{h \in G} \varphi(gh^{-1})\psi(h)$$

Since we have $\delta_k * \delta_l = \delta_{kl}$ for any $k, l \in G$, as you can easily check by using the above formula, the Dirac masses $\delta_g \in C^*(G)$ behave like the group elements $g \in G$. Thus, we can view our algebra as follows, with multiplication given by $g \cdot h = gh$, and linearity:

$$C^*(G) = \left\{ \sum_{g \in G} \lambda_g g \Big| \lambda_g \in \mathbb{C} \right\}$$

(3) Now that we know what the statement is about, let us go for the proof. The first observation is that we have a morphism of algebras as follows:

$$C^*(G) \to C(\widehat{G}) \quad , \quad g \to [\chi \to \chi(g)]$$

Now since on both sides we have vector spaces of dimension $N = |G|$, it is enough to check that this morphism is injective. But this is best done via Theorem 7.23, which shows that the characters $\chi \in \widehat{G}$ separate the points $g \in G$, as desired. □

We can feel that Theorem 7.24 is related to Fourier analysis, and we have:

FACT 7.25. *The following happen, regarding the locally compact abelian groups:*

(1) *What we did in the finite case, namely group characters, and construction and basic properties of the dual, can be extended to them.*
(2) *As basic examples of this, besides what we have in the finite case, and notably $\widehat{\mathbb{Z}_N} = \mathbb{Z}_N$, we have $\widehat{\mathbb{Z}} = \mathbb{T}$, $\widehat{\mathbb{T}} = \mathbb{Z}$, and also $\widehat{\mathbb{R}} = \mathbb{R}$.*
(3) *With some care for analytic aspects, $C^*(G) \simeq C(\widehat{G})$ remains true in this setting, and in the case $G = \mathbb{R}$, this isomorphism is the Fourier transform.*

Obviously, all this is a bit heavy, but you get the point, we have 3 types of Fourier analysis in life, namely the "standard" one that we previously learned in this chapter, corresponding to $G = \mathbb{R}$, then another one that we skipped, and that we encourage you to learn, called the "Fourier series" one, corresponding to $G = \mathbb{Z}, \mathbb{T}$, and finally the "discrete" one that we started to learn, over $G = \mathbb{Z}_N$ and other finite abelian groups.

In practice, all this is a bit complicated, and back now to the finite abelian groups, let us work out a softer version of all the above, which is what is really needed, in practice, when doing discrete Fourier analysis. For $G = \mathbb{Z}_N$, what we need is:

DEFINITION 7.26. *The Fourier matrix F_N is the following matrix, with $w = e^{2\pi i/N}$:*

$$F_N = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & w & w^2 & \ldots & w^{N-1} \\ 1 & w^2 & w^4 & \ldots & w^{2(N-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w^{N-1} & w^{2(N-1)} & \ldots & w^{(N-1)^2} \end{pmatrix}$$

That is, $F_N = (w^{ij})_{ij}$, with indices $i, j \in \{0, 1, \ldots, N-1\}$, taken modulo N.

Observe that this matrix is Hadamard, in the sense that its entries are on the unit circle, and the rows are pairwise orthogonal. In fact, in general, we have:

THEOREM 7.27. *Given a finite abelian group G, with dual group $\widehat{G} = \{\chi : G \to \mathbb{T}\}$, consider the corresponding Fourier coupling, namely:*

$$\mathcal{F}_G : G \times \widehat{G} \to \mathbb{T} \quad , \quad (i, \chi) \to \chi(i)$$

(1) *Via the standard isomorphism $G \simeq \widehat{G}$, this Fourier coupling can be regarded as a square matrix, $F_G \in M_G(\mathbb{T})$, which is a complex Hadamard matrix.*
(2) *In the case of the cyclic group $G = \mathbb{Z}_N$ we obtain in this way, via the standard identification $\mathbb{Z}_N = \{1, \ldots, N\}$, the Fourier matrix F_N.*
(3) *In general, when using a decomposition $G = \mathbb{Z}_{N_1} \times \ldots \times \mathbb{Z}_{N_k}$, the corresponding Fourier matrix is given by $F_G = F_{N_1} \otimes \ldots \otimes F_{N_k}$.*

PROOF. This follows indeed by using the above finite abelian group theory:

(1) With the identification $G \simeq \widehat{G}$ made our matrix is given by $(F_G)_{i\chi} = \chi(i)$, and the scalar products between the rows are computed as follows:

$$< R_i, R_j >= \sum_\chi \chi(i)\overline{\chi(j)} = \sum_\chi \chi(i-j) = |G| \cdot \delta_{ij}$$

Thus, we obtain indeed a complex Hadamard matrix.

(2) This follows from the well-known and elementary fact that, via the identifications $\mathbb{Z}_N = \widehat{\mathbb{Z}_N} = \{1, \ldots, N\}$, the Fourier coupling here is as follows, with $w = e^{2\pi i/N}$:

$$(i, j) \to w^{ij}$$

(3) We use here the following formula that we know, for the duals of products:

$$\widehat{H \times K} = \widehat{H} \times \widehat{K}$$

At the level of the corresponding Fourier couplings, we obtain from this:

$$F_{H \times K} = F_H \otimes F_K$$

Now by decomposing G into cyclic groups, as in the statement, and by using (2) for the cyclic components, we obtain the formula in the statement. □

As a nice application of discrete Fourier analysis, we have:

THEOREM 7.28. *For a matrix $M \in M_N(\mathbb{C})$, the following are equivalent:*

(1) *M is circulant, $M_{ij} = \xi_{j-i}$, for a certain vector $\xi \in \mathbb{C}^N$.*
(2) *M is Fourier-diagonal, $M = F_N Q F_N^*$, for a certain diagonal matrix Q.*

Moreover, if these conditions hold, then $\xi = F_N^* q$, where $q = (Q_{11}, \ldots, Q_{NN})$.

PROOF. This follows from some basic computations with roots of unity, as follows:

(1) $\implies$ (2) Assuming $M_{ij} = \xi_{j-i}$, the matrix $Q = F_N^* M F_N$ is indeed diagonal, as shown by the following computation:

$$\begin{aligned} Q_{ij} &= \sum_{kl} w^{-ik} M_{kl} w^{lj} \\ &= \sum_{kl} w^{jl-ik} \xi_{l-k} \\ &= \sum_{kr} w^{j(k+r)-ik} \xi_r \\ &= \sum_r w^{jr} \xi_r \sum_k w^{(j-i)k} \\ &= N\delta_{ij} \sum_r w^{jr} \xi_r \end{aligned}$$

(2) $\implies$ (1) Assuming $Q = diag(q_1, \ldots, q_N)$, the matrix $M = F_N Q F_N^*$ is indeed circulant, as shown by the following computation:

$$M_{ij} = \sum_k w^{ik} Q_{kk} w^{-jk} = \sum_k w^{(i-j)k} q_k$$

To be more precise, in this formula the last term depends only on $j - i$, and so shows that we have $M_{ij} = \xi_{j-i}$, with ξ being the following vector:

$$\xi_i = \sum_k w^{-ik} q_k = (F_N^* q)_i$$

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

As a basic illustration for the above result, the all-one matrix diagonalizes as follows:

$$\begin{pmatrix} 1 & \ldots & \ldots & 1 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 1 & \ldots & \ldots & 1 \end{pmatrix} = \frac{1}{N} F_N \begin{pmatrix} N & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} F_N^*$$

But you might know this already, since this was a recurrent exercise, in this book.

7d. Independence, limits

With the above theory in hand, let us go back to probability. Our claim is that things are very interesting here, with the real-life notion of independence corresponding to our mathematical notion of convolution, and with the Fourier transform being something very useful in probability, in order to understand the notion of independence.

In order to discuss all this, let us start with the following definition:

DEFINITION 7.29. *Two variables $f, g \in L^\infty(X)$ are called independent when*

$$\mathbb{E}(f^k g^l) = \mathbb{E}(f^k)\,\mathbb{E}(g^l)$$

happens, for any $k, l \in \mathbb{N}$.

This definition hides some non-trivial things. Indeed, by linearity, we would like to have a formula as follows, valid for any polynomials $P, Q \in \mathbb{R}[X]$:

$$\mathbb{E}[P(f)Q(g)] = \mathbb{E}[P(f)]\,\mathbb{E}[Q(g)]$$

By a continuity argument, it is enough to have this formula for characteristic functions χ_I, χ_J of the arbitrary measurable sets of real numbers $I, J \subset \mathbb{R}$:

$$\mathbb{E}[\chi_I(f)\chi_J(g)] = \mathbb{E}[\chi_I(f)]\,\mathbb{E}[\chi_J(g)]$$

Thus, we are led to the usual definition of independence, namely:

$$\mathbb{P}(f \in I, g \in J) = \mathbb{P}(f \in I)\,\mathbb{P}(g \in J)$$

All this might seem a bit abstract, but in practice, the idea is of course that f, g must be independent, in an intuitive, real-life sense. As a first result now, we have:

THEOREM 7.30. *Assuming that $f, g \in L^\infty(X)$ are independent, we have*

$$\mu_{f+g} = \mu_f * \mu_g$$

where $*$ is the convolution of real probability measures.

PROOF. We have the following computation, using the independence of f, g:

$$\begin{aligned} M_k(f+g) &= \mathbb{E}((f+g)^k) \\ &= \sum_r \binom{k}{r} \mathbb{E}(f^r g^{k-r}) \\ &= \sum_r \binom{k}{r} M_r(f) M_{k-r}(g) \end{aligned}$$

On the other hand, by using the Fubini theorem, we have as well:

$$\begin{aligned} \int_{\mathbb{R}} x^k d(\mu_f * \mu_g)(x) &= \int_{\mathbb{R}\times\mathbb{R}} (x+y)^k d\mu_f(x) d\mu_g(y) \\ &= \sum_r \binom{k}{r} \int_{\mathbb{R}} x^r d\mu_f(x) \int_{\mathbb{R}} y^{k-r} d\mu_g(y) \\ &= \sum_r \binom{k}{r} M_r(f) M_{k-r}(g) \end{aligned}$$

Thus μ_{f+g} and $\mu_f * \mu_g$ have the same moments, so they coincide, as desired. □

Here is now a second result on independence, which is something more advanced:

THEOREM 7.31. *Assuming that $f, g \in L^\infty(X)$ are independent, we have*

$$F_{f+g} = F_f F_g$$

where $F_f(x) = \mathbb{E}(e^{ixf})$ is the Fourier transform.

PROOF. We have the following computation, using Theorem 7.30 and Fubini:

$$\begin{aligned} F_{f+g}(x) &= \int_{\mathbb{R}} e^{ixz} d\mu_{f+g}(z) \\ &= \int_{\mathbb{R}} e^{ixz} d(\mu_f * \mu_g)(z) \\ &= \int_{\mathbb{R}\times\mathbb{R}} e^{ix(z+t)} d\mu_f(z) d\mu_g(t) \\ &= \int_{\mathbb{R}} e^{ixz} d\mu_f(z) \int_{\mathbb{R}} e^{ixt} d\mu_g(t) \\ &= F_f(x) F_g(x) \end{aligned}$$

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

In order to work out some basic illustrations, let us go back to the Poisson laws p_t, from chapter 4. We can now say more about them, first with the following result:

THEOREM 7.32. *We have the following formula, for any $s, t > 0$,*

$$p_s * p_t = p_{s+t}$$

so the Poisson laws form a convolution semigroup.

PROOF. By using $\delta_k * \delta_l = \delta_{k+l}$ and the binomial formula, we obtain:

$$\begin{aligned} p_s * p_t &= e^{-s} \sum_k \frac{s^k}{k!} \delta_k * e^{-t} \sum_l \frac{t^l}{l!} \delta_l \\ &= e^{-s-t} \sum_n \delta_n \sum_{k+l=n} \frac{s^k t^l}{k!l!} \\ &= e^{-s-t} \sum_n \frac{\delta_n}{n!} \sum_{k+l=n} \frac{n!}{k!l!} s^k t^l \\ &= e^{-s-t} \sum_n \frac{(s+t)^n}{n!} \delta_n \\ &= p_{s+t} \end{aligned}$$

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

Next in line, we have the following result, which is fundamental as well:

THEOREM 7.33. *The Poisson laws appear as formal exponentials*

$$p_t = \sum_k \frac{t^k(\delta_1 - \delta_0)^{*k}}{k!}$$

with respect to the convolution of measures $*$.

PROOF. By using the binomial formula, the measure on the right is:

$$\begin{aligned}
\mu &= \sum_k \frac{t^k}{k!} \sum_{r+s=k} (-1)^s \frac{k!}{r!s!} \delta_r \\
&= \sum_k t^k \sum_{r+s=k} (-1)^s \frac{\delta_r}{r!s!} \\
&= \sum_r \frac{t^r \delta_r}{r!} \sum_s \frac{(-1)^s}{s!} \\
&= \frac{1}{e} \sum_r \frac{t^r \delta_r}{r!} \\
&= p_t
\end{aligned}$$

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

Regarding now the Fourier transform computation, this is as follows:

THEOREM 7.34. *The Fourier transform of* p_t *is given by*

$$F_{p_t}(y) = \exp\left((e^{iy} - 1)t\right)$$

for any $t > 0$.

PROOF. We have indeed the following computation:

$$\begin{aligned}
F_{p_t}(y) &= e^{-t} \sum_k \frac{t^k}{k!} F_{\delta_k}(y) \\
&= e^{-t} \sum_k \frac{t^k}{k!} e^{iky} \\
&= e^{-t} \sum_k \frac{(e^{iy}t)^k}{k!} \\
&= \exp(-t) \exp(e^{iy}t) \\
&= \exp\left((e^{iy} - 1)t\right)
\end{aligned}$$

Thus, we obtain the formula in the statement. □

Observe that the above formula gives an alternative proof for Theorem 7.32, by the using the fact that the logarithm of the Fourier transform linearizes the convolution. As another application, we can now establish the Poisson Limit Theorem, as follows:

THEOREM 7.35 (PLT). *We have the following convergence, in moments,*

$$\left(\left(1-\frac{t}{n}\right)\delta_0+\frac{t}{n}\delta_1\right)^{*n}\to p_t$$

for any $t>0$.

PROOF. Let us denote by ν_n the measure under the convolution sign, namely:

$$\nu_n=\left(1-\frac{t}{n}\right)\delta_0+\frac{t}{n}\delta_1$$

We have the following computation, for the Fourier transform of the limit:

$$\begin{aligned}
F_{\delta_r}(y)=e^{iry} &\implies F_{\nu_n}(y)=\left(1-\frac{t}{n}\right)+\frac{t}{n}e^{iy}\\
&\implies F_{\nu_n^{*n}}(y)=\left(\left(1-\frac{t}{n}\right)+\frac{t}{n}e^{iy}\right)^n\\
&\implies F_{\nu_n^{*n}}(y)=\left(1+\frac{(e^{iy}-1)t}{n}\right)^n\\
&\implies F(y)=\exp\left((e^{iy}-1)t\right)
\end{aligned}$$

Thus, we obtain indeed the Fourier transform of p_t, as desired. □

We have accumulated so far a lot of probability knowledge, going along with our learning of calculus, in chapter 4, chapter 6, and in the present chapter. All this certainly needs some systematic discussion, and we will be back to it once we will have the full calculus tools that are needed, in chapter 14, which will be dedicated to probability.

7e. Exercises

This was a dense, introductory chapter to Fourier analysis, with many technical things missing, going beyond our scope here, and here are some exercises on all this:

EXERCISE 7.36. *Clarify the proof of $\widehat{fg}=\widehat{f}*\widehat{g}$, say via Fourier inversion.*

EXERCISE 7.37. *Learn about the Plancherel formula, and Fourier over $L^2(\mathbb{R})$.*

EXERCISE 7.38. *Learn about the Schwartz space $\mathcal{S}$, and Fourier over it.*

EXERCISE 7.39. *Further clarify the Fourier transforms over $\mathbb{Z}_N,\mathbb{Z},\mathbb{T},\mathbb{R}$.*

As bonus exercise, learn some probability. Indeed, as mentioned above, we already know some, but we will take a long break from this, until chapter 14 below.

CHAPTER 8

Harmonic functions

8a. Laplace operator

Welcome to theoretical physics. We will discuss in this chapter some applications of the theory of complex functions developed so far, to some questions from physics, along with some more theory, needed in order to reach to these applications. We will be mainly interested in two fundamental equations of physics, namely the wave equation, and the heat equation. We will have a look as well into basic quantum mechanics.

All this sounds very exciting, doesn't it. However, in practice, technically speaking, we are not exactly ready yet for all these things. But the temptation remains high to talk about them, a bit like physicists do. So, what to do. Ask the cat of course, who says:

CAT 8.1. *The more you know, the better that is.*

Thanks cat. So, we will talk indeed about all this, a bit like physicists do, by taking some freedom with surfing on difficult mathematics. In short, this will be a physics chapter, and for a more rigorous treatement of all this, which will be more general too, do not worry, that will come, later in this book, once we will know more calculus.

So, let us start, with some mathematics. We have:

DEFINITION 8.2. *A function $f : X \to \mathbb{C}$ with $X \subset \mathbb{C}$ is called differentiable in the real sense if the following limits, called its partial derivatives,*

$$\frac{df}{dx}(z) = \lim_{t\to 0} \frac{f(z+t) - f(z)}{t}$$

$$\frac{df}{dy}(z) = \lim_{t\to 0} \frac{f(z+it) - f(z)}{t}$$

computed by using $t \in \mathbb{R}$, $t \to 0$, exist at any $z \in X$, and are continuous.

We have already met this notion, in chapter 6, in a purely mathematical context, when taking about holomorphic functions. Our observation there was that a holomorphic function is differentiable in the above real sense, and also that there are further functions, such as the conjugate $f(z) = \bar{z}$ or the modulus $f(z) = |z|$, which are still differentiable in the above real sense, but are not holomorphic. But more on this later.

We will need as well the following definition, complementing Definition 8.2:

DEFINITION 8.3. *A function $f : X \to \mathbb{C}$ is called doubly differentiable if its partial derivatives are both differentiable. These double derivatives are denoted as follows:*

$$\frac{d^2f}{dx^2} = \frac{d}{dx}\left(\frac{df}{dx}\right) \quad , \quad \frac{d^2f}{dxdy} = \frac{d}{dx}\left(\frac{df}{dy}\right)$$

$$\frac{d^2f}{dydx} = \frac{d}{dy}\left(\frac{df}{dx}\right) \quad , \quad \frac{d^2f}{dy^2} = \frac{d}{dy}\left(\frac{df}{dy}\right)$$

We can in fact talk, in a similar way, about order k partial derivatives, for any $k \in \mathbb{N}$.

As before with Definition 8.2, there are many things that can be said here, mathematically speaking, but we will leave this discussion for later, starting with chapter 9 below, when systematically investigating functions of several variables. In fact, the material in this chapter will be, among others, an introduction to that. However, we will need:

THEOREM 8.4. *The double derivatives satisfy the formula*

$$\frac{d^2f}{dxdy} = \frac{d^2f}{dydx}$$

called Clairaut formula.

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

(1) Before pulling out a formal proof, as an intuitive justification for our formula, let us consider a product of power functions, $f(z) = x^py^q$. We have then:

$$\frac{d^2f}{dxdy} = \frac{d}{dx}\left(\frac{dx^py^q}{dy}\right) = \frac{d}{dx}\left(qx^py^{q-1}\right) = pqx^{p-1}y^{q-1}$$

$$\frac{d^2f}{dydx} = \frac{d}{dy}\left(\frac{dx^py^q}{dx}\right) = \frac{d}{dy}\left(px^{p-1}y^q\right) = pqx^{p-1}y^{q-1}$$

Next, let us consider a linear combination of power functions, $f(z) = \sum_{pq} c_{pq}x^py^q$, which can be finite or not. We have then, by using the above computation:

$$\frac{d^2f}{dxdy} = \frac{d^2f}{dydx} = \sum_{pq} c_{pq}pqx^{p-1}y^{q-1}$$

Thus, we can see that our commutation formula for derivatives holds indeed, and this due to the fact that the functions in x, and in y, commute. Of course, all this does not prove our formula, in general. But exercise for you, to have this idea fully working.

(2) Getting now to more standard techniques, given a point in the complex plane, $z = a + ib$, 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 $z=a+ib$, as desired. □

Regarding the double derivatives in one of the variables, here our intuition from the one-variable case can help, and in fact there is no need for more, at least of this stage of the things. But, as already mentioned, we will be back to this, later in this book.

Getting now into physics, which will be our starting point for the considerations in this chapter, let us start with the waves. Their behavior is described by the wave equation, found via experiments, and nothing can of course replace these experiments. However, we can come upon this equation in a purely mathematical way, as follows:

THEOREM 8.5. *The wave equation in the plane $\mathbb{R}^2$ is*

$$\ddot{\varphi}=v^2\Delta\varphi$$

where dots denote time derivatives, Δ is the Laplace operator, given by

$$\Delta\varphi=\frac{d^2\varphi}{dx^2}+\frac{d^2\varphi}{dy^2}$$

and $v>0$ is the propagation speed.

PROOF. We have already met this equation in chapter 3, in one dimension. In what follows we will briefly recall the proof from there, then discuss the extension to 2 dimensions. The idea is that we can "prove" this equation, by discretizing, as follows:

(1) Let us first consider the 1D case. In order to understand the propagation of waves, we will model $\mathbb{R}$ as a network of balls, with springs between them, as follows:

$$\cdots \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \cdots$$

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:

$$\cdots\cdots \bullet_{\varphi(x-l)} \times\!\!\times\!\!\times \bullet_{\varphi(x)} \times\!\!\times\!\!\times \bullet_{\varphi(x+l)} \cdots\cdots$$

We have two forces acting at x, namely the Newton motion force, mass m times acceleration, and the Hooke force, displacement of the spring, times spring constant k. 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 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:

$$m = \frac{M}{N} \quad , \quad k = KN \quad , \quad l = \frac{L}{N}$$

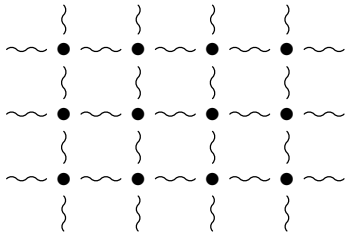
With these changes, our equation of motion found in (1) reads:

$$\ddot{\varphi}(x) = \frac{KN^2}{M}(\varphi(x+l) - 2\varphi(x) + \varphi(x-l))$$

With $N \to \infty$, and therefore $l \to 0$, we obtain in this way:

$$\ddot{\varphi}(x) = \frac{KL^2}{M} \cdot \frac{d^2\varphi}{dx^2}(x)$$

(3) In 2 dimensions now, the same argument carries on. Indeed, we can use here a lattice model as follows, with all the edges standing for small springs:



As before by denoting by $\varphi(x,y)$ the position function, in the limit, we are led to the following equation, with $v=\sqrt{K/M}\cdot L$ being the propagation speed:

$$\ddot{\varphi}(z)=v^2\left(\frac{d^2\varphi}{dx^2}+\frac{d^2\varphi}{dy^2}\right)(z)$$

But we recognize at right the Laplace operator, and we are done. There is of course some more discussion to be made here, arguing that our spring model that we used is indeed the correct one. But do not worry, experiments confirm our findings.

(4) Finally, for completness, let us mention that the same argument, namely a lattice model, carries on in arbitrary N dimensions, and we obtain here the same wave equation as before, namely $\ddot{\varphi}=v^2\Delta\varphi$, with the following straightforward definition for the Laplacian, based on the obvious N-dimensional analogues of Definition 8.2 and Definition 8.3:

$$\Delta\varphi=\sum_{i=1}^{N}\frac{d^2\varphi}{dx_i^2}$$

This is quite nice, because at $N=1$ this Laplace operator is just the second derivative, so what we have here is a unification of what we did in (1,2) and in (3) above. In addition, at $N=3$ we obtain in this way the wave equation in the case of main interest, namely the one of our real-life world. But more on this later in this book. □

Regarding now heat diffusion, we have here a similar equation, as follows:

THEOREM 8.6. *Heat diffusion in $\mathbb{R}^2$ is described by the heat equation*

$$\dot{\varphi}=\alpha\Delta\varphi$$

where $\alpha>0$ is the thermal diffusivity of the medium, and Δ is the Laplace operator.

PROOF. Again, many things can be said here, the idea being that this is what comes out of experiments, but that some mathematical justifications, that we will present in what follows, are possible as well. We actually already met this equation in chapter 3, in one dimension, with some brief explanations. In what follows we will review the 1D case, with full explanations, then discuss the extension to 2D. The idea is as follows:

(1) Let us first assume, for simplifying, that we are in the one-dimensional case, $N=1$. Here our model looks as follows, with distance $l>0$ between neighbors:

$$\text{---}\circ_{x-l}\overset{l}{\text{---}}\circ_{x}\overset{l}{\text{---}}\circ_{x+l}\text{---}$$

In order to model heat diffusion, we have to implement the intuitive mechanism of this phenomenon, namely "the rate of change 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".

(2) In practice, this leads to a condition as follows, expressing the change of the temperature φ, over a small period of time $\delta > 0$:

$$\varphi(x, t+\delta) = \varphi(x,t) + \frac{\alpha\delta}{l^2}\sum_{x\sim y}[\varphi(y,t) - \varphi(x,t)]$$

To be more precise, we have made several assumptions here, as follows:

– General heat diffusion assumption: the change of temperature at any given point x is proportional to the average over neighbors, $y \sim x$, of the differences $\varphi(y,t) - \varphi(x,t)$ between the temperatures at x, and at these neighbors y.

– Infinitesimal time and length conditions: in our model, the change of temperature at a given point x is proportional to small period of time involved, $\delta > 0$, and is inverse proportional to the square of the distance between neighbors, l^2.

(3) Regarding these latter assumptions, the one regarding the proportionality with the time elapsed $\delta > 0$ is something quite natural, physically speaking, and mathematically speaking too, because we can rewrite our equation as follows, making it clear that we have here an equation regarding the rate of change of temperature at x:

$$\frac{\varphi(x, t+\delta) - \varphi(x,t)}{\delta} = \frac{\alpha}{l^2}\sum_{x\sim y}[\varphi(y,t) - \varphi(x,t)]$$

As for the second assumption that we made above, namely inverse proportionality with l^2, this can be justified on physical grounds too, but again, perhaps the best is to do the math, which will show right away where this proportionality comes from.

(4) So, 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:

$$\frac{\varphi(x, t+\delta) - \varphi(x,t)}{\delta} = \alpha \cdot \frac{\varphi(x+l,t) - 2\varphi(x,t) + \varphi(x-l,t)}{l^2}$$

As it was the case when modelling the wave equation in chapter 1, 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:

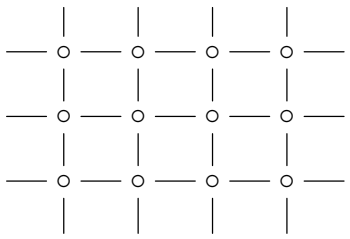
$$\frac{\varphi(x, t+\delta) - \varphi(x,t)}{\delta} = \alpha \cdot \varphi''(x,t)$$

Now with $t \to 0$, we are led in this way to the heat equation, namely:

$$\dot{\varphi}(x,t) = \alpha \cdot \varphi''(x,t)$$

(5) Summarizing, we are done with the 1D case, with our proof being quite similar to the one for the wave equation, from chapter 1. There are of course still a few details to be discussed, in relation with all this, for instance at the end, in relation with the precise order of the limiting operations $l \to 0$ and $\delta \to 0$ to be performed, but these remain minor aspects, because our equation makes it clear, right from the beginning, that time and space are separated, and so that there is no serious issue with all this.

(6) With this done, let us discuss now the case of 2 dimensions. Here we can use a lattice model as follows, with all lengths being $l > 0$, for simplifying:



But, a similar analysis to the one performed above leads to a similar heat equation, with the second derivative φ'' being now replaced by the quantity $\Delta\varphi$.

(7) Finally, in analogy with what happened before for the waves, it is possible in fact to talk about the heat equation in arbitrary N dimensions, as being $\dot{\varphi} = \alpha\Delta\varphi$, with the Laplace operator here being defined in a straightforward way, as follows:

$$\Delta\varphi = \sum_{i=1}^{N} \frac{d^2\varphi}{dx_i^2}$$

As before, this is quite nice, because at $N = 1$ this Laplace operator is just the second derivative, so what we have here is a unification of what we did in the above. In addition, at $N = 3$ we obtain in this way the heat equation in the case of main interest, namely the one of our real-life world. But more on this later in this book. □

8b. Harmonic functions

To summarize, we have been doing some physics, with the conclusion that both the wave and heat equation involve the Laplace operator. So, let us formulate, as a conclusion, the definition of this operator, that we would like to further understand:

DEFINITION 8.7. *The Laplace operator in 2 dimensions is:*

$$\Delta f = \frac{d^2 f}{dx^2} + \frac{d^2 f}{dy^2}$$

A function $f : \mathbb{R}^2 \to \mathbb{C}$ satisfying $\Delta f = 0$ will be called harmonic.

Here the formula of Δ is the one coming from Theorem 8.5 and Theorem 8.6. As for the notion of harmonic function, this is something quite natural. Indeed, we can think of Δ as being a linear operator on the space of functions $f : \mathbb{R}^2 \to \mathbb{C}$, and previous experience with linear operators and linear algebra in general suggests looking first into the eigenvectors of Δ. But the simplest such eigenvectors are those corresponding to the eigenvalue $\lambda = 0$, and these are exactly our harmonic functions, satisfying:

$$\Delta f = 0$$

Getting now to more concrete things, and to some mathematics that we can do, using our knowledge, let us try to find the functions $f : \mathbb{R}^2 \to \mathbb{C}$ which are harmonic. And here, as a good surprise, we have an interesting link with the holomorphic functions:

THEOREM 8.8. *Any holomorphic function $f : \mathbb{C} \to \mathbb{C}$, when regarded as real function*

$$f : \mathbb{R}^2 \to \mathbb{C}$$

is harmonic. Moreover, the conjugates $\bar{f}$ of holomorphic functions are harmonic too.

PROOF. The first assertion follows from the following computation, for the power functions $f(z) = z^n$, with the usual notation $z = x + iy$:

$$\begin{aligned}
\Delta z^n &= \frac{d^2 z^n}{dx^2} + \frac{d^2 z^n}{dy^2} \\
&= \frac{d(nz^{n-1})}{dx} + \frac{d(inz^{n-1})}{dy} \\
&= n(n-1)z^{n-2} - n(n-1)z^{n-2} \\
&= 0
\end{aligned}$$

As for the second assertion, this follows from $\Delta \bar{f} = \overline{\Delta f}$, which is clear from definitions, and which shows that if f is harmonic, than so is its conjugate $\bar{f}$. □

All this is quite interesting, and the idea in what follows will be that of developing a theory of harmonic functions, as a generalization of the theory that we know of the holomorphic functions, covering as well functions of type $\bar{z}$. Them, we will go back to physics, with some applications of this to the wave equation, and the heat equation.

As a first goal, in order to understand the harmonic functions, we can try to find the homogeneous polynomials $P \in \mathbb{R}[x, y]$ which are harmonic. In order to do so, the most convenient is to use the variable $z = x + iy$, and think of these polynomials as being homogeneous polynomials $P \in \mathbb{R}[z, \bar{z}]$. With this convention, the result is as follows:

THEOREM 8.9. *The degree n homogeneous polynomials $P \in \mathbb{R}[x, y]$ which are harmonic are precisely the linear combinations of*

$$P = z^n \quad , \quad P = \bar{z}^n$$

with the usual identification $z = x + iy$.

PROOF. As explained above, any homogeneous polynomial $P \in \mathbb{R}[x,y]$ can be regarded as an homogeneous polynomial $P \in \mathbb{R}[z,\bar{z}]$, with the change of variables $z = x+iy$, and in this picture, the degree n homogeneous polynomials are as follows:

$$P(z) = \sum_{k+l=n} c_{kl} z^k \bar{z}^l$$

In oder to solve now the Laplace equation $\Delta P = 0$, we must compute the quantities $\Delta(z^k\bar{z}^l)$, for any k,l. But the computation here is routine. We first have the following formula, with the derivatives being computed with respect to the variable x:

$$\begin{aligned}\frac{d(z^k\bar{z}^l)}{dx} &= (z^k)'\bar{z}^l + z^k(\bar{z}^l)'\\ &= kz^{k-1}\bar{z}^l + lz^k\bar{z}^{l-1}\end{aligned}$$

By taking one more time the derivative with respect to x, we obtain:

$$\begin{aligned}\frac{d^2(z^k\bar{z}^l)}{dx^2} &= k(z^{k-1}\bar{z}^l)' + l(z^k\bar{z}^{l-1})'\\ &= k\left[(z^{k-1})'\bar{z}^l + z^{k-1}(\bar{z}^l)'\right] + l\left[(z^k)'\bar{z}^{l-1} + z^k(\bar{z}^{l-1})'\right]\\ &= k\left[(k-1)z^{k-2}\bar{z}^l + lz^{k-1}\bar{z}^{l-1}\right] + l\left[kz^{k-1}\bar{z}^{l-1} + (l-1)z^k\bar{z}^{l-2}\right]\\ &= k(k-1)z^{k-2}\bar{z}^l + 2klz^{k-1}\bar{z}^{l-1} + l(l-1)z^k\bar{z}^{l-2}\end{aligned}$$

With respect to the variable y, the computations are similar, but some $\pm i$ factors appear, due to $z' = i$ and $\bar{z}' = -i$, coming from $z = x+iy$. We first have:

$$\begin{aligned}\frac{d(z^k\bar{z}^l)}{dy} &= (z^k)'\bar{z}^l + z^k(\bar{z}^l)'\\ &= ikz^{k-1}\bar{z}^l - ilz^k\bar{z}^{l-1}\end{aligned}$$

By taking one more time the derivative with respect to y, we obtain:

$$\begin{aligned}\frac{d^2(z^k\bar{z}^l)}{dy^2} &= ik(z^{k-1}\bar{z}^l)' - il(z^k\bar{z}^{l-1})'\\ &= ik\left[(z^{k-1})'\bar{z}^l + z^{k-1}(\bar{z}^l)'\right] - il\left[(z^k)'\bar{z}^{l-1} + z^k(\bar{z}^{l-1})'\right]\\ &= ik\left[i(k-1)z^{k-2}\bar{z}^l - ilz^{k-1}\bar{z}^{l-1}\right] - il\left[ikz^{k-1}\bar{z}^{l-1} - i(l-1)z^k\bar{z}^{l-2}\right]\\ &= -k(k-1)z^{k-2}\bar{z}^l + 2klz^{k-1}\bar{z}^{l-1} - l(l-1)z^k\bar{z}^{l-2}\end{aligned}$$

We can now sum the formulae that we found, and we obtain:

$$\begin{aligned}\Delta(z^k\bar{z}^l) &= \frac{d^2(z^k\bar{z}^l)}{dx^2}+\frac{d^2(z^k\bar{z}^l)}{dy^2}\\ &= k(k-1)z^{k-2}\bar{z}^l+2klz^{k-1}\bar{z}^{l-1}+l(l-1)z^k\bar{z}^{l-2}\\ &\quad -k(k-1)z^{k-2}\bar{z}^l+2klz^{k-1}\bar{z}^{l-1}-l(l-1)z^k\bar{z}^{l-2}\\ &= 4klz^{k-1}\bar{z}^{l-1}\end{aligned}$$

In other words, we have reached to the following formula:

$$f=z^k\bar{z}^l\implies\Delta f=\frac{4klf}{|z|^2}$$

Now let us get back to our homogeneous polynomial P, written as follows:

$$P(z)=\sum_{k+l=n}c_{kl}z^k\bar{z}^l$$

By using the above formula, the Laplacian of P is given by:

$$\Delta P(z)=\frac{4}{|z|^2}\sum_{k+l=n}klc_{kl}z^k\bar{z}^l$$

We conclude that the Laplace equation for P reads:

$$\begin{aligned}\Delta P=0 &\iff klc_{kl}=0,\forall k,l\\ &\iff [k,l\neq 0\implies c_{kl}=0]\\ &\iff P=c_{n0}z^n+c_{0n}\bar{z}^n\end{aligned}$$

Thus, we are led to the conclusion in the statement. And with the observation that the real formulation of the final result is something quite complicated, and so, for one more time, the use of the complex variable $z=x+iy$ is something very useful. □

We know that the holomorphic functions are harmonic, and it follows from this that the real and imaginary parts of the holomorphic functions, as well as any linear combinations of these real and imaginary parts, are harmonic too. That is, if f is holomorphic, then the following function is harmonic, for any values of the parameters $\alpha,\beta\in\mathbb{C}$:

$$f_{\alpha\beta}=\alpha Re(f)+\beta Im(f)$$

Observe that this result covers all the examples that we have so far, for instance with the function $\bar{z}$, that we know to be harmonic, appearing as follows:

$$\bar{z}=Re(z)-iIm(z)$$

Our main goal in what follows will be that of proving a converse to this, at least locally. For this purpose, let us start with the following definition:

DEFINITION 8.10. *The Cauchy-Riemann operators are*

$$\partial = \frac{1}{2}\left(\frac{d}{dx} - i\frac{d}{dy}\right) \quad , \quad \bar{\partial} = \frac{1}{2}\left(\frac{d}{dx} + i\frac{d}{dy}\right)$$

where $\frac{d}{dx}$ *and* $\frac{d}{dy}$ *are the usual partial derivatives for complex functions.*

There are many things that can be said about the Cauchy-Riemann operators $\partial, \bar{\partial}$, the idea being that in many contexts, these are better to use than the usual partial derivatives $\frac{d}{dx}, \frac{d}{dy}$, and with this being a bit like the usage of the variables $z, \bar{z}$, instead of the decomposition $z = a + ib$, for many questions regarding the complex numbers.

We have already seen in fact some instances of this, in our computations above. At the general level, the main properties of $\partial, \bar{\partial}$ can be summarized as follows:

PROPOSITION 8.11. *Assume that* $f : X \to \mathbb{C}$ *is differentiable in the real sense.*

(1) f *is holomorphic precisely when it satisfies* $\bar{\partial} f = 0$.
(2) *Moreover, in this case its derivative is* $f' = \partial f$.
(3) *The Laplace operator is given by* $\Delta = 4\partial\bar{\partial}$.
(4) f *is harmonic precisely when it satisfies* $\partial\bar{\partial} f = 0$.

PROOF. We can assume by linearity that we are dealing with differentiability questions at 0. Since our function $f : X \to \mathbb{C}$ is differentiable in the real sense, we have a formula as follows, with $z = x + iy$, and with $a, b \in \mathbb{C}$ being the partial derivatives at 0:

$$f(z) = ax + by + o(z)$$

Now observe that we can write this formula in the following way:

$$\begin{aligned} f(z) &= a \cdot \frac{z + \bar{z}}{2} + b \cdot \frac{z - \bar{z}}{2i} + o(z) \\ &= a \cdot \frac{z + \bar{z}}{2} + b \cdot \frac{i\bar{z} - iz}{2} + o(z) \\ &= \frac{a - ib}{2} \cdot z + \frac{a + ib}{2} \cdot \bar{z} + o(z) \end{aligned}$$

Now by dividing by z, we obtain from this the following formula:

$$\begin{aligned} \frac{f(z)}{z} &= \frac{a - ib}{2} + \frac{a + ib}{2} \cdot \frac{\bar{z}}{z} + o(1) \\ &= \partial f(0) + \bar{\partial} f(0) \cdot \frac{\bar{z}}{z} + o(1) \end{aligned}$$

But this gives the first two assertions, because in order for the derivative $f'(0)$ to exist, appearing as the $z \to 0$ limit of the above quantity, the coefficient of $\bar{z}/z$, which does not

converge, must vanish. Regarding now the third assertion, this follows from:

$$\begin{aligned}\Delta &= \frac{d^2}{dx^2}+\frac{d^2}{dy^2}\\ &= \left(\frac{d}{dx}-i\frac{d}{dy}\right)\left(\frac{d}{dx}+i\frac{d}{dy}\right)\\ &= 4\partial\bar{\partial}\end{aligned}$$

As for the last assertion, this is clear from our formula of Δ. □

In analogy now with the theory of the holomorphic functions, we have:

THEOREM 8.12. *The harmonic functions obey to the same general principles as the holomorphic functions, namely:*

(1) *The maximum modulus principle.*
(2) *The plain mean value formula.*
(3) *The boundary mean value formula.*
(4) *The Liouville theorem.*

Also, locally, the real harmonic functions are the real parts of holomorphic functions.

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

(1) Regarding the maximum modulus principle, the statement here is that any harmonic function $f : X \to \mathbb{C}$ has the property that the maximum of $|f|$ over a domain is attained on its boundary. That is, given a domain D, with boundary γ, we have:

$$\exists x \in \gamma \quad , \quad |f(x)| = \max_{y\in D} |f(y)|$$

(2) Regarding the plain mean value formula, here the statement is that given an harmonic function $f : X \to \mathbb{C}$, and a disk D, the following happens:

$$f(x) = \int_D f(y)dy$$

(3) Regarding the boundary mean value formula, here the statement is that given an harmonic function $f : X \to \mathbb{C}$, and a disk D, with boundary γ, the following happens:

$$f(x) = \int_\gamma f(y)dy$$

(4) Regarding the Liouville theorem, the statement here is that an entire, bounded harmonic function must be constant:

$$f : \mathbb{C} \to \mathbb{C} \quad , \quad |f| \leq M \quad \implies \quad f = \text{constant}$$

(5) Finally, regarding the proofs, these are not exactly trivial. According to our experience with holomorphic functions, (2,3) are equivalent, and imply (1,4) via some simple arguments, and the same happens for the harmonic functions. However, establishing (2)

and the last assertion is not exactly trivial, but hey, remember that we are doing a physics chapter here, so we refer to Rudin [75] for proofs of this, and with the promise that we will be back to this, later in this book, once we will know more calculus. □

Observe that we are in the process of a tactical retreat from mathematics. As a last objective, however, before giving up completely with all this, and getting back to physics, let us try to find the harmonic functions which are radial, in the following sense:

$$f(z) = \varphi(|z|)$$

However, things are quite tricky here, involving a blowup phenomenon at the dimension value $N = 2$, which is precisely the one that we are interested in. So, moving now to N dimensions, with a straightforward definition for Δ there, here is the result:

THEOREM 8.13. *The fundamental radial solutions of $\Delta f = 0$ are*

$$f(x) = \begin{cases} ||x||^{2-N} & (N \neq 2) \\ \log ||x|| & (N = 2) \end{cases}$$

with the log *at $N = 2$ basically coming from $\log' = 1/x$.*

PROOF. Consider indeed a radial function, defined outside the origin $x = 0$:

$$f : \mathbb{R}^N - \{0\} \to \mathbb{C} \quad , \quad f(x) = \varphi(|x|)$$

Here, and in what follows, we denote by $|x|$ instead of $||x||$, for simplifying notations, the norm of the vectors $x \in \mathbb{R}^N$, which is given, as usual, by the following formula:

$$|x| = \sqrt{\sum_{i=1}^{N} x_i^2}$$

Our first goal is that of reformulating the Laplace equation $\Delta f = 0$ in terms of the one-variable function $\varphi : (0, \infty) \to \mathbb{C}$. For this purpose, observe first that we have:

$$\begin{aligned} \frac{d|x|}{dx_i} &= \frac{d\sqrt{\sum_{i=1}^{N} x_i^2}}{dx_i} \\ &= \frac{1}{2} \cdot \frac{1}{\sqrt{\sum_{i=1}^{N} x_i^2}} \cdot \frac{d\left(\sum_{i=1}^{N} x_i^2\right)}{dx_i} \\ &= \frac{1}{2} \cdot \frac{1}{|x|} \cdot 2x_i \\ &= \frac{x_i}{|x|} \end{aligned}$$

By using this formula, we have the following computation:

$$\begin{aligned}\frac{df}{dx_i} &= \frac{d\varphi(|x|)}{dx_i}\\ &= \varphi'(|x|)\cdot\frac{d|x|}{dx_i}\\ &= \varphi'(|x|)\cdot\frac{x_i}{|x|}\end{aligned}$$

By differentiating one more time, we obtain the following formula:

$$\begin{aligned}\frac{d^2f}{dx_i^2} &= \frac{d}{dx_i}\left(\varphi'(|x|)\cdot\frac{x_i}{|x|}\right)\\ &= \frac{d\varphi'(|x|)}{dx_i}\cdot\frac{x_i}{|x|}+\varphi'(|x|)\cdot\frac{d}{dx_i}\left(\frac{x_i}{|x|}\right)\\ &= \left(\varphi''(|x|)\cdot\frac{x_i}{|x|}\right)\cdot\frac{x_i}{|x|}+\varphi'(|x|)\cdot\frac{|x|-x_i\cdot x_i/|x|}{|x|^2}\\ &= \varphi''(|x|)\cdot\frac{x_i^2}{|x|^2}+\varphi'(|x|)\cdot\frac{|x|^2-x_i^2}{|x|^3}\end{aligned}$$

Now by summing over $i\in\{1,\ldots,N\}$, this gives the following formula:

$$\begin{aligned}\Delta f &= \sum_{i=1}^N\varphi''(|x|)\cdot\frac{x_i^2}{|x|^2}+\sum_{i=1}^N\varphi'(|x|)\cdot\frac{|x|^2-x_i^2}{|x|^3}\\ &= \varphi''(|x|)\cdot\frac{|x|^2}{|x|^2}+\varphi'(|x|)\cdot\frac{(N-1)|x|^2}{|x|^3}\\ &= \varphi''(|x|)+\varphi'(|x|)\cdot\frac{N-1}{|x|}\end{aligned}$$

Thus, with $r=|x|$, the Laplace equation $\Delta f=0$ can be reformulated as follows:

$$\varphi''(r)+\frac{(N-1)\varphi'(r)}{r}=0$$

Equivalently, the equation that we want to solve is as follows:

$$r\varphi''+(N-1)\varphi'=0$$

Now observe that we have the following formula:

$$\begin{aligned}(r^{N-1}\varphi')' &= (N-1)r^{N-2}\varphi'+r^{N-1}\varphi''\\ &= r^{N-2}((N-1)\varphi'+r\varphi'')\end{aligned}$$

Thus, the equation to be solved can be simply written as follows:

$$(r^{N-1}\varphi')'=0$$

We conclude that $r^{N-1}\varphi'$ must be a constant K, and so, that we must have:

$$\varphi' = Kr^{1-N}$$

But the fundamental solutions of this latter equation are as follows:

$$\varphi(r) = \begin{cases} r^{2-N} & (N \neq 2) \\ \log r & (N = 2) \end{cases}$$

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

And good news, that is the end. We have learned many things about the Laplace operator and the harmonic functions, but all this clearly becomes too complicated, spilling sometimes in arbitrary N dimensions, so time to stop. We will be back to this.

8c. Light, spectroscopy

Getting back now to physics, a natural question would be that of going back to the equations that we started with, namely the wave and heat equations in 2D, and see if our accumulated knowledge about harmonic functions, which is after all not that bad, can help there. Unfortunately, this is not exactly the case, and we are still a long way to go, from solving that equations. So, obviously, time to ask the cat. And cat says:

CAT 8.14. *If you don't understand physics, do more physics.*

Thanks cat, and this seems wise indeed, let's have some fun with more physics, still related of course to the Laplace operator Δ, and for difficult mathematics, which is still to be solved, we will see later, towards the end of the present book. Getting started now, we will need for what we want to talk about something quite heavy, namely:

THEOREM 8.15 (Maxwell theory). *In regions of space where there is no charge or current present the Maxwell equations for electrodynamics read*

$$< \nabla, E >=< \nabla, B >= 0$$

$$\nabla \times E = -\dot{B} \quad , \quad \nabla \times B = \dot{E}/c^2$$

and both the electric field E and magnetic field B are subject to the wave equation

$$\ddot{\varphi} = c^2\Delta\varphi$$

where $\Delta = \sum_i d^2/dx_i^2$ is the Laplace operator, and $c = 299,792,458$.

PROOF. This is something fundamental, appearing as a tricky mixture of physics facts and mathematical results, the idea being as follows:

(1) To start with, electrodynamics is the science of moving electrical charges. And this is something quite complicated, because unlike in classical mechanics, where the Newton law is good for both the static and the dynamic setting, the Coulomb law, which

is actually very similar to the Newton law, does the job when the charges are static, but no longer describes well the situation when the charges are moving.

(2) The problem comes from the fact that moving charges produce magnetism, and with this being visible when putting together two electric wires, which will attract or repel, depending on orientation. Thus, in contrast with classical mechanics, where static or dynamic problems are described by a unique field, the gravitational one, in electrodynamics we have two fields, namely the electric field E, and the magnetic field B.

(3) Fortunately, there is a full set of equations relating the electric field E and the magnetic field B. These are the Maxwell equations, which look as follows:

$$< \nabla, E >= \frac{\rho}{\varepsilon_0} \quad , \quad < \nabla, B >= 0$$

$$\nabla \times E = -\dot{B} \quad , \quad \nabla \times B = \mu_0 J + \mu_0 \varepsilon_0 \dot{E}$$

(4) To be more precise, regarding first the math, the dots denote derivatives with respect to time, and ∇ is the gradient operator, or space derivative, given by:

$$\nabla = \begin{pmatrix} \frac{d}{dx} \\ \frac{d}{dy} \\ \frac{d}{dz} \end{pmatrix}$$

(5) Regarding the physics, the first formula is the Gauss law, ρ being the charge, and ε_0 being a constant, and with this Gauss law more or less replacing the Coulomb law from electrostatics. The second formula is something basic, and anonymous. The third formula is the Faraday law. As for the fourth formula, this is the Ampère law, as modified by Maxwell, with J being the volume current density, and μ_0 being a constant.

(6) Without bothering too much about the precise meaning of all this, we can see right away that under the circumstances in the statement, namely in the regions of space where there is no charge or current present, the Maxwell equations take a simple and comprehensible form, readable even without much physics background, namely:

$$< \nabla, E >=< \nabla, B >= 0$$

$$\nabla \times E = -\dot{B} \quad , \quad \nabla \times B = \mu_0 \varepsilon_0 \dot{E}$$

(7) Thus, we have reached to the equations in the statement, modulo a discussion about the constant $\mu_0 \varepsilon_0$. And the point here is that, according to a remarkable discovery of Biot and Savart, the main electrodynamics constants μ_0, ε_0 are magically related to the observed speed of light in vacuum $c = 299,792,458$ by the following formula:

$$\mu_0 \varepsilon_0 = \frac{1}{c^2}$$

(8) Summarizing, we have our equations. Leaving aside the first two equations, by applying the curl operator to the last two equations, we obtain:

$$\nabla \times (\nabla \times E) = -\nabla \times \dot{B} = -(\nabla \times B)' = -\ddot{E}/c^2$$

$$\nabla \times (\nabla \times B) = \nabla \times \dot{E}/c^2 = (\nabla \times E)'/c^2 = -\ddot{B}/c^2$$

But the double curl operator is subject to the following formula:

$$\nabla \times (\nabla \times \varphi) = \nabla < \nabla, \varphi > -\Delta\varphi$$

Now by using the first two equations, we are led to the conclusion in the statement. □

So, big question now, what is light? Light is the wave predicted by Theorem 8.15, travelling at speed c, and with the important extra property that it depends on a real positive parameter, that can be called, upon taste, frequency, wavelength, or color:

FACT 8.16. *An accelerating or decelerating charge produces electromagnetic waves, travelling in vacuum at speed* $c = 299,792,458$,

$$\ddot{\varphi} = c^2\Delta\varphi$$

and in non-vacuum at a lower speed $v < c$. *These waves are called light, whose frequency and wavelength can be explicitely computed.*

This phenomenon can be observed is a variety of situations, such as the usual light bulbs, where electrons get decelerated by the filament, acting as a resistor, or in usual fire, which is a chemical reaction, with the electrons moving around, as they do in any chemical reaction, or in more complicated machinery like nuclear plants, particle accelerators, and so on, leading there to all sorts of eerie glows, of various colors.

Let us try now to understand the simplest solutions of the wave equation $\ddot{\varphi} = c^2\Delta\varphi$, or more generally $\ddot{\varphi} = v^2\Delta\varphi$, from Fact 8.16. In 1D, the situation is as follows:

THEOREM 8.17. *The 1D wave equation, with speed* v, *namely*

$$\ddot{\varphi} = v^2\,\frac{d^2\varphi}{dx^2}$$

has as basic solutions the following functions,

$$\varphi(x) = A\cos(kx - wt + \delta)$$

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:

$$\ddot{\varphi} = -w^2\varphi \quad , \quad \frac{d^2\varphi}{dx^2} = -k^2\varphi$$

Thus, the wave equation is indeed satisfied, with speed $v = w/k$:

$$\ddot{\varphi} = \left(\frac{w}{k}\right)^2 \frac{d^2\varphi}{dx^2} = v^2\,\frac{d^2\varphi}{dx^2}$$

(2) Regarding now the other things in the statement, all this is basically terminology, which is very natural, when thinking how $\varphi(x) = A\cos(kx - wt + \delta)$ propagates.

(3) Finally, the last assertion is standard, coming from Fourier analysis, and we will leave this as an instructive exercise, based on the material from chapter 7. □

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:

$$\varphi(x) = Re\left[A\,e^{i(kx-wt+\delta)}\right]$$

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:

$$\varphi = Re(\widetilde{\varphi}) \quad , \quad \widetilde{\varphi} = \widetilde{A}e^{i(kx-wt)}$$

Moving ahead now towards electromagnetism and 3D, let us formulate:

DEFINITION 8.18. *A monochromatic plane wave is a solution of the 3D wave equation which moves in only 1 direction, making it in practice a solution of the 1D wave equation, and which is of the special from found in Theorem 8.17, with no frequencies mixed.*

In other words, we are making here two assumptions on our wave. First is the 1-dimensionality assumption, which gets us into the framework of Theorem 8.17. And second is the assumption, in connection with the Fourier decomposition result from the end of Theorem 8.17, that our solution is of "pure" type, meaning a wave having a well-defined wavelenght and frequency, instead of being a "packet" of such pure waves.

Summarizing, we have now a decent intuition about what light is, and more on this later, and let us discuss now the examples. The idea is that we have various types of light, depending on frequency and wavelength. These are normally referred to as "electromagnetic waves", but for keeping things simple and luminous, we will keep using

the familiar term "light". The classification, in a rough form, is as follows:

Frequency	Type	Wavelength
	–	
$10^{18} - 10^{20}$	γ rays	$10^{-12} - 10^{-10}$
$10^{16} - 10^{18}$	X – rays	$10^{-10} - 10^{-8}$
$10^{15} - 10^{16}$	UV	$10^{-8} - 10^{-7}$
	–	
$10^{14} - 10^{15}$	blue	$10^{-7} - 10^{-6}$
$10^{14} - 10^{15}$	yellow	$10^{-7} - 10^{-6}$
$10^{14} - 10^{15}$	red	$10^{-7} - 10^{-6}$
	–	
$10^{11} - 10^{14}$	IR	$10^{-6} - 10^{-3}$
$10^{9} - 10^{11}$	microwave	$10^{-3} - 10^{-1}$
$1 - 10^{9}$	radio	$10^{-1} - 10^{8}$

Observe the tiny space occupied by the visible light, all colors there, and the many more missing, being squeezed under the $10^{14} - 10^{15}$ frequency banner. Here is a zoom on that part, with of course the remark that all this, colors, is something subjective:

Frequency THz $= 10^{12}$ Hz	Color	Wavelength nm $= 10^{-9}$ m
	–	
$670 - 790$	violet	$380 - 450$
$620 - 670$	blue	$450 - 485$
$600 - 620$	cyan	$485 - 500$
$530 - 600$	green	$500 - 565$
$510 - 530$	yellow	$565 - 590$
$480 - 510$	orange	$590 - 625$
$400 - 480$	red	$625 - 750$

With this in hand, we can now do some basic optics. Light usually comes in "bundles", with waves of several wavelenghts coming at the same time, from the same source, and the first challenge is that of separating these wavelenghts. In order to discuss this, from a practical perspective, let us start with the following fact:

FACT 8.19. *When travelling through a material, and hitting a new material, some of the light gets reflected, at the same angle, and some of it gets refracted, at a different angle, depending both on the old and the new material, and on the wavelength.*

Again, this is something deep, and there are many things that can be said here, ranging from various computations based on the Maxwell equations, to all sorts of considerations belonging to advanced materials theory. As a basic formula, we have the famous Snell

law, which relates the incidence angle θ_1 to the refraction angle θ_2, as follows:

$$\frac{\sin\theta_2}{\sin\theta_1} = \frac{n_1(\lambda)}{n_2(\lambda)}$$

Here $n_i(\lambda)$ are the refraction indices of the two materials, adjusted for the wavelength, and with this adjustment for wavelength being the whole point, which is something quite complicated. Now as a simple consequence of the above, we have:

THEOREM 8.20. *Light can be decomposed, by using a prism.*

PROOF. This follows from Fact 8.19. Indeed, when hitting a piece of glass, provided that the hitting angle is not 90°, the light will decompose over the wavelenghts present, with the corresponding refraction angles depending on these wavelengths. And we can capture these split components at the exit from the piece of glass, again deviated a bit, provided that the exit surface is not parallel to the entry surface. And the simplest device doing the job, that is, having two non-parallel faces, is a prism. □

As an application of this, we can study events via spectroscopy, by capturing the light the event has produced, decomposing it with a prism, carefully recording its "spectral signature", consisting of the wavelenghts present, and their density, and then doing some reverse engineering, consisting in reconstructing the event out of its spectral signature.

8d. Atomic spectrum

Getting now to some truly exciting applications of light and spectroscopy, let us discuss the beginnings of the atomic theory. There is a long story here, involving many discoveries, around 1890-1900, focusing on hydrogen H. We will present here things a bit retrospectively. First on our list is the following discovery, by Lyman in 1906:

FACT 8.21 (Lyman). *The hydrogen atom has spectral lines given by the formula*

$$\frac{1}{\lambda} = R\left(1 - \frac{1}{n^2}\right)$$

where $R \simeq 1.097 \times 10^7$ *and* $n \geq 2$*, which are as follows,*

n	Name	Wavelength	Color
	$-$	$-$	
2	α	121.567	UV
3	β	102.572	UV
4	γ	97.254	UV
$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	limit	91.175	UV

called Lyman series of the hydrogen atom.

Observe that all the Lyman series lies in UV, which is invisible to the naked eye. Due to this fact, this series, while theoretically being the most important, was discovered only second. The first discovery, which was the big one, and the breakthrough, was by Balmer, the founding father of all this, back in 1885, in the visible range, as follows:

FACT 8.22 (Balmer). *The hydrogen atom has spectral lines given by the formula*

$$\frac{1}{\lambda} = R\left(\frac{1}{4} - \frac{1}{n^2}\right)$$

where $R \simeq 1.097 \times 10^7$ *and* $n \geq 3$*, which are as follows,*

n	Name	Wavelength	Color
	–	–	
3	α	656.279	red
4	β	486.135	aqua
5	γ	434.047	blue
6	δ	410.173	violet
7	ε	397.007	UV
$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	limit	346.600	UV

called Balmer series of the hydrogen atom.

So, this was Balmer's original result, which started everything. As a third main result now, this time in IR, due to Paschen in 1908, we have:

FACT 8.23 (Paschen). *The hydrogen atom has spectral lines given by the formula*

$$\frac{1}{\lambda} = R\left(\frac{1}{9} - \frac{1}{n^2}\right)$$

where $R \simeq 1.097 \times 10^7$ *and* $n \geq 4$*, which are as follows,*

n	Name	Wavelength	Color
	–	–	
4	α	1875	IR
5	β	1282	IR
6	γ	1094	IR
$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	limit	820.4	IR

called Paschen series of the hydrogen atom.

Observe the striking similarity between the above three results. In fact, we have here the following fundamental, grand result, due to Rydberg in 1888, based on the Balmer series, and with later contributions by Ritz in 1908, using the Lyman series as well:

CONCLUSION 8.24 (Rydberg, Ritz). *The spectral lines of the hydrogen atom are given by the Rydberg formula, depending on integer parameters $n_1 < n_2$,*

$$\frac{1}{\lambda_{n_1n_2}} = R\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

with R being the Rydberg constant for hydrogen, which is as follows:

$$R \simeq 1.096\ 775\ 83 \times 10^7$$

These spectral lines combine according to the Ritz-Rydberg principle, as follows:

$$\frac{1}{\lambda_{n_1n_2}} + \frac{1}{\lambda_{n_2n_3}} = \frac{1}{\lambda_{n_1n_3}}$$

Similar formulae hold for other atoms, with suitable fine-tunings of R.

Here the first part, the Rydberg formula, generalizes the results of Lyman, Balmer, Paschen, which appear at $n_1 = 1, 2, 3$, at least retrospectively. The Rydberg formula predicts further spectral lines, appearing at $n_1 = 4, 5, 6, \ldots$, and these were discovered later, by Brackett in 1922, Pfund in 1924, Humphreys in 1953, and others aftwerwards, with all these extra lines being in far IR. The simplified complete table is as follows:

n_1	n_2	Series name	Wavelength $n_2 = \infty$	Color $n_2 = \infty$
		–	–	
1	$2 - \infty$	Lyman	91.13 nm	UV
2	$3 - \infty$	Balmer	364.51 nm	UV
3	$4 - \infty$	Paschen	820.14 nm	IR
		–	–	
4	$5 - \infty$	Brackett	1458.03 nm	far IR
5	$6 - \infty$	Pfund	2278.17 nm	far IR
6	$7 - \infty$	Humphreys	3280.56 nm	far IR
⋮	⋮	⋮	⋮	⋮

Regarding the last assertion, concerning other elements, this is something conjectured and partly verified by Ritz, and fully verified and clarified later, via many experiments, the fine-tuning of R being basically $R \to RZ^2$, where Z is the atomic number.

But from a theoretical physics viewpoint, the main result remains the middle assertion, called Ritz-Rydberg combination principle. This is something at the same time extremely simple, and completely puzzling, the informal conclusion being as follows:

THOUGHT 8.25. *The simplest observables of the hydrogen atom, combining via*

$$\frac{1}{\lambda_{n_1n_2}} + \frac{1}{\lambda_{n_2n_3}} = \frac{1}{\lambda_{n_1n_3}}$$

look like quite weird quantities. Why wouldn't they just sum normally.

Fortunately, mathematics comes to the rescue. Indeed, the Ritz-Rydberg combination principle reminds the formula $e_{n_1n_2}e_{n_2n_3} = e_{n_1n_3}$ for the usual matrix units $e_{ij} : e_j \to e_i$. In short, we are in familiar territory here, and we can start dreaming of:

THOUGHT 8.26. *Observables in quantum mechanics should be some sort of infinite matrices, generalizing the Lyman, Balmer, Paschen lines of the hydrogen atom, and multiplying between them as the matrices do, as to produce further observables.*

Time now to put everything together. As a main problem that we would like to solve, we have the understanding the intimate structure of matter, at the atomic level. There is of course a long story here, regarding the intimate structure of matter, going back centuries and even millennia ago, and our presentation here will be quite simplified. As a starting point, since we need a starting point, let us agree on:

CLAIM 8.27. *Ordinary matter is made of small particles called atoms, with each atom appearing as a mix of even smaller particles, namely protons $+$, neutrons 0 and electrons $-$, with the same number of protons $+$ and electrons $-$.*

As a first observation, this is something which does not look obvious at all, with probably lots of work, by many people, being involved, as to lead to this claim. And so it is. The story goes back to the discovery of charges and electricity, which were attributed to a small particle, the electron $-$. Now since matter is by default neutral, this naturally leads to the consideration to the proton $+$, having the same charge as the electron.

But, as a natural question, why should be these electrons $-$ and protons $+$ that small? And also, what about the neutron 0? These are not easy questions, and the fact that it is so came from several clever experiments. Let us first recall that careful experiments with tiny particles are practically impossible. However, all sorts of brutal experiments, such as bombarding matter with other pieces of matter, accelerated to the extremes, or submitting it to huge electric and magnetic fields, do work. And it is such kind of experiments, due to Thomson, Rutherford and others, "peeling off" protons $+$, neutrons 0 and electrons $-$ from matter, and observing them, that led to the conclusion that these small beasts $+, 0, -$ exist indeed, in agreement with Claim 8.27.

Of particular importance here was as well the radioactivity theory of Becquerel and Pierre and Marie Curie, involving this time such small beasts, or perhaps some related radiation, peeling off by themselves, in heavy elements such as uranium ${}_{92}\mathrm{U}$, polonium ${}_{84}\mathrm{Po}$ and radium ${}_{88}\mathrm{Ra}$. And there was also Einstein's work on the photoelectric effect, light interacting with matter, suggesting that even light itself might have associated to it some kind of particle, called photon. All this goes of course beyond Claim 8.27, with further particles involved, and more on this later, but as a general idea, all this deluge of small particle findings, all coming around 1900-1910, further solidified Claim 8.27.

So, taking now Claim 8.27 for granted, how are then the atoms organized, as mixtures of protons $+$, neutrons 0 and electrons $-$? The answer here lies again in the above-mentioned "brutal" experiments of Thomson, Rutherford and others, which not only proved Claim 8.27, but led to an improved version of it, as follows:

CLAIM 8.28. *The atoms are formed by a core of protons $+$ and neutrons 0, surrounded by a cloud of electrons $-$, gravitating around the core.*

This is a considerable advance, because we are now into familiar territory, namely some kind of mechanics. And with this in mind, all the pieces of our puzzle start fitting together, and we are led to the following grand conclusion:

CLAIM 8.29 (Bohr and others). *The atoms are formed by a core of protons and neutrons, surrounded by a cloud of electrons, basically obeying to a modified version of electromagnetism. And with a fine mechanism involved, as follows:*

(1) *The electrons are free to move only on certain specified elliptic orbits, labelled $1, 2, 3, \ldots$, situated at certain specific heights.*
(2) *The electrons can jump or fall between orbits $n_1 < n_2$, absorbing or emitting light and heat, that is, electromagnetic waves, as accelerating charges.*
(3) *The energy of such a wave, coming from $n_1 \to n_2$ or $n_2 \to n_1$, is given, via the Planck viewpoint, by the Rydberg formula, applied with $n_1 < n_2$.*
(4) *The simplest such jumps are those observed by Lyman, Balmer, Paschen. And multiple jumps explain the Ritz-Rydberg formula.*

And isn't this beautiful. Moreover, some further claims, also by Bohr and others, are that the theory can be further extended and fine-tuned as to explain many other phenomena, such as the above-mentioned findings of Einstein, and of Becquerel and Pierre and Marie Curie, and generally speaking, all the physics and chemistry known.

And the story is not over here. Following now Heisenberg, the next claim is that the underlying mathematics in all the above can lead to a beautiful axiomatization of quantum mechanics, as a "matrix mechanics", along the lines of Thought 8.26. We will be back to all this in chapter 16 below, at the end of the present book.

8e. Exercises

We had a tough physics chapter here, and here are some physics exercises:

EXERCISE 8.30. *Find proofs for the main properties of the harmonic functions.*

EXERCISE 8.31. *Learn about soap films, and their relation with harmonic functions.*

EXERCISE 8.32. *Explore a bit the wave and heat equations, in 1 and 2 dimensions.*

EXERCISE 8.33. *Work out the details for the decomposition of wave packets.*

As bonus exercise, try developing quantum mechanics, based on the above. With this being not a joke, Heisenberg did not know much more than that, when he did it.

Part III

Several variables

This is my church
This is where I heal my hurts
For tonight
God is a DJ

CHAPTER 9

Linear maps

9a. Linear maps

In the remainder of this book we study the functions of several variables. Things here are quite tricky, and we will do this slowly, over a whole 200 pages. Let us start with the linear functions, which are the simplest. We have here the following result:

THEOREM 9.1. *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 $A_{ij} =< f(e_j), e_i >$. Similarly, the linear maps $f : \mathbb{C}^N \to \mathbb{C}^N$ are in correspondence with the matrices $A \in M_{M\times N}(\mathbb{C})$.

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 numbers $A_{ij} \in \mathbb{R}$:

$$f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} A_{11}x_1 + \ldots + A_{1N}x_N \\ \vdots \\ A_{M1}x_1 + \ldots + A_{MN}x_N \end{pmatrix}$$

Now the parameters $A_{ij} \in \mathbb{R}$ can be regarded as being the entries of a rectangular matrix $A \in M_{M\times N}(\mathbb{R})$, and with the usual convention for the rectangular matrix multiplication, the above formula is precisely the one in the statement, namely:

$$f(x) = Ax$$

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:

$$f(e_j) = \begin{pmatrix} A_{1j} \\ \vdots \\ A_{Mj} \end{pmatrix}$$

But this gives the second formula, $< f(e_j), e_i >= A_{ij}$, as desired. As for the last assertion, regarding complex maps and matrices, the proof here is similar. □

At the level of examples, let us focus on the linear maps $f : \mathbb{R}^2 \to \mathbb{R}^2$. We have:

PROPOSITION 9.2. *The rotation of angle $t \in \mathbb{R}$, and the symmetry with respect to the Ox axis rotated by an angle $t/2 \in \mathbb{R}$, are given by the matrices*

$$R_t = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \quad , \quad S_t = \begin{pmatrix} \cos t & \sin t \\ \sin t & -\cos t \end{pmatrix}$$

both depending on $t \in \mathbb{R}$ taken modulo 2π.

PROOF. The rotation being linear, it must correspond to a certain matrix:

$$R_t = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

We can guess this matrix, via its action on the basic coordinate vectors $\binom{1}{0}$ and $\binom{0}{1}$. Indeed, a quick picture in the plane shows that we must have:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix} \quad , \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -\sin t \\ \cos t \end{pmatrix}$$

Guessing now the matrix is not complicated, because the first equation gives us the first column, and the second equation gives us the second column:

$$\begin{pmatrix} a \\ c \end{pmatrix} = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix} \quad , \quad \begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} -\sin t \\ \cos t \end{pmatrix}$$

Thus, we can just put together these two vectors, and we obtain our matrix R_t. As for the symmetry, the proof here is similar, again by computing $S_t\binom{1}{0}$ and $S_t\binom{0}{1}$. □

Let us record as well a result regarding the projections, as follows:

PROPOSITION 9.3. *The projection on the Ox axis rotated by an angle $t/2 \in \mathbb{R}$ is*

$$P_t = \frac{1}{2} \begin{pmatrix} 1 + \cos t & \sin t \\ \sin t & 1 - \cos t \end{pmatrix}$$

depending on $t \in \mathbb{R}$ taken modulo 2π.

PROOF. A quick picture in the plane, using similarity of triangles, and the basic trigonometry formulae for the duplication of angles, show that we must have:

$$P_t \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \cos \frac{t}{2} \begin{pmatrix} \cos \frac{t}{2} \\ \sin \frac{t}{2} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 + \cos t \\ \sin t \end{pmatrix}$$

Similarly, another quick picture plus trigonometry show that we must have:

$$P_t \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \sin \frac{t}{2} \begin{pmatrix} \cos \frac{t}{2} \\ \sin \frac{t}{2} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \sin t \\ 1 - \cos t \end{pmatrix}$$

Now by putting together these two vectors, and we obtain our matrix. □

Back to Theorem 9.1, our claim is that, no matter what we want to do with f or A, we will run at some point into their adjoints f^* and A^*, constructed as follows:

THEOREM 9.4. *The adjoint linear map $f^* : \mathbb{C}^N \to \mathbb{C}^N$, which is given by*

$$< f(x), y >=< x, f^*(y) >$$

corresponds to the adjoint matrix $A^* \in M_N(\mathbb{C})$, given by

$$(A^*)_{ij} = \bar{A}_{ji}$$

via the correspondence between linear maps and matrices constructed above.

PROOF. Given a linear map $f : \mathbb{C}^N \to \mathbb{C}^N$, fix $y \in \mathbb{C}^N$, and consider the linear form $\varphi(x) =< f(x), y >$. This form must be as follows, for a certain vector $f^*(y) \in \mathbb{C}^N$:

$$\varphi(x) =< x, f^*(y) >$$

Thus, we have constructed a map $y \to f^*(y)$ as in the statement, which is obviously linear, and that we can call f^*. Now by taking the vectors $x, y \in \mathbb{C}^N$ to be elements of the standard basis of $\mathbb{C}^N$, our defining formula for f^* reads:

$$< f(e_i), e_j >=< e_i, f^*(e_j) >$$

By reversing the scalar product on the right, this formula can be written as:

$$< f^*(e_j), e_i >= \overline{< f(e_i), e_j >}$$

But this means that the matrix of f^* is given by $(A^*)_{ij} = \bar{A}_{ji}$, as desired. □

Getting back to our claim, the adjoints $*$ are indeed ubiquitous, as shown by:

THEOREM 9.5. *The following happen:*

(1) *$f(x) = Ux$ with $U \in M_N(\mathbb{C})$ is an isometry precisely when $U^* = U^{-1}$.*
(2) *$f(x) = Px$ with $P \in M_N(\mathbb{C})$ is a projection precisely when $P = P^2 = P^*$.*

PROOF. Let us first recall that the lengths, or norms, of the vectors $x \in \mathbb{C}^N$ can be recovered from the knowledge of the scalar products, as follows:

$$||x|| = \sqrt{< x, x >}$$

Conversely, we can recover the scalar products out of norms, by using the following difficult to remember formula, called complex polarization identity:

$$\begin{aligned} & ||x+y||^2 - ||x-y||^2 + i||x+iy||^2 - i||x-iy||^2 \\ = \ & ||x||^2 + ||y||^2 - ||x||^2 - ||y||^2 + i||x||^2 + i||y||^2 - i||x||^2 - i||y||^2 \\ & +2Re(< x, y >) + 2Re(< x, y >) + 2iIm(< x, y >) + 2iIm(< x, y >) \\ = \ & 4 < x, y > \end{aligned}$$

Finally, we will use Theorem 9.4, and more specifically the following formula coming from there, valid for any matrix $A \in M_N(\mathbb{C})$ and any two vectors $x, y \in \mathbb{C}^N$:

$$< Ax, y >=< x, A^*y >$$

(1) Given a matrix $U \in M_N(\mathbb{C})$, we have indeed the following equivalences, with the first one coming from the polarization identity, and the other ones being clear:

$$\begin{aligned} ||Ux|| = ||x|| &\iff < Ux, Uy > = < x, y > \\ &\iff < x, U^*Uy > = < x, y > \\ &\iff U^*Uy = y \\ &\iff U^*U = 1 \\ &\iff U^* = U^{-1} \end{aligned}$$

(2) Given a matrix $P \in M_N(\mathbb{C})$, in order for $x \to Px$ to be an oblique projection, we must have $P^2 = P$. Now observe that this projection is orthogonal when:

$$\begin{aligned} < Px - x, Py > = 0 &\iff < P^*Px - P^*x, y > = 0 \\ &\iff P^*Px - P^*x = 0 \\ &\iff P^*P - P^* = 0 \\ &\iff P^*P = P^* \end{aligned}$$

The point now is that by conjugating the last formula, we obtain $P^*P = P$. Thus we must have $P = P^*$, and this gives the result. □

Summarizing, the linear operators come in pairs T, T^*, and the associated matrices come as well in pairs A, A^*. We will keep this in mind, and come back to it later.

9b. Matrix inversion

We have seen so far that most of the interesting maps $f : \mathbb{R}^N \to \mathbb{R}^N$ that we know, such as the rotations, symmetries and projections, are linear, and can be written in the following form, with $A \in M_N(\mathbb{R})$ being a square matrix:

$$f(v) = Av$$

We develop now more general theory for such linear maps. We will be interested in the question of inverting the linear maps $f : \mathbb{R}^N \to \mathbb{R}^N$. And the point is that this is the same question as inverting the corresponding matrices $A \in M_N(\mathbb{R})$, due to:

THEOREM 9.6. *A linear map $f : \mathbb{R}^N \to \mathbb{R}^N$, written as*

$$f(v) = Av$$

is invertible precisely when A is invertible, and in this case we have $f^{-1}(v) = A^{-1}v$.

PROOF. This is something that we basically know, coming from the fact that, with the notation $f_A(v) = Av$, we have the following formula:

$$f_A f_B = f_{AB}$$

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

In order to study invertibility questions, for matrices or linear maps, let us begin with some examples. In the simplest case, in 2 dimensions, the result is as follows:

THEOREM 9.7. *We have the following inversion formula, for the 2×2 matrices:*

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

When $ad - bc = 0$, the matrix is not invertible.

PROOF. We have two assertions to be proved, the idea being as follows:

(1) As a first observation, when $ad - bc = 0$ we must have, for some $\lambda \in \mathbb{R}$:

$$b = \lambda a \quad , \quad d = \lambda c$$

Thus our matrix must be of the following special type:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & \lambda a \\ a & \lambda c \end{pmatrix}$$

But in this case the columns are proportional, and so the linear map associated to the matrix is not invertible, and so the matrix itself is not invertible either.

(2) When $ad - bc \neq 0$, let us look for an inversion formula of the following type:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} * & * \\ * & * \end{pmatrix}$$

We must therefore solve the following equations:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} * & * \\ * & * \end{pmatrix} = \begin{pmatrix} ad - bc & 0 \\ 0 & ad - bc \end{pmatrix}$$

The obvious solution here is as follows:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \begin{pmatrix} ad - bc & 0 \\ 0 & ad - bc \end{pmatrix}$$

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

In order to deal now with the inversion problem in general, for the arbitrary matrices $A \in M_N(\mathbb{R})$, we will use the same method as the one above, at $N = 2$. Let us write indeed our matrix as follows, with $v_1, \ldots, v_N \in \mathbb{R}^N$ being its column vectors:

$$A = [v_1, \ldots, v_N]$$

We know from the above that, in order for A to be invertible, the vectors $v_1, \ldots, v_N$ must be linearly independent. Thus, we are led into the question of understanding when a family of vectors $v_1, \ldots, v_N \in \mathbb{R}^N$ are linearly independent. In order to deal with this latter question, let us introduce the following notion:

DEFINITION 9.8. *Associated to any vectors $v_1, \ldots, v_N \in \mathbb{R}^N$ is the volume*

$$\det{}^+(v_1 \ldots v_N) = vol < v_1, \ldots, v_N >$$

of the parallelepiped made by these vectors.

Here the volume is taken in the standard N-dimensional sense. At $N = 1$ this volume is a length, at $N = 2$ this volume is an area, at $N = 3$ this is the usual 3D volume, and so on. In general, the volume of a body $X \subset \mathbb{R}^N$ is by definition the number $vol(X) \in [0, \infty]$ of copies of the unit cube $C \subset \mathbb{R}^N$ which are needed for filling X. Now with this notion in hand, in relation with our inversion problem, we have the following statement:

PROPOSITION 9.9. *The quantity* $\det^+$ *that we constructed, regarded as a function of the corresponding square matrices, formed by column vectors,*

$$\det{}^+ : M_N(\mathbb{R}) \to \mathbb{R}_+$$

has the property that a matrix $A \in M_N(\mathbb{R})$ is invertible precisely when $\det^+(A) > 0$.

PROOF. This follows from the fact that a matrix $A \in M_N(\mathbb{R})$ is invertible precisely when its column vectors $v_1, \ldots, v_N \in \mathbb{R}^N$ are linearly independent. But this latter condition is equivalent to the fact that we must have the following strict inequality:

$$vol < v_1, \ldots, v_N >> 0$$

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

Summarizing, all this leads us into the explicit computation of $\det^+$. As a first observation, in 1 dimension we obtain the absolute value of the real numbers:

$$\det{}^+(a) = |a|$$

In 2 dimensions now, the computation is non-trivial, and we have the following result, making the link with our main result so far, namely Theorem 9.7:

THEOREM 9.10. *In 2 dimensions we have the following formula,*

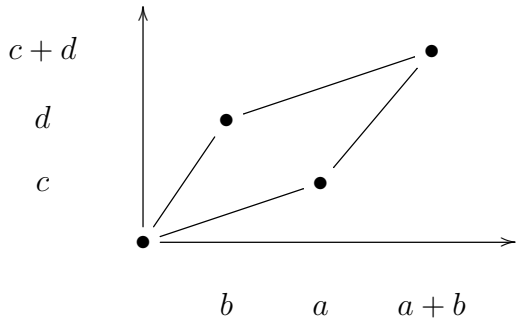
$$\det{}^+ \begin{pmatrix} a & b \\ c & d \end{pmatrix} = |ad - bc|$$

with $\det^+ : M_2(\mathbb{R}) \to \mathbb{R}_+$ *being the function constructed above.*

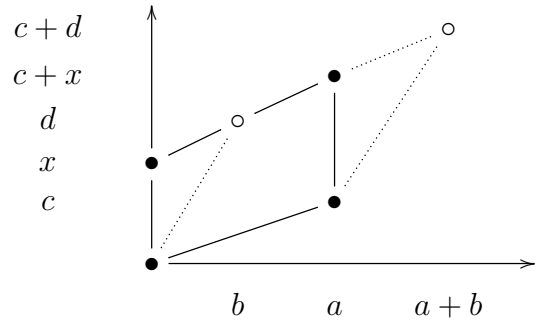
PROOF. We must show that the area of the parallelogram formed by $\binom{a}{c}, \binom{b}{d}$ equals $|ad - bc|$. We can assume $a, b, c, d > 0$ for simplifying, the proof in general being similar. Moreover, by switching if needed the vectors $\binom{a}{c}, \binom{b}{d}$, we can assume that we have:

$$\frac{a}{c} > \frac{b}{d}$$

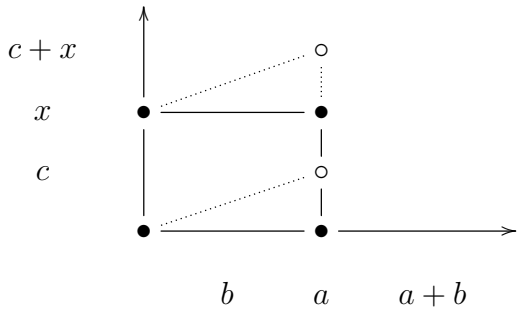
According to these conventions, the picture of our parallelogram is as follows:



Now let us slide the upper side downwards left, until we reach the Oy axis. Our parallelogram, which has not changed its area in this process, becomes:



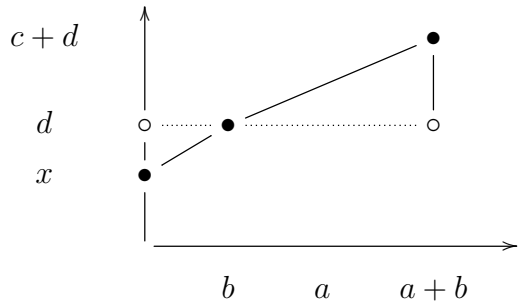
We can further modify this parallelogram, once again by not altering its area, by sliding the right side downwards, until we reach the Ox axis:



Let us compute now the area. Since our two sliding operations have not changed the area of the original parallelogram, this area is given by:

$$A = ax$$

In order to compute the quantity x, observe that in the context of the first move, we have two similar triangles, according to the following picture:



Thus, we are led to the following equation for the number x:

$$\frac{d-x}{b}=\frac{c}{a}$$

By solving this equation, we obtain the following value for x:

$$x=d-\frac{bc}{a}$$

Thus the area of our parallelogram, or rather of the final rectangle obtained from it, which has the same area as the original parallelogram, is given by:

$$ax=ad-bc$$

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

All this is very nice, and obviously we have a beginning of theory here. However, when looking carefully, we can see that our theory has a weakness, because:

(1) In 1 dimension the number a, which is the simplest function of a itself, is certainly a better quantity than the number $|a|$.

(2) In 2 dimensions the number $ad-bc$, which is linear in a,b,c,d, is certainly a better quantity than the number $|ad-bc|$.

So, let us upgrade now our theory, by constructing a better function, which takes signed values. In order to do this, we must come up with a way of splitting the systems of vectors $v_1,\ldots,v_N\in\mathbb{R}^N$ into two classes, call them positive and negative. And here, the answer is quite clear, because a bit of thinking leads to the following definition:

DEFINITION 9.11. *A system of vectors $v_1,\ldots,v_N\in\mathbb{R}^N$ is called:*

(1) *Oriented, if one can continuously pass from the standard basis to it.*
(2) *Unoriented, otherwise.*

The associated sign is $+$ in the oriented case, and $-$ in the unoriented case.

As a first example, in 1 dimension the basis consists of the single vector $e = 1$, which can be continuously deformed into any vector $a > 0$. Thus, the sign is the usual one:

$$sgn(a) = \begin{cases} + & \text{if } a > 0 \\ - & \text{if } a < 0 \end{cases}$$

Thus, in connection with our original question, we are definitely on the good track, because when multiplying $|a|$ by this sign we obtain a itself, as desired:

$$a = sgn(a)|a|$$

In 2 dimensions now, the explicit formula of the sign is as follows:

PROPOSITION 9.12. *We have the following formula, valid for any* 2 *vectors in* $\mathbb{R}^2$*,*

$$sgn\left[\begin{pmatrix} a \\ c \end{pmatrix}, \begin{pmatrix} b \\ d \end{pmatrix}\right] = sgn(ad - bc)$$

with the sign function on the right being the usual one, in 1 *dimension.*

PROOF. According to our conventions, the sign of $\binom{a}{c}, \binom{b}{d}$ is as follows:

(1) The sign is $+$ when these vectors come in this order with respect to the counterclockwise rotation in the plane, around 0.

(2) The sign is $-$ otherwise, meaning when these vectors come in this order with respect to the clockwise rotation in the plane, around 0.

If we assume now $a, b, c, d > 0$ for simplifying, we are left with comparing the angles having the numbers c/a and d/b as tangents, and we obtain in this way:

$$sgn\left[\begin{pmatrix} a \\ c \end{pmatrix}, \begin{pmatrix} b \\ d \end{pmatrix}\right] = \begin{cases} + & \text{if } \frac{c}{a} < \frac{d}{b} \\ - & \text{if } \frac{c}{a} > \frac{d}{b} \end{cases}$$

But this gives the formula in the statement. The proof in general is similar. □

Once again, in connection with our original question, we are on the good track, because when multiplying $|ad - bc|$ by this sign we obtain $ad - bc$ itself, as desired:

$$ad - bc = sgn(ad - bc)|ad - bc|$$

At the level of the general results now, we have:

PROPOSITION 9.13. *The orientation of a system of vectors changes as follows:*

(1) *If we switch the sign of a vector, the associated sign switches.*
(2) *If we permute two vectors, the associated sign switches as well.*

PROOF. Both these assertions are clear from the definition of the sign, because the two operations in question change the orientation of the system of vectors. □

With the above notion in hand, we can now formulate:

DEFINITION 9.14. *The determinant of $v_1, \ldots, v_N \in \mathbb{R}^N$ is the signed volume*

$$\det(v_1 \ldots v_N) = \pm vol < v_1, \ldots, v_N >$$

of the parallelepiped made by these vectors.

In other words, we are upgrading here Definition 9.8, by adding a sign to the quantity $\det^+$ constructed there, as to potentially reach to good additivity properties:

$$\det(v_1 \ldots v_N) = \pm \det^+(v_1 \ldots v_N)$$

In relation with our original inversion problem for the square matrices, this upgrade does not change what we have so far, and we have the following statement:

THEOREM 9.15. *The quantity* det *that we constructed, regarded as a function of the corresponding square matrices, formed by column vectors,*

$$\det : M_N(\mathbb{R}) \to \mathbb{R}$$

has the property that a matrix $A \in M_N(\mathbb{R})$ is invertible precisely when $\det(A) \neq 0$.

PROOF. We know from the above that a matrix $A \in M_N(\mathbb{R})$ is invertible precisely when $\det^+(A) = |\det A|$ is strictly positive, and this gives the result. □

Let us try now to compute the determinant. In 1 dimension we have of course the formula $\det(a) = a$, because the absolute value fits, and so does the sign:

$$\det(a) = sgn(a) \times |a| = a$$

In 2 dimensions now, we have the following result:

THEOREM 9.16. *In 2 dimensions we have the following formula,*

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

with $|\,.\,| = \det$ being the determinant function constructed above.

PROOF. According to our definition, to the computation in Theorem 9.10, and to the sign formula from Proposition 9.12, the determinant of a 2×2 matrix is given by:

$$\begin{aligned} \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} &= sgn \left[\begin{pmatrix} a \\ c \end{pmatrix}, \begin{pmatrix} b \\ d \end{pmatrix} \right] \times \det^+ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \\ &= sgn \left[\begin{pmatrix} a \\ c \end{pmatrix}, \begin{pmatrix} b \\ d \end{pmatrix} \right] \times |ad - bc| \\ &= sgn(ad - bc) \times |ad - bc| \\ &= ad - bc \end{aligned}$$

Thus, we have obtained the formula in the statement. □

9c. The determinant

In order to discuss now arbitrary dimensions, we will need a number of theoretical results. Here is a first series of formulae, coming straight from the definitions:

THEOREM 9.17. *The determinant has the following properties:*

(1) *When multiplying by scalars, the determinant gets multiplied as well:*

$$\det(\lambda_1 v_1, \ldots, \lambda_N v_N) = \lambda_1 \ldots \lambda_N \det(v_1, \ldots, v_N)$$

(2) *When permuting two columns, the determinant changes the sign:*

$$\det(\ldots, u, \ldots, v, \ldots) = -\det(\ldots, v, \ldots, u, \ldots)$$

(3) *The determinant* $\det(e_1, \ldots, e_N)$ *of the standard basis of* $\mathbb{R}^N$ *is* 1.

PROOF. All this is clear from definitions, as follows:

(1) This follows from definitions, and from Proposition 9.13 (1).

(2) This follows as well from definitions, and from Proposition 9.13 (2).

(3) This is clear from our definition of the determinant. □

As an application of the above result, we have:

THEOREM 9.18. *The determinant of a diagonal matrix is given by:*

$$\begin{vmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{vmatrix} = \lambda_1 \ldots \lambda_N$$

That is, we obtain the product of diagonal entries, or of eigenvalues.

PROOF. The formula in the statement is clear by using Theorem 9.17, which gives:

$$\begin{vmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{vmatrix} = \lambda_1 \ldots \lambda_N \begin{vmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{vmatrix} = \lambda_1 \ldots \lambda_N$$

As for the last assertion, this is rather a remark. □

In order to reach to a more advanced theory, let us adopt now the linear map point of view. In this setting, the definition of the determinant reformulates as follows:

THEOREM 9.19. *Given a linear map, written as* $f(v) = Av$*, its "inflation coefficient", obtained as the signed volume of the image of the unit cube, is given by:*

$$I_f = \det A$$

More generally, I_f *is the inflation ratio of any parallelepiped in* $\mathbb{R}^N$*, via the transformation* f*. In particular* f *is invertible precisely when* $\det A \neq 0$.

PROOF. The only non-trivial thing in all this is the fact that the inflation coefficient I_f, as defined above, is independent of the choice of the parallelepiped. But this is a generalization of the Thales theorem, which follows from the Thales theorem itself. □

As a first application of the above linear map viewpoint, we have:

THEOREM 9.20. *We have the following formula, valid for any matrices A, B:*

$$\det(AB) = \det A \cdot \det B$$

In particular, we have $\det(AB) = \det(BA)$.

PROOF. The first formula follows from the formula $f_{AB} = f_A f_B$ for the associated linear maps. As for $\det(AB) = \det(BA)$, this is clear from the first formula. □

Getting back now to explicit computations, we have the following key result:

THEOREM 9.21. *The determinant of a diagonalizable matrix*

$$A \sim \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix}$$

is the product of its eigenvalues, $\det A = \lambda_1 \ldots \lambda_N$.

PROOF. We know that a diagonalizable matrix can be written in the form $A = PDP^{-1}$, with $D = diag(\lambda_1, \ldots, \lambda_N)$. Now by using Theorem 9.20, we obtain:

$$\begin{aligned} \det A &= \det(PDP^{-1}) \\ &= \det(DP^{-1}P) \\ &= \det D \\ &= \lambda_1 \ldots \lambda_N \end{aligned}$$

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

In general now, at the theoretical level, we have the following key result:

THEOREM 9.22. *The determinant has the additivity property*

$$\det(\ldots, u + v, \ldots) = \det(\ldots, u, \ldots) + \det(\ldots, v, \ldots)$$

valid for any choice of the vectors involved.

PROOF. This follows by doing some elementary geometry, in the spirit of the computations in the proof of Theorem 9.10, as follows:

(1) We can either use the Thales theorem, and then compute the volumes of all the parallelepipeds involved, by using basic algebraic formulae.

(2) Or we can solve the problem in "puzzle" style, the idea being to cut the big parallelepiped, and then recover the small ones, after some manipulations.

(3) We can do as well something hybrid, consisting in deforming the parallelepipeds involved, without changing their volumes, and then cutting and gluing. □

As a basic application of the above result, we have:

THEOREM 9.23. *We have the following results:*

(1) *The determinant of a diagonal matrix is the product of diagonal entries.*
(2) *The same is true for the upper triangular matrices.*
(3) *The same is true for the lower triangular matrices.*

PROOF. All this can be deduced by using our various general formulae, as follows:

(1) This is something that we already know, from Theorem 9.21.

(2) This follows by using our various formulae, then (1), as follows:

$$\begin{vmatrix} \lambda_1 & & & * \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{vmatrix} = \begin{vmatrix} \lambda_1 & 0 & & * \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{vmatrix}$$

$$\vdots$$

$$\vdots$$

$$= \begin{vmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{vmatrix}$$

$$= \lambda_1 \ldots \lambda_N$$

(3) This follows as well from our various formulae, then (1), by proceeding this time from right to left, from the last column towards the first column. □

As an important theoretical result now, we have:

THEOREM 9.24. *The determinant of square matrices is the unique map*

$$\det : M_N(\mathbb{R}) \to \mathbb{R}$$

satisfying the conditions found above.

PROOF. Any map $\det' : M_N(\mathbb{R}) \to \mathbb{R}$ satisfying our conditions must indeed coincide with det on the upper triangular matrices, and then all the matrices. □

Here is now another important theoretical result:

THEOREM 9.25. *The determinant is subject to the row expansion formula*

$$\begin{vmatrix} a_{11} & \ldots & a_{1N} \\ \vdots & & \vdots \\ a_{N1} & \ldots & a_{NN} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & \ldots & a_{2N} \\ \vdots & & \vdots \\ a_{N2} & \ldots & a_{NN} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} & \ldots & a_{2N} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N3} & \ldots & a_{NN} \end{vmatrix}$$
$$+ \ldots\ldots + (-1)^{N+1} a_{1N} \begin{vmatrix} a_{21} & \ldots & a_{2,N-1} \\ \vdots & & \vdots \\ a_{N1} & \ldots & a_{N,N-1} \end{vmatrix}$$

and this method fully computes it, by recurrence.

PROOF. This follows from the fact that the formula in the statement produces a certain function $\det : M_N(\mathbb{R}) \to \mathbb{R}$, which has the 4 properties in Theorem 9.24. □

We can expand as well over the columns, as follows:

THEOREM 9.26. *The determinant is subject to the column expansion formula*

$$\begin{vmatrix} a_{11} & \ldots & a_{1N} \\ \vdots & & \vdots \\ a_{N1} & \ldots & a_{NN} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & \ldots & a_{2N} \\ \vdots & & \vdots \\ a_{N2} & \ldots & a_{NN} \end{vmatrix} - a_{21} \begin{vmatrix} a_{12} & \ldots & a_{1N} \\ a_{32} & \ldots & a_{3N} \\ \vdots & & \vdots \\ a_{N2} & \ldots & a_{NN} \end{vmatrix}$$
$$+ \ldots\ldots + (-1)^{N+1} a_{N1} \begin{vmatrix} a_{12} & \ldots & a_{1N} \\ \vdots & & \vdots \\ a_{N-1,2} & \ldots & a_{N-1,N} \end{vmatrix}$$

and this method fully computes it, by recurrence.

PROOF. This follows by using the same argument as for the rows. □

As a first application of the above methods, we can now prove:

THEOREM 9.27. *The determinant of the 3×3 matrices is given by*

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - ceg - bdi - afh$$

which can be memorized by using Sarrus' triangle method, "triangles parallel to the diagonal, minus triangles parallel to the antidiagonal".

PROOF. Here is the computation, using the above results:

$$\begin{aligned}\begin{vmatrix}a&b&c\\d&e&f\\g&h&i\end{vmatrix} &= a\begin{vmatrix}e&f\\h&i\end{vmatrix}-b\begin{vmatrix}d&f\\g&i\end{vmatrix}+c\begin{vmatrix}d&e\\g&h\end{vmatrix}\\ &= a(ei-fh)-b(di-fg)+c(dh-eg)\\ &= aei-afh-bdi+bfg+cdh-ceg\\ &= aei+bfg+cdh-ceg-bdi-afh\end{aligned}$$

Thus, we obtain the formula in the statement. □

Let us discuss now the general formula of the determinant, at arbitrary values $N \in \mathbb{N}$ of the matrix size, generalizing the formulae that we have at $N = 2, 3$. We will need:

DEFINITION 9.28. *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.

There are many possible notations for the permutations, the simplest one consisting in writing the numbers $1, \ldots, N$, and below them, their permuted versions:

$$\sigma = \begin{pmatrix}1&2&3&4&5\\2&1&4&5&3\end{pmatrix}$$

Another method, which is better for most purposes, and faster too, remember that time is money, is by denoting permutations as diagrams, going from top to bottom:

$\sigma =$ 

There are many interesting things that can be said about permutations. In what concerns us, we will need the following key result:

THEOREM 9.29. *The permutations have a signature function*

$$\varepsilon : S_N \to \{\pm 1\}$$

which can be defined in the following equivalent ways:

(1) *As $(-1)^c$, where c is the number of inversions.*
(2) *As $(-1)^t$, where t is the number of transpositions.*
(3) *As $(-1)^o$, where o is the number of odd cycles.*
(4) *As $(-1)^x$, where x is the number of crossings.*
(5) *As the sign of the corresponding permuted basis of $\mathbb{R}^N$.*

PROOF. Let us begin with the precise definition of c, t, o, x, as numbers modulo 2:

(1) The idea here is that given any two numbers $i < j$ among $1, \ldots, N$, the permutation can either keep them in the same order, $\sigma(i) < \sigma(j)$, or invert them:

$$\sigma(j) > \sigma(i)$$

Now by making $i < j$ vary over all pairs of numbers in $1, \ldots, N$, we can count the number of inversions, and call it c. This is an integer, $c \in \mathbb{N}$, which is well-defined.

(2) Here the idea, which is something quite intuitive, is that any permutation appears as a product of switches, also called transpositions:

$$i \leftrightarrow j$$

The decomposition as a product of transpositions is not unique, but the number t of the needed transpositions is unique, when considered modulo 2. This follows for instance from the equivalence of (2) with (1,3,4,5), explained below.

(3) Here the point is that any permutation decomposes, in a unique way, as a product of cycles, which are by definition permutations of the following type:

$$i_1 \to i_2 \to i_3 \to \ldots\ldots \to i_k \to i_1$$

Some of these cycles have even length, and some others have odd length. By counting those having odd length, we obtain a well-defined number $o \in \mathbb{N}$.

(4) Here the method is that of drawing the permutation, as we usually do, and by avoiding triple crossings, and then counting the number of crossings. This number x depends on the way we draw the permutations, but modulo 2, we always get the same number. Indeed, this follows from the fact that we can continuously pass from a drawing to each other, and that when doing so, the number of crossings can only jump by ± 2.

Summarizing, we have 4 different definitions for the signature of the permutations, which all make sense, constructed according to (1-4) above. Regarding now the fact that we always obtain the same number, this can be established as follows:

(1)=(2) This is clear, because any transposition inverts once, modulo 2.

(1)=(3) This is clear as well, because the odd cycles invert once, modulo 2.

(1)=(4) This comes from the fact that the crossings correspond to inversions.

(2)=(3) This follows by decomposing the cycles into transpositions.

(2)=(4) This comes from the fact that the crossings correspond to transpositions.

(3)=(4) This follows by drawing a product of cycles, and counting the crossings.

Finally, in what regards the equivalence of all these constructions with (5), here simplest is to use (2). Indeed, we already know that the sign of a system of vectors switches when interchanging two vectors, and so the equivalence between (2,5) is clear. □

Now back to linear algebra, we can formulate a key result, as follows:

THEOREM 9.30. *We have the following formula for the determinant,*

$$\det A = \sum_{\sigma\in S_N} \varepsilon(\sigma) A_{1\sigma(1)} \ldots A_{N\sigma(N)}$$

with the signature function being the one introduced above.

PROOF. This follows by recurrence over $N \in \mathbb{N}$, as follows:

(1) When developing the determinant over the first column, we obtain a signed sum of N determinants of size $(N-1)\times(N-1)$. But each of these determinants can be computed by developing over the first column too, and so on, and we are led to the conclusion that we have a formula as in the statement, with $\varepsilon(\sigma) \in \{-1, 1\}$ being certain coefficients.

(2) But these latter coefficients $\varepsilon(\sigma) \in \{-1, 1\}$ can only be the signatures of the corresponding permutations $\sigma \in S_N$, with this being something that can be viewed again by recurrence, with either of the definitions (1-5) in Theorem 9.29 for the signature. □

The above result is something quite tricky, and in order to get familiar with it, there is nothing better than doing some computations. As a first, basic example, in 2 dimensions we recover the usual formula of the determinant, the details being as follows:

$$\begin{aligned}\begin{vmatrix} a & b \\ c & d \end{vmatrix} &= \varepsilon(|\,|) \cdot ad + \varepsilon(\diagup\!\diagdown) \cdot cb \\ &= 1 \cdot ad + (-1) \cdot cb \\ &= ad - bc\end{aligned}$$

In 3 dimensions now, we recover the Sarrus formula:

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - ceg - bdi - afh$$

Let us record as well the following general formula, which was not obvious with our previous theory of the determinant, but which follows from Theorem 9.30:

$$\det(A) = \det(A^t)$$

There are countless other applications of the formula in Theorem 9.30, and we will be back to this, on several occassions. But, most importantly, that formula allows us to deal now with the complex matrices too, by formulating the following statement:

THEOREM 9.31. *If we define the determinant of a complex matrix $A \in M_N(\mathbb{C})$ to be*

$$\det A = \sum_{\sigma\in S_N} \varepsilon(\sigma) A_{1\sigma(1)} \ldots A_{N\sigma(N)}$$

then this determinant has the same properties as the determinant of the real matrices.

PROOF. This follows by doing some sort of reverse engineering, with respect to what has been done in this section, and we reach to the conclusion that det has indeed all the good properties that we are familiar with. Except of course for the properties at the very beginning of this section, in relation with volumes, which don't extend well to $\mathbb{C}^N$. □

Good news, this is the end of the general theory that we wanted to develop. We have now in our bag all the needed techniques for computing the determinant.

9d. Diagonalization

Let us discuss now the diagonalization question for linear maps and matrices. The basic diagonalization theory, formulated in terms of matrices, is as follows:

PROPOSITION 9.32. *A vector $v \in \mathbb{C}^N$ is called eigenvector of $A \in M_N(\mathbb{C})$, with corresponding eigenvalue λ, when A multiplies by λ in the direction of v:*

$$Av = \lambda v$$

In the case where $\mathbb{C}^N$ 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:

$$PDP^{-1}v_i = PDe_i = P\lambda_i e_i = \lambda_i P e_i = \lambda_i v_i$$

Thus, the matrices A and PDP^{-1} coincide, as stated. □

In order to study the diagonalization problem, the idea is that the eigenvectors can be grouped into linear spaces, called eigenspaces, as follows:

THEOREM 9.33. *Let $A \in M_N(\mathbb{C})$, and for any eigenvalue $\lambda \in \mathbb{C}$ define the corresponding eigenspace as being the vector space formed by the corresponding eigenvectors:*

$$E_\lambda = \left\{ v \in \mathbb{C}^N \middle| Av = \lambda v \right\}$$

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 $\sum_\lambda \dim(E_\lambda) \leq N$, with the sum being over all the eigenvalues, 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. On the left we obtain:

$$Av_k = \lambda_k v_k = \lambda_k c_1 v_1 + \ldots + \lambda_k c_{k-1} v_{k-1}$$

On the right we obtain something different, as follows:

$$\begin{aligned} A(c_1 v_1 + \ldots + c_{k-1} v_{k-1}) &= c_1 A v_1 + \ldots + c_{k-1} A v_{k-1} \\ &= c_1 \lambda_1 v_1 + \ldots + c_{k-1} \lambda_{k-1} v_{k-1} \end{aligned}$$

We conclude from this that the following equality must hold:

$$\lambda_k c_1 v_1 + \ldots + \lambda_k c_{k-1} v_{k-1} = c_1 \lambda_1 v_1 + \ldots + c_{k-1} \lambda_{k-1} v_{k-1}$$

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 right and at left:

$$\lambda_k c_1 = c_1 \lambda_1$$

$$\vdots$$

$$\lambda_k c_{k-1} = c_{k-1} \lambda_{k-1}$$

Now since at least one c_i must be nonzero, from $\lambda_k c_i = c_i \lambda_i$ we obtain $\lambda_k = \lambda_i$, which is a contradiction. Thus our proof by recurrence of the first assertion is complete. As for the second assertion, this follows from the first one. □

In order to reach now to more advanced results, we can use the characteristic polynomial, which appears via the following fundamental result:

THEOREM 9.34. *Given a matrix $A \in M_N(\mathbb{C})$, consider its characteristic polynomial:*

$$P(x) = \det(A - x1_N)$$

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 $\mathbb{C}^N$, 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. □

Now recall that we are over $\mathbb{C}$. We obtain the following result:

THEOREM 9.35. *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,

$$P(X) = (-1)^N (X - \lambda_1)^{n_1} \dots (X - \lambda_k)^{n_k}$$

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 the above results. By summing the inequalities $\dim(E_\lambda) \leq m_\lambda$ from Theorem 9.34, 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 Theorem 9.33 that our matrix is diagonalizable when we have global equality. Thus, we are led to the conclusion in the statement. □

As an illustration for all this, which is a must-know computation, we have:

PROPOSITION 9.36. *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 indicated, 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. □

At the level of basic examples of diagonalizable matrices, we first have the following result, which provides us with the "generic" examples:

THEOREM 9.37. *For a matrix $A \in M_N(\mathbb{C})$ the following conditions are equivalent,*

(1) *The eigenvalues are different, $\lambda_i \neq \lambda_j$,*
(2) *The characteristic polynomial P has simple roots,*
(3) *The characteristic polynomial satisfies $(P, P') = 1$,*
(4) *The resultant of P, P' is nonzero, $R(P, P') \neq 0$,*
(5) *The discriminant of P is nonzero, $\Delta(P) \neq 0$,*

and in this case, the matrix is diagonalizable.

PROOF. The last assertion holds indeed, due to Theorem 9.35. As for the equivalences in the statement, these are all standard, by using the R, Δ theory from chapter 5. □

As already mentioned, one can prove that the matrices having distinct eigenvalues are "generic", and so the above result basically captures the whole situation. We have in fact the following collection of density results, which are quite advanced:

THEOREM 9.38. *The following happen, inside $M_N(\mathbb{C})$:*

(1) *The invertible matrices are dense.*
(2) *The matrices having distinct eigenvalues are dense.*
(3) *The diagonalizable matrices are dense.*

PROOF. These are quite advanced results, which can be proved as follows:

(1) This is clear, intuitively speaking, because the invertible matrices are given by the condition $\det A \neq 0$. Thus, the set formed by these matrices appears as the complement of the surface $\det A = 0$, and so must be dense inside $M_N(\mathbb{C})$, as claimed.

(2) Here we can use a similar argument, this time by saying that the set formed by the matrices having distinct eigenvalues appears as the complement of the surface given by $\Delta(P_A) = 0$, and so must be dense inside $M_N(\mathbb{C})$, as claimed.

(3) This follows from (2), via the fact that the matrices having distinct eigenvalues are diagonalizable, that we know from Theorem 9.37. There are of course some other proofs as well, for instance by putting the matrix in Jordan form. □

This was for the basic theory of the linear maps, that we will need in what follows. There are of course far more things that can be said, and we will be back to this gradually, when needed, and notably in chapter 12 below, when doing advanced theory.

9e. Exercises

This was an easy chapter, and start of something new, at least when compared with the horrors from chapters 7-8, and here are some easy exercises for you:

EXERCISE 9.39. *Work out all the details for the main properties of the determinant function* $\det : M_N(\mathbb{R}) \to \mathbb{R}$*, by using our geometric approach, and Thales.*

EXERCISE 9.40. *Prove that for* $H \in M_N(\pm 1)$ *we have* $|\det H| \leq N^{N/2}$*, with equality precisely when* H *is Hadamard, in the sense that its rows are pairwise orthogonal.*

EXERCISE 9.41. *Work out all the details for the properties of the determinant function* $\det : M_N(\mathbb{C}) \to \mathbb{C}$*, by using our algebraic approach, and permutations.*

EXERCISE 9.42. *Learn the Jordan form, and come up with an alternative proof for our result stating that the diagonalizable matrices are dense.*

As an important bonus exercise, diagonalize 3×3 matrices, that you can find on the internet, as many as needed, as for the computation to take 20 minutes.

CHAPTER 10

Partial derivatives

10a. Functions, continuity

We have seen so far the theory of linear maps $f : \mathbb{R}^N \to \mathbb{R}^M$, and $f : \mathbb{C}^N \to \mathbb{C}^M$. In what follows we develop the theory of maps $f : \mathbb{R}^N \to \mathbb{R}^M$ and $f : \mathbb{C}^N \to \mathbb{C}^M$ in general, in analogy with what we know about the maps $f : \mathbb{R} \to \mathbb{R}$, from Part I of this book, and with what we know about the maps $f : \mathbb{C} \to \mathbb{C}$, from Part II of this book.

We will rely of course as well on the substantial help from the linear algebra that we learned in chapter 9, which more or less solves all our potential questions regarding analysis, when the map f in question happens to be linear, of the following form:

$$f(x) = Ax$$

However, things in general will be quite tricky, among others because a map $f : \mathbb{C} \to \mathbb{C}$ can be regarded as a map $f : \mathbb{R}^2 \to \mathbb{R}^2$, leading to an obvious dimension mess. So, in order not to mess up things with dimensions, we will first study the real case, that of the maps $f : \mathbb{R}^N \to \mathbb{R}^M$, with minimal reference to the complex numbers, and study the maps $f : \mathbb{C}^N \to \mathbb{C}^M$ only afterwards. And with the remark that this plan, while certainly reasonable, will stumble at some point into some difficulties, coming from the fact that a linear map $f : \mathbb{R}^N \to \mathbb{R}^N$ can have complex eigenvalues, and so staying over $\mathbb{R}$ is not enough, for fully understanding it. Welcome to the several variables mess.

But probably enough talking, let us get started. As a first objective, we would like to talk about the continuity, and other basic analytic properties, of the maps $f : \mathbb{R}^N \to \mathbb{R}^M$. And for this, we will run all the time into the formula of the distance in $\mathbb{R}^N$, namely:

$$d(x, y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$

In order to avoid using all the time this formula, which quite often can lead into complicated computations, and even into following wrong paths, it is convenient to relax a bit, and take an abstract point of view on all this.

So, let us begin by axiomatizing the properties of the distance $d(x, y)$ given above, by generalizing what we know in $\mathbb{R}^N$. This leads us into the following notion:

DEFINITION 10.1. *A metric space is a set X with a distance function $d : X \times X \to \mathbb{R}_+$, having 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) $d(x, y) \leq d(x, z) + d(y, z)$.

As a basic example, we have $\mathbb{R}^N$, as well as any of its subsets $X \subset \mathbb{R}^N$. Indeed, the first two axioms are clear, and for the third axiom, we must prove that:

$$\sqrt{\sum_i (x_i - y_i)^2} \leq \sqrt{\sum_i (x_i - z_i)^2} + \sqrt{\sum_i (y_i - z_i)^2}$$

But with $a = x - z$ and $b = y - z$, this is the same as proving that:

$$\sqrt{\sum_i (a_i + b_i)^2} \leq \sqrt{\sum_i a_i^2} + \sqrt{\sum_i b_i^2}$$

Moreover, by raising to the square, this is the same as proving that:

$$\left(\sum_i a_i b_i\right)^2 \leq \left(\sum_i a_i^2\right)\left(\sum_i b_i^2\right)$$

But this latter inequality is one of the many equivalent formulations of the Cauchy-Schwarz inequality, that we know well from the above, and which follows by using the fact that $f(t) = \sum_i (a_i + tb_i)^2$ being positive, its discriminant must be negative.

As another example, we have $\mathbb{C}^N$, as well as any of its subsets $X \subset \mathbb{C}^N$. Indeed, this follows either from $\mathbb{C}^N \simeq \mathbb{R}^{2N}$, which is an isomorphism of metric spaces, or directly, along the lines of the above proof for $\mathbb{R}^N$, by making changes where needed. To be more precise, after doing the algebra, we are left with proving the following inequality:

$$\left|\sum_i a_i \bar{b}_i\right|^2 \leq \left(\sum_i |a_i|^2\right)\left(\sum_i |b_i|^2\right)$$

But this is the complex version of the Cauchy-Schwarz inequality, that we know well from the above, and which follows by using the fact that the function $f(t) = \sum_i |a_i + twb_i|^2$ with $t \in \mathbb{R}$ and $|w| = 1$ being positive, its discriminant must be negative.

More generally, we have seen in chapter 7 that the basic properties of the distance on $\mathbb{R}^N$ and $\mathbb{C}^N$ have some natural extensions to certain infinite dimensional spaces, real or complex, consisting of certain classes of functions. We will be back to this.

Here is now another example, which at first looks new and interesting, but is in fact not new, because it appears as a subspace of a suitable $\mathbb{R}^N$:

PROPOSITION 10.2. *Given a finite set X, the following function is a metric on it, called discrete metric:*

$$d(x,y)=\begin{cases}1 & \text{if } x\neq y\\ 0 & \text{if } x=y\end{cases}$$

This metric space is in fact the N-simplex, which can be realized as a subspace of $\mathbb{R}^{N-1}$, or, more conveniently, as a subspace of $\mathbb{R}^N$.

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

(1) First of all, all the axioms from Definition 10.1 are trivially satisfied, and with the main axiom, namely the triangle inequality, basically coming from:

$$1+1\leq 1$$

(2) At the level of examples, at $|X|=1$ we obtain a point, at $|X|=2$ we obtain a segment, at $|X|=3$ we obtain an equilateral triangle, at $|X|=4$ we obtain a regular tetrahedron, and so on. Thus, what we have in general, at $|X|=N$, is the arbitrary dimensional generalization of this series of geometric objects, called the N-simplex.

(3) In what regards now the geometric generalization of the N-simplex, our above examples, namely segment, triangle, tetrahedron and so on, suggest to look for an embedding $X\subset\mathbb{R}^{N-1}$. Which is something which is certainly possible, but the computations here are quite complicated, involving a lot of trigonometry, as you can check yourself by studying the problem at $N=4$, that is, parametrizing the regular tetrahedron in $\mathbb{R}^3$.

(4) However, mathematics, or perhaps physics come to the rescue, via the idea "add a dimension, for getting smarter". Indeed, when looking for an embedding $X\subset\mathbb{R}^N$ things drastically simplify, because we can simply take X to be the standard basis of $\mathbb{R}^N$:

$$X=\{e_1,\ldots,e_N\}$$

Indeed, we have by definition $d(e_i,e_j)=1$ for any $i\neq j$. So, we have solved our embedding problem, just like that, without doing any computations or trigonometry. □

Getting back now to Definition 10.1 as it is, the axioms there are satisfied as well for the various spaces of infinite dimensions that we met in chapter 7. Thus, the whole discussion following Definition 10.1 generalizes to arbitrary dimensions. Also, so does Proposition 10.2, to the case of the sets X of arbitrary cardinality, and with the realization of the corresponding metric space being again by using a standard basis, as follows:

$$X=\{e_x\}_{x\in X}\subset l^2(X)$$

However, in what follows we will not insist much on these latter examples, because we will be mainly interested, as planned before, in $\mathbb{R}^N,\mathbb{C}^N$ and their subspaces.

Moving ahead now with some theory, and allowing us a bit of slopiness, we have:

PROPOSITION 10.3. *We can talk about limits inside metric spaces X, by saying that*

$$x_n \to x \iff d(x_n, x) \to 0$$

and we can talk as well about continuous functions $f : X \to Y$, by requiring that

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

and with these notions in hand, all the basic results from the cases $X = \mathbb{R}, \mathbb{C}$ extend.

PROOF. All this is very standard, and we will leave this as an exercise, namely carefully checking what we did so far in this book, in relation with limits and continuity, in the cases $X = \mathbb{R}, \mathbb{C}$, and working out the metric space extensions of this. Of course I can hear you screaming that this is too much work, but believe me, after some thinking, there is in fact not much work to be done. Indeed, all that we have been doing of advanced type requires sums $x+y$ or multiplication by scalars λx, and such operations being not allowed in the general context of Definition 10.1, we are just left with a handful of trivialities, that you can surely work out, as a quick and instructive exercise. □

More interestingly now, we can talk about open and closed sets inside metric spaces X, again in analogy with what we did for $X = \mathbb{R}, \mathbb{C}$, but with a whole lot of interesting new phenomena appearing. So, we will do this in detail. Let us start with:

DEFINITION 10.4. *Let X be a metric space.*

(1) *The open balls are the sets $B_x(r) = \{y \in X | d(x, y) < r\}$.*
(2) *The closed balls are the sets $\bar{B}_x(r) = \{y \in X | d(x, y) \leq r\}$.*
(3) *$E \subset X$ is called open if for any $x \in E$ we have a ball $B_x(r) \subset E$.*
(4) *$E \subset X$ is called closed if its complement $E^c \subset X$ is open.*

At the level of examples, you can quickly convince yourself, by working out a few of them, that our notions above coincide with the usual ones, that we know well, in the case $X = \mathbb{R}, \mathbb{C}$. We will be back to this later, with some general results in this sense, confirming all this. But for the moment, let us work out the basics. We first have the following result, clarifying some terminology issues from Definition 10.4:

PROPOSITION 10.5. *The open balls are open, and the closed balls are closed.*

PROOF. This might sound a bit as a joke, but it is not one, because this is the kind of thing that we have to check. Fortunately, all this is elementary, as follows:

(1) Given an open ball $B_x(r)$ and a point $y \in B_x(r)$, by using the triangle inequality we have $B_y(r') \subset B_x(r)$, with $r' = r - d(x, y)$. Thus, $B_x(r)$ is indeed open.

(2) Given a closed ball $\bar{B}_x(r)$ and a point $y \in B_x(r)^c$, by using the triangle inequality we have $B_y(r') \subset B_x(r)^c$, with $r' = d(x, y) - r$. Thus, $\bar{B}_x(r)$ is indeed closed. □

Here is now something more interesting, making the link with our intuitive understanding of the notion of closedness, coming from our experience so far with analysis:

THEOREM 10.6. *For a set $E \in X$, the following are equivalent:*

(1) *E is closed in our sense, meaning that E^c is open.*
(2) *We have $x_n \to x, x_n \in E \implies x \in E$.*

PROOF. We can prove this by double implication, as follows:

(1) $\implies$ (2) Assume by contradiction $x_n \to x, x_n \in E$ with $x \notin E$. Since we have $x \in E^c$, which is open, we can pick a ball $B_x(r) \subset E^c$. But this contradicts our convergence assumption $x_n \to x$, so we are done with this implication.

(2) $\implies$ (1) Assume by contradiction that E is not closed in our sense, meaning that E^c is not open. Thus, we can find $x \in E^c$ such that there is no ball $B_x(r) \subset E^c$. But with $r = 1/n$ this provides us with a point $x_n \in B_x(1/n) \cap E$, and since we have $x_n \to x$, this contradicts our assumption (2). Thus, we are done with this implication too. □

Here is another basic theorem about open and closed sets:

THEOREM 10.7. *Let X be a metric space.*

(1) *If E_i are open, then $\cup_i E_i$ is open.*
(2) *If F_i are closed, then $\cap_i F_i$ is closed.*
(3) *If $E_1, \ldots, E_n$ are open, then $\cap_i E_i$ is open.*
(4) *If $F_1, \ldots, F_n$ are closed, then $\cup_i F_i$ is closed.*

Moreover, both (3) and (4) can fail for infinite intersections and unions.

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

(1) This is clear from definitions, because any point $x \in \cup_i E_i$ must satisfy $x \in E_i$ for some i, and so has a ball around it belonging to E_i, and so to $\cup_i E_i$.

(2) This follows from (1), by using the following well-known set theory formula:

$$\left(\bigcup_i E_i \right)^c = \bigcap_i E_i^c$$

(3) Given an arbitrary point $x \in \cap_i E_i$, we have $x \in E_i$ for any i, and so we have a ball $B_x(r_i) \subset E_i$ for any i. Now with this in hand, let us set:

$$B = B_x(r_1) \cap \ldots \cap B_x(r_n)$$

As a first observation, this is a ball around x, $B = B_x(r)$, of radius given by:

$$r = \min(r_1, \ldots, r_n)$$

But this ball belongs to all the E_i, and so belongs to their intersection $\cap_i E_i$. We conclude that the intersection $\cap_i E_i$ is open, as desired.

(4) This follows from (3), by using the following well-known set theory formula:

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c$$

(5) Finally, in what regards the counterexamples at the end, these can be both found on $\mathbb{R}$. Indeed, for the infinite intersections of open sets, we can use:

$$\bigcap_n \left(-\frac{1}{n}\ ,\ \frac{1}{n}\right) = \{0\}$$

As for the infinite unions of closed sets, here we can use:

$$\bigcup_n \left[0\ ,\ 1-\frac{1}{n}\right] = [0,1)$$

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

Finally, still in relation with open and closed sets, we have as well:

DEFINITION 10.8. *Let X be a metric space, and $E \subset X$ be a subset.*

(1) *The interior $E^\circ \subset E$ is the set of points $x \in E$ which admit around them open balls $B_x(r) \subset E$.*
(2) *The closure $E \subset \bar{E}$ is the set of points $x \in X$ which appear as limits of sequences $x_n \to x$, with $x \in E$.*

These notions are quite interesting, because they make sense for any set E. That is, when E is open, that is open and end of the story, and when E is closed, that is closed and end of the story. In general, however, a set $E \subset X$ is not open or closed, and what we can best do to it, in order to study with our tools, is to "squeeze" it, as follows:

$$E^\circ \subset E \subset \bar{E}$$

In practice now, in order to use the above notions, we need to know a number of things, including that fact that E open implies $E^\circ = E$, the fact that E closed implies $\bar{E} = E$, and many more such results, not to forget the fact that the closures of the open balls $B_r(x)$ are the closed balls $\bar{B}_x(r)$, clarifying an obvious notational issue which appears with respect to Definition 10.4. But all this can be done, and the useful statement here, summarizing all that we need to know about interiors and closures, is as follows:

THEOREM 10.9. *Let X be a metric space, and $E \subset X$ be a subset.*

(1) *The interior $E^\circ \subset E$ is the biggest open set contained in E.*
(2) *The closure $E \subset \bar{E}$ is the smallest closed set containing E.*

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

(1) Let us first prove that the interior E° is open. For this purpose, pick $x \in E^\circ$. We know that we have a ball $B_x(r) \subset E$, and since this ball is open, it follows that we have $B_x(r) \subset E^\circ$. Thus, the interior E° is open, as claimed.

(2) Let us prove now that the closure $\bar{E}$ is closed. For this purpose, we will prove that the complement $\bar{E}^c$ is open. So, pick $x \in \bar{E}^c$. Then x cannot appear as a limit of a sequence $x_n \to x$ with $x_n \in E$, so we have a ball $B_x(r) \subset \bar{E}^c$, as desired.

(3) Finally, the maximality and minimality assertions regarding E° and $\bar{E}$ are both routine too, coming from definitions, and we will leave them as exercises. □

As an application of the theory developed above, and more specifically of the notion of closure from Definition 10.8, we can talk as well about density, as follows:

DEFINITION 10.10. *We say that a subset $E \subset X$ is dense when:*

$$\bar{E} = X$$

That is, any point of X must appear as a limit of points of E.

Obviously, this is something which is in tune with what we know so far from this book, and with the intuitive notion of density. As a basic example, we have $\bar{\mathbb{Q}} = \mathbb{R}$, that we know well from the beginning of this book. As another example, we have the fact that the diagonalizable matrices are dense inside $M_N(\mathbb{C})$, that we know from chapter 9. There are of course many other examples, and we will be back to this, in what follows.

10b. Compact sets

Moving ahead now, at a more subtle level, again in analogy with what we know about $X = \mathbb{R}, \mathbb{C}$, we can talk about compact sets, and about connected sets. Again things here are quite tricky, in the general metric space framework, actually substantially deviating from what we know, and we will do this in detail. Let us start with:

DEFINITION 10.11. *A set $K \subset X$ is called compact if any cover with open sets*

$$K \subset \bigcup_i E_i$$

has a finite subcover, $K \subset (E_{i_1} \cup \ldots \cup E_{i_n})$.

This definition might seem overly abstract, and perhaps even sound like a joke, but our claim is that this is the correct definition, and that there is no way of doing otherwise. Let us start with some examples, with $X = \mathbb{R}$. The situation here is as follows:

(1) A point is obviously compact, and we can choose that finite subcover with $n = 1$. Similarly, 2 points are compact, and we can choose the subcover with $n = 2$. More generally, N points are compact, and we can choose the subcover with $n = N$.

(2) In contrast, the set $\mathbb{N} \subset \mathbb{R}$ is not compact, because the following open cover of it obviously has no finite subcover:

$$\mathbb{N} \subset \bigcup_n \left(n - \frac{1}{3} \, , \, n + \frac{1}{3} \right)$$

Similarly, the set $\{1/n | n \in \mathbb{N}\}$ is not compact either, with a similar cover, consisting of a suitable union of open intervals around each point, doing the job.

(3) However, and here comes an interesting point, the following set is compact:

$$K = \left\{ \frac{1}{n} \Big| n \in \mathbb{N} \right\} \cup \{0\}$$

Indeed, any open cover of it $\cup_i E_i$ has to cover 0, and by selecting an open set E_i covering 0, this set E_i will cover the whole K, except for finitely many points, due to $1/n \to 0$. But these finitely points left, say N of them, can be covered by suitable sets $E_{i_1}, \ldots, E_{i_N}$, and by adding to this family the set E_i, we have our finite subcover.

As a conclusion to this, Definition 10.11 seems to be in tune with what we know about the compact subsets $K \subset \mathbb{R}$, namely that these are the sets which are closed and bounded. However, and here comes our point, such things are wrong in general, due to:

THEOREM 10.12. *Given an infinite set X with the discrete distance on it, namely $d(p,q) = 1 - \delta_{pq}$, which can be modelled as the basis of a suitable Hilbert space,*

$$X = \{e_x\}_{x \in X} \subset l^2(X)$$

this set is closed and bounded, but not compact.

PROOF. Here the first part, regarding the modelling of X, that we will actually not really need, is something that we already know. Regarding now the second part:

(1) X being the total space, it is by definition closed. As a remark here, that we will need later, since the points of X are obviously open, any subset $E \subset X$ is open, and by taking complements, any set $E \subset X$ is closed as well.

(2) X is also bounded, because all distances are smaller than 1.

(3) However, our set X is not compact, because its points being open, as noted above, $X = \cup_{x \in X} \{x\}$ is an open cover, having no finite subcover. □

The above result is quite interesting, and is the source of all troubles with compactness, in general, and can be subject to some further meditation. First, you might argue that by declaring X to be the total space, we have "cheated" with its closedness, which came like this for free, without computations. But, when using the modelling $X \subset l^2(X)$ mentioned in the statement, X still remains closed, this time with no cheating involved.

As a second thought that you might have, you might perhaps say okay, but why not changing our definition of compactness at to include the above set X, which looks quite nice, being both closed and bounded, into our class of compact sets. But this does not work, because after some thinking, the above set X is in fact not nice at all, and all sorts of things that you can try with it, for "confirming" its compactness, will simply fail.

So, this is the situation, as a conclusion, Definition 10.11 is both in tune with what we know about the compact sets $K \subset \mathbb{R}$, which must be closed and bounded, and also in tune with what we just learned from Theorem 10.12, telling us that closed and bounded is worth nothing, or almost, in general. Thus, Definition 10.11 is the correct definition.

Moving away now from these philosophical thoughts, or rather with the promise to come back to them later, when we will know more, let us develop the theory of compact sets, as axiomatized in Definition 10.11, and see what we get. We first have the following elementary result, confirming that we are on the good way, with Definition 10.11:

PROPOSITION 10.13. *The following hold:*

(1) *Compact implies closed.*
(2) *Closed inside compact is compact.*
(3) *Compact intersected with closed is compact.*

PROOF. These assertions are all clear from definitions, as follows:

(1) Assume that $K \subset X$ is compact, and let us prove that K is closed. For this purpose, we will prove that K^c is open. So, pick $p \in K^c$. For any $q \in K$ we set $r = d(p,q)/3$, and we consider the following balls, separating p and q:

$$U_q = B_p(r) \quad , \quad V_q = B_q(r)$$

We have then $K \subset \cup_{q \in K} V_q$, so we can pick a finite subcover, as follows:

$$K \subset (V_{q_1} \cup \ldots \cup V_{q_n})$$

With this done, consider the following intersection:

$$U = U_{q_1} \cap \ldots \cap U_{q_n}$$

This intersection is then a ball around p, and since this ball avoids $V_{q_1}, \ldots, V_{q_n}$, it avoids the whole K. Thus, we have proved that K^c is open at p, as desired.

(2) Assume that $F \subset K$ is closed, with $K \subset X$ being compact. For proving our result, we can assume, by replacing X with K, that we have $X = K$. In order to prove now that F is compact, consider an open cover of it, as follows:

$$F \subset \bigcup_i E_i$$

By adding the set F^c, which is open, to this cover, we obtain a cover of K. Now since K is compact, we can extract from this a finite subcover Ω, and there are two cases:

– If $F^c \in \Omega$, by removing F^c from Ω we obtain a finite cover of F, as desired.

– If $F^c \notin \Omega$, we are done too, because in this case Ω is a finite cover of F.

(3) This follows from (1) and (2), because if $K \subset X$ is compact, and $F \subset X$ is closed, then $K \cap F \subset K$ is closed inside a compact, so it is compact. □

As a second batch of results, which are useful as well, we have:

PROPOSITION 10.14. *The following hold:*

(1) *If $K_i \subset X$ are compact, satisfying $K_{i_1} \cap \ldots \cap K_{i_n} \neq \emptyset$, then $\cap_i K_i \neq \emptyset$.*
(2) *If $K_1 \supset K_2 \supset K_3 \supset \ldots$ are non-empty compacts, then $\cap_i K_i \neq \emptyset$.*
(3) *If K is compact, and $E \subset K$ is infinite, then E has a limit point in K.*
(4) *If K is compact, any sequence $\{x_n\} \subset K$ has a limit point in K.*
(5) *If K is compact, any $\{x_n\} \subset K$ has a subsequence which converges in K.*

PROOF. Again, these are elementary results, which can be proved as follows:

(1) Assume by contradiction $\cap_i K_i = \emptyset$, and let us pick $K_1 \in \{K_i\}$. Since any $x \in K_1$ is not in $\cap_i K_i$, there is an index i such that $x \in K_i^c$, and we conclude that we have:

$$K_1 \subset \bigcup_{i \neq 1} K_i^c$$

But this can be regarded as being an open cover of K_1, that we know to be compact, so we can extract from it a finite subcover, as follows:

$$K_1 \subset \left(K_{i_1}^c \cup \ldots \cup K_{i_n}^c\right)$$

Now observe that this latter subcover tells us that we have:

$$K_1 \cap K_{i_1} \cap \ldots \cap K_{i_n} = \emptyset$$

But this contradicts our intersection assumption in the statement, and we are done.

(2) This is a particular case of (1), proved above.

(3) We prove this by contradiction. So, assume that E has no limit point in K. This means that any $p \in K$ can be isolated from the rest of E by a certain open ball $V_p = B_p(r)$, and in both the cases that can appear, $p \in E$ or $p \notin E$, we have:

$$|V_p \cap E| = 0, 1$$

Now observe that these sets V_p form an open cover of K, and so of E. But due to $|V_p \cap E| = 0, 1$ and to $|E| = \infty$, this open cover of E has no finite subcover. Thus the same cover, regarded now as cover of K, has no finite subcover either, contradiction.

(4) This follows from (3) that we just proved, with $E = \{x_n\}$.

(5) This is a reformulation of (4), that we just proved. □

Getting now to some more exciting theory, here is a key result about compactness, which is less trivial, and that we will need on a regular basis, in what follows:

THEOREM 10.15. *For a subset $K \subset \mathbb{R}^N$, the following are equivalent:*

(1) *K is closed and bounded.*
(2) *K is compact.*
(3) *Any infinite subset $E \subset K$ has a limiting point in K.*

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

(1) $\implies$ (2) As a first task, in order to establish this implication, let us prove that any product of closed intervals, as follows, is indeed compact:

$$J = \prod_{i=1}^{N} [a_i, b_i] \subset \mathbb{R}^N$$

We can assume by linearity that we are dealing with the unit cube:

$$C_1 = \prod_{i=1}^{N} [0, 1] \subset \mathbb{R}^N$$

In order to prove that C_1 is compact, we proceed by contradiction. So, assume that we have an open cover as follows, having no finite subcover:

$$C_1 \subset \bigcup_i E_i$$

Now let us cut C_1 into 2^N small cubes, in the obvious way, over the N coordinate axes. Then at least one of these small cubes, which are all covered by $\cup_i E_i$ too, has no finite subcover. So, let us call $C_2 \subset C_1$ one of these small cubes, having no finite subcover:

$$C_2 \subset \bigcup_i E_i$$

We can then cut C_2 into 2^N small cubes, and by the same reasoning, we obtain a smaller cube $C_3 \subset C_2$ having no finite subcover. And so on by recurrence, and we end up with a decreasing sequence of cubes, as follows, having no finite subcover:

$$C_1 \supset C_2 \supset C_3 \supset \ldots$$

Now since these decreasing cubes have edge size $1, 1/2, 1/4, \ldots$, their intersection must be a point. So, let us call p this point, defined by the following formula:

$$\{p\} = \bigcap_k C_k$$

But this point p must be covered by $\cup_i E_i$, so we can find an index i such that:

$$p \in E_i$$

Now observe that E_i must contain a whole ball around p, and so starting from a certain $K \in \mathbb{N}$, all the cubes C_k will be contained in this ball, and so in E_i:

$$C_k \subset E_i \quad , \quad \forall k \geq K$$

But this is a contradiction, because C_K, and in fact the smaller cubes C_k with $k > K$ as well, were assumed to have no finite subcover. Thus, we have proved our claim.

(1) $\implies$ (2), continuation. But with this claim in hand, the result is now clear. Indeed, assume that $K \subset \mathbb{R}^N$ is closed and bounded. Then, since K is bounded, we can view it as a subset as a suitable big cube, of the following form:

$$K \subset \prod_{i=1}^{N} [-M, M] \subset \mathbb{R}^N$$

But, what we have here is a closed subset inside a compact set, that follows to be compact, as desired.

(2) $\implies$ (3) This is something that we already know, not needing $K \subset \mathbb{R}^N$.

(3) $\implies$ (1) We have to prove that K as in the statement is both closed and bounded, and we will do both these things by contradiction, as follows:

– Assume first that K is not closed. But this means that we can find a point $x \notin K$ which is a limiting point of K. Now let us pick $x_n \in K$, with $x_n \to x$, and consider the set $E = \{x_n\}$. According to our assumption, E must have a limiting point in K. But this limiting point can only be x, which is not in K, contradiction.

– Assume now that K is not bounded. But this means that we can find points $x_n \in K$ satisfying $||x_n|| \to \infty$, and if we consider the set $E = \{x_n\}$, then again this set must have a limiting point in K, which is impossible, so we have our contradiction, as desired. □

So long for compactness. As a last piece of general topology, in our metric space framework, we can talk as well about connectedness, as follows:

DEFINITION 10.16. *We can talk about connected sets $E \subset X$, as follows:*

(1) *We say that E is connected if it cannot be separated as $E = E_1 \cup E_2$, with the components E_1, E_2 satisfying $E_1 \cap \bar{E}_2 = \bar{E}_1 \cap E_2 = \emptyset$.*
(2) *We say that E is path connected if any two points $p, q \in E$ can be joined by a path, meaning a continuous $f : [0.1] \to X$, with $f(0) = p$, $f(1) = q$.*

All this looks a bit technical, and indeed it is. To start with, (1) is something quite natural, but the separation condition there $E_1 \cap \bar{E}_2 = \bar{E}_1 \cap E_2 = \emptyset$ can be weakened into $E_1 \cap E_2 = \emptyset$, or strengthened into $\bar{E}_1 \cap \bar{E}_2 = \emptyset$, depeding on purposes, and with our (1) as formulated being the good compromise, for most purposes.

As for (2), this condition is obviously something stronger, and we have in fact the following implications, which are both clear:

$$\text{convex} \implies \text{path connected} \implies \text{connected}$$

To be more precise, the first implication is clear, by taking as path as in Definition 10.16 (2) the segment joining p, q, and for the second implication, the idea is that a separation of E as in Definition 10.16 (1) will produce a separation of the path joining p, q, and so ultimately, a separation of the interval $[0, 1]$, which is impossible.

The problem, however, is that connected does not imply path connected, and there are as well various counterexamples in relation with the various versions of (1) that can be formulated, as explained above. Anyway, leaving aside the discussion here, which is something quite technical, once all these questions clarified, the idea is that any set E can be written as a disjoint union of connected components, as follows:

$$E = \bigsqcup_i E_i$$

However, the story is not over here, because when looking at the connected components E_i, or simply at the connected sets E, if you prefer, there are many things that can happen, in relation with the "holes" that E can have or not. Thus, the classification of the connected sets runs into the question of deciding how many holes can have such a set E, and this is something quite subtle, that we will discuss in chapter 11 below.

Getting back now to more concrete things, remember that we are here in this book for studying functions, and doing calculus. And, regarding functions, we have:

THEOREM 10.17. *Assuming that $f : X \to Y$ is continuous, the following happen:*

(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 fundamental, which can be proved as follows:

(1) This is clear from the definition of continuity, written with ε, δ. In fact, the converse holds too, in the sense that if $f^{-1}(\text{open}) = \text{open}$, then f must be continuous.

(2) This follows from (1), by taking complements. And again, the converse holds too, in the sense that if $f^{-1}(\text{closed}) = \text{closed}$, then f must be continuous.

(3) This is something that took us some time to prove for the functions $f : \mathbb{R} \to \mathbb{R}$, earlier in this book, with our compactness technology there, but which is now clear, by using our definition of compactness with open covers. Indeed, given an open cover $f(K) \subset \cup_i E_i$, we have by using (1) an open cover $K \subset \cup_i f^{-1}(E_i)$, and so by compactness

of K, a finite subcover $K \subset f^{-1}(E_{i_1}) \cup \ldots \cup f^{-1}(E_{i_n})$, and so finally a finite subcover $f(K) \subset E_{i_1} \cup \ldots \cup E_{i_n}$, as desired. And isn't this smart, hope you agree with me.

(4) This can be proved via the same trick as for (3). Indeed, any separation of $f(E)$ into two parts can be returned via f^{-1} into a separation of E into two parts, contradiction. □

As a comment here, Theorem 10.17 generalizes, and in a clever way, many things that we know from one-variable calculus. Of particular interest is (3), which shows in particular that any continuous function on a compact space $f : X \to \mathbb{R}$ attains its minimum and its maximum, and then (4), which can be regarded as being a general mean value theorem. As for (1) and (2), these are useful in everyday life, and we will see examples of this.

In addition to this, all useful things, there are a few other things that can be said about the continuity of the functions $f : \mathbb{R}^N \to \mathbb{R}^M$. We will be back to this.

10c. Partial derivatives

Let us discuss now the differentiability in several variables. At order 1, the situation is quite similar to the one in 1 variable, but this time involving matrices. In order to explain this material, let us start with a straightforward definition, as follows:

DEFINITION 10.18. *We say that a map $f : \mathbb{R}^N \to \mathbb{R}^M$ is differentiable at $x \in \mathbb{R}^N$ if*

$$f(x + t) \simeq f(x) + f'(x)t$$

for some linear map $f'(x) : \mathbb{R}^N \to \mathbb{R}^M$, called derivative of f at the point $x \in \mathbb{R}^N$.

But is this the correct definition. I can hear you screaming that we are probably going the wrong way, because for functions $f : \mathbb{R} \to \mathbb{R}$ the derivative is something much simpler, as follows, and that we should try to imitate, in our higher dimensional setting:

$$f'(x) = \lim_{t \to 0} \frac{f(x + t) - f(x)}{t}$$

However, this is not possible, for a number of reasons, that are worth discussing in detail. So, here is the discussion, answering all kinds of questions that you might have:

(1) First of all, the above formula does not make any sense for a function $f : \mathbb{R}^N \to \mathbb{R}^M$ with $N \neq M$, because we cannot divide oranges by apples. And it doesn't make sense either at $N = M \in \mathbb{N}$, because, well, here we have $\mathbb{R}^N$ oranges, I agree with you, but there is no way of dividing these oranges, unless we are in the special cases $N = 1, 2$.

(2) More philosophically know, we have seen that having $f'(x)$ defined as a number is difficult, but the question is, do we really want to have $f'(x)$ defined as a number? And my claim here is that, this would be a pity. Think at the case where $f : \mathbb{R}^N \to \mathbb{R}^M$ is linear. Such a map is just "perfect", and so should equal its own derivative, $f = f'$.

(3) Summarizing, our Definition 10.18 is just perfection, and is waiting for some further study, and this is what we will do. And in case you're still secretly dreaming about having $f'(x)$ defined as some sort of number, wait for it. When $N = M$ at least, there is indeed a lucky number, namely $\det(f'(x))$, called Jacobian, but more on this later.

Getting back now to Definition 10.18 as formulated, and agreed upon, we have there a linear map $f'(x) : \mathbb{R}^N \to \mathbb{R}^M$, waiting to be further understood. So, time now to use our linear algebra knowledge from chapter 9. We know from there that such linear maps correspond to rectangular matrices $A \in M_{M\times N}(\mathbb{R})$, and we are led in this way to:

QUESTION 10.19. *Given a differentiable map $f : \mathbb{R}^N \to \mathbb{R}^M$, in the abstract sense of Definition 10.18, what exactly is its derivative*

$$f'(x) : \mathbb{R}^N \to \mathbb{R}^M$$

regarded as a rectangular matrix, $f'(x) \in M_{M\times N}(\mathbb{R})$?

Again, I might hear scream you here, arguing that you come after a long battle, just agreeing that the derivative is a linear map, and not a number, and now what, we are trying to replace this linear map by a matrix, and so by a bunch of numbers.

Good point, and I must admit that I have no good answer to this. In fact, what we are doing here, namely Definition 10.18, then Question 10.19, and finally Theorem 10.20 to follow in a moment, are quite deep things, that took mankind several centuries to develop, and that we are now presenting in a compressed form. So yes, all this is difficult mathematics, when you first see it, I perfectly agree with you.

In any case, hope that you're still with me, and in order to further clarify all this, here is the answer to Question 10.19:

THEOREM 10.20. *The derivative of a differentiable function $f : \mathbb{R}^N \to \mathbb{R}^M$, making the approximation formula*

$$f(x + t) \simeq f(x) + f'(x)t$$

work, is 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, which does make sense, the idea is 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 general case $M = 1$, with the formula here, obtained via a straightforward recurrence, being as follows:

$$\begin{aligned} f\begin{pmatrix} x_1 + t_1 \\ \vdots \\ x_N + t_N \end{pmatrix} &\simeq f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \frac{df}{dx_1}(x)t_1 + \ldots + \frac{df}{dx_N}(x)t_N \\ &= f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \frac{df}{dx_1}(x) & \ldots & \frac{df}{dx_N}(x) \end{pmatrix}\begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \end{aligned}$$

(4) But this gives the result in the case where both $N, M \in \mathbb{N}$ are arbitrary too. Indeed, consider a function $f : \mathbb{R}^N \to \mathbb{R}^M$, and let us write it as follows:

$$f = \begin{pmatrix} f_1 \\ \vdots \\ f_M \end{pmatrix}$$

We can apply (3) 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 collection of M formulae tells us precisely that the following happens, as an equality, or rather approximation, of vectors in $\mathbb{R}^M$:

$$f\begin{pmatrix}x_1+t_1\\ \vdots\\ x_N+t_N\end{pmatrix}\simeq f\begin{pmatrix}x_1\\ \vdots\\ x_N\end{pmatrix}+\begin{pmatrix}\frac{df_1}{dx_1}(x) & \ldots & \frac{df_1}{dx_N}(x)\\ \vdots & & \vdots\\ \frac{df_M}{dx_1}(x) & \ldots & \frac{df_M}{dx_N}(x)\end{pmatrix}\begin{pmatrix}t_1\\ \vdots\\ t_N\end{pmatrix}$$

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

The above result, while being very nice, clear and useful, does not close the discussion regarding the derivative. Indeed, conversely, we would like to know if the existence of the partial derivatives guarantees the fact that f is differentiable, with derivative $f'(x)$ appearing as the rectangular matrix fomed by the partial derivatives at x.

So, let us discuss now this remaining theoretical question. The result here, which is something quite technical, but which can be useful in practice, is as follows:

THEOREM 10.21. *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 10.20, 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 10.20 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 $s_j \in [0,1]$:

$$f(x + t^{(j)}) - f(x + t^{(j-1)}) = \frac{df}{dx_j}(x + s_j t^{(j)} + (1-s_j)t^{(j-1)}) \cdot t_j$$

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:

$$\left|\frac{df}{dx_j}(x + s_j t^{(j)} + (1-s_j)t^{(j-1)}) - \frac{df}{dx_j}(x)\right| < \frac{\varepsilon}{N}$$

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. □

This was for the basic theory of partial derivatives. In practice, there are far more things that can be said, both at the abstract and the concrete level, including of course many examples. We will be back to this, after developing some more general theory.

Before getting into this, however, let me formulate a definition that you will certainly appreciate, bringing a bit of humanity, and more specifically a good old real number, in this world of vectors, matrices and other beasts which is multivariable calculus:

DEFINITION 10.22. *Given a differentiable function $f:\mathbb{R}^N\to\mathbb{R}^N$, its Jacobian at $x\in\mathbb{R}^N$ is the number*

$$\det(f'(x))\in\mathbb{R}$$

measuring the infinitesimal rate of the volume inflation by f, at the point x.

Here the first part is standard, because when $N=M$, as above, the derivative is a linear map $f'(x):\mathbb{R}^N\to\mathbb{R}^N$, which is the same as a square matrix $f'(x)\in M_N(\mathbb{R})$, and so we can consider the determinant of this matrix, $\det(f'(x))\in\mathbb{R}$. As for the second part, this comes from our knowledge of the determinant from chapter 9.

All this is very nice, and as a first observation, according to our formula of $f'(x)$ as being the matrix formed by the partial derivatives, we have:

$$\det(f'(x)) = \begin{vmatrix} \frac{df_1}{dx_1}(x) & \ldots & \frac{df_1}{dx_N}(x) \\ \vdots & & \vdots \\ \frac{df_N}{dx_1}(x) & \ldots & \frac{df_N}{dx_N}(x) \end{vmatrix}$$

Thus, the Jacobian can be explicitely computed. However, in what regards the practical uses of the Jacobian, these are quite complicated, and this will have to wait a bit, until chapter 14 below. So, sorry for this, not yet time to enjoy Definition 10.22, and stay with me, plenty of further linear algebra, and matrices instead of numbers, to follow.

10d. The chain rule

Generally speaking, Theorem 10.20 is what you need to know for upgrading from calculus to multivariable calculus, of course with some help from linear algebra as well. This is a general principle of multivariable calculus, that we will explore now.

So, let us first recall that we know from Theorem 10.20 that the derivative of a differentiable function $f : \mathbb{R}^N \to \mathbb{R}^M$, making the approximation formula $f(x+t) \simeq f(x) + f'(x)t$ work, is the matrix of partial derivatives at x, namely:

$$f'(x) = \left(\frac{df_i}{dx_j}(x) \right)_{ij} \in M_{M \times N}(\mathbb{R})$$

In practice, unless we are in the special case $N = M$, where our derivative is a square matrix, $f'(x) \in M_N(\mathbb{R})$, and more can be said, as briefly explained at the end of the previous section, we can always write our function as follows, with each of the components $f_1, \ldots, f_M$ being a scalar-valued function, $f_i : \mathbb{R}^N \to \mathbb{R}$:

$$f = \begin{pmatrix} f_1 \\ \vdots \\ f_M \end{pmatrix}$$

With this convention, the derivative, as a rectangular matrix, is given by:

$$f'(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_M(x) \end{pmatrix}$$

Moreover, assuming that the formula $f_i(x+t) \simeq f_i(x) + f_i'(x)t$ works for any i, it follows that the formula $f(x+t) \simeq f(x) + f'(x)t$ works too. Thus, as a conclusion to this, Theorem 10.20 is more of less equivalent to its $M = 1$ particular case.

So, let us record this $M = 1$ particular case of Theorem 10.20, as follows:

THEOREM 10.23. *The derivative of a differentiable function $f : \mathbb{R}^N \to \mathbb{R}$, making the approximation formula $f(x+t) \simeq f(x) + f'(x)t$ work, is given by*

$$f'(x) = (\nabla f(x))^t$$

where ∇f, called gradient of f, is the vector formed by the partial derivatives:

$$\nabla f(x) = \left(\frac{df}{dx_j}(x) \right)_j \in \mathbb{R}^N$$

Equivalently, without reference to f', we have the approximation formula

$$f(x+t) \simeq f(x) + < \nabla f(x), t >$$

where $<,>$ is the usual scalar product on $\mathbb{R}^N$.

PROOF. This is a reformulation of Theorem 10.20, in the particular case $M = 1$. Indeed, we know from there that the derivative of our function $f : \mathbb{R}^N \to \mathbb{R}$, making the approximation formula $f(x+t) \simeq f(x) + f'(x)t$ work, is given by:

$$f'(x) = \left(\frac{df}{dx_j}(x) \right)_j \in M_{1 \times N}(\mathbb{R})$$

But, what we have here is the horizontal vector formed by the partial derivatives at x, which is transpose to the vertical vector formed by the same partial derivatives, which is by definition $\nabla f(x)$. Thus, we are led to the formula in the statement, namely:

$$f'(x) = (\nabla f(x))^t$$

Finally, regarding the last assertion, this follows from this, as follows:

$$\begin{aligned} f(x+t) &\simeq f(x) + f'(x)t \\ &= f(x) + (\nabla f(x))^t t \\ &= f(x) + \sum_j (\nabla f(x))_j t_j \\ &= f(x) + < \nabla f(x), t > \end{aligned}$$

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

In practice, more things can be said about the gradient, which is a notion which is quite intuitive. Also, getting back now to the general functions $f : \mathbb{R}^N \to \mathbb{R}^M$, in the case where we are in small dimension, $N, M \leq 3$, we can do many other geometric things with the partial derivatives. We will be back to all this, gradually, in what follows.

In general now, our first task will be that of extending to several variables our basic results from one-variable calculus. As a standard result here, we have:

THEOREM 10.24. *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 notion of derivative coming from Theorem 10.20. 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 10.20, 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. □

The above formula is quite interesting, and working its main particular cases, when some of parameters M,N,K equal 1, is a very instructive exercice. We typically obtain here some mysterious formulae involving the gradient from Theorem 10.23. But, the best is to rather ignore all these particular cases, and memorize Theorem 10.24 as stated.

Along the same lines, observe also that we can as well reformulate Theorem 10.24 in terms of partial derivatives, by performing the matrix multiplication on the right. But again, the best is to rather ignore all this, and memorize Theorem 10.24 as stated.

This being said, nothing better than working out a few exercises, at this point of learning. As a first result coming from Theorem 10.24, as a particular case, we have:

PROPOSITION 10.25. *We have the chain derivative formula*

$$\frac{d(f\circ g)_i}{dx_j}(x)=\sum_k\frac{df_i}{dx_k}(g(x))\cdot\frac{dg_k}{dx_j}(x)$$

as an equality of numbers.

PROOF. This follows indeed from the formula in Theorem 10.24, namely:

$$(f\circ g)'(x)=f'(g(x))\cdot g'(x)$$

Indeed, what we have here is an equality of matrices, and at the level of the individual entries of these matrices, by performing the multiplication on the right, we obtain:

$$(f\circ g)'(x)_{ij}=\sum_k f'(g(x))_{ik}\cdot g'(x)_{kj}$$

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

As already mentioned, there are many possible variations of the above formula, obtained when some of dimension parameters M, N, K equal 1, and with the final formulae involving tricky sums, or scalar products, whose components are partial derivatives, or entries of the gradient. We will leave them as an instructive exercise.

Moving ahead now, we have several useful consequences of the above result, a standard application, generalizing things that we know in one variable, being as follows:

THEOREM 10.26. *Assuming that $f:X\to\mathbb{R}^M$ is differentiable, with $X\subset\mathbb{R}^N$ being convex, we have the estimate*

$$||f(x)-f(y)||\leq M||x-y||$$

for any $x,y\in X$, where the quantity on the right is given by:

$$M=\sup_{x\in X}||f'(x)||$$

Moreover, this estimate can be sharp, for instance for the linear functions.

PROOF. This is something quite tricky, which in several variables cannot be proved with bare hands. However, we can get it by using our chain derivative formula. Consider indeed the path $\gamma:[0,1]\to\mathbb{R}^M$ given by the following formula:

$$\gamma(t)=tx+(1-t)y$$

Now let us set $g(t)=f(\gamma(t))$. We have then, according to the chain rule formula:

$$\begin{aligned}g'(t)&=f'(\gamma(t))\gamma'(t)\\&=f'(\gamma(t))(x-y)\end{aligned}$$

But this gives the following estimate, with $M>0$ being as in the statement:

$$\begin{aligned}|g'(t)|&\leq||f'(\gamma(t))||\cdot||x-y||\\&\leq M||x-y||\end{aligned}$$

Now by using one-variable results that we know, we obtain from this:

$$||g(1)-g(0)||\leq||M||\cdot||x-y||$$

But since we have $g(1)=f(x), g(0)=f(y)$, this gives the formula in the statement. Finally, the last assertion is clear. □

As a conclusion to all this, we have extended to the case of vector functions most of what we know about the one-variable functions, at the general level, of basic calculus. Still waiting to be discussed are the higher derivatives, and their various applications, and we will come back to this in chapter 12, after a short geometric break, and then of course the theory of integration, that we will discuss in chapters 13-16.

10e. Exercises

This was a quite straightforward chapter, assuming that you have understood well Part I and chapter 9, and as exercises here, mostly straightforward, we have:

EXERCISE 10.27. *Write a short essay about the basic theory of sequences and continuous functions over metric spaces, following the material from Part I.*

EXERCISE 10.28. *Clarify the relation between the various notions of connectedness, with the goal of decomposing each metric space into connected components.*

EXERCISE 10.29. *Enjoy defining the derivative of a function $f : \mathbb{R}^N \to \mathbb{R}^N$ as being a number, namely its Jacobian. What can you do, and what not, with this approach?*

EXERCISE 10.30. *Look up the internet for various other formulations of the chain rule for derivatives, and prove them all, quickly, by using what we know.*

As bonus exercise, think a bit about higher derivatives, and what can you do with them, following the material from Part I. We will be back to this, after a break.

CHAPTER 11

Some geometry

11a. Equations, conics

As an application of the theory of partial derivatives that we developed, and of calculus in general, as we know it at this point, we can now talk about some beautiful things, namely geometry and physics. As a starting point, you surely have noticed that the Sun moves around the Earth on a circle. However, when carefully measuring it, this circle is not exactly a circle, but rather an ellipsis. Also, some further possible trajectories, of one object with respect to another, due to gravity, include parabolas and hyperbolas, that you can observe with a telescope, by looking at asteroids and comets.

So, before even starting to look at the equations of gravity, and having some fun in solving them, we need a mathematical theory of simple curves like ellipses, parabolas and hyperbolas, which are what we can expect to find, from gravity computations. And, good news, this theory exists, since the ancient Greeks. Let us start with:

DEFINITION 11.1. *A conic is a plane algebraic curve of the form*

$$C = \left\{(x,y) \in \mathbb{R}^2 \middle| P(x,y) = 0\right\}$$

with $P \in \mathbb{R}[x,y]$ *being of degree* ≤ 2.

As basic examples of conics, we have the ellipses, parabolas and hyperbolas. The simplest examples of these are as follows, with the ellipsis actually being a circle:

$$x^2 + y^2 = 1 \quad , \quad x^2 = y \quad , \quad xy = 1$$

Observe that, due to our assumption $\deg P \leq 2$, we have as conics some degenerate curves as well, such as lines, $\emptyset$, and $\mathbb{R}^2$ itself, coming from $\deg P \leq 1$, as follows:

$$x = 0 \quad , \quad 1 = 0 \quad , \quad 0 = 0$$

This might suggest to replace our assumption $\deg P \leq 2$ by $\deg P = 2$, but we will not do so, because $\deg P = 2$ does not rule out degenerate situations, such as:

$$x^2 + y^2 = -1 \quad , \quad x^2 + y^2 = 0 \quad , \quad x^2 = 0 \quad , \quad xy = 0$$

In fact, what we get here are $\emptyset$, points, lines, and pairs of lines, so in the end, assuming $\deg P = 2$ instead of $\deg P \leq 2$ would only rule out $\mathbb{R}^2$ itself, which is not worth it.

Summarizing, our notion of conic from Definition 11.1 looks quite reasonable, so let us agree on this notion. Getting now to classification matters, we first have:

THEOREM 11.2. *Up to non-degenerate linear transformations of the plane,*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to A \begin{pmatrix} x \\ y \end{pmatrix}$$

with $\det A \neq 0$*, the conics fall into two classes, as follows:*

(1) *Non-degenerate: circles, parabolas, hyperbolas.*
(2) *Degenerate:* $\emptyset$*, points, lines, pairs of lines,* $\mathbb{R}^2$*.*

PROOF. As a first observation, looks like we forgot the ellipses, but via linear transformations these become circles, so things fine. As for the proof, this goes as follows:

(1) Consider an arbitrary conic, written as follows, with $a, b, c, d, e, f \in \mathbb{R}$:

$$ax^2 + by^2 + cxy + dx + ey + f = 0$$

(2) Assume first $a \neq 0$. By making a square out of ax^2, up to a linear transformation in (x, y), we can get rid of the term cxy, and we are left with:

$$ax^2 + by^2 + dx + ey + f = 0$$

In the case $b \neq 0$ we can make two obvious squares, and again up to a linear transformation in (x, y), we are left with an equation as follows:

$$x^2 \pm y^2 = k$$

In the case of positive sign, $x^2 + y^2 = k$, the solutions are the circle, when $k \geq 0$, the point, when $k = 0$, and $\emptyset$, when $k < 0$. As for the case of negative sign, $x^2 - y^2 = k$, which reads $(x-y)(x+y) = k$, here once again by linearity our equation becomes $xy = l$, which is a hyperbola when $l \neq 0$, and two lines when $l = 0$.

(3) In the case $b \neq 0$ the study is similar, with the same solutions, so we are left with the case $a = b = 0$. Here our conic is as follows, with $c, d, e, f \in \mathbb{R}$:

$$cxy + dx + ey + f = 0$$

If $c \neq 0$, by linearity our equation becomes $xy = l$, which produces a hyperbola or two lines, as explained before. As for the remaining case, $c = 0$, here our equation is:

$$dx + ey + f = 0$$

But this is generically the equation of a line, unless we are in the case $d = e = 0$, where our equation is $f = 0$, having as solutions $\emptyset$ when $f \neq 0$, and $\mathbb{R}^2$ when $f = 0$.

(4) So, this was the study of an arbitrary conic, and by putting now everything together, we are led to the conclusions in the statement. □

In order now to plainly classify the conics, without reference to a linear transformation of the plane, we just need to apply linear transformations to the curves that we found in Theorem 11.2. This leads to the following classification result:

THEOREM 11.3. *The conics fall into two classes, as follows:*

(1) *Non-degenerate: ellipses, parabolas, hyperbolas.*
(2) *Degenerate: $\emptyset$, points, lines, pairs of lines, $\mathbb{R}^2$.*

Also, the compact conics are $\emptyset$, the points, and the ellipses.

PROOF. We have several assertions here, the idea being as follows:

(1) As said above, in order to get such a classification result, we just need to apply linear transformations to the curves that we found in Theorem 11.2. But this leaves the list there unchanged, up to the circles becoming ellipses, as stated above.

(2) In what regards the last assertion, this is clear from the first one, but since this assertion is quite interesting, let us give it a quick, independent proof as well. Consider an arbitary conic, written as follows, with $a, b, c, d, e, f \in \mathbb{R}$:

$$ax^2 + by^2 + cxy + dx + ey + f = 0$$

Compacity rules then out the case $c \neq 0$, and our conic must be in fact:

$$ax^2 + by^2 + dx + ey + f = 0$$

But then with $a, b \neq 0$ we must have by compacity $a, b > 0$ or $a, b < 0$, and we get an ellipsis, then with $a = 0, b \neq 0$ or $a \neq 0, b = 0$ we get by compacity either $\emptyset$ or a point, and finally with $a = b = 0$ the compacity rules out again everything, except for $\emptyset$. □

As a third main result now on the conics, also known since the ancient Greeks, and which justifies the name "conics", coming from "cone", we have:

THEOREM 11.4. *Up to some degenerate cases, the conics are exactly the curves which appear by cutting a 2-sided cone with a plane.*

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

(1) By suitably choosing our coordinate axes (x, y, z), we can assume that our 2-sided cone is given by an equation as follows, with $k > 0$:

$$x^2 + y^2 = kz^2$$

In order to prove the result, we must intersect this cone with an arbitrary plane, which has an equation as follows, with $(a, b, c) \neq (0, 0, 0)$:

$$ax + by + cz = d$$

(2) However, before getting into computations, observe that what we want to find is a certain degree 2 equation in the above plane, for the intersection. Thus, it is convenient to change the coordinates, as for our plane to be given by the following equation:

$$z = 0$$

(3) But with this done, what we have to do is to see how the cone equation $x^2+y^2 = kz^2$ changes, under this change of coordinates, and then set $z = 0$, as to get the (x, y) equation of the intersection. But this leads, via some thinking or computations, to the conclusion that the cone equation $x^2 + y^2 = kz^2$ becomes in this way a degree 2 equation in (x, y), which can be arbitrary, and so to the final conclusion in the statement.

(4) Alternatively, and perhaps more concretely, we can use the original coordinates, with the cone being $x^2 + y^2 = kz^2$, and compute the intersection, with the conclusion that what we get, depending on the slope of the cone, and modulo degenerate cases, is an ellipsis, hyperbola or parabola. So, by invoking Theorem 11.3, we obtain the result.

(5) Summarizing, we have proved the result, modulo some details and interesting computations which are left to you, reader. Left to you as well is the full discussion concerning degree 2 curve degeneracy vs cone cutting degeneracy, with the remark that in what regards the cone cuts, the degenerate cases are very easy to identity and list, with the list consisting of $\emptyset$, the points, the lines, the pairs of lines, and $\mathbb{R}^2$ itself. □

All this is very nice, and as a conclusion to what we have so far about conics, we have the following statement, containing all the needed essentials:

THEOREM 11.5. *The conics, which are the algebraic curves of degree* 2 *in the plane,*

$$C = \left\{(x, y) \in \mathbb{R}^2 \middle| P(x, y) = 0\right\}$$

with $\deg P \leq 2$*, appear modulo degeneration by cutting a* 2*-sided cone with a plane, and can be classified into ellipses, parabolas and hyperbolas.*

PROOF. This follows indeed by putting together all the above results, and with the discussion concerning degeneration being left, as usual, as an instructive exercise. □

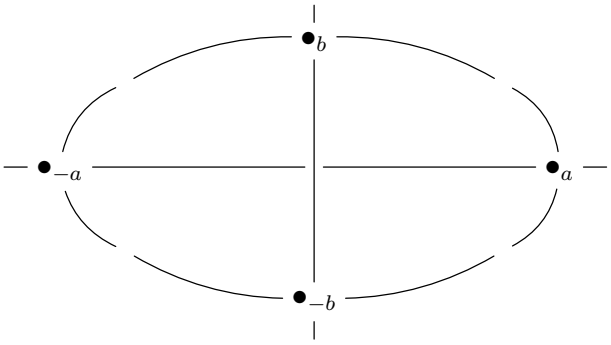
Moving ahead now, the most interesting conics, which are both compact and non-degenerate, are the ellipses. So, let us study them more in detail. As a starting point, we have the following statement, summarizing our knowledge about ellipses:

THEOREM 11.6. *The compact non-degenerate conics are the ellipses, which can be written, modulo rotations and translations in the plane, as*

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$$

with $a, b > 0$ *being half the size of a box containing the ellipsis. These ellipses also appear by compactly cutting a cone with a plane. The area of such an ellipsis is* $A = \pi ab$*.*

PROOF. In this statement most of the mathematics is from above, and with our explanations regarding the parameters $a, b > 0$ coming from the following picture:



As for the formula $A = \pi ab$, this comes from a computation done in chapter 4, but with different parameters, and here is the computation with our present parameters:

$$\begin{aligned} A &= 2\int_{-a}^{a} b\sqrt{1-\frac{x^2}{a^2}}\,dx \\ &= \frac{4b}{a}\int_{0}^{a}\sqrt{a^2-x^2}\,dx \\ &= 4ab\int_{0}^{1}\sqrt{1-y^2}\,dy \\ &= 4ab\cdot\frac{\pi}{4} \\ &= \pi ab \end{aligned}$$

Finally, as a verification, for $a = b = 1$ we get $A = \pi$, as we should. □

The above result is not the end of the story with ellipses, because we have as well, as a complement to it, or even as a rival result, which is just fine on its own:

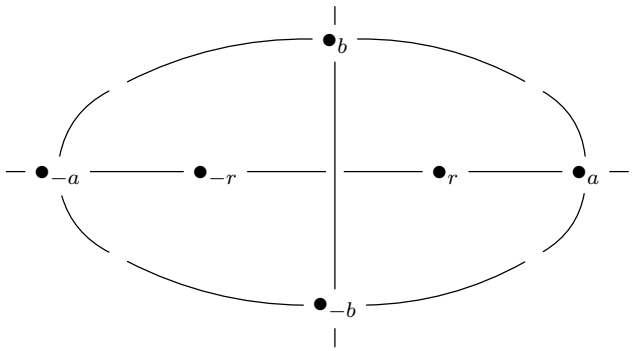
THEOREM 11.7. *The ellipses appear via equations of the following type, with p, q being two points in the plane, and with $l \geq d(p, q)$ being a certain length:*

$$d(z, p) + d(z, q) = l$$

For an ellipsis parametrized as before, $(x/a)^2 + (y/b)^2 = 1$ with $a \geq b \geq 0$, the focal points are $p = (0, -r)$ and $q = (0, r)$, with $r = \sqrt{a^2 - b^2}$, and the length is $l = 2a$.

PROOF. As already mentioned, it is possible to take $d(z, p) + d(z, q) = l$ as a definition for the ellipses, which is nice because all you need for drawing such an ellipsis is a string and a pencil, and then work out all the theory starting from this. In what concerns us, we will rather further build on what we know from Theorem 11.6, as follows:

(1) After some routine thinking, in order to fully prove the result, what we have to do is to take an ellipsis as parametrized in Theorem 11.6, and look for the focal points:



To be more precise, we are looking for a number $r > 0$, and a number $l > 0$, such that our ellipsis appears as $d(z, p) + d(z, q) = l$, with $p = (0, -r)$ and $q = (0, r)$.

(2) Let us first compute these numbers $r, l > 0$. Assuming that our result holds indeed as stated, by taking $z = (0, a)$, we see that the length l is:

$$l = (a - r) + (a + r) = 2a$$

As for the parameter r, by taking $z = (b, 0)$, we conclude that we must have:

$$2\sqrt{b^2 + r^2} = 2a \implies r = \sqrt{a^2 - b^2}$$

(3) With these observations made, let us prove the result. Given $l, r > 0$, and setting $p = (0, -r)$ and $q = (0, r)$, we have the following computation, with $z = (x, y)$:

$$\begin{aligned}
& d(z, p) + d(z, q) = l \\
\iff \quad & \sqrt{(x + r)^2 + y^2} + \sqrt{(x - r)^2 + y^2} = l \\
\iff \quad & \sqrt{(x + r)^2 + y^2} = l - \sqrt{(x - r)^2 + y^2} \\
\iff \quad & (x + r)^2 + y^2 = (x - r)^2 + y^2 + l^2 - 2l\sqrt{(x - r)^2 + y^2} \\
\iff \quad & 2l\sqrt{(x - r)^2 + y^2} = l^2 - 4xr \\
\iff \quad & 4l^2(x^2 + r^2 - 2xr + y^2) = l^4 + 16x^2r^2 - 8l^2xr \\
\iff \quad & 4l^2x^2 + 4l^2r^2 + 4l^2y^2 = l^4 + 16x^2r^2 \\
\iff \quad & (4x^2 - l^2)(4r^2 - l^2) = 4l^2y^2
\end{aligned}$$

(4) Now observe that we can further process the equation that we found as follows:

$$\begin{aligned}
(4x^2-l^2)(4r^2-l^2)=4l^2y^2 &\iff \frac{4x^2-l^2}{l^2}=\frac{4y^2}{4r^2-l^2}\\
&\iff \frac{4x^2-l^2}{l^2}=\frac{y^2}{r^2-l^2/4}\\
&\iff \left(\frac{x}{2l}\right)^2-1=\left(\frac{y}{\sqrt{r^2-l^2/4}}\right)^2\\
&\iff \left(\frac{x}{2l}\right)^2+\left(\frac{y}{\sqrt{r^2-l^2/4}}\right)^2=1
\end{aligned}$$

(5) Thus, our result holds indeed, and with the numbers $l,r>0$ appearing, and no surprise here, via the formulae $l=2a$ and $r=\sqrt{a^2-b^2}$, found in (2) above. □

The above results, which are old as modern mathematics itself, are foundational for algebraic geometry, and for geometry in general. We will be back to them later, after doing some physics, following Kepler and Newton, making the link with calculus.

11b. Kepler and Newton

As a continuation of the above, or rather as a complement, let us do some physics. Theorem 11.5 was the foundational result of modern mathematics, and by a remarkable twist of fate, the foundational result of physics is something related to it, as follows:

THEOREM 11.8. *Planets and other celestial bodies move around the Sun on conics,*

$$C=\left\{(x,y)\in\mathbb{R}^2\,\middle|\,P(x,y)=0\right\}$$

with $P\in\mathbb{R}[x,y]$ being of degree 2, which can be ellipses, parabolas or hyperbolas.

PROOF. This is something quite long, due to Kepler and Newton, but no fear, we know calculus, and threrefore what can resist us. The proof goes 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:

$$\begin{aligned} F &= -||F|| \cdot \frac{x}{||x||} \\ &= -\frac{GMm}{||x||^2} \cdot \frac{x}{||x||} \\ &= -\frac{GMmx}{||x||^3} \end{aligned}$$

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:

$$\ddot{x} = -\frac{GMx}{||x||^3}$$

Obviously, the problem happens in 2 dimensions, and you can even find, as an exercise, a formal proof of that, based on the above equation, if your really want to. Now 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:

$$\ddot{z} = -\frac{Kz}{||z||^3}$$

(2) The idea now is that 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:

$$\ddot{z} = -\frac{Kz}{r^3}$$

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:

$$\ddot{x} = a\cos\theta - b\sin\theta$$

$$\ddot{y} = a\sin\theta + b\cos\theta$$

(3) We can now solve the equation of motion from (1). 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:

$$\left(a + \frac{K}{r^2}\right)\cos\theta = b\sin\theta \quad , \quad \left(a + \frac{K}{r^2}\right)\sin\theta = -b\cos\theta$$

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:

$$a + \frac{K}{r^2} = b = 0$$

(4) Let us first examine the second equation, $b = 0$. This 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 $a + K/r^2 = 0$, this becomes:

$$\ddot{r} - \frac{\lambda^2}{r^3} + \frac{K}{r^2} = 0$$

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:

$$\ddot{r} = \frac{\lambda^2}{r^2}\left(\frac{1}{r} - \frac{1}{c}\right) \quad , \quad \dot{\theta} = \frac{\lambda}{r^2}$$

(5) In order to study the first equation, we use a trick. Let us write:

$$r(t) = \frac{1}{f(\theta(t))}$$

Abbreviated, and by 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 (4) above 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 following formula:

$$r = \frac{c}{1 + \varepsilon\cos\theta + \delta\sin\theta}$$

(6) 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:

$$x = \frac{c\cos\theta}{1 + \varepsilon\cos\theta + \delta\sin\theta}$$

$$y = \frac{c\sin\theta}{1 + \varepsilon\cos\theta + \delta\sin\theta}$$

Now observe that these two functions x, y satisfy the following formula:

$$x^2 + y^2 = \frac{c^2(\cos^2\theta + \sin^2\theta)}{(1 + \varepsilon\cos\theta + \delta\sin\theta)^2} = \frac{c^2}{(1 + \varepsilon\cos\theta + \delta\sin\theta)^2}$$

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, and we are done. □

The above result is not the end of the story, because there is still some discussion to be made, in relation with degeneration. There is as well a discussion as well concerning normalization, because in the Kepler problem we assumed M to be fixed at 0. However, when changing coordinates via a translation, we can obtain in this way all conics.

Finally, from a physical perspective, that of concretely solving the gravity equation, there is a long discussion, and lots of additional formulae, regarding the trajectory and its parameters, as functions of the initial data. Without getting into full details here, let us record however the following result, coming as a useful version of Theorem 11.8:

THEOREM 11.9. *In the context of a 2-body problem, with M fixed at 0, and m starting its movement from Ox, the equation of motion of m, namely*

$$\ddot{z}=-\frac{Kz}{||z||^3}$$

with $K=GM$, and $z=(x,y)$, becomes in polar coordinates, $x=r\cos\theta$, $y=r\sin\theta$,

$$\ddot{r}=\frac{\lambda^2}{r^2}\left(\frac{1}{r}-\frac{1}{c}\right)\quad,\quad\dot{\theta}=\frac{\lambda}{r^2}$$

for some $\lambda, c\in\mathbb{R}$, related by $\lambda^2=Kc$. The value of r in terms of θ is given by

$$r=\frac{c}{1+\varepsilon\cos\theta+\delta\sin\theta}$$

for some $\varepsilon,\delta\in\mathbb{R}$. At the level of the affine coordinates x, y, this means

$$x=\frac{c\cos\theta}{1+\varepsilon\cos\theta+\delta\sin\theta}\quad,\quad y=\frac{c\sin\theta}{1+\varepsilon\cos\theta+\delta\sin\theta}$$

with $\theta=\theta(t)$ being subject to $\dot{\theta}=\lambda^2/r$, as above. Finally, we have

$$x^2+y^2=(\varepsilon x+\delta y-c)^2$$

which is a degree 2 equation, and so the resulting trajectory is a conic.

PROOF. This is a sort of "best of" the formulae found in the proof of Theorem 11.8. And in the hope of course that we have not forgotten anything. Finally, let us mention that the simplest illustration for this is the circular motion, and for details on this, not included in the above, we refer to the proof of Theorem 11.8. □

As a first concrete question, we would like to understand how the various parameters appearing above, namely $\lambda, c, \varepsilon, \delta$, which via some basic math can only tell us more about the shape of the orbit, appear from the initial data. The formulae here are as follows:

PROPOSITION 11.10. *In the context of Theorem 11.9, and in polar coordinates, $x = r\cos\theta$, $y = r\sin\theta$, the initial data is as follows, with $R = r_0$:*

$$r_0 = \frac{c}{1+\varepsilon} \quad , \quad \theta_0 = 0$$

$$\dot{r}_0 = -\frac{\delta\sqrt{K}}{\sqrt{c}} \quad , \quad \dot{\theta}_0 = \frac{\sqrt{Kc}}{R^2}$$

$$\ddot{r}_0 = \frac{\varepsilon K}{R^2} \quad , \quad \ddot{\theta}_0 = \frac{4\delta K}{R^2}$$

The corresponding formulae for the affine coordinates x, y can be deduced from this. Also, the various motion parameters c, ε, δ and $\lambda = \sqrt{Kc}$ can be recovered from this data.

PROOF. We have several assertions here, the idea being as follows:

(1) As mentioned in Theorem 11.9, the object m begins its movement on Ox. Thus we have $\theta_0 = 0$, and from this we get the formula of r_0 in the statement.

(2) Regarding the initial speed now, the formula of $\dot{\theta}_0$ follows from:

$$\dot{\theta} = \frac{\lambda}{r^2} = \frac{\sqrt{Kc}}{r^2}$$

Also, in what concerns the radial speed, the formula of $\dot{r}_0$ follows from:

$$\begin{aligned}
\dot{r} &= \frac{c(\varepsilon\sin\theta - \delta\cos\theta)\dot{\theta}}{(1+\varepsilon\cos\theta+\delta\sin\theta)^2} \\
&= \frac{c(\varepsilon\sin\theta-\delta\cos\theta)}{c^2/r^2}\cdot\frac{\sqrt{Kc}}{r^2} \\
&= \frac{\sqrt{K}(\varepsilon\sin\theta-\delta\cos\theta)}{\sqrt{c}}
\end{aligned}$$

(3) Regarding now the initial acceleration, by using $\dot{\theta} = \sqrt{Kc}/r^2$ we find:

$$\ddot{\theta} = -2\sqrt{Kc}\cdot\frac{2r\dot{r}}{r^3} = -\frac{4\sqrt{Kc}\cdot\dot{r}}{r^2}$$

In particular at $t=0$ we obtain the formula in the statement, namely:

$$\ddot{\theta}_0=-\frac{4\sqrt{Kc}\cdot\dot{r}_0}{R^2}=\frac{4\sqrt{Kc}}{R^2}\cdot\frac{\delta\sqrt{K}}{\sqrt{c}}=\frac{4\delta K}{R^2}$$

(4) Also regarding acceleration, with $\lambda=\sqrt{Kc}$ our main motion formula reads:

$$\ddot{r}=\frac{Kc}{r^2}\left(\frac{1}{r}-\frac{1}{c}\right)$$

In particular at $t=0$ we obtain the formula in the statement, namely:

$$\ddot{r}_0=\frac{Kc}{R^2}\left(\frac{1}{R}-\frac{1}{c}\right)=\frac{Kc}{R^2}\cdot\frac{\varepsilon}{c}=\frac{\varepsilon K}{R^2}$$

(5) Finally, the last assertion is clear, and since the formulae look better anyway in polar coordinates than in affine coordinates, we will not get into details here. □

With the above formulae in hand, which are a precious complement to Theorem 11.9, we can do some reverse engineering at the level of parameters, and work out how various inital speeds and accelerations lead to various types of conics. The computations here are very interesting, and we will leave them as an instructive exercise.

11c. Algebraic manifolds

We have seen so far the mathematical and physical meaning of the conics, which are the simplest curves around, with in both cases, very simple answers. The continuation of the story, however, is more complicated, because beyond conics, things ramify. A first idea, in order to generalize the conics, is to look at zeros of arbitrary polynomials:

$$P(x,y)=0\quad,\quad P\in\mathbb{R}[x,y]$$

More generally, we can look at zeros of polynomials in arbitrary N dimensions:

$$P(x_1,\ldots,x_N)=0\quad,\quad P\in\mathbb{R}[x_1,\ldots,x_N]$$

Observe that, at $N\geq 3$, what we have is not exactly a curve, but rather some sort of $(N-1)$-dimensional surface, called algebraic hypersurface. Due to this, in order to have a full collection of beasts, of all possible dimensions, we must intersect such algebraic hypersurfaces. We are led in this way to zeros of families of polynomials, as follows:

Definition 11.11. *An algebraic manifold is a space of the form*

$$X=\left\{(x_1,\ldots,x_N)\in\mathbb{R}^N\middle|P_i(x_1,\ldots,x_N)=0,\forall i\right\}$$

with $P_i\in\mathbb{R}[x_1,\ldots,x_N]$ being a family of polynomials.

And, good news, this is the good definition, and with the branch of mathematics studying such manifolds being called algebraic geometry. However, it is in fact possible to do even more generally, by looking at algebraic manifolds defined over an arbitrary field k, by using a family of polynomials $P_i \in k[x_1, \ldots, x_N]$, as follows:

$$X = \left\{(x_1, \ldots, x_N) \in k^N \middle| P_i(x_1, \ldots, x_N) = 0, \forall i\right\}$$

These ideas are very old, going back to the old Greeks, and there are many things that can be said about algebraic geometry, especially in its "arithmetic" version, over arbitrary fields k, where the theory really shines, with virtually every known advanced result in number theory having been obtained in this way, via algebraic geometry.

Instead of pursuing with usual, affine geometry, which can quickly escalate into fairly complicated things, let us take a look at projective geometry too, which is something fun, and interesting, and quite often more fun and interesting than affine geometry itself.

You might have heard of not of projective geometry. In case you didn't yet, the general principle is that "this is the wonderland where parallel lines cross". Which might sound a bit crazy, and not very realistic, but take a picture of some railroad tracks, and look at that picture. Do these parallel railroad tracks cross, on the picture? Sure they do. So, we are certainly not into abstractions here, but rather into serious science. QED.

Mathematically now, here are some axioms, to start with:

DEFINITION 11.12. *A projective space is a space consisting of points and lines, subject to the following conditions:*

(1) *Each* 2 *points determine a line.*
(2) *Each* 2 *lines cross, on a point.*

As a basic example we have the usual projective plane $P^2_{\mathbb{R}}$, which is best seen as being the space of lines in $\mathbb{R}^3$ passing through the origin. To be more precise, let us call each of these lines in $\mathbb{R}^3$ passing through the origin a "point" of $P^2_{\mathbb{R}}$, and let us also call each plane in $\mathbb{R}^3$ passing through the origin a "line" of $P^2_{\mathbb{R}}$. Now observe the following:

(1) Each 2 points determine a line. Indeed, 2 points in our sense means 2 lines in $\mathbb{R}^3$ passing through the origin, and these 2 lines obviously determine a plane in $\mathbb{R}^3$ passing through the origin, namely the plane they belong to, which is a line in our sense.

(2) Each 2 lines cross, on a point. Indeed, 2 lines in our sense means 2 planes in $\mathbb{R}^3$ passing through the origin, and these 2 planes obviously determine a line in $\mathbb{R}^3$ passing through the origin, namely their intersection, which is a point in our sense.

Thus, what we have is a projective space in the sense of Definition 11.12. More generally, we have the following construction, in arbitrary dimensions:

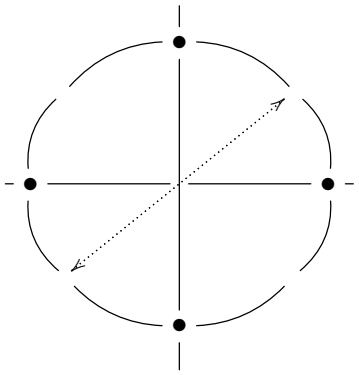
THEOREM 11.13. *We can define the projective space $P_{\mathbb{R}}^{N-1}$ as being the space of lines in $\mathbb{R}^N$ passing through the origin, and in small dimensions:*

(1) *$P_{\mathbb{R}}^1$ is the usual circle.*
(2) *$P_{\mathbb{R}}^2$ is some sort of twisted sphere.*

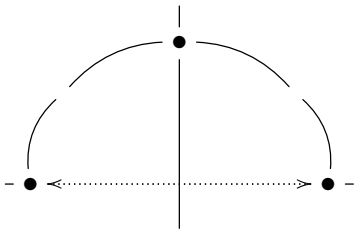
PROOF. We have several assertions here, with all this being of course a bit informal, and self-explanatory, the idea and some further details being as follows:

(1) To start with, the fact that the space $P_{\mathbb{R}}^{N-1}$ constructed in the statement is indeed a projective space in the sense of Definition 11.12 follows from definitions, exactly as in the discussion preceding the statement, regarding the case $N = 3$.

(2) At $N = 2$ now, a line in $\mathbb{R}^2$ passing through the origin corresponds to 2 opposite points on the unit circle $\mathbb{T} \subset \mathbb{R}^2$, according to the following scheme:

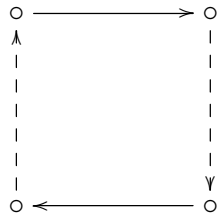


Thus, $P_{\mathbb{R}}^1$ corresponds to the upper semicircle of $\mathbb{T}$, with the endpoints identified, and so we obtain a circle, $P_{\mathbb{R}}^1 = \mathbb{T}$, according to the following scheme:

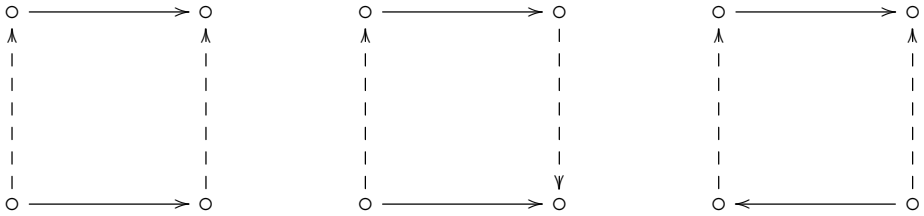


(3) At $N = 3$, the space $P_{\mathbb{R}}^2$ corresponds to the upper hemisphere of the sphere $S_{\mathbb{R}}^2 \subset \mathbb{R}^3$, with the points on the equator identified via $x = -x$. Topologically speaking, we can deform if we want the hemisphere into a square, with the equator becoming the boundary of this square, and in this picture, the $x = -x$ identification corresponds to a

"identify opposite edges, with opposite orientations" folding method for the square:



(4) Thus, we have our space. In order to understand now what this beast is, let us look first at the other 3 possible methods of folding the square, which are as follows:



Regarding the first space, the one on the left, things here are quite simple. Indeed, when identifying the solid edges we get a cylinder, and then when further identifying the dotted edges, what we get is some sort of closed cylinder, which is a torus.

(5) Regarding the second space, the one in the middle, things here are more tricky. Indeed, when identifying the solid edges we get again a cylinder, but then when further identifying the dotted edges, we obtain some sort of "impossible" closed cylinder, called Klein bottle. This Klein bottle obviously cannot be drawn in 3 dimensions, but with a bit of imagination, you can see it, in its full splendor, in 4 dimensions.

(6) Finally, regarding the third space, the one on the right, we know by symmetry that this must be the Klein bottle too. But we can see this as well via our standard folding method, namely identifying solid edges first, and dotted edges afterwards. Indeed, we first obtain in this way a Möbius strip, and then, well, the Klein bottle.

(7) With these preliminaries made, and getting back now to the projective space $P_{\mathbb{R}}^2$, we can see that this is something more complicated, of the same type, reminding the torus and the Klein bottle. So, we will call it "sort of twisted sphere", as in the statement, and exercise for you to imagine how this beast looks like, in 4 dimensions. □

All this is quite exciting, and reminds childhood and primary school, but is however a bit tiring for our neurons, guess that is pure mathematics. It is possible to come up with some explicit formulae for the embedding $P_{\mathbb{R}}^2 \subset \mathbb{R}^4$, which are useful in practice, allowing us to do some analysis over $P_{\mathbb{R}}^2$, and we will leave this as an instructive exercise.

There is some linear algebra to be done here too, by identifying the lines in $\mathbb{R}^N$ with the corresponding rank 1 projections, along with many other things, and we have:

THEOREM 11.14. *The projective space $P_{\mathbb{R}}^{N-1}$ can be thought of as being the space of rank 1 projections in the matrix algebra $M_N(\mathbb{R})$, given by*

$$P_x = \frac{1}{||x||^2}(x_i x_j)_{ij}$$

by identifying the lines in $\mathbb{R}^N$ passing through the origin with the corresponding rank 1 projections in $M_N(\mathbb{R})$, in the obvious way.

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

(1) The main assertion is more or less clear from definitions, the point being that the lines in $\mathbb{R}^N$ passing through the origin are obviously in bijection with the corresponding rank 1 projections. Thus, we obtain the interpretation of $P_{\mathbb{R}}^{N-1}$ in the statement.

(2) Regarding now the formula of the rank 1 projections, which is a must-know, for this, and in everyday life, consider a vector $y \in \mathbb{R}^N$. Its projection on $\mathbb{R}x$ must be a certain multiple of x, and we are led in this way to the following formula:

$$P_x y = \frac{< y, x >}{< x, x >} x = \frac{1}{||x||^2} < y, x > x$$

(3) But with this in hand, we can now compute the entries of P_x, as follows:

$$\begin{aligned}(P_x)_{ij} &= < P_x e_j, e_i > \\ &= \frac{1}{||x||^2} < e_j, x >< x, e_i > \\ &= \frac{x_j x_i}{||x||^2}\end{aligned}$$

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

All this is very interesting, but we will pause our study here, because we still have many other things to say. Getting now to finite fields, we have:

THEOREM 11.15. *Given a field k, we can talk about the projective space P_k^{N-1}, as being the space of lines in k^N passing through the origin, having cardinality*

$$|P_k^{N-1}| = q^2 + q + 1$$

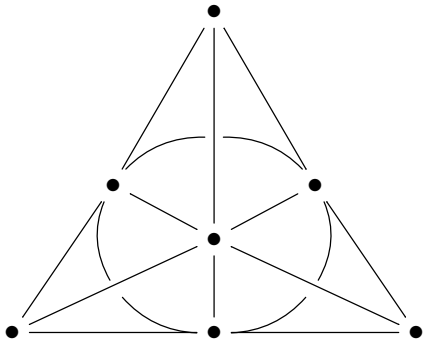
where $q = |k|$, in the case where our field k is finite.

PROOF. This is indeed clear from definitions, with the cardinality coming from:

$$|P_k^{N-1}| = \frac{|k^3 - \{0\}|}{|k - \{0\}|} = \frac{q^3 - 1}{q - 1} = q^2 + q + 1$$

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

As an example, let us see what happens for the simplest finite field that we know, namely $k = \mathbb{Z}_2$. Here our projective plane, having $4 + 2 + 1 = 7$ points, and 7 lines, is a famous combinatorial object, called Fano plane, which is depicted as follows:



Here the circle in the middle is by definition a line, and with this convention, the basic axioms in Definition 11.12 are satisfied, in the sense that any two points determine a line, and any two lines determine a point. And isn't this beautiful. Let us record:

CONCLUSION 11.16. *For getting started with geometry, all you need is* 7 *points.*

So long for algebraic geometry, real, complex or over arbitrary fields, and affine or projective. For more, a good reference here is the book by Harris [45].

Finally, no discussion about algebraic geometry would be complete without a look into algebraic topology. We have already seen, in the proof of Theorem 11.13, that "shape", taken in a basic topological sense, matters. So, let us further explore this.

Forgetting about manifolds, let us start with something that we know, namely:

DEFINITION 11.17. *A topological space X is called connected when any two points $x, y \in X$ can be connected by a path. That is, given any two points $x, y \in X$, we can find a continuous function $f : [0, 1] \to X$ such that $f(0) = x$ and $f(1) = y$.*

The problem is now, given a connected space X, how to count its "holes". And this is quite subtle problem, because as examples of such spaces we have:

(1) The sphere, the donut, the double-holed donut, the triple-holed donut, and so on. These spaces are quite simple, and intuition suggests to declare that the number of holes of the N-holed donut is, and you guessed right, N.

(2) However, we have as well as example the empty sphere, I mean just the crust of the sphere, and while this obviously falls into the class of "one-holed spaces", this is not the same thing as a donut, its hole being of different nature.

(3) As another example, consider again the sphere, but this time with two tunnels drilled into it, in the shape of a cross. Whether that missing cross should account for 1 hole, or for 2 holes, or for something in between, I will leave it up to you.

Summarizing, things are quite tricky, suggesting that the "number of holes" of a topological space X is not an actual number, but rather something more complicated. Now with this in mind, let us formulate the following definition:

DEFINITION 11.18. *The homotopy group $\pi_1(X)$ of a connected space X is the group of loops based at a given point $* \in X$, with the following conventions,*

(1) *Two such loops are identified when one can pass continuously from one loop to the other, via a family of loops indexed by $t \in [0, 1]$,*
(2) *The composition of two such loops is the obvious one, namely is the loop obtaining by following the first loop, then the second loop,*
(3) *The unit loop is the null loop at $*$, which stays there, and the inverse of a given loop is the loop itself, followed backwards,*

with the remark that the group $\pi_1(X)$ defined in this way does not depend on the choice of the given point $* \in X$, where the loops are based.

Here the fact that $\pi_1(X)$ defined in this way is indeed a group is obvious, and obvious as well is the fact that, since X is assumed to be connected, this group does not depend on the choice of the given point $* \in X$, where the loops are based.

As basic examples, for spaces having "no holes", such as $\mathbb{R}$ itself, or $\mathbb{R}^N$, and so on, we have $\pi_1 = \{1\}$. In fact, having no holes can only mean, by definition, that $\pi_1 = \{1\}$. As further illustrations, here are now a few basic computations:

THEOREM 11.19. *We have the following computations of homotopy groups:*

(1) *For the circle, we obtain $\pi_1 = \mathbb{Z}$.*
(2) *For the torus, we obtain $\pi_1 = \mathbb{Z} \times \mathbb{Z}$.*
(3) *For the disk minus 2 points, we have $\pi_1 = \mathbb{Z} * \mathbb{Z}$.*

PROOF. These results are all standard, as follows:

(1) The first assertion is clear, because a loop on the circle must wind $n \in \mathbb{Z}$ times around the center, and this parameter $n \in \mathbb{Z}$ uniquely determines the loop, up to the identification in Definition 11.18. Thus, the homotopy group of the circle is the group of such parameters $n \in \mathbb{Z}$, which is of course the group $\mathbb{Z}$ itself.

(2) In what regards now the second assertion, the torus being a product of two circles, we are led to the conclusion that its homotopy group must be some kind of product of $\mathbb{Z}$ with itself. But pictures show that the two standard generators of $\mathbb{Z}$, and so the two copies of $\mathbb{Z}$ themselves, commute, $gh = hg$, and so we obtain the product of $\mathbb{Z}$ with itself, subject to commutation, which is the usual product $\mathbb{Z} \times \mathbb{Z}$.

(3) This is quite clear, because the homotopy group is generated by the 2 loops around the 2 missing points, which are obviously free, algebrically speaking. Thus, we obtain a free product of the group $\mathbb{Z}$ with itself, also known as free group on 2 generators. □

There are many other interesting things that can be said about homotopy groups. Also, another thing that can be done with the arbitrary spaces X, again in relation with studying their "shape", is that of looking at the fiber bundles over them, again up to continuous deformation. We are led in this way into a group, called $K_0(X)$. Moreover, both $\pi_1(X)$ and $K_0(X)$ have in fact higher analogues $\pi_n(X)$ and $K_n(X)$ as well, and the general goal of algebraic topology is that of understanding all these groups.

But all this, obviously, starts to become too complicated. So, leaving now the general manifolds and topological spaces aside, let us focus now on the simplest objects of topology, namely the knots, with this meaning the smooth closed curves in $\mathbb{R}^3$:

DEFINITION 11.20. *A knot is a smooth closed curve in $\mathbb{R}^3$, regarded modulo smooth transformations.*

And isn't this a beautiful definition. We are here at the core of everything that can be called "geometry", and in fact, thinking a bit on how knots can be knotted, in so many fascinating ways, we are led to the following philosophical conclusion:

CONCLUSION 11.21. *Knots are to geometry what prime numbers are to number theory.*

At the level of questions now, once we have a closed curve, say given via its algebraic equations, can we decide if is tied or not, and if tied, how complicated is it tied, how to untie it, and so on? But these are, obviously, quite difficult questions.

Perhaps simpler now, experience with cables and ropes shows that a random closed curve is usually tied. But can we really prove this? Once again, difficult question. So, we will stop here, and exercise for you, to say something non-trivial about knots.

11d. Differential manifolds

Back now to applied mathematics, involving analysis and physics, the situation here is a bit different. Although many interesting algebraic manifolds appear at the advanced level, making algebraic geometry a key tool in advanced physics, in what concerns the basics, here we are mostly in need of a different definition, as follows:

DEFINITION 11.22. *A smooth manifold is a space X which is locally isomorphic to $\mathbb{R}^N$. To be more precise, this space X must be covered by charts, bijectively mapping open pieces of it to open pieces of $\mathbb{R}^N$, with the changes of charts being C^∞ functions.*

As a basic example, we have $\mathbb{R}^N$ itself, or any open subset $X \subset \mathbb{R}^N$. Another example is the circle, or curves like ellipses and so on, for obvious reasons. To be more precise, the unit circle can be covered by 2 charts as above, by using polar coordinates, in the obvious

way, and then by applying dilations, translations and other such transformations, namely bijections which are smooth, we obtain a whole menagery of circle-looking manifolds.

In particular, we can see from this that Definition 11.22 is a serious rival to Definition 11.11, because both generalize, in a natural way, the conics that we know well.

So, this is the situation, we have two geometric disciplines inspired by the old findings of the old Greeks, regarding the conics, called algebraic geometry and differential geometry. With algebraic geometry being substantially older, but in what concerns us, we will rather go for differential geometry, which is more relevant to basic modern physics.

Going back now to Definition 11.22 as stated, let us first explore the basic examples. We have already talked about them in the above, but all that discussion needed to be accompanied by some proofs, and here is a more precise statement in this sense:

THEOREM 11.23. *The following are smooth manifolds, in the plane:*

(1) *The circles.*
(2) *The ellipses.*
(3) *The non-degenerate conics.*
(4) *Smooth deformations of these.*

PROOF. All this is quite intuitive, the idea being as follows:

(1) Consider the unit circle, $x^2 + y^2 = 1$. We can write then $x = \cos t$, $y = \sin t$, with $t \in [0, 2\pi)$, and we seem to have here the solution to our problem, just using 1 chart. But this is of course wrong, because $[0, 2\pi)$ is not open, and we have a problem at 0. In practice we need to use 2 such charts, say with the first one being with $t \in (0, 3\pi/2)$, and the second one being with $t \in (\pi, 5\pi/2)$. As for the fact that the change of charts is indeed smooth, this comes by writing down the formulae, or just thinking a bit, and arguing that this change of chart being actually a translation, it is automatically linear.

(2) This follows from (1), by pulling the circle in both the Ox and Oy directions, and the formulae here, based on Theorem 11.6 or Theorem 11.7, are left to you reader.

(3) We already have the ellipses, and the case of the parabolas and hyperbolas is elementary as well, and in fact simpler than the case of the ellipses. Indeed, a parablola is clearly homeomorphic to $\mathbb{R}$, and a hyperbola, to two copies of $\mathbb{R}$.

(4) This is something which is clear too, depending of course on what exactly we mean by "smooth deformation", and by using a bit of multivariable calculus if needed. □

In higher dimensions now, as basic examples here, we have the unit sphere in $\mathbb{R}^N$, and smooth deformations of it, once again, somehow by obvious reasons. In case you are wondering on how to construct explicit charts for the sphere, the answer comes from:

THEOREM 11.24. *We have spherical coordinates in N dimensions,*

$$\begin{cases} x_1 &= r\cos t_1 \\ x_2 &= r\sin t_1\cos t_2 \\ \vdots & \\ x_{N-1} &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\cos t_{N-1} \\ x_N &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\sin t_{N-1} \end{cases}$$

with this guaranteeing that the sphere is indeed a smooth manifold.

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

(1) The fact that we have indeed spherical coordinates is clear, with the only point to be clarified being the identification of the precise ranges of the angles, which follows from some geometric thinking, first at $N = 2, 3$, and then in general.

(2) With the remark that we use here mathematicians' convention for the angles, which works nicely in N arbitrary dimensions, as opposed to physicists' convention, which works best at $N = 3$. And with this being something quite subjective, because mathematicians' convention is based on physicists' finding that the more dimensions, the better.

(3) Finally, in what regards the last assertion, this can proved a bit like for the circle, as we did in the proof of Theorem 11.23 (1), basically by cutting the sphere into 2^N parts, and we will leave the details here as an instructive exercise. □

Summarizing, we have the spheres as main examples of differential manifolds, as expected, but we can also see from the above that the technical verification of the manifold axioms is something quite complicated. Welcome to differential geometry, where nothing is really trivial, I mean even things which are supposed to be trivial are not.

In relation with these questions, namely parametrizing the spheres, we have the stereographic projection as well, which works more directly, as follows:

THEOREM 11.25. *The stereographic projection is given by inverse maps*

$$\Phi : \mathbb{R}^N \to S^N_{\mathbb{R}} - \{\infty\} \quad , \quad \Psi : S^N_{\mathbb{R}} - \{\infty\} \to \mathbb{R}^N$$

given by the following formulae,

$$\Phi(v) = (1, 0) + \frac{2}{1 + ||v||^2}(-1, v) \quad , \quad \Psi(c, x) = \frac{x}{1 - c}$$

with the convention $\mathbb{R}^{N+1} = \mathbb{R} \times \mathbb{R}^N$, and with the coordinate of $\mathbb{R}$ denoted x_0, and with the coordinates of $\mathbb{R}^N$ denoted $x_1, \ldots, x_N$.

PROOF. We are looking for the formulae of the isomorphism $\mathbb{R}^N \simeq S^N_{\mathbb{R}} - \{\infty\}$, obtained by identifying $\mathbb{R}^N = \mathbb{R}^N \times \{0\} \subset \mathbb{R}^{N+1}$ with the unit sphere $S^N_{\mathbb{R}} \subset \mathbb{R}^{N+1}$, with the convention that the point which is added is $\infty = (1, 0, \ldots, 0)$, via the stereographic projection. That is, we need the precise formulae of two inverse maps, as follows:

$$\Phi : \mathbb{R}^N \to S^N_{\mathbb{R}} - \{\infty\} \quad , \quad \Psi : S^N_{\mathbb{R}} - \{\infty\} \to \mathbb{R}^N$$

In one sense, according to our conventions above, we must have a formula as follows for our map Φ, with the parameter $t \in (0, 1)$ being such that $||\Phi(v)|| = 1$:

$$\Phi(v) = t(0, v) + (1 - t)(1, 0)$$

The equation for the parameter $t \in (0, 1)$ can be solved as follows:

$$\begin{aligned} (1 - t)^2 + t^2||v||^2 = 1 &\iff t^2(1 + ||v||^2) = 2t \\ &\iff t = \frac{2}{1 + ||v||^2} \end{aligned}$$

We conclude that the formula of the map Φ is as follows:

$$\Phi(v) = (1, 0) + \frac{2}{1 + ||v||^2}(-1, v)$$

In the other sense now we must have, for a certain $\alpha \in \mathbb{R}$:

$$(0, \Psi(c, x)) = \alpha(c, x) + (1 - \alpha)(1, 0)$$

But from $\alpha c + 1 - \alpha = 0$ we get the following formula for the parameter α:

$$\alpha = \frac{1}{1 - c}$$

We conclude that the formula of the map Ψ is as follows:

$$\Psi(c, x) = \frac{x}{1 - c}$$

Here, as before, we use the convention in the statement, namely $\mathbb{R}^{N+1} = \mathbb{R} \times \mathbb{R}^N$, with the coordinate of $\mathbb{R}$ denoted x_0, and with the coordinates of $\mathbb{R}^N$ denoted $x_1, \ldots, x_N$. □

There are of course many other possible parametrizations of the sphere, such as the one using cylindrical coordinates, or the one peeling the sphere as an orange, and so on. All this is quite interesting, and as question here, of practical interest, we have:

QUESTION 11.26. *What is the best parametrization of the unit sphere in $\mathbb{R}^3$, for purely mathematical reasons? What about for cartography reasons?*

To be more precise, the first question, which is quite challenging, is that of finding the simplest proof ever for the fact that the sphere in $\mathbb{R}^3$ is indeed a smooth manifold. As for the second question, which is even more challenging, and not really solved by mankind, despite centuries of work, and many bright ideas, the problem here is to have your charts reflecting as nice as possible useful things such as lengths, angles and areas.

By the way, speaking lengths, angles and areas, observe that the general differential manifold formalism from Definition 11.22 is obviously too broad for talking about these. In order to do so, several more axioms must be added, and we end up with something called Riemannian manifold, which is the main object of study of advanced differential geometry. And, regarding such these manifolds, there are many deep theorems, including a key result of Nash, stating that we can always find an embedding, as follows:

$$X \subset \mathbb{R}^N$$

Thus, all in all, all this leads us into the good old $\mathbb{R}^N$, and multivariable calculus. But the story here is quite long and technical, and it's getting late, and we will stop here.

So long for geometry, in a large sense. As a conclusion to all this, geometry comes in many flavors, algebraic or differential, Riemannian or not, over $\mathbb{R}$, or $\mathbb{C}$, or some other field k, and in addition to this we can talk about affine or projective geometry, or about discrete or continuous geometry, and so on. Many interesting things, and if excited by all this, orient yourself towards physics, where geometry in all its flavors is needed.

11e. Exercises

This was a pure mathematics chapter, save for some old school computations coming from Kepler and Newton, and as exercises, of pure mathematics type, we have:

EXERCISE 11.27. *Learn some algebraic geometry.*

EXERCISE 11.28. *Learn some differential geometry.*

EXERCISE 11.29. *Learn some Riemannian geometry.*

EXERCISE 11.30. *Learn some symplectic geometry too.*

With apologies for this, but that is how things go in pure mathematics, "learn" being the keyword. As bonus exercise, more concrete, have some fun with cartography.

CHAPTER 12

Higher derivatives

12a. Higher derivatives

Welcome back to analysis, and in the hope that you survived the previous chapter. Goow news, in this chapter we go for the real thing, namely the computation of maxima and minima of functions. Want to optimize the characteristics of your building, bridge, plane, engine, medicine, computer, or casino and other racketeering operations? It all comes down, you guessed right, to maximizing or minimizing a certain function.

In one variable things are quite easy, at least from a theoretical viewpoint, due to the following result, that we know well from Part I of the present book:

THEOREM 12.1. *The one-variable smooth functions are subject to the Taylor formula*

$$f(x+t)=\sum_{k=0}^{\infty}\frac{f^{(k)}(x)}{k!}\,t^k$$

which allows, via suitable truncations, to determine the local maxima and minima.

PROOF. This is a compact summary of what we know from Part I, with everything being in fact quite technical, and with the idea being as follows:

(1) In order to compute the local maxima and minima, a first method is by using the following formula, which comes straight from the definition of the derivative:

$$f(x+t)\simeq f(x)+f'(x)t$$

Indeed, this formula shows 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. Thus, in order to find the local minima and maxima, we must compute first the points x satisfying $f'(x)=0$, and then perform a more detailed study of each solution x that we found.

(2) In relation with the problems left, the second derivative comes to the rescue. Indeed, we can use the following more advanced formula, coming via l'Hôpital's rule:

$$f(x+t)\simeq f(x)+f'(x)t+\frac{f''(x)}{2}\,t^2$$

To be more precise, assume that we have $f'(x) = 0$, as required by the study in (1). Then this second order formula simply reads:

$$f(x+t) \simeq f(x) + \frac{f''(x)}{2}\,t^2$$

But this is something very useful, telling us that when $f''(x) < 0$, what we have is a local maximum, and when $f''(x) > 0$, what we have is a local minimum. As for the remaining case, that when $f''(x) = 0$, things here remain open.

(3) All this is very useful in practice, and with what we have in (1), complemented if needed with what we have in (2), we can in principle compute the local minima and maxima, without much troubles. However, if really needed, more tools are available. Indeed, we can use if we want the order 3 Taylor formula, which is as follows:

$$f(x+t) \simeq f(x) + f'(x)t + \frac{f''(x)}{2}\,t^2 + \frac{f'''(x)}{6}\,t^3$$

To be more precise, assume that we are in the case $f'(x) = f''(x) = 0$, which is where our joint algorithm coming from (1) and (2) fails. In this case, our formula becomes:

$$f(x+t) \simeq f(x) + \frac{f'''(x)}{6}\,t^3$$

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 formulate now an abstract result about this, we can use the Taylor formula at order n:

$$f(x+t) \simeq \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\,t^k$$

Indeed, assume that we started to compute the derivatives $f'(x), f''(x), f''(x), \ldots$ of our function at the point x, with the goal of finding the first such derivative which does not vanish, and we found this derivative, as being the order n one:

$$f'(x) = f''(x) = \ldots = f^{(n-1)}(x) = 0 \quad , \quad f^{(n)}(x) \neq 0$$

Then, the Taylor formula at x at order n takes the following form:

$$f(x+t) \simeq f(x) + \frac{f^{(n)}(x)}{n!}\,t^n$$

But this is exactly what we need, in order to fully solve our local extremum problem. Indeed, when n is even, if $f^{(n)}(x) < 0$ what we have is a local maximum, and if $f^{(n)}(x) > 0$, what we have is a local minimum. As for the case where n is odd, here we cannot have a local minimum or maximum, due to $t \to -t$ which switches growth. $\square$

All the above, Theorem 12.1 and its full proof, must be of course perfectly known, when looking for applications of such things. However, for theoretical purposes, let us record as well, in a very compact form, what is basically to be remembered:

THEOREM 12.2. *Given a differentiable function $f : \mathbb{R} \to \mathbb{R}$, we can always write*

$$f(x+t) \simeq f(x) + \frac{f^{(n)}(x)}{n!}\, t^n$$

with $f^{(n)}(x) \neq 0$, and this tells us if x is a local minimum, or maximum of f.

PROOF. This was the conclusion of the proof of Theorem 12.1, and with the extra remark that local extremum means that n is even, with in this case $f^{(n)}(x) < 0$ corresponding to local maximum, and $f^{(n)}(x) > 0$ corresponding to local minimum. □

In several variables now, things will be quite tricky, making full use of the material that we learned in the previous 3 chapters, and even more, requiring some continuations of that. Indeed, we need a lot of knowledge, in order to solve our problems:

(1) The linear algebra from chapter 9 is the backbone of multivariable calculus, so we will surely need all of that, and more. For instance, since the first derivatives $f'(x)$ are now matrices, expect some sort of positivity theory for the matrices, to be needed.

(2) Regarding the partial derivatives from chapter 10, no question about it, all that material is certainly useful, and the more theory we have there, the better it will be, for our questions. The problem is that of iterating that partial derivative operations.

(3) Finally, regarding the geometry from chapter 11, the functions that we want to study will be naturally defined either on $\mathbb{R}^N$, or on spheres, ellipses, cylinders and other differential submanifolds $X \subset \mathbb{R}^N$, so we will certainly need that material too.

Getting started now, we can talk about higher derivatives, in the obvious way, simply by performing the operation of taking derivatives recursively. As result here, we have:

THEOREM 12.3. *Given a continuous function $f : \mathbb{R}^N \to \mathbb{R}$, we can talk about its higher derivatives, defined recursively as*

$$\frac{d^k f}{dx_{i_1} \ldots dx_{i_k}} = \frac{d}{dx_{i_1}} \cdots \frac{d}{dx_{i_k}}(f)$$

provided that all 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. There are several things going on here, the idea being as follows:

(1) First of all, we can talk about the quantities in the statement, with the remark however that at each step of our recursion, the corresponding partial derivative can exist of not. We will say in what follows that our function is k times differentiable if the quantities in the statement exist at any $l \leq k$, and smooth, if this works with $k = \infty$.

(2) Regarding now the second assertion, this is something more tricky. Let us first recall from chapter 8 that the second derivatives of a twice differentiable function of two variable $f : \mathbb{R}^2 \to \mathbb{R}$ are subject to the Clairaut formula, namely:

$$\frac{d^2 f}{dxdy} = \frac{d^2 f}{dydx}$$

(3) But this result clearly extends to our function $f : \mathbb{R}^N \to \mathbb{R}$, simply by ignoring the unneeded variables, so we have the Clairaut formula in general, also called Schwarz formula, which is the one in the statement, namely:

$$\frac{d^2 f}{dx_i dx_j} = \frac{d^2 f}{dx_j dx_i}$$

(4) Now observe that this tells us that the order in which the higher derivatives are computed is irrelevant. That is, we can permute the order of our partial derivative computations, and a standard way of doing this is by differentiating first with respect to x_1, as many times as needed, then with respect to x_2, and so on. Thus, the collection of our partial derivatives can be written, in a more convenient form, as follows:

$$\frac{d^k f}{dx_1^{k_1} \ldots dx_N^{k_N}} = \frac{d^{k_1}}{dx_1^{k_1}} \cdots \frac{d^{k_N}}{dx_{i_N}^{k_N}}(f)$$

(5) To be more precise, here $k \in \mathbb{N}$ is as usual the global order of our derivatives, the exponents $k_1, \ldots, k_N \in \mathbb{N}$ are subject to the condition $k_1 + \ldots + k_N = k$, and the operations on the right are the familiar one-variable higher derivative operations.

(6) This being said, for certain tricky questions it is more convenient not to order the indices, or rather to order them according to what order best fits your computation, so what we have in the statement is the good formula, and (4-5) are mere remarks.

(7) And with the remark that for trivial questions too, what we have in the statement is the good formula, simply because there are less indices to be written, when compared to what we have to write when using the ordering procedure in (4-5) above. □

All this is very nice, and as an illustration for the above, let us work out the case $k = 2$. Here things are quite special, and we can formulate the following definition:

DEFINITION 12.4. *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 our previous definition of the second derivative f'', and this because in this case, the 1×1 matrix from Definition 12.4 is:

$$f''(x) = (f''(x)) \in M_{1\times 1}(\mathbb{R})$$

As a word of warning, however, never use Definition 12.4 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 business, with these notions, we have the following question to be solved:

QUESTION 12.5. *What is the Taylor formula for a function*

$$f : \mathbb{R}^N \to \mathbb{R}$$

and how can this be used for computing the local minima and maxima of f?

We will solve this slowly, a bit as we did in the proof of Theorem 12.1, in the $N = 1$ case. Let us start with something that we know well from chapter 10, namely:

$$f(x + t) \simeq f(x) + f'(x)t$$

To be more precise, we know that this formula holds indeed, with the derivative $f'(x)$ being by definition the horizontal vector formed by the partial derivatives, and with $t \in \mathbb{R}^N$ being regarded as usual as column vector, the formula of $f'(x)t$ being:

$$\begin{aligned} f'(x)t &= \left(\frac{df}{dx_1} \ \cdots \ \frac{df}{dx_N} \right) \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \\ &= \sum_{i=1}^{N} \frac{df}{dx_i} t_i \\ &\in \mathbb{R} \end{aligned}$$

Here we have of course identified the 1×1 matrices with their numeric content. As a consequence, in analogy with what we know in 1 variable, we can formulate:

THEOREM 12.6. *The Taylor formula at order 1 for a function $f : \mathbb{R}^N \to \mathbb{R}$ is*

$$f(x+t) \simeq f(x) + f'(x)t$$

and in particular, in order for x to be a local extremum, we must have $f'(x) = 0$.

PROOF. Here the first assertion is something that we know, as explained above, and the second assertion follows from it. Indeed, let us look at the order 1 term, given by:

$$f'(x)t = \sum_{i=1}^{N} \frac{df}{dx_i}\, t_i$$

Now since this linear combination of the entries of $t \in \mathbb{R}^N$ can range among positives and negatives, unless all the coefficients are zero, which means $f'(x) = 0$, we are led to the conclusion that local extremum needs $f'(x) = 0$ to hold, as stated. □

Let us discuss now the Taylor formula at order 2. We have here:

THEOREM 12.7. *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 more tricky, the idea being as follows:

(1) As a first observation, at $N = 1$ the Hessian matrix as constructed in Definition 12.4 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 Part I, 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:

$$\begin{aligned} g''(r) &= (f'(x + ry) \cdot y)' \\ &= \left(\sum_i \frac{df}{dx_i}(x + ry) \cdot y_i \right)' \\ &= \sum_i \sum_j \frac{d^2 f}{dx_i dx_j}(x + ry) \cdot \frac{d(x + ry)_j}{dr} \cdot y_i \\ &= \sum_i \sum_j \frac{d^2 f}{dx_i dx_j}(x + ry) \cdot y_i y_j \\ &= < f''(x + ry)y, y > \end{aligned}$$

(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.

(5) Finally, for completness, let us record as well a more numeric formulation of what we found. According to our usual rules for matrix calculus, what we found is:

$$f(x + t) \simeq f(x) + \sum_{i=1}^{N} \frac{df}{dx_i} t_i + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{d^2 f}{dx_i dx_j} t_i t_j$$

Observe that, since the Hessian matrix $f''(x)$ is symmetric, most of the terms on the right will appear in pairs, making it clear what the $1/2$ is there for, namely avoiding redundancies. However, this is only true for the off-diagonal terms, so instead of further messing up our numeric formula above, we will just leave it like this. □

We can now go back to local extrema, and we have, improving Theorem 12.6:

THEOREM 12.8. *In order for a twice differentiable function $f : \mathbb{R}^N \to \mathbb{R}$ to have a local minimum or maximum at $x \in \mathbb{R}^N$, the first derivative must vanish there,*

$$f'(x) = 0$$

and the Hessian must be positive or negative, in the sense that the quantities

$$< f''(x)t, t > \in \mathbb{R}$$

must keep a constant sign, positive or negative, when $t \in \mathbb{R}^N$ varies.

PROOF. This is clear from Theorem 12.7. Consider indeed the formula there, namely:

$$f(x + t) \simeq f(x) + f'(x)t + \frac{< f''(x)t, t >}{2}$$

It is clear then that, in order for our function to have a local minimum or maximum at $x \in \mathbb{R}^N$, the first derivative must vanish there, $f'(x) = 0$. Moreover, with this assumption made, the approximation that we have around x becomes:

$$f(x + t) \simeq f(x) + \frac{< f''(x)t, t >}{2}$$

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

As a conclusion to our study so far, our analytic questions lead us into a linear algebra question, regarding the square matrices of type $f''(x) \in M_N(\mathbb{R})$, and more specifically the positivity properties of the following quantities, when $t \in \mathbb{R}^N$ varies:

$$< f''(x)t, t > \in \mathbb{R}$$

This is actually a quite subtle question, and many things can be said here, after some linear algebra work. We will be back to this in a moment.

Finally, at higher order things become more complicated, as follows:

THEOREM 12.9. *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 12.7, with everything coming from the usual Taylor formula, applied on:

$$I = [x, x + t]$$

We will leave this as an instructive exercise, for your long Summer nights. □

12b. Matrices, positivity

According to what we know so far, the local extrema of a function $f : \mathbb{R}^N \to \mathbb{R}$ come under the assumption $f'(x) = 0$, and depend on the positivity properties of the Hessian matrix $f''(x) \in M_N(\mathbb{R})$. In view of this, and of the Taylor formula at order 2 in general, the best method in order to deal with a function $f : \mathbb{R}^N \to \mathbb{R}$ satisfying $f'(x) = 0$ is that of changing the basis around $x \in \mathbb{R}^N$ as to have the Hessian matrix $f''(x) \in M_N(\mathbb{R})$ diagonalized, if this diagonalization is of course possible. Thus, forgetting now about functions and analysis, we are led into the following linear algebra question:

QUESTION 12.10. *Which symmetric matrices $A \in M_N(\mathbb{R})$ have the property*

$$< At, t > \geq 0$$

for any $t \in \mathbb{R}^N$? Also, when is a matrix $A \in M_N(\mathbb{R})$ diagonalizable?

Observe that in the second question we have lifted the assumption that A is symmetric, and this for good reason. Indeed, besides the local extremum problematics for the functions $f : \mathbb{R}^N \to \mathbb{R}$, discussed above, we have as well the quite interesting problem of studying the functions $f : \mathbb{R}^N \to \mathbb{R}^N$, whose derivatives are square matrices, $f'(x) \in M_N(\mathbb{R})$, waiting to be diagonalized too. So, we are trying here two shoot 2 rabbits at the same time, with our Question 12.10, as formulated above.

Obviously, hunting matters, so time to ask the expert. And the expert says:

CAT 12.11. *In the lack of speed and claws, yes, develop as much linear algebra as you can. Without even caring for applications, these will come naturally.*

Thanks cat, so this will be our plan for this section, develop as much linear algebra as we can. And for applications, we will most likely leave them to you, reader, for later in life, depending on the precise physics and engineering questions that you will be interested in. And with the advice of course to follow the feline way, relax, and no mercy.

With this plan made, let us go back to the diagonalization question, from chapter 9. We will need here diagonalization results which are far more powerful. We first have:

THEOREM 12.12. *Any matrix $A \in M_N(\mathbb{C})$ which is self-adjoint, $A = A^*$, is diagonalizable, with the diagonalization being of the following type,*

$$A = UDU^*$$

with $U \in U_N$, and with $D \in M_N(\mathbb{R})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because if we take a matrix of the form $A = UDU^*$, with U unitary and D diagonal and real, then we have:

$$A^* = (UDU^*)^* = UD^*U^* = UDU^* = A$$

In the other sense now, assume that A is self-adjoint, $A = A^*$. 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 now 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 $\lambda, \mu \in \mathbb{R}$:

$$\begin{aligned}\lambda < v, w > &= < \lambda v, w > \\ &= < Av, w > \\ &= < v, Aw > \\ &= < v, \mu w > \\ &= \mu < v, w >\end{aligned}$$

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 $\mathbb{C}^N$. 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. Moreover, since a matrix A is self-adjoint precisely when $< Av, v > \in \mathbb{R}$ for any vector $v \in \mathbb{C}^N$, as one can see by expanding the scalar product, the restriction of A to the subspace $v^\perp$ is self-adjoint. Thus, we can proceed by recurrence, and we obtain the result. □

Observe that, as a consequence of the above result, that you certainly might have heard of, any symmetric matrix $A \in M_N(\mathbb{R})$ is diagonalizable. In fact, we have:

PROPOSITION 12.13. *Any matrix $A \in M_N(\mathbb{R})$ which is symmetric, $A = A^t$, is diagonalizable, with the diagonalization being of the following type,*

$$A = UDU^t$$

with $U \in O_N$, and with $D \in M_N(\mathbb{R})$ diagonal. The converse holds too.

PROOF. As before, the converse trivially holds, because if we take a matrix of the form $A = UDU^t$, with U orthogonal and D diagonal and real, then we have $A^t = A$. In the other sense now, this follows from Theorem 12.12, and its proof. □

As basic examples of self-adjoint matrices, we have the orthogonal projections:

PROPOSITION 12.14. *The matrices $P \in M_N(\mathbb{C})$ which are projections, $P^2 = P = P^*$, are precisely those which diagonalize as follows,*

$$P = UDU^*$$

with $U \in U_N$, and with $D \in M_N(0,1)$ being diagonal.

PROOF. Since we have $P = P^*$, by using Theorem 12.12, the eigenvalues must be real. Then, by using $P^2 = P$, assuming that we have $Pv = \lambda v$, we obtain:

$$\begin{aligned}
\lambda < v, v > &= < \lambda v, v > \\
&= < Pv, v > \\
&= < P^2 v, v > \\
&= < Pv, Pv > \\
&= < \lambda v, \lambda v > \\
&= \lambda^2 < v, v >
\end{aligned}$$

We therefore have $\lambda \in \{0, 1\}$, as claimed, and as a final conclusion here, the diagonalization of the self-adjoint matrices is as follows, with $e_i \in \{0, 1\}$:

$$P \sim \begin{pmatrix} e_1 & & \\ & \ddots & \\ & & e_N \end{pmatrix}$$

To be more precise, the number of 1 values is the dimension of the image of P. □

In the real case, the result regarding the projections is as follows:

PROPOSITION 12.15. *The matrices $P \in M_N(\mathbb{R})$ which are projections, $P^2 = P = P^t$, are precisely those which diagonalize as follows,*

$$P = UDU^t$$

with $U \in O_N$, and with $D \in M_N(0,1)$ being diagonal.

PROOF. This follows indeed from Proposition 12.14, and its proof. □

An important class of self-adjoint matrices, which includes for instance all the projections, are the positive matrices. The theory here is as follows:

THEOREM 12.16. *For a matrix $A \in M_N(\mathbb{C})$ the following conditions are equivalent, and if they are satisfied, we say that A is positive:*

(1) $A = B^2$, *with* $B = B^*$.
(2) $A = CC^*$, *for some* $C \in M_N(\mathbb{C})$.
(3) $< Ax, x > \geq 0$, *for any vector* $x \in \mathbb{C}^N$.
(4) $A = A^*$, *and the eigenvalues are positive,* $\lambda_i \geq 0$.
(5) $A = UDU^*$, *with* $U \in U_N$ *and with* $D \in M_N(\mathbb{R}_+)$ *diagonal.*

PROOF. The idea is that the equivalences in the statement basically follow from some elementary computations, with only Theorem 12.12 needed, at some point:

(1) $\implies$ (2) This is clear, because we can take $C = B$.

(2) $\implies$ (3) This follows from the following computation:

$$\begin{aligned} < Ax, x > &= < CC^*x, x > \\ &= < C^*x, C^*x > \\ &\geq 0 \end{aligned}$$

(3) $\implies$ (4) By using the fact that $< Ax, x >$ is real, we have:

$$\begin{aligned} < Ax, x > &= < x, A^*x > \\ &= < A^*x, x > \end{aligned}$$

Thus we have $A = A^*$, and the remaining assertion, regarding the eigenvalues, follows from the following computation, assuming $Ax = \lambda x$:

$$\begin{aligned} < Ax, x > &= < \lambda x, x > \\ &= \lambda < x, x > \\ &\geq 0 \end{aligned}$$

(4) $\implies$ (5) This follows indeed by using Theorem 12.12.

(5) $\implies$ (1) Assuming $A = UDU^*$ with $U \in U_N$, and with $D \in M_N(\mathbb{R}_+)$ diagonal, we can set $B = U\sqrt{D}U^*$. Then B is self-adjoint, and its square is given by:

$$\begin{aligned} B^2 &= U\sqrt{D}U^* \cdot U\sqrt{D}U^* \\ &= UDU^* \\ &= A \end{aligned}$$

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

Let us record as well the following technical version of the above result:

THEOREM 12.17. *For a matrix $A \in M_N(\mathbb{C})$ the following conditions are equivalent, and if they are satisfied, we say that A is strictly positive:*

(1) $A = B^2$, *with* $B = B^*$, *invertible.*
(2) $A = CC^*$, *for some* $C \in M_N(\mathbb{C})$ *invertible.*
(3) $< Ax, x >> 0$, *for any nonzero vector* $x \in \mathbb{C}^N$.
(4) $A = A^*$, *and the eigenvalues are strictly positive,* $\lambda_i > 0$.
(5) $A = UDU^*$, *with* $U \in U_N$ *and with* $D \in M_N(\mathbb{R}_+^*)$ *diagonal.*

PROOF. This follows either from Theorem 12.16, by adding the above various extra assumptions, or from the proof of Theorem 12.16, by modifying where needed. □

Let us discuss now the case of the unitary matrices. We have here:

THEOREM 12.18. *Any matrix $U \in M_N(\mathbb{C})$ which is unitary, $U^* = U^{-1}$, is diagonalizable, with the eigenvalues on $\mathbb{T}$. More precisely we have*

$$U = VDV^*$$

with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because given a matrix of type $U = VDV^*$, with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ being diagonal, we have:

$$\begin{aligned} U^* &= (VDV^*)^* \\ &= VD^*V^* \\ &= VD^{-1}V^{-1} \\ &= (V^*)^{-1}D^{-1}V^{-1} \\ &= (VDV^*)^{-1} \\ &= U^{-1} \end{aligned}$$

Let us prove now the first assertion, stating that the eigenvalues of a unitary matrix $U \in U_N$ belong to $\mathbb{T}$. Indeed, assuming $Uv = \lambda v$, we have:

$$\begin{aligned} < v, v > &= < U^*Uv, v > \\ &= < Uv, Uv > \\ &= < \lambda v, \lambda v > \\ &= |\lambda|^2 < v, v > \end{aligned}$$

Thus we obtain $\lambda \in \mathbb{T}$, as claimed. Our next claim now is that the eigenspaces corresponding to different eigenvalues are pairwise orthogonal. Assume indeed that:

$$Uv = \lambda v \quad , \quad Uw = \mu w$$

We have then the following computation, using $U^* = U^{-1}$ and $\lambda, \mu \in \mathbb{T}$:

$$\begin{aligned}
\lambda < v, w > &= < \lambda v, w > \\
&= < Uv, w > \\
&= < v, U^* w > \\
&= < v, U^{-1} w > \\
&= < v, \mu^{-1} w > \\
&= \mu < v, w >
\end{aligned}$$

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 U span the whole space $\mathbb{C}^N$. For this purpose, we will use a recurrence method. Let us pick an eigenvector of our matrix:

$$Uv = \lambda v$$

Assuming that we have a vector w orthogonal to it, $v \perp w$, we have:

$$\begin{aligned}
< Uw, v > &= < w, U^* v > \\
&= < w, U^{-1} v > \\
&= < w, \lambda^{-1} v > \\
&= \lambda < w, v > \\
&= 0
\end{aligned}$$

Thus, if v is an eigenvector, then the vector space $v^\perp$ is invariant under U. Now since U is an isometry, so is its restriction to this space $v^\perp$. Thus this restriction is a unitary, and so we can proceed by recurrence, and we obtain the result. □

Let us record as well the real version of the above result, in a weak form:

PROPOSITION 12.19. *Any matrix $U \in M_N(\mathbb{R})$ which is orthogonal, $U^t = U^{-1}$, is diagonalizable, with the eigenvalues on $\mathbb{T}$. More precisely we have*

$$U = VDV^*$$

with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ being diagonal.

PROOF. This follows indeed from Theorem 12.18. □

Observe that the above result does not provide us with a complete characterization of the matrices $U \in M_N(\mathbb{R})$ which are orthogonal. To be more precise, the question left is that of understanding when the matrices of type $U = VDV^*$, with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ being diagonal, are real, and this is something non-trivial.

As an illustration, for the simplest unitaries that we know, namely the rotations in the real plane, we have the following formula, that we know well from chapter 9:

$$\begin{pmatrix}\cos t & -\sin t\\ \sin t & \cos t\end{pmatrix}=\frac{1}{2}\begin{pmatrix}1 & 1\\ i & -i\end{pmatrix}\begin{pmatrix}e^{-it} & 0\\ 0 & e^{it}\end{pmatrix}\begin{pmatrix}1 & -i\\ 1 & i\end{pmatrix}$$

Back to generalities, the self-adjoint matrices and the unitary matrices are particular cases of the general notion of a "normal matrix", and we have here:

THEOREM 12.20. *Any matrix $A\in M_N(\mathbb{C})$ which is normal, $AA^*=A^*A$, is diagonalizable, with the diagonalization being of the following type,*

$$A=UDU^*$$

with $U\in U_N$, and with $D\in M_N(\mathbb{C})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because if we take a matrix of the form $A=UDU^*$, with U unitary and D diagonal, then we have:

$$\begin{aligned}AA^*&=UDU^*\cdot UD^*U^*\\ &=UDD^*U^*\\ &=UD^*DU^*\\ &=UD^*U^*\cdot UDU^*\\ &=A^*A\end{aligned}$$

In the other sense now, this is something more technical. Our first claim is that a matrix A is normal precisely when the following happens, for any vector v:

$$||Av||=||A^*v||$$

Indeed, the above equality can be written as follows:

$$<AA^*v,v>=<A^*Av,v>$$

But this is equivalent to $AA^*=A^*A$, by expanding the scalar products. Our claim now is that A,A^* have the same eigenvectors, with conjugate eigenvalues:

$$Av=\lambda v\implies A^*v=\bar{\lambda}v$$

Indeed, this follows from the following computation, and from the trivial fact that if A is normal, then so is any matrix of type $A-\lambda 1_N$:

$$\begin{aligned}||(A^*-\bar{\lambda}1_N)v||&=||(A-\lambda 1_N)^*v||\\ &=||(A-\lambda 1_N)v||\\ &=0\end{aligned}$$

Let us prove now, by using this, that the eigenspaces of A are pairwise orthogonal. Assume that we have two eigenvectors, corresponding to different eigenvalues, $\lambda\neq\mu$:

$$Av=\lambda v\quad,\quad Aw=\mu w$$

We have the following computation, which shows that $\lambda \neq \mu$ implies $v \perp w$:

$$\begin{aligned}\lambda < v, w > &= < \lambda v, w > \\ &= < Av, w > \\ &= < v, A^* w > \\ &= < v, \bar{\mu} w > \\ &= \mu < v, w >\end{aligned}$$

In order to finish, it remains to prove that the eigenspaces of A span the whole $\mathbb{C}^N$. This is something that we have already seen for the self-adjoint matrices, and for unitaries, and we will use here these results, in order to deal with the general normal case. As a first observation, given an arbitrary matrix A, the matrix AA^* is self-adjoint:

$$(AA^*)^* = AA^*$$

Thus, we can diagonalize this matrix AA^*, as follows, with the passage matrix being a unitary, $V \in U_N$, and with the diagonal form being real, $E \in M_N(\mathbb{R})$:

$$AA^* = VEV^*$$

Now observe that, for matrices of type $A = UDU^*$, which are those that we supposed to deal with, we have the following formulae:

$$V = U \quad , \quad E = D\bar{D}$$

In particular, the matrices A and AA^* have the same eigenspaces. So, this will be our idea, proving that the eigenspaces of AA^* are eigenspaces of A. In order to do so, let us pick two eigenvectors v, w of the matrix AA^*, corresponding to different eigenvalues, $\lambda \neq \mu$. The eigenvalue equations are then as follows:

$$AA^* v = \lambda v \quad , \quad AA^* w = \mu w$$

We have the following computation, using the normality condition $AA^* = A^* A$, and the fact that the eigenvalues of AA^*, and in particular μ, are real:

$$\begin{aligned}\lambda < Av, w > &= < \lambda Av, w > \\ &= < A\lambda v, w > \\ &= < AAA^* v, w > \\ &= < AA^* Av, w > \\ &= < Av, AA^* w > \\ &= < Av, \mu w > \\ &= \mu < Av, w >\end{aligned}$$

We conclude that we have $< Av, w >= 0$. But this reformulates as follows:

$$\lambda \neq \mu \implies A(E_\lambda) \perp E_\mu$$

Now since the eigenspaces of AA^* are pairwise orthogonal, and span the whole $\mathbb{C}^N$, we deduce from this that these eigenspaces are invariant under A:

$$A(E_\lambda) \subset E_\lambda$$

But with this result in hand, we can finish. Indeed, we can decompose the problem, and the matrix A itself, following these eigenspaces of AA^*, which in practice amounts in saying that we can assume that we only have 1 eigenspace. By rescaling, this is the same as assuming that we have $AA^* = 1$, and so we are now into the unitary case, that we know how to solve, as explained in Theorem 12.18. □

As a first application of all this, we have the following result:

THEOREM 12.21. *Given a matrix $A \in M_N(\mathbb{C})$, we can construct a matrix $|A|$ as follows, by using the fact that A^*A is diagonalizable, with positive eigenvalues:*

$$|A| = \sqrt{A^*A}$$

This matrix $|A|$ is then positive, and its square is $|A|^2 = A$. In the case $N = 1$, we obtain in this way the usual absolute value of the complex numbers.

PROOF. Consider indeed the matrix A^*A, which is normal. According to Theorem 12.20, we can diagonalize this matrix as follows, with $U \in U_N$, and with D diagonal:

$$A = UDU^*$$

From $A^*A \geq 0$ we obtain $D \geq 0$. But this means that the entries of D are real, and positive. Thus we can extract the square root $\sqrt{D}$, and then set:

$$\sqrt{A^*A} = U\sqrt{D}U^*$$

Thus, we are basically done. Indeed, if we call this latter matrix $|A|$, then we are led to the conclusions in the statement. Finally, the last assertion is clear from definitions. □

We can now formulate a first polar decomposition result, as follows:

THEOREM 12.22. *Any invertible matrix $A \in M_N(\mathbb{C})$ decomposes as*

$$A = U|A|$$

with $U \in U_N$, and with $|A| = \sqrt{A^*A}$ as above.

PROOF. This is routine, and follows by comparing the actions of $A, |A|$ on the vectors $v \in \mathbb{C}^N$, and deducing from this the existence of a unitary $U \in U_N$ as above. □

Observe that at $N = 1$ we obtain in this way the usual polar decomposition of the nonzero complex numbers. More generally now, we have the following result:

THEOREM 12.23. *Any square matrix $A \in M_N(\mathbb{C})$ decomposes as*

$$A = U|A|$$

with U being a partial isometry, and with $|A| = \sqrt{A^*A}$ as above.

PROOF. Again, this follows by comparing the actions of $A, |A|$ on the vectors $v \in \mathbb{C}^N$, and deducing from this the existence of a partial isometry U as above. Alternatively, we can get this from Theorem 12.22, applied on the complement of the 0-eigenvectors. □

So long for advanced linear algebra. There are actually many other decomposition results for the real matrices, quite often in relation with positivity, and the Jordan form too, which are all useful for questions in analysis, via derivatives and Hessians.

Good luck of course in learning all this, when needed later in life, for your various math problems at that time. And always have in mind expert's advice, Cat 12.11.

12c. Harmonic functions

As another application of the second derivatives, we can now talk about harmonic functions in general, extending what we know from chapter 8. We first have:

DEFINITION 12.24. *The Laplace operator in N dimensions is:*

$$\Delta f = \sum_{i=1}^{N} \frac{d^2 f}{dx_i^2}$$

A function $f : \mathbb{R}^N \to \mathbb{C}$ satisfying $\Delta f = 0$ will be called harmonic.

Observe that this generalizes what we know in two dimensions, from chapter 8. In order to discuss this, let us first recall the handful of results that we know from there, in N dimensions, with the comment of course that chapter 8 was a physics chapter, and now that we learned mathematics in N dimensions, you can check all that stuff, which was coming a bit in advance, and correct all that stuff is. First is the validity of the wave equation in N dimensions, which is as follows, $v > 0$ being the propagation speed:

$$\ddot{\varphi} = v^2 \Delta \varphi$$

Then second is the validity of the heat equation in N dimensions, which is as follows, with $\alpha > 0$ being the thermal diffusivity of the medium:

$$\dot{\varphi} = \alpha \Delta \varphi$$

And finally, we had as well some mathematics in chapter 8, but mostly 2-dimensional, and not very useful for our purposes here. Let us recall however a key computation, that we did in N dimensions, stating that the fundamental radial solutions of $\Delta f = 0$ are as follows, with the log at $N = 2$ basically coming from $\log' = 1/x$:

$$f(x) = \begin{cases} ||x||^{2-N} & (N \neq 2) \\ \log ||x|| & (N = 2) \end{cases}$$

Not bad for a start, all this. In analogy now with the one complex variable theory that we know from chapter 6 and chapter 8, we have the following result:

THEOREM 12.25. *The harmonic functions in N dimensions obey to the same general principles as the holomorphic functions, namely:*

(1) *The plain mean value formula.*
(2) *The boundary mean value formula.*
(3) *The maximum modulus principle.*
(4) *The Liouville theorem.*

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

(1) Regarding the plain mean value formula, here the statement is that given an harmonic function $f : X \to \mathbb{C}$, and a ball B, the following happens:

$$f(x) = \int_B f(y)dy$$

In order to prove this, we can assume that B is centered at 0, of radius $r > 0$. if we denote by χ_r the characteristic function of this ball, nomalized as to integrate up to 1, in terms of the convolution operation from chapter 7, we want to prove that we have:

$$f = f * \chi_r$$

For doing so, let us pick a number $0 < s < r$, and a solution w of the following equation, supported on B, which can be constructed explicitely:

$$\Delta w = \chi_r - \chi_s$$

By using now the properties of the convolution $*$ from chapter 7, we have:

$$\begin{aligned} f * \chi_r - f * \chi_s &= f * (\chi_r - \chi_s) \\ &= f * \Delta w \\ &= \Delta f * w \\ &= 0 \end{aligned}$$

Thus $f * \chi_r = f * \chi_s$, and by letting now $s \to 0$, we get $f * \chi_r = f$, as desired.

(2) Regarding the boundary mean value formula, here the statement is that given an harmonic function $f : X \to \mathbb{C}$, and a ball B, with boundary γ, the following happens:

$$f(x) = \int_\gamma f(y)dy$$

But this follows as a consequence of the plain mean value formula in (1), with our two mean value formulae, the one there and the one here, being in fact equivalent, by using annuli and radial integration for the proof of the equivalence, in the obvious way.

(3) Regarding the maximum modulus principle, the statement here is that any holomorphic function $f : X \to \mathbb{C}$ has the property that the maximum of $|f|$ over a domain is

attained on its boundary. That is, given a domain D, with boundary γ, we have:

$$\exists x \in \gamma \quad , \quad |f(x)| = \max_{y \in D} |f(y)|$$

But this is something which follows again from the mean value formula in (1), first for the balls, and then in general, by using a standard division argument.

(4) Finally, regarding the Liouville theorem, the statement here is that an entire, bounded harmonic function must be constant:

$$f : \mathbb{R}^N \to \mathbb{C} \quad , \quad \Delta f = 0 \quad , \quad |f| \leq M \quad \implies \quad f = \text{constant}$$

As a slightly weaker statement, again called Liouville theorem, we have the fact that an entire harmonic function which vanishes at ∞ must vanish globally:

$$f : \mathbb{R}^N \to \mathbb{C} \quad , \quad \Delta f = 0 \quad , \quad \lim_{x \to \infty} f(x) = 0 \quad \implies \quad f = 0$$

But can view these as a consequence of the mean value formula in (1), because given two points $x \neq y$, we can view the values of f at these points as averages over big balls centered at these points, say $B = B_x(R)$ and $C = B_y(R)$, with $R >> 0$:

$$f(x) = \int_B f(z) dz \quad , \quad f(y) = \int_C f(z) dz$$

Indeed, the point is that when the radius goes to ∞, these averages tend to be equal, and so we have $f(x) \simeq f(y)$, which gives $f(x) = f(y)$ in the limit, as desired. □

12d. Lagrange multipliers

We discuss now some more specialized topics, in relation with optimization. First of all, thinking well, the functions that we have to minimize or maximize, in the real life, are often defined on a manifold, instead of being defined on the whole $\mathbb{R}^N$. Fortunately, the good old principle $f'(x) = 0$ can be adapted to the manifold case, as follows:

PRINCIPLE 12.26. *In order for a function $f : X \to \mathbb{R}$ defined on a manifold X to have a local extremum at $x \in X$, we must have, as usual*

$$f'(x) = 0$$

but with this taking into account the fact that the equations defining the manifold count as well as "zero", and so must be incorporated into the formula $f'(x) = 0$.

Obviously, we are punching here about our weight, because our discussion about manifolds from chapter 11 was quite amateurish, and we have no tools in our bag for proving such things, or even for properly formulating them. However, we can certainly talk about all this, a bit like physicists do. So, our principle is the one above, and in practice, the

idea is that we must have a formula as follows, with g_i being the constraint functions for our manifold X, and with $\lambda_i \in \mathbb{R}$ being certain scalars, called Lagrange multipliers:

$$f'(x) = \sum_i \lambda_i g_i'(x)$$

As a basic illustration for this, our claim is that, by using a suitable manifold, and a suitable function, and Lagrange multipliers, we can prove in this way the Hölder inequality, that we know well of course, but without any computation. Let us start with:

PROPOSITION 12.27. *For any exponent $p > 1$, the following set*

$$S_p = \left\{ x \in \mathbb{R}^N \middle| \sum_i |x_i|^p = 1 \right\}$$

is a submanifold of $\mathbb{R}^N$.

PROOF. We know from chapter 11 that the unit sphere in $\mathbb{R}^N$ is a manifold. In our terms, this solves our problem at $p = 2$, because this unit sphere is:

$$S_2 = \left\{ x \in \mathbb{R}^N \middle| \sum_i x_i^2 = 1 \right\}$$

Now observe that we have a bijection $S_p \simeq S_2$, at least on the part where all the coordinates are positive, $x_i > 0$, given by the following function:

$$x_i \to x_i^{2/p}$$

Thus we obtain that S_p is indeed a manifold, as claimed. □

We already know that the manifold S_p constructed above is the unit sphere, in the case $p = 2$. In order to have a better geometric picture of what is going on, in general, observe that S_p can be constructed as well at $p = 1$, as follows:

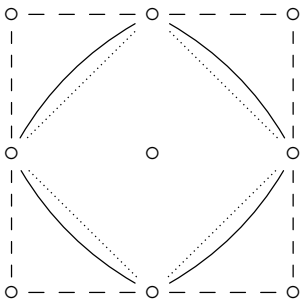
$$S_1 = \left\{ x \in \mathbb{R}^N \middle| \sum_i |x_i| = 1 \right\}$$

However, this is no longer a manifold, as we can see for instance at $N = 2$, where we obtain a square. Now observe that we can talk as well about $p = \infty$, as follows:

$$S_\infty = \left\{ x \in \mathbb{R}^N \middle| \sup_i |x_i| = 1 \right\}$$

This letter set is no longer a manifold either, as we can see for instance at $N = 2$, where we obtain a again a square, containing the previous square, the one at $p = 1$.

With these limiting constructions in hand, we can now have a better geometric picture of what is going on, in the general context of Proposition 12.27. Indeed, let us draw, at $N = 2$ for simplifying, our sets S_p at the values $p = 1, 2, \infty$ of the exponent:



We can see that what we have is a small square, at $p = 1$, becoming smooth and inflating towards the circle, in the parameter range $p \in (1, 2]$, and then further inflating, in the parameter range $p \in [2, \infty)$, towards the big square appearing at $p = \infty$.

With these preliminaries in hand, we can formulate our result, as follows:

THEOREM 12.28. *The local extrema over S_p of the function*

$$f(x) = \sum_i x_i y_i$$

can be computed by using Lagrange multipliers, and this gives

$$\left| \sum_i x_i y_i \right| \leq \left(\sum_i |x_i|^p \right)^{1/p} \left(\sum_i |y_i|^q \right)^{1/q}$$

with $1/p + 1/q = 1$, that is, the Hölder inequality, with a purely geometric proof.

PROOF. We can restrict the attention to the case where all the coordinates are positive, $x_i > 0$ and $y_i > 0$. The derivative of the function in the statement is:

$$f'(x) = (y_1, \ldots, y_N)$$

On the other hand, we know that the manifold S_p appears by definition as the set of zeroes of the function $\varphi(x) = \sum_i x_i^p - 1$, having derivative as follows:

$$\varphi'(x) = p(x_1^{p-1}, \ldots, x_N^{p-1})$$

Thus, by using Lagrange multipliers, the critical points of f must satisfy:

$$(y_1, \ldots, y_N) \sim (x_1^{p-1}, \ldots, x_N^{p-1})$$

In other words, the critical points must satisfy $x_i = \lambda y_i^{1/(p-1)}$, for some $\lambda > 0$, and by using now $\sum_i x_i^p = 1$ we can compute the precise value of λ, and we get:

$$\lambda = \left(\sum_i y_i^{p/(p-1)} \right)^{-1/p}$$

Now let us see what this means. Since the critical point is unique, this must be a maximum of our function, and we conclude that for any $x \in S_p$, we have:

$$\sum_i x_i y_i \leq \sum_i \lambda y_i^{1/(p-1)} \cdot y_i = \left(\sum_i y_i^{p/(p-1)} \right)^{1-1/p} = \left(\sum_i y_i^q \right)^{1/q}$$

Thus we have Hölder, and the general case follows from this, by rescaling. □

As a second illustration for the method of Lagrange multipliers, this time in relation with certain questions from linear algebra, we have the following result:

THEOREM 12.29. *Consider a differentiable function $\varphi : [0, \infty) \to \mathbb{R}$. An orthogonal matrix having nonzero entries, $U \in O_N^*$, is then a critical point of the function*

$$F(U) = \sum_{ij} \varphi(|U_{ij}|)$$

precisely when the matrix WU^t is symmetric, where:

$$W_{ij} = \text{sgn}(U_{ij}) \varphi'(|U_{ij}|)$$

In particular, for $F(U) = ||U||_1$ we need SU^t to be symmetric, where $S_{ij} = \text{sgn}(U_{ij})$.

PROOF. We regard O_N as a real algebraic manifold, with coordinates U_{ij}. This manifold consists by definition of the zeroes of the following polynomials:

$$A_{ij} = \sum_k U_{ik} U_{jk} - \delta_{ij}$$

Thus $U \in O_N$ is a critical point of F when the following condition is satisfied:

$$dF \in span(dA_{ij})$$

Regarding the space $span(dA_{ij})$, this consists of the following quantities:

$$\begin{aligned} \sum_{ij} M_{ij} dA_{ij} &= \sum_{ijk} M_{ij} (U_{ik} dU_{jk} + U_{jk} dU_{ik}) \\ &= \sum_{jk} (M^t U)_{jk} dU_{jk} + \sum_{ik} (MU)_{ik} dU_{ik} \\ &= \sum_{ij} (M^t U)_{ij} dU_{ij} + \sum_{ij} (MU)_{ij} dU_{ij} \end{aligned}$$

In order to compute dF, observe first that, with $S_{ij} = sgn(U_{ij})$, we have:

$$d|U_{ij}| = d\sqrt{U_{ij}^2} = \frac{U_{ij}dU_{ij}}{|U_{ij}|} = S_{ij}dU_{ij}$$

Thus, with $W_{ij} = sgn(U_{ij})\varphi'(|U_{ij}|)$ as in the statement, we have:

$$\begin{aligned} dF &= \sum_{ij} d\left(\varphi(|U_{ij}|)\right) \\ &= \sum_{ij} \varphi'(|U_{ij}|)d|U_{ij}| \\ &= \sum_{ij} W_{ij}dU_{ij} \end{aligned}$$

We conclude that $U \in O_N$ is a critical point of F precisely when there exists a matrix $M \in M_N(\mathbb{R})$ such that the following two conditions are satisfied:

$$W = M^tU \quad , \quad W = MU$$

Now observe that these two equations can be written as follows:

$$M^t = WU^t \quad , \quad M = WU^t$$

Thus, the matrix WU^t must be symmetric, as claimed. □

Finally, have you ever been hiking, in the fog or under the rain, trying to reach the summit? What you have to do is to follow the steepest slope, and then climb and climb, only to discover at the end, when on the summit, in an instant where the fog briefly dissapears, that there is in fact a second summit, a bit higher, not far away. And so that you have to climb down, and then go all over again, with your steepest slope method.

In mathematics, this is called the gradient method. The idea is that of following the gradient in order to reach to local maxima or minima, exactly as above, when hiking. And there are a number of things that can be said here, and for more on all this, and for optimization questions in general, we recommend a serious, applied analysis book.

12e. Exercises

This chapter was tough analysis as it comes, and as exercises, we have:

EXERCISE 12.30. *Clarify the Taylor formula at order* 3*, and its applications.*

EXERCISE 12.31. *Learn the Jordan form, as a complement to the linear algebra here.*

EXERCISE 12.32. *Harmonic functions in* 3 *dimensions, and their applications.*

EXERCISE 12.33. *Clarify all the details, in relation with Lagrange multipliers.*

As bonus exercise, learn the gradient method. You won't escape from that.

Part IV

Integration theory

I'm the left eye
You're the right
Would it not be madness to fight
We come one

CHAPTER 13

Multiple integrals

13a. Multiple integrals

Welcome to advanced calculus. We have kept the best for the end, and in this whole last part of the present book we will learn how to integrate functions of several variables. The general, fascinating question that we will be interested in is as follows:

QUESTION 13.1. *Given a function $f : \mathbb{R}^N \to \mathbb{R}$, how to compute*

$$\int_{\mathbb{R}^N} f(x)dx_1 \ldots dx_N$$

or at least, what are the rules satisfied by such integrals?

Here we adopt, somehow by definition, the convention that the above integral is constructed a bit like the one-variable integrals, via Riemann sums as in chapter 4, by using this time divisions of the space $\mathbb{R}^N$ into small N-dimensional cubes. There are of course some theory and details to be worked out here, not exactly trivial, but since we have not really done this in chapter 4, nor we will do this here. Let's focus on computations.

At the first glance solving Question 13.1 this looks like an easy task, because we can iterate one-variable integrations, which are something that we know well, from chapter 4. For instance the integral of a function $f : \mathbb{R}^2 \to \mathbb{R}$ can be computed by using:

$$\int_{\mathbb{R}^2} f(z)dz = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y)dxdy$$

This being said, we are faced right away with a difficulty, when using this method. Indeed, we can use as well the following rival method, yielding the same answer:

$$\int_{\mathbb{R}^2} f(z)dz = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y)dydx$$

So, which method is the best? Depends on f, of course. However, things do not stop here, because in certain situations it is better to use polar coordinates, as follows:

$$\int_{\mathbb{R}^2} f(z)dz = \int_0^{2\pi} \int_0^{\infty} f(r\cos t, r\sin t)Jdrdt$$

Here the factor on the left is $J = dxdy/drdt$, which remains to be computed. And for the picture to be complete, we have as well the following fourth formula:

$$\int_{\mathbb{R}^2} f(z)dz = \int_0^\infty \int_0^{2\pi} f(r\cos t, r\sin t)Jdtdr$$

In short, you got my point, I hope, things are quite complicated in several variables, with the complications starting already in 2 variables. And actually, if you think a bit about 3 variables, it is quite clear that the above complications become true nightmares, leading to long nights spent in computing simple integrals, there in 3 variables.

So, what to do? Work and patience, of course, and here is our plan:

(1) In this chapter we will get used to the multiple integrals, with some general rules for their computation, including a rule for computing the above factor $J = dxdy/drdt$. Then we will enjoy all this by computing some integrals over the spheres in $\mathbb{R}^N$.

(2) In chapter 14 we will review probability theory, and further develop it, notably with the theory of normal variables. And finally, in chapters 15-16 we will get back to physics, with some sharp results in 3 dimensions, and then in infinite dimensions.

This sounds good, but as a matter of doublechecking what we are doing, make sure that it is wise indeed, let us ask the cat about what he thinks. And cat answers:

CAT 13.2. *Not quite sure about your formula*

$$\int_{\mathbb{R}} \int_{\mathbb{R}} f(x,y)dxdy = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x,y)dydx$$

and I doubt too that you can properly compute $J = dxdy/drdt$. Read Rudin.

Oh dear. What can I say. Sure I read Rudin, as a Gen X mathematician, but we are now well into the 21st century, and shall I go ahead with heavy measure theory, for having all this properly developed? Not quite sure, I'd rather stick to my plan.

So, ignoring what cat says, but get however a copy of Rudin's red book, and don't forget about Mao's too, and getting back to our plan, as a first goal, we would like to compute the factor $J = dxdy/drdt$. Let us start with something that we know, in 1D:

PROPOSITION 13.3. *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 $f = F'$, via the following differentiation rule:

$$(F\varphi)'(t) = F'(\varphi(t))\varphi'(t)$$

Indeed, by integrating between c and d, we obtain the result. □

In several variables now, we can only expect the above $\varphi'(t)$ factor to be replaced by something similar, a sort of "derivative of φ, arising as a real number". But this can only be the Jacobian $\det(\varphi'(t))$, and with this in mind, we are led to:

THEOREM 13.4. *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 formula from Proposition 13.3.

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 Proposition 13.3, 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 definition of the determinant as a volume, as in chapter 9.

(4) The details and computations however are quite non-trivial, and can be found for instance in Rudin [74]. So, please read that. With this, reading the complete proof of the present theorem from Rudin, being part of the standard math experience. □

Finally, speaking Rudin, and getting back to Cat 13.2, there are some deep truths there. But, remember our Advice 1.3, from the beginning of this book. Full rigor does not guarantee that your computation is correct, you always have to doublecheck. So an alternative method is that of using less rigor, and more doublechecks at the end. And this will be our philosophy in what follows, with all our formulae below being correct.

13b. Spherical coordinates

Time now do some exciting computations, with the technology that we have. In what regards the applications of Theorem 13.4, these often come via:

PROPOSITION 13.5. *We have polar coordinates in 2 dimensions,*

$$\begin{cases} x &= r\cos t\\ y &= r\sin t \end{cases}$$

the corresponding Jacobian being $J = r$.

PROOF. This is elementary, the Jacobian being:

$$\begin{aligned} J &= \begin{vmatrix} \frac{d(r\cos t)}{dr} & \frac{d(r\cos t)}{dt}\\ \\ \frac{d(r\sin t)}{dr} & \frac{d(r\sin t)}{dt} \end{vmatrix}\\ &= \begin{vmatrix} \cos t & -r\sin t\\ \sin t & r\cos t \end{vmatrix}\\ &= r\cos^2 t + r\sin^2 t\\ &= r \end{aligned}$$

Thus, we have indeed the formula in the statement. □

We can now compute the Gauss integral, which is the best calculus formula ever:

THEOREM 13.6. *We have the following formula,*

$$\int_{\mathbb{R}} e^{-x^2}dx = \sqrt{\pi}$$

called Gauss integral formula.

PROOF. Let I be the above integral. By using polar coordinates, we obtain:

$$\begin{aligned} I^2 &= \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-x^2-y^2}dxdy\\ &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2}rdrdt\\ &= 2\pi\int_0^{\infty}\left(-\frac{e^{-r^2}}{2}\right)'dr\\ &= 2\pi\left[0-\left(-\frac{1}{2}\right)\right]\\ &= \pi \end{aligned}$$

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

Moving now to 3 dimensions, we have here the following result:

PROPOSITION 13.7. *We have spherical coordinates in* 3 *dimensions,*

$$\begin{cases} x &= r\cos s \\ y &= r\sin s\cos t \\ z &= r\sin s\sin t \end{cases}$$

the corresponding Jacobian being $J(r,s,t) = r^2\sin s$.

PROOF. The fact that we have indeed spherical coordinates is clear. Regarding now the Jacobian, this is given by the following formula:

$$\begin{aligned} & J(r,s,t) \\ = & \begin{vmatrix} \cos s & -r\sin s & 0 \\ \sin s\cos t & r\cos s\cos t & -r\sin s\sin t \\ \sin s\sin t & r\cos s\sin t & r\sin s\cos t \end{vmatrix} \\ = & r^2\sin s\sin t\begin{vmatrix} \cos s & -r\sin s \\ \sin s\sin t & r\cos s\sin t \end{vmatrix} + r\sin s\cos t\begin{vmatrix} \cos s & -r\sin s \\ \sin s\cos t & r\cos s\cos t \end{vmatrix} \\ = & r\sin s\sin^2 t\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} + r\sin s\cos^2 t\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} \\ = & r\sin s(\sin^2 t+\cos^2 t)\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} \\ = & r\sin s\times 1\times r \\ = & r^2\sin s \end{aligned}$$

Thus, we have indeed the formula in the statement. □

Let us work out now the general spherical coordinate formula, in arbitrary N dimensions. The formula here, which generalizes those at $N = 2, 3$, is as follows:

THEOREM 13.8. *We have spherical coordinates in* N *dimensions,*

$$\begin{cases} x_1 &= r\cos t_1 \\ x_2 &= r\sin t_1\cos t_2 \\ \vdots & \\ x_{N-1} &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\cos t_{N-1} \\ x_N &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\sin t_{N-1} \end{cases}$$

the corresponding Jacobian being given by the following formula,

$$J(r,t) = r^{N-1}\sin^{N-2}t_1\sin^{N-3}t_2\ \ldots\ \sin^2 t_{N-3}\sin t_{N-2}$$

and with this generalizing the known formulae at $N = 2, 3$.

PROOF. As before, the fact that we have spherical coordinates is clear. Regarding now the Jacobian, also as before, by developing over the last column, we have:

$$\begin{aligned} J_N &= r\sin t_1\ldots\sin t_{N-2}\sin t_{N-1}\times\sin t_{N-1}J_{N-1}\\ &+ r\sin t_1\ldots\sin t_{N-2}\cos t_{N-1}\times\cos t_{N-1}J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}(\sin^2 t_{N-1}+\cos^2 t_{N-1})J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}J_{N-1}\end{aligned}$$

Thus, we obtain the formula in the statement, by recurrence. □

As a comment here, the above convention for spherical coordinates is one among many, designed to best work in arbitrary N dimensions. Also, in what regards the precise range of the angles $t_1,\ldots,t_{N-1}$, we will leave this to you, as an instructive exercise.

As an application, let us compute the volumes of spheres. For this purpose, we must understand how the products of coordinates integrate over spheres. Let us start with the case $N=2$. Here the sphere is the unit circle $\mathbb{T}$, and with $z=e^{it}$ the coordinates are $\cos t,\sin t$. We can first integrate arbitrary powers of these coordinates, as follows:

PROPOSITION 13.9. *We have the following formulae,*

$$\int_0^{\pi/2}\cos^p t\,dt=\int_0^{\pi/2}\sin^p t\,dt=\left(\frac{\pi}{2}\right)^{\varepsilon(p)}\frac{p!!}{(p+1)!!}$$

where $\varepsilon(p)=1$ if p is even, and $\varepsilon(p)=0$ if p 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:

$$I_p=\int_0^{\pi/2}\cos^p t\,dt$$

We do this by partial integration. We have the following formula:

$$\begin{aligned}(\cos^p t\sin t)' &= p\cos^{p-1}t(-\sin t)\sin t+\cos^p t\cos t\\ &= p\cos^{p+1}t-p\cos^{p-1}t+\cos^{p+1}t\\ &= (p+1)\cos^{p+1}t-p\cos^{p-1}t\end{aligned}$$

By integrating between 0 and $\pi/2$, we obtain the following formula:

$$(p+1)I_{p+1}=pI_{p-1}$$

Thus we can compute I_p by recurrence, and we obtain:

$$\begin{aligned} I_p &= \frac{p-1}{p} I_{p-2} \\ &= \frac{p-1}{p} \cdot \frac{p-3}{p-2} I_{p-4} \\ &= \frac{p-1}{p} \cdot \frac{p-3}{p-2} \cdot \frac{p-5}{p-4} I_{p-6} \\ &\vdots \\ &= \frac{p!!}{(p+1)!!} I_{1-\varepsilon(p)} \end{aligned}$$

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. □

We can now compute the volume of the sphere, as follows:

THEOREM 13.10. *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. Let us denote by B^+ the positive part of the unit sphere, or rather unit ball B, obtained by cutting this unit ball in 2^N parts. At the level of volumes, we have:

$$V = 2^N V^+$$

We have the following computation, using spherical coordinates, and Fubini:

$$\begin{aligned} V^+ &= \int_{B^+} 1 \\ &= \int_0^1 \int_0^{\pi/2} \ldots \int_0^{\pi/2} r^{N-1} \sin^{N-2} t_1 \ldots \sin t_{N-2}\, dr dt_1 \ldots dt_{N-1} \\ &= \int_0^1 r^{N-1} dr \int_0^{\pi/2} \sin^{N-2} t_1\, dt_1 \ldots \int_0^{\pi/2} \sin t_{N-2} dt_{N-2} \int_0^{\pi/2} 1 dt_{N-1} \\ &= \frac{1}{N} \times \left(\frac{\pi}{2}\right)^{[N/2]} \times \frac{(N-2)!!}{(N-1)!!} \cdot \frac{(N-3)!!}{(N-2)!!} \ldots \frac{2!!}{3!!} \cdot \frac{1!!}{2!!} \cdot 1 \\ &= \frac{1}{N} \times \left(\frac{\pi}{2}\right)^{[N/2]} \times \frac{1}{(N-1)!!} \\ &= \left(\frac{\pi}{2}\right)^{[N/2]} \frac{1}{(N+1)!!} \end{aligned}$$

Here we have used the following formula, for computing the exponent of $\pi/2$:

$$\begin{aligned}\varepsilon(0)+\varepsilon(1)+\varepsilon(2)+\ldots+\varepsilon(N-2) &= 1+0+1+\ldots+\varepsilon(N-2)\\ &= \left[\frac{N-2}{2}\right]+1\\ &= \left[\frac{N}{2}\right]\end{aligned}$$

Thus, we obtain the formula in the statement. □

As main particular cases of the above formula, we have:

THEOREM 13.11. *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. Some of these results are well-known, but we can obtain all of them as particular cases of the general formula in Theorem 13.10, as follows:

(1) At $N=1$ we obtain $V=1\cdot\frac{2}{1}=2$.

(2) At $N=2$ we obtain $V=\frac{\pi}{2}\cdot\frac{4}{2}=\pi$.

(3) At $N=3$ we obtain $V=\frac{\pi}{2}\cdot\frac{8}{3}=\frac{4\pi}{3}$.

(4) At $N=4$ we obtain $V=\frac{\pi^2}{4}\cdot\frac{16}{8}=\frac{\pi^2}{2}$. □

There are many other applications of the above, as we will see in what follows.

13c. Stirling estimates

The formula in Theorem 13.10 is certainly nice, but in practice, we would like for instance to have estimates for that volumes of the spheres. We will need:

THEOREM 13.12. *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:

$$\begin{aligned}\log(N!) &= \sum_{k=1}^{N}\log k\\ &\approx \int_1^N \log x\,dx\\ &= N\log N - N + 1\end{aligned}$$

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:

$$\begin{aligned}\log(N!) &= \sum_{k=1}^{N}\log k\\ &\approx \int_1^N \log x\,dx + \frac{\log 1 + \log N}{2}\\ &= N\log N - N + 1 + \frac{\log N}{2}\end{aligned}$$

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) However, all this remains a bit complicated, so we would like to present now an alternative approach to (3), which also misses some details, but better does the job, explaining where the $\sqrt{2\pi}$ factor comes from. First, by partial integration we have:

$$N! = \int_0^\infty x^N e^{-x} dx$$

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:

$$\begin{aligned}
\log(x^N e^{-x}) &= N\log x - x\\
&= N\log(N+y) - (N+y)\\
&= N\log N + N\log\left(1+\frac{y}{N}\right) - (N+y)\\
&\simeq N\log N + N\left(\frac{y}{N} - \frac{y^2}{2N^2}\right) - (N+y)\\
&= N\log N - N - \frac{y^2}{2N}
\end{aligned}$$

By exponentiating, we obtain from this the following estimate:

$$x^N e^{-x} \simeq \left(\frac{N}{e}\right)^N e^{-y^2/2N}$$

Now by integrating, and using the Gauss formula, we obtain from this:

$$\begin{aligned}
N! &= \int_0^\infty x^N e^{-x} dx\\
&\simeq \int_{-N}^N \left(\frac{N}{e}\right)^N e^{-y^2/2N}\, dy\\
&\simeq \left(\frac{N}{e}\right)^N \int_{\mathbb{R}} e^{-y^2/2N}\, dy\\
&= \left(\frac{N}{e}\right)^N \sqrt{2N} \int_{\mathbb{R}} e^{-z^2}\, dz\\
&= \left(\frac{N}{e}\right)^N \sqrt{2\pi N}
\end{aligned}$$

Thus, we have proved the Stirling formula, as formulated in the statement. □

With the above formula in hand, we have many useful applications, such as:

PROPOSITION 13.13. *We have the following estimate for binomial coefficients,*

$$\binom{N}{K} \simeq \left(\frac{1}{t^t(1-t)^{1-t}}\right)^N \frac{1}{\sqrt{2\pi t(1-t)N}}$$

in the $K \simeq tN \to \infty$ limit, with $t \in (0,1]$. In particular we have

$$\binom{2N}{N} \simeq \frac{4^N}{\sqrt{\pi N}}$$

in the $N \to \infty$ limit, for the central binomial coefficients.

PROOF. All this is very standard, by using the Stirling formula etablished above, for the various factorials which appear, the idea being as follows:

(1) This follows from the definition of the binomial coefficients, namely:

$$\begin{aligned}\binom{N}{K} &= \frac{N!}{K!(N-K)!}\\ &\simeq \left(\frac{N}{e}\right)^N\sqrt{2\pi N}\left(\frac{e}{K}\right)^K\frac{1}{\sqrt{2\pi K}}\left(\frac{e}{N-K}\right)^{N-K}\frac{1}{\sqrt{2\pi(N-K)}}\\ &= \frac{N^N}{K^K(N-K)^{N-K}}\sqrt{\frac{N}{2\pi K(N-K)}}\\ &\simeq \frac{N^N}{(tN)^{tN}((1-t)N)^{(1-t)N}}\sqrt{\frac{N}{2\pi tN(1-t)N}}\\ &= \left(\frac{1}{t^t(1-t)^{1-t}}\right)^N\frac{1}{\sqrt{2\pi t(1-t)N}}\end{aligned}$$

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

(2) This estimate follows from a similar computation, as follows:

$$\begin{aligned}\binom{2N}{N} &= \frac{(2N)!}{N!N!}\\ &\simeq \left(\frac{2N}{e}\right)^{2N}\sqrt{4\pi N}\left(\frac{e}{N}\right)^{2N}\frac{1}{2\pi N}\\ &= \frac{4^N}{\sqrt{\pi N}}\end{aligned}$$

Alternatively, we can take $t=1/2$ in (1), then rescale. Indeed, we have:

$$\begin{aligned}\binom{N}{[N/2]} &\simeq \left(\frac{1}{(\frac{1}{2})^{1/2}(\frac{1}{2})^{1/2}}\right)^N\frac{1}{\sqrt{2\pi\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot N}}\\ &= 2^N\sqrt{\frac{2}{\pi N}}\end{aligned}$$

Thus with the change $N\to 2N$ we obtain the formula in the statement. □

Summarizing, we have so far complete estimate for the factorials. Regarding now the double factorials, that we will need as well, the result here is as follows:

PROPOSITION 13.14. *We have the following estimate for the double factorials,*

$$N!! \simeq \left(\frac{N}{e}\right)^{N/2} C$$

with $C = \sqrt{2}$ *for* N *even, and* $C = \sqrt{\pi}$ *for* N *odd. Alternatively, we have*

$$(N+1)!! \simeq \left(\frac{N}{e}\right)^{N/2} D$$

with $D = \sqrt{\pi N}$ *for* N *even, and* $D = \sqrt{2N}$ *for* N *odd.*

PROOF. Once again this is standard, the idea being as follows:

(1) When $N = 2K$ is even, we have the following computation:

$$\begin{aligned} N!! &= (2K-1)(2K-3)\ldots 1 \\ &= \frac{(2K)!}{2^K K!} \\ &\simeq \frac{1}{2^K}\left(\frac{2K}{e}\right)^{2K}\sqrt{4\pi K}\left(\frac{e}{K}\right)^K \frac{1}{\sqrt{2\pi K}} \\ &= \left(\frac{2K}{e}\right)^K \sqrt{2} \\ &= \left(\frac{N}{e}\right)^{N/2}\sqrt{2} \end{aligned}$$

(2) When $N = 2K+1$ is odd, we have the following computation:

$$\begin{aligned} N!! &= (2K)(2K-2)\ldots 2 \\ &= 2^K K! \\ &\simeq \left(\frac{2K}{e}\right)^K \sqrt{2\pi K} \\ &= \left(\frac{2K+1}{e}\right)^{K+1/2}\sqrt{\frac{e}{2K+1}}\left(\frac{2K}{2K+1}\right)^K \sqrt{2\pi K} \\ &\simeq \left(\frac{N}{e}\right)^{N/2}\sqrt{\frac{e}{2K}}\cdot\frac{1}{\sqrt{e}}\cdot\sqrt{2\pi K} \\ &= \left(\frac{N}{e}\right)^{N/2}\sqrt{\pi} \end{aligned}$$

(3) Back to the case where $N = 2K$ is even, by using (2) we obtain:

$$\begin{aligned}(N+1)!! &\simeq \left(\frac{N+1}{e}\right)^{(N+1)/2}\sqrt{\pi}\\ &= \left(\frac{N+1}{e}\right)^{N/2}\sqrt{\frac{N+1}{e}}\cdot\sqrt{\pi}\\ &= \left(\frac{N}{e}\right)^{N/2}\left(\frac{N+1}{N}\right)^{N/2}\sqrt{\frac{N+1}{e}}\cdot\sqrt{\pi}\\ &\simeq \left(\frac{N}{e}\right)^{N/2}\sqrt{e}\cdot\sqrt{\frac{N}{e}}\cdot\sqrt{\pi}\\ &= \left(\frac{N}{e}\right)^{N/2}\sqrt{\pi N}\end{aligned}$$

(4) Finally, back to the case where $N = 2K+1$ is odd, by using (1) we obtain:

$$\begin{aligned}(N+1)!! &\simeq \left(\frac{N+1}{e}\right)^{(N+1)/2}\sqrt{2}\\ &= \left(\frac{N+1}{e}\right)^{N/2}\sqrt{\frac{N+1}{e}}\cdot\sqrt{2}\\ &= \left(\frac{N}{e}\right)^{N/2}\left(\frac{N+1}{N}\right)^{N/2}\sqrt{\frac{N+1}{e}}\cdot\sqrt{2}\\ &\simeq \left(\frac{N}{e}\right)^{N/2}\sqrt{e}\cdot\sqrt{\frac{N}{e}}\cdot\sqrt{2}\\ &= \left(\frac{N}{e}\right)^{N/2}\sqrt{2N}\end{aligned}$$

Thus, we have proved the estimates in the statement. □

We can now estimate the volumes of the spheres, as follows:

THEOREM 13.15. *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. We use the formula for V found in Theorem 13.10, namely:

$$V = \left(\frac{\pi}{2}\right)^{[N/2]}\frac{2^N}{(N+1)!!}$$

In the case where N is even, the estimate goes as follows:

$$\begin{aligned} V &= \left(\frac{\pi}{2}\right)^{N/2}\frac{2^N}{(N+1)!!}\\ &\simeq \left(\frac{\pi}{2}\right)^{N/2}2^N\left(\frac{e}{N}\right)^{N/2}\frac{1}{\sqrt{\pi N}}\\ &= \left(\frac{2\pi e}{N}\right)^{N/2}\frac{1}{\sqrt{\pi N}}\end{aligned}$$

In the case where N is odd, 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. □

Getting back now to our main result so far, Theorem 13.10, we can compute in the same way the area of the sphere, the result being as follows:

THEOREM 13.16. *The area of the unit sphere in $\mathbb{R}^N$ is given by*

$$A = \left(\frac{\pi}{2}\right)^{[N/2]}\frac{2^N}{(N-1)!!}$$

with the our usual convention for double factorials, namely:

$$N!! = (N-1)(N-3)(N-5)\ldots$$

In particular, at $N = 2, 3, 4$ we obtain respectively $A = 2\pi, 4\pi, 2\pi^2$.

PROOF. Regarding the first assertion, there is no need to compute again, because the formula in the statement can be deduced from Theorem 13.10, as follows:

(1) We can either use the "pizza" argument from chapter 1, which shows that the area and volume of the sphere in $\mathbb{R}^N$ are related by the following formula:

$$A = N \cdot V$$

Together with the formula in Theorem 13.10 for V, this gives the result.

(2) Or, we can start the computation in the same way as we started the proof of Theorem 13.10, the beginning of this computation being as follows:

$$vol(S^+) = \int_0^{\pi/2} \ldots \int_0^{\pi/2} \sin^{N-2} t_1 \ldots \sin t_{N-2}\, dt_1 \ldots dt_{N-1}$$

Now by comparing with the beginning of the proof of Theorem 13.10, the only thing that changes is the following quantity, which now dissapears:

$$\int_0^1 r^{N-1}\, dr = \frac{1}{N}$$

Thus, we have $vol(S^+) = N \cdot vol(B^+)$, and so we obtain the following formula:

$$vol(S) = N \cdot vol(B)$$

But this means $A = N \cdot V$, and together with the formula in Theorem 13.10 for V, this gives the result. As for the last assertion, this can be either worked out directly, or deduced from the results for volumes that we have so far, by multiplying by N. □

13d. Spherical integrals

Let us discuss now the computation of the arbitrary integrals over the sphere. We will need a technical result extending Proposition 13.9, as follows:

THEOREM 13.17. *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, 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. We use the same idea as in Proposition 13.9. 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 Proposition 13.9, 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:

$$\begin{aligned} I_{pq} &= \frac{q-1}{p+1} I_{p+2,q-2} \\ &= \frac{q-1}{p+1} \cdot \frac{q-3}{p+3} I_{p+4,q-4} \\ &= \frac{q-1}{p+1} \cdot \frac{q-3}{p+3} \cdot \frac{q-5}{p+5} I_{p+6,q-6} \\ &= \vdots \\ &= \frac{p!!q!!}{(p+q-1)!!} I_{p+q-1,1} \end{aligned}$$

In order to compute the last term, observe that we have:

$$\begin{aligned} I_{p1} &= \int_0^{\pi/2} \cos^p t \sin t \, dt \\ &= -\frac{1}{p+1} \int_0^{\pi/2} (\cos^{p+1} t)' \, dt \\ &= \frac{1}{p+1} \end{aligned}$$

Thus, we can finish our computation in the case p, q odd, as follows:

$$\begin{aligned} I_{pq} &= \frac{p!!q!!}{(p+q-1)!!} I_{p+q-1,1} \\ &= \frac{p!!q!!}{(p+q-1)!!} \cdot \frac{1}{p+q} \\ &= \frac{p!!q!!}{(p+q+1)!!} \end{aligned}$$

Thus, we obtain the formula in the statement, the exponent of $\pi/2$ appearing there being $\varepsilon(p)\varepsilon(q) = 0 \cdot 0 = 0$ in the present case, and this finishes the proof. □

We can now integrate over the spheres, as follows:

THEOREM 13.18. *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 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 the exponents k_i are even. As a first observation, the result holds indeed at $N = 2$, due to the formula from Theorem 13.17, which reads:

$$\int_0^{\pi/2} \cos^p t \sin^q t \, dt = \left(\frac{\pi}{2}\right)^{\varepsilon(p)\varepsilon(q)} \frac{p!!q!!}{(p+q+1)!!} = \frac{p!!q!!}{(p+q+1)!!}$$

In the general case now, where the dimension $N \in \mathbb{N}$ is arbitrary, the integral in the statement can be written in spherical coordinates, as follows:

$$I = \frac{2^N}{A} \int_0^{\pi/2} \ldots \int_0^{\pi/2} x_1^{k_1} \ldots x_N^{k_N} \, J \, dt_1 \ldots dt_{N-1}$$

Here A is the area of the sphere, J is the Jacobian, and the 2^N factor comes from the restriction to the $1/2^N$ part of the sphere where all the coordinates are positive. According to Theorem 13.16, the normalization constant in front of the integral is:

$$\frac{2^N}{A} = \left(\frac{2}{\pi}\right)^{[N/2]} (N-1)!!$$

As for the unnormalized integral, this is given by:

$$\begin{aligned} I' = \int_0^{\pi/2} \dots \int_0^{\pi/2} \quad & (\cos t_1)^{k_1} (\sin t_1 \cos t_2)^{k_2} \\ & \vdots \\ & (\sin t_1 \sin t_2 \dots \sin t_{N-2} \cos t_{N-1})^{k_{N-1}} \\ & (\sin t_1 \sin t_2 \dots \sin t_{N-2} \sin t_{N-1})^{k_N} \\ & \sin^{N-2} t_1 \sin^{N-3} t_2 \dots \sin^2 t_{N-3} \sin t_{N-2} \\ & dt_1 \dots dt_{N-1} \end{aligned}$$

By rearranging the terms, we obtain:

$$\begin{aligned} I' \;=\; & \int_0^{\pi/2} \cos^{k_1} t_1 \sin^{k_2+\dots+k_N+N-2} t_1 \, dt_1 \\ & \int_0^{\pi/2} \cos^{k_2} t_2 \sin^{k_3+\dots+k_N+N-3} t_2 \, dt_2 \\ & \vdots \\ & \int_0^{\pi/2} \cos^{k_{N-2}} t_{N-2} \sin^{k_{N-1}+k_N+1} t_{N-2} \, dt_{N-2} \\ & \int_0^{\pi/2} \cos^{k_{N-1}} t_{N-1} \sin^{k_N} t_{N-1} \, dt_{N-1} \end{aligned}$$

Now by using the above-mentioned formula at $N = 2$, this gives:

$$\begin{aligned} I' \;=\; & \frac{k_1!!(k_2+\dots+k_N+N-2)!!}{(k_1+\dots+k_N+N-1)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(N-2)} \\ & \frac{k_2!!(k_3+\dots+k_N+N-3)!!}{(k_2+\dots+k_N+N-2)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(N-3)} \\ & \vdots \\ & \frac{k_{N-2}!!(k_{N-1}+k_N+1)!!}{(k_{N-2}+k_{N-1}+l_N+2)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(1)} \\ & \frac{k_{N-1}!!k_N!!}{(k_{N-1}+k_N+1)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(0)} \end{aligned}$$

Now let F be the part involving the double factorials, and P be the part involving the powers of $\pi/2$, so that $I' = F \cdot P$. Regarding F, by cancelling terms we have:

$$F = \frac{k_1!! \ldots k_N!!}{(\Sigma k_i + N - 1)!!}$$

As in what regards P, by summing the exponents, we obtain $P = \left(\frac{\pi}{2}\right)^{[N/2]}$. We can now put everything together, and we obtain:

$$\begin{aligned} I &= \frac{2^N}{A} \times F \times P \\ &= \left(\frac{2}{\pi}\right)^{[N/2]} (N-1)!! \times \frac{k_1!! \ldots k_N!!}{(\Sigma k_i + N - 1)!!} \times \left(\frac{\pi}{2}\right)^{[N/2]} \\ &= \frac{(N-1)!! k_1!! \ldots k_N!!}{(\Sigma k_i + N - 1)!!} \end{aligned}$$

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

We have the following useful generalization of the above formula:

THEOREM 13.19. *We have the following integration formula over the sphere $S^{N-1}_{\mathbb{R}} \subset \mathbb{R}^N$, with respect to the normalized measure, valid for any exponents $k_i \in \mathbb{N}$,*

$$\int_{S^{N-1}_{\mathbb{R}}} |x_1^{k_1} \ldots x_N^{k_N}| \, dx = \left(\frac{2}{\pi}\right)^{\Sigma(k_1,\ldots,k_N)} \frac{(N-1)!! k_1!! \ldots k_N!!}{(N + \Sigma k_i - 1)!!}$$

with $\Sigma = [odds/2]$ if N is odd and $\Sigma = [(odds+1)/2]$ if N is even, where "odds" denotes the number of odd numbers in the sequence $k_1, \ldots, k_N$.

PROOF. As before, the formula holds at $N = 2$, due to Theorem 13.17. In general, the integral in the statement can be written in spherical coordinates, as follows:

$$I = \frac{2^N}{A} \int_0^{\pi/2} \ldots \int_0^{\pi/2} x_1^{k_1} \ldots x_N^{k_N} J \, dt_1 \ldots dt_{N-1}$$

Here A is the area of the sphere, J is the Jacobian, and the 2^N factor comes from the restriction to the $1/2^N$ part of the sphere where all the coordinates are positive. The normalization constant in front of the integral is, as before:

$$\frac{2^N}{A} = \left(\frac{2}{\pi}\right)^{[N/2]} (N-1)!!$$

As for the unnormalized integral, this can be written as before, as follows:

$$\begin{aligned} I' &= \int_0^{\pi/2} \cos^{k_1} t_1 \sin^{k_2+\ldots+k_N+N-2} t_1 \, dt_1 \\ &\quad \int_0^{\pi/2} \cos^{k_2} t_2 \sin^{k_3+\ldots+k_N+N-3} t_2 \, dt_2 \\ &\quad \vdots \\ &\quad \int_0^{\pi/2} \cos^{k_{N-2}} t_{N-2} \sin^{k_{N-1}+k_N+1} t_{N-2} \, dt_{N-2} \\ &\quad \int_0^{\pi/2} \cos^{k_{N-1}} t_{N-1} \sin^{k_N} t_{N-1} \, dt_{N-1} \end{aligned}$$

Now by using the formula at $N = 2$, we get:

$$\begin{aligned} I' &= \frac{\pi}{2} \cdot \frac{k_1!!(k_2 + \ldots + k_N + N - 2)!!}{(k_1 + \ldots + k_N + N - 1)!!} \left(\frac{2}{\pi}\right)^{\delta(k_1, k_2+\ldots+k_N+N-2)} \\ &\quad \frac{\pi}{2} \cdot \frac{k_2!!(k_3 + \ldots + k_N + N - 3)!!}{(k_2 + \ldots + k_N + N - 2)!!} \left(\frac{2}{\pi}\right)^{\delta(k_2, k_3+\ldots+k_N+N-3)} \\ &\quad \vdots \\ &\quad \frac{\pi}{2} \cdot \frac{k_{N-2}!!(k_{N-1} + k_N + 1)!!}{(k_{N-2} + k_{N-1} + k_N + 2)!!} \left(\frac{2}{\pi}\right)^{\delta(k_{N-2}, k_{N-1}+k_N+1)} \\ &\quad \frac{\pi}{2} \cdot \frac{k_{N-1}!!k_N!!}{(k_{N-1} + k_N + 1)!!} \left(\frac{2}{\pi}\right)^{\delta(k_{N-1}, k_N)} \end{aligned}$$

In order to compute this quantity, let us denote by F the part involving the double factorials, and by P the part involving the powers of $\pi/2$, so that we have:

$$I' = F \cdot P$$

Regarding F, there are many cancellations there, and we end up with:

$$F = \frac{k_1!! \ldots k_N!!}{(\Sigma k_i + N - 1)!!}$$

As in what regards P, the δ exponents on the right sum up to the following number:

$$\Delta(k_1, \ldots, k_N) = \sum_{i=1}^{N-1} \delta(k_i, k_{i+1} + \ldots + k_N + N - i - 1)$$

In other words, with this notation, the above formula reads:

$$\begin{aligned} I' &= \left(\frac{\pi}{2}\right)^{N-1} \frac{k_1!!k_2!!\ldots k_N!!}{(k_1+\ldots+k_N+N-1)!!} \left(\frac{2}{\pi}\right)^{\Delta(k_1,\ldots,k_N)} \\ &= \left(\frac{2}{\pi}\right)^{\Delta(k_1,\ldots,k_N)-N+1} \frac{k_1!!k_2!!\ldots k_N!!}{(k_1+\ldots+k_N+N-1)!!} \\ &= \left(\frac{2}{\pi}\right)^{\Sigma(k_1,\ldots,k_N)-[N/2]} \frac{k_1!!k_2!!\ldots k_N!!}{(k_1+\ldots+k_N+N-1)!!} \end{aligned}$$

Here the formula relating Δ to Σ follows from a number of simple observations, the first of which being the act that, due to obvious parity reasons, the sequence of δ numbers appearing in the definition of Δ cannot contain two consecutive zeroes. Together with $I = (2^N/V)I'$, this gives the formula in the statement. □

Let us discuss as well the complex versions of the above results. We have the following variation of the formula in Theorem 13.19, dealing with the complex sphere:

THEOREM 13.20. *We have the following integration formula over the complex sphere $S_{\mathbb{C}}^{N-1} \subset \mathbb{C}^N$, with respect to the normalized uniform measure,*

$$\int_{S_{\mathbb{C}}^{N-1}} |z_1|^{2k_1} \ldots |z_N|^{2k_N}\, dz = \frac{(N-1)!k_1!\ldots k_n!}{(N+\sum k_i-1)!}$$

valid for any exponents $k_i \in \mathbb{N}$. As for the other polynomial integrals in $z_1, \ldots, z_N$ and their conjugates $\bar{z}_1, \ldots, \bar{z}_N$, these all vanish.

PROOF. Consider an arbitrary polynomial integral over $S_{\mathbb{C}}^{N-1}$, written as follows:

$$I = \int_{S_{\mathbb{C}}^{N-1}} z_{i_1}\bar{z}_{i_2} \ldots z_{i_{2k-1}}\bar{z}_{i_{2k}}\, dz$$

By using transformations of type $p \to \lambda p$ with $|\lambda| = 1$, we see that this integral I vanishes, unless each z_a appears as many times as $\bar{z}_a$ does, and this gives the last assertion. So, assume now that we are in the non-vanishing case. Then the k_a copies of z_a and the k_a copies of $\bar{z}_a$ produce by multiplication a factor $|z_a|^{2k_a}$, so we have:

$$I = \int_{S_{\mathbb{C}}^{N-1}} |z_1|^{2k_1} \ldots |z_N|^{2k_N}\, dz$$

Now by using the standard identification $S_{\mathbb{C}}^{N-1} \simeq S_{\mathbb{R}}^{2N-1}$, we obtain:

$$\begin{aligned} I &= \int_{S_{\mathbb{R}}^{2N-1}} (x_1^2+y_1^2)^{k_1} \ldots (x_N^2+y_N^2)^{k_N}\, d(x,y) \\ &= \sum_{r_1\ldots r_N} \binom{k_1}{r_1} \ldots \binom{k_N}{r_N} \int_{S_{\mathbb{R}}^{2N-1}} x_1^{2k_1-2r_1} y_1^{2r_1} \ldots x_N^{2k_N-2r_N} y_N^{2r_N}\, d(x,y) \end{aligned}$$

By using the formula in Theorem 13.19, we obtain:

$$\begin{aligned} I &= \sum_{r_1\ldots r_N}\binom{k_1}{r_1}\ldots\binom{k_N}{r_N}\frac{(2N-1)!!(2r_1)!!\ldots(2r_N)!!(2k_1-2r_1)!!\ldots(2k_N-2r_N)!!}{(2N+2\sum k_i-1)!!}\\ &= \sum_{r_1\ldots r_N}\binom{k_1}{r_1}\ldots\binom{k_N}{r_N}\frac{2^{N-1}(N-1)!\prod(2r_i)!/(2^{r_i}r_i!)\prod(2k_i-2r_i)!/(2^{k_i-r_i}(k_i-r_i)!)}{2^{N+\sum k_i-1}(N+\sum k_i-1)!}\\ &= \sum_{r_1\ldots r_N}\binom{k_1}{r_1}\ldots\binom{k_N}{r_N}\frac{(N-1)!(2r_1)!\ldots(2r_N)!(2k_1-2r_1)!\ldots(2k_N-2r_N)!}{4^{\sum k_i}(N+\sum k_i-1)!r_1!\ldots r_N!(k_1-r_1)!\ldots(k_N-r_N)!} \end{aligned}$$

Now observe that can rewrite this quantity in the following way:

$$\begin{aligned} I &= \sum_{r_1\ldots r_N}\frac{k_1!\ldots k_N!(N-1)!(2r_1)!\ldots(2r_N)!(2k_1-2r_1)!\ldots(2k_N-2r_N)!}{4^{\sum k_i}(N+\sum k_i-1)!(r_1!\ldots r_N!(k_1-r_1)!\ldots(k_N-r_N)!)^2}\\ &= \sum_{r_1}\binom{2r_1}{r_1}\binom{2k_1-2r_1}{k_1-r_1}\ldots\sum_{r_N}\binom{2r_N}{r_N}\binom{2k_N-2r_N}{k_N-r_N}\frac{(N-1)!k_1!\ldots k_N!}{4^{\sum k_i}(N+\sum k_i-1)!}\\ &= 4^{k_1}\times\ldots\times 4^{k_N}\times\frac{(N-1)!k_1!\ldots k_N!}{4^{\sum k_i}(N+\sum k_i-1)!}\\ &= \frac{(N-1)!k_1!\ldots k_N!}{(N+\sum k_i-1)!} \end{aligned}$$

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

We will see applications of all this in the next chapter, with some quite conceptual results regarding the spherical coordinates, in the $N\to\infty$ limit, obtained by processing the above results, by using standard tools from probability.

13e. Exercises

Here are some exercises on the above, often insisting on missing details:

EXERCISE 13.21. *Clarify the details of the change of variable formula.*

EXERCISE 13.22. *Find the ranges of angles in the spherical coordinate formula.*

EXERCISE 13.23. *Further refine the Stirling formula, with more terms.*

EXERCISE 13.24. *Prove* $\sum_r\binom{2r}{r}\binom{2k-2r}{k-r}=4^k$*, used in the above, at the end.*

As bonus exercise, compute areas and volumes, as many as you can.

CHAPTER 14

Normal variables

14a. Probability basics

Good news, with the calculus that we know, we can have a crash course in advanced probability. We already met some probability in this book, scattered at the end of chapters 4, 6, 7, and we will first review that material. Then, helped by the Gauss integral formula $\int e^{-x^2} = \sqrt{\pi}$ that we just learned, and by multivariable integration in general, we will develop the theory of normal variables, real and complex, which is central to advanced probability. And then, there will certainly be more, all interesting things.

Sounds exciting, doesn't it. Cat however seems unfazed, and declares:

CAT 14.1. *Probability is the same thing as measure theory. Read Rudin.*

Damn cat, seems like we disagree more are more as time goes by, especially on these multivariable integration topics. So, let me ask you cat, how many times did you come upon a function which is not integrable? How many times did you fail catching a mouse, due to a failure of Fubini? What about catching birds, does the Zorn lemma really help there? Things are nice and smooth in life, or at least that's my belief.

CAT 14.2. *Yes for mice and birds, but electrons can be quite tricky.*

Humm, good point, but let's leave electrons for later, for chapter 16. So, forgetting now about philosophy, and pedagogy matters, but dear reader feel free to have your own opinion here, and why not agreeing with cat, and going ahead with our plan, let us first review the few things that we know about probability, from chapters 4, 6, 7.

There will be of course a bit of redundancy, but always good to talk again about that things, with a bit less details of course, this time. As a starting point, we have:

DEFINITION 14.3. *Let X be a probability space, that is, a space with a probability measure, and with the corresponding integration denoted $\mathbb{E}$, and called expectation.*

(1) *The random variables are the real functions $f \in L^\infty(X)$.*
(2) *The moments of such a variable are the numbers $M_k(f) = \mathbb{E}(f^k)$.*
(3) *The law of such a variable is the measure given by $M_k(f) = \int_{\mathbb{R}} x^k d\mu_f(x)$.*

Here, as explained in chapter 4, 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:

$$\mathbb{E}(\varphi(f)) = \int_{\mathbb{R}} \varphi(x) d\mu_f(x)$$

Indeed, having this for monomials $\varphi(x) = x^n$, as in the statement, is the same as having it for polynomials $\varphi \in \mathbb{R}[X]$, which in turn is the same as having it for characteristic functions $\varphi = \chi_I$ of measurable sets $I \subset \mathbb{R}$. Thus, in the end, what we need is:

$$\mathbb{P}(f \in I) = \mu_f(I)$$

But this formula can serve as a definition for μ_f, and we are done.

Regarding now independence, as explained in chapter 7, we can formulate here:

DEFINITION 14.4. *Two variables $f, g \in L^\infty(X)$ are called independent when*

$$\mathbb{E}(f^k g^l) = \mathbb{E}(f^k)\,\mathbb{E}(g^l)$$

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:

$$\mathbb{E}[\varphi(f)\psi(g)] = \mathbb{E}[\varphi(f)]\,\mathbb{E}[\psi(g)]$$

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:

$$\mathbb{P}(f \in I, g \in J) = \mathbb{P}(f \in I)\,\mathbb{P}(g \in J)$$

As a first result now, that we know well from chapter 7, we have:

THEOREM 14.5. *Assuming that $f, g \in L^\infty(X)$ are independent, we have*

$$\mu_{f+g} = \mu_f * \mu_g$$

where $*$ is the convolution of real probability measures.

PROOF. We have the following computation, using the independence of f, g:

$$\int_{\mathbb{R}} x^k d\mu_{f+g}(x) = \mathbb{E}((f+g)^k) = \sum_r \binom{k}{r} M_r(f) M_{k-r}(g)$$

On the other hand, by using the Fubini theorem, we have as well:

$$\begin{aligned} \int_{\mathbb{R}} x^k d(\mu_f * \mu_g)(x) &= \int_{\mathbb{R}\times\mathbb{R}} (x+y)^k d\mu_f(x) d\mu_g(y) \\ &= \sum_r \binom{k}{r} M_r(f) M_{k-r}(g) \end{aligned}$$

Thus μ_{f+g} and $\mu_f * \mu_g$ have the same moments, so they coincide, as desired. □

As a second result on independence, which is more advanced, we have:

THEOREM 14.6. *Assuming that $f, g \in L^\infty(X)$ are independent, we have*

$$F_{f+g} = F_f F_g$$

where $F_f(x) = \mathbb{E}(e^{ixf})$ is the Fourier transform.

PROOF. This is something that we know too from chapter 7, coming from:

$$\begin{aligned} F_{f+g}(x) &= \int_{\mathbb{R}} e^{ixz} d(\mu_f * \mu_g)(z) \\ &= \int_{\mathbb{R}\times\mathbb{R}} e^{ix(z+t)} d\mu_f(z) d\mu_g(t) \\ &= \int_{\mathbb{R}} e^{ixz} d\mu_f(z) \int_{\mathbb{R}} e^{ixt} d\mu_g(t) \\ &= F_f(x) F_g(x) \end{aligned}$$

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

So long for laws, moments and independence, at the theoretical level. In what follows Definitions 14.3 and 14.4 and Theorems 14.5 and 14.6 are what we need, I mean we will certainly get away with that, as general theory, and this no matter what cat says.

However, still staying general, we will need as well, from time to time, some key complex analysis results that we established in chapter 6. First, we have:

THEOREM 14.7. *The density of a real probability measure μ can be recaptured from the sequence of moments $\{M_k\}_{k\geq 0}$ via the Stieltjes inversion formula*

$$d\mu(x) = \lim_{t\searrow 0} -\frac{1}{\pi} Im\left(G(x+it)\right) \cdot dx$$

where the function on the right, given in terms of moments by

$$G(\xi) = \xi^{-1} + M_1\xi^{-2} + M_2\xi^{-3} + \ldots$$

is the Cauchy transform of the measure μ.

PROOF. This is something quite subtle and heavy, and for the full proof, along with some basic applications, we refer to chapter 6, the idea being as follows:

(1) Regarding the proof, the Cauchy transform of our measure μ is given by:

$$G(\xi) = \xi^{-1} \sum_{k=0}^{\infty} M_k \xi^{-k} = \int_{\mathbb{R}} \frac{1}{\xi - y}\, d\mu(y)$$

Now with $\xi = x + it$, we obtain from this the following formula:

$$\begin{aligned} Im(G(x+it)) &= \int_{\mathbb{R}} Im\left(\frac{1}{x-y+it}\right) d\mu(y) \\ &= -\int_{\mathbb{R}} \frac{t}{(x-y)^2+t^2}\, d\mu(y) \end{aligned}$$

By integrating over $[a, b]$ we obtain, with the change of variables $x = y + tz$:

$$\begin{aligned} \int_a^b Im(G(x+it))dx &= -\int_{\mathbb{R}} \int_{(a-y)/t}^{(b-y)/t} \frac{t}{(tz)^2+t^2}\, t\, dz\, d\mu(y) \\ &= -\int_{\mathbb{R}} \int_{(a-y)/t}^{(b-y)/t} \frac{1}{1+z^2}\, dz\, d\mu(y) \\ &= -\int_{\mathbb{R}} \left(\arctan\frac{b-y}{t} - \arctan\frac{a-y}{t}\right) d\mu(y) \end{aligned}$$

(2) The point now is that with $t \searrow 0$ we have the following estimates:

$$\lim_{t\searrow 0}\left(\arctan\frac{b-y}{t} - \arctan\frac{a-y}{t}\right) = \begin{cases} \frac{\pi}{2} - \frac{\pi}{2} = 0 & (y < a) \\ \frac{\pi}{2} - 0 = \frac{\pi}{2} & (y = a) \\ \frac{\pi}{2} - (-\frac{\pi}{2}) = \pi & (a < y < b) \\ 0 - (-\frac{\pi}{2}) = \frac{\pi}{2} & (y = b) \\ -\frac{\pi}{2} - (-\frac{\pi}{2}) = 0 & (y > b) \end{cases}$$

We therefore obtain the following formula, which proves our result:

$$\lim_{t\searrow 0} \int_a^b Im(G(x+it))dx = -\pi\left(\mu(a,b) + \frac{\mu(a)+\mu(b)}{2}\right)$$

(3) As applications, the first thing that we did in chapter 6 was to find the measure having as even moments the Catalan numbers, $C_k = \frac{1}{k+1}\binom{2k}{k}$, and having all odd moments 0. We found the following measure, called Wigner semicircle law on $[-2, 2]$:

$$\gamma_1 = \frac{1}{2\pi}\sqrt{4-x^2}dx$$

(4) We also computed the measure having as moments the Catalan numbers, $M_k = C_k$. We found the following measure, called Marchenko-Pastur law on $[0, 4]$:

$$\pi_1 = \frac{1}{2\pi}\sqrt{4x^{-1}-1}\, dx$$

(5) Next, we found that the measure having as moments the central binomial coefficients, $D_k = \binom{2k}{k}$, is the following measure, called arcsine law on $[0, 4]$:

$$\alpha_1 = \frac{1}{\pi\sqrt{x(4-x)}}\,dx$$

(6) Finally, for the middle binomial coefficients, $E_k = \binom{k}{[k/2]}$, we found the following law on $[-2, 2]$, called "square root" of the arcsine law on $[0, 4]$:

$$\sigma_1 = \frac{1}{2\pi}\sqrt{\frac{2+x}{2-x}}\,dx$$

(7) All this is very nice, but I can hear you screaming, what is the point with all these fringe measures. In answer, these measures are not fringe at all, quite the opposite, they are the main laws in Random Matrix Theory (RMT). More on them later. □

We have as well the following result, also from chapter 6, which can be useful too:

THEOREM 14.8. *A sequence of numbers $M_0, M_1, M_2, M_3, \ldots \in \mathbb{R}$, with $M_0 = 1$, is the series of moments of a real probability measure μ precisely when:*

$$\left|M_0\right| \geq 0 \quad , \quad \begin{vmatrix} M_0 & M_1 \\ M_1 & M_2 \end{vmatrix} \geq 0 \quad , \quad \begin{vmatrix} M_0 & M_1 & M_2 \\ M_1 & M_2 & M_3 \\ M_2 & M_3 & M_4 \end{vmatrix} \geq 0 \quad , \quad \ldots$$

That is, the associated Hankel determinants must be all positive.

PROOF. This is something a bit heavier, and as a first observation, the positivity conditions in the statement tell us that the following linear forms must be positive:

$$\sum_{i,j=1}^{n} c_i\bar{c}_j M_{i+j} \geq 0$$

But this is something very classical, in one sense the result being elementary, coming from the following computation, which shows that we have positivity indeed:

$$\int_{\mathbb{R}} \left|\sum_{i=1}^{n} c_i x^i\right|^2 d\mu(x) = \int_{\mathbb{R}} \sum_{i,j=1}^{n} c_i\bar{c}_j x^{i+j} d\mu(x) = \sum_{i,j=1}^{n} c_i\bar{c}_j M_{i+j}$$

As for the other sense, here the result comes once again from the above formula, this time via some standard study, inspired from the positivity results from chapter 12. □

All this is very nice, and we have some interesting theory going on. Let us discuss now some illustrations. We will first talk about discrete probability. We have:

DEFINITION 14.9. *The Poisson law of parameter $t > 0$ is the measure*

$$p_t = e^{-t} \sum_{k \in \mathbb{N}} \frac{t^k}{k!} \, \delta_k$$

with the letter "p" standing for Poisson.

We have already talked about these laws, first in chapter 4, with an elementary discussion, using binomials and factorials, and then in chapter 7 too. So, let us quickly review what we know. Going directly for the kill, Fourier transform computation, we have:

THEOREM 14.10. *The Fourier transform of p_t is given by:*

$$F_{p_t}(y) = \exp\big((e^{iy} - 1)t\big)$$

In particular we have $p_s * p_t = p_{s+t}$, called convolution semigroup property.

PROOF. We have indeed the following computation:

$$\begin{aligned} F_{p_t}(y) &= e^{-t} \sum_k \frac{t^k}{k!} F_{\delta_k}(y) \\ &= e^{-t} \sum_k \frac{t^k}{k!} \, e^{iky} \\ &= e^{-t} \sum_k \frac{(e^{iy}t)^k}{k!} \\ &= \exp\big((e^{iy} - 1)t\big) \end{aligned}$$

As for the second assertion, this follows from the fact that $\log F_{p_t}$ is linear in t, via the linearization property for the convolution from Theorem 14.6. □

The above result suggests that the laws p_t should appear as some sort of exponentials with respect to convolution, and yes indeed, this is the case, as shown by:

PROPOSITION 14.11. *The Poisson laws appear as formal exponentials*

$$p_t = \sum_k \frac{t^k (\delta_1 - \delta_0)^{*k}}{k!}$$

with respect to the convolution of measures $*$.

PROOF. By using the binomial formula, the measure on the right is:

$$\begin{aligned}\mu &= \sum_k \frac{t^k}{k!} \sum_{r+s=k} (-1)^s \frac{k!}{r!s!} \delta_r \\ &= \sum_r \frac{t^r \delta_r}{r!} \sum_s \frac{(-1)^s}{s!} \\ &= \frac{1}{e} \sum_r \frac{t^r \delta_r}{r!} \\ &= p_t\end{aligned}$$

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

As a main result now, we have the Poisson Limit Theorem, as follows:

THEOREM 14.12 (PLT). *We have the following convergence, in moments,*

$$\left(\left(1 - \frac{t}{n} \right) \delta_0 + \frac{t}{n} \delta_1 \right)^{*n} \to p_t$$

for any $t > 0$.

PROOF. Indeed, if we denote by ν_n the measure under the convolution sign, we have the following computation, for the Fourier transform of the limit:

$$\begin{aligned} F_{\delta_r}(y) = e^{iry} &\implies F_{\nu_n}(y) = \left(1 - \frac{t}{n} \right) + \frac{t}{n} e^{iy} \\ &\implies F_{\nu_n^{*n}}(y) = \left(\left(1 - \frac{t}{n} \right) + \frac{t}{n} e^{iy} \right)^n \\ &\implies F_{\nu_n^{*n}}(y) = \left(1 + \frac{(e^{iy} - 1)t}{n} \right)^n \\ &\implies F(y) = \exp\left((e^{iy} - 1)t \right)\end{aligned}$$

Thus, we obtain indeed the Fourier transform of p_t, as desired. □

Finally, one more thing that we know about the Poisson laws are some interesting formulae for their moments, from chapter 4. The result here was as follows:

THEOREM 14.13. *The moments of p_1 are the Bell numbers,*

$$M_k(p_1) = |P(k)|$$

where $P(k)$ is the set of partitions of $\{1, \ldots, k\}$. More generally, we have

$$M_k(p_t) = \sum_{\pi \in P(k)} t^{|\pi|}$$

for any $t > 0$, where $|.|$ is the number of blocks.

PROOF. We know that the moments of p_1 are given by the following formula:

$$M_k = \frac{1}{e}\sum_r \frac{r^k}{r!}$$

We therefore have the following recurrence formula for these moments:

$$\begin{aligned} M_{k+1} &= \frac{1}{e}\sum_r \frac{r^k}{r!}\left(1+\frac{1}{r}\right)^k \\ &= \frac{1}{e}\sum_r \frac{r^k}{r!}\sum_s \binom{k}{s} r^{-s} \\ &= \sum_s \binom{k}{s} M_{k-s} \end{aligned}$$

But the Bell numbers $B_k = |P(k)|$ satisfy the same recurrence, so we have $M_k = B_k$, as claimed. Next, we know that the moments of p_t with $t > 0$ are given by:

$$N_k = e^{-t}\sum_r \frac{t^r r^k}{r!}$$

We therefore have the following recurrence formula for these moments:

$$\begin{aligned} N_{k+1} &= e^{-t}\sum_r \frac{t^{r+1} r^k}{r!}\left(1+\frac{1}{r}\right)^k \\ &= e^{-t}\sum_r \frac{t^{r+1} r^k}{r!}\sum_s \binom{k}{s} r^{-s} \\ &= t\sum_s \binom{k}{s} N_{k-s} \end{aligned}$$

But the numbers $S_k = \sum_{\pi\in P(k)} t^{|\pi|}$ are easily seen to satisfy the same recurrence, with the same initial values, namely t and $t + t^2$, so we have $N_k = S_k$, as claimed. □

Summarizing, we know so far what probability theory is, modulo perhaps some annoying foundational details, that do not seem needed for computations. Then we have some powerful tools, namely convolution, Fourier, Stieltjes, and Hankel determinants if needed. Then, as a main achievement, we have a full theory in the main discrete case, that of the Poisson laws. And finally, in the continuous case, we came across a number of bizarre laws, namely Wigner, Marchenko-Pastur, arcsine, and its square root.

This is not bad, as a start. In what follows we will first do the necessary, namely mirror what we know about the Poisson laws, with a complete study of the normal laws, which are the central objects in continuous probability. And then, we will go for more.

14b. Normal variables

In order to introduce the normal variables, we will need the Gauss integral formula, established in chapter 13. So, let us recall from there that we have:

THEOREM 14.14. *We have the following formula,*

$$\int_{\mathbb{R}} e^{-x^2} dx = \sqrt{\pi}$$

called Gauss integral formula.

PROOF. Let I be the above integral. By using polar coordinates, we obtain:

$$\begin{aligned} I^2 &= \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-x^2-y^2} dxdy \\ &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2} r dr dt \\ &= 2\pi \times \frac{1}{2} \end{aligned}$$

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

As a main application of the Gauss formula, we can now formulate:

DEFINITION 14.15. *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 $x = \sqrt{2t}\, y$:

$$\begin{aligned} \int_{\mathbb{R}} e^{-x^2/2t} dx &= \int_{\mathbb{R}} e^{-y^2} \sqrt{2t}\, dy \\ &= \sqrt{2t} \int_{\mathbb{R}} e^{-y^2} dy \\ &= \sqrt{2t} \times \sqrt{\pi} \\ &= \sqrt{2\pi t} \end{aligned}$$

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 14.16. *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 $V = M_2 = t$. □

Here is another result, which is the key one for the study of the normal laws:

THEOREM 14.17. *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:

$$\begin{aligned} F_{g_t}(x) &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-y^2/2t + ixy} dy \\ &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-(y/\sqrt{2t} - \sqrt{t/2}ix)^2 - tx^2/2} dy \\ &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-z^2 - tx^2/2} \sqrt{2t} dz \\ &= \frac{1}{\sqrt{\pi}} e^{-tx^2/2} \int_{\mathbb{R}} e^{-z^2} dz \\ &= \frac{1}{\sqrt{\pi}} e^{-tx^2/2} \cdot \sqrt{\pi} \\ &= e^{-tx^2/2} \end{aligned}$$

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 14.18 (CLT). *Given random variables $f_1, f_2, f_3, \ldots \in L^\infty(X)$ 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, having as density $\frac{1}{\sqrt{2\pi t}} e^{-y^2/2t} dy$.

PROOF. We use the Fourier transform, which is by definition given by:

$$F_f(x) = \mathbb{E}(e^{ixf})$$

In terms of moments, we have the following formula:

$$\begin{aligned} F_f(x) &= \mathbb{E}\left(\sum_{k=0}^{\infty} \frac{(ixf)^k}{k!}\right) \\ &= \sum_{k=0}^{\infty} \frac{(ix)^k \mathbb{E}(f^k)}{k!} \\ &= \sum_{k=0}^{\infty} \frac{i^k M_k(f)}{k!} x^k \end{aligned}$$

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. □

Let us discuss now some further properties of the normal law. We first have:

PROPOSITION 14.19. *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 $k \in \mathbb{N}$:

$$\begin{aligned} M_k &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} y^k e^{-y^2/2t} dy \\ &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} (ty^{k-1}) \left(-e^{-y^2/2t}\right)' dy \\ &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} t(k-1) y^{k-2} e^{-y^2/2t} dy \\ &= t(k-1) \times \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} y^{k-2} e^{-y^2/2t} dy \\ &= t(k-1) M_{k-2} \end{aligned}$$

Now recall from the proof of Proposition 14.16 that we have $M_0 = 1$, $M_1 = 0$. Thus by recurrence, we are led to the formula in the statement. □

We have the following alternative formulation of the above result:

PROPOSITION 14.20. *The moments of the normal law are the numbers*

$$M_k(g_t) = t^{k/2} |P_2(k)|$$

where $P_2(k)$ is the set of pairings of $\{1, \ldots, k\}$.

PROOF. Let us count the pairings of $\{1, \ldots, k\}$. In order to have such a pairing, we must pair 1 with one of the numbers $2, \ldots, k$, and then use a pairing of the remaining $k-2$ numbers. Thus, we have the following recurrence formula:

$$|P_2(k)| = (k-1)|P_2(k-2)|$$

As for the initial data, this is $P_1 = 0$, $P_2 = 1$. Thus, we are led to the result. □

We are not done yet, and here is one more improvement of the above:

THEOREM 14.21. *The moments of the normal law are the numbers*

$$M_k(g_t) = \sum_{\pi \in P_2(k)} t^{|\pi|}$$

where $P_2(k)$ is the set of pairings of $\{1, \ldots, k\}$, and $|.|$ is the number of blocks.

PROOF. This follows indeed from Proposition 14.20, because the number of blocks of a pairing of $\{1, \ldots, k\}$ is trivially $k/2$, independently of the pairing. □

Let us discuss now the complex analogues of all this, with a notion of complex normal, or Gaussian law. To start with, we have the following definition:

DEFINITION 14.22. *The complex normal, or Gaussian law of parameter $t > 0$ is*

$$G_t = law\left(\frac{1}{\sqrt{2}}(a + ib)\right)$$

where a, b are independent, each following the law g_t.

In short, the complex normal laws appear as natural complexifications of the real normal laws. As in the real case, these measures form convolution semigroups:

PROPOSITION 14.23. *The complex Gaussian laws have the property*

$$G_s * G_t = G_{s+t}$$

for any $s, t > 0$, and so they form a convolution semigroup.

PROOF. This follows indeed from the real result, namely $g_s * g_t = g_{s+t}$, established in Theorem 14.17, simply by taking real and imaginary parts. □

We have as well the following complex analogue of the CLT:

THEOREM 14.24 (CCLT). *Given complex variables $f_1, f_2, f_3, \ldots \in L^\infty(X)$ which are i.i.d., centered, and with common variance $t > 0$, we have*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f_i \sim G_t$$

with $n \to \infty$, in moments.

PROOF. This follows indeed from the real CLT, established in Theorem 14.18, simply by taking the real and imaginary parts of all the variables involved. □

Regarding now the moments, the situation here is more complicated than in the real case, because in order to have good results, we have to deal with both the complex variables, and their conjugates. Let us formulate the following definition:

DEFINITION 14.25. *The moments a complex variable $f \in L^\infty(X)$ are the numbers*

$$M_k = \mathbb{E}(f^k)$$

depending on colored integers $k = \circ \bullet \bullet \circ \ldots$, with the conventions

$$f^\emptyset = 1 \quad , \quad f^\circ = f \quad , \quad f^\bullet = \bar{f}$$

and multiplicativity, in order to define the colored powers f^k.

Observe that, since $f, \bar{f}$ commute, we can permute terms, and restrict the attention to exponents of type $k = \ldots \circ \circ \circ \bullet \bullet \bullet \bullet \ldots$, if we want to. However, our results about the complex Gaussian laws, and other complex laws, later on, will actually look better without doing is, and so we will use Definition 14.25 as stated. We first have:

THEOREM 14.26. *The moments of the complex normal law are given by*

$$M_k(G_t) = \begin{cases} t^p p! & (k \text{ uniform, of length } 2p) \\ 0 & (k \text{ not uniform}) \end{cases}$$

where $k = \circ \bullet \bullet \circ \ldots$ *is called uniform when it contains the same number of* $\circ$ *and* $\bullet$*.*

PROOF. We must compute the moments, with respect to colored integer exponents $k = \circ \bullet \bullet \circ \ldots$, of the variable from Definition 14.22, namely:

$$f = \frac{1}{\sqrt{2}}(a + ib)$$

We can assume that we are in the case $t = 1$, and the proof here goes as follows:

(1) As a first observation, in the case where our exponent $k = \circ \bullet \bullet \circ \ldots$ is not uniform, a standard rotation argument shows that the corresponding moment of f vanishes. To be more precise, the variable $f' = wf$ is complex Gaussian too, for any complex number $w \in \mathbb{T}$, and from $M_k(f) = M_k(f')$ we obtain $M_k(f) = 0$, in this case.

(2) In the uniform case now, where the exponent $k = \circ \bullet \bullet \circ \ldots$ consists of p copies of $\circ$ and p copies of $\bullet$, the corresponding moment can be computed as follows:

$$\begin{aligned} M_k &= \int (f\bar{f})^p \\ &= \frac{1}{2^p} \int (a^2 + b^2)^p \\ &= \frac{1}{2^p} \sum_r \binom{p}{r} \int a^{2r} \int b^{2p-2r} \\ &= \frac{1}{2^p} \sum_r \binom{p}{r} (2r)!!(2p-2r)!! \\ &= \frac{1}{2^p} \sum_r \frac{p!}{r!(p-r)!} \cdot \frac{(2r)!}{2^r r!} \cdot \frac{(2p-2r)!}{2^{p-r}(p-r)!} \\ &= \frac{p!}{4^p} \sum_r \binom{2r}{r} \binom{2p-2r}{p-r} \end{aligned}$$

(3) In order to finish now the computation, let us recall that we have the following formula, coming from the generalized binomial formula, or from the Taylor formula:

$$\frac{1}{\sqrt{1+t}} = \sum_{q=0}^{\infty} \binom{2q}{q} \left(\frac{-t}{4}\right)^q$$

By taking the square of this series, we obtain the following formula:

$$\frac{1}{1+t}=\sum_p\left(\frac{-t}{4}\right)^p\sum_r\binom{2r}{r}\binom{2p-2r}{p-r}$$

Now by looking at the coefficient of t^p on both sides, we conclude that the sum on the right equals 4^p. Thus, we can finish the moment computation in (2), as follows:

$$M_k=\frac{p!}{4^p}\times 4^p=p!$$

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

As before with the real Gaussian laws, a better-looking statement is in terms of partitions. Given a colored integer $k=\circ\bullet\bullet\circ\ldots$, we say that a pairing $\pi\in P_2(k)$ is matching when it pairs $\circ-\bullet$ symbols. With this convention, we have the following result:

THEOREM 14.27. *The moments of the complex normal law are the numbers*

$$M_k(G_t)=\sum_{\pi\in\mathcal{P}_2(k)}t^{|\pi|}$$

where $\mathcal{P}_2(k)$ are the matching pairings of $\{1,\ldots,k\}$, and $|.|$ is the number of blocks.

PROOF. This is a reformulation of Theorem 14.26. Indeed, we can assume that we are in the case $t=1$, and here we know from Theorem 14.26 that the moments are:

$$M_k=\begin{cases}(|k|/2)! & (k\text{ uniform})\\ 0 & (k\text{ not uniform})\end{cases}$$

On the other hand, the numbers $|\mathcal{P}_2(k)|$ are given by exactly the same formula. Indeed, in order to have a matching pairing of k, our exponent $k=\circ\bullet\bullet\circ\ldots$ must be uniform, consisting of p copies of $\circ$ and p copies of $\bullet$, with $p=|k|/2$. But then the matching pairings of k correspond to the permutations of the $\bullet$ symbols, as to be matched with $\circ$ symbols, and so we have $p!$ such pairings. Thus, we have the same formula as for the moments of f, and we are led to the conclusion in the statement. □

In practice, we also need to know how to compute joint moments. We have here:

THEOREM 14.28 (Wick formula). *Given independent variables f_i, each following the complex normal law G_t, with $t>0$ being a fixed parameter, we have the formula*

$$\mathbb{E}\left(f_{i_1}^{k_1}\ldots f_{i_s}^{k_s}\right)=t^{s/2}\#\left\{\pi\in\mathcal{P}_2(k)\Big|\pi\leq\ker i\right\}$$

where $k=k_1\ldots k_s$ and $i=i_1\ldots i_s$, for the joint moments of these variables, where $\pi\leq\ker i$ means that the indices of i must fit into the blocks of π, in the obvious way.

PROOF. This is something well-known, which can be proved as follows:

(1) Let us first discuss the case where we have a single variable f, which amounts in taking $f_i = f$ for any i in the formula in the statement. What we have to compute here are the moments of f, with respect to colored integer exponents $k = \circ\bullet\bullet\circ\ldots$, and the formula in the statement tells us that these moments must be:

$$\mathbb{E}(f^k) = t^{|k|/2}|\mathcal{P}_2(k)|$$

But this is the formula in Theorem 14.27, so we are done with this case.

(2) In general now, when expanding the product $f_{i_1}^{k_1}\ldots f_{i_s}^{k_s}$ and rearranging the terms, we are left with doing a number of computations as in (1), and then making the product of the expectations that we found. But this amounts in counting the partitions in the statement, with the condition $\pi \leq \ker i$ there standing for the fact that we are doing the various type (1) computations independently, and then making the product. □

The above statement is one of the possible formulations of the Wick formula, and there are many more formulations, which are all useful. For instance, we have:

THEOREM 14.29 (Wick formula 2). *Given independent variables f_i, each following the complex normal law G_t, with $t > 0$ being a fixed parameter, we have the formula*

$$\mathbb{E}\left(f_{i_1}\ldots f_{i_k}f_{j_1}^*\ldots f_{j_k}^*\right) = t^k\#\left\{\pi \in S_k\middle| i_{\pi(r)} = j_r, \forall r\right\}$$

for the non-vanishing joint moments of these variables.

PROOF. This follows from the usual Wick formula, from Theorem 14.28. With some changes in the indices and notations, the formula there reads:

$$\mathbb{E}\left(f_{I_1}^{K_1}\ldots f_{I_s}^{K_s}\right) = t^{s/2}\#\left\{\sigma \in \mathcal{P}_2(K)\middle|\sigma \leq \ker I\right\}$$

Now observe that we have $\mathcal{P}_2(K) = \emptyset$, unless the colored integer $K = K_1\ldots K_s$ is uniform, in the sense that it contains the same number of $\circ$ and $\bullet$ symbols. Up to permutations, the non-trivial case, where the moment is non-vanishing, is the case where the colored integer $K = K_1\ldots K_s$ is of the following special form:

$$K = \underbrace{\circ\circ\ldots\circ}_{k}\ \underbrace{\bullet\bullet\ldots\bullet}_{k}$$

So, let us focus on this case, which is the non-trivial one. Here we have $s = 2k$, and we can write the multi-index $I = I_1\ldots I_s$ in the following way:

$$I = i_1\ldots i_k\ j_1\ldots j_k$$

With these changes made, the above usual Wick formula reads:

$$\mathbb{E}\left(f_{i_1}\ldots f_{i_k}f_{j_1}^*\ldots f_{j_k}^*\right) = t^k\#\left\{\sigma \in \mathcal{P}_2(K)\middle|\sigma \leq \ker(ij)\right\}$$

The point now is that the matching pairings $\sigma \in \mathcal{P}_2(K)$, with $K = \circ \ldots \circ \bullet \ldots \bullet$, of length $2k$, as above, correspond to the permutations $\pi \in S_k$, in the obvious way. With this identification made, the above modified usual Wick formula becomes:

$$\mathbb{E}\left(f_{i_1} \ldots f_{i_k} f_{j_1}^* \ldots f_{j_k}^*\right) = t^k \# \left\{\pi \in S_k \middle| i_{\pi(r)} = j_r, \forall r\right\}$$

Thus, we have reached to the formula in the statement, and we are done. □

Finally, here is one more formulation of the Wick formula, useful as well:

THEOREM 14.30 (Wick formula 3). *Given independent variables f_i, each following the complex normal law G_t, with $t > 0$ being a fixed parameter, we have the formula*

$$\mathbb{E}\left(f_{i_1} f_{j_1}^* \ldots f_{i_k} f_{j_k}^*\right) = t^k \# \left\{\pi \in S_k \middle| i_{\pi(r)} = j_r, \forall r\right\}$$

for the non-vanishing joint moments of these variables.

PROOF. This follows from our second Wick formula, from Theorem 14.29, simply by permuting the terms, as to have an alternating sequence of plain and conjugate variables. Alternatively, we can start with Theorem 14.28, and then perform the same manipulations as in the proof of Theorem 14.29, but with the exponent being this time as follows:

$$K = \underbrace{\circ \bullet \circ \bullet \ldots\ldots \circ \bullet}_{2k}$$

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

So long for the normal laws, real and complex. These appear pretty much everywhere in mathematics, and with the above, you are fully armed for dealing with them. In particular the Wick formula is what you need for Random Matrix Theory (RMT), shall you ever get interested in that, and with a good reference there being Mehta [68].

14c. Hyperspherical laws

We can do many things with the probability theory that we learned so far, and especially with the real and complex normal laws. As a first application, let us go back to the spherical integrals, studied in chapter 13. In the real case, we have:

THEOREM 14.31. *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. The moment formula in the statement follows from the general formulae in chapter 13. As a consequence, with $N \to \infty$ we have the following estimate:

$$\begin{aligned}\int_{S_{\mathbb{R}}^{N-1}} x_i^p dx &\simeq N^{-p/2} \times p!! \\ &= N^{-p/2} M_p(g_1)\end{aligned}$$

Thus, the rescaled variables $\sqrt{N}x_i$ become normal with $N \to \infty$, as claimed. As for the proof of the asymptotic independence, this is standard too, once again by using the formulae in chapter 13. 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. □

We can recover the normal laws as well in connection with the rotation groups. Indeed, we have the following reformulation of Theorem 14.31:

THEOREM 14.32. *We have the integration formula*

$$\int_{O_N} U_{ij}^p\, dU = \frac{(N-1)!!p!!}{(N+p-1)!!}$$

and the rescaled variables $V_{ij} = \sqrt{N}U_{ij}$ become normal and independent with $N \to \infty$.

PROOF. We use the basic fact that the rotations $U \in O_N$ act on the points of the real sphere $z \in S_{\mathbb{R}}^{N-1}$, with the stabilizer of $z = (1, 0, \ldots, 0)$ being the subgroup $O_{N-1} \subset O_N$. In algebraic terms, this gives an identification as follows:

$$S_{\mathbb{R}}^{N-1} = O_N / O_{N-1}$$

In functional analytic terms, this result provides us with an embedding as follows, for any i, which makes correspond the respective integration functionals:

$$C(S_{\mathbb{R}}^{N-1}) \subset C(O_N) \quad , \quad x_i \to U_{1i}$$

With this identification made, the result follows from Theorem 14.31. □

In the complex case, the analogues of the above results are as follows:

THEOREM 14.33. *The rescalings $\sqrt{N}z_i$ of the unit complex sphere coordinates*

$$z_i : S_{\mathbb{C}}^{N-1} \to \mathbb{C}$$

as well as the rescalings $\sqrt{N}U_{ij}$ of the unitary group coordinates

$$U_{ij} : U_N \to \mathbb{C}$$

become complex Gaussian and independent with $N \to \infty$.

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

(1) According to the formulae in chapter 13, the polynomials integrals in $z_i, \bar{z}_i$ vanish, unless the number of $z_i, \bar{z}_i$ is the same. In this latter case these terms can be grouped together, by using $z_i\bar{z}_i = |z_i|^2$, and the relevant integration formula is:

$$\int_{S_{\mathbb{C}}^{N-1}} |z_i|^{2k}\, dz = \frac{(N-1)!k!}{(N+k-1)!}$$

Now with $N \to \infty$, we obtain from this the following estimate:

$$\int_{S_{\mathbb{C}}^{N-1}} |z_i|^{2k} dx \simeq N^{-k} \times k!$$

Thus, the rescaled variables $\sqrt{N}z_i$ become normal with $N \to \infty$, as claimed.

(2) As for the proof of the asymptotic independence, this is standard too, again by using the formulae in chapter 13. Indeed, the joint moments of $z_1, \ldots, z_N$ are given by:

$$\begin{aligned}\int_{S_{\mathbb{R}}^{N-1}} |z_1|^{2k_1} \ldots |z_N|^{2k_N}\, dx &= \frac{(N-1)!k_1! \ldots k_n!}{(N+\sum k_i-1)!}\\ &\simeq N^{-\Sigma k_i} \times k_1! \ldots k_N!\end{aligned}$$

By rescaling, the joint moments of the variables $y_i = \sqrt{N}z_i$ are given by:

$$\int_{S_{\mathbb{R}}^{N-1}} |y_1|^{2k_1} \ldots |y_N|^{2k_N}\, dx \simeq k_1! \ldots k_N!$$

Thus, we have multiplicativity, and so independence with $N \to \infty$, as claimed.

(3) Regarding the last assertion, we can use the basic fact that the rotations $U \in U_N$ act on the points of the sphere $z \in S_{\mathbb{C}}^{N-1}$, with the stabilizer of $z = (1, 0, \ldots, 0)$ being the subgroup $U_{N-1} \subset U_N$. In algebraic terms, this gives an identification as follows:

$$S_{\mathbb{C}}^{N-1} = U_N/U_{N-1}$$

In functional analytic terms, this result provides us with an embedding as follows, for any i, which makes correspond the respective integration functionals:

$$C(S_{\mathbb{C}}^{N-1}) \subset C(U_N) \quad , \quad x_i \to U_{1i}$$

With this identification made, the result follows from (1,2). □

14d. Rotations, reflections

All the above is quite nice, and suggests that there are deep relations between group theory and probability. So, let us discuss this now, by starting with the discrete case, which is the simplest one. We have here the following beautiful, well-known result:

THEOREM 14.34. *The probability for a random $\sigma \in S_N$ to have no fixed points is*

$$P \simeq \frac{1}{e}$$

in the $N \to \infty$ limit, where $e = 2.718\ldots$ is the usual constant from analysis.

PROOF. This is best viewed by using the inclusion-exclusion principle. Let us set:

$$S_N^k = \left\{\sigma \in S_N \middle| \sigma(k) = k\right\}$$

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:

$$\begin{aligned}
|X_N| &= \left|\left(\bigcup_k S_N^k\right)^c\right| \\
&= |S_N| - \sum_k |S_N^k| + \sum_{k<l} |S_N^k \cap S_N^l| - \ldots + (-1)^N \sum_{k_1<\ldots<k_N} |S_N^{k_1} \cup \ldots \cup S_N^{k_N}| \\
&= N! - N(N-1)! + \binom{N}{2}(N-2)! - \ldots + (-1)^N \binom{N}{N}(N-N)! \\
&= \sum_{r=0}^{N} (-1)^r \binom{N}{r}(N-r)!
\end{aligned}$$

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:

$$P = \frac{|X_N|}{N!} = \sum_{r=0}^{N} \frac{(-1)^r}{r!}$$

Since on the right we have the expansion of $1/e$, this gives the result. □

The above result looks very exciting. In order to further explore and refine it, we will need some notions from group theory, and more specifically, the following definition:

DEFINITION 14.35. *Given a closed subgroup $G \subset U_N$, the function*

$$\chi : G \to \mathbb{C} \quad , \quad \chi(g) = \sum_{i=1}^{N} g_{ii}$$

is called main character of G.

We should mention that there is a long story with these characters, the idea being that a compact group G can have several representations $\pi : G \subset U_N$, which can be studied via their characters $\chi_\pi : G \to \mathbb{C}$, and with all this coming from Weyl [97], [98].

All this is quite heavy, but technically, we will not need all this here. Indeed, in relation with our questions, we can formulate the following nice, elementary result:

THEOREM 14.36. *Consider the symmetric group S_N, regarded as the permutation group, $S_N \subset O_N$, of the N coordinate axes of $\mathbb{R}^N$.*

(1) *The main character $\chi \in C(S_N)$ counts the number of fixed points.*
(2) *The law of $\chi \in C(S_N)$ becomes Poisson (1), in the $N \to \infty$ limit.*

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

(1) The permutation matrices $\sigma \in O_N$, which give the embedding $S_N \subset O_N$ in the statement, being given by $\sigma_{ij} = \delta_{i\sigma(j)}$, we have the following computation:

$$\chi(\sigma) = \sum_i \delta_{\sigma(i)i} = \#\left\{i \in \{1, \ldots, N\} \middle| \sigma(i) = i\right\}$$

(2) In order to establish now the asymptotic result in the statement, we must prove 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 14.34, 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:

$$\frac{\#\left\{\sigma\in S_N\Big|\chi(\sigma)=r\right\}}{N!}=\frac{1}{r!}\times\frac{\#\left\{\sigma\in S_{N-r}\Big|\chi(\sigma)=0\right\}}{(N-r)!}$$

Now by using the computation at $r=0$, that we already have, from Theorem 14.34, 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. □

As a next step, let us try now to generalize what we have, namely Theorem 14.36, as to reach to the Poisson laws of arbitrary parameter $t>0$. Again we will need some notions from group theory, and more specifically, the following definition:

DEFINITION 14.37. *Given a closed subgroup $G\subset U_N$, the function*

$$\chi_t:G\to\mathbb{C}\quad,\quad\chi_t(g)=\sum_{i=1}^{[tN]}g_{ii}$$

is called main truncated character of G, of parameter $t\in(0,1]$.

As before with plain characters, there is some theory behind this definition, involving this time Random Matrix Theory (RMT), but technically, we will not need this here. Indeed, we can now formulate the following result, which is nice and elementary, generalizing Theorem 14.36, and which will be our final saying on the subject:

THEOREM 14.38. *Consider the symmetric group S_N, regarded as the permutation group, $S_N\subset O_N$, of the N coordinate axes of $\mathbb{R}^N$.*

(1) *The variable χ_t counts the number of fixed points among $1,\ldots,[tN]$.*
(2) *The law of this variable χ_t becomes Poisson (t), in the $N\to\infty$ limit.*

PROOF. We already know from Theorem 14.36 that the results hold at $t=1$. In general, the proof is similar, the idea being as follows:

(1) We have indeed the following computation, coming from definitions:

$$\chi_t(\sigma)=\sum_{i=1}^{[tN]}\delta_{\sigma(i)i}=\#\left\{i\in\{1,\ldots,[tN]\}\Big|\sigma(i)=i\right\}$$

(2) Consider indeed the following sets, as in the proof of Theorem 14.34:

$$S_N^k=\left\{\sigma\in S_N\Big|\sigma(k)=k\right\}$$

The set of permutations having no fixed points among $1,\ldots,[tN]$ is then:

$$X_N=\left(\bigcup_{k\leq[tN]}S_N^k\right)^c$$

As before in the proof of Theorem 14.36, we obtain by inclusion-exclusion that:

$$\begin{aligned}P(\chi_t=0)&=\frac{1}{N!}\sum_{r=0}^{[tN]}(-1)^r\sum_{k_1<\ldots<k_r<[tN]}|S_N^{k_1}\cap\ldots\cap S_N^{k_r}|\\&=\frac{1}{N!}\sum_{r=0}^{[tN]}(-1)^r\binom{[tN]}{r}(N-r)!\\&=\sum_{r=0}^{[tN]}\frac{(-1)^r}{r!}\cdot\frac{[tN]!(N-r)!}{N!([tN]-r)!}\end{aligned}$$

Now with $N\to\infty$, we obtain from this the following estimate:

$$\begin{aligned}P(\chi_t=0)&\simeq\sum_{r=0}^{[tN]}\frac{(-1)^r}{r!}\cdot t^r\\&=\sum_{r=0}^{[tN]}\frac{(-t)^r}{r!}\\&\simeq e^{-t}\end{aligned}$$

More generally, by counting the permutations $\sigma\in S_N$ having exactly r fixed points among $1,\ldots,[tN]$, as in the proof of Theorem 14.36, 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. □

It is possible to establish as well continuous analogues of the above results, involving the groups O_N,U_N, and the real and complex normal laws. However, the proofs here are more technical. For an elementary introduction to this, you have my book [9].

Finally, let us end all this with some low-dimensional magics, involving physicists' favorite group, SU_2, and one of their favorite laws, the Wigner law γ_1. We have:

THEOREM 14.39. *The main character of SU_2 follows the following law,*

$$\gamma_1=\frac{1}{2\pi}\sqrt{4-x^2}dx$$

which is the Wigner law of parameter 1.

PROOF. The group SU_2 is by definition the group of unitary rotations $U \in U_2$ of determinant one, and by solving the equation $U^* = U^{-1}$, we are led to:

$$SU_2 = \left\{ \begin{pmatrix} a+ib & c+id \\ -c+id & a-ib \end{pmatrix} \;\middle|\; a^2+b^2+c^2+d^2=1 \right\}$$

In this picture, the main character is given by the following formula:

$$\chi \begin{pmatrix} a+ib & c+id \\ -c+id & a-ib \end{pmatrix} = 2a$$

We are therefore left with computing the law of the following variable:

$$a \in C(S^3_{\mathbb{R}})$$

But this is something very familiar, namely a hyperspherical variable at $N = 4$, so we can use here Theorem 14.31. We obtain the following moment formula:

$$\begin{aligned} \int_{S^3_{\mathbb{R}}} a^{2k} &= \frac{3!!(2k)!!}{(2k+3)!!} \\ &= 2 \cdot \frac{(2k)!}{2^k k! 2^{k+1}(k+1)!} \\ &= \frac{1}{4^k} \cdot \frac{1}{k+1} \binom{2k}{k} \\ &= \frac{C_k}{4^k} \end{aligned}$$

Thus the variable $2a \in C(S^3_{\mathbb{R}})$ follows the Wigner semicircle law γ_1, as claimed. □

There are many more things that can be said, in relation with the above, all beautiful mathematics, useful in relation with physics. If interested, have a look at my book [9].

14e. Exercises

Probability theory is a big business, and many other things can be said, in relation with the above, which are all useful, and good to know. As exercises, we have:

EXERCISE 14.40. *Learn about compound Poisson laws.*

EXERCISE 14.41. *Study more in detail the CLT, and the convergence there.*

EXERCISE 14.42. *Learn about the heat kernel, and solving the heat equation.*

EXERCISE 14.43. *Recover the Marchenko-Pastur law π_1, in relation with SO_3.*

As bonus exercise, learn some compact group theory, and more specifically the Peter-Weyl theory for such groups. This is quite tough, but is an excellent investment.

CHAPTER 15

Partial integration

15a. Vector products

Good news, done with mathematics, and in the remainder of this book, last 50 pages, we will get into physics. We already met some physics in this book, namely gravity, waves and heat in one dimension, in chapter 3, then waves and heat again, in two dimensions, along with some basic electrodynamics, light and early atomic theory, in chapter 8, and finally waves and heat again, this time in arbitrary N dimensions, in chapter 12.

Obviously, time to have some cleanup here, review what we know, and do more, in a more systematic way. We have divided what we have two say in two parts, as follows:

(1) In the present chapter we will discuss physics in low dimensions, $N = 2, 3, 4$, with on the menu more classical mechanics, notably with the Coriolis force, then some relativity following Einstein, and finally the basics of electrostatics, following Gauss and others, which are quite interesting, mathematically speaking. We will be mostly interested in $N = 3$, and the common feature of everything that we want to talk about will be the vector product $\times$, which exists only in 3 dimensions, and is something quite magic.

(2) In the next and final chapter we will discuss physics in infinite dimensions, $N = \infty$. Our goal here will be that of having some quantum mechanics theory started, along the lines suggested at the end of chapter 8, and more specifically, solving the hydrogen atom. There are actually several ways of proceeding here, following Heisenberg, Schrödinger and others, and matter of annoying my cat, who seems to be a big fan of measure theory and Hilbert spaces, we will opt here for the Schrödinger approach, which is elementary.

Getting now to what we want to do in this chapter, namely 3 dimensions, vector products $\times$, and all sorts of related physics, there is some important mathematical interest in doing this, because we will reach in this way to answers to the following question:

QUESTION 15.1. *What are the higher dimensional analogues of the formula*

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

that is, of the fundamental theorem of calculus?

And isn't this a fundamental question for us, mathematicians, because we have so far in our bag multivariable extensions of all the main theorems of one-variable calculus, except for this. So, definitely something to be solved, before the end of this book.

So, let us discuss this first. The fundamental theorem of calculus tells us that the integral of a function on an interval $[a,b]$ can be suitably recaptured from what happens at the boundary $\{a,b\}$ of this interval. Obviously, this is something quite magic, and thinking now at what we can expect in $N=2,3$ or more dimensions, that can only be quite complicated, involving curves, surfaces, solid bodies and so on, and with all this vaguely reminding all sorts of physics things, such as field lines for gravity, magnets and so on. In short, despite having no formal proof yet for all this, let us formulate:

ANSWER 15.2. *Partial integration in several dimensions most likely means physics, and we will probably only get some partial results here, at $N=2,3$.*

Of course, all this remains to be confirmed, but assuming that you trust me a bit, here we are now at the plan that we made before, for this chapter. That is, do physics in low dimensions, guided by the beauty of the world surrounding us, and once this physics done, record some mathematical corollaries too, in relation with Question 15.1.

Before getting started, however, as usual when struggling with pedagogical matters, and other delicate dillemas, let us ask the cat. But here, unfortunately, no surprise:

CAT 15.3. *Read Rudin.*

Thanks cat, and guess this sort of discussion, that we started in chapter 13, looks more and more cyclic. So, I'll just go my way, on your side have a good hunt, and by the way make sure to catch enough mice and birds, using your measure theory and differential forms techniques, because there is a bit of shortage of cat food today, sorry for that.

Getting started now, here is what we need:

DEFINITION 15.4. *The vector product of two vectors in $\mathbb{R}^3$ is given by*

$$x\times y=||x||\cdot||y||\cdot\sin\theta\cdot n$$

where $n\in\mathbb{R}^3$ with $n\perp x,y$ and $||n||=1$ is constructed using the right-hand rule:

$$\begin{array}{c}\uparrow_{x\times y}\\ \leftarrow_{x}\\ \swarrow_{y}\end{array}$$

Alternatively, in usual vertical linear algebra notation for all vectors,

$$\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\times\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}x_2y_3-x_3y_2\\x_3y_1-x_1y_3\\x_1y_2-x_2y_1\end{pmatrix}$$

the rule being that of computing 2×2 determinants, and adding a middle sign.

Obviously, this definition is something quite subtle, and also something very annoying, because you always need this, and always forget the formula. Here are my personal methods. With the first definition, what I always remember is that:

$$||x \times y|| \sim ||x||, ||y|| \quad , \quad x \times x = 0 \quad , \quad e_1 \times e_2 = e_3$$

So, here's how it works. We are looking for a vector $x \times y$ whose length is proportional to those of x, y. But the second formula tells us that the angle θ between x, y must be involved via $0 \to 0$, and so the factor can only be $\sin\theta$. And with this we are almost there, it's just a matter of choosing the orientation, and this comes from $e_1 \times e_2 = e_3$.

As with the second definition, that I like the most, what I remember here is simply:

$$\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} = ?$$

Indeed, when trying to compute this determinant, by developing over the first column, what you get as coefficients are the entries of $x \times y$. And with the good middle sign.

In practice now, in order to get familiar with the vector products, nothing better than doing some classical mechanics. We have here the following key result:

THEOREM 15.5. *In the gravitational 2-body problem, the angular momentum*

$$J = x \times p$$

with $p = mv$ being the usual momentum, is conserved.

PROOF. There are several things to be said here, the idea being as follows:

(1) First of all the usual momentum, $p = mv$, is not conserved, because the simplest solution is the circular motion, where the moment gets turned around. But this suggests precisely that, in order to fix the lack of conservation of the momentum p, what we have to do is to make a vector product with the position x. Leading to J, as above.

(2) Regarding now the proof, consider indeed a particle m moving under the gravitational force of a particle M, assumed, as usual, to be fixed at 0. By using the fact that for two proportional vectors, $p \sim q$, we have $p \times q = 0$, we obtain:

$$\begin{aligned} \dot{J} &= \dot{x} \times p + x \times \dot{p} \\ &= v \times mv + x \times ma \\ &= m(v \times v + x \times a) \\ &= m(0 + 0) \\ &= 0 \end{aligned}$$

Now since the derivative of J vanishes, this quantity is constant, as stated. □

While the above principle looks like something quite trivial, the mathematics behind it is quite interesting, and has several notable consequences, as follows:

THEOREM 15.6. *In the context of a 2-body problem, the following happen:*

(1) *The fact that the direction of J is fixed tells us that the trajectory of one body with respect to the other lies in a plane.*
(2) *The fact that the magnitude of J is fixed tells us that the Kepler 2 law holds, namely that we have same areas sweeped by Ox over the same times.*

PROOF. This follows indeed from Theorem 15.5, as follows:

(1) We have by definition $J = m(x \times v)$, and since a vector product is orthogonal on both the vectors it comes from, we deduce from this that we have:

$$J \perp x, v$$

But this can be written as follows, with $J^\perp$ standing for the plane orthogonal to J:

$$x, v \in J^\perp$$

Now since J is fixed by Theorem 15.5, we conclude that both x, v, and in particular the position x, and so the whole trajectory, lie in this fixed plane $J^\perp$, as claimed.

(2) Conversely now, forget about Theorem 15.5, and assume that the trajectory lies in a certain plane E. Thus $x \in E$, and by differentiating we have $v \in E$ too, and so $x, v \in E$. Thus $E = J^\perp$, and so $J = E^\perp$, so the direction of J is fixed, as claimed.

(3) Regarding now the last assertion, we already know from the various formulae from chapter 11, or rather from the exercises there, that Kepler 2 is more or less equivalent to $\dot{\theta} = \lambda/r^2$. However, the derivation of $\dot{\theta} = \lambda/r^2$ was something tricky, and what we want to prove now is that this appears as a simple consequence of $||J|| = \text{constant}$.

(4) In order to to so, let us compute J, according to its definition $J = x \times p$, but in polar coordinates, which will change everything. Since $p = m\dot{x}$, we have:

$$J = r \begin{pmatrix} \cos\theta \\ \sin\theta \\ 0 \end{pmatrix} \times m \begin{pmatrix} \dot{r}\cos\theta - r\sin\theta \cdot \dot{\theta} \\ \dot{r}\sin\theta + r\cos\theta \cdot \dot{\theta} \\ 0 \end{pmatrix}$$

Now recall from the definition of the vector product that we have:

$$\begin{pmatrix} a \\ b \\ 0 \end{pmatrix} \times \begin{pmatrix} c \\ d \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ ad - bc \end{pmatrix}$$

Thus J is a vector of the above form, with its last component being:

$$\begin{aligned} J_z &= rm\begin{vmatrix}\cos\theta & \dot{r}\cos\theta - r\sin\theta\cdot\dot{\theta}\\ \sin\theta & \dot{r}\sin\theta + r\cos\theta\cdot\dot{\theta}\end{vmatrix}\\ &= rm\cdot r(\cos^2\theta+\sin^2\theta)\dot{\theta}\\ &= r^2m\cdot\dot{\theta}\end{aligned}$$

(5) Now with the above formula in hand, our claim is that the magnitude $||J||$ is constant precisely when $\dot{\theta}=\lambda/r^2$, for some $\lambda\in\mathbb{R}$. Indeed, up to the obvious fact that the orientation of J is a binary parameter, who cannot just switch like that, let us just agree on this, knowing J is the same as knowing J_z, and is also the same as knowing $||J||$. Thus, our claim is proved, and this leads to the conclusion in the statement. □

As another basic application of the vector products, still staying with classical mechanics, we have all sorts of useful formulae regarding rotating frames. We first have:

THEOREM 15.7. *Assume that a 3D body rotates along an axis, with angular speed w. For a fixed point of the body, with position vector x, the usual 3D speed is*

$$v=\omega\times x$$

where $\omega=wn$, with n unit vector pointing North. When the point moves on the body

$$V=\dot{x}+\omega\times x$$

is its speed computed by an inertial observer O on the rotation axis.

PROOF. We have two assertions here, both requiring some 3D thinking, as follows:

(1) Assuming that the point is fixed, the magnitude of $\omega\times x$ is the good one, due to the following computation, with r being the distance from the point to the axis:

$$||\omega\times x||=w||x||\sin t=wr=||v||$$

As for the orientation of $\omega\times x$, this is the good one as well, because the North pole rule used above amounts in applying the right-hand rule for finding n, and so ω, and this right-hand rule was precisely the one used in defining the vector products $\times$.

(2) Next, when the point moves on the body, the inertial observer O can compute its speed by using a frame (u_1,u_2,u_3) which rotates with the body, as follows:

$$\begin{aligned} V &= \dot{x}_1u_1+\dot{x}_2u_2+\dot{x}_3u_3+x_1\dot{u}_1+x_2\dot{u}_2+x_3\dot{u}_3\\ &= \dot{x}+(x_1\cdot\omega\times u_1+x_2\cdot\omega\times u_2+x_3\cdot\omega\times u_3)\\ &= \dot{x}+w\times(x_1u_1+x_2u_2+x_3u_3)\\ &= \dot{x}+\omega\times x\end{aligned}$$

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

In what regards now the acceleration, the result, which is famous, is as follows:

THEOREM 15.8. *Assuming as before that a 3D body rotates along an axis, the acceleration of a moving point on the body, computed by O as before, is given by*

$$A = a + 2\omega \times v + \omega \times (\omega \times x)$$

with $\omega = wn$ being as before. In this formula the second term is called Coriolis acceleration, and the third term is called centripetal acceleration.

PROOF. This comes by using twice the formulae in Theorem 15.7, as follows:

$$\begin{aligned} A &= \dot{V} + \omega \times V \\ &= (\ddot{x} + \dot{\omega} \times x + \omega \times \dot{x}) + (\omega \times \dot{x} + \omega \times (\omega \times x)) \\ &= \ddot{x} + \omega \times \dot{x} + \omega \times \dot{x} + \omega \times (\omega \times x) \\ &= a + 2\omega \times v + \omega \times (\omega \times x) \end{aligned}$$

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

The truly famous result is actually the one regarding forces, obtained by multiplying everything by a mass m, and writing things the other way around, as follows:

$$ma = mA - 2m\omega \times v - m\omega \times (\omega \times x)$$

Here the second term is called Coriolis force, and the third term is called centrifugal force. These forces are both called apparent, or fictious, because they do not exist in the inertial frame, but they exist however in the non-inertial frame of reference, as explained above. And with of course the terms centrifugal and centripetal not to be messed up.

In fact, even more famous is the terrestrial application of all this, as follows:

THEOREM 15.9. *The acceleration of an object m subject to a force F is given by*

$$ma = F - mg - 2m\omega \times v - m\omega \times (\omega \times x)$$

with g pointing upwards, and with the last terms being the Coriolis and centrifugal forces.

PROOF. This follows indeed from the above discussion, by assuming that the acceleration A there comes from the combined effect of a force F, and of the usual g. □

We refer to any standard undergraduate mechanics book, such as Feynman [31], Kibble [53] or Taylor [89] for more on the above, including various numerics on what happens here on Earth, the Foucault pendulum, history of all this, and many other things. Let us just mention here, as a basic illustration for all this, that a rock dropped from 100m deviates about 1cm from its intended target, due to the formula in Theorem 15.9.

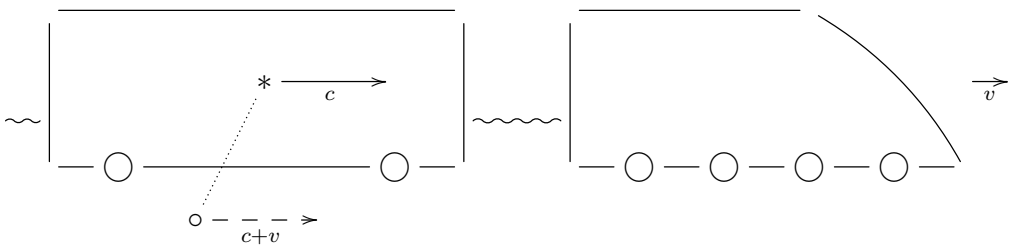
15b. Einstein addition

As another application of the vector products, let us discuss now the speed addition in relativity. Based on experiments by Fizeau, then Michelson-Morley and others, and some physics by Maxwell and Lorentz, Einstein came upon the following principles:

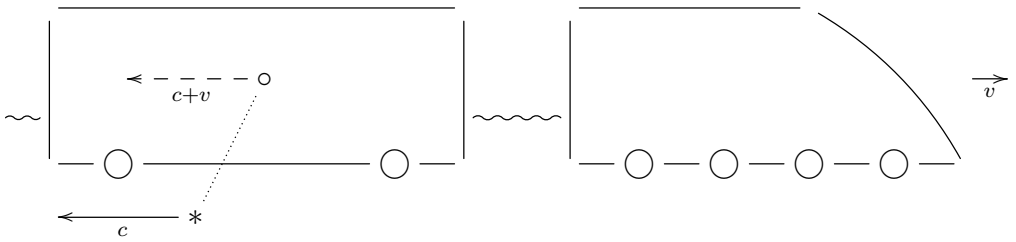
FACT 15.10 (Einstein principles). *The following happen:*

(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 travelling 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 15.10, while physically true, implies $c + v = c$, so contradicts classical mechanics, which needs a fix. In the classical case, to start with, we have:

PROPOSITION 15.11. *The classical speeds add according to the Galileo formula*

$$v_{AC} = v_{AB} + v_{BC}$$

where v_{AB} denotes the relative speed of A with respect to B.

PROOF. This is clear indeed from the definition of speed, and very intuitive. □

In order to find the fix, we will first discuss the 1D case, and leave the 3D case, which is a bit more complicated, for later. We will use two tricks. First, let us forget about absolute speeds, with respect to a given frame, and talk about relative speeds only. In this case we are allowed to sum only quantities of type v_{AB}, v_{BC}, and we denote by $v_{AB} +_g v_{BC}$ the corresponding sum v_{AC}. With this convention, the Galileo formula becomes:

$$u +_g v = u + v$$

As a second trick now, observe that this Galileo formula holds in any system of units. In order now to deal with our problems, basically involving high speeds, it is convenient to change the system of units, as to have $c = 1$. With this convention our $c + v = c$ problem becomes $1 + v = 1$, and the solution to it is quite obvious, as follows:

THEOREM 15.12. *If we define the Einstein sum $+_e$ of relative speeds by*

$$u +_e v = \frac{u + v}{1 + uv}$$

in $c = 1$ units, then we have the formula $1 +_e v = 1$, valid for any v.

PROOF. This is obvious indeed from our definition of $+_e$, because if we plug in $u = 1$ in the above formula, we obtain as result:

$$1 +_e v = \frac{1 + v}{1 + v} = 1$$

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

Summarizing, we have solved our problem. In order now to formulate a final result, we must do some reverse engineering, by waiving the above two tricks. First, by getting back to usual units, $v \to v/c$, our new addition formula becomes:

$$\frac{u}{c} +_e \frac{v}{c} = \frac{\frac{u}{c} + \frac{v}{c}}{1 + \frac{u}{c} \cdot \frac{v}{c}}$$

By multiplying by c, we can write this formula in a better way, as follows:

$$u +_e v = \frac{u + v}{1 + uv/c^2}$$

In order now to finish, it remains to get back to absolute speeds, as in Proposition 15.11. And by doing so, we are led to the following result:

THEOREM 15.13. *If we sum the speeds according to the Einstein formula*

$$v_{AC} = \frac{v_{AB} + v_{BC}}{1 + v_{AB}v_{BC}/c^2}$$

then the Galileo formula still holds, approximately, for low speeds

$$v_{AC} \simeq v_{AB} + v_{BC}$$

and if we have $v_{AB} = c$ or $v_{BC} = c$, the resulting sum is $v_{AC} = c$.

PROOF. We have two assertions here, which are both clear, as follows:

(1) Regarding the first assertion, if we are at low speeds, $v_{AB}, v_{BC} << c$, the correction term $v_{AB}v_{BC}/c^2$ dissapears, and we are left with the Galileo formula, as claimed.

(2) As for the second assertion, this follows from the above discussion. □

The Einstein summation formula, while looking very simple, is in fact quite subtle, and must be handled with care. Indeed, getting back to $c = 1$ conventions, we have:

PROPOSITION 15.14. *The Einstein speed summation, written in $c = 1$ units as*

$$u +_e v = \frac{u + v}{1 + uv}$$

has the following properties:

(1) $u, v < 1$ *implies* $u + v < 1$.
(2) $\lambda u +_e \lambda v = \lambda(u + v)$ *fails.*
(3) $(u +_e v) +_e w = u +_e (v +_e w)$ *fails too.*

PROOF. All these assertions are elementary, as follows:

(1) This follows from the following formula, valid for any speeds u, v:

$$1 - u +_e v = 1 - \frac{u + v}{1 + uv} = \frac{(1 - u)(1 - v)}{1 + uv}$$

(2) This is clear, with the remark however that it works at $\lambda = -1, 0, 1$.

(3) This follows indeed from the following computation:

$$\begin{aligned}(u +_e v) +_e w &= \frac{u + v + w + uv}{1 + uv + uw + vw}\\ &\neq \frac{u + v + w + vw}{1 + uv + uw + vw}\\ &= u +_e (v +_e w)\end{aligned}$$

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

All the above is very nice, but remember, takes place in 1D. So, time now to get seriously to work, and see what all this becomes in 3D. Expect of course a lot a vector calculus, as usual in relation with 3D problems, and in the hope that you love that. As a main goal, we must review the Einstein speed summation formula:

QUESTION 15.15. *What is the correct analogue of the Einstein summation formula*

$$u +_e v = \frac{u + v}{1 + uv}$$

in 2 and 3 dimensions?

In order to discuss this question, let us attempt to construct $u +_e v$ in arbitrary dimensions, just by using our common sense and intuition. When the vectors $u, v \in \mathbb{R}^N$ are proportional, we are basically in 1D, and so our addition formula must satisfy:

$$u \sim v \implies u +_e v = \frac{u + v}{1 + < u, v >}$$

However, the formula on the right will not work as such in general, for arbitrary speeds $u, v \in \mathbb{R}^N$, and this because we have, as main requirement for our operation, in analogy with the $1 +_e v = 1$ formula from 1D, the following condition:

$$||u|| = 1 \implies u +_e v = u$$

Equivalently, in analogy with $u +_e 1 = 1$ from 1D, we would like to have:

$$||v|| = 1 \implies u +_e v = v$$

Summarizing, our $u \sim v$ formula above is not bad, as a start, but we must add a correction term to it, for the above requirements to be satisfied, and of course with the correction term vanishing when $u \sim v$. So, we are led to a math puzzle:

PUZZLE 15.16. *What vanishes when $u \sim v$, and then how to correctly define*

$$u +_e v = \frac{u + v + \gamma_{uv}}{1 + < u, v >}$$

as for the correction term γ_{uv} to vanish when $u \sim v$?

But the solution to the first question is well-known in 3D. Indeed, here we can use the vector product $u \times v$, that we met before, which notoriously satisfies:

$$u \sim v \implies u \times v = 0$$

Thus, our correction term γ_{uv} must be something containing $w = u \times v$, which vanishes when this vector w vanishes, and in addition arranged such that $||u|| = 1$ produces a simplification, with $u +_e v = u$ as end result, and with $||v|| = 1$ producing a simplification too, with $u +_e v = v$ as end result. Thus, our vector calculus puzzle becomes:

PUZZLE 15.17. *How to correctly define the Einstein summation in 3 dimensions,*

$$u +_e v = \frac{u + v + \gamma_{uvw}}{1 + < u, v >}$$

with $w = u \times v$, in such a way as for the correction term γ_{uvw} to satisfy

$$w = 0 \implies \gamma_{uvw} = 0$$

and also such that $||u|| = 1 \implies u +_e v = u$, and $||v|| \implies u +_e v = v$?

In order to solve this latter puzzle, the first observation is that $\gamma_{uvw} = w$ will not do, and this for several reasons. First, this vector points in the wrong direction, orthogonal to the plane spanned by u, v, and we certainly don't want to leave this plane, with our correction. Also, as a technical remark to be put on top of this, the choice $\gamma_{uvw} = w$ will not bring any simplifications, as required above, in the cases $||u|| = 1$ or $||v|| = 1$. Thus, certainly wrong choice, and we must invent something more complicated.

Moving ahead now, as obvious task, we must "transport" the vector w to the plane spanned by u, v. But this is simplest done by taking the vector product with any vector in this plane, and so as a reasonable candidate for our correction term, we have:

$$\gamma_{uvw} = (\alpha u + \beta v) \times w$$

Here $\alpha, \beta \in \mathbb{R}$ are some scalars to be determined, but let us take a break, and leave the computations for later. We did some good work, time to update our puzzle:

PUZZLE 15.18. *How to define the Einstein summation in* 3 *dimensions,*

$$u +_e v = \frac{u + v + \gamma_{uvw}}{1 + < u, v >}$$

with the correction term being of the following form, with $w = u \times v$*, and* $\alpha, \beta \in \mathbb{R}$*,*

$$\gamma_{uvw} = (\alpha u + \beta v) \times w$$

in such a way as to have $||u|| = 1 \implies u +_e v = u$*, and* $||v|| \implies u +_e v = v$*?*

In order to investigate what happens when $||u|| = 1$ or $||v|| = 1$, we must compute the vector products $u \times w$ and $v \times w$. So, pausing now our study for consulting the vector calculus database, and then coming back, here is the formula that we need:

$$u \times (u \times v) = < u, v > u - < u, u > v$$

As for the formula of $v \times w$, that I forgot to record, we can recover it from the one above of $u \times w$, by using the basic properties of the vector products, as follows:

$$\begin{aligned} v \times (u \times v) &= -v \times (v \times u) \\ &= -(< v, u > v - < v, v > u) \\ &= < v, v > u - < u, v > v \end{aligned}$$

With these formulae in hand, we can now compute the correction term, with the result here, that we will need several times in what comes next, being as follows:

PROPOSITION 15.19. *The correction term* $\gamma_{uvw} = (\alpha u + \beta v) \times w$ *is given by*

$$\gamma_{uvw} = (\alpha < u, v > + \beta < v, v >)u - (\alpha < u, u > + \beta < u, v >)v$$

for any values of the scalars $\alpha, \beta \in \mathbb{R}$*.*

PROOF. According to our vector product formulae above, we have:

$$\begin{aligned}\gamma_{uvw} &= (\alpha u+\beta v)\times w\\ &= \alpha(<u,v>u-<u,u>v)+\beta(<v,v>u-<u,v>v)\\ &= (\alpha<u,v>+\beta<v,v>)u-(\alpha<u,u>+\beta<u,v>)v\end{aligned}$$

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

Time now to get into the real thing, see what happens when $||u||=1$ and $||v||=1$, if we can get indeed $u+_e v=u$ and $u+_e v=v$. It is convenient here to do some reverse engineering. Regarding the first desired formula, namely $u+_e v=u$, we have:

$$\begin{aligned}u+_e v=u &\iff u+v+\gamma_{uvw}=(1+<u,v>)u\\ &\iff \gamma_{uvw}=<u,v>u-v\\ &\iff \alpha=1,\ \beta=0,\ ||u||=1\end{aligned}$$

Thus, with the parameter choice $\alpha=1,\beta=0$, we will have, as desired:

$$||u||=1\implies u+_e v=u$$

In what regards now the second desired formula, namely $u+_e v=v$, here the computation is almost identical, save for a sign switch, which after some thinking comes from our choice $w=u\times v$ instead of $w=v\times u$, clearly favoring u, as follows:

$$\begin{aligned}u+_e v=v &\iff u+v+\gamma_{uvw}=(1+<u,v>)v\\ &\iff \gamma_{uvw}=-u+<u,v>v\\ &\iff \alpha=0,\ \beta=-1,\ ||v||=1\end{aligned}$$

Thus, with the parameter choice $\alpha=0,\beta=-1$, we will have, as desired:

$$||v||=1\implies u+_e v=v$$

All this is mixed news, because we managed to solve both our problems, at $||u||=1$ and at $||v||=1$, but our solutions are different. So, time to breathe, decide that we did enough interesting work for the day, and formulate our conclusion as follows:

PROPOSITION 15.20. *When defining the Einstein speed summation in* 3D *as*

$$u+_e v=\frac{u+v+u\times(u\times v)}{1+<u,v>}$$

in $c=1$ *units, the following happen:*

(1) *When* $u\sim v$*, we recover the previous* 1D *formula.*
(2) *When* $||u||=1$*, speed of light, we have* $u+_e v=u$*.*
(3) *However,* $||v||=1$ *does not imply* $u+_e v=v$*.*
(4) *Also, the formula* $u+_e v=v+_e u$ *fails.*

PROOF. Here (1) and (2) follow from the above discussion, with the following choice for the correction term, by favoring the $||u|| = 1$ problem over the $||v|| = 1$ one:

$$\gamma_{uvw} = u \times w$$

In fact, with this choice made, the computation is very simple, as follows:

$$\begin{aligned} ||u|| = 1 &\implies \gamma_{uvw} =< u, v > u - v \\ &\implies u + v + \gamma_{uvw} = u + < u, v > u \\ &\implies \frac{u + v + \gamma_{uvw}}{1 + < u, v >} = u \end{aligned}$$

As for (3) and (4), these are also clear from the above discussion, coming from the obvious lack of symmetry of our summation formula. □

Looking now at Proposition 15.20 from an abstract, mathematical perspective, there are still many things missing from there, which can be summarized as follows:

QUESTION 15.21. *Can we fine-tune the Einstein speed summation in* 3D *into*

$$u +_e v = \frac{u + v + \lambda \cdot u \times (u \times v)}{1 + < u, v >}$$

with $\lambda \in \mathbb{R}$*, chosen such that* $||u|| = 1 \implies \lambda = 1$*, as to have:*

(1) $||u||, ||v|| < 1 \implies ||u +_e v|| < 1$.
(2) $||v|| = 1 \implies ||u +_e v|| = 1$.

All this is quite tricky, and deserves some explanations. First, if we add a scalar $\lambda \in \mathbb{R}$ into our formula, as above, we will still have, exactly as before:

$$u \sim v \implies u +_e v = \frac{1 + uv}{1 + < u, v >}$$

On the other hand, we already know from our previous computations, those preceding Proposition 15.20, that if we ask for $\lambda \in \mathbb{R}$ to be a plain constant, not depending on u, v, then $\lambda = 1$ is the only good choice, making the following formula happen:

$$||u|| = 1 \implies u +_e v = u$$

But, and here comes our point, $\lambda = 1$ is not an ideal choice either, because it would be nice to have the properties (1,2) in the statement, and these properties have no reason to be valid for $\lambda = 1$, as you can check for instance by yourself by doing some computations. Thus, the solution to our problem most likely involves a scalar $\lambda \in \mathbb{R}$ depending on u, v, and satisfying the following condition, as to still have $||u|| = 1 \implies u +_e v = u$:

$$||u|| = 1 \implies \lambda = 1$$

Obviously, as simplest answer, λ must be some well-chosen function of $||u||$, or rather of $||u||^2$, because it is always better to use square norms, when possible. But then, with this idea in mind, after a few computations we are led to the following solution:

$$\lambda = \frac{1}{1+\sqrt{1-||u||^2}}$$

Summarizing, final correction done, and with this being the end of mathematics, we did a nice job, and we can now formulate our findings as a theorem, as follows:

THEOREM 15.22. *When defining the Einstein speed summation in* 3D *as*

$$u +_e v = \frac{1}{1+<u,v>}\left(u+v+\frac{u\times(u\times v)}{1+\sqrt{1-||u||^2}}\right)$$

in $c=1$ *units, the following happen:*

(1) *When* $u\sim v$*, we recover the previous* 1D *formula.*
(2) *We have* $||u||,||v||<1 \implies ||u+_e v||<1$*.*
(3) *When* $||u||=1$*, we have* $u+_e v=u$*.*
(4) *When* $||v||=1$*, we have* $||u+_e v||=1$*.*
(5) *However,* $||v||=1$ *does not imply* $u+_e v=v$*.*
(6) *Also, the formula* $u+_e v=v+_e u$ *fails.*

PROOF. This follows from the above discussion, as follows:

(1) This is something that we know from Proposition 15.20.

(2) In order to simplify notation, let us set $\delta=\sqrt{1-||u||^2}$, which is the inverse of the quantity $\gamma=1/\sqrt{1-||u||^2}$. With this convention, we have:

$$\begin{aligned} u+_e v &= \frac{1}{1+<u,v>}\left(u+v+\frac{<u,v>u-||u||^2v}{1+\delta}\right)\\ &= \frac{(1+\delta+<u,v>)u+(1+\delta-||u||^2)v}{(1+<u,v>)(1+\delta)}\end{aligned}$$

Taking now the squared norm and computing gives the following formula:

$$||u+_e v||^2=\frac{(1+\delta)^2||u+v||^2+(||u||^2-2(1+\delta))(||u||^2||v||^2-<u,v>^2)}{(1+<u,v>)^2(1+\delta)^2}$$

But this formula can be further processed by using $\delta=\sqrt{1-||u||^2}$, and by navigating through the various quantities which appear, we obtain, as a final product:

$$||u+_e v||^2=\frac{||u+v||^2-||u||^2||v||^2+<u,v>^2}{(1+<u,v>)^2}$$

But this type of formula is exactly what we need, for what we want to do. Indeed, by assuming $||u||, ||v|| < 1$, we have the following estimate:

$$\begin{aligned} ||u +_e v||^2 < 1 &\iff ||u+v||^2 - ||u||^2||v||^2 + < u, v >^2 < (1 + < u, v >)^2 \\ &\iff ||u+v||^2 - ||u||^2||v||^2 < 1 + 2 < u, v > \\ &\iff ||u||^2 + ||v||^2 - ||u||^2||v||^2 < 1 \\ &\iff (1 - ||u||^2)(1 - ||v||^2) > 0 \end{aligned}$$

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

(3) This is something that we know from Proposition 15.20.

(4) This comes from the squared norm formula established in the proof of (2) above, because when assuming $||v|| = 1$, we obtain:

$$\begin{aligned} ||u +_e v||^2 &= \frac{||u+v||^2 - ||u||^2 + < u, v >^2}{(1 + < u, v >)^2} \\ &= \frac{||u||^2 + 1 + 2 < u, v > - ||u||^2 + < u, v >^2}{(1 + < u, v >)^2} \\ &= \frac{1 + 2 < u, v > + < u, v >^2}{(1 + < u, v >)^2} \\ &= 1 \end{aligned}$$

(5) This is clear, from the obvious lack of symmetry of our formula.

(6) This is again clear, from the obvious lack of symmetry of our formula. □

That was nice, all this mathematics, and hope you're still with me. And good news, the formula in Theorem 15.22 is the good one, confirmed by experimental physics.

15c. Charges and flux

Time for electricity. In analogy with the usual study of gravity, let us start with:

DEFINITION 15.23. *Given charges $q_1, \ldots, q_k \in \mathbb{R}$ located at positions $x_1, \ldots, x_k \in \mathbb{R}^3$, we define their electric field to be the vector function*

$$E(x) = K \sum_i \frac{q_i(x - x_i)}{||x - x_i||^3}$$

so that their force applied to a charge $Q \in \mathbb{R}$ positioned at $x \in \mathbb{R}^3$ is given by $F = QE$.

More generally, we will be interested in electric fields of various non-discrete configurations of charges, such as charged curves, surfaces and solid bodies. Indeed, things like wires or metal sheets or solid bodies coming in all sorts of shapes, tailored for their purpose, play a key role, so this extension is essential. So, let us go ahead with:

DEFINITION 15.24. *The electric field of a charge configuration $L \subset \mathbb{R}^3$, with charge density function $\rho : L \to \mathbb{R}$, is the vector function*

$$E(x) = K \int_L \frac{\rho(z)(x-z)}{||x-z||^3}\, dz$$

so that the force of L applied to a charge Q positioned at x is given by $F = QE$.

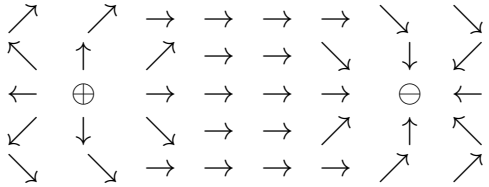
With the above definitions in hand, it is most convenient now to forget about the charges, and focus on the study of the corresponding electric fields E.

These fields are by definition vector functions $E : \mathbb{R}^3 \to \mathbb{R}^3$, with the convention that they take $\pm\infty$ values at the places where the charges are located, and intuitively, are best represented by their field lines, which are constructed as follows:

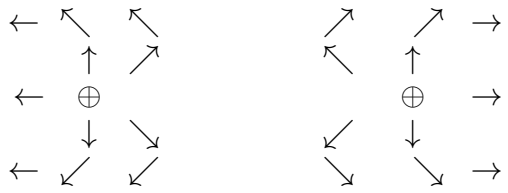
DEFINITION 15.25. *The field lines of an electric field $E : \mathbb{R}^3 \to \mathbb{R}^3$ are the oriented curves $\gamma \subset \mathbb{R}^3$ pointing at every point $x \in \mathbb{R}^3$ at the direction of the field, $E(x) \in \mathbb{R}^3$.*

As a basic example here, for one charge the field lines are the half-lines emanating from its position, oriented according to the sign of the charge:

For two charges now, if these are of opposite signs, $+$ and $-$, you get a picture that you are very familiar with, namely that of the field lines of a bar magnet:



If the charges are $+,+$ or $-,-$, you get something of similar type, but repulsive this time, with the field lines emanating from the charges being no longer shared:



These pictures, and notably the last one, with $+,+$ charges, are quite interesting, because the repulsion situation does not appear in the context of gravity. Thus, we can only expect our geometry here to be far more complicated than that of gravity.

The field lines, as constructed in Definition 15.25, obviously do not encapsulate the whole information about the field, with the direction of each vector $E(x) \in \mathbb{R}^3$ being there, but with the magnitude $||E(x)|| \geq 0$ of this vector missing. However, say when drawing, when picking up uniformly radially spaced field lines around each charge, and with the number of these lines proportional to the magnitude of the charge, and then completing the picture, the density of the field lines around each point $x \in \mathbb{R}$ will give you then the magnitude $||E(x)|| \geq 0$ of the field there, up to a scalar.

Let us summarize these observations as follows:

PROPOSITION 15.26. *Given an electric field $E : \mathbb{R}^3 \to \mathbb{R}^3$, the knowledge of its field lines is the same as the knowledge of the composition*

$$nE : \mathbb{R}^3 \to \mathbb{R}^3 \to S$$

where $S \subset \mathbb{R}^3$ is the unit sphere, and $n : \mathbb{R}^3 \to S$ is the rescaling map, namely:

$$n(x) = \frac{x}{||x||}$$

However, in practice, when the field lines are accurately drawn, the density of the field lines gives you the magnitude of the field, up to a scalar.

PROOF. We have two assertions here, the idea being as follows:

(1) The first assertion is clear from definitions, with of course our usual convention that the electric field and its problematics take place outside the locations of the charges, which makes everything in the statement to be indeed well-defined.

(2) Regarding now the last assertion, which is of course a bit informal, this follows from the above discussion. It is possible to be a bit more mathematical here, with a definition, formula and everything, but we will not need this, in what follows. □

Let us introduce now a key definition, as follows:

DEFINITION 15.27. *The flux of an electric field $E : \mathbb{R}^3 \to \mathbb{R}^3$ through a surface $S \subset \mathbb{R}^3$, assumed to be oriented, is the quantity*

$$\Phi_E(S) = \int_S < E(x), n(x) > dx$$

with $n(x)$ being unit vectors orthogonal to S, following the orientation of S. Intuitively, the flux measures the signed number of field lines crossing S.

Here by orientation of S we mean precisely the choice of unit vectors $n(x)$ as above, orthogonal to S, which must vary continuously with x. For instance a sphere has two possible orientations, one with all these vectors $n(x)$ pointing inside, and one with all these vectors $n(x)$ pointing outside. More generally, any surface has locally two possible

orientations, so if it is connected, it has two possible orientations. In what follows the convention is that the closed surfaces are oriented with each $n(x)$ pointing outside.

As a first illustration, let us do a basic computation, as follows:

PROPOSITION 15.28. *For a point charge $q \in \mathbb{R}$ at the center of a sphere S,*

$$\Phi_E(S) = \frac{q}{\varepsilon_0}$$

where the constant is $\varepsilon_0 = 1/(4\pi K)$, independently of the radius of S.

PROOF. Assuming that S has radius r, we have the following computation:

$$\begin{aligned}
\Phi_E(S) &= \int_S < E(x), n(x) > dx \\
&= \int_S \left\langle \frac{Kqx}{r^3}, \frac{x}{r} \right\rangle dx \\
&= \int_S \frac{Kq}{r^2}\, dx \\
&= \frac{Kq}{r^2} \times 4\pi r^2 \\
&= 4\pi Kq
\end{aligned}$$

Thus with $\varepsilon_0 = 1/(4\pi K)$ as above, we obtain the result. □

More generally now, we have the following result:

THEOREM 15.29. *The flux of a field E through a sphere S is given by*

$$\Phi_E(S) = \frac{Q_{enc}}{\varepsilon_0}$$

where Q_{enc} is the total charge enclosed by S, and $\varepsilon_0 = 1/(4\pi K)$.

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

(1) Before jumping into computations, let us do some manipulations. First, by discretizing the problem, we can assume that we are dealing with a system of point charges. Moreover, by additivity, we can assume that we are dealing with a single charge. And if we denote by $q \in \mathbb{R}$ this charge, located at $v \in \mathbb{R}^3$, we want to prove that we have the following formula, where $B \subset \mathbb{R}^3$ denotes the ball enclosed by S:

$$\Phi_E(S) = \frac{q}{\varepsilon_0}\, \delta_{v \in B}$$

(2) By linearity we can assume that we are dealing with the unit sphere S. Moreover, by rotating we can assume that our charge q lies on the Ox axis, that is, that we have

$v = (r, 0, 0)$ with $r \geq 0$, $r \neq 1$. The formula that we want to prove becomes:

$$\Phi_E(S) = \frac{q}{\varepsilon_0}\,\delta_{r<1}$$

(3) Let us start now the computation. With $u = (x, y, z)$, we have:

$$\begin{aligned}
\Phi_E(S) &= \int_S < E(u), u > du \\
&= \int_S \left\langle \frac{Kq(u-v)}{||u-v||^3}, u \right\rangle du \\
&= Kq \int_S \frac{< u-v, u >}{||u-v||^3}\, du \\
&= Kq \int_S \frac{1 - < v, u >}{||u-v||^3}\, du \\
&= Kq \int_S \frac{1 - rx}{(1 - 2xr + r^2)^{3/2}}\, du
\end{aligned}$$

(4) In order to compute the above integral, we will use spherical coordinates for the unit sphere S, which are as follows, with $s \in [0, \pi]$ and $t \in [0, 2\pi]$:

$$\begin{cases}
x &= \cos s \\
y &= \sin s \cos t \\
z &= \sin s \sin t
\end{cases}$$

We recall that the corresponding Jacobian, computed before, is given by:

$$J = \sin s$$

(5) With the above change of coordinates, our integral from (3) becomes:

$$\begin{aligned}
\Phi_E(S) &= Kq \int_S \frac{1 - rx}{(1 - 2xr + r^2)^{3/2}}\, du \\
&= Kq \int_0^{2\pi} \int_0^{\pi} \frac{1 - r\cos s}{(1 - 2r\cos s + r^2)^{3/2}} \cdot \sin s\, ds\, dt \\
&= 2\pi Kq \int_0^{\pi} \frac{(1 - r\cos s)\sin s}{(1 - 2r\cos s + r^2)^{3/2}}\, ds \\
&= \frac{q}{2\varepsilon_0} \int_0^{\pi} \frac{(1 - r\cos s)\sin s}{(1 - 2r\cos s + r^2)^{3/2}}\, ds
\end{aligned}$$

(6) The point now is that the integral on the right can be computed with the change of variables $x = \cos s$. Indeed, we have $dx = -\sin s\, ds$, and we obtain:

$$\begin{aligned}
\int_0^\pi \frac{(1 - r\cos s)\sin s}{(1 - 2r\cos s + r^2)^{3/2}}\, ds &= \int_{-1}^1 \frac{1 - rx}{(1 - 2rx + r^2)^{3/2}}\, dx \\
&= \left[\frac{x - r}{\sqrt{1 - 2rx + r^2}}\right]_{-1}^1 \\
&= \frac{1 - r}{\sqrt{1 - 2r + r^2}} - \frac{-1 - r}{\sqrt{1 + 2r + r^2}} \\
&= \frac{1 - r}{|1 - r|} + 1 \\
&= 2\delta_{r<1}
\end{aligned}$$

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

More generally now, we have the following key result, due to Gauss:

THEOREM 15.30 (Gauss law). *The flux of a field E through a surface S is given by*

$$\Phi_E(S) = \frac{Q_{enc}}{\varepsilon_0}$$

where Q_{enc} is the total charge enclosed by S, and $\varepsilon_0 = 1/(4\pi K)$.

PROOF. This basically follows from Theorem 15.29, or even from Proposition 15.28, by adding to the results there a number of new ingredients, as follows:

(1) Our first claim is that given a closed surface S, with no charges inside, the flux through it of any choice of external charges vanishes:

$$\Phi_E(S) = 0$$

This claim is indeed supported by the intuitive interpretation of the flux, as corresponding to the signed number of field lines crossing S. Indeed, any field line entering as $+$ must exit somewhere as $-$, and vice versa, so when summing we get 0.

(2) In practice now, in order to prove this rigorously, there are several ways. A standard argument, which is quite elementary, is the one used by Feynman in [32], based on the fact that, due to $F \sim 1/d^2$, local deformations of S will leave invariant the flux, and so in the end we are left with a rotationally invariant surface, where the result is clear.

(3) The point now is that, with this and Proposition 15.28 in hand, we can finish by using a standard math trick. Let us assume indeed, by discretizing, that our system of charges is discrete, consisting of enclosed charges $q_1, \ldots, q_k \in \mathbb{R}$, and an exterior total

charge Q_{ext}. We can surround each of $q_1, \ldots, q_k$ by small disjoint spheres $U_1, \ldots, U_k$, chosen such that their interiors do not touch S, and we have:

$$\begin{aligned}\Phi_E(S) &= \Phi_E(S - \cup U_i) + \Phi_E(\cup U_i)\\ &= 0 + \Phi_E(\cup U_i)\\ &= \sum_i \Phi_E(U_i)\\ &= \sum_i \frac{q_i}{\varepsilon_0}\\ &= \frac{Q_{enc}}{\varepsilon_0}\end{aligned}$$

(4) To be more precise, in the above the union $\cup U_i$ is a usual disjoint union, and the flux is of course additive over components. As for the difference $S - \cup U_i$, this is by definition the disjoint union of S with the disjoint union $\cup(-U_i)$, with each $-U_i$ standing for U_i with orientation reversed, and since this difference has no enclosed charges, the flux through it vanishes by (2). Finally, the end makes use of Proposition 15.28. □

So long for the Gauss law. We will be back to it in a moment, with a new, better proof, by using some advanced 3D calculus, that we will have to learn first.

15d. Gauss, Green, Stokes

Let us start with a standard definition, immersing us into 3D problematics, as follows:

DEFINITION 15.31. *Given a function $f : \mathbb{R}^3 \to \mathbb{R}$, its usual derivative $f'(u) \in \mathbb{R}^3$ can be written as $f'(u) = \nabla f(u)$, where the gradient operator ∇ is given by:*

$$\nabla = \begin{pmatrix} \frac{d}{dx} \\ \frac{d}{dy} \\ \frac{d}{dz} \end{pmatrix}$$

By using ∇, we can talk about the divergence of a function $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$, as being

$$< \nabla, \varphi >= \left\langle \begin{pmatrix} \frac{d}{dx} \\ \frac{d}{dy} \\ \frac{d}{dz} \end{pmatrix}, \begin{pmatrix} \varphi_x \\ \varphi_y \\ \varphi_z \end{pmatrix} \right\rangle = \frac{d\varphi_x}{dx} + \frac{d\varphi_y}{dy} + \frac{d\varphi_z}{dz}$$

as well as about the curl of the same function $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$, as being

$$\nabla \times \varphi = \begin{vmatrix} u_x & \frac{d}{dx} & \varphi_x \\ u_y & \frac{d}{dy} & \varphi_y \\ u_z & \frac{d}{dz} & \varphi_z \end{vmatrix} = \begin{pmatrix} \frac{d\varphi_z}{dy} - \frac{d\varphi_y}{dz} \\ \frac{d\varphi_x}{dz} - \frac{d\varphi_z}{dx} \\ \frac{d\varphi_y}{dx} - \frac{d\varphi_x}{dy} \end{pmatrix}$$

where u_x, u_y, u_z are the unit vectors along the coordinate directions x, y, z.

All this might seem a bit abstract, but is in fact very intuitive. The gradient ∇f points in the direction of the maximal increase of f, with $|\nabla f|$ giving you the rate of increase of f, in that direction. As for the divergence and curl, these measure the divergence and curl of the vectors $\varphi(u+v)$ around a given point $u \in \mathbb{R}^3$, in a usual, real-life sense.

Getting back now to calculus tools, what was missing from our picture was the higher dimensional analogue of the fundamental theorem of calculus, and more generally of the partial integration formula. In 3 dimensions, we have the following result:

THEOREM 15.32. *The following results hold, in* 3 *dimensions:*

(1) *Fundamental theorem for gradients, namely*

$$\int_a^b < \nabla f, dx >= f(b) - f(a)$$

(2) *Fundamental theorem for divergences, or Gauss or Green formula,*

$$\int_B < \nabla, \varphi >= \int_S < \varphi(x), n(x) > dx$$

(3) *Fundamental theorem for curls, or Stokes formula,*

$$\int_A < (\nabla \times \varphi)(x), n(x) > dx = \int_P < \varphi(x), dx >$$

where S is the boundary of the body B, and P is the boundary of the area A.

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

(1) This is something that we know well in 1D, namely the fundamental theorem of calculus, and the general, N-dimensional formula follows from that.

(2) This is something more subtle, and we had a taste of it when dealing with the Gauss law, and its various proofs. In general, the proof is similar, by using the various ideas from the proof of the Gauss law, and this can be worked out.

(3) This is again something subtle, and again with a flavor of things that we know, from the proof of the Gauss law, and which can be again worked out. □

Getting back now to electrostatics, as a main application of the above, we have the following new point of view on the Gauss formula, which is more conceptual:

THEOREM 15.33 (Gauss). *Given an electric potential E, its divergence is given by*

$$< \nabla, E >= \frac{\rho}{\varepsilon_0}$$

where ρ denotes as usual the charge distribution. Also, we have

$$\nabla \times E = 0$$

meaning that the curl of E vanishes.

PROOF. We have several assertions here, the idea being as follows:

(1) The first formula, called Gauss law in differential form, follows from:

$$\begin{aligned}\int_B <\nabla, E> &= \int_S <E(x), n(x)> dx\\ &= \Phi_E(S)\\ &= \frac{Q_{enc}}{\varepsilon_0}\\ &= \int_B \frac{\rho}{\varepsilon_0}\end{aligned}$$

Now since this must hold for any B, this gives the formula in the statement.

(2) As a side remark, the Gauss law in differential form can be established as well directly, with the computation, involving a Dirac mass, being as follows:

$$\begin{aligned}<\nabla, E>(x) &= \left\langle \nabla, K\int_{\mathbb{R}^3} \frac{\rho(z)(x-z)}{||x-z||^3}\, dz \right\rangle\\ &= K\int_{\mathbb{R}^3} \left\langle \nabla, \frac{x-z}{||x-z||^3} \right\rangle \rho(z)\, dz\\ &= K\int_{\mathbb{R}^3} 4\pi\delta_x \cdot \rho(z) dz\\ &= 4\pi K\int_{\mathbb{R}^3} \delta_x\, \rho(z) dz\\ &= \frac{\rho(x)}{\varepsilon_0}\end{aligned}$$

And with this in hand, we have via (1) a new proof of the usual Gauss law.

(3) Regarding the curl, by discretizing and linearity we can assume that we are dealing with a single charge q, positioned at 0. We have, by using spherical coordinates r, s, t:

$$\begin{aligned}\int_a^b < E(x), dx > &= \int_a^b \left\langle \frac{Kqx}{||x||^3}, dx \right\rangle \\ &= \int_a^b \left\langle \frac{Kq}{r^2} \cdot \frac{x}{||x||}, dx \right\rangle \\ &= \int_a^b \frac{Kq}{r^2}\, dr \\ &= \left[-\frac{Kq}{r} \right]_a^b \\ &= Kq \left(\frac{1}{r_a} - \frac{1}{r_b} \right)\end{aligned}$$

In particular the integral of E over any closed loop vanishes, and by using now Stokes' theorem, we conclude that the curl of E vanishes, as stated.

(4) Finally, as a side remark, both the formula of the divergence and the vanishing of the curl are somewhat clear by looking at the field lines of E. However, as all the above mathematics shows, there is certainly something to be understood, in all this. □

So long for electrostatics, which provide a good motivation and illustration for our mathematics. When upgrading to electrodynamics, things become even more interesting, because our technology can be used in order to understand the Maxwell equations.

15e. Exercises

We have a tough chapter here, and as exercises, we have:

EXERCISE 15.34. *Work out some numerics for the Coriolis force, on Earth.*

EXERCISE 15.35. *Learn about Einstein summation via Lorentz transformation.*

EXERCISE 15.36. *Prove the Gauss law by counting the flux lines. Can you?*

EXERCISE 15.37. *Find some further applications of Gauss, Green, Stokes.*

As bonus exercise, learn full electrodynamics, say from Griffiths [41].

CHAPTER 16

Infinite dimensions

16a. Operators, matrices

Welcome to quantum mechanics, and in the hope that we will survive. We already talked in chapter 8 about the main idea of Heisenberg, namely using infinite matrices in order to axiomatize quantum mechanics, based on the following key fact:

FACT 16.1 (Rydberg, Ritz). *The spectral lines of the hydrogen atom are given by the Rydberg formula, as follows, depending on integer parameters $n_1 < n_2$:*

$$\frac{1}{\lambda_{n_1n_2}} = R\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

These spectral lines combine according to the Ritz-Rydberg principle, as follows:

$$\frac{1}{\lambda_{n_1n_2}} + \frac{1}{\lambda_{n_2n_3}} = \frac{1}{\lambda_{n_1n_3}}$$

Similar formulae hold for other atoms, with suitable fine-tunings of the constant R.

We refer to chapter 8 for the full story with all this, which is theory based on some key observations of Lyman, Balmer, Paschen, around 1890-1900. The point now is that the above combination principle reminds the multiplication formula $e_{n_1n_2}e_{n_2n_3} = e_{n_1n_3}$ for the elementary matrices $e_{ij} : e_j \to e_i$, which leads to the following principle:

PRINCIPLE 16.2 (Heisenberg). *Observables in quantum mechanics should be some sort of infinite matrices, generalizing the Lyman, Balmer, Paschen lines of the hydrogen atom, and multiplying between them as the matrices do, as to produce further observables.*

All this is quite deep, and needs a number of comments, as follows:

(1) First of all, our matrices must be indeed infinite, because so are the series observed by Lyman, Balmer, Paschen, corresponding to $n_1 = 1, 2, 3$ in the Rydberg formula, and making it clear that the range of the second parameter $n_2 > n_1$ is up to ∞.

(2) Although this was not known to Ritz-Rydberg and Heisenberg, let us mention too that some later results of Brackett, Pfund, Humphreys and others, at $n_1 = 4, 5, 6, \ldots$, confirmed the fact that the range of the first parameter n_1 is up to ∞ too.

(3) As a more tricky comment now, going beyond what Principle 16.2 says, our infinite matrices must be in fact complex. This was something known to Heisenberg, and later Schrödinger came with proof that quantum mechanics naturally lives over $\mathbb{C}$.

(4) But all this leads us into some tricky mathematics, because the infinite matrices $A \in M_\infty(\mathbb{C})$ do not act on the vectors $v \in \mathbb{C}^\infty$ just like that. For instance the all-one matrix $A_{ij} = 1$ does not act on the all-one vector $v_i = 1$, for obvious reasons.

Summarizing, in order to get to some mathematical theory going, out of Principle 16.2, we must assume that our matrices $A \in M_\infty(\mathbb{C})$ must be "bounded" in some sense. Or perhaps the vectors $v \in \mathbb{C}^\infty$ must be bounded. Or perhaps, both.

In order to fix all this, let us start with $\mathbb{C}^\infty$. We would like to replace it with its subspace $H = l^2(\mathbb{N})$ consisting of vectors having finite norm, as for our various computations to converge. But this being said, taking a look at what Schrödinger was saying too, a bit later, why not including right away in our theory spaces like $H = L^2(\mathbb{R}^3)$ too, which are perhaps a bit more relevant than Heisenberg's $l^2(\mathbb{N})$. We are led in this way into:

DEFINITION 16.3. *A Hilbert space is a complex vector space H with a scalar product $< x, y >$, which will be linear at left and antilinear at right,*

$$< \lambda x, y >= \lambda < x, y > \quad , \quad < x, \lambda y >= \bar{\lambda} < x, y >$$

and which is complete with respect to corresponding norm

$$||x|| = \sqrt{< x, x >}$$

in the sense that any sequence $\{x_n\}$ which is a Cauchy sequence, having the property $||x_n - x_m|| \to 0$ with $n, m \to \infty$, has a limit, $x_n \to x$.

Here our convention for the scalar products, written $< x, y >$ and being linear at left, is one among others, often used by mathematicians, and we will just use this, in the lack of a physicist with an axe around. As further comments now on Definition 16.3, there is some mathematics encapsulated there, needing some discussion. First, we have:

THEOREM 16.4. *Given an index set I, which can be finite or not, the space of square-summable vectors having indices in I, namely*

$$l^2(I) = \left\{ (x_i)_{i \in I} \Big| \sum_i |x_i|^2 < \infty \right\}$$

is a Hilbert space, with scalar product as follows:

$$< x, y >= \sum_i x_i \bar{y}_i$$

When I is finite, $I = \{1, \ldots, N\}$, we obtain in this way the usual space $H = \mathbb{C}^N$.

PROOF. We have already met such things in chapter 7, but let us recall all this:

(1) We know that $l^2(I) \subset \mathbb{C}^I$ is the space of vectors satisfying $||x|| < \infty$. We want to prove that $l^2(I)$ is a vector space, that $< x, y >$ is a scalar product on it, that $l^2(I)$ is complete with respect to $||.||$, and finally that for $|I| < \infty$ we have $l^2(I) = \mathbb{C}^{|I|}$.

(2) The last assertion, $l^2(I) = \mathbb{C}^{|I|}$ for $|I| < \infty$, is clear, because in this case the sums are finite, so the condition $||x|| < \infty$ is automatic. So, we know at least one thing.

(3) Regarding the rest, our claim here, which will more or less prove everything, is that for any two vectors $x, y \in l^2(I)$ we have the Cauchy-Schwarz inequality:

$$| < x, y > | \leq ||x|| \cdot ||y||$$

But this follows from the positivity of the following degree 2 quantity, depending on a real variable $t \in \mathbb{R}$, and on a variable on the unit circle, $w \in \mathbb{T}$:

$$f(t) = ||twx + y||^2$$

(4) Now with Cauchy-Schwarz proved, everything is straightforward. We first obtain, by raising to the square and expanding, that for any $x, y \in l^2(I)$ we have:

$$||x + y|| \leq ||x|| + ||y||$$

Thus $l^2(I)$ is indeed a vector space, the other vector space conditions being trivial.

(5) Also, $< x, y >$ is surely a scalar product on this vector space, because all the conditions for a scalar product are trivially satisfied.

(6) Finally, the fact that our space $l^2(I)$ is indeed complete with respect to its norm $||.||$ follows in the obvious way, the limit of a Cauchy sequence $\{x_n\}$ being the vector $y = (y_i)$ given by $y_i = \lim_{n\to\infty} x_{ni}$, with all the verifications here being trivial. □

Going now a bit abstract, we have, more generally, the following result, which shows that our formalism covers as well the Schrödinger spaces of type $L^2(\mathbb{R}^3)$:

THEOREM 16.5. *Given an arbitrary space X with a positive measure μ on it, the space of square-summable complex functions on it, namely*

$$L^2(X) = \left\{ f : X \to \mathbb{C} \Big| \int_X |f(x)|^2 \, d\mu(x) < \infty \right\}$$

is a Hilbert space, with scalar product as follows:

$$< f, g > = \int_X f(x)\overline{g(x)} \, d\mu(x)$$

When $X = I$ is discrete, meaning that the measure μ on it is the counting measure, $\mu(\{x\}) = 1$ for any $x \in X$, we obtain in this way the previous spaces $l^2(I)$.

PROOF. This is something routine, remake of Theorem 16.4, as follows:

(1) The proof of the first, and main assertion is something perfectly similar to the proof of Theorem 16.4, by replacing everywhere the sums by integrals.

(2) With the remark that we forgot to say in the statement that the L^2 functions are by definition taken up to equality almost everywhere, $f = g$ when $||f - g|| = 0$.

(3) As for the last assertion, when μ is the counting measure all our integrals here become usual sums, and so we recover in this way Theorem 16.4. □

As a third and last theorem about Hilbert spaces, that we will need, we have:

THEOREM 16.6. *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 worse, depends only on H, and is called dimension of H. We have

$$H \simeq l^2(I)$$

in the obvious way, mapping $\sum \lambda_i e_i \to (\lambda_i)$. The Hilbert spaces with $\dim H = |I|$ being countable, including $l^2(\mathbb{N})$ and $L^2(\mathbb{R})$, 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 $\{x_i\}_{i\in I}$, and using the 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 scary tools from logic, and more specifically the Zorn lemma.

(4) Finally, everything at the end is clear from definitions, except perhaps for the fact that $L^2(\mathbb{R})$ is separable. But here we can argue that, since functions can be approximated by polynomials, we have a countable algebraic basis, namely $\{x^n\}_{n\in\mathbb{N}}$, called the Weierstrass basis, that we can orthogonalize afterwards by using Gram-Schmidt. □

Moving ahead, now that we know what our vector spaces are, we can talk about infinite matrices with respect to them. And the situation here is as follows:

THEOREM 16.7. *Given a Hilbert space H, consider the linear operators $T : H \to H$, and for each such operator define its norm by the following formula:*

$$||T|| = \sup_{||x||=1} ||Tx||$$

The operators which are bounded, $||T|| < \infty$, form then a complex algebra $B(H)$, which is complete with respect to $||.||$. When H comes with a basis $\{e_i\}_{i\in I}$, we have

$$B(H) \subset \mathcal{L}(H) \subset M_I(\mathbb{C})$$

where $\mathcal{L}(H)$ is the algebra of all linear operators $T : H \to H$, and $\mathcal{L}(H) \subset M_I(\mathbb{C})$ is the correspondence $T \to M$ obtained via the usual linear algebra formulae, namely:

$$T(x) = Mx \quad , \quad M_{ij} =< Te_j, e_i >$$

In infinite dimensions, none of the above two inclusions is an equality.

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

(1) The fact that we have indeed an algebra, satisfying the product condition in the statement, follows from the following estimates, which are all elementary:

$$||S + T|| \leq ||S|| + ||T|| \quad , \quad ||\lambda T|| = |\lambda| \cdot ||T|| \quad , \quad ||ST|| \leq ||S|| \cdot ||T||$$

(2) Regarding now the completness assertion, if $\{T_n\} \subset B(H)$ is Cauchy then $\{T_n x\}$ is Cauchy for any $x \in H$, so we can define the limit $T = \lim_{n\to\infty} T_n$ by setting:

$$Tx = \lim_{n\to\infty} T_n x$$

Let us first check that the application $x \to Tx$ is linear. We have:

$$\begin{aligned} T(x + y) &= \lim_{n\to\infty} T_n(x + y) \\ &= \lim_{n\to\infty} T_n(x) + T_n(y) \\ &= \lim_{n\to\infty} T_n(x) + \lim_{n\to\infty} T_n(y) \\ &= T(x) + T(y) \end{aligned}$$

Similarly, we have $T(\lambda x) = \lambda T(x)$, and we conclude that $T \in \mathcal{L}(H)$.

(3) With this done, it remains to prove now that we have $T \in B(H)$, and that $T_n \to T$ in norm. For this purpose, observe that we have:

$$\begin{aligned} ||T_n - T_m|| \leq \varepsilon \ , \ \forall n, m \geq N \quad &\implies \quad ||T_n x - T_m x|| \leq \varepsilon \ , \ \forall ||x|| = 1 \ , \ \forall n, m \geq N \\ &\implies \quad ||T_n x - Tx|| \leq \varepsilon \ , \ \forall ||x|| = 1 \ , \ \forall n \geq N \\ &\implies \quad ||T_N x - Tx|| \leq \varepsilon \ , \ \forall ||x|| = 1 \\ &\implies \quad ||T_N - T|| \leq \varepsilon \end{aligned}$$

But this gives both $T \in B(H)$, and $T_N \to T$ in norm, and we are done.

(4) Regarding the embeddings, the correspondence $T \to M$ in the statement is indeed linear, and its kernel is $\{0\}$, so we have indeed an embedding as follows, as claimed:

$$\mathcal{L}(H) \subset M_I(\mathbb{C})$$

In finite dimensions we have an isomorphism, because any $M \in M_N(\mathbb{C})$ determines an operator $T : \mathbb{C}^N \to \mathbb{C}^N$, given by $< Te_j, e_i >= M_{ij}$. However, in infinite dimensions, we have matrices not producing operators, as for instance the all-one matrix.

(5) As for the examples of linear operators which are not bounded, these are more complicated, coming from logic, and we will not need them in what follows. □

Finally, as a second and last result regarding the operators, we will need:

THEOREM 16.8. *Each operator $T \in B(H)$ has an adjoint $T^* \in B(H)$, given by:*

$$< Tx, y >=< x, T^*y >$$

The operation $T \to T^*$ is antilinear, antimultiplicative, involutive, and satisfies:

$$||T|| = ||T^*|| \quad , \quad ||TT^*|| = ||T||^2$$

When H comes with a basis $\{e_i\}_{i \in I}$, the operation $T \to T^*$ corresponds to

$$(M^*)_{ij} = \overline{M}_{ji}$$

at the level of the associated matrices $M \in M_I(\mathbb{C})$.

PROOF. This is standard too, and can be proved in 3 steps, as follows:

(1) The existence of the adjoint operator T^*, given by the formula in the statement, comes from the fact that the function $\varphi(x) =< Tx, y >$ being a linear map $H \to \mathbb{C}$, we must have a formula as follows, for a certain vector $T^*y \in H$:

$$\varphi(x) =< x, T^*y >$$

Moreover, since this vector is unique, T^* is unique too, and we have as well:

$$(S + T)^* = S^* + T^* \quad , \quad (\lambda T)^* = \bar{\lambda} T^* \quad , \quad (ST)^* = T^* S^* \quad , \quad (T^*)^* = T$$

Observe also that we have indeed $T^* \in B(H)$, because:

$$\begin{aligned} ||T|| &= \sup_{||x||=1} \sup_{||y||=1} < Tx, y > \\ &= \sup_{||y||=1} \sup_{||x||=1} < x, T^*y > \\ &= ||T^*|| \end{aligned}$$

(2) Regarding now $||TT^*|| = ||T||^2$, which is a key formula, observe that we have:

$$||TT^*|| \leq ||T|| \cdot ||T^*|| = ||T||^2$$

On the other hand, we have as well the following estimate:

$$\begin{aligned} ||T||^2 &= \sup_{||x||=1} |<Tx, Tx>| \\ &= \sup_{||x||=1} |<x, T^*Tx>| \\ &\leq ||T^*T|| \end{aligned}$$

By replacing $T \to T^*$ we obtain from this $||T||^2 \leq ||TT^*||$, as desired.

(3) Finally, when H comes with a basis, the formula $<Tx, y>=<x, T^*y>$ applied with $x = e_i$, $y = e_j$ translates into the formula $(M^*)_{ij} = \overline{M}_{ji}$, as desired. □

So long for Hilbert spaces and operators. For more, you can check my book [10].

16b. Schrödinger equation

Time now to get into quantum mechanics. This is a bit of a nightmare, and I am afraid that I will have to ask again the cat. With him, things going better since last time, we sort of agree now on multivariable calculus, following a long philosophical debate, and my willingness to start sharing his diet, namely raw mice and birds. But let's see now in relation with this highly divisive topic which is quantum mechanics:

CAT 16.9. *Schrödinger.*

Thanks cat, this is actually quite surprising, and I am particularly pleased to hear this. So, before getting back to what Heisenberg was saying, based on Lyman, Balmer, Paschen, namely developing some sort of "matrix mechanics", let us hear as well the point of view of Schrödinger, which came a few years later. His idea was to forget about exact things, and try to investigate the hydrogen atom statistically. Let us start with:

QUESTION 16.10. *In the context of the hydrogen atom, assuming that the proton is fixed, what is the probability density $\varphi_t(x)$ of the position of the electron e, at time t,*

$$P_t(e \in V) = \int_V \varphi_t(x)dx$$

as function of an intial probability density $\varphi_0(x)$? Moreover, can the corresponding equation be solved, and will this prove the Bohr claims for hydrogen, statistically?

In order to get familiar with this question, let us first look at examples coming from classical mechanics. In the context of a particle whose position at time t is given by $x_0 + \gamma(t)$, the evolution of the probability density will be given by:

$$\varphi_t(x) = \varphi_0(x) + \gamma(t)$$

However, such examples are somewhat trivial, of course not in relation with the computation of γ, usually a difficult question, but in relation with our questions, and do not apply to the electron. The point indeed is that, in what regards the electron, we have:

FACT 16.11. *In respect with various simple interference experiments:*

(1) *The electron is definitely not a particle in the usual sense.*
(2) *But in most situations it behaves exactly like a wave.*
(3) *But in other situations it behaves like a particle.*

Getting back now to the Schrödinger question, all this suggests to use, as for the waves, an amplitude function $\psi_t(x) \in \mathbb{C}$, related to the density $\varphi_t(x) > 0$ by the formula $\varphi_t(x) = |\psi_t(x)|^2$. Not that a big deal, you would say, because the two are related by simple formulae as follows, with $\theta_t(x)$ being an arbitrary phase function:

$$\varphi_t(x) = |\psi_t(x)|^2 \quad , \quad \psi_t(x) = e^{i\theta_t(x)}\sqrt{\varphi_t(x)}$$

However, such manipulations can be crucial, raising for instance the possibility that the amplitude function satisfies some simple equation, while the density itself, maybe not. And this is what happens indeed. Schrödinger was led in this way to:

CLAIM 16.12 (Schrödinger). *In the context of the hydrogen atom, the amplitude function of the electron $\psi = \psi_t(x)$ is subject to the Schrödinger equation*

$$ih\dot{\psi} = -\frac{h^2}{2m}\Delta\psi + V\psi$$

m being the mass, $h = h_0/2\pi$ the reduced Planck constant, and V the Coulomb potential of the proton. The same holds for movements of the electron under any potential V.

Observe the similarity with the wave equation $\ddot{\varphi} = v^2\Delta\varphi$, and with the heat equation $\dot{\varphi} = \alpha\Delta\varphi$ too. Many things can be said here. Following now Heisenberg and Schrödinger, and then especially Dirac, who did the axiomatization work, we have:

DEFINITION 16.13. *In quantum mechanics the states of the system are vectors of a Hilbert space H, and the observables of the system are linear operators*

$$T : H \to H$$

which can be densely defined, and are taken self-adjoint, $T = T^*$. The average value of such an observable T, evaluated on a state $\xi \in H$, is given by:

$$< T >=< T\xi, \xi >$$

In the context of the Schrödinger mechanics of the hydrogen atom, the Hilbert space is the space $H = L^2(\mathbb{R}^3)$ where the wave function ψ lives, and we have

$$< T >= \int_{\mathbb{R}^3} T(\psi) \cdot \bar{\psi}\, dx$$

which is called "sandwiching" formula, with the operators

$$x \quad , \quad -\frac{ih}{m}\nabla \quad , \quad -ih\nabla \quad , \quad -\frac{h^2\Delta}{2m} \quad , \quad -\frac{h^2\Delta}{2m} + V$$

representing the position, speed, momentum, kinetic energy, and total energy.

In other words, we are doing here two things. First, we are declaring by axiom that various "sandwiching" formulae found before by Heisenberg, involving the operators at the end, that we will not get into in detail here, hold true. And second, we are raising the possibility for other quantum mechanical systems, more complicated, to be described as well by the mathematics of the operators on a certain Hilbert space H, as above.

So, this was the story of early quantum mechanics, over-simplified as to fit here in a few pages. For more, you can check Feynman [33] for foundations, and everything, including for some nice pictures and explanations regarding Fact 16.11. You have as well Griffiths [42] or Weinberg [94], for further explanations on Definition 16.13, not to forget Dirac's original text [22], and all this is discussed as well in my book [11].

16c. Spherical coordinates

In order to solve now the hydrogen atom, the idea will be that of reformulating the Schrödinger equation in spherical coordinates. And for this purpose, we will need:

THEOREM 16.14. *The Laplace operator in spherical coordinates is*

$$\Delta = \frac{1}{r^2} \cdot \frac{d}{dr}\left(r^2 \cdot \frac{d}{dr}\right) + \frac{1}{r^2 \sin s} \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d}{ds}\right) + \frac{1}{r^2 \sin^2 s} \cdot \frac{d^2}{dt^2}$$

with our standard conventions for these coordinates, in 3D.

PROOF. There are several proofs here, a short, elementary one being as follows:

(1) Let us first see how Δ behaves under a change of coordinates $\{x_i\} \to \{y_i\}$, in arbitrary N dimensions. Our starting point is the chain rule for derivatives:

$$\frac{d}{dx_i} = \sum_j \frac{d}{dy_j} \cdot \frac{dy_j}{dx_i}$$

By using this rule, then Leibnitz for products, then again this rule, we obtain:

$$\begin{aligned}
\frac{d^2 f}{dx_i^2} &= \sum_j \frac{d}{dx_i}\left(\frac{df}{dy_j} \cdot \frac{dy_j}{dx_i}\right) \\
&= \sum_j \frac{d}{dx_i}\left(\frac{df}{dy_j}\right) \cdot \frac{dy_j}{dx_i} + \frac{df}{dy_j} \cdot \frac{d}{dx_i}\left(\frac{dy_j}{dx_i}\right) \\
&= \sum_j \left(\sum_k \frac{d}{dy_k} \cdot \frac{dy_k}{dx_i}\right)\left(\frac{df}{dy_j}\right) \cdot \frac{dy_j}{dx_i} + \frac{df}{dy_j} \cdot \frac{d^2 y_j}{dx_i^2} \\
&= \sum_{jk} \frac{d^2 f}{dy_k dy_j} \cdot \frac{dy_k}{dx_i} \cdot \frac{dy_j}{dx_i} + \sum_j \frac{df}{dy_j} \cdot \frac{d^2 y_j}{dx_i^2}
\end{aligned}$$

(2) Now by summing over i, we obtain the following formula, with A being the derivative of $x \to y$, that is to say, the matrix of partial derivatives dy_i/dx_j:

$$\begin{aligned}\Delta f &= \sum_{ijk} \frac{d^2f}{dy_k dy_j} \cdot \frac{dy_k}{dx_i} \cdot \frac{dy_j}{dx_i} + \sum_{ij} \frac{df}{dy_j} \cdot \frac{d^2y_j}{dx_i^2} \\ &= \sum_{ijk} A_{ki}A_{ji} \frac{d^2f}{dy_k dy_j} + \sum_{ij} \frac{d^2y_j}{dx_i^2} \cdot \frac{df}{dy_j} \\ &= \sum_{jk} (AA^t)_{jk} \frac{d^2f}{dy_k dy_j} + \sum_j \Delta(y_j) \frac{df}{dy_j}\end{aligned}$$

(3) So, this will be the formula that we will need. Observe that this formula can be further compacted as follows, with all the notations being self-explanatory:

$$\Delta f = Tr(AA^t H_y(f)) + < \Delta(y), \nabla_y(f) >$$

(4) Getting now to spherical coordinates, $(x, y, z) \to (r, s, t)$, the derivative of the inverse, obtained by differentiating x, y, z with respect to r, s, t, is given by:

$$A^{-1} = \begin{pmatrix} \cos s & -r\sin s & 0 \\ \sin s \cos t & r\cos s \cos t & -r\sin s \sin t \\ \sin s \sin t & r\cos s \sin t & r\sin s \cos t \end{pmatrix}$$

The product $(A^{-1})^t A^{-1}$ of the transpose of this matrix with itself is then:

$$\begin{pmatrix} \cos s & \sin s \cos t & \sin s \sin t \\ -r\sin s & r\cos s \cos t & r\cos s \sin t \\ 0 & -r\sin s \sin t & r\sin s \cos t \end{pmatrix} \begin{pmatrix} \cos s & -r\sin s & 0 \\ \sin s \cos t & r\cos s \cos t & -r\sin s \sin t \\ \sin s \sin t & r\cos s \sin t & r\sin s \cos t \end{pmatrix}$$

But everything simplifies here, and we have the following remarkable formula, which by the way is something very useful, worth to be memorized:

$$(A^{-1})^t A^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2 s \end{pmatrix}$$

Now by inverting, we obtain the following formula, in relation with the above:

$$AA^t = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/r^2 & 0 \\ 0 & 0 & 1/(r^2 \sin^2 s) \end{pmatrix}$$

(5) Let us compute now the Laplacian of r, s, t. We first have the following formula, that we will use many times in what follows, and is worth to be memorized:

$$\begin{aligned}\frac{dr}{dx} &= \frac{d}{dx}\sqrt{x^2+y^2+z^2}\\ &= \frac{1}{2}\cdot\frac{2x}{\sqrt{x^2+y^2+z^2}}\\ &= \frac{x}{r}\end{aligned}$$

Of course the same computation works for y, z too, and we therefore have:

$$\frac{dr}{dx}=\frac{x}{r}\quad,\quad\frac{dr}{dy}=\frac{y}{r}\quad,\quad\frac{dr}{dz}=\frac{z}{r}$$

(6) By using the above formulae, twice, we can compute the Laplacian of r:

$$\begin{aligned}\Delta(r) &= \Delta\left(\sqrt{x^2+y^2+z^2}\right)\\ &= \frac{d}{dx}\left(\frac{x}{r}\right)+\frac{d}{dy}\left(\frac{y}{r}\right)+\frac{d}{dz}\left(\frac{z}{r}\right)\\ &= \frac{r^2-x^2}{r^3}+\frac{r^2-y^2}{r^3}+\frac{r^2-z^2}{r^3}\\ &= \frac{2}{r}\end{aligned}$$

(7) In what regards now s, the computation here goes as follows:

$$\begin{aligned}\Delta(s) &= \Delta\left(\arccos\left(\frac{x}{r}\right)\right)\\ &= \frac{d}{dx}\left(-\frac{\sqrt{r^2-x^2}}{r^2}\right)+\frac{d}{dy}\left(\frac{xy}{r^2\sqrt{r^2-x^2}}\right)+\frac{d}{dz}\left(\frac{xz}{r^2\sqrt{r^2-x^2}}\right)\\ &= \frac{2x\sqrt{r^2-x^2}}{r^4}+\frac{r^2(z^2-2y^2)+2x^2y^2}{r^4\sqrt{r^2-x^2}}+\frac{r^2(y^2-2z^2)+2x^2z^2}{r^4\sqrt{r^2-x^2}}\\ &= \frac{2x\sqrt{r^2-x^2}}{r^4}+\frac{x(2x^2-r^2)}{r^4\sqrt{r^2-x^2}}\\ &= \frac{x}{r^2\sqrt{r^2-x^2}}\\ &= \frac{\cos s}{r^2\sin s}\end{aligned}$$

(8) Finally, in what regards t, the computation here goes as follows:

$$\begin{aligned}\Delta(t) &= \Delta\left(\arctan\left(\frac{z}{y}\right)\right)\\ &= \frac{d}{dx}(0)+\frac{d}{dy}\left(-\frac{z}{y^2+z^2}\right)+\frac{d}{dz}\left(\frac{y}{y^2+z^2}\right)\\ &= 0-\frac{2yz}{(y^2+z^2)^2}+\frac{2yz}{(y^2+z^2)^2}\\ &= 0\end{aligned}$$

(9) We can now plug the data from (4) and (6,7,8) in the general formula that we found in (2) above, and we obtain in this way:

$$\begin{aligned}\Delta f &= \frac{d^2f}{dr^2}+\frac{1}{r^2}\cdot\frac{d^2f}{ds^2}+\frac{1}{r^2\sin^2 s}\cdot\frac{d^2f}{dt^2}+\frac{2}{r}\cdot\frac{df}{dr}+\frac{\cos s}{r^2\sin s}\cdot\frac{df}{ds}\\ &= \frac{2}{r}\cdot\frac{df}{dr}+\frac{d^2f}{dr^2}+\frac{\cos s}{r^2\sin s}\cdot\frac{df}{ds}+\frac{1}{r^2}\cdot\frac{d^2f}{ds^2}+\frac{1}{r^2\sin^2 s}\cdot\frac{d^2f}{dt^2}\\ &= \frac{1}{r^2}\cdot\frac{d}{dr}\left(r^2\cdot\frac{df}{dr}\right)+\frac{1}{r^2\sin s}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{df}{ds}\right)+\frac{1}{r^2\sin^2 s}\cdot\frac{d^2f}{dt^2}\end{aligned}$$

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

Still with me, I hope, and do not worry, one day you will have such computations for breakfast. We can now reformulate the Schrödinger equation in spherical coordinates, and separate the variables, which leads to a radial and angular equation, as follows:

THEOREM 16.15. *The time-independent Schrödinger equation in spherical coordinates separates, for solutions of type $\phi=\rho(r)\alpha(s,t)$, into two equations, as follows,*

$$\frac{d}{dr}\left(r^2\cdot\frac{d\rho}{dr}\right)-\frac{2mr^2}{h^2}(V-E)\rho=K\rho$$

$$\sin s\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{d^2\alpha}{dt^2}=-K\sin^2 s\cdot\alpha$$

with K being a constant, called radial equation, and angular equation.

PROOF. By using the formula in Theorem 16.14, the time-independent Schrödinger equation reformulates in spherical coordinates as follows:

$$(V-E)\phi=\frac{h^2}{2m}\left[\frac{1}{r^2}\cdot\frac{d}{dr}\left(r^2\cdot\frac{d\phi}{dr}\right)+\frac{1}{r^2\sin s}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\phi}{ds}\right)+\frac{1}{r^2\sin^2 s}\cdot\frac{d^2\phi}{dt^2}\right]$$

Let us look now for separable solutions for this latter equation, consisting of a radial part and an angular part, as in the statement, namely:

$$\phi(r,s,t)=\rho(r)\alpha(s,t)$$

By plugging this function into our equation, we obtain:

$$(V-E)\rho\alpha=\frac{h^2}{2m}\left[\frac{\alpha}{r^2}\cdot\frac{d}{dr}\left(r^2\cdot\frac{d\rho}{dr}\right)+\frac{\rho}{r^2\sin s}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{\rho}{r^2\sin^2 s}\cdot\frac{d^2\alpha}{dt^2}\right]$$

In order to solve this equation, we will do two manipulations. First, by multiplying everything by $2mr^2/(h^2\rho\alpha)$, this equation takes the following more convenient form:

$$\frac{2mr^2}{h^2}(V-E)=\frac{1}{\rho}\cdot\frac{d}{dr}\left(r^2\cdot\frac{d\rho}{dr}\right)+\frac{1}{\alpha\sin s}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{1}{\alpha\sin^2 s}\cdot\frac{d^2\alpha}{dt^2}$$

Now observe that by moving the radial terms to the left, and the angular terms to the right, this latter equation can be written as follows:

$$\frac{2mr^2}{h^2}(V-E)-\frac{1}{\rho}\cdot\frac{d}{dr}\left(r^2\cdot\frac{d\rho}{dr}\right)=\frac{1}{\alpha\sin^2 s}\left[\sin s\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{d^2\alpha}{dt^2}\right]$$

Since this latter equation is now separated between radial and angular variables, both sides must be equal to a certain constant $-K$, as follows:

$$\frac{2mr^2}{h^2}(V-E)-\frac{1}{\rho}\cdot\frac{d}{dr}\left(r^2\cdot\frac{d\rho}{dr}\right)=-K$$

$$\frac{1}{\alpha\sin^2 s}\left[\sin s\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{d^2\alpha}{dt^2}\right]=-K$$

But this leads to the conclusion in the statement. □

Let us first study the angular equation. We first have the following result:

PROPOSITION 16.16. *The angular equation that we found before, namely*

$$\sin s\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{d^2\alpha}{dt^2}=-K\sin^2 s\cdot\alpha$$

separates, for solutions of type $\alpha=\sigma(s)\theta(t)$, into two equations, as follows,

$$\frac{1}{\theta}\cdot\frac{d^2\theta}{dt^2}=-m^2$$

$$\frac{\sin s}{\sigma}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\sigma}{ds}\right)+K\sin^2 s=m^2$$

with m being a constant, called azimuthal equation, and polar equation.

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

(1) Let us first recall that $r\in[0,\infty)$ is the radius, $s\in[0,\pi]$ is the polar angle, and $t\in[0,2\pi]$ is the azimuthal angle. Be said in passing, there are several conventions and notations here, and the above ones, that we use here, come from the general ones in N dimensions, because further coordinates can be easily added, in the obvious way.

(2) Getting back now to our question, by plugging $\alpha = \sigma(s)\theta(t)$ into the angular equation, we obtain:

$$\sin s \cdot \theta \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\sigma}{ds}\right) + \sigma \cdot \frac{d^2\theta}{dt^2} = -K \sin^2 s \cdot \sigma\theta$$

By dividing everything by $\sigma\theta$, this equation can be written as follows:

$$-\frac{1}{\theta} \cdot \frac{d^2\theta}{dt^2} = \frac{\sin s}{\sigma} \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\sigma}{ds}\right) + K \sin^2 s$$

Since the variables are separated, we must have, for a certain constant m:

$$\frac{1}{\theta} \cdot \frac{d^2\theta}{dt^2} = -m^2$$

$$\frac{\sin s}{\sigma} \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\sigma}{ds}\right) + K \sin^2 s = m^2$$

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

Regarding the azimuthal equation, things here are quickly settled, as follows:

PROPOSITION 16.17. *The solutions of the azimuthal equation, namely*

$$\frac{1}{\theta} \cdot \frac{d^2\theta}{dt^2} = -m^2$$

are the functions as follows, with $a, b \in \mathbb{C}$ being parameters,

$$\theta(t) = ae^{imt} + be^{-imt}$$

and with only the case $m \in \mathbb{Z}$ being acceptable, on physical grounds.

PROOF. The first assertion is clear, because we have a second order equation, and two obvious solutions for it, $e^{\pm imt}$, and then their linear combinations, and that's all. Regarding the last assertion, the point here is that by using $\theta(t) = \theta(t + 2\pi)$, which is a natural physical assumption on the wave function, we are led to $m \in \mathbb{Z}$, as stated. □

We are now about to solve the angular equation, with only the polar equation remaining to be studied. However, in practice, this polar equation is 10 times more difficult that everything what we did so far, so be patient. We first have:

PROPOSITION 16.18. *The polar equation that we found before, namely*

$$\frac{\sin s}{\sigma} \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\sigma}{ds}\right) + K \sin^2 s = m^2$$

with $m \in \mathbb{Z}$, translates via $\sigma(s) = f(\cos s)$ into the following equation,

$$(1 - x^2)f''(x) - 2xf'(x) = \left(\frac{m^2}{1 - x^2} - K\right) f(x)$$

where $x = \cos s$, called Legendre equation.

PROOF. Let us first do a number of manipulations on our equation, before making the change of variables. By multiplying by σ, our equation becomes:

$$\sin s \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\sigma}{ds}\right) = \left(m^2 - K\sin^2 s\right)\sigma$$

By differentiating at left, this equation becomes:

$$\sin s\left(\cos s \cdot \sigma' + \sin s \cdot \sigma''\right) = \left(m^2 - K\sin^2 s\right)\sigma$$

Finally, by dividing everything by $\sin^2 s$, our equation becomes:

$$\sigma'' + \frac{\cos s}{\sin s} \cdot \sigma' = \left(\frac{m^2}{\sin^2 s} - K\right)\sigma$$

Now let us set $\sigma(s) = f(\cos s)$. With this change of variables, we have:

$$\sigma = f(\cos s)$$

$$\sigma' = -\sin s \cdot f'(\cos s)$$

$$\sigma'' = -\cos s \cdot f'(\cos s) + \sin^2 s \cdot f''(\cos s)$$

By plugging this data, our radial equation becomes:

$$\sin^2 s \cdot f''(\cos s) - 2\cos s \cdot f'(\cos s) = \left(\frac{m^2}{\sin^2 s} - K\right) f(\cos s)$$

Now with $x = \cos s$, which is our new variable, this equation reads:

$$(1 - x^2)f''(x) - 2xf'(x) = \left(\frac{m^2}{1 - x^2} - K\right) f(x)$$

But this is the Legendre equation, as stated. □

Here comes now the difficult point. We have the following non-trivial result:

THEOREM 16.19. *The solutions of the Legendre equation, namely*

$$(1 - x^2)f''(x) - 2xf'(x) = \left(\frac{m^2}{1 - x^2} - K\right) f(x)$$

can be explicitely computed, via complicated math, and only the case

$$K = l(l + 1) \quad : \quad l \in \mathbb{N}$$

is acceptable, on physical grounds.

PROOF. The first part is something quite complicated, involving the hypergeometric functions ${}_2F_1$, that you don't want to hear about, believe me. As for the second part, analysis and physics, this is something not trivial either. See Griffiths [42]. □

In order to construct the solutions, we will need:

THEOREM 16.20. *The orthonormal basis of $L^2[-1,1]$ obtained by starting with the Weierstrass basis $\{x^l\}$, and doing Gram-Schmidt, is the family of polynomials $\{P_l\}$, with each P_l being of degree l, and with positive leading coefficient, subject to:*

$$\int_{-1}^{1} P_k(x)P_l(x)\,dx = \delta_{kl}$$

These polynomials, called Legendre polynomials, satisfy the equation

$$(1-x^2)P_l''(x) - 2xP_l'(x) + l(l+1)P_l(x) = 0$$

which is the Legendre equation at $m = 0$, and with $K = l(l+1)$. Moreover,

$$P_l(x) = \frac{1}{2^l l!}\left(\frac{d}{dx}\right)^l (x^2-1)^l$$

which is called the Rodrigues formula for Legendre polynomials.

PROOF. As a first observation, we are not lost somewhere in abstract math, because of the occurrence of the Legendre equation. As for the proof, this goes as follows:

(1) The first assertion is clear, because the Gram-Schmidt procedure applied to the Weierstrass basis $\{x^l\}$ can only lead to a certain family of polynomials $\{P_l\}$, with each P_l being of degree l, and also unique, if we assume that it has positive leading coefficient, with this $\pm$ choice being needed, as usual, at each step of Gram-Schmidt.

(2) In order to have now an idea about these beasts, here are the first few of them, which can be obtained say via a straightforward application of Gram-Schmidt:

$$\begin{aligned}
P_0 &= 1\\
P_1 &= x\\
P_2 &= (3x^2-1)/2\\
P_3 &= (5x^3-3x)/2\\
P_4 &= (35x^4-30x^2+3)/8\\
P_5 &= (63x^5-70x^3+15x)/8
\end{aligned}$$

(3) Now thinking about what Gram-Schmidt does, this is certainly something by recurrence. And examining the recurrence leads to the Legendre equation, as stated.

(4) As for the Rodrigues formula, by uniqueness no need to try to understand where this formula comes from, and we have two choices here, either by verifying that $\{P_l\}$ is orthonormal, or by verifying the Legendre equation. And both methods work. □

Going ahead now, we can solve in fact the Legendre equation at any m, as follows:

PROPOSITION 16.21. *The general Legendre equation, with parameters $m \in \mathbb{N}$ and $K = l(l+1)$ with $l \in \mathbb{N}$, namely*

$$(1-x^2)f''(x) - 2xf'(x) = \left(\frac{m^2}{1-x^2} - l(l+1)\right) f(x)$$

is solved by the following functions, called Legendre functions,

$$P_l^m(x) = (-1)^m (1-x^2)^{m/2} \left(\frac{d}{dx}\right)^m P_l(x)$$

where P_l are as before the Legendre polynomials. Also, we have

$$P_l^m(x) = (-1)^m \frac{(1-x^2)^{m/2}}{2^l l!} \left(\frac{d}{dx}\right)^{l+m} (x^2-1)^l$$

called Rodrigues formula for Legendre functions.

PROOF. The first assertion is something elementary, coming by differentiating m times the Legendre equation, which leads to the general Legendre equation. As for the second assertion, this follows from the Rodrigues formula for Legendre polynomials. □

And this is the end of our study. Eventually. By putting together all the above results, we are led to the following conclusion:

THEOREM 16.22. *The separated solutions $\alpha = \sigma(s)\theta(t)$ of the angular equation,*

$$\sin s \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\alpha}{ds}\right) + \frac{d^2\alpha}{dt^2} = -K \sin^2 s \cdot \alpha$$

are given by the following formulae, where $l \in \mathbb{N}$ is such that $K = l(l+1)$,

$$\sigma(s) = P_l^m(\cos s) \quad , \quad \theta(t) = e^{imt}$$

and where $m \in \mathbb{Z}$ is a constant, and with P_l^m being the Legendre function,

$$P_l^m(x) = (-1)^m (1-x^2)^{m/2} \left(\frac{d}{dx}\right)^m P_l(x)$$

where P_l are the Legendre polynomials, given by the following formula:

$$P_l(x) = \frac{1}{2^l l!} \left(\frac{d}{dx}\right)^l (x^2-1)^l$$

These solutions $\alpha = \sigma(s)\theta(t)$ are called spherical harmonics.

PROOF. This follows indeed from all the above, and with the comment that everything is taken up to linear combinations. We will normalize the wave function later. □

16d. The hydrogen atom

Hydrogen, eventually. In order now to finish our study, and eventually get to conclusions about hydrogen, it remains to solve the radial equation, for the Coulomb potential V of the proton. Let us begin with some generalities, valid for any time-independent potential V. As a first manipulation on the radial equation, we have:

PROPOSITION 16.23. *The radial equation, written with $K = l(l+1)$,*

$$(r^2\rho')' - \frac{2mr^2}{h^2}(V - E)\rho = l(l+1)\rho$$

takes with $\rho = u/r$ the following form, called modified radial equation,

$$Eu = -\frac{h^2}{2m} \cdot u'' + \left(V + \frac{h^2 l(l+1)}{2mr^2}\right) u$$

which is a time-independent 1D Schrödinger equation.

PROOF. With $\rho = u/r$ as in the statement, we have:

$$\rho = \frac{u}{r} \quad , \quad \rho' = \frac{u'r - u}{r^2} \quad , \quad (r^2\rho')' = u''r$$

By plugging this data into the radial equation, this becomes:

$$u''r - \frac{2mr}{h^2}(V - E)u = \frac{l(l+1)}{r} \cdot u$$

By multiplying everything by $h^2/(2mr)$, this latter equation becomes:

$$\frac{h^2}{2m} \cdot u'' - (V - E)u = \frac{h^2 l(l+1)}{2mr^2} \cdot u$$

But this gives the formula in the statement. As for the interpretation, as time-independent 1D Schrödinger equation, this is clear as well, and with the comment here that the term added to the potential V is some sort of centrifugal term. □

Getting back now to the Coulomb potential of the proton, we have here:

FACT 16.24. *The Coulomb potential of the hydrogen atom proton, acting on the electron by attraction, is given according to the Coulomb law by*

$$V = -\frac{Kep}{r}$$

where p is the charge of the proton, and K is the Coulomb constant. In practice however we have $p \simeq e$ up to order 10^{-7}, and so our formula can be written as

$$V \simeq -\frac{Ke^2}{r}$$

and we will use this latter formula, and with $=$ sign, for simplifying.

Getting back now to math, it remains to solve the modified radial equation, for the above potential V. And we have here the following result, which does not exactly solve this radial equation, but provides us instead with something far better, namely the proof of the original claim by Bohr, which was at the origin of everything:

THEOREM 16.25 (Schrödinger). *In the case of the hydrogen atom, where V is the Coulomb potential of the proton, the modified radial equation, which reads*

$$Eu = -\frac{h^2}{2m} \cdot u'' + \left(-\frac{Ke^2}{r} + \frac{h^2 l(l+1)}{2mr^2} \right) u$$

leads to the Bohr formula for allowed energies,

$$E_n = -\frac{m}{2} \left(\frac{Ke^2}{h} \right)^2 \cdot \frac{1}{n^2}$$

with $n \in \mathbb{N}$, the binding energy being

$$E_1 \simeq -2.177 \times 10^{-18}$$

with means $E_1 \simeq -13.591$ eV.

PROOF. This is again something non-trivial, and we will be following Griffiths [42], with some details missing. The idea is as follows:

(1) By dividing our modified radial equation by E, this becomes:

$$-\frac{h^2}{2mE} \cdot u'' = \left(1 + \frac{Ke^2}{Er} - \frac{h^2 l(l+1)}{2mEr^2} \right) u$$

In terms of $\alpha = \sqrt{-2mE}/h$, this equation takes the following form:

$$\frac{u''}{\alpha^2} = \left(1 + \frac{Ke^2}{Er} + \frac{l(l+1)}{(\alpha r)^2} \right) u$$

In terms of the new variable $p = \alpha r$, this latter equation reads:

$$u'' = \left(1 + \frac{\alpha Ke^2}{Ep} + \frac{l(l+1)}{p^2} \right) u$$

Now let us introduce a new constant S for our problem, as follows:

$$S = -\frac{\alpha Ke^2}{E}$$

In terms of this new constant, our equation reads:

$$u'' = \left(1 - \frac{S}{p} + \frac{l(l+1)}{p^2} \right) u$$

(2) The idea will be that of looking for a solution written as a power series, but before that, we must "peel off" the asymptotic behavior. Which is something that can be done,

of course, heuristically. With $p \to \infty$ we are led to $u'' = u$, and ignoring the solution $u = e^p$ which blows up, our approximate asymptotic solution is:

$$u \sim e^{-p}$$

Similarly, with $p \to 0$ we are led to $u'' = l(l+1)u/p^2$, and ignoring the solution $u = p^{-l}$ which blows up, our approximate asymptotic solution is:

$$u \sim p^{l+1}$$

(3) The above heuristic considerations suggest writing our function u as follows:

$$u = p^{l+1}e^{-p}v$$

So, let us do this. In terms of v, we have the following formula:

$$u' = p^l e^{-p}\left[(l+1-p)v + pv'\right]$$

Differentiating a second time gives the following formula:

$$u'' = p^l e^{-p}\left[\left(\frac{l(l+1)}{p} - 2l - 2 + p\right)v + 2(l+1-p)v' + pv''\right]$$

Thus the radial equation, as modified in (1) above, reads:

$$pv'' + 2(l+1-p)v' + (S - 2(l+1))v = 0$$

(4) We will be looking for a solution v appearing as a power series:

$$v = \sum_{j=0}^{\infty} c_j p^j$$

But our equation leads to the following recurrence formula for the coefficients:

$$c_{j+1} = \frac{2(j+l+1) - S}{(j+1)(j+2l+2)} \cdot c_j$$

(5) We are in principle done, but we still must check that, with this choice for the coefficients c_j, our solution v, or rather our solution u, does not blow up. And the whole point is here. Indeed, at $j >> 0$ our recurrence formula reads, approximately:

$$c_{j+1} \simeq \frac{2c_j}{j}$$

But, surprisingly, this leads to $v \simeq c_0 e^{2p}$, and so to $u \simeq c_0 p^{l+1}e^p$, which blows up.

(6) As a conclusion, the only possibility for u not to blow up is that where the series defining v terminates at some point. Thus, we must have for a certain index j:

$$2(j+l+1) = S$$

In other words, we must have, for a certain integer $n > l$:

$$S = 2n$$

(7) We are almost there. Recall from (1) above that S was defined as follows:

$$S = -\frac{\alpha K e^2}{E} \quad : \quad \alpha = \frac{\sqrt{-2mE}}{h}$$

Thus, we have the following formula for the square of S:

$$S^2 = \frac{\alpha^2 K^2 e^4}{E^2} = -\frac{2mE}{h^2} \cdot \frac{K^2 e^4}{E^2} = -\frac{2mK^2 e^4}{h^2 E}$$

Now by using the formula $S = 2n$ from (6), the energy E must be of the form:

$$E = -\frac{2mK^2 e^4}{h^2 S^2} = -\frac{mK^2 e^4}{2h^2 n^2}$$

Calling this energy E_n, depending on $n \in \mathbb{N}$, we have, as claimed:

$$E_n = -\frac{m}{2}\left(\frac{Ke^2}{h}\right)^2 \cdot \frac{1}{n^2}$$

(8) Thus, we proved the Bohr formula. Regarding numerics, the data is as follows:

$$K = 8.988 \times 10^9 \quad , \quad e = 1.602 \times 10^{-19}$$

$$h = 1.055 \times 10^{-34} \quad , \quad m = 9.109 \times 10^{-31}$$

But this gives the formula of E_1 in the statement. □

As a first remark, all this agrees with the Rydberg formula, due to:

THEOREM 16.26. *The Rydberg constant for hydrogen is given by*

$$R = -\frac{E_1}{h_0 c}$$

where E_1 is the Bohr binding energy, and the Rydberg formula itself, namely

$$\frac{1}{\lambda_{n_1 n_2}} = R\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

simply reads, via the energy formula in Theorem 16.25,

$$\frac{1}{\lambda_{n_1 n_2}} = \frac{E_{n_2} - E_{n_1}}{h_0 c}$$

which is in agreement with the Planck formula $E = h_0 c/\lambda$.

PROOF. Here the first assertion is something numeric, coming from the fact that the formula in the statement gives, when evaluated, the Rydberg constant:

$$R = \frac{-E_1}{h_0 c} = \frac{2.177 \times 10^{-18}}{6.626 \times 10^{-34} \times 2.998 \times 10^8} = 1.096 \times 10^7$$

As a consequence, and passed now what the experiments exactly say, we can define the Rydberg constant of hydrogen abstractly, by the following formula:

$$R = \frac{m}{2h_0c}\left(\frac{Ke^2}{h}\right)^2$$

Regarding now the second assertion, by dividing $R = -E_1/(h_0c)$ by any number of type n^2 we obtain, according to the energy convention in Theorem 16.25:

$$\frac{R}{n^2} = -\frac{E_n}{h_0c}$$

But these are exactly the numbers which are subject to substraction in the Rydberg formula, and so we are led to the conclusion in the statement. □

Let us go back now to our study of the Schrödinger equation. Our conclusions are:

THEOREM 16.27. *The wave functions of the hydrogen atom are the following functions, labelled by three quantum numbers, n, l, m,*

$$\phi_{nlm}(r, s, t) = \rho_{nl}(r)\alpha_l^m(s, t)$$

where $\rho_{nl}(r) = p^{l+1}e^{-p}v(p)/r$ with $p = \alpha r$ as before, with the coefficients of v subject to

$$c_{j+1} = \frac{2(j + l + 1 - n)}{(j + 1)(j + 2l + 2)} \cdot c_j$$

and $\alpha_l^m(s, t)$ being the spherical harmonics found before.

PROOF. This follows indeed by putting together all the results obtained so far, and with the remark that everything is up to the normalization of the wave function. □

In what regards the main wave function, that of the ground state, we have:

THEOREM 16.28. *With the hydrogen atom in its ground state, the wave function is*

$$\phi_{100}(r, s, t) = \frac{1}{\sqrt{\pi a^3}}\, e^{-r/a}$$

where $a = 1/\alpha$ is the inverse of the parameter appearing in our computations above,

$$\alpha = \frac{\sqrt{-2mE}}{h}$$

called Bohr radius of the hydrogen atom. This Bohr radius is the mean distance between the electron and the proton, in the ground state, and is given by the formula

$$a = \frac{h^2}{mKe^2}$$

which numerically means $a \simeq 5.291 \times 10^{-11}$.

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

(1) According to the various formulae in the proof of Theorem 16.25, taken at $n = 1$, the parameter α appearing in the computations there is given by:

$$\alpha = \frac{\sqrt{-2mE}}{h} = \frac{1}{h} \cdot m \cdot \frac{Ke^2}{h} = \frac{mKe^2}{h^2}$$

Thus, the inverse $\alpha = 1/a$ is indeed given by the formula in the statement.

(2) Regarding the wave function, according to Theorem 16.27 this consists of:

$$\rho_{10}(r) = \frac{2e^{-r/a}}{\sqrt{a^3}} \quad , \quad \alpha_0^0(s,t) = \frac{1}{2\sqrt{\pi}}$$

By making the product, we obtain the formula of ϕ_{100} in the statement.

(3) But this formula of ϕ_{100} shows in particular that the Bohr radius a is indeed the mean distance between the electron and the proton, in the ground state.

(4) Finally, in what regards the numerics, these are as follows:

$$a = \frac{1.055^2 \times 10^{-68}}{9.109 \times 10^{-31} \times 8.988 \times 10^9 \times 1.602^2 \times 10^{-38}} = 5.297 \times 10^{-11}$$

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

Getting back now to the general setting of Theorem 16.25, the point is that the polynomials $v(p)$ appearing there are well-known objects in mathematics, as follows:

PROPOSITION 16.29. *The polynomials $v(p)$ are given by the formula*

$$v(p) = L_{n-l-1}^{2l+1}(p)$$

where the polynomials on the right, called associated Laguerre polynomials, are given by

$$L_q^p(x) = (-1)^p \left(\frac{d}{dx}\right)^p L_{p+q}(x)$$

with L_{p+q} being the Laguerre polynomials, given by the following formula:

$$L_q(x) = \frac{e^x}{q!} \left(\frac{d}{dx}\right)^q (e^{-x}x^q)$$

PROOF. The story here is very similar to that of the Legendre polynomials. Consider the Hilbert space $H = L^2[0,\infty)$, with the following scalar product on it:

$$< f, g >= \int_0^\infty f(x)g(x)e^{-x}\,dx$$

(1) The orthogonal basis obtained by applying Gram-Schmidt to the Weierstrass basis $\{x^q\}$ is then the basis formed by the Laguerre polynomials $\{L_q\}$.

(2) We have the explicit formula for L_q in the statement, which is analogous to the Rodrigues formula for the Legendre polynomials.

(3) The first assertion follows from the fact that the coefficients of the associated Laguerre polynomials satisfy the equation for the coefficients of $v(p)$.

(4) Alternatively, the first assertion follows as well by using an equation for the Laguerre polynomials, which is very similar to the Legendre equation. □

With the above result in hand, we can now improve Theorem 16.25, as follows:

THEOREM 16.30. *The wave functions of the hydrogen atom are given by*

$$\phi_{nlm}(r,s,t)=\sqrt{\left(\frac{2}{na}\right)^3\frac{(n-l-1)!}{2n(n+l)!}}e^{-r/na}\left(\frac{2r}{na}\right)^l L_{n-l-1}^{2l+1}\left(\frac{2r}{na}\right)\alpha_l^m(s,t)$$

with $\alpha_l^m(s,t)$ being the spherical harmonics found before.

PROOF. This follows indeed by putting together what we have, namely Theorem 16.25 and Proposition 16.29, and then doing some remaining work, concerning the normalization of the wave function, which leads to the normalization factor appearing above. □

And good news, that is all. The above formula is all you need, in everyday life.

16e. Exercises

Congratulations for having read this book, and no exercises for this final chapter. However, if you enjoyed this book, and looking for more to read, have a look at the various books referenced below. Normally these are all good books, which all sort of go to the point, without bothering much with annoying details, and normally you should like them too. So, have a look at them, and start with the one that you like the most.

Before everything, however, learn more calculus, and especially do many exercises, as many as needed, first in order to be at ease with calculus, and then in order to really love calculus. In fact, and in the hope that you got it, the present book was just an introduction to calculus, not calculus itself. Just a beginning.

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, Linear algebra and group theory (2023).

[10] T. Banica, Principles of operator algebras (2023).

[11] T. Banica, Introduction to modern physics (2023).

[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, CBMS Lecture Notes (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] B.M. Peterson and B. Ryden, Foundations of astrophysics, Cambridge Univ. Press (2010).

[74] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[75] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[76] W. Rudin, Functional analysis, McGraw-Hill (1973).

[77] W. Rudin, Fourier analysis on groups, Dover (1974).

[78] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[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

abelian group, 166
absolute value, 289
adjoint map, 204
adjoint matrix, 204
adjoint operator, 374
algebraic curve, 249
algebraic manifold, 261
alternating series, 27, 29
analytic function, 140, 146
angular equation, 380
arctan, 63
area, 81
area of sphere, 312
argument of complex number, 110
atom, 200
average of function, 84
azimuthal equation, 381

barycenter, 127
Bell numbers, 327
Bernoulli laws, 176, 327
binomial coefficient, 11
binomial coefficients, 308
binomial formula, 13, 48, 74
Bohr formula, 387
Bohr radius, 390
boundary of domain, 141
bounded operator, 372
bounded sequence, 23

cartography, 271
Catalan numbers, 49, 75
Cauchy formula, 143, 144, 146
Cauchy sequence, 25, 129
Cauchy-Schwarz, 71
CCLT, 333
central binomial coefficients, 49, 75, 308
chain rule, 61, 97, 245
change of variable, 97, 300, 301
characteristic polynomial, 221, 222
charge, 359
chart, 268
closed and bounded, 41, 235
closed set, 39, 40
colored integers, 333
colored moments, 333
column expansion, 216
common roots, 122
compact set, 41, 235
compact support, 155
complete space, 25, 129
Complex CLT, 333
complex conjugate, 113, 138, 184
complex eigenvalues, 223
complex function, 116, 137
complex matrix, 219
complex number, 107, 109
complex plane, 129
complex roots, 121
concave function, 70
conic, 249, 255
connected set, 41, 137, 236
continuous function, 33, 39, 116, 237
continuously differentiable, 241
convergent sequence, 22, 129
convergent series, 25
convex function, 70
convolution, 155, 173, 322
convolution exponential, 174, 326
convolution semigroup, 326
coordinates, 268

cos, 21, 35, 59, 76, 133
cosh, 136
cover, 41
curl, 365
cyclic group, 167

decimal writing, 18
decreasing sequence, 23
Dedekind cut, 16
degree 2 equation, 17, 108, 113, 249
density, 223
derivative, 57, 58, 238
determinant, 211
determinant formula, 219
determinant of products, 214
deterrminant, 219
diagonalizable matrix, 214, 223, 287
diagonalization, 220, 222
differentiable function, 57, 137, 238
differential equation, 78
differential manifold, 268
discriminant, 123, 223
distance, 47, 129, 225
distribution, 147, 321
divergence, 365
double derivatives, 177
double factorial, 305
double factorials, 304, 309, 315
double root, 123

e, 29, 53
eigenvalue, 220, 222
eigenvector, 220, 222
electric field, 359
electron, 376
ellipsis, 250
equal almost everywhere, 154
exchange of hat, 161
exp, 59, 76, 133
expectation, 86
exponential, 53, 60, 116

factorials, 11
Fano plane, 266
field, 130
field lines, 360
finite field, 265
flux, 361, 362, 364

Fourier inversion, 165
Fourier transform, 158, 174, 323, 326
fraction, 62
fractions, 11
free group, 267
function, 33
function space, 154
fundamental theorem of calculus, 93, 94

Gauss formula, 366
Gauss law, 364
generalized binomial formula, 48, 74
generalized binomial numbers, 48
geometric series, 26, 117, 132
gradient, 365
gradient method, 296
Gram-Schmidt, 372
gravity, 78, 255
Green formula, 366
ground state, 390

Hölder inequality, 71, 153
harmonic function, 183, 290
heat equation, 79, 181
Hessian, 278
higher derivative, 98, 275
higher derivatives, 72
Hilbert space, 370
hole, 267
holomorphic function, 137, 146
homogeneous polynomial, 184
homotopy group, 267
Hooke law, 78, 179
hyperbola, 250

i, 107
increasing sequence, 23
independence, 173, 174, 322, 323
infinite matrix, 372
infinitely differentiable, 138, 146
infinitesimal, 93
inflation coefficient, 213
integrable function, 82
integral, 81
integral over curve, 143
integration by parts, 96
intermediate value, 42, 137
inverse function, 43

inversion formula, 207
invertible matrix, 206, 208, 212

Jacobian, 243, 301, 362
Jensen inequality, 70

Kepler laws, 255
Klein bottle, 262
knot, 268

L'Hôpital's rule, 68, 72
Lagrange multipliers, 295
Laplace equation, 189
Laplace operator, 179, 181, 183, 290, 377
lattice model, 78, 79, 179
law, 147, 321
Legendre function, 385
lim inf, 25
lim sup, 25
limit of continuous functions, 46
limit of sequence, 22
limit of series, 25
linear function, 203
linear map, 203
linear operator, 372
Liouville theorem, 146, 188, 290
local maximum, 64, 69, 278
local minimum, 64, 69, 278
locally affine, 58
log, 59, 76, 133
loop, 267
lower triangular matrix, 215

Möbius strip, 262
main value formula, 142, 146
matching pairings, 335
matrix, 203
maximum, 42, 64, 278
maximum modulus, 188, 290
maximum principle, 141, 146
Maxwell equations, 191
mean, 101
mean value formula, 188, 290
mean value property, 89
metric space, 225
minimum, 42, 64, 278
Minkowski inequality, 153
modulus, 57, 138
modulus of complex number, 110
modulus of matrix, 289
moment, 101
moments, 147, 321
monotone function, 42
Monte Carlo integration, 84
multiplication of complex numbers, 120

Nash theorem, 272
Netwon law, 255
Newton law, 78, 179
normal matrix, 287
normed space, 154
number of inversions, 217

observable, 376
open set, 39, 40
oriented system of vectors, 210
orthonormal basis, 372

p-norm, 154
parabola, 250
parallelogram rule, 109
partial derivatives, 239
partial isometry, 289
partitions, 327
Pascal triangle, 14
passage matrix, 220
permutation, 217
pi, 20
piecewise continuous, 87
piecewise linear, 82
piecewise monotone, 87
Planck constant, 376
Planck formula, 389
PLT, 176, 327
point charge, 362
pointwise convergence, 45, 46, 130
Poisson law, 325
Poisson Limit Theorem, 176, 327
polar coordinates, 110, 120, 269, 302
polar decomposition, 289
polar equation, 381
polar writing, 119
polarization formula, 205
pole, 130
polynomial, 73, 130
positive eigenvalues, 284

positive matrix, 278, 284
power function, 43, 58, 136
powers of complex number, 111
probability 0, 19
probability density, 101
probability space, 147, 321
product of cyclic groups, 167
product of eigenvalues, 214
projection, 204, 205, 283
projective space, 262
Pythagoras' theorem, 21

radial equation, 380, 386
radial function, 189
radial limit, 139
random number, 84
random variable, 86, 147, 321
rank 1 projection, 264
rational function, 130, 138
real number, 16
rectangular matrix, 203
reflection, 113
remainder, 98
resultant, 122, 223
Riemann integration, 83
Riemann series, 26
Riemann sum, 90, 306
Riemann-Lebesgue property, 164
right-hand rule, 346
roots of polynomial, 43, 121, 130
roots of unity, 126, 127
rotation, 203, 223
row expansion, 215
Rydberg formula, 389

Sarrus formula, 216, 219
scalar product, 205, 370
Schrödinger equation, 376
second derivative, 68
second order derivative, 278
self-adjoint matrix, 281
self-dual group, 167
separable space, 372
separation of variables, 380, 385
sequence, 22
sequence of functions, 45
series, 25
sign of system of vectors, 210
signature, 217
signed volume, 211
sin, 21, 35, 59, 76, 133
single roots, 123
sinh, 136
smooth manifold, 268
spectral theorem, 287
spherical coordinates, 269, 303, 377
spherical harmonics, 385, 390, 392
spherical integral, 315, 317, 319
spiral, 117, 132
square root, 16, 17, 43, 49, 75, 108, 113, 284
step function, 45
stereographic projection, 270
Stirling formula, 306
Stokes formula, 366
strictly positive matrix, 284
subcover, 41
subsequence, 23, 25
sum of vectors, 109
support, 155
symmetric group, 217
symmetry, 203

tan, 63
Taylor formula, 68, 72, 73, 76, 98
time-independent equation, 380
totally discontinuous, 45
transposition, 217
trigonometric integral, 304, 313

uniform convergence, 46, 130
union of intervals, 40
unitary, 205, 285
unoriented system of vectors, 210
upper triangular matrix, 215

variance, 101
vector, 109
vector product, 346
volume, 81
volume inflation, 243
volume of parallelepiped, 207
volume of sphere, 305, 306, 311

wave equation, 78, 179
wave function, 376