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, in one or several variables, real or complex, and its applications to basic questions from physics. We first discuss the theory of functions $f : \mathbb{R} \to \mathbb{R}$, notably with the notion of continuity, and with the construction and main properties 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 of the holomorphic functions, and harmonic functions. Then, we discuss 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 the general theory (notably change of variable formula, coming with full proof), applications to probability (Gaussian laws), advanced integration results (Green, Stokes, Gauss), maximization questions (Hessian, Lagrange multipliers), and basic applications to physics (ODE and PDE). Finally, we discuss a number of more advanced topics, such as Fourier analysis and waves.

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 [67], [68] 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 [60], [61].

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.

We will first discuss the theory of functions $f : \mathbb{R} \to \mathbb{R}$, with the construction and main properties of the derivative $f'(x)$ and of the integral $\int_a^b f(x)dx$. Then we will investigate the case of the complex functions $f : \mathbb{C} \to \mathbb{C}$, and notably of the holomorphic functions, and harmonic functions. Then, we will discuss 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 the general theory (notably change of variable formula, coming with full proof), applications to probability (Gaussian laws), advanced integration results (Green, Stokes, Gauss), maximization questions (Hessian, Lagrange multipliers), and basic applications to physics (ODE and PDE). Finally, we will discuss a number of more advanced topics, such as Fourier analysis and waves.

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

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. □

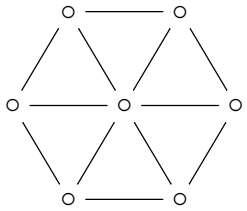
As a comment here, all the above is of course quite tricky, and a bit bordeline in respect to what can be called "rigorous mathematics". But we will be back to this, namely general probability theory, and in particular meaning of the mysterious formula $P = 0$, countable sets, infinite sums and so on,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

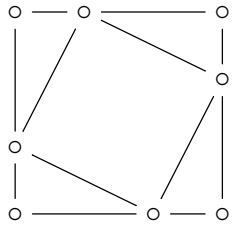
THEOREM 1.15. *The following happen:*

(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.

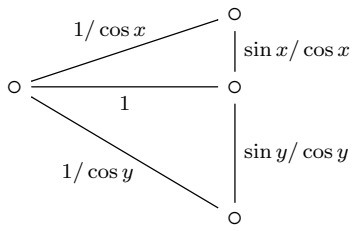
(2) The statement here, which generalizes (1), follows exactly as (1), by using the various findings from Theorem 2.2.

(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.

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 2.4. *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 2.5. *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.

"$\Longleftarrow$" Assume that $C \subset \mathbb{R}$ is open, and let $O = \mathbb{R} - C$. In order to prove that O is open, let $x \in O$, and consider the intervals $(x - 1/n, x + 1/n)$, with $n \in \mathbb{N}$. If one of these intervals lies in O, we are done. Otherwise, this would mean that for any $n \in \mathbb{N}$ we have at least one point $x_n \in (x - 1/n, x + 1/n)$ satisfying $x_n \notin O$, and so $x_n \in C$. But since C is closed and $x_n \to x$, we get $x \in C$, and so $x \notin O$, contradiction, and we are done. □

As a basic 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 2.6. *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 2.6, we will need to know more about open and closed sets. Let us begin with a useful result, as follows:

PROPOSITION 2.7. *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 2.8. *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 2.7 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 2.5. □

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 2.9. *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 intervals $[a,b]$ with $a,b\neq\pm\infty$ are compact, 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 is obviously impossible to come up with more examples. In fact, we have the following result:

THEOREM 2.10. *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, it is clear that any closed interval $[a,b]$ with $a,b\neq\pm\infty$ 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 2.9 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 2.11. *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 2.9. 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. □

You might perhaps ask at this point, was Theorem 2.11 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 2.12. *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, $f(c)=0$ has a solution.*
(4) *There is $c\in[a,b]$ such that $f(c)=(f(b)-f(a))/(b-a)$.*

PROOF. All these statements are related, and are called altogether "intemediate value theorem". Regarding now the proof, one way of viewing things is that since $[a,b]$ is

connected, then $f([a,b])$ is connected too, and this gives the results. However, this is based on rather advanced technology, and we can prove (1-4) directly as well. □

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

THEOREM 2.13. *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 2.12. 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 2.13, we have:

PROPOSITION 2.14. *The various usual inverse functions, such as the inverse trigonometric functions* arcsin*,* arccos*,* arctan*,* arccot*, are all continuous.*

PROOF. This follows indeed from Theorem 2.13, 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 2.15. *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 2.12, as follows:

(1) This is clear from Theorem 2.12 (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 2.12 (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 2.12. We will be back to this fundamental question of analysis on numerous occasions, in what follows.

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

THEOREM 2.16. *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 2.15, 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 2.15:

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

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 2.15. 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 2.17. *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 2.16, by exchanging the variables, $x \leftrightarrow a$. To be more precise, the situation is as follows:

(1) We know from Theorem 2.16 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 2.16 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

We can talk about the convergence of sequences of functions $f_n \to f$, with a main result here being the fact that if the functions f_n are continuous, and the convergence is uniform, is some suitable sense, then the limit f is continuous too.

Importantly, the uniformity asumption is really needed, because it is easy to see on pictures that discontinuous functions can be approximated by continuous functions.

Regarding now series, we have again some general theory to be developed here, 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 2.18. *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 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 2.18 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 $a=\pm1/2$, to start with, we have the following result:

THEOREM 2.19. *The generalized binomial formula, namely*

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

holds as well at $a=\pm1/2$. 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:

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

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

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

(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:

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

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

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

(4) Let us first prove the formula on the right. By using the geometric series formula for $1/(1-4t)$, we are led into proving the following formula:

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

But this is something standard, by doing some combinatorics with the central binomial coefficients $D_k = \binom{2k}{k}$. As for the formula on the left, involving the Catalan numbers $C_k = \frac{1}{k+1}\binom{2k}{k}$, the proof here is similar, by doing some combinatorics. □

As a next step, we can unify Theorem 2.18 and Theorem 2.19, as follows:

THEOREM 2.20. *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 $a \in \mathbb{Z}/2$, and any $|x| < 1$.

PROOF. This is something quite trickly, which can be done as follows:

(1) As in the proof of Theorem 2.19, let us start by first computing the binomial coefficients at $a \in \mathbb{Z} + 1/2$, whose knowledge is useful. For $n \in \mathbb{N}$ we have:

$$\begin{aligned}\binom{n-1/2}{k} &= \frac{(n-1/2)(n-3/2)\ldots(n-k+1/2)}{k!}\\ &= \frac{(2n-1)(2n-3)\ldots(2n-2k+1)}{2^k k!}\end{aligned}$$

We can see that we have here two cases, $k \leq n$ and $k > n$.

(2) Assume first $k \leq n$. We have then the following computation:

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

(3) Assume now $k > n$. We have then the following computation:

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

(4) Summarizing, we have computed so far the binonial coeffients at $a = n-1/2$, with $n \in \mathbb{N}$. Observe that we can write a partly uniform formula, as follows:

$$\binom{n-1/2}{k} = 4^{-k}\cdot\frac{(2n)!}{n!k!}\times\begin{cases}\frac{(n-k)!}{(2n-2k)!} & (k\leq n)\\ (-1)^{n-k}\frac{(2k-2n)!}{(k-n)!} & (k>n)\end{cases}$$

(5) In order to prove now that the binomial formula holds indeed at $a = n-1/2$, we can use our previous knowledge at $n, -1/2$, along with the following formula:

$$(1+t)^{n-1/2} = (1+t)^n(1+t)^{-1/2}$$

We are led in this way to the result, after some combinatorics.

(6) It remains to discuss the case $a = -n-1/2$, with $n \in \mathbb{N}$. Again, let us begin by computing the corresponding binomial coefficients. We have:

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

(7) In order to prove now that the binomial formula holds indeed at $a = -n-1/2$, we can use our previous knowledge at $-n, -1/2$, along with the following formula:

$$(1+t)^{-n-1/2} = (1+t)^{-n}(1+t)^{-1/2}$$

We are led in this way to the result, after some combinatorics. □

As already mentioned, the binomial formula holds in fact for any exponent $a \in \mathbb{R}$, but this is non-trivial, and the elementary study stops with Theorem 2.20. Indeed, the

next step, before getting into arbitrary rationals $a \in \mathbb{Q}$, which would solve in fact the problem for any $a \in \mathbb{R}$, by continuity, would be that of looking at general exponents of type $a = p/2$, with $p \in \mathbb{N}$. But here we are in trouble, because we have:

$$\begin{aligned}\binom{1/p}{k} &= \frac{1/p(1/p-1)\ldots(1/p-k+1)}{k!}\\ &= (-1)^{k-1}\frac{(p-1)(2p-1)\ldots((k-1)p-1)}{p^k k!}\end{aligned}$$

Thus, we are no longer dealing here with binomial coeffients, or products of binomial coefficients, and the combinatorics can only be quite complicated, not to say extreme, and we will stop here. However, we will see in chapter 3 below that the problem can be solved, and in a very elegant way, but 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 2.21. *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, 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 2.22. *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 $x = 1$, some combinatorics and analysis show indeed that we have the following equality, between the sum of a series, and a limit of a sequence:

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

Thus, we obtain in this way the result at $x = 1$.

(2) 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.

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

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

But this can be checked indeed, by doing some combinatorics.

(4) 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

Exercises.

CHAPTER 3

Derivatives

3a. Derivatives, rules

We have the following result, which is the starting point for everything in analysis:

THEOREM 3.1. *Any function of one variable $f : \mathbb{R} \to \mathbb{R}$ is locally affine,*

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

with $f'(x) \in \mathbb{R}$ being the derivative of f at the point x, given by

$$f'(x) = \lim_{t \to 0} \frac{f(x+t) - f(x)}{t}$$

provided that this latter limit converges indeed.

PROOF. Assume indeed that the limit in the statement converges. 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. □

As an illustration, the derivatives of the power functions are as follows:

PROPOSITION 3.2. *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:

$$a^n - b^n = (a-b)(a^{n-1} + a^{n-2}b + \ldots + b^{n-1})$$

To be more precise, we will use this formula, written as follows:

$$a-b=\frac{a^n-b^n}{a^{n-1}+a^{n-2}b+\ldots+b^{n-1}}$$

We set in this formula $a=(x+t)^{m/n}$ and $b=x^{m/n}$. We obtain in this way, by using the binomial formula for approximating on top, as desired:

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

(3) In the general case now, where $p\in\mathbb{R}$ is real, the same formula holds, namely $(x^p)'=px^{p-1}$, by using what we found above, and a continuity argument. □

There are many applications of the derivative, and we have for instance:

PROPOSITION 3.3. *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 is clear from the formula in Theorem 3.1, namely:

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

As for the converse, the simplest counterexample is $f(x)=x^3$, at $x=0$. □

3b. Second derivatives

At a more advanced level now, we have the following result:

THEOREM 3.4. *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 3.1 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 3.1. 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 3.1, and there are many applications of this. We can improve for instance Proposition 3.3, as follows:

PROPOSITION 3.5. *The local minima and 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 3.4, 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. □

We have the following result:

THEOREM 3.6. *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 something quite routine. □

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

THEOREM 3.7. *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 is again something quite routine. □

Observe that at $p = 2$ we obtain the Cauchy-Schwarz inequality.

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 3.8. *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 3.4. 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. □

While fully proved in the above, the Taylor formula remains something quite mysterious. Here is a related interesting statement, inspired from the proof:

PROPOSITION 3.9. *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. □

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 3.10. *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, with $|t| < 1$, one proof being by multiplying everything by $(1+t)^m$, expanded by using the usual binomial formula. Indeed, by working out the combinatorics, on the right side, we eventually obtain 1.

(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 $|t| < 1$. $\square$

We can see from the above the power of the Taylor formula. As an application now of our generalized binomial formula, we can extract square roots, as follows:

PROPOSITION 3.11. *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. The above formulae both follow from Theorem 3.10, as follows:

(1) At $p = 1/2$, the generalized binomial coefficients are:

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

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

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

Thus, we obtain the formulae in the statement. □

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

THEOREM 3.12. *We have the following formulae,*

$$\sin t = \sum_l (-1)^l \frac{t^{2l+1}}{(2l+1)!} \quad , \quad \cos t = \sum_l (-1)^l \frac{t^{2l}}{(2l)!}$$

as well as the following formulae,

$$e^x = \sum_k \frac{x^k}{k!} \quad , \quad \log(1+y) = \sum_k (-1)^{k+1} \frac{y^k}{k}$$

as Taylor series, and in general as well, with $|y| < 1$ *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 well-known formulae:

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

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

With these formulae in hand we can appproximate both sin and cos, and we get:

$$(\sin t)' = \cos t$$

$$(\cos t)' = -\sin t$$

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

Basic differential equations, 1D gravity.

3e. Exercises

CHAPTER 4

Integration

4a. Integration theory

In this chapter we discuss an important topic, namely the integration of the functions $f : \mathbb{R} \to \mathbb{R}$. 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 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 mention that for basic functions like the continuous ones, or the increasing ones, this area can be indeed computed, and so our formalism is quite general. More on this later.

In order to compute areas, and so integrals of functions, we can use of course our geometric knowledge. Here are some basic results of this type:

PROPOSITION 4.2. *We have the following results:*

(1) *When f is linear, we have $\int_a^b f(x)dx = (b-a)(f(a)+f(b))/2$.*
(2) *A similar formula holds for the piecewise linear functions.*
(3) *We have $\int_{-1}^1 \sqrt{1-x^2}\,dx = \pi/2$.*

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, by additivity. It is possible of course to write down a general formula here, but let us not bother with that, this being not that useful anyway.

(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 $\pi/2$. □

Here are as well some general results regarding the integrals:

PROPOSITION 4.3. *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 4.4. *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 4.5. *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 4.4. □

The point of view in Theorem 4.5 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 4.2, 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 4.5. The same goes for the results in Proposition 4.3, which become even more obvious with the viewpoint from Theorem 4.5. 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 4.2 (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 4.6. *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 4.4 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 $N\to N+1$. Also, the Monte Carlo integration works better for functions having various symmetries.

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

THEOREM 4.7. *The integral of a function $f:[a,b]\to\mathbb{R}$ is given by*

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

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

PROOF. This is some sort of fancy reformulation of Theorem 4.5 above. □

Summarizing, we have so far a deep knowledge of what the integral is, philosophically speaking, but unfortunately, not many concrete results about it.

4b. Riemann sums

Our purpose now will be to understand which functions are integrable, and how to compute their integrals. For this purpose, the Riemann formula in Theorem 4.4 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 4.8. *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. □

Going ahead with more theory, we have:

THEOREM 4.9. *Behavior of integrals with respect to:*

(1) *Limits.*
(2) *Infinite sums.*

PROOF. This is again something quite standard, by using the theory developed in chapter 3 above for the sequences and series of functions. □

At the level of examples now, we have:

THEOREM 4.10. *We have the formula*

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

valid for any exponent $k \in \mathbb{R} - \{1\}$.

PROOF. This is something quite tricky. At $k = 0, 1, 2, 3$ the Riemann sums can be explicitely computed, as explained below, with explicit formulae, and this gives the result. In general, however, this is something non-trivial. □

As a basic application, we can sum powers, with exact results at small exponents, and with estimates in general. We have indeed the following result:

THEOREM 4.11. *The basic Riemann sums, namely $1^k + 2^k + \ldots + N^k$, depending on an exponent $k \in \mathbb{N}$, are as follows:*

(1) *At $k = 1$ we have $1 + 2 + \ldots + N = \frac{N(N+1)}{2}$.*

(2) *At $k = 2$ we have $1^2 + 2^2 + \ldots + N^2 = \frac{N(N+1)(2N+1)}{6}$.*

(3) *At $k = 3$ we have $1^3 + 2^3 + \ldots + N^3 = \left[\frac{N(N+1)}{2}\right]^2$.*

(4) *At $k \in \mathbb{N}$ we have $1^k + 2^k + \ldots + N^k \simeq \frac{N^{k+1}}{k+1}$.*

PROOF. This is something well-known, the idea being as follows:

(1) The average of $1, 2, \ldots, N$ being the number in the middle, we have:

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

Thus, we obtain the formula in the statement, without any computation.

(2) A well-known proof here is by drawing certain squares in the plane.

(3) A well-known proof here is by drawing certain cubes in the space.

(4) Here there is no trick, and we must use Riemann integration. We have:

$$\int_0^1 x^k dx = \frac{1}{k + 1}$$

But this gives the estimate in the statement, by using a Riemann sum. □

4c. Basic results

Getting back now to the basics, the formula in Theorem 4.10 suggests that integration might have something to do with differentiation. And fortunately, this is indeed the case, as shown by the following result, called fundamental theorem of calculus:

THEOREM 4.12. *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. This follows indeed from the Riemann integration picture. □

There are many other equivalent formulations of the fundamental theorem of calculus, and countless applications as well. Staying theoretical, given a function f, we can call primitive of f any function F satisfying $F' = f$. The primitives are unique up to an additive constant, and are denoted $\int f$, up to this indeterminacy.

With this convention, we have the following result:

THEOREM 4.13. *We have the formula*

$$\int f'g + \int fg' = fg$$

called integration by parts.

PROOF. This follows indeed by integrating the Leibnitz formula, namely:

$$(fg)' = f'g + fg'$$

Once this known, it is of course possible to pass to usual, definite integrals, and we obtain a formula here as well, as follows, also called integration by parts:

$$\int_a^b f'g + \int_a^b fg' = \Big[fg\Big]_a^b$$

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. □

We have as well the following result:

THEOREM 4.14. *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. □

As a main application now, we have:

THEOREM 4.15. *Taylor formula with integral formula for the remainder.*

PROOF. This is something very standard, by using the fact that integrals of functions are weighted averages. □

There are of course many concrete applications of all the above.

4d. Some probability

We discuss here the alternative point of view on integration coming from Theorem 4.7. Among others, at the theoretical level, this will lead us into integrating functions with respect to other measures than the usual, uniform one.

4e. Exercises

Exercises.

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 $f : \mathbb{C} \to \mathbb{C}$. 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*

$$x = a + ib$$

which add in the obvious way, and multiply according to the following rule:

$$i^2 = -1$$

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 key 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:

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

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.

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. □

Observe that in the above picture, the real numbers correspond to the numbers on the Ox axis. As for the purely imaginary numbers, these lie on the Oy axis, with:

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

All this is very nice, but in order to understand now the multiplication, we must do something more complicated, namely using polar coordinates. Let us start with:

DEFINITION 5.4. *The complex numbers $x=a+ib$ can be written in polar coordinates,*

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

with the connecting formulae being

$$a=r\cos t$$

$$b=r\sin t$$

and in the other sense being

$$r=\sqrt{a^2+b^2}$$

$$\tan t=b/a$$

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.

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 5.5. *Two complex numbers written in polar coordinates,*

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

$$y=p(\cos t+i\sin t)$$

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 well-known trigonometry formulae, which can be proved by doing some plane geometry:

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

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

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 5.6. *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.

5b. Exponential writing

Let us discuss now the final and most convenient writing of the complex numbers, which is a well-known variation on the polar writing, as follows:

$$x = re^{it}$$

In what follows we will not really need the true power of this formula, which is of analytic nature, due to occurrence of the number e. However, we would like to use the notation $x = re^{it}$, as everyone does, among others because it simplifies the writing.

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

THEOREM 5.7. *We have the following formula, valid for any $t \in \mathbb{R}$,*

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

where $e = 2.7182\ldots$ is the usual constant from analysis.

PROOF. In order to prove such a result, we must first recall what e is, and what e^x is. One way of viewing things is that e^x is the unique function satisfying:

$$(e^x)' = e^x \quad , \quad e^0 = 1$$

Then, we can set $e = e^1$, and then prove that e^x equals indeed e to the power x. This is a bit abstract, but is convenient for our purposes. Indeed, the solution to the above derivative problem is easy to work out, as a series, by recurrence, the answer being:

$$e^x = \sum_k \frac{x^k}{k!}$$

Now let us plug $x = it$ in this formula. We obtain the following formula:

$$\begin{aligned} e^{it} &= \sum_k \frac{(it)^k}{k!} \\ &= \sum_{k=2l} \frac{(it)^k}{k!} + \sum_{k=2l+1} \frac{(it)^k}{k!} \\ &= \sum_l (-1)^l \frac{t^{2l}}{(2l)!} + i \sum_l (-1)^l \frac{t^{2l+1}}{(2l+1)!} \end{aligned}$$

Our claim now, which will complete the proof, is that we have:

$$\cos t = \sum_l (-1)^l \frac{t^{2l}}{(2l)!}$$

$$\sin t = \sum_l (-1)^l \frac{t^{2l+1}}{(2l+1)!}$$

In order to prove this claim, let us compute the Taylor series of cos and sin. By using the formulae for sums of angles, used in the proof of Theorem 5.5, we have:

$$\sin' = \cos \quad , \quad \cos' = -\sin$$

Thus, we know how to differentiate sin and cos, once, then twice, then as many times as we want to, and with this we can compute the corresponding Taylor series, the answers being those given above. Now by putting everything together, we have:

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

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

All this was quite brief, but we will be back to it with full details in chapters 5-8 below, when doing analysis. For the moment, let us just enjoy all this. We first have:

THEOREM 5.8. *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 5.7, 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 [28], [29], [30]. □

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 5.9. *The complex numbers $x = a + ib$ can be written in polar coordinates,*

$$x = re^{it}$$

with the connecting formulae being

$$a = r\cos t$$

$$b = r\sin t$$

and in the other sense being

$$r = \sqrt{a^2 + b^2}$$

$$\tan t = b/a$$

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

PROOF. This is a reformulation of Definition 5.4, by using the formula $e^{it} = \cos t + i\sin t$ from Theorem 5.7, and multiplying everything by r. □

We can now go back to the basics, namely the addition and multiplication of the complex numbers, and we have the following result:

THEOREM 5.10. *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 5.5, reformulated by using the notations from Theorem 5.9. Observe that this follows as well directly, from the fact that we have $e^{a+b} = e^a e^b$, 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 now investigate more complicated operations, as follows:

THEOREM 5.11. *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:* $(re^{it})^a = r^a e^{ita}$.

PROOF. This is something that we already know, from Theorem 5.6, 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 5.10:

$$\begin{aligned}(re^{it})(r^{-1}e^{-it}) &= rr^{-1}\cdot e^{i(t-t)}\\ &= 1\cdot 1\\ &= 1\end{aligned}$$

(2) Once again by using Theorem 5.10, we have:

$$\begin{aligned}(\pm\sqrt{r}e^{it/2})^2 &= (\sqrt{r})^2e^{i(t/2+t/2)}\\ &= re^{it}\end{aligned}$$

(3) Given an arbitrary number $a\in\mathbb{R}$, we can define, as stated:

$$(re^{it})^a=r^ae^{ita}$$

Due to Theorem 5.10, this operation $x\to x^a$ is indeed the correct one. □

5c. Equations, roots

With the above results in hand, and notably with the square root formula from Theorem 5.11 (2), we can now go back to the degree 2 equations, and we have:

THEOREM 5.12. *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 Theorem 5.11 (2) that any complex number has a square root, we are led to the conclusion in the statement. □

More generally now, we can prove that any polynomial equation, of arbitrary degree $N\in\mathbb{N}$, has exactly N complex solutions, counted with multiplicities:

THEOREM 5.13. *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.

5d. Roots of unity

As a last topic regarding the complex numbers, we have:

THEOREM 5.14. *The equation $x^N = 1$ has N complex solutions, namely*

$$\left\{ w^k \middle| k = 0, 1, \ldots, N-1 \right\} \quad , \quad w = e^{2\pi i/N}$$

which are called roots of unity of order N.

PROOF. This follows from Theorem 5.10. Indeed, with $x = re^{it}$ our equation reads:

$$r^N e^{itN} = 1$$

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 $w=e^{2\pi i/3}$.

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 5.6 and Theorem 5.11, the statement being as follows:

THEOREM 5.15. *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, to be used many times in what follows:

THEOREM 5.16. *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. □

This was for our basic presentation of the theory of complex numbers. We will be back to more things regarding the complex numbers, and the roots of unity, later on.

5e. Exercises

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. Before starting, two remarks:

(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 6.1. *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, when talking about continuity of real functions of several variables.

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

PROPOSITION 6.2. *The basic functions that we know, namely* x^a*, and* sin*,* cos*,* exp*,* log *have suitable complex extensions, which are continuous.*

PROOF. This is quite clear, as in the real case, by using the various formulae for the complex numbers that we learned in chapter 5. To be more precise, we can set:

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

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

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

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

There are however a number of subtleties to be discussed, in relation with the powers x^a, and with the logarithm $\log x$, which cannot be defined everywhere. We will be actually back to these two issues, a bit later, when we will have more tools. □

Let us record as well the following result, that we will be of of interest later:

PROPOSITION 6.3. *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 of Q*, called poles of* f*.*

PROOF. This is again clear from definitions, and from Proposition 6.2. We will be back to rational functions later, when we will have more tools for studying them. □

Let us point out now 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:

PROPOSITION 6.4. *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. □

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 intervals of complex numbers. However, the intermediate value theorem in its advanced formulation, that with connected sets, extends of course, and we have the following result:

THEOREM 6.5. *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 $\mathbb{C}$. □

Summarizing, the theory of continuous functions $f : \mathbb{R} \to \mathbb{R}$ extends well into a theory of continuous functions $f : \mathbb{C} \to \mathbb{C}$, modulo a few minor subtleties. We will see right next that things drastically change when looking at the notion of differentiability.

6b. Holomorphic functions

Let us study now differentiability. We have here:

DEFINITION 6.6. *We say that a function $f : X \to \mathbb{C}$ is differentiable in the complex sense when the following limit is defined for any $x \in X$:*

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

In this case, we also say that f is holomorphic, and we write $f \in H(X)$.

As basic examples, we have for instance the power functions $f(x) = x^n$, and more examples in a moment. The general theory from the real case extends well, as follows:

PROPOSITION 6.7. *We have the following results:*

(1) $(f+g)' = f' + g'$.
(2) $(\lambda f)' = \lambda f'$.
(3) $(f \circ g)' = f'(g)g'$.

PROOF. These formulae are all clear from definitions, exactly as in the real case. Thus, we are led to the conclusions in the statement. □

At the level of the examples now, we have:

PROPOSITION 6.8. *All basic functions that we know are holomorphic, and in fact are infinitely differentiable. However, functions like $\bar{z}$ or $|z|$ are not holomorphic.*

PROOF. Here the first assertion is standard. Regarding now $f(z) = \bar{z}$, we have:

$$\frac{f(x+h) - f(x)}{h} = \frac{\bar{x} + \bar{h} - \bar{x}}{h} = \frac{\bar{h}}{h}$$

But this limit does not converge with $h \to 0$, for instance because with $h \in \mathbb{R}$ we obtain 1 as limit, while with $h \in i\mathbb{R}$ we obtain -1 as limit. In fact, with $h = rw$ with $|w| = 1$ fixed and $r \in \mathbb{R}$, $r \to 0$, we can obtain as limit any number on the unit circle:

$$\lim_{r \to 0} \frac{f(x+rw) - f(x)}{rw} = \lim_{r \to 0} \frac{r\bar{w}}{rw} = \bar{w}^2$$

The situation for the function $f(z) = |z|$ 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 $\bar{w}$, and now by making vary w we can obtain in this way limits pointing in all possible directions. □

The above result might seem quite surprising, for a number of reasons. First, $z \to \bar{z}$ looks like a "smooth" operation, and so smoothness is not the same thing as differentiability, in the complex setting. We will be back to functions of type $\bar{z}$ later, first in chapter 8 below, with the results that they are however "harmonic", in some suitable sense, and then in chapters 9-16 below, when talking about functions in several real variables.

Another interesting feature of Proposition 6.8 concerns the positive results, regarding the functions which are holomorphic. The examples there are all 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 6.9. *Assuming that a function $f : X \to \mathbb{C}$ is analytic, in the sense that it is a series, around each point $x \in X$,*

$$f(x+h) = \sum_{n=0}^{\infty} c_n h^n$$

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:

$$f'(x+h) = \sum_{n=1}^{\infty} n c_n h^{n-1}$$

Thus f' exists and is analytic too, and this gives the result. □

We will see later that, conversely, holomorphic implies analytic. In order to get into this, let us forget for the moment the general holomorphic functions, and study instead the analytic functions. We already know from chapter 5 that even in the polynomial case, $P \in \mathbb{C}[X]$, some interesting things happen, because any such polynomial has a root. Kepping looking at polynomials, with the same methods, we are led to:

THEOREM 6.10. *The maximum principle holds for polynomials $P \in \mathbb{C}[X]$.*

PROOF. It is enough indeed to prove this for circles, and in the circle case, the proof is standard, similar to the proof of the existence of a root, from chapter 5 above. □

Thinking a bit at Theorem 6.10, one 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 domain can be recovered from its values on the boundary. And fortunately, this is indeed the case, and we have the following result:

THEOREM 6.11. *The Cauchy formula holds for polynomials $P \in \mathbb{C}[X]$.*

PROOF. It is enough indeed to prove this for circles, and in the circle case, the proof is standard, somewhat by generalizing the computations from the proof of Theorem 6.10. □

All this is quite interesting, and we are now into some serious mathematics here. Importantly, Theorem 6.10 and Theorem 6.11 provide us with a path for proving the converse of Theorem 6.9. 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, this will be our plan, in what follows.

6c. Cauchy formula

Our aim here is to prove the Cauchy formula for any holomorphic function $f : X \to \mathbb{C}$. Let us start with some preliminaries, putting the integration results from the previous section, for polynomials, into a more conceptual framework. We first have:

PROPOSITION 6.12. *We can integrate functions f over curves γ according to*

$$\int_\gamma f(z)dz = \int_a^b f(\gamma(t))\gamma'(t)dt$$

with this quantity being independent on the parametrization $\gamma : [a, b] \to \mathbb{C}$.

PROOF. This follows indeed from the chain rule for derivatives. □

The main properties of the above integration method are as follows:

PROPOSITION 6.13. *We have the following formula, for a union of paths:*

$$\int_{\gamma\cup\eta} f(z)dz = \int_\gamma f(z)dz + \int_\eta f(z)dz$$

Also, when reversing the path, the integral changes its sign.

PROOF. All these assertions are indeed clear from definitions. □

We can now state and prove, as a main result:

THEOREM 6.14. *The Cauchy formula holds for any holomorphic function $f : X \to \mathbb{C}$.*

PROOF. This can be proved for instance for triangles, via a standard computation, and then by using triangles we can have it for convex sets, and then in general. □

As a main application of the Cauchy formula, we have:

THEOREM 6.15. *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. Here (3) $\implies$ (2) $\implies$ (1) are elementary results, that we know from Theorem 6.9, the implication (1) $\implies$ (4) is non-trivial, but we know this from Theorem 6.14, and the remaining implication, namely (4) $\implies$ (3), is clear. □

As another application of the Cauchy formula, we have:

THEOREM 6.16. *The maximum principle holds for any holomorphic function f.*

PROOF. This follows indeed from the Cauchy formula. Observe that the converse is not true, for instance because functions like $\bar{z}$ satisfy too the maximum principle. We will be back to this later, when talking about harmonic functions. □

As yet another application of the Cauchy formula, we have:

THEOREM 6.17. *An entire, bounded holomorphic function must be constant.*

PROOF. This follows indeed from the Cauchy formula. □

Again, we will be back to this when talking about harmonic functions.

6d. Poles, residues

Poles, residues, and some applications.

6e. Exercises

Exercises.

CHAPTER 7

Fourier analysis

7a. Function spaces

In this chapter we go back to the functions of one real variable, $f : \mathbb{R} \to \mathbb{R}$, that we will regard as functions $f : \mathbb{R} \to \mathbb{C}$, with some applications, by using our complex number technology. We will be mostly interested in constructing the Fourier transform, which is an operation $f \to \widehat{f}$, on such functions, which can solve various concrete questions.

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 7.1. *Given two functions $f, g : \mathbb{R} \to \mathbb{C}$ and an exponent $p \geq 1$, we have*

$$\left(\int |f+g|^p\right)^{1/p} \leq \left(\int |f|^p\right)^{1/p} + \left(\int |g|^p\right)^{1/p}$$

called Minkowski inequality. Also, assuming that $p, q \geq 1$ satisfy $1/p + 1/q = 1$, we have

$$\int |fg| \leq \left(\int |f|^p\right)^{1/p} \left(\int |g|^q\right)^{1/q}$$

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 $p = 2$ we have $q = 2$ as well, and the Minkowski and Hölder inequalities are as follows:

$$\left(\int |f+g|^2\right)^{1/2} \leq \left(\int |f|^2\right)^{1/2} + \left(\int |g|^2\right)^{1/2}$$

$$\int |fg| \leq \left(\int |f|^2\right)^{1/2} \left(\int |g|^2\right)^{1/2}$$

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 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, and the same goes for the Hölder inequality at $p = \infty, q = 1$, or at $p = 1, q = \infty$. □

As a consequence of the above results, we can formulate:

THEOREM 7.2. *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*

$$L^p(I) = \left\{ f : I \to \mathbb{C} \middle| \int |f(x)|^p dx < \infty \right\}$$

is a vector space, and the following quantity

$$||f||_p = \left(\int |f(x)|^p \right)^{1/p}$$

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 7.1, 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 7.1, 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:

$$< f, g >= \int f(x)\, \overline{g(x)}\, dx$$

(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 7.1.

(3) Finally, the extension at $p = \infty$ is clear too, coming from definitions. □

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:

$$L^p(I)^* = L^q(I)$$

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 7.3. *The convolution of two functions $f, g : \mathbb{R} \to \mathbb{C}$ is the function*

$$f * g\,(x) = \int f(x-y)g(y)dy$$

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 7.4. *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 $\mathbb{R}$:

$$\begin{aligned}
\int |f(x-y)g(y)|dy &= \int_{supp(g)} |f(x-y)g(y)|dy \\
&\leq \max(g) \int_{supp(g)} |f(x-y)|dy \\
&\leq \max(g) \cdot l(supp(g)) \max(f) \\
&< \infty
\end{aligned}$$

(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 absolute continuity for the function $f\in C_c(\mathbb{R})$:

$$\begin{aligned}|f*g(x+\varepsilon)-f*g(x)| &= \left|\int f(x+\varepsilon-y)g(y)dy-\int f(x-y)g(y)dy\right|\\ &= \left|\int\left(f(x+\varepsilon-y)-f(x-y)\right)g(y)dy\right|\\ &\leq \int|f(x+\varepsilon-y)-f(x-y)|\cdot|g(y)|dy\\ &\leq K_f\cdot\varepsilon\cdot\int|g|\end{aligned}$$

(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 $f*g$:

$$\begin{aligned}f*g(x) &= \int f(x-y)g(y)dy\\ &= \int_{supp(g)}f(x-y)g(y)dy\end{aligned}$$

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 7.5. *The following hold, for functions in $C_c(\mathbb{R})$:*

(1) $f*g=g*f$.
(2) $f*(g*h)=(f*g)*h$.
(3) $f*(ag+bh)=af*g+bf*h$.

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

(1) This follows from the following computation, with $y=x-t$:

$$\begin{aligned}f*g(x) &= \int f(x-y)g(y)dy\\ &= \int f(t)g(x-t)dt\\ &= \int g(x-t)f(t)dt\\ &= g*f(x)\end{aligned}$$

(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 7.3, we have the following result:

THEOREM 7.6. *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, this follows from the following computation, with the change of variables $y = x - t$, then the change of variables $t = x - y$:

$$\begin{aligned}
(f * g)'(x) &= \frac{d}{dx}\int f(x-y)g(y)dy \\
&= \frac{d}{dx}\int f(t)g(x-t)dt \\
&= \int f(t)g'(x-t)dt \\
&= \int f(x-y)g'(y)dy \\
&= (f * g')(x)
\end{aligned}$$

As for the second assertion, this follows form the first one, by recurrence. □

Finally, getting beyond the compactly supported continous functions, we have:

THEOREM 7.7. *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:

$$\begin{aligned}\int|f*g(x)| &\leq \int\int|f(x-y)g(y)|dydx\\ &= \int\int|f(x-y)g(y)|dxdy\\ &\leq \int|f|\int|g|\end{aligned}$$

As for the second assertion, this follows from the first one, and from Theorem 7.6. □

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 7.8. *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 $\widehat{f}(0)=\int f$, which shows that $\widehat{f}$ encodes interesting information about f, being quite illustrating for what we want to do. Here are some basic properties of the Fourier transform, all providing some good motivations:

PROPOSITION 7.9. *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) This is indeed clear from definitions.

(2) The continuity of $\widehat{f}$ follows indeed from:

$$\begin{aligned}|\widehat{f}(\xi+\varepsilon)-\widehat{f}(\xi)| &\leq \int \left|(e^{ix(\xi+\varepsilon)}-e^{ix\xi})f(x)\right|dx\\ &= \int \left|e^{ix\xi}(e^{ix\varepsilon}-1)f(x)\right|dx\\ &\leq |e^{ix\varepsilon}-1|\int |f|\end{aligned}$$

As for the boundedness of $\widehat{f}$, this is clear as well.

(3) This follows from the following computation, assuming that f is even:

$$\begin{aligned}\widehat{f}(-\xi) &= \int e^{-ix\xi}f(x)dx\\ &= \int e^{ix\xi}f(-x)dx\\ &= \int e^{ix\xi}f(x)dx\\ &= \widehat{f}(\xi)\end{aligned}$$

(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 7.10. *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 7.9:

$$F(x) = f(-x) \implies \widehat{F}(\xi) = \widehat{f}(-\xi)$$

(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 7.9. □

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 7.9, and providing more motivations for the Fourier transform:

PROPOSITION 7.11. *Given $f, g \in L^1(\mathbb{R})$ we have $\widehat{f}g, f\widehat{g} \in L^1(\mathbb{R})$ and*

$$\int f(\xi)\widehat{g}(\xi)d\xi = \int \widehat{f}(x)g(x)dx$$

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:

$$\int f(\xi)\widehat{g}(\xi)d\xi = \int\int f(\xi)e^{ix\xi}g(x)dxd\xi$$

On the other hand, we have as well the following formula:

$$\int \widehat{f}(x)g(x)dx = \int\int e^{ix\xi}f(x)g(\xi)dxd\xi$$

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, we have:

THEOREM 7.12. *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:

$$\begin{aligned}\widehat{f'}(\xi) &= \int e^{ix\xi}f'(x)dx\\ &= -\int i\xi e^{ix\xi}f(x)dx\\ &= -i\xi\int e^{ix\xi}f(x)dx\\ &= -i\xi\widehat{f}(\xi)\end{aligned}$$

(2) Regarding the higher derivatives, the formula here follows by recurrence. □

Importantly, we have a converse statement as well, as follows:

THEOREM 7.13. *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, assuming that $F_k(x) = x^k f(x)$ belongs to $L^1(\mathbb{R})$, for $k = 0, 1, \ldots, n$, the function $\widehat{f}$ is n times differentiable, with derivatives given by

$$(\widehat{f})^{(k)}(\xi) = i^k\widehat{F_k}(\xi)$$

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:

$$\begin{aligned}(\widehat{f})'(\xi) &= \frac{d}{d\xi}\int e^{ix\xi}f(x)dx\\ &= \int ixe^{ix\xi}f(x)dx\\ &= i\int e^{ix\xi}xf(x)dx\\ &= i\widehat{F}(\xi)\end{aligned}$$

(2) As for the second assertion, this follows from the first one, by recurrence. □

As a conclusion to all this, we have:

CONCLUSION 7.14. *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, and this even for courageous people us, loving and mastering calculus. Here is now another useful result, of the same type, this time regarding convolutions:

THEOREM 7.15. *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, which can be proved as follows, with the change of variables $x = y + t$:

$$\begin{aligned}\widehat{f * g}(\xi) &= \int e^{ix\xi}(f * g)(x)dx\\ &= \int\int e^{ix\xi}f(x-y)g(y)dxdy\\ &= \int e^{iy\xi}\left(\int e^{i(x-y)\xi}f(x-y)dx\right)g(y)dy\\ &= \int e^{iy\xi}\left(\int e^{it\xi}f(t)dt\right)g(y)dy\\ &= \int e^{iy\xi}\widehat{f}(\xi)g(y)dy\\ &= \widehat{f}(\xi)\widehat{g}(\xi)\end{aligned}$$

(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:

$$\begin{aligned}
\widehat{f} * \widehat{g}(\xi) &= \int \widehat{f}(\xi-\eta)\widehat{g}(\eta)d\eta \\
&= \int\int\int e^{ix(\xi-\eta)}f(x)e^{iy\eta}g(y)dxdyd\eta \\
&= \int\int\int e^{ix\eta}e^{i(y-x)\eta}f(x)g(y)dxdyd\eta \\
&= \int e^{ix\eta}f(x)g(x)dx \\
&= \widehat{fg}(\eta)
\end{aligned}$$

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 do not have yet much evidence for this argument, but we will be back to this later, with a full justification. □

As an updated conclusion to all this, we have, modulo a few bugs, still to be fixed:

CONCLUSION 7.16. *The Fourier transform converts the derivatives into multiplications by the variable, and convolutions into products, and vice versa.*

Obviously, this is something quite powerful, which can be used for a wide range of purposes. We will see applications later, after developing some more general theory.

7c. Inversion formula

We develop now more theory for the Fourier transform. We first have:

THEOREM 7.17. *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) Let us begin with something which is of independent interest. 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 7.18. *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 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 7.17 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 here 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 restricted to the space $\mathcal{S}$ of functions all whose derivatives are rapidly decreasing, called Schwartz space.

7d. Groups, extensions

Groups, extensions.

7e. Exercises

Exercises.

CHAPTER 8

Harmonic functions

8a. The Laplacian

We discuss in this chapter some applications of the theory developed so far, along with some more theory, needed in order to really reach to these applications. We will be mainly interested in two fundamental equations of physics, namely the wave equation, and the heat equation. 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 8.1. *The wave equation in $\mathbb{R}^N$ is*

$$\ddot{\varphi} = v^2 \Delta \varphi$$

where Δ is the Laplace operator, given by

$$\Delta f(x) = \sum_{i=1}^{N} \frac{d^2 f}{dx_i^2}$$

and where $v > 0$ is the propagation speed.

PROOF. As already mentioned, the equation in the statement is what comes out of experiments. However, allowing us a bit of imagination, and trust in this imagination, we can mathematically "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. First is the Newton motion force, mass times acceleration, which is as follows, with m being the mass of each ball:

$$F_n = m \cdot \ddot{\varphi}(x)$$

And second is the Hooke force, displacement of the spring, times spring constant. Since we have two springs at x, this is as follows, k being the spring constant:

$$\begin{aligned} F_h &= F_h^r - F_h^l \\ &= k(\varphi(x+l) - \varphi(x)) - k(\varphi(x) - \varphi(x-l)) \\ &= k(\varphi(x+l) - 2\varphi(x) + \varphi(x-l)) \end{aligned}$$

We conclude that the equation of motion, in our model, is as follows:

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

(2) Now let us take the limit of our model, as to reach to continuum. For this purpose we will assume that our system consists of $N >> 0$ balls, having a total mass M, and spanning a total distance L. Thus, our previous infinitesimal parameters are as follows, with K being the spring constant of the total system, which is of course lower than k:

$$m = \frac{M}{N} \quad , \quad k = KN \quad , \quad l = \frac{L}{N}$$

With these changes, our equation of motion found in (1) reads:

$$\ddot{\varphi}(x) = \frac{KN^2}{M}(\varphi(x+l) - 2\varphi(x) + \varphi(x-l))$$

Now observe that this equation can be written, more conveniently, as follows:

$$\ddot{\varphi}(x) = \frac{KL^2}{M} \cdot \frac{\varphi(x+l) - 2\varphi(x) + \varphi(x-l)}{l^2}$$

With $N \to \infty$, and therefore $l \to 0$, we obtain in this way:

$$\ddot{\varphi}(x) = \frac{KL^2}{M} \cdot \frac{d^2\varphi}{dx^2}(x)$$

(3) In arbitrary N dimensions now, the same argument carries on, and we are led to the following equation, with $v = \sqrt{K/M} \cdot L$ being the propagation speed:

$$\ddot{\varphi}(x) = v^2 \sum_i \frac{d^2\varphi}{dx_i^2}(x)$$

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 in (1) is indeed the correct one. But do not worry, experiments confirm our findings. □

Regarding now heat, we have here a similar equation, as follows:

THEOREM 8.2. *Heat diffusion in $\mathbb{R}^N$ 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 are possible as well. □

As a conclusion now, we have been doing some physics, with the conclusion that both the wave and heat equation involve the Laplace operator, namely:

$$\Delta f(x) = \sum_{i=1}^{N} \frac{d^2 f}{dx_i^2}$$

In what follows we will be mainly interested in the 2D case, where the technology that we have can be applied. So, let us formulate, as a mathematical conclusion:

DEFINITION 8.3. *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 8.1 and Theorem 8.2, at $N = 2$. 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 the harmonic functions, $\Delta f = 0$.

Getting now to more concrete things, and to 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 8.4. *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, with $z = x + iy$:

$$\begin{aligned}
\Delta z^n &= \frac{d^2 z^n}{dx^2} + \frac{d^2 z^n}{dy^2} \\
&= \frac{d(nz^{n-1})}{dx} + \frac{d(inz^{n-1})}{dy} \\
&= n(n-1)z^{n-2} - n(n-1)z^{n-2} \\
&= 0
\end{aligned}$$

As for the second assertion, this follows from $\Delta \bar{f} = \overline{\Delta f}$. □

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.

8b. Harmonic functions

As a first goal, 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 $P \in \mathbb{R}[z, \bar{z}]$. Now regarding these latter polynomials, a direct computation shows that we have:

$$f = z^k \bar{z}^l \implies \Delta f = \frac{4klf}{|z|^2}$$

We conclude by linearity that $z^n, \bar{z}^n$ are the only solutions. 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.

As another goal, let us try now to find the radial harmonic functions $f : \mathbb{R}^2 \to \mathbb{C}$. Things are quite tricky here, involving two interesting phenomena, namely a blowup phenomenon at the origin $z = 0$, and also a blowup phenomenon at the dimension value $N = 2$, the one that we are interested in, when considering the problem in the more general context of the radial harmonic functions $f : \mathbb{R}^N \to \mathbb{C}$. In view of these phenomena, that will become clear later on, let us try to find the radial harmonic functions, as follows:

$$f : \mathbb{R}^N - \{0\} \to \mathbb{C}$$

Let us first look at the 1D case. Here the Laplacian is the second derivative, $\Delta f = f''$, and we obtain as solutions the symmetrizations of the linear functions $f : \mathbb{R}_+ \to \mathbb{R}$:

$$f(x) = a|x| + b$$

In more dimensions things are more complicated, the result being as follows:

THEOREM 8.5. *The fundamental radial solutions of* $\Delta f = 0$ *are*

$$f = \begin{cases} ||x||^{2-N} & (N \neq 2) \\ \log|x| & (N = 2) \end{cases}$$

with the log *at* $N = 2$ *basically coming from* $\log' = 1/x$.

PROOF. This follows indeed by writing $f(x) = u(||x||)$, then reformulating the Laplace equation $\Delta f = 0$ into a certain differential equation on $u : \mathbb{R} \to \mathbb{C}$, and then solving this latter differential equation, with the log appearing at $N = 2$ as indicated. □

Going ahead now with some theory, we have:

THEOREM 8.6. *The harmonic functions in 2 dimensions, and in fact in any N dimensions, obey to the same general principles as the holomorphic functions, such as the mean value formula, the maximum modulus principle, and the Liouville theorem.*

PROOF. This is something quite tricky, with the study here partly bringing us into some subtle calculus results in N real variables, that we will investigate more in detail later, in the second part of the present book, starting with chapter 9 below. □

There are many more things that can be said about the harmonic functions, for instance of geometric nature, in connection with soap films, which are harmonic.

Also, the mean value formula, which is our main tool, leads us into the computation of the volume of unit sphere in $\mathbb{R}^N$. We will solve this question much later, in chapter 14 below, after developing a suitable theory of spherical coordinates.

8c. Waves and heat

With the above theory developed, we can go back to waves and heat in 2D.

8d. Higher dimensions

In higher dimensions, and especially in 3D, which is of practical interest, the wave and heat equations become quite complicated, but we can say a few things about them.

8e. Exercises

Exercises.

Part III

Several variables

This is my church
This is where I heal my hurts
For tonight
God is a DJ

CHAPTER 9

Linear algebra

9a. Linear maps

In the remainder of this book we study the functions of several variables. Let us start with the linear functions, which are the simplest. We have 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 $A_{ij} \in \mathbb{R}$:

$$f\begin{pmatrix} x_1 \\ \vdots \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} A_{11}x_1 + \ldots + A_{1N}x_N \\ \vdots \\ \vdots \\ A_{M1}x_1 + \ldots + A_{MN}x_N \end{pmatrix}$$

Now the parameters $A_{ij} \in \mathbb{R}$ can be regarded as being the entries of a 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:

$$f(e_j) = \begin{pmatrix} A_{1j} \\ \vdots \\ \vdots \\ A_{Mj} \end{pmatrix}$$

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 $t \in \mathbb{R}$ is given by the matrix*

$$R_t = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}$$

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 $\binom{0}{1}$. A quick picture shows that we must have:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$$

Also, by paying attention to positives and negatives, we must have:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -\sin t \\ \cos t \end{pmatrix}$$

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. □

Regarding now the symmetries, the formula here is as follows:

PROPOSITION 9.3. *The symmetry with respect to the Ox axis rotated by an angle $t/2 \in \mathbb{R}$ is given by the matrix*

$$S_t = \begin{pmatrix} \cos t & \sin t \\ \sin t & -\cos t \end{pmatrix}$$

depending on $t \in \mathbb{R}$ taken modulo 2π.

PROOF. The symmetry being linear, it must correspond to a certain matrix:

$$S_t = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

As before, we can guess this matrix via its action on the basic coordinate vectors $\binom{1}{0}$ and $\binom{0}{1}$. A quick picture shows that we must have:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$$

Also, by paying attention to positives and negatives, we must have:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \sin t \\ -\cos t \end{pmatrix}$$

Guessing now the matrix is not complicated, because we must have:

$$\begin{pmatrix} a \\ c \end{pmatrix} = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix} \quad , \quad \begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} \sin t \\ -\cos t \end{pmatrix}$$

Thus, we can just put together these two vectors, and we obtain our matrix. □

Finally, regarding the projections, the formula here is as follows:

PROPOSITION 9.4. *The projection on the Ox axis rotated by an angle $t/2 \in \mathbb{R}$ is given by the matrix*

$$P_t = \frac{1}{2} \begin{pmatrix} 1 + \cos t & \sin t \\ \sin t & 1 - \cos t \end{pmatrix}$$

depending on $t \in \mathbb{R}$ taken modulo 2π.

PROOF. A quick picture, 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 theory, our claim is that, no matter what we want to do with T or A, we will run at some point into their adjoints T^* and A^*, constructed as follows:

THEOREM 9.5. *The adjoint operator $T^* : \mathbb{C}^N \to \mathbb{C}^N$, which is given by*

$$< Tx, y > = < x, T^* y >$$

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 $T : \mathbb{C}^N \to \mathbb{C}^N$, fix $y \in \mathbb{C}^N$, and consider the linear form $\varphi(x) = < Tx, y >$. This form must be as follows, for a certain vector $T^* y \in \mathbb{C}^N$:

$$\varphi(x) = < x, T^* y >$$

Thus, we have constructed a map $y \to T^*y$ as in the statement, which is obviously linear, and that we can call T^*. 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 T^* reads:

$$< Te_i, e_j >=< e_i, T^*e_j >$$

By reversing the scalar product on the right, this formula can be written as:

$$< T^*e_j, e_i >= \overline{< Te_i, e_j >}$$

But this means that the matrix of T^* is given by $(A^*)_{ij} = \bar{A}_{ji}$, as desired. □

Getting back to our claim, the adjoints $*$ are indeed ubiquitous, as shown by:

THEOREM 9.6. *The following happen:*

(1) $T(x) = Ux$ *with* $U \in M_N(\mathbb{C})$ *is an isometry precisely when* $U^* = U^{-1}$.
(2) $T(x) = Px$ *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:

$$4 < x, y >= ||x + y||^2 - ||x - y||^2 + i||x + iy||^2 - i||x - iy||^2$$

The proof of this latter formula is indeed elementary, as follows:

$$\begin{aligned} & ||x + y||^2 - ||x - y||^2 + i||x + iy||^2 - i||x - iy||^2 \\ = \; & ||x||^2 + ||y||^2 - ||x||^2 - ||y||^2 + i||x||^2 + i||y||^2 - i||x||^2 - i||y||^2 \\ & +2Re(< x, y >) + 2Re(< x, y >) + 2iIm(< x, y >) + 2iIm(< x, y >) \\ = \; & 4 < x, y > \end{aligned}$$

Finally, we will use Theorem 9.5, 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:

$$\begin{aligned} ||Ux|| = ||x|| \quad & \iff \quad < Ux, Uy >=< x, y > \\ & \iff \quad < x, U^*Uy >=< x, y > \\ & \iff \quad U^*Uy = y \\ & \iff \quad U^*U = 1 \\ & \iff \quad U^* = U^{-1} \end{aligned}$$

(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:

$$f(v) = Av$$

In this chapter we develop more general theory for such linear maps. We are interested in the question of inverting the linear maps $f : \mathbb{R}^N \to \mathbb{R}^N$. And this is the same question as inverting the corresponding matrices $A \in M_N(\mathbb{R})$, due to:

THEOREM 9.7. *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 9.8. *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 9.9. *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.

In relation with our inversion problem, we have the following statement:

PROPOSITION 9.10. *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 9.8:

THEOREM 9.11. *In* 2 *dimensions we have the following formula,*

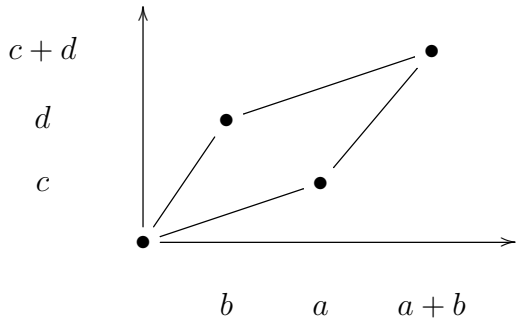
$$\det^+ \begin{pmatrix} a & b \\ c & d \end{pmatrix} = |ad - bc|$$

with $\det^+ : M_2(\mathbb{R}) \to \mathbb{R}_+$ *being the function constructed above.*

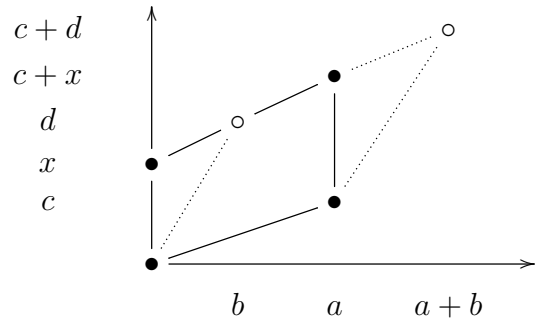
PROOF. We must show that the area of the parallelogram formed by $\binom{a}{c}, \binom{b}{d}$ equals $|ad - bc|$. We can assume $a, b, c, d > 0$ for simplifying, the proof in general being similar. Moreover, by switching if needed the vectors $\binom{a}{c}, \binom{b}{d}$, we can assume that we have:

$$\frac{a}{c} > \frac{b}{d}$$

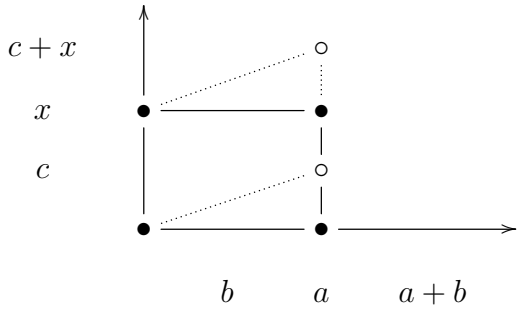
According to these conventions, the picture of our parallelogram is as follows:



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:



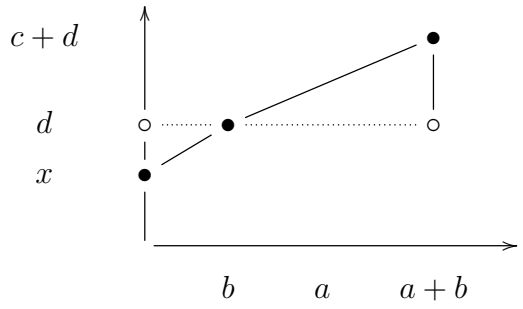
We can further modify this parallelogram, once again by not altering its area, by sliding the right side downwards, until we reach the Ox axis:



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 we obviously 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 9.12. *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 9.13. *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 9.14. *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 9.15. *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 9.9, 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 9.16. *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 9.17. *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 9.11, and to the sign formula from Proposition 9.13, 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 9.18. *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 9.14 (1).

(2) This follows as well from definitions, and from Proposition 9.14 (2).

(3) This is clear from our definition of the determinant. □

As an application of the above result, we have:

THEOREM 9.19. *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 the rules (1) and (3) in Theorem 9.18, which in matrix terms give:

$$\begin{aligned} \begin{vmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{vmatrix} &= \lambda_1 \ldots \lambda_N \begin{vmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{vmatrix} \\ &= \lambda_1 \ldots \lambda_N \end{aligned}$$

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 9.20. *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 9.21. *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 decomposition formula in the statement follows by using the associated linear maps, which multiply as follows:

$$f_{AB} = f_A f_B$$

As for the formula $\det(AB) = \det(BA)$, this is clear from the first formula. □

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

THEOREM 9.22. *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 9.21, 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. □

Here is another important result, which is very useful for diagonalization:

THEOREM 9.23. *The eigenvalues of a matrix* $A \in M_N(\mathbb{R})$ *are the roots of*

$$P(x) = \det(A - x1_N)$$

called characteristic polynomial of the matrix.

PROOF. We have the following computation, using the fact that a linear map is bijective precisely when the determinant of the associated matrix is nonzero:

$$\begin{aligned} \exists v, Av = \lambda v &\iff \exists v, (A - \lambda 1_N)v = 0 \\ &\iff \det(A - \lambda 1_N) = 0 \end{aligned}$$

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

In general now, at the theoretical level, we have the following key result:

THEOREM 9.24. *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 9.11, 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 9.25. *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 9.22.

(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 9.26. *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 9.27. *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 9.26. □

We can expand as well over the columns, as follows:

THEOREM 9.28. *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 final statement, here is an alternative reformulation of Theorem 9.26:

THEOREM 9.29. *The determinant of the systems of vectors*

$$\det : \mathbb{R}^N \times \ldots \times \mathbb{R}^N \to \mathbb{R}$$

is multilinear, alternate and unital, and unique with these properties.

PROOF. This is a fancy reformulation of Theorem 9.26, with the various properties of det from the statement being those found above. □

9d. Sarrus and beyond

As a first application of the above methods, we can now prove:

THEOREM 9.30. *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,

$$\begin{aligned}\det &= \begin{pmatrix} * & & \\ & * & \\ & & *\end{pmatrix}+\begin{pmatrix} & * & \\ & & *\\ *& & \end{pmatrix}+\begin{pmatrix} & & *\\ *& & \\ & *& \end{pmatrix}\\ &-\begin{pmatrix} & & *\\ & * & \\ *& & \end{pmatrix}+\begin{pmatrix} & * & \\ *& & \\ & & *\end{pmatrix}+\begin{pmatrix} *& & \\ & & *\\ & *& \end{pmatrix}\end{aligned}$$

"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. □

As a first application, let us go back to the inversion problem for the 3×3 matrices, that we left open in the above. We can now solve this problem, as follows:

THEOREM 9.31. *The inverses of the* 3×3 *matrices are given by*

$$\begin{pmatrix} a & b & c\\ d & e & f\\ g & h & i\end{pmatrix}^{-1} = \frac{1}{D}\begin{pmatrix} ei-fh & ch-bi & bf-ce\\ fg-di & ai-cg & cd-af\\ dh-eg & bg-ah & ae-bd\end{pmatrix}$$

with D *being the determimant. When* $D=0$*, the matrix is not invertible.*

PROOF. We can use here the same method as for the 2×2 matrices. To be more precise, in order for the matrix to be invertible, we must have:

$$D\neq 0$$

The trick now is to look for solutions of the following problem:

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} * & * & * \\ * & * & * \\ * & * & * \end{pmatrix} = \begin{pmatrix} D & 0 & 0 \\ 0 & D & 0 \\ 0 & 0 & D \end{pmatrix}$$

We know from Theorem 9.30 that the determinant is given by:

$$D = aei + bfg + cdh - ceg - bdi - afh$$

But this leads, via some obvious choices, to the following solution:

$$\begin{pmatrix} * & * & * \\ * & * & * \\ * & * & * \end{pmatrix} = \begin{pmatrix} ei - fh & ch - bi & bf - ce \\ fg - di & ai - cg & cd - af \\ dh - eg & bg - ah & ae - bd \end{pmatrix}$$

Thus, by rescaling, we obtain the formula in the statement. □

In fact, we can now fully solve the inversion problem, as follows:

THEOREM 9.32. *The inverse of a square matrix, having nonzero determinant,*

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1N} \\ \vdots & & \vdots \\ a_{N1} & \ldots & a_{NN} \end{pmatrix}$$

is given by the following formula,

$$A^{-1} = \frac{1}{\det A} \begin{pmatrix} \det A^{(11)} & -\det A^{(21)} & \det A^{(31)} & \ldots \\ -\det A^{(12)} & \det A^{(22)} & -\det A^{(32)} & \ldots \\ \det A^{(13)} & -\det A^{(23)} & \det A^{(33)} & \ldots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

where $A^{(ij)}$ is the matrix A, with the i-th row and j-th column removed.

PROOF. This follows indeed by using the row expansion formula, which in terms of the matrix A^{-1} in the statement reads $AA^{-1} = 1$. □

Let us discuss now the general formula of the determinant, at arbitrary values $N \in \mathbb{N}$ of the matrix size, generalizing those that we have at $N = 2, 3, 4$. We will need:

DEFINITION 9.33. *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, is by denoting the permutations as diagrams, going from top to bottom:

$\sigma =$ 

Here are some basic properties of the permutations:

THEOREM 9.34. *The permutations have the following properties:*

(1) *There are $N!$ of them.*
(2) *They are stable by composition, and inversion.*

PROOF. In order to construct a permutation $\sigma \in S_N$, we have:

– N choices for the value of $\sigma(N)$.
– $(N-1)$ choices for the value of $\sigma(N-1)$.
– $(N-2)$ choices for the value of $\sigma(N-2)$.
$\vdots$
– and so on, up to 1 choice for the value of $\sigma(1)$.

Thus, we have $N!$ choices, as claimed. As for the second assertion, this is clear. □

We will need the following key result:

THEOREM 9.35. *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, and the fact that these numbers are well-defined 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. □

We can now formulate a key result, as follows:

THEOREM 9.36. *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 9.35 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:

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - ceg - bdi - afh$$

Observe that the triangles in the Sarrus formula correspond to the permutations of $\{1, 2, 3\}$, and their signs correspond to the signatures of these permutations:

$$\begin{aligned}\det &= \begin{pmatrix} * & & \\ & * & \\ & & * \end{pmatrix} + \begin{pmatrix} & * & \\ & & * \\ * & & \end{pmatrix} + \begin{pmatrix} & & * \\ * & & \\ & * & \end{pmatrix} \\ &- \begin{pmatrix} & & * \\ & * & \\ * & & \end{pmatrix} + \begin{pmatrix} & * & \\ * & & \\ & & * \end{pmatrix} + \begin{pmatrix} * & & \\ & & * \\ & * & \end{pmatrix}\end{aligned}$$

As a basic application, we have the following key result:

THEOREM 9.37. *We have the formula*

$$\det A = \det A^t$$

valid for any square matrix A.

PROOF. This follows from the formula in Theorem 9.36. Indeed, we have:

$$\begin{aligned}
\det A^t &= \sum_{\sigma\in S_N} \varepsilon(\sigma)(A^t)_{1\sigma(1)}\ldots(A^t)_{N\sigma(N)} \\
&= \sum_{\sigma\in S_N} \varepsilon(\sigma)A_{\sigma(1)1}\ldots A_{\sigma(N)N} \\
&= \sum_{\sigma\in S_N} \varepsilon(\sigma)A_{1\sigma^{-1}(1)}\ldots A_{N\sigma^{-1}(N)} \\
&= \sum_{\sigma\in S_N} \varepsilon(\sigma^{-1})A_{1\sigma^{-1}(1)}\ldots A_{N\sigma^{-1}(N)} \\
&= \sum_{\sigma\in S_N} \varepsilon(\sigma)A_{1\sigma(1)}\ldots A_{N\sigma(N)} \\
&= \det A
\end{aligned}$$

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

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.

9e. Exercises

CHAPTER 10

Continuity

10a. Open and closed sets

We have seen so far the theory of maps $f : \mathbb{R}^N \to \mathbb{R}^M$ which are linear. In what follows we develop the theory of such maps in general, in analogy with what we know about the arbitrary maps $f : \mathbb{R} \to \mathbb{R}$, from part I of the present book.

This is something quite technical, which will take some time, even for truly getting started. As a first objective, we would like to talk about continuity, and other basic analytic properties, of the maps $f : \mathbb{R}^N \to \mathbb{R}^M$. And here, things are more tricky than in 1D, because no matter what kind of analysis we want to do in $\mathbb{R}^N$, we will run all the time into the formula of the distance there, which is something quite complicated, 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 $d(x, y)$ constructed above. This leads us into:

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.

Moving ahead now with some theory, we have:

PROPOSITION 10.2. *We can talk about limits, continuity, open and closed sets inside metric spaces X, and in particular inside $\mathbb{R}^N$, exactly as we did before, inside $\mathbb{R}$.*

PROOF. All this is very standard, and with the remark that our previous experience with the subject includes, besides $\mathbb{R}$, the complex plane $\mathbb{C}$ too, with its usual distance $d(x, y) = |x - y|$, which is the same as the distance on the real plane $\mathbb{R}^2$. □

As a more subtle level now, we have:

THEOREM 10.3. *We can talk about compact sets $K \subset X$, as follows:*

(1) *K compact means that any open cover has a finite subcover.*
(2) *Compact implies closed. Also, closed inside compact implies compact.*
(3) *If f is continuous and K is compact, then $f(K)$ is compact.*
(4) *In a compact, any infinite subset $E \subset K$ has a limit point.*
(5) *For subsets $K \subset \mathbb{R}^N$, compact means closed and bounded.*

PROOF. The idea here is that (1) is an abstract definition for the notion of compactness, then (2,3,4) are elementary properties, and then (5) needs some study. To be more precise, in order to establish (5), the first observation, which is quite standard to prove, is that a product of closed intervals $\prod_i [a_i, b_i] \subset \mathbb{R}^N$ is indeed compact. But then each closed ball $B \subset \mathbb{R}^N$ follows to be compact, and this shows that closed and bounded is compact. As for the converse, compact is closed and bounded, this is clear. □

We can talk as well about connectedness:

THEOREM 10.4. *We can talk about connected sets $E \subset X$, with as basic result here, the fact that if f is continuous and E is connected, then $f(E)$ is connected.*

PROOF. This is again something quite clear, as in the $X = \mathbb{R}$ case. □

Finally, there are actually some subtleties in connection with the notion of connectedness, because even if we assume that $E \subset X$ is connected, the question of understanding how many "holes" does E have appears at $N \geq 2$, and many things can be said here.

10b. Functions, continuity

With the above general theory developed, we can now take $X = \mathbb{R}^N$, and talk about continuous functions $f : \mathbb{R}^N \to \mathbb{R}^M$, and their basic properties. We first have:

THEOREM 10.5. *Assuming that $f : \mathbb{R}^N \to \mathbb{R}^M$ 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 follows indeed from the general theory developed above. □

In addition to this, there are a few other things that can be said about the continuity of the functions $f : \mathbb{R}^N \to \mathbb{R}^M$.

10c. Inverse functions

An interesting question is the study of the functions $f : \mathbb{R}^N \to \mathbb{R}^M$ which are locally or globally bijective, and their local or global inverses. Obviously, we must have $M = N$. There are a number of general results that can be established here.

10d. Some geometry

In what follows we will be mainly interested in developing some advanced analysis theory for the functions $f : \mathbb{R}^N \to \mathbb{R}^M$, for instance with results regarding differentiation and integration, but before that, as a first application of the theory developed above, we can talk about curves, surfaces or other manifolds $X \subset \mathbb{R}^N$. Indeed, such manifolds can be constructed, locally or globally, by using suitable functions $f : \mathbb{R}^N \to \mathbb{R}^M$.

10e. Exercises

Exercises.

CHAPTER 11

Differentiation

11a. Partial derivatives

Let us discuss what happens in several variables. At order 1, the situation is quite similar to the one in 1 variable, but this time involving matrices, as follows:

THEOREM 11.1. *Any function of several variables can be locally approximated as*

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

with A being the matrix of partial derivatives at x,

$$A = \left(\frac{\partial f_i}{\partial x_j}(x) \right)_{ij}$$

acting on the vectors $t \in \mathbb{C}^N$ by usual multiplication.

PROOF. This is indeed standard, by using a recurrence method, with respect to the number of dimensions. To be more precise, consider the matrix in the statement:

$$A = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \ldots & \frac{\partial f_1}{\partial x_N}(x) \\ \vdots & & \vdots \\ \frac{\partial f_N}{\partial x_1}(x) & \ldots & \frac{\partial f_N}{\partial x_N}(x) \end{pmatrix}$$

The idea is then that around x, our function behaves exactly as $t \to At$:

$$f \begin{pmatrix} x_1 + t_1 \\ \vdots \\ x_N + t_N \end{pmatrix} \simeq f \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \ldots & \frac{\partial f_1}{\partial x_N}(x) \\ \vdots & & \vdots \\ \frac{\partial f_N}{\partial x_1}(x) & \ldots & \frac{\partial f_N}{\partial x_N}(x) \end{pmatrix} \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix}$$

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

11b. Basic examples

Basic examples. Functions $f : \mathbb{R}^2 \to \mathbb{R}^2$ vs functions $f : \mathbb{C} \to \mathbb{C}$.

11c. The chain rule

Generally speaking, Theorem 11.1 is what you need to know for upgrading from calculus to multivariable calculus. As a standard result here, we have:

THEOREM 11.2. *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 11.1. To be more precise, consider a composition of functions, as follows:

$$f:\mathbb{R}^M\to\mathbb{R}^N\quad,\quad g:\mathbb{R}^K\to\mathbb{R}^M\quad,\quad f\circ g:\mathbb{R}^K\to\mathbb{R}^N$$

According to Theorem 11.1, the derivatives of these functions are certain linear maps, corresponding to certain rectangular matrices, as follows:

$$f'(g(x))\in M_{N\times M}(\mathbb{R})\quad,\quad g'(x)\in M_{M\times K}(\mathbb{R})\qquad(f\circ g)'(x)\in M_{N\times K}(\mathbb{R})$$

Thus, the formula in the statement makes sense. As for proof, this is done first in the one variable case, and then in general, with this being something quite routine. □

Regarding the higher derivatives, the situation here is more complicated.

Let us record, however, the following fundamental result, happening at order 2, and which does the job, the job in analysis being usually that of finding the minima or maxima:

THEOREM 11.3. *Associated to any numeric function of several variables*

$$f:\mathbb{R}^N\to\mathbb{R}$$

is its Hessian function, given by the following formula,

$$H(x)=\left(\frac{\partial^2 f}{\partial x_j\partial x_j}(x)\right)_{ij}$$

whose positivity properties are related to the local minima and maxima of f.

PROOF. This is quite standard, by using the same method as in the 1D case, namely building on Theorem 11.1, and approximating the function at order 2. □

We will be back to this in the next chapter.

11d. Kepler and Newton

As a main application of all the above, we have:

THEOREM 11.4. *The following happen:*

(1) *Planets and other celestial bodies move around the Sun on conics, that is, curves given by $P(x,y)=0$, with $P\in\mathbb{R}[x,y]$ being of degree 2.*
(2) *The conics are the curves which appear by cutting a 2-sided cone with a plane, and can be classified into ellipses, parabolas and hyperbolas.*

PROOF. Here (1) is something tough, due to Kepler and Newton, and (2), due to the ancient Greeks, is elementary. In what follows we will only need (2), but we will prove everything, meaning (1) as well, as a matter of principle. Let us mention too that the above statement is a bit informal, with the 3 viewpoints on the conics, coming from gravity, cutting cones and classification, agreeing in the non-degenerate case, $\deg P=2$, modulo some normalizations, and with some disagreements in the degenerate case, $\deg P\leq 1$. But the complete statement, including a full discussion of the normalizations and of the degenerate cases, being too long, we have preferred to formulate things as above, and for more we refer to the comments at the end of the proof. Getting started now:

(1) According to observations and calculations performed over the centuries, since the ancient times, and first formalized by Newton, following some groundbreaking work of Kepler, the force of attraction between two bodies of masses M,m is given by:

$$||F||=G\cdot\frac{Mm}{d^2}$$

Here d is the distance between the two bodies, and $G\simeq 6.674\times 10^{-11}$ is a constant. Now assuming that M is fixed at $0\in\mathbb{R}^3$, the force exterted on m positioned at $x\in\mathbb{R}^3$, regarded as a vector $F\in\mathbb{R}^3$, is given by the following formula:

$$F=-||F||\cdot\frac{x}{||x||}=-\frac{GMm}{||x||^2}\cdot\frac{x}{||x||}=-\frac{GMmx}{||x||^3}$$

But $F=ma=m\ddot{x}$, with $a=\ddot{x}$ being the acceleration, second derivative of the position, so the equation of motion of m, assuming that M is fixed at 0, is:

$$\ddot{x}=-\frac{GMx}{||x||^3}$$

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:

$$\left(a + \frac{K}{r^2}\right)\cos\theta = b\sin\theta$$

$$\left(a + \frac{K}{r^2}\right)\sin\theta = -b\cos\theta$$

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:

$$\begin{aligned} x^2 + y^2 &= \frac{c^2(\cos^2\theta + \sin^2\theta)}{(1 + \varepsilon \cos\theta + \delta \sin\theta)^2} \\ &= \frac{c^2}{(1 + \varepsilon \cos\theta + \delta \sin\theta)^2} \end{aligned}$$

On the other hand, these two functions satisfy as well the following formula:

$$\begin{aligned} (\varepsilon x + \delta y - c)^2 &= \frac{c^2\big(\varepsilon \cos\theta + \delta \sin\theta - (1 + \varepsilon \cos\theta + \delta \sin\theta)\big)^2}{(1 + \varepsilon \cos\theta + \delta \sin\theta)^2} \\ &= \frac{c^2}{(1 + \varepsilon \cos\theta + \delta \sin\theta)^2} \end{aligned}$$

We conclude that our coordinates x, y satisfy the following equation:

$$x^2 + y^2 = (\varepsilon x + \delta y - c)^2$$

But what we have here is an equation of a conic, and this ends the proof of the first assertion of the theorem. Which is not bad at all, because you probably know now more classical mechanics than the average math nerd. And if you don't believe me, ask around your fellow math students, or even my fellow math professors, about who knows how to do the Kepler problem on the spot, in 20 minutes. You will be surprised.

(7) Before getting into the mathematics of conics, a bit more physics. Astronomy and Kepler tell us that for planets the trajectory should be an ellipsis, and this can be deduced from what we have, the missing piece of math, which is elementary, being that of proving that a bounded non-degenerate conic must be an ellipsis. However, this will follow as well from the classification results below, so we will stop physics here.

(8) The classification of the conics, going back to the ancient Greeks, is standard. Consider indeed one of these conics:

$$C = \left\{(x, y) \in \mathbb{R}^2 \middle| P(x, y) = 0, \deg P \leq 2\right\}$$

By doing some suitable manipulations on the degree 2 polynomial $P \in \mathbb{R}[x, y]$, up to affine transformations of the curve, we can have this curve written in some simple, "standard" form, with standard depending a bit on you, matter of taste. But this standard form can only lead to the 3 cases in the statement, namely ellipses, parabolas and hyperbolas, up to degeneration, with the degenerate cases being the lines, double lines, points, empty set, and $\mathbb{R}^2$ itself, basically appearing when $\deg P \leq 1$.

(9) The fact that the conics appear by cutting a 2-sided cone with a plane is also elementary, and also known since the ancient Greeks. A first proof is by doing some abstract algebra, and verifying that the cut must be indeed a curve of degree 2. A second proof is by computing the cut in the various cases that might appear, depending on the angle of the plane with respect to the cone, with this leading to the curves found in (8), namely ellipses, parabolas and hyperbolas, up to degeneration.

(10) Summarizing, we are done with what was announced in the theorem, but there is still some discussion to be made, in relation with degeneration. With respect to what was found in (8), when cutting cones with a plane, if you want to get exactly the same list of conics, you have to allow the degenerate cone, of angle $180°$, which in practice means a plane, in order to have as examples the empty set, and $\mathbb{R}^2$ itself.

(11) Also in relation to what was found in (8), what comes out of gravity basically agrees, namely ellipses, parabolas and hyperbolas. However, at the level of degenerate examples, these are different, consisting of the point, and then of the segment, which corresponds to the object m falling into the object M on a perfectly straight line.

(12) Finally, there is as well a discussion 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 the ellipses, parabolas and hyperbolas. □

11e. Exercises

CHAPTER 12

Optimization

12a. 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 12.1. *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:

$$\begin{aligned} PDP^{-1}v_i &= PDe_i \\ &= P\lambda_i e_i \\ &= \lambda_i Pe_i \\ &= \lambda_i v_i \end{aligned}$$

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 12.2. *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. $\square$

In order to reach now to more advanced results, we can use the characteristic polynomial, which appears via the following fundamental result:

THEOREM 12.3. *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:

$$\begin{aligned} P_A &= P_{\lambda 1_{d_\lambda}} P_B \\ &= (\lambda - x)^{d_\lambda} P_B \end{aligned}$$

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 12.4. *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 12.3, 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 12.2 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 12.5. *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 12.6. *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 12.5. As for the equivalences in the statement, these are all standard, the idea for their proofs, along with some more theory, needed for using in practice the present result, being as follows:

(1) $\iff$ (2) This follows from Theorem 12.5.

(2) $\iff$ (3) This is standard, the double roots of P being roots of P'.

(3) $\iff$ (4) The idea here is that associated to any two polynomials P, Q is their resultant $R(P, Q)$, which checks whether P, Q have a common root. Let us write:

$$P = c(X - a_1) \dots (X - a_k)$$

$$Q = d(X - b_1) \dots (X - b_l)$$

We can define then the resultant as being the following quantity:

$$R(P, Q) = c^l d^k \prod_{ij} (a_i - b_j)$$

The point now, that we will explain as well, is that this is a polynomial in the coefficients of P, Q, with integer coefficients. Indeed, this can be checked as follows:

– We can expand the formula of $R(P, Q)$, and in what regards $a_1, \dots, a_k$, 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.

– We can then look what happens with respect to the remaining variables $b_1, \dots, b_l$, 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.

– Thus, we are led to the above conclusion, that $R(P, Q)$ is a polynomial in the coefficients of P, Q, with integer coefficients, and with the remark that the $c^l d^k$ factor is there for these latter coefficients to be indeed integers, instead of rationals.

Alternatively, let us write our two polynomials in usual form, as follows:

$$P = p_k X^k + \dots + p_1 X + p_0$$

$$Q = q_l X^l + \dots + q_1 X + q_0$$

The corresponding resultant appears then as the determinant of an associated matrix, having size $k + l$, and having 0 coefficients at the blank spaces, as follows:

$$R(P, Q) = \begin{vmatrix} p_k & & & q_l & & \\ \vdots & \ddots & & \vdots & \ddots & \\ p_0 & & p_k & q_0 & & q_k \\ & \ddots & \vdots & & \ddots & \vdots \\ & & p_0 & & & q_0 \end{vmatrix}$$

(4) $\iff$ (5) Once again this is something standard, the idea here being that the discriminant $\Delta(P)$ of a polynomial $P \in \mathbb{C}[X]$ is, modulo scalars, the resultant $R(P, P')$. To be more precise, let us write our polynomial as follows:

$$P(X) = cX^N + dX^{N-1} + \dots$$

Its discriminant is then defined as being the following quantity:

$$\Delta(P) = \frac{(-1)^{\binom{N}{2}}}{c} R(P, P')$$

This is a polynomial in the coefficients of P, with integer coefficients, 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. □

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 12.7. *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 12.6. There are of course some other proofs as well, for instance by putting the matrix in Jordan form. □

Let us go back to the main problem raised by the diagonalization procedure, namely the computation of the roots of characteristic polynomials. We have here:

THEOREM 12.8. *The complex eigenvalues of a matrix $A \in M_N(\mathbb{C})$, counted with multiplicities, have the following properties:*

(1) *Their sum is the trace.*
(2) *Their product is the determinant.*

PROOF. Consider indeed the characteristic polynomial P of the matrix:

$$\begin{aligned} P(X) &= \det(A - X1_N) \\ &= (-1)^N X^N + (-1)^{N-1} Tr(A) X^{N-1} + \ldots + \det(A) \end{aligned}$$

We can factorize this polynomial, by using its N complex roots, and we obtain:

$$\begin{aligned} P(X) &= (-1)^N(X-\lambda_1)\ldots(X-\lambda_N) \\ &= (-1)^N X^N+(-1)^{N-1}\left(\sum_i\lambda_i\right)X^{N-1}+\ldots+\prod_i\lambda_i \end{aligned}$$

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

Regarding now the intermediate terms, we have here:

THEOREM 12.9. *Assume that $A\in M_N(\mathbb{C})$ has eigenvalues $\lambda_1,\ldots,\lambda_N\in\mathbb{C}$, counted with multiplicities. The basic symmetric functions of these eigenvalues, namely*

$$c_k=\sum_{i_1<\ldots<i_k}\lambda_{i_1}\ldots\lambda_{i_k}$$

are then given by the fact that the characteristic polynomial of the matrix is:

$$P(X)=(-1)^N\sum_{k=0}^{N}(-1)^kc_kX^k$$

Moreover, all symmetric functions of the eigenvalues, such as the sums of powers

$$d_s=\lambda_1^s+\ldots+\lambda_N^s$$

appear as polynomials in these characteristic polynomial coefficients c_k.

PROOF. These results can be proved by doing some algebra, as follows:

(1) Consider indeed the characteristic polynomial P of the matrix, factorized by using its N complex roots, taken with multiplicities. By expanding, we obtain:

$$\begin{aligned} P(X) &= (-1)^N(X-\lambda_1)\ldots(X-\lambda_N) \\ &= (-1)^N X^N+(-1)^{N-1}\left(\sum_i\lambda_i\right)X^{N-1}+\ldots+\prod_i\lambda_i \\ &= (-1)^N X^N+(-1)^{N-1}c_1X^{N-1}+\ldots+(-1)^0c_N \\ &= (-1)^N\left(X^N-c_1X^{N-1}+\ldots+(-1)^Nc_N\right) \end{aligned}$$

With the convention $c_0=1$, we are led to the conclusion in the statement.

(2) This is something standard, coming by doing some abstract algebra. Working out the formulae for the sums of powers $d_s=\sum_i\lambda_i^s$, at small values of the exponent $s\in\mathbb{N}$, is an excellent exercise, which shows how to proceed in general, by recurrence. □

Let us go back now to the diagonalization question. Here is a key result:

THEOREM 12.10. *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:

$$\begin{aligned} A^* &= (UDU^*)^* \\ &= UD^*U^* \\ &= UDU^* \\ &= A \end{aligned}$$

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. □

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 12.11. *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 12.10 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 12.10.

(5) $\implies$ (1) Assuming $A = UDU^*$ is as in the statement, with $U \in U_N$, and with $D \in M_N(\mathbb{R}_+)$ being 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 12.12. *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 12.11, by adding the above various extra assumptions, or from the proof of Theorem 12.11, by modifying where needed. □

Let us discuss now the case of the unitary matrices. We have here:

THEOREM 12.13. *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. □

The self-adjoint matrices and the unitary matrices are particular cases of the general notion of a "normal matrix", and we have here:

THEOREM 12.14. *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 12.13. □

As a first application, we have the following result:

THEOREM 12.15. *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 12.14, 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 12.16. *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 12.17. *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 12.16, applied on the complement of the 0-eigenvectors. □

12b. The Hessian, positivity

We recall from the previous chapter that we have:

THEOREM 12.18. *Associated to any numeric function of several variables*

$$f : \mathbb{R}^N \to \mathbb{R}$$

is its Hessian function, given by the following formula,

$$H(x) = \left(\frac{\partial^2 f}{\partial x_j \partial x_j}(x) \right)_{ij}$$

whose positivity properties are related to the local minima and maxima of f.

PROOF. This is indeed standard, by using the same method as in the 1D case. □

With our linear algebra knowledge, we can say more about this.

12c. The gradient method

12d. Lagrange multipliers

12e. Exercises

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

Measure theory

13a. Discrete measures

13b. Continuous measures

13c. Integration, Fubini

13d. Probability basics

With the idea in mind of doing things a bit abstractly, our starting point will be:

DEFINITION 13.1. *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 $f : X \to \mathbb{R}$.

Next in line, we need to talk about independence. We can do this as follows:

DEFINITION 13.2. *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. 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:

PROPOSITION 13.3. *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 13.4. *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 Proposition 13.3 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. □

More concretely now, we have the following definition:

DEFINITION 13.5. *The Poisson law of parameter* 1 *is the following measure,*

$$p_1 = \frac{1}{e} \sum_{k\in\mathbb{N}} \frac{\delta_k}{k!}$$

and the Poisson law of parameter $t > 0$ *is the following measure,*

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

with the letter "p" standing for Poisson.

Observe that these laws have indeed mass 1, as they should, due to:

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

We will see in the moment why these measures appear a bit everywhere, in discrete contexts, the reasons behind this coming from the Poisson Limit Theorem (PLT). Let us first develop some general theory. We first have the following result:

THEOREM 13.6. *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 13.7. *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 13.8. *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 13.6, 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 13.9 (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. □

13e. Exercises

CHAPTER 14

Integration theory

14a. Multiple integrals

14b. Change of variables

We recall that we have the following key result:

PROPOSITION 14.1. *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 well, 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. □

In several variables now, the result is as follows:

THEOREM 14.2. *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*

$$J_\varphi(t) = \det\left[\left(\frac{\partial\varphi_i}{\partial x_j}(x)\right)_{ij}\right]$$

and with this generalizing the formula from Proposition 14.1.

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 14.1, 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, the proof is quite similar, the idea being that of assuming that the change of variables is linear, and by using the definition of the determinant as a volume. More specifically, this follows from the material developed in chapter 9. □

14c. Spherical coordinates

In what regards the applications of Theorem 14.2, these often come via:

THEOREM 14.3. *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:

$$\begin{aligned} J &= \begin{vmatrix} \frac{\partial r\cos t}{\partial r} & \frac{\partial r\cos t}{\partial t} \\ \frac{\partial r\sin t}{\partial r} & \frac{\partial r\sin t}{\partial t} \end{vmatrix} \\ &= \begin{vmatrix} \cos t & -r\sin t \\ \sin t & r\cos t \end{vmatrix} \\ &= r\cos^2 t + r\sin^2 t \\ &= r \end{aligned}$$

Thus, we have indeed the formula in the statement. □

In 3 dimensions the formula is similar, as follows:

THEOREM 14.4. *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 spherical coordinate formula in N dimensions. The formula here, which generalizes those at $N=2,3$, is as follows:

THEOREM 14.5. *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:

$$J(r,t)=r^{N-1}\sin^{N-2}t_1\sin^{N-3}t_2\ \ldots\ \sin^2 t_{N-3}\sin t_{N-2}$$

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 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 14.6. *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:

$$\begin{aligned}I_p &= \frac{p-1}{p} I_{p-2}\\ &= \frac{p-1}{p}\cdot\frac{p-3}{p-2} I_{p-4}\\ &= \frac{p-1}{p}\cdot\frac{p-3}{p-2}\cdot\frac{p-5}{p-4} I_{p-6}\\ &\vdots\\ &= \frac{p!!}{(p+1)!!} I_{1-\varepsilon(p)}\end{aligned}$$

On the other hand, at $p = 0$ we have the following formula:

$$I_0 = \int_0^{\pi/2} 1\, dt = \frac{\pi}{2}$$

Also, at $p = 1$ we have the following formula:

$$I_1 = \int_0^{\pi/2} \cos t \, dt = 1$$

Thus, we obtain the result, by recurrence. As for the second formula, regarding $\sin t$, this follows from the first formula, with the following change of variables:

$$t = \frac{\pi}{2} - s$$

Thus, we have proved both formulae in the statement. □

We can now compute the volume of the sphere, as follows:

THEOREM 14.7. *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 the convention

$$N!! = (N-1)(N-3)(N-5)\ldots$$

with the product ending at 2 if N is odd, and ending at 1 if N is even.

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:

$$\begin{aligned}
& V^+ \\
= \ & \int_{B^+} 1 \\
= \ & \int_0^1 \int_0^{\pi/2} \ldots \int_0^{\pi/2} r^{N-1} \sin^{N-2} t_1 \ldots \sin t_{N-2} \, dr dt_1 \ldots dt_{N-1} \\
= \ & \int_0^1 r^{N-1} \, dr \int_0^{\pi/2} \sin^{N-2} t_1 \, dt_1 \ldots \int_0^{\pi/2} \sin t_{N-2} dt_{N-2} \int_0^{\pi/2} 1 dt_{N-1} \\
= \ & \frac{1}{N} \times \left(\frac{\pi}{2}\right)^{[N/2]} \times \frac{(N-2)!!}{(N-1)!!} \cdot \frac{(N-3)!!}{(N-2)!!} \cdots \frac{2!!}{3!!} \cdot \frac{1!!}{2!!} \cdot 1 \\
= \ & \frac{1}{N} \times \left(\frac{\pi}{2}\right)^{[N/2]} \times \frac{1}{(N-1)!!} \\
= \ & \left(\frac{\pi}{2}\right)^{[N/2]} \frac{1}{(N+1)!!}
\end{aligned}$$

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. □

There are many applications of the above formula, as we will see in what follows. As main particular cases of the above formula, we have:

THEOREM 14.8. *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 14.7, 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 $V=\frac{\pi^2}{4}\cdot\frac{16}{8}=\frac{\pi^2}{2}$. □

In order to obtain estimates for the volumes, in the large N limit, we can use:

THEOREM 14.9. *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 to use the integral of the logarithm. Indeed, by using a Riemann sum, we have:

$$\begin{aligned}\int_0^1 \log x\,dx &\simeq \frac{1}{N}\sum_{k=1}^{N}\log\left(\frac{k}{N}\right)\\ &= \frac{1}{N}\left(\sum_{k=1}^{N}\log k\right)-\log N\\ &= \frac{\log N!}{N}-\log N\end{aligned}$$

On the other hand, our integral can be explicitely computed, as follows:

$$\begin{aligned}\int_0^1 \log x\,dx &= \Big[x\log x-x\Big]_0^1\\ &= -1-0\\ &= -1\end{aligned}$$

We are therefore led to the following formula:

$$\frac{\log N!}{N}\simeq\log N-1$$

Now by exponentiating, this leads to the following estimate:

$$\big(N!\big)^{1/N}\simeq\frac{N}{e}$$

With a bit more care, we obtain, a bit in the same way:

$$N!\sim\left(\frac{N}{e}\right)^N$$

The point now is that, with even more care, by reviewing the Riemann sums used above, we can obtain the formula in the statement. □

With the above formula in hand, we have many useful applications, such as:

PROPOSITION 14.10. *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 14.11. *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 14.12. *The volume of the unit sphere in $\mathbb{R}^N$ is given by*

$$V \simeq \left(\frac{2\pi e}{N}\right)^{N/2}\frac{1}{\sqrt{\pi N}}$$

in the $N \to \infty$ limit.

PROOF. We use the formula for V found in Theorem 14.7, 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 14.7, we can compute in the same way the area of the sphere, the result being as follows:

THEOREM 14.13. *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 14.7, as follows:

(1) We can either use the "pizza" argument from 1 dimension, 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 14.7 for V, this gives the result.

(2) Or, we can start the computation in the same way as we started the proof of Theorem 14.7, the beginning of this computation being as follows:

$$\begin{aligned} vol(S^+) &= \int_{S^+} 1 \\ &= \int_0^{\pi/2} \ldots \int_0^{\pi/2} \sin^{N-2} t_1 \ldots \sin t_{N-2}\, dt_1 \ldots dt_{N-1} \end{aligned}$$

Now by comparing with the beginning of the proof of Theorem 14.7, 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:

$$vol(S) = N \cdot vol(B)$$

But this means $A = N \cdot V$, and together with the formula in Theorem 14.7 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. □

14d. Gaussian laws

As a main application of all the theory developed in the above, and as the best calculus formula ever, we can now compute the Gauss integral, as follows:

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:

$$\begin{aligned} I^2 &= \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-x^2-y^2} dxdy \\ &= \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r dr dt \\ &= 2\pi \int_0^{\infty} \left(-\frac{e^{-r^2}}{2} \right)' dr \\ &= 2\pi \left[0 - \left(-\frac{1}{2} \right) \right] \\ &= \pi \end{aligned}$$

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

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. At the level of basic facts, we have:

THEOREM 14.16. *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 14.17 (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 the computation of the arbitrary integrals over the sphere, and that of the hyperspherical laws.

We will need a technical result extending Proposition 14.6, as follows:

THEOREM 14.18. *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. 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 14.6, and we obtain the result:

$$\begin{aligned}I_{pq} &= \frac{p!!q!!}{(p+q)!!} I_{p+q}\\ &= \frac{p!!q!!}{(p+q)!!}\left(\frac{\pi}{2}\right)^{\varepsilon(p+q)}\frac{(p+q)!!}{(p+q+1)!!}\\ &= \left(\frac{\pi}{2}\right)^{\varepsilon(p)\varepsilon(q)}\frac{p!!q!!}{(p+q+1)!!}\end{aligned}$$

Observe that this gives the result for p even as well, by symmetry. Indeed, we have $I_{pq} = I_{qp}$, by using the following change of variables:

$$t = \frac{\pi}{2} - s$$

In the remaining case now, where both p, q are odd, we can use once again the formula $pI_{p-1,q+1} = qI_{p+1,q-1}$ established above, and the recurrence goes as follows:

$$\begin{aligned} I_{pq} &= \frac{q-1}{p+1} I_{p+2,q-2} \\ &= \frac{q-1}{p+1} \cdot \frac{q-3}{p+3} I_{p+4,q-4} \\ &= \frac{q-1}{p+1} \cdot \frac{q-3}{p+3} \cdot \frac{q-5}{p+5} I_{p+6,q-6} \\ &= \vdots \\ &= \frac{p!!q!!}{(p+q-1)!!} I_{p+q-1,1} \end{aligned}$$

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 14.19. *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 14.18, which reads:

$$\begin{aligned}\int_0^{\pi/2} \cos^p t \sin^q t \, dt &= \left(\frac{\pi}{2}\right)^{\varepsilon(p)\varepsilon(q)} \frac{p!!q!!}{(p+q+1)!!} \\ &= \frac{p!!q!!}{(p+q+1)!!}\end{aligned}$$

Indeed, this formula computes the integral in the statement over the first quadrant. But since the exponents $p, q \in \mathbb{N}$ are assumed to be even, the integrals over the other quadrants are given by the same formula, so when averaging we obtain the result.

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 14.13, 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:

$$\begin{aligned}I' = \int_0^{\pi/2} \ldots \int_0^{\pi/2} \quad & (\cos t_1)^{k_1} (\sin t_1 \cos t_2)^{k_2} \\ & \vdots \\ & (\sin t_1 \sin t_2 \ldots \sin t_{N-2} \cos t_{N-1})^{k_{N-1}} \\ & (\sin t_1 \sin t_2 \ldots \sin t_{N-2} \sin t_{N-1})^{k_N} \\ & \sin^{N-2} t_1 \sin^{N-3} t_2 \ldots \sin^2 t_{N-3} \sin t_{N-2} \\ & dt_1 \ldots dt_{N-1}\end{aligned}$$

By rearranging the terms, we obtain:

$$\begin{aligned} I' &= \int_0^{\pi/2} \cos^{k_1} t_1 \sin^{k_2+\ldots+k_N+N-2} t_1 \, dt_1 \\ &\int_0^{\pi/2} \cos^{k_2} t_2 \sin^{k_3+\ldots+k_N+N-3} t_2 \, dt_2 \\ &\vdots \\ &\int_0^{\pi/2} \cos^{k_{N-2}} t_{N-2} \sin^{k_{N-1}+k_N+1} t_{N-2} \, dt_{N-2} \\ &\int_0^{\pi/2} \cos^{k_{N-1}} t_{N-1} \sin^{k_N} t_{N-1} \, dt_{N-1} \end{aligned}$$

Now by using the above-mentioned formula at $N = 2$, this gives:

$$\begin{aligned} I' &= \frac{k_1!!(k_2+\ldots+k_N+N-2)!!}{(k_1+\ldots+k_N+N-1)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(N-2)} \\ &\frac{k_2!!(k_3+\ldots+k_N+N-3)!!}{(k_2+\ldots+k_N+N-2)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(N-3)} \\ &\vdots \\ &\frac{k_{N-2}!!(k_{N-1}+k_N+1)!!}{(k_{N-2}+k_{N-1}+l_N+2)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(1)} \\ &\frac{k_{N-1}!!k_N!!}{(k_{N-1}+k_N+1)!!} \left(\frac{\pi}{2}\right)^{\varepsilon(0)} \end{aligned}$$

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. □

As an application of all this, we have the following result:

THEOREM 14.20. *The moments of the hyperspherical variables are*

$$\int_{S_{\mathbb{R}}^{N-1}} x_i^k dx = \frac{(N-1)!!k!!}{(N+k-1)!!}$$

and the normalized hyperspherical variables

$$y_i = \frac{x_i}{\sqrt{N}}$$

become normal and independent with $N \to \infty$.

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

(1) The formula in the statement follows from the general integration formula over the sphere, established above. Indeed, this formula gives:

$$\int_{S_{\mathbb{R}}^{N-1}} x_i^k dx = \frac{(N-1)!!k!!}{(N+k-1)!!}$$

Now observe that with $N \to \infty$ we have the following estimate:

$$\begin{aligned}\int_{S_{\mathbb{R}}^{N-1}} x_i^k dx &= \frac{(N-1)!!}{(N+k-1)!!} \times k!! \\ &\simeq N^{k/2} k!! \\ &= N^{k/2} M_k(g_1)\end{aligned}$$

Thus, the variables $y_i = \frac{x_i}{\sqrt{N}}$ become normal with $N \to \infty$.

(2) As for the asymptotic independence result, this is standard as well, once again by using Theorem 14.19, for computing mixed moments, and taking the $N \to \infty$ limit. □

14e. Exercises

CHAPTER 15

Partial integration

15a. Vector products

Our goal here is to discuss a number of remarkable integration results in $N = 1, 2, 3$ dimensions, which are useful in relation with various questions from physics.

Let us start with a standard 3D construction, as follows:

DEFINITION 15.1. *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:

$$||x \times y|| \sim ||x||, ||y||$$

$$x \times x = 0$$

$$e_1 \times e_2 = e_3$$

So, here's how it works. We're looking for a vector $x \times y$ whose length is proportional to those of x, y. But now 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're almost there, it's just a matter of choosing the orientation, and this comes either from the right-hand rule (or perhaps left-hand rule, do I remember right?) or 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.

It is also good to know that $x \times y$ exists only in 3 dimensions, with our only tool in $N \neq 3$ dimensions being the usual $< x, y >$. This is actually quite interesting for us, in relation with various questions from physics, but more on this later.

15b. Functions, derivatives

Let us start with a standard definition, immersing us into 3D problematics, as follows:

DEFINITION 15.2. *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:*

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix}$$

By using ∇, we can talk about the divergence of a function $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$, as being

$$< \nabla, \varphi >= \left\langle \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix}, \begin{pmatrix} \varphi_x \\ \varphi_y \\ \varphi_z \end{pmatrix} \right\rangle = \frac{\partial \varphi_x}{\partial x} + \frac{\partial \varphi_y}{\partial y} + \frac{\partial \varphi_z}{\partial z}$$

as well as about the curl of the same function $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$, as being

$$\nabla \times \varphi = \begin{vmatrix} u_x & \frac{\partial}{\partial x} & \varphi_x \\ u_y & \frac{\partial}{\partial y} & \varphi_y \\ u_z & \frac{\partial}{\partial z} & \varphi_z \end{vmatrix} = \begin{pmatrix} \frac{\partial \varphi_z}{\partial y} - \frac{\partial \varphi_y}{\partial z} \\ \frac{\partial \varphi_x}{\partial z} - \frac{\partial \varphi_z}{\partial x} \\ \frac{\partial \varphi_y}{\partial x} - \frac{\partial \varphi_x}{\partial y} \end{pmatrix}$$

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.

15c. Gauss, Green, Stokes

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 15.3. *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 found in any calculus book.

(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 found in any calculus book. □

15d. Magnetic fields

All the above applies to basic questions in electromagnetism.

15e. Exercises

Exercises.

CHAPTER 16

Infinite dimensions

16a. Quantum mechanics

Quantum mechanics.

16b. Operators, matrices

Heisenberg's idea was to use infinite matrices, with some sort of frequencies $\pm 1/\lambda \in \mathbb{R}$ as entries. However, in view of many other things, including the Schrödinger finding that quantum mechanics naturally lives over $\mathbb{C}$, we would like in fact our infinite matrices to be over the complex numbers. So, we are led into the following question:

QUESTION 16.1. *How do the matrices $A \in M_\infty(\mathbb{C})$ act on the vectors $v \in \mathbb{C}^\infty$?*

This is something quite tricky. The problem is that the matrices $A \in M_\infty(\mathbb{C})$ do not always act on the vectors $v \in \mathbb{C}^\infty$. As an example, check this out:

$$\begin{pmatrix} 1 & 2 & 3 & \ldots \\ 2 & 3 & 4 & \ldots \\ 3 & 4 & 5 & \ldots \\ \vdots & \vdots & \vdots & \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{pmatrix} = ?$$

This is actually something doubly wrong, first because the vector Av is not well-defined, but then also since v itself is not a good vector, its norm being $||v|| = \infty$.

In order to fix all this, let us start with fixing the vector space $\mathbb{C}^\infty$. We would like to replace it with its subspace $H = l^2(\mathbb{N})$ consisting of vectors having finite norm. But this being said, taking a look at what Schrödinger was saying too, why not including in our theory spaces like $H = L^2(\mathbb{R}^3)$ too. We are led in this way into:

DEFINITION 16.2. *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 scalar products, written $< x, y >$ and being linear at left, is one among others, quite often used by mathematicians.

Getting back now to work, and to Definition 16.2, in its entirety, there is some mathematics encapsulated there, needing some discussion. First, we have:

THEOREM 16.3. *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. This can be done in several steps, as follows:

(1) Given a vector $x \in \mathbb{C}^I$, let us define its norm by the following formula:

$$||x|| = \sqrt{\sum_i |x_i|^2}$$

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:

$$| < x, y > | \leq ||x|| \cdot ||y||$$

(4) In order to prove this inequality, consider the following quantity, depending on a real variable $t \in \mathbb{R}$, and on a variable on the unit circle, $w \in \mathbb{T}$:

$$f(t) = ||twx + y||^2$$

By developing f, we see that this is a degree 2 polynomial in t:

$$\begin{aligned} f(t) &= < twx + y, twx + y > \\ &= t^2 < x, x > + tw < x, y > + t\bar{w} < y, x > + < y, y > \\ &= t^2||x||^2 + 2tRe(w < x, y >) + ||y||^2 \end{aligned}$$

Since f is obviously positive, its discriminant must be negative:

$$4Re(w < x, y >)^2 - 4||x||^2 \cdot ||y||^2 \leq 0$$

But this is equivalent to the following condition:

$$|Re(w < x, y >)| \leq ||x|| \cdot ||y||$$

Now the point is that we can arrange for the number $w \in \mathbb{T}$ to be such that the quantity $w < x, y >$ is real. Thus, we obtain, as desired:

$$| < x, y > | \leq ||x|| \cdot ||y||$$

(5) As a side remark here, observe that the equality case happens precisely when the discriminant of f vanishes, so when f has a root, and so when x, y are proportional.

(6) 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.

(7) Also, $< x, y >$ is surely a scalar product on this vector space, because all the conditions for a scalar product, which are as follows, are satisfied:

$* < x, y >$ is linear in x, and antilinear in y.

$* \overline{< x, y >} =< y, x >$, for any x, y.

$* < x, x >> 0$, for any $x \neq 0$.

(8) 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 16.4. *Given an arbitrary space X with a positive measure μ on it, the space of square-summable complex functions on it, namely*

$$L^2(X) = \left\{ f : X \to \mathbb{C} \middle| \int_X |f(x)|^2 \, d\mu(x) < \infty \right\}$$

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 16.3, as follows:

(1) The proof of the first, and main assertion is something perfectly similar to the proof of Theorem 16.3, 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 $||f - g|| = 0$.

(2) 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 16.3. □

As a third and last theorem about Hilbert spaces, that we will need, we have:

THEOREM 16.5. *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 one 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 for instance at the standard basis of $l^2(\mathbb{N})$. 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.

(3) 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. □

Observe that, while the space $L^2(\mathbb{R})$ is in theory separable, in practice this is not really the case, because the orthogonalization of the Weierstrass basis $\{x^n\}_{n\in\mathbb{N}}$ is something quite complicated. More on this later, when we will really need such things.

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 16.6. *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, well-known by linear algebra in finite dimensions, and the proof in general is similar. As for the last assertion, which is more tricky, in finite dimensions we have of course $B(H) = \mathcal{L}(H) = M_I(\mathbb{C})$. However, in infinite dimensions, we have matrices not producing operators, as for instance:

$$M = \begin{pmatrix} 1 & 1 & \ldots \\ 1 & 1 & \ldots \\ \vdots & \vdots & \end{pmatrix}$$

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 16.7. *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 $x=e_i$, $y=e_j$ gives the formula $(M^*)_{ij}=\overline{M}_{ji}$ in the statement. □

So, this was for the basics of operator theory, extending the basics of linear algebra. For more on all this, including full proofs for certain things in the above, you can check any book labelled functional analysis, or operator theory, or operator algebras.

16c. Schrödinger equation

We recall that Schrödinger came upon:

Claim 16.8 (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 the modified Planck constant, and V the Coulomb potential of the proton. The same holds for movements of the electron under an arbitrary 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. There might be of course some speculations to be made here, but passed that, this is certainly not your easy to decipher equation. This is how things are.

Following Heisenberg and Schrödinger, and then especially Dirac, who did the axiomatization work, we have the following abstract definition, based on the above:

DEFINITION 16.9. *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 our previous "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 our previous "sandwiching" formula holds true, and with this having all sorts of interesting consequences, already discussed before. 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.

16d. The hydrogen atom

In order to solve the hydrogen atom, we first must reformulate the Schrödinger equation in spherical coordinates. And for this purpose, we will need:

THEOREM 16.10. *The Laplace operator in spherical coordinates is:*

$$\Delta = \frac{1}{r^2} \cdot \frac{d}{dr}\left(r^2 \cdot \frac{d}{dr}\right) + \frac{1}{r^2 \sin s} \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d}{ds}\right) + \frac{1}{r^2 \sin^2 s} \cdot \frac{d^2}{dt^2}$$

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:

$$\begin{aligned}\frac{d^2f}{dx_i^2} &= \sum_j \frac{d}{dx_i}\left(\frac{df}{dy_j}\cdot\frac{dy_j}{dx_i}\right)\\ &= \sum_j \frac{d}{dx_i}\left(\frac{df}{dy_j}\right)\cdot\frac{dy_j}{dx_i}+\frac{df}{dy_j}\cdot\frac{d}{dx_i}\left(\frac{dy_j}{dx_i}\right)\\ &= \sum_j \left(\sum_k \frac{d}{dy_k}\cdot\frac{dy_k}{dx_i}\right)\left(\frac{df}{dy_j}\right)\cdot\frac{dy_j}{dx_i}+\frac{df}{dy_j}\cdot\frac{d^2y_j}{dx_i^2}\\ &= \sum_{jk}\frac{d^2f}{dy_kdy_j}\cdot\frac{dy_k}{dx_i}\cdot\frac{dy_j}{dx_i}+\sum_j\frac{df}{dy_j}\cdot\frac{d^2y_j}{dx_i^2}\end{aligned}$$

(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:

$$\begin{aligned}\Delta f &= \sum_{ijk}\frac{d^2f}{dy_kdy_j}\cdot\frac{dy_k}{dx_i}\cdot\frac{dy_j}{dx_i}+\sum_{ij}\frac{df}{dy_j}\cdot\frac{d^2y_j}{dx_i^2}\\ &= \sum_{ijk}A_{ki}A_{ji}\frac{d^2f}{dy_kdy_j}+\sum_{ij}\frac{d^2y_j}{dx_i^2}\cdot\frac{df}{dy_j}\\ &= \sum_{jk}(AA^t)_{jk}\frac{d^2f}{dy_kdy_j}+\sum_j\Delta(y_j)\frac{df}{dy_j}\end{aligned}$$

(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:

$$\Delta f = Tr(AA^tH_y(f))+<\Delta(y),\nabla_y(f)>$$

(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 $(A^{-1})^tA^{-1}$ 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. □

We can now reformulate the Schrödinger equation in spherical coordinates, and then separate the variables, which leads to a radial and angular equation, as follows:

THEOREM 16.11. *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 16.10, 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. □

Next, the idea is that both the radial and angular equations can be solved, and this leads to the standard solution of the hydrogen atom.

16e. Exercises

Exercises.

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] N.W. Ashcroft and N.D. Mermin, Solid state physics, Saunders College Publ. (1976).

[7] M.F. Atiyah, The geometry and physics of knots, Cambridge Univ. Press (1990).

[8] T. Banica, Linear algebra and group theory (2022).

[9] T. Banica, Introduction to quantum mechanics (2023).

[10] G.K. Batchelor, An introduction to fluid dynamics, Cambridge Univ. Press. (1967).

[11] R.J. Baxter, Exactly solved models in statistical mechanics, Academic Press (1982).

[12] I. Bengtsson and K. Życzkowski, Geometry of quantum states, Cambridge Univ. Press (2006).

[13] S.M. Carroll, Spacetime and geometry, Cambridge Univ. Press (2004).

[14] P.M. Chaikin and T.C. Lubensky, Principles of condensed matter physics, Cambridge Univ. Press (1995).

[15] V. Chari and A. Pressley, A guide to quantum groups, Cambridge Univ. Press (1994).

[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] P.A. Davidson, Introduction to magnetohydrodynamics, Cambridge Univ. Press (2001).

[20] P.A.M. Dirac, Principles of quantum mechanics, Oxford Univ. Press (1930).

[21] M.P. do Carmo, Differential geometry of curves and surfaces, Dover (1976).

[22] M.P. do Carmo, Riemannian geometry, Birkhäuser (1992).

[23] S. Dodelson, Modern cosmology, Academic Press (2003).

[24] R. Durrett, Probability: theory and examples, Cambridge Univ. Press (1990).

[25] L.C. Evans, Partial differential equations, AMS (1998).

[26] W. Feller, An introduction to probability theory and its applications, Wiley (1950).

[27] E. Fermi, Thermodynamics, Dover (1937).

[28] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics I: mainly mechanics, radiation and heat, Caltech (1963).

[29] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics II: mainly electromagnetism and matter, Caltech (1964).

[30] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics III: quantum mechanics, Caltech (1966).

[31] R.P. Feynman and A.R. Hibbs, Quantum mechanics and path integrals, Dover (1965).

[32] H. Goldstein, C. Safko and J. Poole, Classical mechanics, Addison-Wesley (1980).

[33] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).

[34] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).

[35] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).

[36] D.J. Griffiths, Revolutions in twentieth-century physics, Cambridge Univ. Press (2012).

[37] S.J. Gustafson and I.M. Sigal, Mathematical concepts of quantum mechanics, Springer (2011).

[38] J. Harris, Algebraic geometry, Springer (1992).

[39] W.A. Harrison, Solid state theory, Dover (1970).

[40] W.A. Harrison, Electronic structure and the properties of solids, Dover (1980).

[41] R.A. Horn and C.R. Johnson, Matrix analysis, Cambridge Univ. Press (1985).

[42] K. Huang, Introduction to statistical physics, CRC Press (2001).

[43] K. Huang, Quantum field theory, Wiley (1998).

[44] K. Huang, Quarks, leptons and gauge fields, World Scientific (1982).

[45] K. Huang, Fundamental forces of nature, World Scientific (2007).

[46] J.E. Humphreys, Introduction to Lie algebras and representation theory, Springer (1972).

[47] J.D. Jackson, Classical electrodynamics, Wiley (1962).

[48] V.F.R. Jones, Subfactors and knots, CBMS Lecture Notes (1991).

[49] L.P. Kadanoff, Statistical physics: statics, dynamics and renormalization, World Scientific (2000).

[50] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).

[51] C. Kittel, Introduction to solid state physics, Wiley (1953).

[52] M. Kumar, Quantum: Einstein, Bohr, and the great debate about the nature of reality, Norton (2009).

[53] L.D. Landau and E.M. Lifshitz, Mechanics, Pergamon Press (1960).

[54] L.D. Landau and E.M. Lifshitz, The classical theory of fields, Addison-Wesley (1951).

[55] L.D. Landau and E.M. Lifshitz, Quantum mechanics: non-relativistic theory, Pergamon Press (1959).

[56] V.B. Berestetskii, E.M. Lifshitz and L.P. Pitaevskii, Quantum electrodynamics, Butterworth-Heinemann (1982).

[57] S. Lang, Algebra, Addison-Wesley (1993).

[58] P. Lax, Linear algebra and its applications, Wiley (2007).

[59] P. Lax, Functional analysis, Wiley (2002).

[60] P. Lax and M.S. Terrell, Calculus with applications, Springer (2013).

[61] P. Lax and M.S. Terrell, Multivariable calculus with applications, Springer (2018).

[62] M.L. Mehta, Random matrices, Elsevier (2004).

[63] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge Univ. Press (2000).

[64] R.K. Pathria and and P.D. Beale, Statistical mechanics, Elsevier (1972).

[65] A. Peres, Quantum theory: concepts and methods, Kluwer (1993).

[66] M. Peskin and D.V. Schroeder, An introduction to quantum field theory, CRC press (1995).

[67] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[68] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[69] W. Rudin, Functional analysis, McGraw-Hill (1973).

[70] W. Rudin, Fourier analysis on groups, Dover (1974).

[71] B.M. Peterson and B. Ryden, Foundations of astrophysics, Cambridge Univ. Press (2010).

[72] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[73] B. Ryden and R.W. Pogge, Interstellar and intergalactic medium, Cambridge Univ. Press (2021).

[74] J.J. Sakurai and J. Napolitano, Modern quantum mechanics, Cambridge Univ. Press (1985).

[75] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).

[76] M. Schwartz, Principles of electrodynamics, Dover (1972).

[77] J. Schwinger, L.L. DeRaad Jr., K.A. Milton and W.Y. Tsai, Classical electrodynamics, CRC Press (1998).

[78] J. Schwinger and B.H. Englert, Quantum mechanics: symbolism of atomic measurements, Springer (2001).

[79] J.P. Serre, Linear representations of finite groups, Springer (1977).

[80] I.R. Shafarevich, Basic algebraic geometry, Springer (1974).

[81] R. Shankar, Fundamentals of physics I: mechanics, relativity, and thermodynamics, Yale Univ. Press (2014).

[82] R. Shankar, Fundamentals of physics II: electromagnetism, optics, and quantum mechanics, Yale Univ. Press (2016).

[83] R. Shankar, Principles of quantum mechanics, Springer (1980).

[84] R. Shankar, Quantum field theory and condensed matter: an introduction, Cambridge Univ. Press (2017).

[85] J.H. Smith, Introduction to special relativity, Dover (1965).

[86] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[87] D.V. Voiculescu, K.J. Dykema and A. Nica, Free random variables, AMS (1992).

[88] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[89] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[90] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[91] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).

[92] S. Weinberg, Cosmology, Oxford Univ. Press (2008).

[93] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).

[94] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).

[95] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).

[96] J.M. Yeomans, Statistical mechanics of phase transitions, Oxford Univ. Press (1992).

[97] K. Yosida, Functional analysis, Springer (1965).

[98] J. Zinn-Justin, Path integrals in quantum mechanics, Oxford Univ. Press (2004).

[99] J. Zinn-Justin, Phase transitions and renormalization group, Oxford Univ. Press (2005).

[100] B. Zwiebach, A first course in string theory, Cambridge Univ. Press (2004).