# The study of functions

Teo Banica

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

2010 *Mathematics Subject Classification.* 97I20

*Key words and phrases.* Functions, Derivatives

Abstract. This is an introduction to the theory of mathematical functions. We first discuss various motivations and examples, ways of representing functions, and with a detailed look into the basic functions, namely polynomials, and $\sin, \cos, \exp, \log$. Then we discuss continuity, with the standard results on the subject, followed by derivatives, again with the standard results on the subject, notably with the Taylor formula, and its applications. Finally, we discuss integrals, with emphasis on what can be done with Riemann sums, and their relation with derivatives. We have a look as well, at the end, at the functions of several variables, whose study is more complicated.

# Preface

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

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

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

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

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

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

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

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

Needless all say, we will only provide an introduction to this. For more, that is, more theory of the real functions $f : \mathbb{R} \to \mathbb{R}$, or theory of the complex functions $f : \mathbb{C} \to \mathbb{C}$, or theory of the real multivariable functions $f : \mathbb{R}^N \to \mathbb{R}^M$, or even theory of the complex multivariable functions $f : \mathbb{C}^N \to \mathbb{C}^M$, that you should master too, in order for you science or engineering to be truly cutting edge, we will recommend some further reading.

Many thanks to my cats, and to the other cats in this world. You have no idea how many functions and theorems must be used, in order to properly catch a mouse, and in fact, the theory of functions $f : \mathbb{R} \to \mathbb{R}$ was formalized only quite late, in the history of mankind, after domesticating the cat, and observing his methods. Later, Newton discovered calculus too after closely monitoring his cat. And so on. And this book is no exception to the rule, the few points in the presentation which are original and useful, I hope, came in the same way, by observing those of us which are smarter, and faster.

*Cergy, May 2025*
*Teo Banica*

# Contents

# Part I

# Functions

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

CHAPTER 1

# Real numbers

## 1a. Numbers

Welcome to functions, and mathematical analysis. We denote by $\mathbb{N}$ the set of positive integers, $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$, with $\mathbb{N}$ standing for "natural". We will certainly 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 here for "zahlen", which is German for "numbers".

As you surely know, there are many questions in mathematics involving fractions, or quotients, which are defined as follows, and 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 of sets as follows, with each integer $a \in \mathbb{Z}$ being identified with the corresponding fraction $a/1$:

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

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, as follows:

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

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

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

In particular that any nonzero rational $r = a/b$ can be inverted, according to:

$$\left(\frac{a}{b}\right)^{-1} = \frac{b}{a}$$

In more formal terms, we can say that $\mathbb{Q}$ is the smallest field containing $\mathbb{N}$. But more on fields later, for the moment, let us not bother with such abstract things.

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 the following formula, which is obviously the wrong answer:

$$N = \frac{3!}{1!} = \frac{1 \cdot 2 \cdot 3}{1} = 6$$

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 that we have:

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

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 the following answer, which is correct:

$$\binom{3}{2} = \frac{3!}{2!1!} = \frac{6}{2} = 3$$

Thus, eventually, theorem proved, and doublechecked as well. □

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.

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 in final form (trust me), we have:

THEOREM 1.5. *We have the binomial formula*

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

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

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

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

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

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

In order to finish, it remains to compute the coefficients $C_k$. But, according to our product formula, $C_k$ is the number of choices for the $k$ needed $a$ variables among the $n$ available $a$ variables. Thus, according to Theorem 1.2, 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:

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

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

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

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

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

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

$$\vdots$$

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

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & , & 1 & & & & \\
 & & & 1 & , & 2 & , & 1 & & & \\
 & & 1 & , & 3 & , & 3 & , & 1 & & \\
 & 1 & , & 4 & , & 6 & , & 4 & , & 1 & \\
1 & , & 5 & , & 10 & , & 10 & , & 5 & , & 1 \\
 & & & & & \vdots & & & & &
\end{array}$$

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

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

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

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

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

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

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

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

By using the binomial formula, this formula becomes:

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

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

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*

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

*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,*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

THEOREM 1.11. *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 abstract definition for the real numbers, that we can certainly do some mathematics with. As a first general result now, which is something very useful, and puts us back into real life, and science and engineering, we have:

THEOREM 1.12. *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 non-trivial, even for the rationals $x\in\mathbb{Q}$ themselves, which require some work in order to be put in decimal form, the idea being as follows:

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

(2) Moreover, we have a second claim as well, namely that any expression of type $x=\pm a_1\ldots a_n.b_1b_2b_3\ldots\ldots$ 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 $\mathbb{Q}=A_x\sqcup B_x$. Since the set $A_x\cap\mathbb{Z}$ consists of integers, and is bounded from above by any element $q\in B_x$ of your choice, this set has a maximal element, that we can denote $[x]$:

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

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 $\mathbb{Q} = A_x \sqcup B_x$ comes from the following set, and its complement:

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

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

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

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

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

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

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

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

THEOREM 1.13. *A real number $r \in \mathbb{R}$ is rational precisely when*

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

*that is, when its decimal writing is periodic.*

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

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

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

At a more advanced level, passed the rationals, our problem remains the same, namely how to recognize the arithmetic properties of the real numbers $r\in\mathbb{R}$, as for instance being square roots of rationals, and so on, when written in decimal form. Many things can be said here, and we will be back to this on several occasions, in this book.

Getting back now to Theorem 1.12, that was definitely something quite difficult. Alternatively, we have the following definition for the real numbers:

THEOREM 1.14. *The field of real numbers $\mathbb{R}$ can be defined as well as the completion of $\mathbb{Q}$ with respect to the usual distance on the rationals, namely*

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

*and with the operations on $\mathbb{R}$ coming from those on $\mathbb{Q}$, via Cauchy sequences.*

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

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

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

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

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

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

(3) With this notion in hand, the idea will be to define the reals $x\in\mathbb{R}$ as being the limits of the Cauchy sequences $\{r_n\}\subset\mathbb{Q}$. But since these limits are not known yet to

exist to us, precisely because they are real, we must employ a trick. So, let us define instead the reals $x \in \mathbb{R}$ as being the Cauchy sequences $\{r_n\} \subset \mathbb{Q}$ themselves.

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

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

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

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

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

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

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

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

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

THEOREM 1.15. *The real numbers $\mathbb{R}$ can be defined as well via the decimal form*

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

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

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

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

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

*and then summing and multiplying, in the obvious way.*

PROOF. This is something which looks quite intuitive, but which in practice, and we insist here, is not exactly beginner level, the idea with this being as follows:

(1) Let us first forget about the precise decimal writing in the statement, and define the real numbers $x \in \mathbb{R}$ as being formal sums as follows, with the sum being over integers $k \in \mathbb{Z}$ assumed to be greater than a certain integer, $k \geq k_0$:

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

(2) Now by truncating, we can see that what we have here are certain Cauchy sequences of rationals, and with a bit more work, we conclude that the $\mathbb{R}$ that we constructed is precisely the $\mathbb{R}$ that we constructed in Theorem 1.14. Thus, we get the result.

(3) Alternatively, by getting back to Theorem 1.12 and its proof, we can argue, based on that, that the $\mathbb{R}$ that we constructed coincides with the old $\mathbb{R}$ from Definition 1.9, the one constructed via Dedekind cuts, and this gives again all the assertions. □

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 1.16. *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 following well-known formula, which comes by dividing the interval $[0,1]$ 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:

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

Thus, getting back now to the above, we have indeed $P\leq c$, and since the number $c>0$ was arbitrary, we conclude that we have $P=0$, as desired. □

As a comment here, all the above was 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.

## 1c. Sequences, limits

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

DEFINITION 1.17. *We say that a sequence $\{x_n\}_{n\in\mathbb{N}}\subset\mathbb{R}$ converges to $x\in\mathbb{R}$ when:*

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

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

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

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

(2) That is, this tells us that "when $n$ is big enough, $x_n$ gets arbitrarily close to $x$".

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

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

(5) Thus, we are naturally led to the above definition, and end of the story.

Time perhaps for some illustrations. As a basic example for all this, we have:

PROPOSITION 1.18. *We have* $1/n \to 0$.

PROOF. This is obvious, but let us prove it by using Definition 1.17. We have:

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

Thus we can take $N=[1/\varepsilon]+1$ in Definition 1.17, and we are done. □

There are countless other examples of limits, and more on this in a moment. Going ahead with more theory, let us complement Definition 1.17 with:

DEFINITION 1.19. *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. We will leave some thinking here, along the lines of the discussion following Definition 1.17, as an instructive exercise.

As a basic illustration for Definition 1.19, we have:

PROPOSITION 1.20. *We have* $n^2\to\infty$.

PROOF. As before, this is obvious, but let us prove it using Definition 1.19. We have:

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

Thus we can take $N=[\sqrt{K}]+1$ in Definition 1.19, and we are done. □

We can unify and generalize Proposition 1.18 and Proposition 1.20, as follows:

PROPOSITION 1.21. *We have the following convergence,*

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

*with* $n\to\infty$.

PROOF. This follows indeed by using the same method as in the proof of Proposition 1.18 and Proposition 1.20, first for $a$ rational, and then for $a$ real as well. We will leave working out the details here as an instructive exercise. □

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

THEOREM 1.22. *The following happen:*

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

PROOF. All this is elementary, coming from definitions:

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

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

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

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

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

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

(4) This is clear from definitions. □

Here are as well some general rules for computing limits:

THEOREM 1.23. *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 1.24. *The $n\to\infty$ limits of quotients of polynomials are given by*

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

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

PROOF. The first assertion comes from the following computation:

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

As for the second assertion, this comes from Proposition 1.21. □

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 1.25. *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 $\infty$ and $\infty$. 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 1.26. *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 1.27. *$\mathbb{R}$ 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:*

$$x_n\sim y_n\iff|x_n-y_n|\to0$$

*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 1.26 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 1.26. □

## 1d. Sums and series

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

DEFINITION 1.28. *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 1.29. *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 1.30. *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 1.31. *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 1.30, 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 1.32. *The following hold, with the converses of* (1) *and* (2) *being wrong, and with* (3) *not holding when the assumption $x_n\geq 0$ is removed:*

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

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

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

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

As for the simplest counterexample for the converse, this comes from the following tricky formula of Riemann, that we know well from Theorem 1.30:

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

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

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

As for the simplest counterexample for the converse, this comes from Theorem 1.31. Indeed, let us have a look at the formula established there, namely:

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

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

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

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

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

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

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

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

THEOREM 1.33. *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 1.31, we have:

THEOREM 1.34. *If $x_n \searrow 0$ then $\sum_n (-1)^n x_n$ converges.*

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

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

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

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

## 1e. Exercises

Exercises:

Exercise 1.35.

Exercise 1.36.

Exercise 1.37.

Exercise 1.38.

Exercise 1.39.

Exercise 1.40.

Exercise 1.41.

Exercise 1.42.

Bonus exercise.

CHAPTER 2

# Functions

## 2a. Functions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$f : X \to Y$$

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

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

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

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

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

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

You certainly know how to do this, draw such graphs, so let us record here:

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

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

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

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

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

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

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

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

We will see in what follows many other applications of the graphs of functions, for countless other questions that we can have, about them. However, as a word of warning, the graph of a function is not everything. For instance the very basic function $f(x) = 2x$ remains best thought of as it comes, in 1D, as being the function which elongates all the distances by 2, and with this property being harder to see on its graph.

Let us record this latter finding, which is something important, as follows:

WARNING 2.7. *The graph is not everything, with for instance $f(x) = 2x$ being best thought of as it comes, as being the function which elongates all the distances by 2.*

Which brings us to the question, after all, what is the best way to represent a function. Not clear, so time to ask the cat. And cat declares:

CAT 2.8. *The best is to think as $f : \mathbb{R} \to \mathbb{R}$ as being yourself. In this way, you will have no troubles with the math, just look at yourself, and record your findings.*

Which sounds quite interesting and conceptual, and reminds a bit about what artists say about their music, you have to be that music, for creating it, and then later, as a listener, for understanding it too. I actually had some experience with this, during my postdoc years at UC Berkeley, a very long time ago. What a crazy place California is, and the Bay Area in particular, and nothing more pleasant, after coming back home from one of these crazy nights out, than relaxing a bit before going to sleep, and listening to John Coltrane. You have this feeling of being just there, as being yourself the saxophone of Coltrane, and producing, and even being, yourself the music, and good musical understanding that is. Good times back then, and please of course do not do like me, unless perhaps for serious things, like your study of functions, along the lines of Cat 2.8.

Back to more concrete mathematics now, so to say, let us have a closer look at the simplest functions that we know, namely the degree 2 ones. One interesting question regards solving the equation $f(x) = 0$, and here we know from chapter 1 that we have:

THEOREM 2.9. *For $f(x) = ax^2 + bx + c$, the solutions of $f(x) = 0$ 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. This is something that we know well from chapter 1, coming as follows:

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

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

Very nice all this, and you would probably say that the story is over here with degree 2, matters to be relegated to the elementary school. However, not really. Have you noticed that at the university, professors are usually faster than students, in dealing with degree 2? I am probably not supposed to talk about our secrets, but here are our tricks:

TRICKS 2.10. *The following happen:*

(1) *The roots of $x^2-ax+b$ can be computed by using $r+s=a$, $rs=b$.*
(2) *The eigenvalues of $A\in M_2(\mathbb{C})$ are given by $r+s=Tr(A)$, $rs=\det A$.*

To be more precise, (1) is clear, and the equations there are usually the fastest way for computing, via instant thinking, the roots $r,s$, provided of course that these roots are simple numbers, say integers. As for (2), consider indeed a $2\times 2$ matrix:

$$A=\begin{pmatrix}m&n\\p&q\end{pmatrix}$$

In order to find the eigenvalues $r,s$, you are certainly very used to compute the characteristic polynomial, then apply Theorem 2.9. But my point is that this characteristic polynomial is of the form $x^2-ax+b$, with $a=Tr(a)$ and $b=\det A$, so we can normally apply the trick in (1), provided of course that $r,s$ are simple numbers, say integers.

Finally, for this discussion to be complete, let us mention too:

WARNING 2.11. *The above tricks work in pure mathematics, where the numbers $r,s$ that we can meet are usually integers, or rationals. In applied mathematics, however, the numbers that we meet are integers or rationals with probability $P=0$, so no tricks.*

I am saying this of course in view of the fact that in applied mathematics the numbers that can appear, say via reading certain scientific instruments, are quite "random", and to be more precise, oscillating in a random way around an average value. Thus, we are dealing here with the continuum, and the probability of being rational is $P=0$.

## 2b. Polynomials, roots

Still in relation with degree 2 equations and functions, what to do when the discriminant is negative? In order to solve $x^2 = -1$, we must trick, in the following way:

DEFINITION 2.12. *The complex numbers are variables of the form*

$$x = a + ib$$

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

$$i^2 = -1$$

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

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

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

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

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

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

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

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

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

$$x = \pm i$$

This is of course very good news. More generally, we have the following result, regarding the arbitrary degree 2 equations, with real coefficients:

THEOREM 2.13. *The complex solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{R}$ are*

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

*with the square root of negative real numbers being defined as*

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

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

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

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

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

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

PROPOSITION 2.14. *The complex numbers, written as usual*

$$x=a+ib$$

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

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

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

PROOF. Consider indeed two arbitrary complex numbers:

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

Their sum is then by definition the following complex number:

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

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

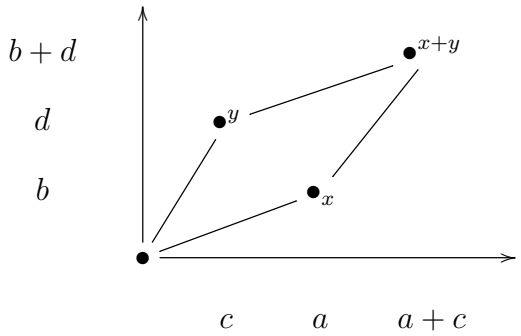
$$x=\begin{pmatrix}a\\b\end{pmatrix}\quad,\quad y=\begin{pmatrix}c\\d\end{pmatrix}$$

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

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

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

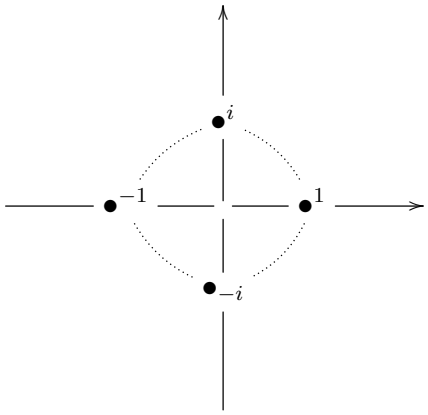
Here we have assumed that you are a bit familiar with vector calculus. If not, no problem, the idea is simply that vectors add by forming a parallelogram, as follows:




Observe that in our geometric picture from Proposition 2.14, the real numbers correspond to the numbers on the $Ox$ axis. As for the purely imaginary numbers, these lie on the $Oy$ axis, with the number $i$ itself being given by the following formula:

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

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




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

ADVICE 2.15. *When drawing complex numbers, always begin with the coordinate axes $Ox, Oy$, and with a copy of the unit circle.*

And more on this later, in chapter 3 below, when talking trigonometry.

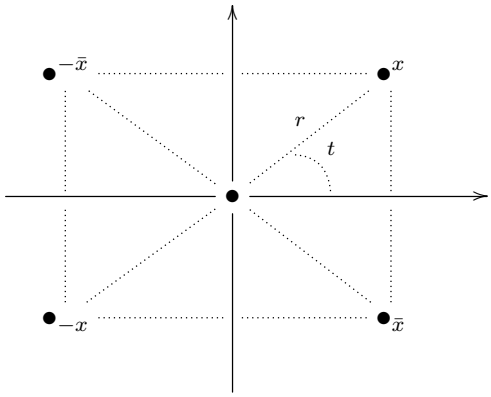
As a last general topic regarding the complex numbers, let us discuss conjugation. This is something quite tricky, complex number specific, as follows:

DEFINITION 2.16. *The complex conjugate of $x = a + ib$ is the following number,*

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

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

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




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

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

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

THEOREM 2.17. *The conjugation operation $x \to \bar{x}$ has the following properties:*

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

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

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

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

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

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

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

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

(5) This comes from the formula of the solutions, that we know from Theorem 2.13, but we can deduce this as well directly, without computations. Indeed, by using our assumption that the coefficients are real, $a, b, c \in \mathbb{R}$, we have:

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

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

Now back to polynomials, as already mentioned, we will see later that any degree 2 complex equation has solutions over the complex numbers, and that more generally, any polynomial equation, real or complex, has solutions over the complex numbers.

## 2c. Degree 3 equations

Moving now to degree 3 and higher, things here are far more complicated, and as a first objective, we would like to understand what the analogue of the discriminant $\Delta = b^2 - 4ac$ is. But even this is something quite tricky, because we would like to have $\Delta = 0$ precisely when $(P, P') \neq 1$, which leads us into the question of deciding, given two polynomials $P, Q \in \mathbb{C}[X]$, if these polynomials have a common root, $(P, Q) \neq 1$, or not.

Fortunately this latter question has a nice answer. We will need:

THEOREM 2.18. *Given a monic polynomial $P \in \mathbb{C}[X]$, factorized as*

$$P = (X - a_1) \ldots (X - a_k)$$

*the following happen:*

(1) *The coefficients of $P$ are symmetric functions in $a_1, \ldots, a_k$.*
(2) *The symmetric functions in $a_1, \ldots, a_k$ are polynomials in the coefficients of $P$.*

PROOF. This is something quite standard, requiring a bit of abstract mathematical thinking, and a few abstract verifications too, the idea being as follows:

(1) By expanding our polynomial, we have the following formula:

$$P=\sum_{r=0}^{k}(-1)^r\sum_{i_1<\ldots<i_r}a_{i_1}\ldots a_{i_r}\cdot X^{k-r}$$

Thus the coefficients of $P$ are, up to some signs, the following functions:

$$f_r=\sum_{i_1<\ldots<i_r}a_{i_1}\ldots a_{i_r}$$

But these are indeed symmetric functions in $a_1,\ldots,a_k$, as claimed.

(2) Conversely now, let us look at the symmetric functions in the roots $a_1,\ldots,a_k$. These appear as linear combinations of the basic symmetric functions, given by:

$$S_r=\sum_i a_i^r$$

Moreover, when allowing polynomials instead of linear combinations, we need in fact only the first $k$ such sums, namely $S_1,\ldots,S_k$. That is, the symmetric functions $\mathcal{F}$ in our variables $a_1,\ldots,a_k$, with integer coefficients, appear as follows:

$$\mathcal{F}=\mathbb{Z}[S_1,\ldots,S_k]$$

(3) The point now is that, alternatively, the symmetric functions in our variables $a_1,\ldots,a_k$ appear as well as linear combinations of the functions $f_r$ that we found in (1), and that when allowing polynomials instead of linear combinations, we need in fact only the first $k$ functions, namely $f_1,\ldots,f_k$. That is, we have as well:

$$\mathcal{F}=\mathbb{Z}[f_1,\ldots,f_k]$$

But this gives the result, because we can pass from $\{S_r\}$ to $\{f_r\}$, and vice versa.

(4) This was for the idea, and in practice now up to you to clarify all the details. In fact, we will also need in what follows the extension of all this to the case where $P$ is no longer assumed to be monic, and with this being, again, exercise for you. □

So, symmetric functions understood, and do not hesitate of course, besides working out the details of the above proof, to work out some examples too. There is a lot of useful algebra know-how to be learned here, and the more exercises you solve, the better.

Getting back now to our original question, namely that of deciding whether two polynomials $P,Q\in\mathbb{C}[X]$ have a common root or not, this has the following nice answer:

THEOREM 2.19. *Given two polynomials $P, Q \in \mathbb{C}[X]$, written as*

$$P = c(X - a_1) \ldots (X - a_k) \quad , \quad Q = d(X - b_1) \ldots (X - b_l)$$

*the following quantity, which is called resultant of $P, Q$,*

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

*is a certain polynomial in the coefficients of $P, Q$, with integer coefficients, and we have $R(P, Q) = 0$ precisely when $P, Q$ have a common root.*

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

(1) Given two polynomials $P, Q \in \mathbb{C}[X]$, we can certainly construct the quantity $R(P, Q)$ in the statement, with the role of the normalization factor $c^l d^k$ to become clear later on, and then we have $R(P, Q) = 0$ precisely when $P, Q$ have a common root:

$$R(P, Q) = 0 \iff \exists i, j, a_i = b_j$$

(2) As bad news, however, this quantity $R(P, Q)$, defined in this way, is a priori not very useful in practice, because it depends on the roots $a_i, b_j$ of our polynomials $P, Q$, that we cannot compute in general. However, and here comes our point, as we will prove below, it turns out that $R(P, Q)$ is in fact a polynomial in the coefficients of $P, Q$, with integer coefficients, and this is where the power of $R(P, Q)$ comes from.

(3) You might perhaps say, nice, but why not doing things the other way around, that is, formulating our theorem with the explicit formula of $R(P, Q)$, in terms of the coefficients of $P, Q$, and then proving that we have $R(P, Q) = 0$, via roots and everything. Good point, but this is not exactly obvious, the formula of $R(P, Q)$ in terms of the coefficients of $P, Q$ being something terribly complicated. In short, trust me, let us prove our theorem as stated, and for alternative formulae of $R(P, Q)$, we will see later.

(4) Getting started now, let us expand the formula of $R(P, Q)$, by making all the multiplications there, abstractly, in our head. Everything being symmetric in $a_1, \ldots, a_k$, we obtain in this way certain symmetric functions in these variables, which will be therefore certain polynomials in the coefficients of $P$. Moreover, due to our normalization factor $c^l$, these polynomials in the coefficients of $P$ will have integer coefficients.

(5) With this done, let us look now what happens with respect to the remaining variables $b_1, \ldots, b_l$, which are the roots of $Q$. Once again what we have here are certain symmetric functions in these variables $b_1, \ldots, b_l$, and these symmetric functions must be certain polynomials in the coefficients of $Q$. Moreover, due to our normalization factor $d^k$, these polynomials in the coefficients of $Q$ will have integer coefficients.

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

All the above might seem a bit complicated, so as an illustration, let us work out an example. Consider the case of a polynomial of degree 2, and a polynomial of degree 1:

$$P = ax^2 + bx + c \quad , \quad Q = dx + e$$

In order to compute the resultant, let us factorize our polynomials:

$$P = a(x - p)(x - q) \quad , \quad Q = d(x - r)$$

The resultant can be then computed as follows, by using the method above:

$$\begin{aligned} R(P, Q) &= ad^2(p - r)(q - r) \\ &= ad^2(pq - (p + q)r + r^2) \\ &= cd^2 + bd^2r + ad^2r^2 \\ &= cd^2 - bde + ae^2 \end{aligned}$$

Finally, observe that $R(P, Q) = 0$ corresponds indeed to the fact that $P, Q$ have a common root. Indeed, the root of $Q$ is $r = -e/d$, and we have:

$$\begin{aligned} P(r) &= \frac{ae^2}{d^2} - \frac{be}{d} + c \\ &= \frac{R(P, Q)}{d^2} \end{aligned}$$

We will be back to more such examples and computations later.

Regarding now the explicit formula of the resultant $R(P, Q)$, this is something quite complicated, and there are several methods for dealing with this problem. We have:

THEOREM 2.20. *The resultant of two polynomials, written as*

$$P = p_kX^k + \ldots + p_1X + p_0 \quad , \quad Q = q_lX^l + \ldots + q_1X + q_0$$

*appears as the determinant of an associated matrix, as follows,*

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

*with the matrix having size $k + l$, and having $0$ coefficients at the blank spaces.*

PROOF. This is something clever, due to Sylvester, as follows:

(1) Consider the vector space $\mathbb{C}_k[X]$ formed by the polynomials of degree $< k$:

$$\mathbb{C}_k[X] = \left\{ P \in \mathbb{C}[X] \middle| \deg P < k \right\}$$

This is a vector space of dimension $k$, having as basis the monomials $1, X, \ldots, X^{k-1}$. Now given polynomials $P, Q$ as in the statement, consider the following linear map:

$$\Phi : \mathbb{C}_l[X] \times \mathbb{C}_k[X] \to \mathbb{C}_{k+l}[X] \quad , \quad (A, B) \to AP + BQ$$

(2) Our first claim is that with respect to the standard bases for all the vector spaces involved, namely those consisting of the monomials $1, X, X^2, \ldots$, the matrix of $\Phi$ is the matrix in the statement. But this is something which is clear from definitions.

(3) Our second claim is that $\det \Phi = 0$ happens precisely when $P, Q$ have a common root. Indeed, our polynomials $P, Q$ having a common root means that we can find $A, B$ such that $AP + BQ = 0$, and so that $(A, B) \in \ker \Phi$, which reads $\det \Phi = 0$.

(4) Finally, our claim is that we have $\det \Phi = R(P, Q)$. But this follows from the uniqueness of the resultant, up to a scalar, and with this uniqueness property being elementary to establish, along the lines of the proofs of Theorems 2.18 and 2.19. □

In what follows we will not really need the above formula, so let us just check now that this formula works indeed. Consider our favorite polynomials, as before:

$$P = ax^2 + bx + c \quad , \quad Q = dx + e$$

According to the above result, the resultant should be then, as it should:

$$R(P, Q) = \begin{vmatrix} a & d & 0 \\ b & e & d \\ c & 0 & e \end{vmatrix} = ae^2 - bde + cd^2$$

We can go back now to our original question, and we have:

THEOREM 2.21. *Given a polynomial $P \in \mathbb{C}[X]$, written as*

$$P(X) = aX^N + bX^{N-1} + cX^{N-2} + \ldots$$

*its discriminant, defined as being the following quantity,*

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

*is a polynomial in the coefficients of $P$, with integer coefficients, and $\Delta(P) = 0$ happens precisely when $P$ has a double root.*

PROOF. The fact that the discriminant $\Delta(P)$ is a polynomial in the coefficients of $P$, with integer coefficients, comes from Theorem 2.19, coupled with the fact that the division by the leading coefficient $a$ is indeed possible, under $\mathbb{Z}$, as being shown by the

following formula, which is of course a bit informal, coming from Theorem 2.20:

$$R(P,P')=\begin{vmatrix} a & & & Na & & \\ \vdots & \ddots & & \vdots & \ddots & \\ z & & a & y & & Na \\ & \ddots & \vdots & & \ddots & \vdots \\ & & z & & & y \end{vmatrix}$$

Also, the fact that we have $\Delta(P)=0$ precisely when $P$ has a double root is clear from Theorem 2.19. Finally, let us mention that the sign $(-1)^{\binom{N}{2}}$ is there for various reasons, including the compatibility with some well-known formulae, at small values of $N\in\mathbb{N}$, such as $\Delta(P)=b^2-4ac$ in degree 2, that we will discuss in a moment. □

As already mentioned, by using Theorem 2.20, we have an explicit formula for the discriminant, as the determinant of a certain matrix. There is a lot of theory here, and in order to get into this, let us first see what happens in degree 2. Here we have:

$$P=aX^2+bX+c\quad,\quad P'=2aX+b$$

Thus, the resultant is given by the following formula:

$$\begin{aligned} R(P,P') &= ab^2-b(2a)b+c(2a)^2\\ &= 4a^2c-ab^2\\ &= -a(b^2-4ac) \end{aligned}$$

It follows that the discriminant of our polynomial is, as it should:

$$\Delta(P)=b^2-4ac$$

Alternatively, we can use the formula in Theorem 2.20, and we obtain:

$$\begin{aligned} \Delta(P) &= -\frac{1}{a}\begin{vmatrix} a & 2a & \\ b & b & 2a \\ c & & b \end{vmatrix}\\ &= -\begin{vmatrix} 1 & 2 & \\ b & b & 2a \\ c & & b \end{vmatrix}\\ &= -b^2+2(b^2-2ac)\\ &= b^2-4ac \end{aligned}$$

Let us discuss now what happens in degree 3. Here the result is as follows:

THEOREM 2.22. *The discriminant of a degree* 3 *polynomial,*

$$P=aX^3+bX^2+cX+d$$

*is the number* $\Delta(P)=b^2c^2-4ac^3-4b^3d-27a^2d^2+18abcd$.

PROOF. We have two methods available, based on Theorem 2.19 and Theorem 2.20, and both being instructive, we will try them both. The computations are as follows:

(1) Let us first go the pedestrian way, based on the definition of the resultant, from Theorem 2.19. Consider two polynomials, of degree 3 and degree 2, written as follows:

$$P = aX^3 + bX^2 + cX + d$$

$$Q = eX^2 + fX + g = e(X - s)(X - t)$$

The resultant of these two polynomials is then given by:

$$\begin{aligned} R(P,Q) &= a^2e^3(p-s)(p-t)(q-s)(q-t)(r-s)(r-t) \\ &= a^2 \cdot e(p-s)(p-t) \cdot e(q-s)(q-t) \cdot e(r-s)(r-t) \\ &= a^2Q(p)Q(q)Q(r) \\ &= a^2(ep^2+fp+g)(eq^2+fq+g)(er^2+fr+g) \end{aligned}$$

By expanding, we obtain the following formula for this resultant:

$$\begin{aligned} \frac{R(P,Q)}{a^2} &= e^3p^2q^2r^2 + e^2f(p^2q^2r + p^2qr^2 + pq^2r^2) \\ &+ e^2g(p^2q^2 + p^2r^2 + q^2r^2) + ef^2(p^2qr + pq^2r + pqr^2) \\ &+ efg(p^2q + pq^2 + p^2r + pr^2 + q^2r + qr^2) + f^3pqr \\ &+ eg^2(p^2 + q^2 + r^2) + f^2g(pq + pr + qr) \\ &+ fg^2(p + q + r) + g^3 \end{aligned}$$

Note in passing that we have 27 terms on the right, as we should, and with this kind of check being mandatory, when doing such computations. Next, we have:

$$p + q + r = -\frac{b}{a} \quad , \quad pq + pr + qr = \frac{c}{a} \quad , \quad pqr = -\frac{d}{a}$$

By using these formulae, we can produce some more, as follows:

$$p^2 + q^2 + r^2 = (p + q + r)^2 - 2(pq + pr + qr) = \frac{b^2}{a^2} - \frac{2c}{a}$$

$$p^2q + pq^2 + p^2r + pr^2 + q^2r + qr^2 = (p + q + r)(pq + pr + qr) - 3pqr = -\frac{bc}{a^2} + \frac{3d}{a}$$

$$p^2q^2 + p^2r^2 + q^2r^2 = (pq + pr + qr)^2 - 2pqr(p + q + r) = \frac{c^2}{a^2} - \frac{2bd}{a^2}$$

By plugging now this data into the formula of $R(P,Q)$, we obtain:

$$\begin{aligned}R(P,Q) &= a^2e^3\cdot\frac{d^2}{a^2}-a^2e^2f\cdot\frac{cd}{a^2}+a^2e^2g\left(\frac{c^2}{a^2}-\frac{2bd}{a^2}\right)+a^2ef^2\cdot\frac{bd}{a^2}\\ &+ a^2efg\left(-\frac{bc}{a^2}+\frac{3d}{a}\right)-a^2f^3\cdot\frac{d}{a}\\ &+ a^2eg^2\left(\frac{b^2}{a^2}-\frac{2c}{a}\right)+a^2f^2g\cdot\frac{c}{a}-a^2fg^2\cdot\frac{b}{a}+a^2g^3\end{aligned}$$

Thus, we have the following formula for the resultant:

$$\begin{aligned}R(P,Q) &= d^2e^3-cde^2f+c^2e^2g-2bde^2g+bdef^2-bcefg+3adefg\\ &- adf^3+b^2eg^2-2aceg^2+acf^2g-abfg^2+a^2g^3\end{aligned}$$

Getting back now to our discriminant problem, with $Q=P'$, which corresponds to $e=3a$, $f=2b$, $g=c$, we obtain the following formula:

$$\begin{aligned}R(P,P') &= 27a^3d^2-18a^2bcd+9a^2c^3-18a^2bcd+12ab^3d-6ab^2c^2+18a^2bcd\\ &- 8ab^3d+3ab^2c^2-6a^2c^3+4ab^2c^2-2ab^2c^2+a^2c^3\end{aligned}$$

By simplifying terms, and dividing by $a$, we obtain the following formula:

$$-\Delta(P)=27a^2d^2-18abcd+4ac^3+4b^3d-b^2c^2$$

But this gives the formula in the statement, namely:

$$\Delta(P)=b^2c^2-4ac^3-4b^3d-27a^2d^2+18abcd$$

(2) Let us see as well how the computation does, by using Theorem 2.20, which is our most advanced tool, so far. Consider a polynomial of degree 3, and its derivative:

$$P=aX^3+bX^2+cX+d$$

$$P'=3aX^2+2bX+c$$

By using now Theorem 2.20 and computing the determinant, we obtain:

$$\begin{aligned}
R(P,P') &= \begin{vmatrix} a & & 3a & & \\ b & a & 2b & 3a & \\ c & b & c & 2b & 3a \\ d & c & & c & 2b \\ & d & & & c \end{vmatrix} \\
&= \begin{vmatrix} a & & & & \\ b & a & -b & 3a & \\ c & b & -2c & 2b & 3a \\ d & c & -3d & c & 2b \\ & d & & & c \end{vmatrix} \\
&= a\begin{vmatrix} a & -b & 3a & \\ b & -2c & 2b & 3a \\ c & -3d & c & 2b \\ d & & & c \end{vmatrix} \\
&= -ad\begin{vmatrix} -b & 3a & \\ -2c & 2b & 3a \\ -3d & c & 2b \end{vmatrix} + ac\begin{vmatrix} a & -b & 3a \\ b & -2c & 2b \\ c & -3d & c \end{vmatrix} \\
&= -ad(-4b^3 - 27a^2d + 12abc + 3abc) \\
&\quad +ac(-2ac^2 - 2b^2c - 9abd + 6ac^2 + b^2c + 6abd) \\
&= a(4b^3d + 27a^2d^2 - 15abcd + 4ac^3 - b^2c^2 - 3abcd) \\
&= a(4b^3d + 27a^2d^2 - 18abcd + 4ac^3 - b^2c^2)
\end{aligned}$$

Now according to Theorem 2.21, the discriminant of our polynomial is given by:

$$\begin{aligned}
\Delta(P) &= -\frac{R(P,P')}{a} \\
&= -4b^3d - 27a^2d^2 + 18abcd - 4ac^3 + b^2c^2 \\
&= b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd
\end{aligned}$$

Thus, we have again obtained the formula in the statement. □

Still talking degree 3 equations, let us try now to solve such an equation $P = 0$, with $P = aX^3 + bX^2 + cX + d$ as above. By linear transformations we can assume $a = 1, b = 0$, and then it is convenient to write $c = 3p, d = 2q$. Thus, our equation becomes:

$$x^3 + 3px + 2q = 0$$

Regarding such equations, many things can be said, and to start with, we have the following famous result, dealing with real roots, due to Cardano:

THEOREM 2.23. *For a normalized degree 3 equation, namely*

$$x^3+3px+2q=0$$

*the discriminant is* $\Delta=-108(p^3+q^2)$*. Assuming* $p,q\in\mathbb{R}$ *and* $\Delta<0$*, the number*

$$x=\sqrt[3]{-q+\sqrt{p^3+q^2}}+\sqrt[3]{-q-\sqrt{p^3+q^2}}$$

*is a real solution of our equation.*

PROOF. The formula of $\Delta$ is clear from definitions, and with $108=4\times 27$. Now with $x$ as in the statement, by using $(a+b)^3=a^3+b^3+3ab(a+b)$, we have:

$$\begin{aligned}x^3&=\left(\sqrt[3]{-q+\sqrt{p^3+q^2}}+\sqrt[3]{-q-\sqrt{p^3+q^2}}\right)^3\\&=-2q+3\sqrt[3]{-q+\sqrt{p^3+q^2}}\cdot\sqrt[3]{-q-\sqrt{p^3+q^2}}\cdot x\\&=-2q+3\sqrt[3]{q^2-p^3-q^2}\cdot x\\&=-2q-3px\end{aligned}$$

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

Regarding the other roots, it is easy to see that these are both real when $\Delta<0$, and complex conjugate when $\Delta<0$. Thus, in the context of Theorem 2.23, the other two roots are complex conjugate, the formula for them being as follows:

PROPOSITION 2.24. *For a normalized degree 3 equation, namely*

$$x^3+3px+2q=0$$

*with* $p,q\in\mathbb{R}$ *and discriminant* $\Delta=-108(p^3+q^2)$ *negative,* $\Delta<0$*, the numbers*

$$z=w\sqrt[3]{-q+\sqrt{p^3+q^2}}+w^2\sqrt[3]{-q-\sqrt{p^3+q^2}}$$

$$\bar{z}=w^2\sqrt[3]{-q+\sqrt{p^3+q^2}}+w\sqrt[3]{-q-\sqrt{p^3+q^2}}$$

*with* $w=e^{2\pi i/3}$ *are the complex conjugate solutions of our equation.*

PROOF. As before, by using $(a+b)^3=a^3+b^3+3ab(a+b)$, we have:

$$\begin{aligned}z^3&=\left(w\sqrt[3]{-q+\sqrt{p^3+q^2}}+w^2\sqrt[3]{-q-\sqrt{p^3+q^2}}\right)^3\\&=-2q+3\sqrt[3]{-q+\sqrt{p^3+q^2}}\cdot\sqrt[3]{-q-\sqrt{p^3+q^2}}\cdot z\\&=-2q+3\sqrt[3]{q^2-p^3-q^2}\cdot z\\&=-2q-3pz\end{aligned}$$

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

As a conclusion, we have the following statement, unifying the above:

THEOREM 2.25. *For a normalized degree* 3 *equation, namely*

$$x^3+3px+2q=0$$

*the discriminant is* $\Delta=-108(p^3+q^2)$. *Assuming* $p,q\in\mathbb{R}$ *and* $\Delta<0$, *the numbers*

$$x=w\sqrt[3]{-q+\sqrt{p^3+q^2}}+w^2\sqrt[3]{-q-\sqrt{p^3+q^2}}$$

*with* $w=1,e^{2\pi i/3},e^{4\pi i/3}$ *are the solutions of our equation.*

PROOF. This follows indeed from Theorem 2.23 and Proposition 2.24. Alternatively, we can redo the computation in their proof, which was nearly identical anyway, in the present setting, with $x$ being given by the above formula, by using $w^3=1$. □

## 2d. Degree 4 equations

In higher degree things become quite complicated. In degree 4, to start with, we first have the following result, dealing with the discriminant and its applications:

THEOREM 2.26. *The discriminant of* $P=ax^4+bx^3+cx^2+dx+e$ *is given by the following formula:*

$$\begin{aligned}\Delta &= 256a^3e^3-192a^2bde^2-128a^2c^2e^2+144a^2cd^2e-27a^2d^4\\ &\quad+144ab^2ce^2-6ab^2d^2e-80abc^2de+18abcd^3+16ac^4e\\ &\quad-4ac^3d^2-27b^4e^2+18b^3cde-4b^3d^3-4b^2c^3e+b^2c^2d^2\end{aligned}$$

*In the case* $\Delta<0$ *we have* 2 *real roots and* 2 *complex conjugate roots, and in the case* $\Delta>0$ *the roots are either all real or all complex.*

PROOF. The formula of $\Delta$ follows from the definition of the discriminant, from Theorem 2.21, with the resultant computed via Theorem 2.20, as follows:

$$\Delta=\frac{1}{a}\begin{vmatrix}a&&&4a&&&\\b&a&&3b&4a&&\\c&b&a&2c&3b&4a&\\d&c&b&d&2c&3b&4a\\e&d&c&&d&2c&3b\\&e&d&&&d&2c\\&&e&&&&d\end{vmatrix}$$

As for the last assertion, the study here is routine, a bit as in degree 3. □

In practice, as in degree 3, we can do first some manipulations on our polynomials, as to have them in simpler form, and we have the following version of Theorem 2.26:

PROPOSITION 2.27. *The discriminant of $P=x^4+cx^2+dx+e$, normalized degree 4 polynomial, is given by the following formula:*

$$\Delta=16c^4e-4c^3d^2-128c^2e^2+144cd^2e-27d^4+256e^3$$

*As before, if $\Delta<0$ we have 2 real roots and 2 complex conjugate roots, and if $\Delta>0$ the roots are either all real or all complex.*

PROOF. This is a consequence of Theorem 2.26, with $a=1,b=0$, but we can deduce this as well directly. Indeed, the formula of $\Delta$ follows, quite easily, from:

$$\Delta=\begin{vmatrix}1&&&4&&&\\&1&&&4&&\\c&&1&2c&&4&\\d&c&&d&2c&&4\\e&d&c&&d&2c&\\&e&d&&&d&2c\\&&e&&&&d\end{vmatrix}$$

As for the last assertion, this is something that we know, from Theorem 2.26. □

We still have some work to do. Indeed, looking back at what we did in degree 3, the passage there from Theorem 2.22 to Theorem 2.23 was made of two operations, namely "depressing" the equation, that is, getting rid of the next-to-highest term, and then rescaling the coefficients, as for the formula of $\Delta$ to become as simple as possible.

In our present setting now, degree 4, with the depressing done as above, in Proposition 2.27, it remains to rescale the coefficients, as for the formula of $\Delta$ to become as simple as possible. And here, a bit of formula hunting, in relation with $2,3$ powers, leads to:

THEOREM 2.28. *The discriminant of a normalized degree 4 polynomial, written as*

$$P=x^4+6px^2+4qx+3r$$

*is given by the following formula:*

$$\Delta=256\times27\times\left(9p^4r-2p^3q^2-6p^2r^2+6pq^2r-q^4+r^3\right)$$

*In the case $\Delta<0$ we have 2 real roots and 2 complex conjugate roots, and in the case $\Delta>0$ the roots are either all real or all complex.*

PROOF. This follows from Proposition 2.27, with $c = 6p, d = 4q, e = 3r$, but we can deduce this as well directly. Indeed, the formula of $\Delta$ follows, quite easily, from:

$$\Delta = \begin{vmatrix} 1 & & & 4 & & & \\ & 1 & & & 4 & & \\ 6p & & 1 & 12p & & 4 & \\ 4q & 6p & & 4q & 12p & & 4 \\ 3r & 4q & 6p & & 4q & 12p & \\ & 3r & 4q & & & 4q & 12p \\ & & 3r & & & & 4q \end{vmatrix}$$

As for the last assertion, this is something that we know from Theorem 2.26. □

Time now to get to the real thing, solving the equation. We have here:

THEOREM 2.29. *The roots of a normalized degree 4 equation, written as*

$$x^4 + 6px^2 + 4qx + 3r = 0$$

*are as follows, with $y$ satisfying the equation $(y^2 - 3r)(y - 3p) = 2q^2$,*

$$x_1 = \frac{1}{\sqrt{2}}\left(-\sqrt{y - 3p} + \sqrt{-y - 3p + \frac{4q}{\sqrt{2y - 6p}}}\right)$$

$$x_2 = \frac{1}{\sqrt{2}}\left(-\sqrt{y - 3p} - \sqrt{-y - 3p + \frac{4q}{\sqrt{2y - 6p}}}\right)$$

$$x_3 = \frac{1}{\sqrt{2}}\left(\sqrt{y - 3p} + \sqrt{-y - 3p - \frac{4q}{\sqrt{2y - 6p}}}\right)$$

$$x_4 = \frac{1}{\sqrt{2}}\left(\sqrt{y - 3p} - \sqrt{-y - 3p - \frac{4q}{\sqrt{2y - 6p}}}\right)$$

*and with $y$ being computable via the Cardano formula.*

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

(1) To start with, let us write our equation in the following form:

$$x^4 = -6px^2 - 4qx - 3r$$

Now assume that we have a number $y$ satisfying the following equation:

$$(y^2 - 3r)(y - 3p) = 2q^2$$

With this magic number $y$ in hand, our equation takes the following form:

$$\begin{aligned}(x^2+y)^2 &= x^4+2x^2y+y^2\\ &= -6px^2-4qx-3r+2x^2y+y^2\\ &= (2y-6p)x^2-4qx+y^2-3r\\ &= (2y-6p)x^2-4qx+\frac{2q^2}{y-3p}\\ &= \left(\sqrt{2y-6p}\cdot x-\frac{2q}{\sqrt{2y-6p}}\right)^2\end{aligned}$$

(2) Which looks very good, leading us to the following degree 2 equations:

$$x^2+y+\sqrt{2y-6p}\cdot x-\frac{2q}{\sqrt{2y-6p}}=0$$

$$x^2+y-\sqrt{2y-6p}\cdot x+\frac{2q}{\sqrt{2y-6p}}=0$$

Now let us write these two degree 2 equations in standard form, as follows:

$$x^2+\sqrt{2y-6p}\cdot x+\left(y-\frac{2q}{\sqrt{2y-6p}}\right)=0$$

$$x^2-\sqrt{2y-6p}\cdot x+\left(y+\frac{2q}{\sqrt{2y-6p}}\right)=0$$

(3) Regarding the first equation, the solutions there are as follows:

$$x_1=\frac{1}{2}\left(-\sqrt{2y-6p}+\sqrt{-2y-6p+\frac{8q}{\sqrt{2y-6p}}}\right)$$

$$x_2=\frac{1}{2}\left(-\sqrt{2y-6p}-\sqrt{-2y-6p+\frac{8q}{\sqrt{2y-6p}}}\right)$$

As for the second equation, the solutions there are as follows:

$$x_3=\frac{1}{2}\left(\sqrt{2y-6p}+\sqrt{-2y-6p-\frac{8q}{\sqrt{2y-6p}}}\right)$$

$$x_4=\frac{1}{2}\left(\sqrt{2y-6p}-\sqrt{-2y-6p-\frac{8q}{\sqrt{2y-6p}}}\right)$$

(4) Now by cutting a $\sqrt{2}$ factor from everything, this gives the formulae in the statement. As for the last claim, regarding the nature of $y$, this comes from Cardano. □

We still have to compute the number $y$ appearing in the above via Cardano, and the result here, adding to what we already have in Theorem 2.29, is as follows:

THEOREM 2.30 (continuation). *The value of $y$ in the previous theorem is*

$$y = t + p + \frac{a}{t}$$

*where the number $t$ is given by the formula*

$$t = \sqrt[3]{b + \sqrt{b^2 - a^3}}$$

*with $a = p^2 + r$ and $b = 2p^2 - 3pr + q^2$.*

PROOF. The legend has it that this is what comes from Cardano, but depressing and normalizing and solving $(y^2 - 3r)(y - 3p) = 2q^2$ makes it for too many operations, so the most pragmatic way is to simply check this equation. With $y$ as above, we have:

$$\begin{aligned} y^2 - 3r &= t^2 + 2pt + (p^2 + 2a) + \frac{2pa}{t} + \frac{a^2}{t^2} - 3r \\ &= t^2 + 2pt + (3p^2 - r) + \frac{2pa}{t} + \frac{a^2}{t^2} \end{aligned}$$

With this in hand, we have the following computation:

$$\begin{aligned} (y^2 - 3r)(y - 3p) &= \left(t^2 + 2pt + (3p^2 - r) + \frac{2pa}{t} + \frac{a^2}{t^2}\right)\left(t - 2p + \frac{a}{t}\right) \\ &= t^3 + (a - 4p^2 + 3p^2 - r)t + (2pa - 6p^3 + 2pr + 2pa) \\ &\quad + (3p^2a - ra - 4p^2a + a^2)\frac{1}{t} + \frac{a^3}{t^3} \\ &= t^3 + (a - p^2 - r)t + 2p(2a - 3p^2 + r) + a(a - p^2 - r)\frac{1}{t} + \frac{a^3}{t^3} \\ &= t^3 + 2p(-p^2 + 3r) + \frac{a^3}{t^3} \end{aligned}$$

Now by using the formula of $t$ in the statement, this gives:

$$\begin{aligned} (y^2 - 3r)(y - 3p) &= b + \sqrt{b^2 - a^3} - 4p^2 + 6pr + \frac{a^3}{b + \sqrt{b^2 - a^3}} \\ &= b + \sqrt{b^2 - a^3} - 4p^2 + 6pr + b - \sqrt{b^2 - a^3} \\ &= 2b - 4p^2 + 6pr \\ &= 2(2p^2 - 3pr + q^2) - 4p^2 + 6pr \\ &= 2q^2 \end{aligned}$$

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

## 2e. Exercises

Exercises:

Exercise 2.31.

Exercise 2.32.

Exercise 2.33.

Exercise 2.34.

Exercise 2.35.

Exercise 2.36.

Exercise 2.37.

Exercise 2.38.

Bonus exercise.

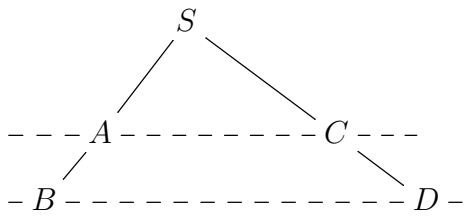
CHAPTER 3

# Sin and cos

## 3a. Angles, triangles

Welcome to geometry. It all started with triangles, drawn on sand. In order to get started, with some basic plane geometry, we first have the following key result:

THEOREM 3.1 (Thales). *Proportions are kept, along parallel lines. That is, given a configuration as follows, consisting of two parallel lines, and of two extra lines,*




*the following equality holds:*

$$\frac{SA}{SB} = \frac{SC}{SD}$$

*Moreover, the converse of this holds too, in the sense that, in the context of a picture as above, if this equality is satisfied, then the lines $AC$ and $BD$ must be parallel.*

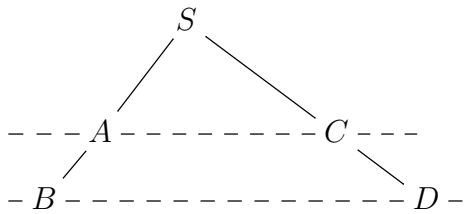
PROOF. We have indeed the following computation, based on the usual area formula for the triangles, that is, half of side times height, used multiple times:

$$\begin{aligned}
\frac{SA}{SB} &= \frac{area(CSA)}{area(CSB)}\\
&= \frac{area(CSA)}{area(CSA)+area(CAB)}\\
&= \frac{area(CSA)}{area(CSA)+area(CAD)}\\
&= \frac{area(ASC)}{area(ASD)}\\
&= \frac{SC}{SD}
\end{aligned}$$

As for the converse, we will leave the proof here as an instructive exercise. □

There are some other useful versions of the Thales theorem. First, we have:

THEOREM 3.2 (Thales 2). *In the context of the Thales theorem configuration,*

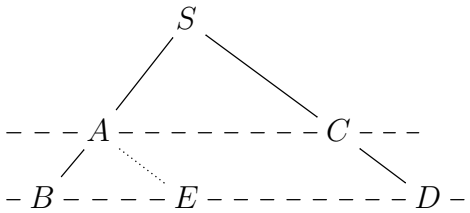



*the following equality, involving the same number, holds as well:*

$$\frac{SA}{SB} = \frac{AC}{BD}$$

*However, the converse of this does not necessarily hold.*

PROOF. In order to prove the formula in the statement, instead of getting lost into some new area computations, let us draw a tricky parallel, as follows:



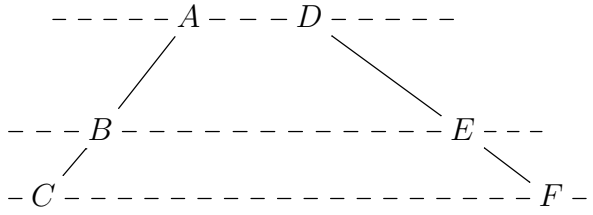

By using Theorem 3.1, we have then the following computation, as desired:

$$\frac{SA}{SB} = \frac{DE}{DB} = \frac{AC}{DB}$$

As for the converse, we will leave the proof here as an instructive exercise. □

As a third Thales theorem now, which is something beautiful too, we have:

THEOREM 3.3 (Thales 3). *Given a configuration as follows, consisting of three parallel lines, and of two extra lines, which can cross or not,*




*the following equality holds:*

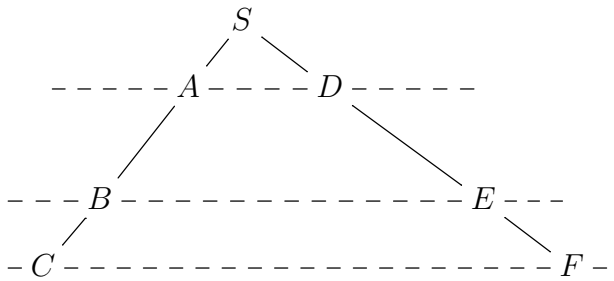
$$\frac{AB}{BC} = \frac{DE}{EF}$$

*That is, once again, the proportions are kept, along parallel lines.*

PROOF. We have two cases here, as follows:

(1) When the two extra lines are parallel, the result is clear, because we have plenty of parallelograms there, and the fractions in question are plainly equal.

(2) When the two lines cross, let us call $S$ their intersection:




Now by using Theorem 3.1 several times, we obtain:

$$\begin{aligned}\frac{AB}{BC} &= \frac{SB-SA}{SC-SB}\\ &= \frac{1-\frac{SA}{SB}}{\frac{SC}{SB}-1}\\ &= \frac{1-\frac{SD}{SE}}{\frac{SF}{SE}-1}\\ &= \frac{SE-SD}{SF-SE}\\ &= \frac{DE}{EF}\end{aligned}$$

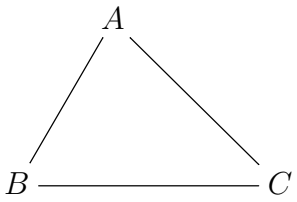
Thus, we are led to the formula in the statement. □

Before getting to angles, let us record as well the following key result:

THEOREM 3.4. *Given a triangle $ABC$, the following happen:*

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

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




Allowing us the freedom to play with some tricks, as advanced mathematicians, both students and professors, are allowed to, here is how the proof goes:

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

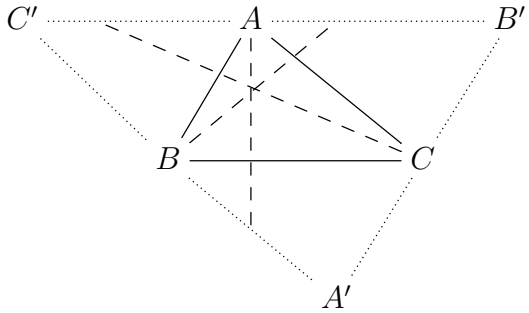
(2) This requires different techniques. Let us call $A, B, C \in \mathbb{C}$ the coordinates of $A, B, C$, and consider the average $P = (A + B + C)/3$. We have then:

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

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

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

(4) This is tougher, and I must admit that, when writing this book, I first struggled a bit with this, then ended looking it up on the internet. So, here is the trick. Draw a parallel to $BC$ at $A$, and similarly, parallels to $AB$ and $AC$ at $C$ and $B$. You will get in this way a bigger triangle, upside-down, $A'B'C'$. But then, the circumcenter of $A'B'C'$, that we know to exist from (3), will be the orthocenter of $ABC$:

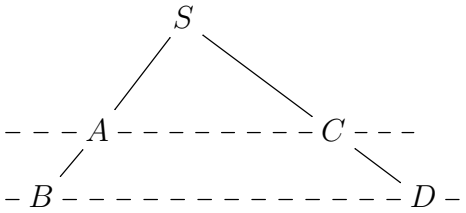



Thus, we are led to the conclusions in the statement. □

Many other things can be said about triangles, and we will be back to this. Importantly, we can now talk about angles, in the obvious way, by using triangles:

FACT 3.5. *We can talk about the angle between two crossing lines, and have some basic theory for the angles going, by using triangles, and Thales, in the obvious way.*

To be more precise here, let us go back to the configuration from the Thales theorem, which was as follows, with two parallel lines, and two other lines:




In this situation, we can say that the two triangles $SAC$ and $SBD$ are similar, and witn an equivalent formulation of similarity being the fact that the angles are equal:

DEFINITION 3.6. *We say that two triangles are similar, and we write*

$$SAC \sim SBD$$

*when their respective angles are equal.*

The point now is that, in this situation, we can have some mathematics going, for the lengths, coming from the following formula, which is the Thales theorem:

$$\frac{SA}{SB} = \frac{SC}{SD} = \frac{AC}{BD}$$

At the philosophical level now, you might wonder of course what the values of these angles, that we have been heavily using in the above, should be, say as real numbers. But this is something quite tricky, that will take us some time to understand. In the lack of something bright, for the moment, let us formulate the following definition:

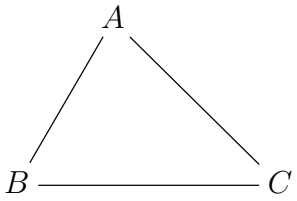
DEFINITION 3.7. *We can talk about the numeric value of angles, as follows:*

(1) *The right angle has value* $90°$.
(2) *We can double angles, in the obvious way.*
(3) *Thus, the half right angle has value* $45°$*, and the flat angle has value* $180°$.
(4) *We can also triple, quadruple and so on, again in the obvious way.*
(5) *Thus, we can talk about arbitrary rational multiples of* $90°$.
(6) *And, with a bit of analysis helping, we can in fact measure any angle.*

So, this will be our starting definition for the numeric values of the angles. Of course, all this might seem a bit improvized, but do not worry, we will come back later to this, with a better, more advanced definition for these numeric values of the angles.

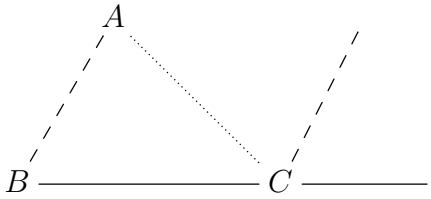
Getting back to work now, theorems and proofs, in relation with the above, here is a key result, which will be our main tool for the study of the angles:

THEOREM 3.8. *In an arbitrary triangle*




*the sum of all three angles is* 180°.

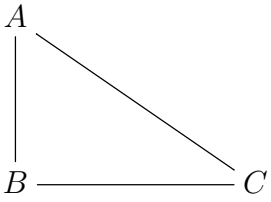
PROOF. This does not seem obvious to prove, with bare hands, but as usual, in such situations, some tricky parallels can come to the rescue. Let us prolong indeed the segment $BC$ a bit, on the $C$ side, and then draw a parallel at $C$, to the line $AB$, as follows:




But now, we can see that the three angles around $C$, summing up to the flat angle 180°, are in fact the 3 angles of our triangle. Thus, theorem proved, just like that. □

Going ahead now with our study of angles, as a continuation of the above, let us first talk about the simplest angle of them all, which is the right angle, denoted 90°. In relation with it, let us formulate the following definition, making the link with triangles:

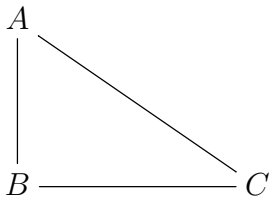
DEFINITION 3.9. *We call right triangle a triangle of type*




*having one of the angles equal to* 90°.

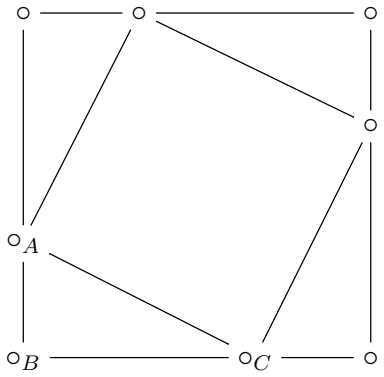
Many things can be said about right triangles, in particular with:

THEOREM 3.10 (Pythagoras). *In a right triangle $ABC$,*




*we have $AB^2 + BC^2 = AC^2$.*

PROOF. This comes from the following picture, consisting of two squares, and four triangles which are identical to $ABC$, as indicated:




Indeed, let us compute the area $S$ of the outer square. This can be done in two ways. First, since the side of this square is $AB + BC$, we obtain:

$$\begin{aligned} S &= (AB + BC)^2 \\ &= AB^2 + BC^2 + 2 \times AB \times BC \end{aligned}$$

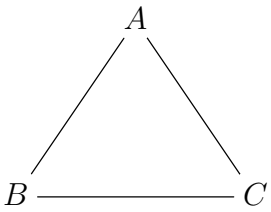
On the other hand, the outer square is made of the smaller square, having side $AC$, and of four identical right triangles, having sizes $AB, BC$. Thus:

$$\begin{aligned} S &= AC^2 + 4 \times \frac{AB \times BC}{2} \\ &= AC^2 + 2 \times AB \times BC \end{aligned}$$

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

As a second important angle, we have the $60^\circ$ angle, which usually appears via:

THEOREM 3.11. *In an equilateral triangle, having all sides equal,*

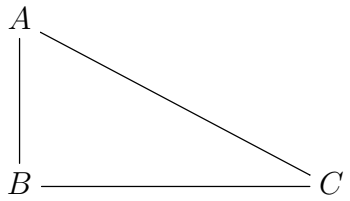



*all angles equal* $60°$*.*

PROOF. This is clear indeed from the fact that the sum is $180°$. □

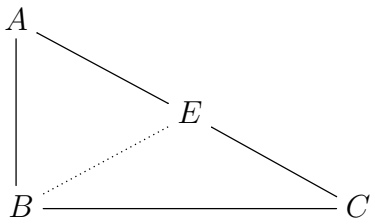
Another interesting angle is the $30°$ one. About it, we have:

THEOREM 3.12. *In a right triangle having small angles* $30°, 60°$*,*




*we have* $AB = AC/2$*.*

PROOF. This is clear by drawing an equilateral triangle, as follows:



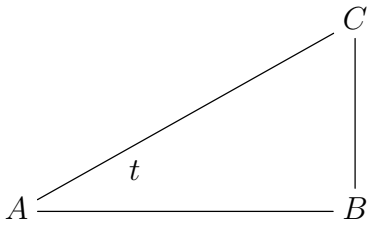

Thus, we are led to the conclusion in the statement. □

We will be back to such things in a moment, when doing trigonometry.

## 3b. Sine and cosine

Now that we know about angles, and about Pythagoras' theorem too, it is tempting at this point to start talking about trigonometry. Let us begin with:

DEFINITION 3.13. *Given a right triangle $ABC$,*




*we define the sine and cosine of the angle at $A$, denoted $t$, by the following formulae:*

$$\sin t = \frac{BC}{AC} \quad , \quad \cos t = \frac{AB}{BC}$$

*We call the sine and cosine basic trigonometric functions.*

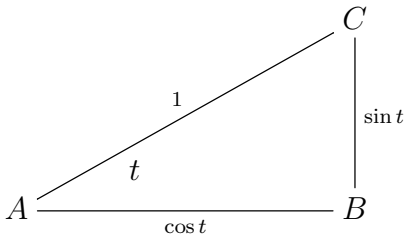
As a first observation, the sine and cosine do not depend on the choice of the given right triangle $ABC$ having an angle $t$ at $A$, and this due to the Thales theorem. In view of this, whenever possible, we will choose the right triangle $ABC$ as to have:

$$AC = 1$$

In this case, the formulae defining the sine and cosine simplify, as follows:

$$\sin t = BC \quad , \quad \cos t = AB$$

Equivalently, we can encode all this in a single picture, as follows:




As a few basic examples now, for the sine, coming from things that we know well, about right triangles, from the previous section, we have:

$$\sin 0^\circ = 0 \quad , \quad \sin 30^\circ = \frac{1}{2} \quad , \quad \sin 45^\circ = \frac{1}{\sqrt{2}} \quad , \quad \sin 60^\circ = \frac{\sqrt{3}}{2} \quad , \quad \sin 90^\circ = 1$$

Let us record as well the list of corresponding cosines. These are as follows:

$$\cos 0^\circ = 1 \quad , \quad \cos 30^\circ = \frac{\sqrt{3}}{2} \quad , \quad \cos 45^\circ = \frac{1}{\sqrt{2}} \quad , \quad \cos 60^\circ = \frac{1}{2} \quad , \quad \cos 90^\circ = 0$$

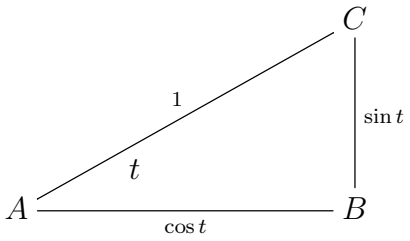
Observe that the numbers in the above two lists are the same, but written backwards in the second list. In fact, we have the following result, regarding this:

THEOREM 3.14. *The sines and cosines are subject to the formulae*

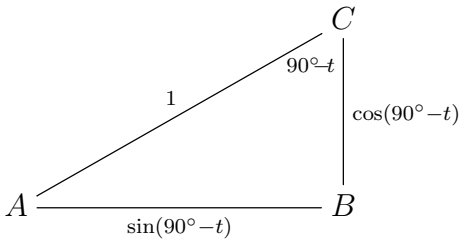
$$\sin(90^\circ - t) = \cos t \quad , \quad \cos(90^\circ - t) = \sin t$$

*valid for any angle* $t \in [0^\circ, 90^\circ]$.

PROOF. In order to understand this, the best is to choose our right triangle $ABC$ with $AC = 1$. In this case, the picture coming from Definition 3.13 is as follows:



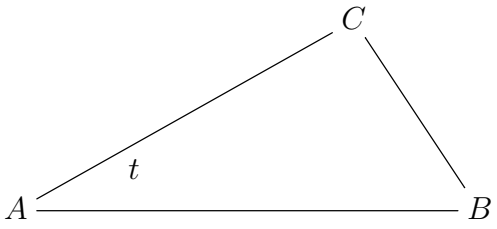

On the other hand, by focusing now at the angle at $C$, and perhaps twisting a bit our minds too, we have as well the following picture, for the same triangle:




Thus, we are led to the conclusion in the statement, and by the way congratulations, with this being our first trigonometry theorem. Many more to come. □

Before going ahead with more trigonometry, a question that you might have, why bothering with sine and cosine? Not clear, and in the lack of a bright idea here, and believe me, I asked my colleagues too, we will have to ask the cat. And cat declares:

CAT 3.15. *The area of an arbitrary triangle, having an angle* $t$ *at* $A$,

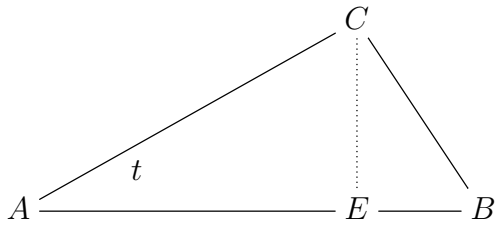



*is given by the following formula, making appear the sine:*

$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

*As for the need for cosines, homework for you buddy.*

Thanks cat, interesting all this, let us try to understand it. To start with, the formula of cat looks like some sort of mathematical theorem, that we must prove. But, in order to do so, the simplest is to draw an altitude of our triangle, as follows:




Now with this altitude drawn, we have the following computation:

$$\begin{aligned} area(ABC) &= \frac{basis \times height}{2} \\ &= \frac{AB \times CE}{2} \\ &= \frac{AB \times AC \times \sin t}{2} \end{aligned}$$

Thus, theorem proved, so the sine is definitely a good and useful thing, as cat says. As for the cosine, damn cat has assigned this to us as an exercise, so we will have to think about it, and come back to it, in due time. And no late homework, of course.

Moving forward now, still in relation with Cat 3.15, we have the following question:

QUESTION 3.16. *What happens to the cat formula,*

$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

*when the angle at $A$ is obtuse, $t > 90^\circ$?*

Which looks like a very good question. In answer now, given a triangle which is obtuse at $A$, we can simply rotate the $AC$ side to the right, as for that obtuse angle to become acute, $t' = 180^\circ - t$, and the area of the triangle obviously remains the same, and this since both the basis and height remain unchanged. Thus, the correct definition for $\sin t$ for obtuse angles should be the one making the following formula work:

$$\frac{AB \times AC \times \sin t}{2} = \frac{AB \times AC \times \sin(180^\circ - t)}{2}$$

Now by simplifying, we are led to the following formula:

$$\sin t = \sin(180^\circ - t)$$

Thus, Question 3.16 answered, with our conclusions being as follows:

THEOREM 3.17. *We can talk about the sine of any angle $t \in [0°, 180°]$, according to*

$$\sin t = \sin(180° - t)$$

*and with this, the cat formula for the area of a triangle, namely*

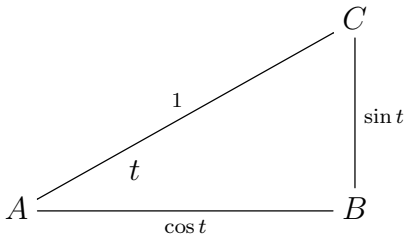
$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

*holds for any triangle, without any assumption on it.*

PROOF. This follows indeed from the above discussion. □

Moving ahead now, defining sines as in Definition 3.13 for $t \in [0°, 90°]$, and as above for $t \in [90°, 180°]$ certainly does the job, as explained above, but is not very elegant. So, let us try to improve this. We have here the following obvious speculation:

SPECULATION 3.18. *The sine of any angle $t \in [0°, 180°]$ can be defined geometrically, according to the usual picture*




*with the convention that for $t > 90°$, the triangle is drawn at the left of $A$.*

Which sounds quite good, but when thinking some more, things fine of course with the sine, but what about the cosine? The problem indeed is that, in the case $t > 90°$, when the triangle is drawn at the left of $A$, the lower side $AB$ changes orientation:

$$AB \to BA$$

But, as we know well from triangle geometry, from various considerations regarding segments and orientation, this would amount in saying that we are replacing:

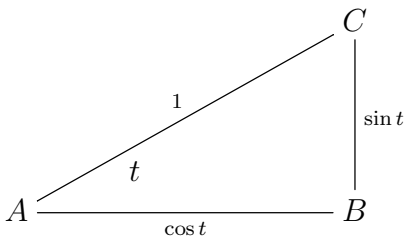
$$AB \to -AB$$

And so, we are led to the following formula for the cosine, in this case:

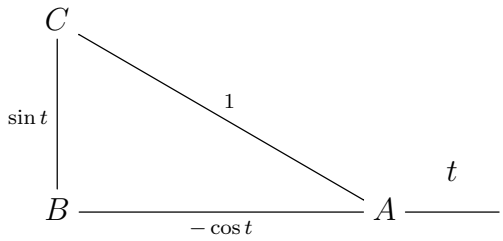
$$\cos t = -\cos(180° - t)$$

Very good all this, so let us update now Theorem 3.17, and by incorporating as well Speculation 3.18, in the form of a grand result, in the following way:

THEOREM 3.19 (update). *We can talk about the sine and cosine of any angle $t \in [0°, 180°]$, according to the following picture,*




*which in the case of obtuse angles becomes by definition as follows,*




*and with this, we have the following formulae, valid for any $t \in [0°, 180°]$:*

$$\sin t = \sin(180° - t) \quad , \quad \cos t = -\cos(180° - t)$$

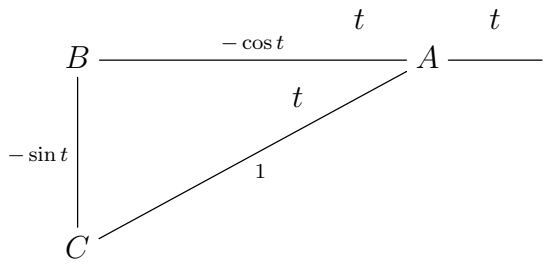
*Moreover, the cat formula for the area of a triangle, namely*

$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

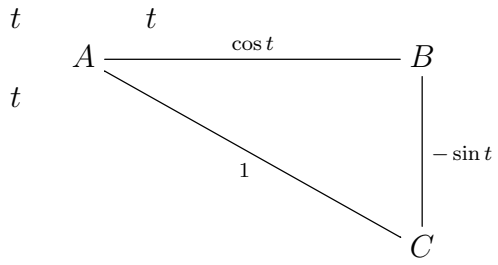
*holds for any triangle, without any assumption on it.*

PROOF. This follows indeed by putting together all the above. □

Which sounds quite good, and normally end of the story, but let us be crazy now, and try to talk as well about the sine or cosine of angles $t < 0°$, or $t > 180°$. Indeed, we know the recipe, namely suitably drawing our right triangle, with attention to positive and negatives. Thus, for $t \in [180°, 270°]$, our picture should be as follows:

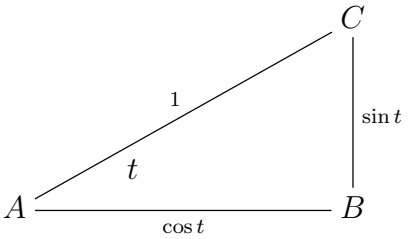


As for the next case, $t \in [270^\circ, 360^\circ]$, here our picture should be as follows:




But with this, we are done, because adding or substracting $360^\circ$ to our angles won't change the corresponding right triangle, and so won't change the sine and cosine. So, good work that we did, and time now to further improve Theorem 3.19, as follows:

THEOREM 3.20 (final update). *We can talk about the sine and cosine of any angle $t \in \mathbb{R}$, according to the following picture,*




*suitably drawn for angles $t < 0^\circ$, or $t > 90^\circ$, with attention to positive and negative lengths, as explained above. With this, all the basic formulae still hold, for any $t \in \mathbb{R}$.*

PROOF. This follows indeed by putting together all the above, and with the basic formulae in question being as follows, and in the hope that I forgot none:

$$\sin(90^\circ - t) = \cos t \quad , \quad \cos(90^\circ - t) = \sin t$$
$$\sin(90^\circ + t) = \cos t \quad , \quad \cos(90^\circ + t) = -\sin t$$

$$\sin(180^\circ - t) = \sin t \quad , \quad \cos(180^\circ - t) = -\cos t$$
$$\sin(180^\circ + t) = -\sin t \quad , \quad \cos(180^\circ + t) = -\cos t$$

$$\sin(270^\circ - t) = -\cos t \quad , \quad \cos(270^\circ - t) = -\sin t$$
$$\sin(270^\circ + t) = -\cos t \quad , \quad \cos(270^\circ + t) = \sin t$$

$$\sin(360^\circ - t) = -\sin t \quad , \quad \cos(360^\circ - t) = \cos t$$
$$\sin(360^\circ + t) = \sin t \quad , \quad \cos(360^\circ + t) = \cos t$$

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

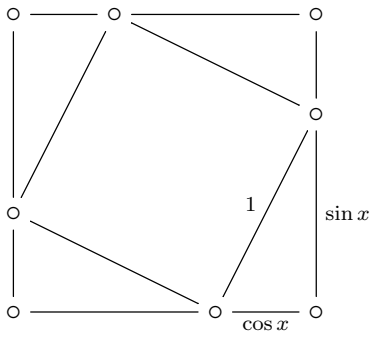
Getting now to more advanced theory, we first have:

THEOREM 3.21. *The sines and cosines are subject to the formula*

$$\sin^2 x + \cos^2 x = 1$$

*coming from Pythagoras' theorem.*

PROOF. This is something which is certainly true, and for pure mathematical pleasure, let us reproduce the picture leading to Pythagoras, in the trigonometric setting:




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 the tangent:

$$\tan x = \frac{\sin x}{\cos x}$$

But more on this, such as various analytic aspects, later in this book, once we will have some appropriate tools, beyond basic geometry, in order to discuss this.

Still at the level of the basics, we have the following result:

THEOREM 3.22. *The sines and cosines of sums are given by*

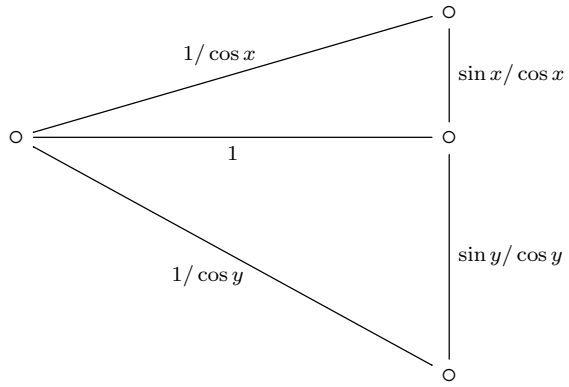
$$\sin(x + y) = \sin x \cos y + \cos x \sin y$$

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

*and these formulae give a formula for* $\tan(x + y)$ *too.*

PROOF. This is something quite tricky, using the same idea as in the proof of Pythagoras' theorem, that is, computing certain areas, the idea being as follows:

(1) Consider the following picture, consisting of a length 1 line segment, with angles $x, y$ drawn on each side, and with the lengths being computed, as indicated:




Now let us compute the area of the big triangle, or rather the double of that area. We can do this in two ways, either directly, with a formula involving $\sin(x+y)$, 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 $\sin(x+y)$ from the statement.

(2) By using $\sin(x+y)$ we can deduce a formula for $\cos(x+y)$, as follows:

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

(3) Finally, in what regards the tangents, we have, according to the above:

$$\tan(x+y)=\frac{\sin x\cos y+\cos x\sin y}{\cos x\cos y-\sin x\sin y}$$

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

There are many applications of Theorem 3.22. Observe in particular that with $x=y$ we obtain some interesting formulae for the duplication of angles, namely:

$$\sin(2x)=2\sin x\cos x$$

$$\cos(2x)=\cos^2x-\sin^2x$$

Regarding the sines and cosines of triples of angles, or higher, things here are more complicated. We will be back to such questions later, with better tools.

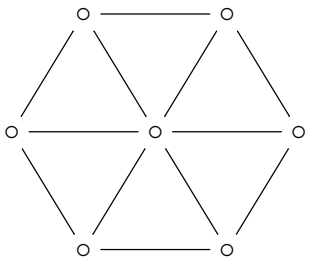
## 3c. Pi, trigonometry

Let us get now into a more advanced study of the angles. For this purpose, the best is to talk first about circles, and the number $\pi$. And here, to start with, we have:

THEOREM 3.23. *The following two definitions of $\pi$ 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 $\pi$, the above picture at $N = 6$ shows that we have $\pi > 3$, but not by much. The precise figure is as follows:

$$\pi = 3.14159\ldots$$

Getting now to what we wanted to do in this section, in relation with the angles, and their numeric measuring, let us formulate the following definition:

DEFINITION 3.24. *The value of an angle is obtained by putting that angle on the center of a circle of radius* 1*, and measuring the corresponding arc length.*

And this, which is something quite smart, will replace our previous conventions for the measuring of angles, with the basic conversion formulae being as follows:

$$0^\circ = 0 \quad , \quad 90^\circ = \frac{\pi}{2} \quad , \quad 180^\circ = \pi \quad , \quad 270^\circ = \frac{3\pi}{2}$$

Let us record as well the conversion formulae for the halves of these angles:

$$45^\circ = \frac{\pi}{4} \quad , \quad 135^\circ = \frac{3\pi}{4} \quad , \quad 225^\circ = \frac{5\pi}{4} \quad , \quad 315^\circ = \frac{7\pi}{4}$$

Finally, let us record as well the formulae for the thirds of the basic angles:

$$30^\circ = \frac{\pi}{6} \quad , \quad 60^\circ = \frac{\pi}{3} \quad , \quad 120^\circ = \frac{2\pi}{3} \quad , \quad 150^\circ = \frac{5\pi}{6}$$

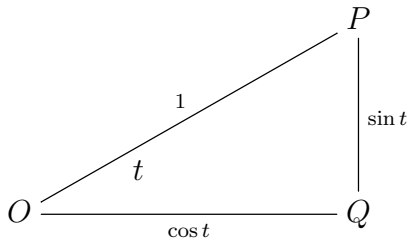
$$210^\circ = \frac{7\pi}{6} \quad , \quad 240^\circ = \frac{4\pi}{3} \quad , \quad 300^\circ = \frac{5\pi}{3} \quad , \quad 330^\circ = \frac{11\pi}{6}$$

In relation now with sin and cos, we are led in this way to the following alternate definitions, which better explain the various sign conventions made before:

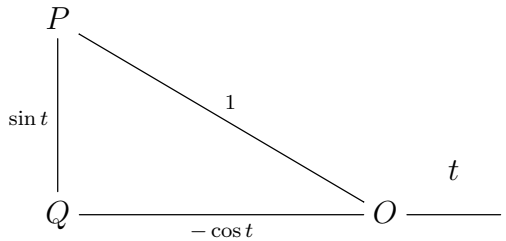
THEOREM 3.25. *The sine and cosine of an angle are obtained by putting the angle on the unit circle, as above, then projecting on the vertical and the horizontal, and then measuring the oriented segments that we get, on the vertical and horizontal.*

PROOF. This is clear from definitions, but for full clarity here, let us review now the detailed construction of the sine and cosine, for the arbitrary angles, from the previous section, with attention to signs, in the present setting. We have 4 cases, as follows:

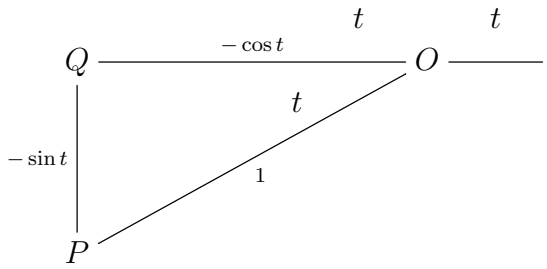
(1) In the simplest case, namely $t \in [0, \pi/2]$, the sine and cosine are indeed computed according to the following picture, which is the one in the statement:



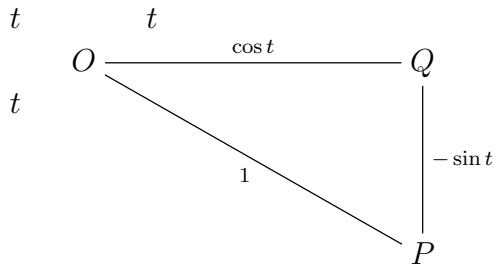

(2) In the case of obtuse angles, $t \in [\pi/2, \pi]$, the picture becomes as follows:



(3) In the next case, namely $t \in [\pi, 3\pi/2]$, the picture becomes as follows:




(4) As for the last case, namely $t \in [3\pi/2, 2\pi]$, here our picture is as follows:




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

As an interesting fact, we can complement Theorem 3.25 with a statement regarding the tangent, the trigonometric function that we often forgot so far, as follows:

THEOREM 3.26. *The tangent of an angle can be obtained by putting the angle on the unit circle, as before, and then measuring the oriented segment that we get, on the vertical, outside the circle, on the vertical tangent at right.*

PROOF. This is, again, something quite self-explanatory. □

With the circle in hand, we have the following estimates, which are both clear:

$$\sin x \leq x \leq \tan x$$

Moreover, we can now establish some useful estimates, as follows:

THEOREM 3.27. *The following happen, for small angles, $x \simeq 0$:*

(1) $\sin x \simeq x$.
(2) $\cos x \simeq 1 - x^2/2$.
(3) $\tan x \simeq x$.

PROOF. Here (1) is clear on the circle, (2) comes from (1) and from Pythagoras, by computing the quantity doing the job, and (3) is clear on the circle too. □

## 3d. Complex numbers

We have a quite good understanding of their complex numbers, and their addition. In order to understand now the multiplication operation, we must do something more complicated, namely using polar coordinates. Let us start with:

DEFINITION 3.28. *The complex numbers $x = a+ib$ can be written in polar coordinates,*

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

*with the connecting formulae being as follows,*

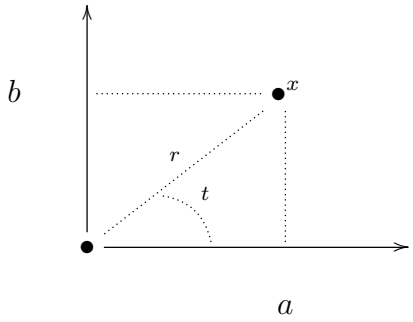
$$a = r \cos t \quad , \quad b = r \sin t$$

*and in the other sense being as follows,*

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

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

There is a clear relation here with the vector notation from chapter 2, because $r$ is the length of the vector, and $t$ is the angle made by the vector with the $Ox$ axis. To be more precise, the picture for what is going on in Definition 3.28 is as follows:




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

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

The point now is that in polar coordinates, the multiplication formula for the complex numbers, which was so far something quite opaque, takes a very simple form:

THEOREM 3.29. *Two complex numbers written in polar coordinates,*

$$x = r(\cos s + i \sin s) \quad , \quad y = p(\cos t + i \sin t)$$

*multiply according to the following formula:*

$$xy = rp(\cos(s + t) + i \sin(s + t))$$

*In other words, the moduli multiply, and the arguments sum up.*

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

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

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

Indeed, we can assume that we have $r = p = 1$, by dividing everything by these numbers. Now with this assumption made, we have the following computation:

$$\begin{aligned} xy &= (\cos s + i \sin s)(\cos t + i \sin t) \\ &= (\cos s \cos t - \sin s \sin t) + i(\cos s \sin t + \sin s \cos t) \\ &= \cos(s+t) + i \sin(s+t) \end{aligned}$$

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

The above result, which is based on some non-trivial trigonometry, is quite powerful. As a basic application of it, we can now compute powers, as follows:

THEOREM 3.30. *The powers of a complex number, written in polar form,*

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

*are given by the following formula, valid for any exponent $k \in \mathbb{N}$:*

$$x^k = r^k(\cos kt + i \sin kt)$$

*Moreover, this formula holds in fact for any $k \in \mathbb{Z}$, and even for any $k \in \mathbb{Q}$.*

PROOF. Given a complex number $x$, written in polar form as above, and an exponent $k \in \mathbb{N}$, we have indeed the following computation, with $k$ terms everywhere:

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

Thus, we are done with the case $k \in \mathbb{N}$. Regarding now the generalization to the case $k \in \mathbb{Z}$, it is enough here to do the verification for $k = -1$, where the formula is:

$$x^{-1} = r^{-1}(\cos(-t) + i \sin(-t))$$

But this number $x^{-1}$ is indeed the inverse of $x$, as shown by:

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

Finally, regarding the generalization to the case $k \in \mathbb{Q}$, it is enough to do the verification for exponents of type $k = 1/n$, with $n \in \mathbb{N}$. The claim here is that:

$$x^{1/n} = r^{1/n}\left[\cos\left(\frac{t}{n}\right) + i\sin\left(\frac{t}{n}\right)\right]$$

In order to prove this, let us compute the $n$-th power of this number. We can use the power formula for the exponent $n \in \mathbb{N}$, that we already established, and we obtain:

$$\begin{aligned}(x^{1/n})^n &= (r^{1/n})^n\left[\cos\left(n\cdot\frac{t}{n}\right) + i\sin\left(n\cdot\frac{t}{n}\right)\right]\\ &= r(\cos t + i\sin t)\\ &= x\end{aligned}$$

Thus, we have indeed a $n$-th root of $x$, and our proof is now complete. □

As a basic application of Theorem 3.30, we have the following result:

PROPOSITION 3.31. *Each complex number, written in polar form,*

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

*has two square roots, given by the following formula:*

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

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

PROOF. The first assertion is clear indeed from the general formula in Theorem 3.30, at $k = 1/2$. As for its particular cases with $x \in \mathbb{R}$, these are clear from it. □

With the above results in hand, and notably with the square root formula from Proposition 3.31, we can now go back to the degree 2 equations, and we have:

THEOREM 3.32. *The complex solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{C}$ are*

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

*with the square root of complex numbers being defined as above.*

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

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

Now since we know from Proposition 3.31 that any complex number has a square root, we are led to the conclusion in the statement. □

We can investigate as well more complicated operations, as follows:

THEOREM 3.33. *We have the following operations on the complex numbers*

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

*written in short polar form* $x = re^{it}$*:*

(1) *Inversion:* $(re^{it})^{-1} = r^{-1}e^{-it}$.
(2) *Square roots:* $\sqrt{re^{it}} = \pm\sqrt{r}e^{it/2}$.
(3) *Powers:* $(re^{it})^a = r^a e^{ita}$.
(4) *Conjugation:* $\overline{re^{it}} = re^{-it}$.

PROOF. As a first observation, with the short polar form convention $x = re^{it}$ from the statement, the multiplication formula from Theorem 3.29 takes the following very simple form, with the arguments $s, t$ being both taken modulo $2\pi$:

$$re^{is} \cdot pe^{it} = rp\, e^{i(s+t)}$$

Getting now to what is to be proved, we basically already know all this, but we can now rediscuss this, from a more conceptual viewpoint, the idea being as follows:

(1) We have indeed the following computation, using the above multiplication formula:

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

(2) Once again by using the above multiplication formula, we have:

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

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

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

We conclude that this operation $x \to x^a$ is indeed the correct one.

(4) This comes from the fact, that we know from the above, that the conjugation operation $x \to \bar{x}$ keeps the modulus, and switches the sign of the argument. □

We kept the best for the end. As a last topic regarding the complex numbers, which is something really beautiful, we have the roots of unity. Let us start with:

THEOREM 3.34. *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 the general multiplication formula for complex numbers from above. 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 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 3.33, the statement here being as follows:

THEOREM 3.35. *Any $x=re^{it}$ 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:

$$z^N=y^N$$

We conclude from this that the solutions $z$ appear by multiplying $y$ by the solutions of $t^N=1$, which are the $N$-th roots of unity, as claimed. □

## 3e. Exercises

Exercises:

EXERCISE 3.36.

EXERCISE 3.37.

EXERCISE 3.38.

EXERCISE 3.39.

EXERCISE 3.40.

EXERCISE 3.41.

EXERCISE 3.42.

EXERCISE 3.43.

Bonus exercise.

CHAPTER 4

# Exp and log

## 4a. The number e

Time now to get into the truly scary things, namely exp and log. These are quite basic functions in mathematics and science, but in order to introduce them, we will have to work a bit. Indeed, the idea will be that the exponential will be something of type $\exp x = e^x$, and the logarithm $\log x$ will be its inverse, but the whole point lies in understanding what the number $e \in \mathbb{R}$ that we really want to use is, and this is something non-trivial.

So, be patient, things here will be non-trivial, and we will have to prove some preliminary results first, with no clear goal. But, as analysts, we are supposed to enjoy everything analysis, so take what comes next like this, analysts enjoying analysis.

Regarding $e$, we have the following remarkable result, to start with:

THEOREM 4.1. *We have the following convergence*

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

*where* $e = 2.71828\ldots$ *is a certain number.*

PROOF. This is something quite tricky, as follows:

(1) Our first claim is that the following sequence is increasing:

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

In order to prove this, 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. □

More generally now, we have the following result, which is a bit more general:

THEOREM 4.2. *We have the following formula,*

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

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

PROOF. We already know from Theorem 4.1 that the result holds at $x = 1$, and this because the number $e$ was by definition given by the following formula:

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

By taking inverses, we obtain as well the result at $x = -1$, namely:

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

In general now, when $\in \mathbb{R}$ is arbitrary, the best is to proceed as follows:

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

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

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

THEOREM 4.3. *We have the formula*

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

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

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

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

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

(2) For this purpose, the first observation is that we have the following estimate:

$$2 < \sum_{k=0}^{\infty} \frac{1}{k!} < \sum_{k=0}^{\infty} \frac{1}{2^{k-1}} = 3$$

Thus, the series $\sum_{k=0}^{\infty} \frac{1}{k!}$ converges indeed, towards a limit in $(2, 3)$.

(3) In order to prove now that this limit is $e$, observe that we have:

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

Thus, with $n \to \infty$, we get that the limit of the series $\sum_{k=0}^{\infty} \frac{1}{k!}$ belongs to $[e, 3)$.

(4) For the reverse inequality, we use the following computation:

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

(5) In order to estimate the above expression that we found, we can use the following trivial inequality, valid for any number $x \in (0, 1)$:

$$1 - x^k = (1 - x)(1 + x + x^2 + \ldots + x^{k-1}) \leq (1 - x)k$$

Indeed, we can use this with $x = 1 - k/n$, and we obtain in this way:

$$\begin{aligned}
\sum_{k=0}^{n} \frac{1}{k!} - \left(1 + \frac{1}{n}\right)^n &\leq \sum_{k=2}^{n} \frac{\frac{k}{n} \cdot k}{k!} \\
&= \frac{1}{n} \sum_{k=2}^{n} \frac{k}{(k-1)!} \\
&= \frac{1}{n} \sum_{k=2}^{n} \frac{k}{k-1} \cdot \frac{1}{(k-2)!} \\
&\leq \frac{1}{n} \sum_{k=2}^{n} \frac{2}{2^{k-2}} \\
&< \frac{4}{n}
\end{aligned}$$

Now since with $n \to \infty$ this goes to 0, we obtain that the limit of the series $\sum_{k=0}^{\infty} \frac{1}{k!}$ is the same as the limit of the sequence $\left(1 + \frac{1}{n}\right)^n$, manely $e$. Thus, getting back now to what we wanted to prove, our theorem, we are done in this way with the case $x = 1$.

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

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

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

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

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

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

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

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

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

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

Observe that we used in the above a few things about functions, which are all intuitive, but not exactly trivial to prove. We will be back to this, with details, in Part II.

As another observation, the proof of Theorem 4.3 leads in fact to:

THEOREM 4.4. *We have the following formula,*

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

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

PROOF. This is indeed something that we know, from the proof of Theorem 4.3, and we will take the present statement as a conclusion to what we did so far. □

## 4b. More about e

Many things can be said about $e$, and we will be back to this on a regular basis, in this book. As a basic result here, which is more advanced, we have:

THEOREM 4.5. *The number $e$ from analysis, given by*

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

*which numerically means $e=2.7182818284\ldots$, is irrational.*

PROOF. Many things can be said here, as follows:

(1) To start with, there are several possible definitions for the number $e$, with the old style one, that we have used in the above, being via a simple limit, as follows:

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

The definition in the statement is the modern one, explained also in the above.

(2) Getting now to numerics, the series of $e$ converges very fast, when compared to the old style sequence in (1), so if you are in a hurry, this series is for you. We have:

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

Thus, the error term in the approximation is really tiny, the estimate being:

$$\sum_{k=0}^{N}\frac{1}{k!}<e<\sum_{k=0}^{N}\frac{1}{k!}+\frac{1}{N\cdot N!}$$

(3) Now by using this, you can easily compute the decimals of $e$. Actually, you can't call yourself mathematician, or scientist, if you haven't done this by hand, just for the fun, but just in case, here is how the approximation goes, for small values of $N$:

$$N=2 \implies 2.5<e<2.75$$

$$N=3 \implies 2.666\ldots<e<2.722\ldots$$

$$N=4 \implies 2.70833\ldots<e<2.71875\ldots$$

$$N=5 \implies 2.71666\ldots<e<2.71833\ldots$$

$$N=6 \implies 2.71805\ldots<e<2.71828\ldots$$

$$N=7 \implies 2.71825\ldots<e<2.71828\ldots$$

Thus, first 4 decimals computed, $e=2.7182\ldots$, and I would leave the continuation to you. With the remark that, when carefully looking at the above, the estimate on the right works much better than the one on the left, so before getting into more serious numerics, try to find a better lower estimate for $e$, that can help you in your work.

(4) Getting now to irrationality, a look at $e=2.7182818284\ldots$ might suggest that the $81, 82, 84\ldots$ values might eventually, after some internal fight, decide for a winner, and so that $e$ might be rational. However, this is wrong, and $e$ is in fact irrational.

(5) So, let us prove now this, that $e$ is irrational. Following Fourier, we will do this by contradiction. So, assume $e = m/n$, and let us look at the following number:

$$x = n!\left(e - \sum_{k=0}^{n} \frac{1}{k!}\right)$$

As a first observation, $x$ is an integer, as shown by the following computation:

$$\begin{aligned} x &= n!\left(\frac{m}{n} - \sum_{k=0}^{n} \frac{1}{k!}\right) \\ &= m(n-1)! - \sum_{k=0}^{n} n(n-1)\ldots(n-k+1) \\ &\in \mathbb{Z} \end{aligned}$$

On the other hand $x > 0$, and we have as well the following estimate:

$$\begin{aligned} x &= n! \sum_{k=n+1}^{\infty} \frac{1}{k!} \\ &= \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \ldots \\ &< \frac{1}{n+1} + \frac{1}{(n+1)^2} + \ldots \\ &= \frac{1}{n} \end{aligned}$$

Thus $x \in (0, 1)$, which contradicts our previous finding $x \in \mathbb{Z}$, as desired. □

Still talking about $e$, I don't know about you, but personally, for some peace of mind, I would like to have as well a combinatorial interpretation of it. So, let us start with some combinatorics. We first have the following well-known, and useful formula:

$$\left(\bigcup_i A_i\right)^c = \bigcap_i A_i^c$$

We have as well a reverse formula, of the same type, as follows:

$$\left(\bigcap_i A_i\right)^c = \bigcup_i A_i^c$$

At a more advanced level, we have the inclusion-exclusion principle, which has many concrete applications. This inclusion-exclusion principle is as follows:

PROPOSITION 4.6. *We have the following formula,*

$$\left|\left(\bigcup_i A_i\right)^c\right| = |A| - \sum_i |A_i| + \sum_{i<j} |A_i \cap A_j| - \sum_{i<j<k} |A_i \cap A_j \cap A_k| + \ldots$$

*called inclusion-exclusion principle.*

PROOF. This is indeed quite clear, by thinking a bit, as follows:

(1) In order to count $(\cup_i A_i)^c$, we certainly have to start with $|A|$.

(2) Then, we obviously have to remove each $|A_i|$, and so remove $\sum_i |A_i|$.

(3) But then, we have to put back each $|A_i \cap A_j|$, and so put back $\sum_{i<j} |A_i \cap A_j|$.

(4) Then, we must remove each $|A_i \cap A_j \cap A_k|$, so remove $\sum_{i<j<k} |A_i \cap A_j \cap A_k|$.

$\vdots$

(5) And so on, which leads to the formula in the statement. □

Getting now towards what we wanted to do, in relation with $e$, let us start with the following definition, which is something very standard:

DEFINITION 4.7. *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 basic 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 certainly faster, and which is actually my personal favorite, is by denoting the permutations as diagrams, acting from top to bottom:

$\sigma =$ 

Here are some basic properties of the permutations:

PROPOSITION 4.8. *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. □

With this discussed, here is now the application of the inclusion-exclusion principle that we were having in mind, making appear $e$, in a nice combinatorial way:

THEOREM 4.9. *The probability for a random permutation $\sigma \in S_N$ to be a derangement, that is, to have no fixed points, is given by the following formula:*

$$P = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \ldots + (-1)^N \frac{1}{N!}$$

*Thus we have the following asymptotic formula, in the $N \to \infty$ limit,*

$$P \simeq \frac{1}{e}$$

*with $e = 2.7182\ldots$ being the usual constant from analysis.*

PROOF. This is something very classical, which is best viewed by using the inclusion-exclusion principle. Consider indeed the following sets:

$$S_N^i = \left\{ \sigma \in S_N \middle| \sigma(i) = i \right\}$$

The set of permutations having no fixed points, called derangements, is then:

$$X_N = \left( \bigcup_k S_N^k \right)^c$$

Now the inclusion-exclusion principle tells us that we have:

$$\begin{aligned}
|X_N| &= \left| \left( \bigcup_k S_N^k \right)^c \right| \\
&= |S_N| - \sum_k |S_N^k| + \sum_{k<l} |S_N^k \cap S_N^l| - \ldots + (-1)^N \sum_{k_1<\ldots<k_N} |S_N^{k_1} \cup \ldots \cup S_N^{k_N}| \\
&= N! - N(N-1)! + \binom{N}{2}(N-2)! - \ldots + (-1)^N \binom{N}{N}(N-N)! \\
&= \sum_{r=0}^{N} (-1)^r \binom{N}{r} (N-r)!
\end{aligned}$$

Thus, the probability that we are interested in, for a random permutation $\sigma \in S_N$ to have no fixed points, is given by the following formula:

$$P = \frac{|X_N|}{N!} = \sum_{r=0}^{N} \frac{(-1)^r}{r!}$$

Since on the right we have the expansion of $1/e$, this gives the result. □

The above is nice, but we can do even better. Let us introduce, indeed:

DEFINITION 4.10. *The Poisson law of parameter* 1 *is the following measure,*

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

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

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

*with the letter "p" standing for Poisson.*

Generally speaking, these measures appear a bit everywhere, in discrete probability contexts, the reasons for this coming from the Poisson Limit Theorem (PLT). However, this is advanced, and for the moment, what we have in Definition 4.10 will do. Observe that our laws have indeed mass 1, as they should, due to the following key formula:

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

Getting back now to what we wanted to do, generalize Theorem 4.9, we have:

THEOREM 4.11. *The main character of* $S_N$*, which counts the fixed points, given by*

$$\chi = \sum_i \sigma_{ii}$$

*via the standard embedding* $S_N \subset O_N$*, follows the Poisson law* $p_1$*, in the* $N \to \infty$ *limit. More generally, the truncated characters of* $S_N$*, given by*

$$\chi_t = \sum_{i=1}^{[tN]} \sigma_{ii}$$

*with* $t \in (0, 1]$*, follow the Poisson laws* $p_t$*, in the* $N \to \infty$ *limit.*

PROOF. Many things going on here, the idea being as follows:

(1) Let us construct the main character of $S_N$, as in the statement. The permutation matrices being given by $\sigma_{ij}=\delta_{i\sigma(j)}$, we have the following formula:

$$\chi(\sigma)=\sum_i\delta_{\sigma(i)i}=\#\left\{i\in\{1,\ldots,N\}\Big|\sigma(i)=i\right\}$$

In order to establish now the asymptotic result in the statement, regarding these characters, we must prove the following formula, for any $r\in\mathbb{N}$, in the $N\to\infty$ limit:

$$P(\chi=r)\simeq\frac{1}{r!e}$$

We already know, from Theorem 4.9, that this formula holds at $r=0$. In the general case now, we have to count the permutations $\sigma\in S_N$ having exactly $r$ points. Now since having such a permutation amounts in choosing $r$ points among $1,\ldots,N$, and then permuting the $N-r$ points left, without fixed points allowed, we have:

$$\begin{aligned}\#\left\{\sigma\in S_N\Big|\chi(\sigma)=r\right\}&=\binom{N}{r}\#\left\{\sigma\in S_{N-r}\Big|\chi(\sigma)=0\right\}\\&=\frac{N!}{r!(N-r)!}\#\left\{\sigma\in S_{N-r}\Big|\chi(\sigma)=0\right\}\\&=N!\times\frac{1}{r!}\times\frac{\#\left\{\sigma\in S_{N-r}\Big|\chi(\sigma)=0\right\}}{(N-r)!}\end{aligned}$$

By dividing everything by $N!$, we obtain from this the following formula:

$$\frac{\#\left\{\sigma\in S_N\Big|\chi(\sigma)=r\right\}}{N!}=\frac{1}{r!}\times\frac{\#\left\{\sigma\in S_{N-r}\Big|\chi(\sigma)=0\right\}}{(N-r)!}$$

Now by using the computation at $r=0$, that we already have, from Theorem 4.9, it follows that with $N\to\infty$ we have the following estimate:

$$P(\chi=r)\simeq\frac{1}{r!}\cdot P(\chi=0)\simeq\frac{1}{r!}\cdot\frac{1}{e}$$

Thus, we obtain as limiting measure the Poisson law of parameter 1, as stated.

(2) Let us construct now the truncated characters of $S_N$, as in the statement. As before in the case $t=1$, we have the following computation, coming from definitions:

$$\chi_t(\sigma)=\sum_{i=1}^{[tN]}\delta_{\sigma(i)i}=\#\left\{i\in\{1,\ldots,[tN]\}\Big|\sigma(i)=i\right\}$$

Also as before, we obtain by inclusion-exclusion that we have:

$$\begin{aligned}
P(\chi_t = 0) &= \frac{1}{N!}\sum_{r=0}^{[tN]}(-1)^r\sum_{k_1<\ldots<k_r<[tN]}|S_N^{k_1}\cap\ldots\cap S_N^{k_r}| \\
&= \frac{1}{N!}\sum_{r=0}^{[tN]}(-1)^r\binom{[tN]}{r}(N-r)! \\
&= \sum_{r=0}^{[tN]}\frac{(-1)^r}{r!}\cdot\frac{[tN]!(N-r)!}{N!([tN]-r)!}
\end{aligned}$$

Now with $N\to\infty$, we obtain from this the following estimate:

$$P(\chi_t = 0)\simeq\sum_{r=0}^{[tN]}\frac{(-1)^r}{r!}\cdot t^r\simeq e^{-t}$$

More generally, by counting the permutations $\sigma\in S_N$ having exactly $r$ fixed points among $1,\ldots,[tN]$, as in the proof of (2), we obtain:

$$P(\chi_t = r)\simeq\frac{t^r}{r!e^t}$$

Thus, we obtain in the limit a Poisson law of parameter $t$, as stated. □

Many other things can be said, as a continuation of the above. We will be back to this on several occassions, once we will be better at calculus, which is needed here.

## 4c. Complex powers

Now that we know about $e=2.7182\ldots$, we would like to understand the meaning of the formula $x=re^{it}$ for the complex numbers, that we used in chapter 3. However, this is no easy task, and we will be punching here a bit above our weight.

Nevermind. So, this will be some kind of physics class. Let us start with:

THEOREM 4.12. *We can exponentiate the complex numbers, according to*

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

*and the function $x\to e^x$ satisfies $e^{x+y}=e^xe^y$.*

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

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

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

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

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

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

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

$$\lim_{t\to 0} e^{x+t} = \lim_{t\to 0} e^x e^t = e^x \lim_{t\to 0} e^t = e^x \cdot 1 = e^x$$

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

Next, we have the following deep result, regarding the complex exponential:

THEOREM 4.13. *We have the following formula,*

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

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

PROOF. Let us first recall from Theorem 4.12 that we have the following formula, for the exponential of an arbitrary complex number $x \in \mathbb{C}$:

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

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

$$\begin{aligned} e^{it} &= \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \\ &= \sum_{k=2l} \frac{(it)^k}{k!} + \sum_{k=2l+1} \frac{(it)^k}{k!} \\ &= \sum_{l=0}^{\infty} (-1)^l \frac{t^{2l}}{(2l)!} + i \sum_{l=0}^{\infty} (-1)^l \frac{t^{2l+1}}{(2l+1)!} \end{aligned}$$

The point now is that, according to the findings of our colleagues in theoretical physics, we have the following two formulae, for the cosine and sine:

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

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

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

THEOREM 4.14. *The complex numbers $x = a + ib$ can be written in polar coordinates,*

$$x = re^{it}$$

*with the connecting formulae being*

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

*and in the other sense being*

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

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

PROOF. This is a reformulation of our previous polar writing notions, by using the formula $e^{it} = \cos t + i \sin t$ from Theorem 4.13, and multiplying everything by $r$. □

With this in hand, we can now go back to the basics, namely the addition and multiplication of the complex numbers. We have here the following result:

THEOREM 4.15. *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 $2\pi$.*

PROOF. This is something that we already know, from chapter 3, reformulated by using the notations from Theorem 4.14. 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. □

As a basic application of Theorem 4.15, we have:

THEOREM 4.16. *We have the following operations on the complex numbers, written in polar form, as above:*

(1) *Inversion:* $(re^{it})^{-1} = r^{-1}e^{-it}$.
(2) *Square roots:* $\sqrt{re^{it}} = \pm\sqrt{r}e^{it/2}$.
(3) *Powers:* $(re^{it})^a = r^a e^{ita}$.
(4) *Conjugation:* $\overline{re^{it}} = re^{-it}$.

PROOF. This is something that we already know, from chapter 3, but we can now discuss all this, from a more conceptual viewpoint, the idea being as follows:

(1) We have indeed the following computation, using Theorem 4.15:

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

(2) Once again by using Theorem 4.15, we have:

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

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

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

Due to Theorem 4.15, this operation $x \to x^a$ is indeed the correct one.

(4) This comes from the fact, that we know from chapter 3, that the conjugation operation $x \to \bar{x}$ keeps the modulus, and switches the sign of the argument. □

Let us rewrite as well the theory of roots of unity, in this way. We first have:

THEOREM 4.17. *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 the general multiplication formula for complex numbers from Theorem 4.15. 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 $w = e^{2\pi i/3}$.

$\underline{N = 7}$. Here, with $w = e^{2\pi i/7}$, the roots of unity are $1, w, w^2, w^3, w^4, w^5, w^6$.

$\underline{N = 8}$. Here the roots of unity, read as usual counterclockwise, are the numbers $1, w, i, iw, -1, -w, -i, -iw$, with $w = e^{\pi i/4}$, which is also given by $w = (1 + i)/\sqrt{2}$.

The roots of unity are very useful variables, and have many interesting properties. As a first application, we can now solve the ambiguity questions related to the extraction of $N$-th roots, from Theorem 4.16, the statement here being as follows:

THEOREM 4.18. *Any nonzero complex number, written as*

$$x = re^{it}$$

*has exactly $N$ roots of order $N$, which appear as*

$$y = r^{1/N} e^{it/N}$$

*multiplied by the $N$ roots of unity of order $N$.*

PROOF. We must solve the equation $z^N = x$, over the complex numbers. Since the number $y$ in the statement clearly satisfies $y^N = x$, our equation is equivalent to:

$$z^N = y^N$$

Now observe that we can write this equation as follows:

$$\left(\frac{z}{y}\right)^N = 1$$

We conclude that the solutions $z$ appear by multiplying $y$ by the solutions of $t^N = 1$, which are the $N$-th roots of unity, as claimed. □

The roots of unity appear in connection with many other interesting questions, and there are many useful formulae relating them, which are good to know. Here is a basic such formula, very beautiful, to be used many times in what follows:

THEOREM 4.19. *The roots of unity, $\{w^k\}$ with $w = e^{2\pi i/N}$, have the property*

$$\sum_{k=0}^{N-1}(w^k)^s = N\delta_{N|s}$$

*for any exponent $s \in \mathbb{N}$, where on the right we have a Kronecker symbol.*

PROOF. The numbers in the statement, when written more conveniently as $(w^s)^k$ with $k = 0, \ldots, N-1$, form a certain regular polygon in the plane $P_s$. Thus, if we denote by $C_s$ the barycenter of this polygon, we have the following formula:

$$\frac{1}{N}\sum_{k=0}^{N-1} w^{ks} = C_s$$

Now observe that in the case $N\not|\, s$ our polygon $P_s$ is non-degenerate, circling around the unit circle, and having center $C_s = 0$. As for the case $N|s$, here the polygon is degenerate, lying at 1, and having center $C_s = 1$. Thus, we have the following formula:

$$C_s = \delta_{N|s}$$

Thus, we obtain the formula in the statement. □

As an interesting philosophical fact, regarding the roots of unity, and the complex numbers in general, we can now solve the following equation, in a "uniform" way:

$$x_1 + \ldots + x_N = 0$$

With this being not a joke. Frankly, can you find some nice-looking family of real numbers $x_1, \ldots, x_N$ satisfying $x_1 + \ldots + x_N = 0$? Certainly not. But with complex numbers we have now our answer, the sum of the $N$-th roots of unity being zero.

As an interesting consequence now of the above results, which is of practical interest, we have the following useful method, for remembering the basic math formulae:

METHOD 4.20. *Knowing $e^x = \sum_k x^k/k!$ and $e^{ix} = \cos x + i\sin x$ gives you*

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

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

*right away, in case you forgot these formulae, as well as*

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

*again, right away, in case you forgot these formulae.*

To be more precise, assume that we forgot everything trigonometry, which is something that can happen to everyone, in the real life, but still know the formulae $e^x = \sum_k x^k/k!$ and $e^{ix} = \cos x + i \sin x$. Then, we can recover the formulae for sums, as follows:

$$e^{i(x+y)} = e^{ix}e^{iy} \implies \cos(x+y) + i\sin(x+y) = (\cos x + i \sin x)(\cos y + i \sin y)$$
$$\implies \begin{cases} \cos(x+y) = \cos x \cos y - \sin x \sin y \\ \sin(x+y) = \sin x \cos y + \cos x \sin y \end{cases}$$

And isn't this smart. Also, and even more impressively, we can recover the physics formulae for $\sin, \cos$, which are certainly difficult to memorize, as follows:

$$e^{ix} = \sum_k \frac{(ix)^k}{k!} \implies \cos x + i \sin x = \sum_k \frac{(ix)^k}{k!}$$
$$\implies \begin{cases} \cos x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l}}{(2l)!} \\ \sin x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l+1}}{(2l+1)!} \end{cases}$$

Finally, in what regards log, there is a trick here too, which is partial, namely:

$$\log(\exp x) = x \implies \log\left(1 + x + \frac{x^2}{2} + \ldots\right) = x$$
$$\implies \log(1+y) = y - \frac{y^2}{2} + \ldots$$

To be more precise, $\log(1+y) \simeq y$ is clear, and with a bit more work, that we will leave here as an instructive exercise, you can recover $\log(1+y) = y - y^2/2$ too. Of course, the higher terms can be recovered too, with enough work involved, at each step.

## 4d. Hyperbolic functions

Time now for some heavier stuff, which will bring us, hang on, into Einstein and relativity, no less than that. We have the following result, to start with:

THEOREM 4.21. *The following functions, called hyperbolic sine and cosine,*

$$\sinh x = \frac{e^x - e^{-x}}{2} \quad , \quad \cosh x = \frac{e^x + e^{-x}}{2}$$

*are subject to the following formulae:*

(1) $e^x = \cosh x + \sinh x$.
(2) $\sinh(ix) = i \sin x$, $\cosh(ix) = \cos x$, *for* $x \in \mathbb{R}$.
(3) $\sinh(x+y) = \sinh x \cosh y + \cosh x \sinh y$.
(4) $\cosh(x+y) = \cosh x \cosh y + \sinh x \sinh y$.
(5) $\sinh x = \sum_l \frac{x^{2l+1}}{(2l+1)!}$, $\cosh x = \sum_l \frac{x^{2l}}{(2l)!}$.

PROOF. The formula (1) follows from definitions. As for (2), this follows from:

$$\sinh(ix)=\frac{e^{ix}-e^{-ix}}{2}=\frac{\cos x+i\sin x}{2}-\frac{\cos x-i\sin x}{2}=i\sin x$$

$$\cosh(ix)=\frac{e^{ix}+e^{-ix}}{2}=\frac{\cos x+i\sin x}{2}+\frac{\cos x-i\sin x}{2}=\cos x$$

Regarding now (3,4), observe first that the formula $e^{x+y}=e^x+e^y$ reads:

$$\cosh(x+y)+\sinh(x+y)=(\cosh x+\sinh x)(\cosh y+\sinh y)$$

Thus, we have some good explanation for (3,4), and in practice, these formulae can be checked by direct computation, as follows:

$$\frac{e^{x+y}-e^{-x-y}}{2}=\frac{e^x-e^{-x}}{2}\cdot\frac{e^y+e^{-y}}{2}+\frac{e^x+e^{-x}}{2}\cdot\frac{e^y-e^{-y}}{2}$$

$$\frac{e^{x+y}+e^{-x-y}}{2}=\frac{e^x+e^{-x}}{2}\cdot\frac{e^y+e^{-y}}{2}+\frac{e^x-e^{-x}}{2}\cdot\frac{e^y-e^{-y}}{2}$$

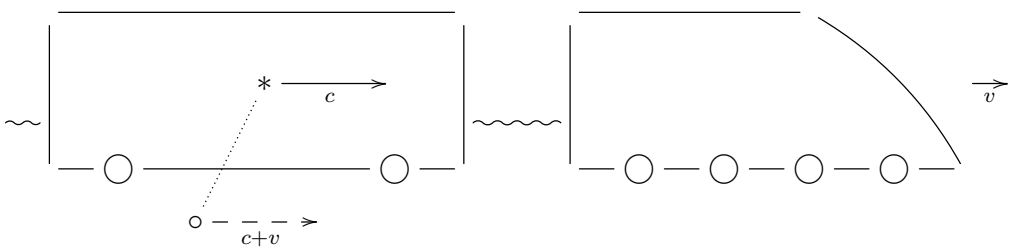
Finally, (5) is clear from the definition of sinh, cosh, and from $e^x=\sum_k\frac{x^k}{k!}$. □

Ready for some physics? Based on experiments by Fizeau, then Michelson-Morley and others, and some physics by Maxwell and Lorentz too, Einstein came upon:

FACT 4.22 (Einstein principles). *The following happen:*

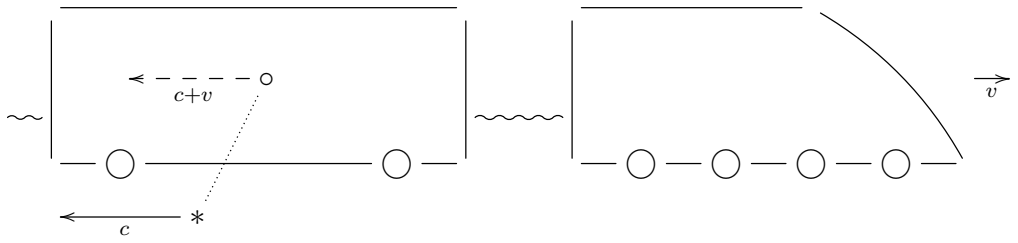
(1) *Light travels in vacuum at a finite speed, $c<\infty$.*
(2) *This speed $c$ is the same for all inertial observers.*
(3) *In non-vacuum, the light speed is lower, $v<c$.*
(4) *Nothing can travel faster than light, $v\not>c$.*

The point now is that, obviously, something is wrong here. Indeed, assuming for instance that we have a train, running in vacuum at speed $v>0$, and someone on board lights a flashlight $*$ towards the locomotive, then an observer $\circ$ on the ground will see the light travelling at speed $c+v>c$, which is a contradiction:




Equivalently, with the same train running, in vacuum at speed $v>0$, if the observer on the ground lights a flashlight $*$ towards the back of the train, then viewed from the

train, that light will travel at speed $c+v>c$, which is a contradiction again:




Summarizing, Fact 4.22 implies $c+v=c$, so contradicts classical mechanics, which therefore needs a fix. By dividing all speeds by $c$, as to have $c=1$, and by restricting the attention to the 1D case, to start with, we are led to the following puzzle:

PUZZLE 4.23. *How to define speed addition on the space of* 1D *speeds, which is*

$$I=[-1,1]$$

*with our* $c=1$ *convention, as to have* $1+c=1$*, as required by physics?*

In view of our geometric knowledge so far, a natural idea here would be that of wrapping $[-1,1]$ into a circle, and then stereographically projecting on $\mathbb{R}$. Indeed, we can then "import" to $[-,1,1]$ the usual addition on $\mathbb{R}$, via the inverse of this map.

So, let us see where all this leads us. First, the formula of our map is as follows:

PROPOSITION 4.24. *The map wrapping* $[-1,1]$ *into the unit circle, and then stereographically projecting on* $\mathbb{R}$ *is given by the formula*

$$\varphi(u)=\tan\left(\frac{\pi u}{2}\right)$$

*with the convention that our wrapping is the most straightforward one, making correspond* $\pm1\to i$*, with negatives on the left, and positives on the right.*

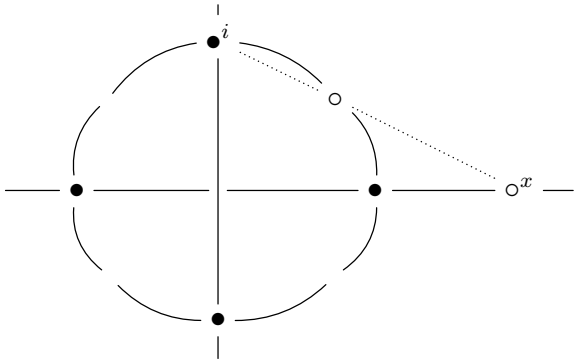
PROOF. Regarding the wrapping, as indicated, this is given by:

$$u\to e^{it}\quad,\quad t=\pi u-\frac{\pi}{2}$$

Indeed, this correspondence wraps $[-1,1]$ as above, the basic instances of our correspondence being as follows, and with everything being fine modulo $2\pi$:

$$-1\to\frac{\pi}{2}\quad,\quad -\frac{1}{2}\to-\pi\quad,\quad 0\to-\frac{\pi}{2}\quad,\quad \frac{1}{2}\to 0\quad,\quad 1\to\frac{\pi}{2}$$

Regarding now the stereographic projection, the picture here is as follows:




Thus, by Thales, the formula of the stereographic projection is as follows:

$$\frac{\cos t}{x} = \frac{1-\sin t}{1} \implies x = \frac{\cos t}{1-\sin t}$$

Now if we compose our wrapping operation above with the stereographic projection, what we get is, via the above Thales formula, and some trigonometry:

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

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

The above result is very nice, but when it comes to physics, things do not work, for instance because of the wrong slope of the function $\varphi(u) = \tan\left(\frac{\pi u}{2}\right)$ at the origin, which makes our summing on $[-1,1]$ not compatible with the Galileo addition, at low speeds.

So, what to do? Obviously, trash Proposition 4.24, and start all over again. Getting back now to Puzzle 4.23, this has in fact a simpler solution, based this time on algebra, and which in addition is the good, physically correct solution, as follows:

THEOREM 4.25. *If we sum the speeds according to the Einstein formula*

$$u +_e v = \frac{u + v}{1 + uv}$$

*then the Galileo formula still holds, approximately, for low speeds*

$$u +_e v \simeq u + v$$

*and if we have* $u = 1$ *or* $v = 1$*, the resulting sum is* $u +_e v = 1$*.*

PROOF. All this is self-explanatory, and clear from definitions, and with the Einstein formula of $u +_e v$ itself being just an obvious solution to Puzzle 4.23, provided that, importantly, we know 0 geometry, and rely on very basic algebra only. □

So, very nice, problem solved, at least in 1D. But, shall we give up with geometry, and the stereographic projection? Certainly not, let us try to recycle that material. In order to do this, let us recall that the usual trigonometric functions are given by:

$$\sin x = \frac{e^{ix} - e^{-ix}}{2i} \quad , \quad \cos x = \frac{e^{ix} + e^{-ix}}{2} \quad , \quad \tan x = \frac{e^{ix} - e^{-ix}}{i(e^{ix} + e^{-ix})}$$

The point now is that, and you might know this from calculus, the above functions have some natural "hyperbolic" or "imaginary" analogues, constructed as follows:

$$\sinh x = \frac{e^{x} - e^{-x}}{2} \quad , \quad \cosh x = \frac{e^{x} + e^{-x}}{2} \quad , \quad \tanh x = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

But the function on the right, tanh, starts reminding the formula of Einstein addition, from Theorem 4.25. So, we have our idea, and we are led to the following result:

THEOREM 4.26. *The Einstein speed summation in 1D is given by*

$$\tanh x +_e \tanh y = \tanh(x + y)$$

*with* $\tanh : [-\infty, \infty] \to [-1, 1]$ *being the hyperbolic tangent function.*

PROOF. This follows by putting together our various formulae above, but it is perhaps better, for clarity, to prove this directly. Our claim is that we have:

$$\tanh(x + y) = \frac{\tanh x + \tanh y}{1 + \tanh x \tanh y}$$

But this can be checked via direct computation, from the definitions, as follows:

$$\begin{aligned}
&\frac{\tanh x+\tanh y}{1+\tanh x\tanh y}\\
=\ &\left(\frac{e^x-e^{-x}}{e^x+e^{-x}}+\frac{e^y-e^{-y}}{e^y+e^{-y}}\right)\Big/\left(1+\frac{e^x-e^{-x}}{e^x+e^{-x}}\cdot\frac{e^y-e^{-y}}{e^y+e^{-y}}\right)\\
=\ &\frac{(e^x-e^{-x})(e^y+e^{-y})+(e^x+e^{-x})(e^y-e^{-y})}{(e^x+e^{-x})(e^y+e^{-y})+(e^x-e^{-x})(e^y+e^{-y})}\\
=\ &\frac{2(e^{x+y}-e^{-x-y})}{2(e^{x+y}+e^{-x-y})}\\
=\ &\tanh(x+y)
\end{aligned}$$

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

Very nice all this, hope you agree. As a conclusion, passing from the Riemann stereographic projection sum to the Einstein summation basically amounts in replacing:

$$\tan\to\tanh$$

Let us formulate as well this finding more philosophically, as follows:

Conclusion 4.27. *The Einstein speed summation in 1D is the imaginary analogue of the summation on* $[-1,1]$ *obtained via Riemann's stereographic projection.*

Which looks quite deep, and we will stop here. More on this later in this book, when discussing curved spacetime, in full generality, and with more advanced tools.

## 4e. Exercises

Exercises:

Exercise 4.28.

Exercise 4.29.

Exercise 4.30.

Exercise 4.31.

Exercise 4.32.

Exercise 4.33.

Exercise 4.34.

Exercise 4.35.

Bonus exercise.

# Part II

# Continuity

*Here we are*
*To celebrate a party*
*In this hot summer night*
*While the moon is shining bright*

CHAPTER 5

# Continuous functions

## 5a. Continuity

Welcome to advanced function theory. In order to say a number of non-trivial things, we will focus now our study on the functions $f : \mathbb{R} \to \mathbb{R}$ which are suitably regular, with the hope that these regularity properties will lead to some interesting theorems.

In what regards these regularity properties, the most basic of them, which is something quite intuitive, that we are somehow already familiar with, is continuity:

DEFINITION 5.1. *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 be mostly interested in what follows in the functions which are continuous everywhere, but continuity at a given point is something useful too.

Regarding now the basic examples of countinous functions, there are many of them, and we will discuss them in a moment, once we will have some basic tools, in order to prove that this or that function is continuous or not, without much pain. As a matter, however, of having a first illustration for Definition 5.1, let us record here:

THEOREM 5.2. *The basic power functions, namely*

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

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

PROOF. According to Definition 5.1, we want to prove that we have:

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

Which looks quite clear, but go now prove this, with full rigor, that does not look totally obvious. So, non-trivial question, and here are two possible solutions:

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
|(x+t)^k - x^k| &= \left|\sum_{s=0}^{k} \binom{k}{s} x^{k-s}t^s - x^k\right| \\
&= \left|\sum_{s=1}^{k} \binom{k}{s} x^{k-s}t^s\right| \\
&\leq \sum_{s=1}^{k} \binom{k}{s} |x|^{k-s}|t|^s \\
&\leq |t| \sum_{s=1}^{k} \binom{k}{s} |x|^{k-s} \\
&\leq |t| \sum_{s=0}^{k} \binom{k}{s} |x|^{k-s} \\
&= |t|(1+|x|)^k
\end{aligned}$$

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

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

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

Again as an illustration for Definition 5.1, let us record as well a counterexample:

THEOREM 5.3. *The basic inverse function, namely*

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

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

$$f(0) = \alpha$$

*the function will be not continuous at* 0*.*

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

(1) Let us begin with the positive statement, saying that $f$ is continuous at any $x \neq 0$. In order to prove this, the situation is a bit similar to what we had in Theorem 5.2. Indeed, we can use the following formula for sequences, that we know from chapter 1:

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

Alternatively, and once again a bit similarly to what we did for Theorem 5.2, we have the 1-neuron proof, based on the following estimate, valid for $|t| < |x|/2$:

$$\begin{aligned}
\left|\frac{1}{x} - \frac{1}{x+t}\right| &= \left|\frac{t}{x(x+t)}\right| \\
&= \frac{|t|}{|x| \cdot |x+t|} \\
&< \frac{|t|}{|x|(|x| - |t|)} \\
&< \frac{|t|}{2|x|^2}
\end{aligned}$$

Thus, either way, we conclude that $f$ is continuous at any $x \neq 0$.

(2) Regarding now the second assertion, non-continuity at 0, this is something more tricky, requiring some thinking. You would say, how to prove such things, without even knowing what $f(0) = \alpha$ is. But there is in fact no problem here, because the Devil invented this method called "proof by contradiction", that us mahematicians are allowed to use. So, let us assume that, by setting $f(0) = \alpha$, our function $f$ becomes continuous at 0. In view of Definition 5.1, this means that the following must happen:

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

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

We will see in a moment that many other familiar functions are continuous, save perhaps at some special points, a bit as in Theorem 5.2, and Theorem 5.3.

Getting back now to the general theory, and to Definition 5.1 as stated, many things can be said. Indeed, there are many other equivalent formulations of the notion of continuity, with a well-known, and much feared one, being as follows:

THEOREM 5.4. *A function $f : X \to \mathbb{R}$ is continuous when*

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

*holds.*

PROOF. Let us prove this, with no fear. According to Definition 5.1, in order for our function $f$ to be continuous, the following must happen, for any $x \in X$:

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

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

As already mentioned, at the level of examples, basically all the functions that we 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 a useful theoretical result regarding continuity, as follows:

THEOREM 5.5. *If $f, g$ are continuous, then so are:*

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

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

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

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

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

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

(3) Here the claim is that if both $f, g$ are continuous at $x$, with $g(x) \neq 0$, then $f/g$ is continuous at $x$. In order to prove this, observe that by continuity, $g(x) \neq 0$ shows that

$g(y) \neq 0$ for $|x-y|$ small enough. Thus we can assume $g \neq 0$, and with this assumption made, the result follows from the similar result for sequences, namely:

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

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

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

Alternatively, let us prove this as well by using that scary $\varepsilon, \delta$ condition from Theorem 5.4. 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 $\varepsilon, \delta$ 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 $\varepsilon, \delta$ 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 $\varepsilon, \delta$ condition. Good question, and in answer:

(1) It is usually said that "for doing advanced math, you must use the $\varepsilon, \delta$ 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 $\varepsilon, \delta$ condition. So, you cannot really run away from this $\varepsilon, \delta$ 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 $\varepsilon, \delta$ 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 $\varepsilon, \delta$ terms.

## 5b. Basic examples

Back to work now, let us have a closer look at the various functions that we know, from the perspective of continuity. At the level of the very basic examples, we have:

THEOREM 5.6. *The following functions are continuous:*

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

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

(1) Since $f(x) = x$ is continuous, by using Theorem 5.5 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 5.5.

(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:

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

Indeed, 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:

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

(4) 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:

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

(5) 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:

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

(6) The fact that $\tan x$ is continuous is clear from the fact that $\sin x$, $\cos x$ are continuous, by using the result regarding quotients from Theorem 5.5. Equivalently, we can deduce this directly, from the following formula for the tangents of sums:

$$\tan(x+y)=\frac{\tan x+\tan y}{1-\tan x\tan y}$$

Indeed, this formula holds indeed, by dividing the formulae for sines and cosines, established above. And with this in hand, the continuity check goes as follows:

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

(7) In what regards now the exponential, the continuity here is something very simple, coming from the following elementary computation:

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

(8) Finally, in what regards the logarithm, recall that for $t$ small we have:

$$e^t \simeq 1 + t$$

By applying the logarithm, we conclude that for $t$ small we have:

$$t \simeq \log(1 + t)$$

But this gives the continuity of log at $x = 1$, because with $t \to 0$ we have:

$$\begin{aligned} |\log(1 + t) - \log 1| &= |\log(1 + t)| \\ &\simeq |t| \\ &\to 0 \end{aligned}$$

(9) In general now, at an arbitrary $x > 0$, we can use a similar argument. Indeed, with $t \to 0$ we have the following estimate:

$$\begin{aligned} |\log(x + t) - \log x| &= \left|\log\left(\frac{x + t}{x}\right)\right| \\ &= \left|\log\left(1 + \frac{t}{x}\right)\right| \\ &\simeq \left|\frac{t}{x}\right| \\ &\to 0 \end{aligned}$$

Thus log is continuous at our arbitrary point $x > 0$, as desired. □

At the level of more specialized examples, we have as well:

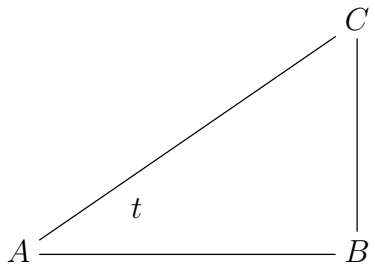
PROPOSITION 5.7. *The following functions are continuous too:*

(1) $\sec x$.
(2) $\csc x$.
(3) $\cot x$.

PROOF. Bad news, regarding these extra trigonometric functions $\sec x$, $\csc x$, $\cot x$, I must admit that I actually forgot their definition. So, allowing me a short break, for a quick Wikipedia read, and feel free in the meantime to consult your Tiktok notifications, or deal with whatever other urgent internet matters, here are the formulae:

$$\sec x = \frac{1}{\sin x} \quad , \quad \csc x = \frac{1}{\cos x} \quad , \quad \cot x = \frac{1}{\tan x}$$

Which might sound quite silly, indeed, but there is in fact some logic behind these notions. Consider indeed a right triangle, as we love them in mathematics, namely:




We have then 6 possible quotients of sides to be considered, and the point is that, in terms of the angle $t$ at the vertex $A$, these are given by the following formulae:

$$\sin t = \frac{BC}{AC} \quad , \quad \cos t = \frac{AB}{AC} \quad , \quad \tan t = \frac{BC}{AB}$$

$$\sec t = \frac{AC}{BC} \quad , \quad \csc t = \frac{AC}{AB} \quad , \quad \cot t = \frac{AB}{BC}$$

Thus, it makes sense, for symmetry reasons, to enlarge the family $\sin x$, $\cos x$, $\tan x$ with the functions $\sec x$, $\csc x$, $\cot x$. Now by getting back to what we wanted to prove, continuity, this is clear from what we have in Theorem 5.6, by using Theorem 5.5. □

As a continuation of the above, still in relation with the basic functions, and more specifically with the exponentials and logarithms, remember from chapter 4 that we had some difficulties in order to have a fully rigourous theory of $\exp x$ and $\log x$ running. But now, with our notion of continuity, we can say more about all this.

To be more precise, let us start with something quite intuitive, as follows:

PRINCIPLE 5.8 (Intermediate value property). *The following happen:*

(1) *If $f : [a, b] \to \mathbb{R}$ is continuous, then $f([a, b])$ is an interval.*
(2) *If $f$ is continuous and invertible, then $f^{-1}$ is continuous.*

Observe that (1) is indeed something very intuitive, basically telling us that $f$ cannot jump, and with this being of course prevented by its continuity property, which is there for preventing such things. So, we should definitely trust this, and in what regards a formal proof, well, this is something a bit technical, that can be done too, and we will discuss this with full details in chapter 6 below, no hurry with that.

As for the assertion (2), this is again something quite intuitive, the point being that if a function $f$ is continuous and invertible, then it must be obviously monotone, and it follows that $f^{-1}$ must be monotone too, and continuous. So, we should trust this too, no question about it, and in what regards the formal proof, the idea is that this comes as an application of (1), and we will discuss this also in chapter 6 below.

Long story short, we have the above Principle 5.8, which is something very intuitive, and that we can trust in, and whose formal proof exists, and is to come soon. Now the point is that, with this in hand, we can come back to $\exp x$ and $\log x$, and we have:

THEOREM 5.9. *The exponential function, defined as usual by*

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

*is indeed given by* $\exp x = e^x$*, with* $e = \exp 1$*. This function is increasing, and continuous. Its inverse, the logarithm, can be defined by one of the following equivalent formulae,*

$$\exp(\log x) = x \quad , \quad \log(\exp x) = x$$

*and is again increasing, and continuous.*

PROOF. These are all things that we basically know, from chapter 4 and from here, but via proofs which, while certainly correct, lack a bit of rigor. Now the point is that, with Principle 5.8 in hand, we can make all this fully rigorous. We will leave this for now as an instructive exercise, and we will come back to it in chapter 6 below, after proving Principle 5.8, as a matter of having all things 100% rigorously done, at that time. □

Still with me I hope, mathematics students I mean, and this because in what regards physicists, engineers and other scientists, dear, do I know well, after a lifetime of teaching, and of research too, that anything messy mathematics always brings you joy.

This being said, perhaps time to ask the cat, about this. And cat, Vladimir by his name, who quite often has a look at what I'm doing, and does not hesitate to warn me, whenever my computations go wrong due to lack of matematical rigor, declares:

CAT 5.10. *Mathematics is the part of physics where experiments are cheap.*

Thanks cat, so we have here confirmation, we are learning good science in this book, in the good old scientific way, which has always worked well, giving us mankind countless interesting results, and this no matter what some mathematicians might say about it. So, Theorem 5.9 proved, and more about it, as promised, in chapter 6 below.

Along the same lines now, with exp and log understood, what about looking at other familiar inverse functions. And here, we have the following result:

THEOREM 5.11. *The following functions are well-defined and continuous,*

(1) $\arcsin x$*.*
(2) $\arccos x$*,*
(3) $\arctan x$*,*

*and the same goes for the functions* $\operatorname{arcsec} x$*,* $\operatorname{arccsc} x$*,* $\operatorname{arccot} x$*.*

PROOF. The situation here is quite similar to that for the logarithm, with the existence and continuity of the functions in the statement coming from what we have in Theorem 5.6 and Proposition 5.7, via Principle 5.8. So, we will leave this as an exercise for you, and we will come back to it, with details, in chapter 6 below, after Principle 5.8 proved. □

We will be back to more examples of continuous functions later, and in particular to the functions of type $x^a$ and $a^x$ with $a \in \mathbb{R}$, which are more tricky to define.

## 5c. Discontinuities, jumps

There are many functions which are not continuous everywhere, such as $f(x) = 1/x$ at $x = 0$. And the question is, what to do with them? That is, can we have some mathematical theory going on for them as well, inspired by what we did in the above?

In answer, here is something that we can do, in general:

DEFINITION 5.12. *Given a function $f : X \to \mathbb{R}$ and $x \in X$, we set*

$$f(x_-) = \lim_{y \nearrow x} f(y) \quad , \quad f(x_+) = \lim_{y \searrow x} f(y)$$

*provided that these two limits exist indeed, and we call the quantity*

$$J_f(x) = f(x_+) - f(x_-)$$

*which does not depend on $f(x)$, the jump of $f$ at the given point $x \in X$.*

As a first observation, assuming that a function $f : X \to \mathbb{R}$ is continuous at $x \in X$, its jump there is zero, so that we have the following implications:

$$f \text{ continuous at } x \implies J_f(x) = 0$$

$$f \text{ continuous} \implies J_f(x) = 0, \ \forall x \in X$$

Observe also that the converses of these implications do not necessarily hold, and this because the jump $J_f(x)$, as constructed above, does not depend on $f(x)$, so we can easily construct counterexamples, just by modifying the value $f(x)$. More on this later.

But are we here for talking about continuous functions. In the non-continuous case, which is the one that we are interested in, here are some basic computations:

THEOREM 5.13. *For the basic step function, given by*

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \end{cases}$$

*we have $J_f(0) = 1$, as we should. Also, for the inverse function*

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

*we have $J_g(0) = \infty$, again as we should.*

PROOF. Both formulae are clear from definitions. Indeed, we have:

$$\begin{aligned} J_f(0) &= f(0_+) - f(0_-) \\ &= \lim_{y\searrow 0} f(y) - \lim_{y\nearrow 0} f(y) \\ &= \lim_{y\searrow 0} 1 - \lim_{y\nearrow 0} 0 \\ &= 1 - 0 \\ &= 1 \end{aligned}$$

As for the second formula, the computation here is similar, as follows:

$$\begin{aligned} J_g(0) &= g(0_+) - g(0_-) \\ &= \lim_{y\searrow 0} g(y) - \lim_{y\nearrow 0} g(y) \\ &= \lim_{y\searrow 0} \frac{1}{y} - \lim_{y\nearrow 0} \frac{1}{y} \\ &= \infty - (-\infty) \\ &= \infty \end{aligned}$$

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

There are of course many other interesting examples of discountinuous functions, quite often coming from physics, and more on them on several occasions, in what follows.

Going ahead now with more theory, we can complement the basic notions introduced in Definition 5.12 with some more notions, which are equally useful. First we have:

DEFINITION 5.14. *Given a function $f : X \to \mathbb{R}$ and $x \in X$, we say that:*

(1) *$f$ is left continuous at $x$, if $f(x_-) = f(x)$.*
(2) *$f$ is right continuous at $x$, if $f(x_+) = f(x)$.*

As before with Definition 5.12, this is something quite self-explanatory. We will see some examples and computations, in relation with this, in a moment.

As a first observation, a function $f : X \to \mathbb{R}$ is continuous at $x$ precisely when it is left and right continuous there, so that we have the following equivalences:

$$f \text{ continuous at } x \iff f(x_-) = f(x) = f(x_+)$$

$$f \text{ continuous} \iff f(x_-) = f(x) = f(x_+),\ \forall x \in X$$

Which sounds quite interesting, in relation with the issues with the jump, discussed after Definition 5.12. So, let us update as well the definition of the jump, as follows:

THEOREM 5.15. *Given a function $f : X \to \mathbb{R}$ and $x \in X$, we call the quantities*

$$J_f(x_-) = f(x) - f(x_-) \quad , \quad J_f(x_+) = f(x_+) - f(x)$$

*the left and right jumps at $x$, so that the total jump there is given by:*

$$J_f(x) = J_f(x_+) + J_f(x_-)$$

*The function $f$ is then continuous at $x$ when both its jumps there vanish,*

$$f \text{ continuous at } x \iff J(x_-) = J(x_+) = 0$$

*and globally continuous, when we have $J(x_-) = J(x_+) = 0$, for any $x \in X$.*

PROOF. This is something quite self-explanatory, with a lot of talking, basically definitions, and with the jump formula being something obvious, as follows:

$$\begin{aligned} J_f(x) &= f(x_+) - f(x_-) \\ &= (f(x_+) - f(x)) + (f(x) - f(x_-)) \\ &= J_f(x_+) + J_f(x_-) \end{aligned}$$

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

As an illustration now, for the same functions as in Theorem 5.13, we have:

THEOREM 5.16. *For the basic step function, given by*

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

*we have $J_f(0_-) = 1$ and $J_f(0_+) = 0$, as we should. Also, for the inverse function*

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

*with the convention $g(0) = \alpha \in \mathbb{R}$, we have $J_g(0-) = J_g(0_+) = \infty$, again as we should.*

PROOF. This is quite obvious, but let us do all four computations:

(1) For the left jump of the step function, we have:

$$\begin{aligned} J_f(0_-) &= f(0) - f(0_-) \\ &= 1 - \lim_{y \nearrow 0} f(y) \\ &= 1 - \lim_{y \nearrow 0} 0 \\ &= 1 - 0 \\ &= 1 \end{aligned}$$

(2) For the right jump of the step function, we have:

$$\begin{aligned} J_f(0_+) &= f(0_+) - f(0) \\ &= \lim_{y\searrow 0} f(y) - 1 \\ &= \lim_{y\searrow 0} 1 - 1 \\ &= 1 - 1 \\ &= 0 \end{aligned}$$

(3) For the left jump of the inverse function, we have:

$$\begin{aligned} J_g(0_-) &= g(0) - g(0_-) \\ &= \alpha - \lim_{y\nearrow 0} g(y) \\ &= \alpha - \lim_{y\nearrow 0} \frac{1}{y} \\ &= \alpha - (-\infty) \\ &= \infty \end{aligned}$$

(4) For the right jump of the inverse function, we have:

$$\begin{aligned} J_g(0_+) &= g(0_+) - g(0) \\ &= \lim_{y\searrow 0} g(y) - \alpha \\ &= \lim_{y\searrow 0} \frac{1}{y} - \alpha \\ &= \infty - \alpha \\ &= \infty \end{aligned}$$

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

All the above was of course quite trivial, but we will be back to this later, once we will have some interesting discontinuous functions, with jumps waiting to be computed.

As a last topic of discussion, let us go back to our original notion of jump, from Definition 5.12. We already know from the discussion there that we have:

$$f \text{ continuous at } x \implies J_f(x) = 0$$

We also know from the discussion above that the converse of this implication obviously does not hold, and this because the jump $J_f(x)$ does not depend on $f(x)$, so we can easily construct counterexamples, just by modifying the value $f(x)$.

However, we can say something interesting here, as follows:

THEOREM 5.17. *Assuming that a function $f : X \to \mathbb{R}$ does not jump at $x \in X$,*

$$J_f(x) = 0$$

*we can modify our function by forgetting the old value $f(x)$, and setting*

$$f(x) = f(x_-) = f(x_+)$$

*and we obtain in this way a function which is continuous at $x$.*

PROOF. This is something which is clear from Theorem 5.15, because our modified function $f$ has both its left and right jumps vanishing at $x$:

$$J_f(x_-) = J_f(x_+) = 0$$

Thus Theorem 5.15 applies, and tells us that our modified $f$ is continuous at $x$. □

The above result is quite interesting, and obviously can be applied to the various points where $f$ is discontinous, provided that these points are "isolated" from each other. In the case where $f$ needs as "fix" on a more substantial set of points, such as a whole interval $(a, b) \subset \mathbb{R}$, things are more complicated, because the solution will obviously be not unique, and so that we have to choose one, depending on what exactly we want our function $f$ to look like, and perform. More on such questions, later in this book.

## 5d. Uniform continuity

Back now to plain continuity, and going ahead with some more theory, some functions are "obviously" continuous, with a basic result here being as follows:

THEOREM 5.18. *If a function $f : X \to \mathbb{R}$ has the Lipschitz property*

$$|f(x) - f(y)| \leq K|x - y|$$

*for some $K > 0$, then it is continuous.*

PROOF. This is indeed clear from our definition of continuity. □

There are many interesting examples of Lipschitz functions, both concrete and abstract. In what regards the usual functions, the situation is as follows:

THEOREM 5.19. *The following functions are Lipschitz, on suitable domains:*

(1) $x^n$, *with* $n \in \mathbb{N}$.
(2) $x^{-n}$, *with* $n \in \mathbb{N}$.
(3) $P(x)/Q(x)$, *with* $P, Q \in \mathbb{R}[X]$.
(4) $\sin x$, $\cos x$, $\tan x$.
(5) $\exp x$, $\log x$.

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

(1) The function $f(x) = x$ is certainly Lipschitz everywhere, with constant $K = 1$. In what regards now $f(x) = x^2$, we have here the following equivalence:

$$|x^2 - y^2| \leq K|x - y| \iff |x + y| \leq K$$

Thus $f(x) = x^2$ is Lipschitz on any interval $[a, b]$, with constant as follows:

$$K = max(|a|, |b|)$$

As for the study of the function $f(x) = x^n$ with $n \in \mathbb{N}$, in general, this is similar, and we will leave the computations here as an instructive exercise.

(2) In what regards the function $f(x) = x^{-1}$, we have here the following equivalence:

$$\begin{aligned}\left|\frac{1}{x} - \frac{1}{y}\right| \leq K|x - y| &\iff \frac{1}{|xy|} \leq K\\ &\iff |xy| \geq \frac{1}{K}\end{aligned}$$

Thus $f(x) = x^{-1}$ is Lipschitz outside $[-\varepsilon, \varepsilon]$, with constant as follows:

$$K = \frac{1}{\varepsilon^2}$$

As for the study of the function $f(x) = x^{-n}$ with $n \in \mathbb{N}$, in general, this is similar.

(3) The functions of type $f(x) = P(x)/Q(x)$, with $P, Q \in \mathbb{R}[X]$, appear as a joint generalization of what we have in (1) and (2), and the same methods as there apply.

(4) Again, the study here is standard, and we will leave this as an exercise.

(5) Similar situation here, standard study, that we will leave as an exercise. □

We will be back to this, Lipschitz functions, on several occassions, in what follows, and notably in Part III below, when talking derivatives, which can help with this.

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

DEFINITION 5.20. *A function $f : X \to \mathbb{R}$ is called uniformly continuous when:*

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

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

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

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

Indeed, it is clear by looking at the graph of $f$ that, the further our point $x \in \mathbb{R}$ is from 0, the smaller our $\delta > 0$ must be, compared to $\varepsilon > 0$, in our $\varepsilon, \delta$ definition of continuity.

Thus, given an $\varepsilon > 0$, we have no $\delta > 0$ doing the $|x - y| < \delta \implies |f(x) - f(y)| < \varepsilon$ job at any $x \in \mathbb{R}$, and so our function is indeed not uniformly continuous.

As before with the Lipschitz functions, all this seems to have something to do with the slope of the graph of $f$, computed at various points. We will be back to this later, in Part III below, when talking slopes, or derivatives, which can help with this.

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

THEOREM 5.21. *Any continuous function defined on a closed, bounded interval*

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

*is automatically uniformly continuous.*

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

(1) Given $\varepsilon > 0$, for any $x \in [a, b]$ we know that we have a $\delta_x > 0$ such that:

$$|x - y| < \delta_x \implies |f(x) - f(y)| < \frac{\varepsilon}{2}$$

So, consider the following open intervals, centered at the various points $x \in [a, b]$:

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

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

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

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

(2) So, let us prove first that the cover in (1) has a finite subcover. For this purpose, we proceed by contradiction. So, assume that $[a, b]$ has no finite subcover, and let us cut this interval in half. Then one of the halves must have no finite subcover either, and we can repeat the procedure, by cutting this smaller interval in half. And so on.

(3) But this leads to a contradiction, because the limiting point $x \in [a, b]$ that we obtain in this way, as the intersection of these smaller and smaller intervals, must be covered by something, and so one of these small intervals leading to it must be covered too, contradiction. Thus, we have proved that the cover in (1) has a finite subcover.

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

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

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

We will be back to such things, which are quite subtle, in the next chapter, with more details, in what regards the above proof, and with some applications too.

We would like to end this introductory chapter on continuity with something nice, very useful and refreshing, namely a basic application of continuity, to the foundations of linear algebra. Let us start with something basic and intuitive, namely:

DEFINITION 5.22. *A map $f : \mathbb{R}^2 \to \mathbb{R}^2$ is called affine when it maps lines to lines,*

$$f(tx + (1-t)y) = tf(x) + (1-t)f(y)$$

*for any $x, y \in \mathbb{R}^2$ and any $t \in \mathbb{R}$. If in addition $f(0) = 0$, we call $f$ linear.*

As a first observation, our "maps lines to lines" interpretation of the equation in the statement assumes that the points are degenerate lines, and this in order for our interpretation to work when $x = y$, or when $f(x) = f(y)$. Also, what we call line is not exactly a set, but rather a dynamic object, think trajectory of a point on that line. We will be back to this later, once we will know more about such maps.

Here are some basic examples of symmetries, all being linear in the above sense:

PROPOSITION 5.23. *The symmetries with respect to $Ox$ and $Oy$ are:*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x \\ -y \end{pmatrix} \quad , \quad \begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} -x \\ y \end{pmatrix}$$

*The symmetries with respect to the $x = y$ and $x = -y$ diagonals are:*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} y \\ x \end{pmatrix} \quad , \quad \begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} -y \\ -x \end{pmatrix}$$

*All these maps are linear, in the above sense.*

PROOF. The fact that all these maps are linear is clear, because they map lines to lines, in our sense, and they also map 0 to 0. As for the explicit formulae in the statement, these are clear as well, by drawing pictures for each of the maps involved. □

Here are now some basic examples of rotations, once again all being linear:

PROPOSITION 5.24. *The rotations of angle* $0^\circ$ *and of angle* $90^\circ$ *are:*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x \\ y \end{pmatrix} \quad , \quad \begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} -y \\ x \end{pmatrix}$$

*The rotations of angle* $180^\circ$ *and of angle* $270^\circ$ *are:*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} -x \\ -y \end{pmatrix} \quad , \quad \begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} y \\ -x \end{pmatrix}$$

*All these maps are linear, in the above sense.*

PROOF. As before, these rotations are all linear, for obvious reasons. As for the formulae in the statement, these are clear as well, by drawing pictures. □

Here are some basic examples of projections, once again all being linear:

PROPOSITION 5.25. *The projections on* $Ox$ *and* $Oy$ *are:*

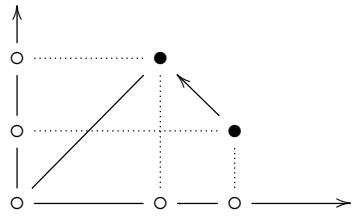
$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x \\ 0 \end{pmatrix} \quad , \quad \begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} 0 \\ y \end{pmatrix}$$

*The projections on the* $x = y$ *and* $x = -y$ *diagonals are:*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \frac{1}{2}\begin{pmatrix} x+y \\ x+y \end{pmatrix} \quad , \quad \begin{pmatrix} x \\ y \end{pmatrix} \to \frac{1}{2}\begin{pmatrix} x-y \\ y-x \end{pmatrix}$$

*All these maps are linear, in the above sense.*

PROOF. Again, these projections are all linear, and the formulae are clear as well, by drawing pictures, with only the last 2 formulae needing some explanations. In what regards the projection on the $x = y$ diagonal, the picture here is as follows:



But this gives the result, since the $45^\circ$ triangle shows that this projection leaves invariant $x + y$, so we can only end up with the average $(x + y)/2$, as double coordinate. As for the projection on the $x = -y$ diagonal, the proof here is similar. □

Finally, we have the translations, which are as follows:

PROPOSITION 5.26. *The translations are exactly the maps of the form*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x+p \\ y+q \end{pmatrix}$$

*with* $p, q \in \mathbb{R}$*, and these maps are all affine, in the above sense.*

PROOF. A translation $f : \mathbb{R}^2 \to \mathbb{R}^2$ is clearly affine, because it maps lines to lines. Also, such a translation is uniquely determined by the following vector:

$$f\begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}p\\q\end{pmatrix}$$

To be more precise, $f$ must be the map which takes a vector $\binom{x}{y}$, and adds this vector $\binom{p}{q}$ to it. But this gives the formula in the statement. □

Summarizing, we have many interesting examples of linear and affine maps. Let us develop now some general theory, for such maps. As a first result, we have:

THEOREM 5.27. *For a map $f : \mathbb{R}^2 \to \mathbb{R}^2$, the following are equivalent:*

(1) *$f$ is linear in our sense, mapping lines to lines, and $0$ to $0$.*
(2) *$f$ maps sums to sums, $f(x+y) = f(x) + f(y)$, and satisfies $f(\lambda x) = \lambda f(x)$.*

PROOF. This is something which comes from definitions, as follows:

(1) $\implies$ (2) We know that $f$ satisfies the following equation, and $f(0) = 0$:

$$f(tx + (1-t)y) = tf(x) + (1-t)f(y)$$

By setting $y = 0$, and by using our assumption $f(0) = 0$, we obtain, as desired:

$$f(tx) = tf(x)$$

As for the first condition, regarding sums, this can be established as follows:

$$\begin{aligned} f(x+y) &= f\left(2 \cdot \frac{x+y}{2}\right) \\ &= 2f\left(\frac{x+y}{2}\right) \\ &= 2 \cdot \frac{f(x)+f(y)}{2} \\ &= f(x) + f(y) \end{aligned}$$

(2) $\implies$ (1) Conversely now, assuming that $f$ satisfies $f(x+y) = f(x) + f(y)$ and $f(\lambda x) = \lambda f(x)$, then $f$ must map lines to lines, as shown by:

$$\begin{aligned} f(tx + (1-t)y) &= f(tx) + f((1-t)y) \\ &= tf(x) + (1-t)f(y) \end{aligned}$$

Also, we have $f(0) = f(2 \cdot 0) = 2f(0)$, which gives $f(0) = 0$, as desired. □

The above result is very useful, and in practice, we will often use the condition (2) there, somewhat as a new definition for the linear maps.

Let us record this finding as an updated definition, as follows:

DEFINITION 5.28 (upgrade). *A map $f : \mathbb{R}^2 \to \mathbb{R}^2$ is called:*

(1) *Linear, when it satisfies $f(x + y) = f(x) + f(y)$ and $f(\lambda x) = \lambda f(x)$.*
(2) *Affine, when it is of the form $f = g + x$, with $g$ linear, and $x \in \mathbb{R}^2$.*

Before getting into the mathematics of linear maps, let us comment a bit more on the "maps lines to lines" feature of such maps. As mentioned after Definition 5.22, this feature requires thinking at lines as being "dynamic" objects, the point being that, when thinking at lines as being sets, this interpretation fails, as shown by the following map:

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x^3 \\ 0 \end{pmatrix}$$

However, in relation with all this we have the following useful result:

THEOREM 5.29. *For a continuous injective $f : \mathbb{R}^2 \to \mathbb{R}^2$, the following are equivalent:*

(1) *$f$ is affine in our sense, mapping lines to lines.*
(2) *$f$ maps set-theoretical lines to set-theoretical lines.*

PROOF. By composing $f$ with a translation, we can assume that we have $f(0) = 0$. With this assumption made, the proof goes as follows:

(1) $\implies$ (2) This is clear from definitions.

(2) $\implies$ (1) Let us first prove that we have $f(x + y) = f(x) + f(y)$. We do this first in the case where our vectors are not proportional, $x \not\sim y$. In this case we have a proper parallelogram $(0, x, y, x + y)$, and since $f$ was assumed to be injective, it must map parallel lines to parallel lines, and so must map our parallelogram into a parallelogram $(0, f(x), f(y), f(x + y))$. But this latter parallelogram shows that we have:

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

In the remaining case where our vectors are proportional, $x \sim y$, we can pick a sequence $x_n \to x$ satisfying $x_n \not\sim y$ for any $n$, and we obtain, as desired:

$$\begin{aligned} x_n \to x, x_n \not\sim y, \forall n &\implies f(x_n + y) = f(x_n) + f(y), \forall n \\ &\implies f(x + y) = f(x) + f(y) \end{aligned}$$

Regarding now $f(\lambda x) = \lambda f(x)$, since $f$ maps lines to lines, it must map the line $0 - x$ to the line $0 - f(x)$, so we have a formula as follows, for any $\lambda, x$:

$$f(\lambda x) = \varphi_x(\lambda) f(x)$$

But since $f$ maps parallel lines to parallel lines, by Thales the function $\varphi_x : \mathbb{R} \to \mathbb{R}$ does not depend on $x$. Thus, we have a formula as follows, for any $\lambda, x$:

$$f(\lambda x) = \varphi(\lambda) f(x)$$

We know that we have $\varphi(0) = 0$ and $\varphi(1) = 1$, and we must prove that we have $\varphi(\lambda) = \lambda$ for any $\lambda$. For this purpose, we use a trick. On one hand, we have:

$$f((\lambda + \mu)x) = \varphi(\lambda + \mu)f(x)$$

On the other hand, since $f$ maps sums to sums, we have as well:

$$\begin{aligned} f((\lambda + \mu)x) &= f(\lambda x) + f(\mu x) \\ &= \varphi(\lambda)f(x) + \varphi(\mu)f(x) \\ &= (\varphi(\lambda) + \varphi(\mu))f(x) \end{aligned}$$

Thus our rescaling function $\varphi : \mathbb{R} \to \mathbb{R}$ satisfies the following conditions:

$$\varphi(0) = 0 \quad , \quad \varphi(1) = 1 \quad , \quad \varphi(\lambda + \mu) = \varphi(\lambda) + \varphi(\mu)$$

But with these conditions in hand, it is clear that we have $\varphi(\lambda) = \lambda$, first for all the inverses of integers, $\lambda = 1/n$ with $n \in \mathbb{N}$, then for all rationals, $\lambda \in \mathbb{Q}$, and finally by continuity for all reals, $\lambda \in \mathbb{R}$. Thus, we have proved the following formula:

$$f(\lambda x) = \lambda f(x)$$

But this finishes the proof of (2) $\implies$ (1), and we are done. □

All this is very nice, linear algebra axiomatized. We will be back to this, later.

## 5e. Exercises

Exercises:

Exercise 5.30.

Exercise 5.31.

Exercise 5.32.

Exercise 5.33.

Exercise 5.34.

Exercise 5.35.

Exercise 5.36.

Exercise 5.37.

Bonus exercise.

CHAPTER 6

# Intermediate values

## 6a. Sets, topology

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 to 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 the following definition, which is certainly new to you:

DEFINITION 6.1. *The open and closed sets are defined as follows:*

(1) *Open means that there is a small interval around each point.*
(2) *Closed means that our set is closed under taking limits.*

As basic illustrations here, the open intervals are open, and the closed intervals are closed, as you would expect, as shown by the following result:

PROPOSITION 6.2. *The following happen:*

(1) *The open intervals $(a,b)$ are open.*
(2) *The closed intervals $[a,b]$ are closed.*

PROOF. This is something fairly easy, as follows:

(1) Consider indeed an open interval $(a,b)$, and a point inside, $x\in(a,b)$. We have then an inclusion as follows, which does the job, as in Definition 6.1 (1):

$$x\in\left(\frac{a+x}{2},\frac{x+b}{2}\right)\subset(a,b)$$

(2) Consider indeed a closed interval $[a,b]$, and a sequence inside $\{x_n\}\subset[a,b]$. Assuming that this sequence converges, $x_n\to x$, we have:

$$x_n\geq a\implies x\geq a$$

$$x_n\leq b\implies x\leq b$$

Thus we obtain $x\in[a,b]$, as required by Definition 6.1 (2). □

Further basic examples, or rather results which are easy to establish, producing further examples, include the fact that the finite unions or intersections of open or closed sets are open or closed. We will be back to this later, with some precise results in this sense.

For the moment, let us develop some general theory. In order to get truly started, with our study, we will need the following theoretical result:

THEOREM 6.3. *A set $O \subset \mathbb{R}$ is open precisely when its complement $C \subset \mathbb{R}$ is closed, and vice versa.*

PROOF. It is enough to prove the first assertion, since the "vice versa" part will follow from it, by taking complements. But this can be done as follows:

" $\implies$ " Assume that $O \subset \mathbb{R}$ is open, and let $C = \mathbb{R} - O$. In order to prove that $C$ is closed, assume that $\{x_n\}_{n\in\mathbb{N}} \subset C$ converges to $x \in \mathbb{R}$. We must prove that $x \in C$, and we will do this by contradiction. So, assume $x \notin C$. Thus $x \in O$, and since $O$ is open we can find a small interval $(x-\varepsilon, x+\varepsilon) \subset O$. But since $x_n \to x$ this shows that $x_n \in O$ for $n$ big enough, which contradicts $x_n \in C$ for all $n$, and we are done.

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

As a basic illustration for the above result, a disjoint union of two infinite open intervals is open, due to the fact that its complement is closed:

$$(-\infty, a) \cup (b, \infty) = \mathbb{R} - [a, b]$$

As another basic illustration, a disjoint union of two infinite closed intervals is open, due to the fact that its complement is open:

$$(-\infty, a] \cup [b, \infty) = \mathbb{R} - (a, b)$$

There are many other such illustrations, and we will be back in a moment to all this, with a systematic study of the various unions of open and closed intervals.

Getting now to the functions, we have the following key result about them, which makes it clear that the notions from Definition 6.1 are of interest for us:

THEOREM 6.4. *A function is continuous precisely when $f^{-1}(O)$ is open, for any $O$ open. Equivalently, $f^{-1}(C)$ must be closed, for any $C$ closed.*

PROOF. This is something coming from definitions, the idea being as follows:

(1) The first assertion follows from definitions, and more specifically from the $\varepsilon, \delta$ 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 a function $f$ satisfies this condition, it is then clear that if a set $O$ is open, then the set $f^{-1}(O)$ is open too. Moreover, the converse clearly holds too.

(2) As for the second assertion, regarding the closed sets, this can be proved directly, by using the $f(x_n) \to f(x)$ definition of continuity, or can be deduced from what we already know about the open sets, by taking complements. □

As a test for the above criterion, let us reprove the fact, that we know from chapter 5, 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 $\varepsilon, \delta$.

Let us record this as a philosophical conclusion, as follows:

Conclusion 6.5. *The open and closed sets are no joke, because with them, we can see right away that a composition of continuous functions is continuous.*

And, many more illustrations for this general principle coming soon. Wait for it.

In order to reach now to some true applications of Theorem 6.4, we will need to know more about the open and closed sets. Let us begin with a useful result, as follows:

Theorem 6.6. *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:

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

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

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

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 6.7. *The open and closed sets are as follows:*

(1) *The open sets are the disjoint unions of open intervals.*
(2) *The closed sets are the complements of these unions.*

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

(1) We know that the open intervals are those of type $(a,b)$ with $a<b$, with the values $a,b=\pm\infty$ allowed, and by Theorem 6.6 a union of such intervals is open.

(2) Conversely, given $O\subset\mathbb{R}$ open, we can cover each point $x\in O$ with an open interval $I_x\subset O$, and we have $O=\cup_x I_x$, so $O$ is a union of open intervals.

(3) In order to finish the proof of the first assertion, it remains to prove that the union $O=\cup_x I_x$ in (2) can be taken to be disjoint. For this purpose, our first observation is that, by approximating points $x\in O$ by rationals $y\in\mathbb{Q}\cap O$, we can make our union to be countable. But once our union is countable, we can start merging intervals, whenever they meet, and we are left in the end with a countable, disjoint union, as desired.

(4) Finally, the second assertion comes from Theorem 6.3. □

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 simple description of the open sets $O\subset\mathbb{R}$.

Finally, a word about $\mathbb{R}^N$. We are for the moment doing one variable functions, so a priori no need for this, but since we will be talking soon about complex functions, defined on $\mathbb{C}=\mathbb{R}^2$, some vague knowledge about what happens in $\mathbb{R}^N$ would be welcome.

And here, very briefly, again we can talk about open and closed sets, with most of the above results extending, but with our last result, Theorem 6.7, not extending, for instance due to all sorts of open connected surfaces in $\mathbb{R}^2$, or open connected bodies in $\mathbb{R}^3$, which obviously cannot all count as "intervals". More on this later, when needed.

Moving towards more concrete things, and applications, let us formulate the following key definition, which is actually one of the most important definitions in analysis:

DEFINITION 6.8. *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 here, the closed bounded intervals $[a, b]$ are compact, as we know from chapter 5, from the proof of the Heine-Cantor theorem, which used the notion of compactness, as formalized above, and so are the finite unions of such intervals.

As for connected sets, the basic examples here are the various types of intervals, namely $(a, b)$, $(a, b]$, $[a, b)$, $[a, b]$, with the endpoints being allowed to be finite, or not.

Thinking a bit at what we have above, as examples for both the compact and the connected sets, it looks impossible to come up with more examples. In fact, we have:

THEOREM 6.9. *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) To start with, the fact that compact implies both closed and bounded is clear from our definition of compactness, because assuming non-closedness or non-boundedness obviously leads to an open cover having no finite subcover.

(2) As for the converse, closed and bounded implies compact, we know from chapter 5 that any closed bounded interval $[a, b]$ is compact, and it follows that any $K \subset \mathbb{R}$ closed and bounded is a closed subset of a compact set, which follows to be compact.

(3) Regarding now the second assertion, in relation with the connexted sets, this is something which is obvious, and this regardless of what "cannot be broken into parts" in Definition 6.8 exactly means, mathematically speaking, with several possible definitions being possible here, and all these definitions being equivalent.

(4) Indeed, skipping some discussion here, $E \subset \mathbb{R}$ having this property is obviously equivalent to $a, b \in E \implies [a, b] \subset E$, and this gives the result. □

We will be back to all this later, when looking at open, closed, compact and connected sets in $\mathbb{R}^N$, or more general spaces, where things are more complicated than in $\mathbb{R}$. And with the comment that, in that setting, the notions of open, closed, compact and connected sets are essential, in order to understand what is going on.

Now with this discussed, let us go back to the continuous functions. We have here the following result, in the spirit of what we already know, from Theorem 6.4:

THEOREM 6.10. *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 6.8. 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. □

Next in line, let us record as well the following useful generalization of the Heine-Cantor theorem, that we know from chapter 5:

THEOREM 6.11. *Any continuous function defined on a compact set*

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

*is automatically uniformly continuous.*

PROOF. We can prove this exactly as we proved the Heine-Cantor theorem in chapter 5, by using the compactness of $X$, as we did there, as follows:

(1) Given $\varepsilon > 0$, for any $x \in X$ we know that we have a $\delta_x > 0$ such that:

$$|x - y| < \delta_x \implies |f(x) - f(y)| < \frac{\varepsilon}{2}$$

So, consider the following open intervals, centered at the various points $x \in X$:

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

(2) These intervals then obviously cover $X$, in the sense that we have:

$$X \subset \bigcup_{x \in X} U_x$$

By compactness of $X$, this cover must have a certain finite subcover, as follows:

$$X \subset \bigcup_i U_{x_i}$$

(3) With this done, we are ready to finish our proof. Let us set indeed:

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

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

## 6b. Intermediate values

Very nice all the above, good mathematical learning that was, but you might perhaps ask at this point, were our main concrete findings, Theorem 6.10 and Theorem 6.11 really worth all this abstract excursion into the 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 6.12. *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, $f(x) = 0$ has a solution.*

PROOF. All these statements are related, and are called altogether "intermediate value theorem". Regarding now the proof, this goes as follows:

(1) One elegant way of viewing things, based on what we learned above, is that since the interval $[a, b]$ is compact and connected, the set $f([a, b])$ is compact and connected too, and so it is a certain closed bounded interval $[c, d]$, and this gives all the results.

(2) However, all this is based on rather advanced technology, as developed above, and we should mention here that it is possible to prove (1-3) as well directly. We will leave finding the needed tricks here as an instructive exercise for you, reader. □

Along the same lines, we have as well the following result, which is also useful for a wide range of purposes, including fixing some previous bugs in the opresent book:

THEOREM 6.13. *Assuming that a function $f$ is continuous and invertible:*

(1) *This function must be monotone.*
(2) *Its inverse function $f^{-1}$ must be monotone and continuous too.*
(3) *Moreover, this statement holds both locally, and globally.*

PROOF. The fact that both $f$ and $f^{-1}$ are monotone follows from Theorem 6.12. 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. □

More concretely now, as a basic application of Theorem 6.13, we have:

THEOREM 6.14. *The exponential function, defined as usual by*

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

*is indeed given by* $\exp x = e^x$*, with* $e = \exp 1$*. This function is increasing, and continuous. Its inverse, the logarithm, can be defined by one of the following equivalent formulae,*

$$\exp(\log x) = x \quad , \quad \log(\exp x) = x$$

*and is again increasing, and continuous.*

PROOF. These are all things that we basically know, from chapter 4 and from here, but via proofs which, while certainly correct, lack a bit of rigor. Now the point is that, with Theorems 6.12 and 6.13 in hand, we can make all this fully rigorous:

(1) Regarding the exponential, let us go back to what we did in chapter 4, when first talking about it. According to our general theory there, the number $e = 2.71828\ldots$ is defined by the following formula, with the limit shown there to converge indeed:

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

In fact, as also explained there, we have the following formula, valid for any $x \in \mathbb{R}$:

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

(2) Coming next, still by following the material in chapter 4, some manipulations on the binomials and factorials producing $e$ lead the following alternative formula for $e$:

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

And for completing now the picture, we have a fourth formula, which is as follows, generalizing the previous one, valid for any $x \in \mathbb{R}$:

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

(3) In what regards now the proof of this latter formula, still by following the material in chapter 4, the idea is to consider the following function:

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

By using the binomial formula, we have then the following computation:

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

(4) In order to finish now our study, 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.

(5) Finally, in what regards the logarithm, its existence and continuity, and property of being increasing, all follow from Theorem 6.13, applied to the exponential. □

Along the same lines now, with exp and log understood, what about looking at other familiar inverse functions. And here, we have the following result:

THEOREM 6.15. *The following functions are well-defined and continuous,*

(1) $\arcsin x$.
(2) $\arccos x$,
(3) $\arctan x$,

*and the same goes for the functions* $\operatorname{arcsec} x$, $\operatorname{arccsc} x$, $\operatorname{arccot} x$.

PROOF. The situation here is quite similar to that for the logarithm, with the existence and continuity of the functions in the statement coming from what we know about sin, cos, tan and sec, csc, cot from chapter 5, via Theorem 6.13. We will leave the verifications here, and the discussion about domains and ranges, as an instructive exercise. □

Getting now towards more concrete things, still solving questions that we previously had open, as another basic application of the intermediate value theorem, we have:

THEOREM 6.16. *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 6.12, as follows:

(1) This is clear from Theorem 6.12 (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 6.12 (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 6.12. And with all this being quite tricky, quite often being rather advanced mathematics, requiring quite a deal of new theory and methods. We will be back to these questions later in this book.

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

THEOREM 6.17. *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 Theorem 6.16, 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 Theorem 6.16:

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

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

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

(3) Assume first $a\in\mathbb{Q}_{even}$, $a>0$. This means $a=p/q$ with $p,q\in\mathbb{N}$, $p\neq 0$ and $q$ even, and as before in (1), we can set $x^a=\sqrt[q]{x^p}$ for $x\geq 0$, by using Theorem 6.16. 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 6.18. *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 6.17, by exchanging the variables, $x \leftrightarrow a$. To be more precise, the situation is as follows:

(1) We know from Theorem 6.17 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 6.17 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. □

As a comment here, both Theorem 6.17 and Theorem 6.18 obviously hide some non-trivial mathematics, and might remain quite mysterious. We will be back to this.

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 see at the end of the present chapter how to extend some of our results to the complex variable case, and with this actually bringing some more light on our last results above, regarding the power functions. More on this in a moment.

## 6c. Separation results

Getting back now to general topology, as developed in the beginning of this chapter, for a number of various technical reasons, we would like to squeeze the arbitrary sets $E \subset X$ between compact sets $K \subset X$ and open sets $U \subset X$, as follows:

$$K \subset E \subset U$$

In order to do this, which is something quite tricky, we will need a number of technical preliminaries, as a continuation of our general topology material. First, we have:

PROPOSITION 6.19. *Given $K \subset U$, compact inside open, we can find inclusions*

$$K \subset V \subset \bar{V} \subset U$$

*with the set $V$ being open, with compact closure $\bar{V}$.*

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

(1) For any $x \in K$ pick a neighborhood $W_x$ having compact closure. Since $K$ is compact, we can find finitely many such neighborhoods, covering it:

$$K \subset W_{x_1} \cup \ldots \cup W_{x_n}$$

With this done, consider now the set on the right, namely:

$$W = W_{x_1} \cup \ldots \cup W_{x_n}$$

This is then an open set containing our compact set $K$, having compact closure, that we will use in what follows, in our constructions.

(2) In order to prove now our result, observe first that if we are in the case $U = X$ we are done, because here we can simply take $V = W$. So, assume $U \neq X$.

(3) Given a point $x \in U^c$, since this point is compact, and our set $K$ not containing it is compact as well, we can find an open set $V_x$ having the following properties:

$$K \subset V_x \quad , \quad x \notin \bar{V}_x$$

Now consider the following family of sets, with $W$ being the open set in (1):

$$\left\{ K \cap \bar{W} \cap \bar{V}_x \middle| x \in U^c \right\}$$

This is then a family of compact sets having empty intersection, so we can find a finite subfamily having empty intersection too. That is, we can find points $x_i \in U^c$ such that:

$$K \cap \bar{W} \cap \bar{V}_{x_1} \cap \ldots \cap \bar{V}_{x_m} = \emptyset$$

(4) With this done, consider now the following open set:

$$V = W \cap V_{x_1} \cap \ldots \cap V_{x_m}$$

Then $V$ has compact closure, and we have the following inclusion:

$$\bar{V} \subset \bar{W} \cap \bar{V}_{x_1} \cap \ldots \cap \bar{V}_{x_m}$$

Thus, $V$ is the open set we are looking for, and we are done. □

The above result, separating $K \subset U$, compact set inside open set, via an open set $V$ with compact closure $\bar{V}$, is something quite useful, in practice. However, for our purposes here, we will need more, along the same lines.

To be more precise, under the same assumptions, $K \subset U$, compact inside open, we would like to come up with a continuous, compactly supported function $f$ such that:

$$\chi_K \leq f \leq \chi_U$$

Which is something quite non-trivial, when thinking a bit, and so again, and with our apologies here, we are in need of some more technical preliminaries.

To be more precise, we will need the following standard notions:

DEFINITION 6.20. *A function $f : X \to \mathbb{R}$, with $X$ topological space, is called:*

(1) *Upper semicontinuous, if $\{x \in X | f(x) < a\}$ is open, for any $a \in \mathbb{R}$.*
(2) *Lower semicontinuous, if $\{x \in X | f(x) > a\}$ is open, for any $a \in \mathbb{R}$.*

These notions are actually quite interesting on their own, having many uses in many contexts in analysis, and as basic illustrations for them, we have:

(1) A characteristic function $\chi_E$ is upper semicontinuous when $E \subset X$ is closed. This follows indeed from the above definition of the upper semicontinuity.

(2) Also, a characteristic function $\chi_E$ is lower semicontinuous when $E \subset X$ is open. Again, this follows from definitions, or simply from (1), by using the set $E^c$.

(3) Observe also that a continuous function is trivially both upper and lower semicontinuous. The converse of this holds too, and more on this in a moment.

Many things can be said about the upper and lower semicontinuous functions, which some knowledge here being something quite useful, when doing in analysis in general. In what follows we will only need a handful of basic results on the subject, as follows:

PROPOSITION 6.21. *The upper and lower semicontinuous functions $f : X \to \mathbb{R}$ have the following properties:*

(1) *If $F \subset X$ is closed, $\chi_F$ is upper semicontinuous.*
(2) *If $U \subset X$ is open, $\chi_U$ is lower semicontinuous.*
(3) *Infimum of upper semicontinuous functions is upper semicontinuous.*
(4) *Supremum of lower semicontinuous functions is lower semicontinuous.*
(5) *$f$ is continuous precisely when it is upper and lower semicontinuous.*

PROOF. All this is elementary, the idea being as follows:

(1) This assertion, already mentioned in the above, follows from definitions.

(2) This assertion, also mentioned in the above, also follows from definitions.

(3) This assertion is also trivial, also coming from definitions.

(4) And this assertion is trivial too, also coming from definitions.

(5) In one sense, this is clear, as already mentioned in the above. In the other sense, assuming that $f : X \to \mathbb{R}$ is both upper and lower semicontinuous, we can see that the preimage of any open interval $(a, b) \subset \mathbb{R}$ is open, due to the following formula:

$$f^{-1}(a, b) = f^{-1}(a, \infty] \cap f^{-1}[-\infty, b)$$

Now since any open set $U \subset \mathbb{R}$ can be written a union of open intervals $(a, b) \subset \mathbb{R}$, it follows that $f^{-1}(U)$ is open, and so that $f$ is continuous, as desired. □

Many other things can be said, about the upper and lower semicontinuous functions, notably with many examples. However, for our purposes here, the above will do.

Good news, we can now formulate the main technical result that we will need, for our measure theory purposes here. With the convention that the support of a function $f$ is the closure of the set $\{x|f(x)\neq 0\}$, this key statement is as follows:

THEOREM 6.22 (Urysohn). *Given $K\subset U$, compact inside open, we can find*

$$\chi_K\leq f\leq\chi_U$$

*with $f$ being continuous, and compactly supported.*

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

(1) Given sets $K\subset U$ as in the statement, by using Proposition 6.19 we can find an open set $V_0$, having compact closure $\bar{V}_0$, such that:

$$K\subset V_0\subset\bar{V}_0\subset U$$

But then, by using again Proposition 6.19, applied this time to the inclusion $K\subset V_0$, we can find a second open set $V_1$, having compact closure $\bar{V}_1$, such that:

$$K\subset V_1\subset\bar{V}_1\subset V_0\subset\bar{V}_0\subset U$$

And so on, would be the idea. In practice now, by using the fact that the rational numbers are countable, we can construct in this way a whole family of open sets $V_r$, having compact closures $\bar{V}_r$, one for each rational number $r\in[0,1]$, such that:

$$r<s\implies\bar{V}_s\subset V_r$$

(2) Time now to construct our function $f$. Let us set, for any $r\in[0,1]$ rational:

$$f_r(x)=\begin{cases}r & \text{if } x\in V_r\\ 0 & \text{otherwise}\end{cases}$$

Then, we can construct our function $f$ in the following way:

$$f=\sup_r f_r$$

It is clear then that $f$ is lower semicontinuous, that we have $\chi_K\leq f\leq\chi_U$, and also that $f$ is compactly supported, with support included in the compact set $\bar{V}_0$. Thus, we are almost there, and it remains to prove that $f$ is upper semicontinuous as well.

(3) For this purpose, let us set as well, for any $r\in[0,1]$ rational:

$$g_s(x)=\begin{cases}1 & \text{if } x\in\bar{V}_s\\ 0 & \text{otherwise}\end{cases}$$

Then, we can construct another function $g$, in the following way:

$$g = \inf_s g_s$$

It is then clear, exactly as in (2), that $g$ is lower semicontinuous, and also that we have $\chi_K \leq g \leq \chi_U$, and that $g$ is compactly supported. Of interest for us is the lower semicontinuity of $g$, because in order to finish the proof, it is enough to show that:

$$f = g$$

(4) So, let us prove this, $f = g$. By definition of the functions $f_r$ and $g_s$, we have:

$$f_r < g_s$$

Thus $f \leq g$. Now assume that we have somewhere a strict inequality, $f(x) < g(x)$. Then, we can find two rational numbers $r, s$ in between, as follows:

$$f(x) < r < s < g(x)$$

But the first inequality tells us that $x \notin V_r$, and the last inequality tells us that $x \in \bar{V}_s$, and this contradicts the condition in (1) on the sets $V_r$, as desired. □

Very nice all this, but in fact, contrary to what was advertised before, we will actually need in what follows, besides the Urysohn theorem, another technical result, which is something useful too, and actually of independent interest too, as follows:

THEOREM 6.23. *Given a compact set inside a union of open sets*

$$K \subset U_1 \cup \ldots \cup U_n$$

*we can find an associated partition of unity, that is, a decomposition of type*

$$f_1(x) + \ldots + f_n(x) = 1 \quad , \quad \forall x \in K$$

*with each $f_i$ being continuous, supported on $U_i$.*

PROOF. This follows by using the Urysohn theorem, as follows:

(1) For any $x \in K$, let us pick a neighborhood $V_x$ such that $\bar{V}_x \subset U_i$, for some $i$. Since $K$ is compact, we can find finitely many points $x_1, \ldots, x_m \in K$ such that:

$$K \subset V_{x_1} \cup \ldots \cup V_{x_m}$$

Now for any index $i \in \{1, \ldots, n\}$ let us consider the following union:

$$K_i = \bigcup_{\bar{V}_{x_i} \subset U_i} \bar{V}_{x_i}$$

We can apply then the Urysohn lemma to the inclusion $K_i \subset U_i$, and we obtain in this way a continuous, compactly supported function $g_i$, such that:

$$\chi_{K_i} \leq g_i \leq \chi_{U_i}$$

(2) With this done, consider the following sequence of functions:

$$f_1 = g_1$$

$$f_2 = (1 - g_1)g_2$$

$$\vdots$$

$$f_n = (1 - g_1) \ldots (1 - g_{n-1})g_n$$

Then each $f_i$ is continuous, supported on $U_i$, and we have:

$$f_1 + \ldots + f_n = 1 - (1 - g_1) \ldots (1 - g_n)$$

(3) On the other hand, recall from the construction of $K_i$ above that we have:

$$K \subset K_1 \cup \ldots \cup K_n$$

We conclude from this that we have the following formula:

$$f_1 + \ldots + f_n = 1$$

Thus, we have our partition of the unity, as desired. □

Still with me, I hope. All the above results are very useful, in practice, for instance when developing measure theory. We will be back to this at the end of this book.

## 6d. Complex functions

Switching topics now, and going towards the complex functions, at the level of the general theory, the main tool for dealing with the continuous functions $f : \mathbb{R} \to \mathbb{R}$ is the above intermediate value theorem. In the complex setting, that of the functions $f : \mathbb{C} \to \mathbb{C}$, we do not have such a theorem, at least in its basic formulation, because there is no order relation for the complex numbers, or things like complex intervals.

However, the intermediate value theorem in its advanced formulation, that with connected sets, extends of course, and this will be our main result, on the subject.

We will need, in order to get started with our discussion:

DEFINITION 6.24. *The distance between two complex numbers is the usual distance in the plane between them, namely:*

$$d(x, y) = |x - y|$$

*With this, we can talk about convergence, by saying that $x_n \to x$ when $d(x_n, x) \to 0$.*

Here the fact that $d(x,y)=|x-y|$ is indeed the usual distance in the plane is clear for $y=0$, because we have $d(x,0)=|x|$, by definition of the modulus $|x|$. As for the general case, $y\in\mathbb{C}$, this comes from the fact that the distance in the plane is given by:

$$d(x,y)=d(x-y,0)=|x-y|$$

Observe that in real coordinates, the distance formula is quite complicated, namely:

$$\begin{aligned}d(a+ib,c+id)&=|(a+ib)-(c+id)|\\&=|(a-c)+i(b-d)|\\&=\sqrt{(a-c)^2+(b-d)^2}\end{aligned}$$

However, for most computations, we will not need this formula, and we can get away with the various tricks regarding complex numbers that we know. As a first result now, regarding $\mathbb{C}$ and its distance, that we will need in what follows, we have:

THEOREM 6.25. *The complex plane $\mathbb{C}$ is complete, in the sense that any Cauchy sequence converges.*

PROOF. Consider indeed a Cauchy sequence $\{x_n\}_{n\in\mathbb{N}}\subset\mathbb{C}$. If we write $x_n=a_n+ib_n$ for any $n\in\mathbb{N}$, then we have the following estimates:

$$|a_n-a_m|\leq\sqrt{(a_n-a_m)^2+(b_n-b_m)^2}=|x_n-x_m|$$

$$|b_n-b_m|\leq\sqrt{(a_n-a_m)^2+(b_n-b_m)^2}=|x_n-x_m|$$

Thus both the sequences $\{a_n\}_{n\in\mathbb{N}}\subset\mathbb{R}$ and $\{b_n\}_{n\in\mathbb{N}}\subset\mathbb{R}$ are Cauchy, and since we know that $\mathbb{R}$ itself is complete, we can consider the limits of these sequences:

$$a_n\to a\quad,\quad b_n\to b$$

With $x=a+ib$, our claim is that $x_n\to x$. Indeed, we have:

$$\begin{aligned}|x_n-x|&=\sqrt{(a_n-a)^2+(b_n-b)^2}\\&\leq|a_n-a|+|b_n-b|\end{aligned}$$

It follows that we have $x_n\to x$, as claimed, and this gives the result. □

Talking complex functions now, we have the following definition:

DEFINITION 6.26. *A complex function $f:\mathbb{C}\to\mathbb{C}$, or more generally $f:X\to\mathbb{C}$, with $X\subset\mathbb{C}$ being a subset, is called continuous when, for any $x_n,x\in X$:*

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

*Also, we can talk about pointwise convergence of functions, $f_n\to f$, and about uniform convergence too, $f_n\to_u f$, exactly as for the real functions.*

Observe that, since $x_n \to x$ in the complex sense means that $(a_n, b_n) \to (a, b)$ in the usual, real plane sense, a function $f : \mathbb{C} \to \mathbb{C}$ is continuous precisely when it is continuous when regarded as real function, $f : \mathbb{R}^2 \to \mathbb{R}^2$. Which is something good to know.

Regarding what is said at the end of Definition 6.26, that is certainly quite vague, I know, and exercise for you to figure out what can be done there, and what not.

Let us point out now the fact that, contrary to what the above might suggest, everything does not always extend trivally from real to complex. For instance, we have:

PROPOSITION 6.27. *We have the following formula, valid for any $|x| < 1$,*

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

*but, for $x \in \mathbb{C} - \mathbb{R}$, the geometric meaning of this formula is quite unclear.*

PROOF. Here the formula in the statement holds indeed, by multiplying and cancelling terms, and with the convergence being justified by the following estimate:

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

As for the last assertion, this is something quite informal, the idea being as follows:

(1) To start with, for the simplest possible value of our parameter, $x = 1/2$, our formula is clear, by cutting the interval $[0, 2]$ into half, and so on:

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

(2) More generally, for $x \in (-1, 1)$ the meaning of the formula in the statement is something quite clear and intuitive, geometrically speaking, by using a similar argument, and we will leave the verifications and thinking here as an instructive exercise.

(3) However, when $x$ is complex, and not real, we are led into a kind of mysterious spiral there, and the only case where the formula is "obvious", geometrically speaking, is that when $x = rw$, with $r \in [0, 1)$, and with $w$ being a root of unity.

(4) To be more precise here, by anticipating a bit, assume that we have a complex number $w \in \mathbb{C}$ satisfying the root of unity equation $w^N = 1$, for some $N \in \mathbb{N}$. We have

then the following formula, for our infinite sum:

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

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

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

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

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

(7) Summarizing, as claimed above, the geometric series formula can be understood, in a purely geometric way, for variables of type $x=rw$, with $r\in[0,1)$, and with $w$ being a root of unity. In general, however, this formula tells us that the numbers on a certain infinite spiral sum up to a certain number, which remains something quite mysterious. □

Getting now to less mysterious mathematics, we have the question of understanding what our open and closed set technology becomes, in the complex setting. And here, as good news, the intermediate value theorem in its advanced formulation, that with connected sets, extends of course, and we have the following result:

THEOREM 6.28. *Assuming that $f:X\to\mathbb{C}$ with $X\subset\mathbb{C}$ is continuous, if the domain $X$ 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 $\mathbb{C}$. □

We will be back to this, with applications, later in this book.

## 6e. Exercises

Exercises:

Exercise 6.29.

Exercise 6.30.

Exercise 6.31.

Exercise 6.32.

Exercise 6.33.

Exercise 6.34.

Exercise 6.35.

Exercise 6.36.

Bonus exercise.

CHAPTER 7

# Sequences and series

## 7a. Pointwise convergence

Our goal now will be to extend the material from chapter 1 regarding the numeric sequences and series, to the case of the sequences and series of functions.

To start with, with our study, we can talk about the convergence of sequences of functions, $f_n \to f$, in a quite straightforward way, as follows:

DEFINITION 7.1. *We say that $f_n$ converges pointwise to $f$, and write $f_n \to f$, if*

$$f_n(x) \to f(x)$$

*for any $x$. Equivalently, $\forall x, \forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, |f_n(x) - f(x)| < \varepsilon$.*

The question is now, assuming that $f_n$ are continuous, does it follow that $f$ is continuous? I am pretty much sure that you think that the answer is "yes", based on:

$$\begin{aligned}\lim_{y\to x} f(y) &= \lim_{y\to x}\lim_{n\to\infty} f_n(y)\\ &= \lim_{n\to\infty}\lim_{y\to x} f_n(y)\\ &= \lim_{n\to\infty} f_n(x)\\ &= f(x)\end{aligned}$$

However, this proof is wrong, because we know well from chapter 1 that we cannot intervert limits, with this being a common beginner mistake. In fact, the result itself is wrong in general, because if we consider the functions $f_n : [0,1] \to \mathbb{R}$ given by $f_n(x) = x^n$, which are obviously continuous, their limit is discontinuous, given by:

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

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

PROPOSITION 7.2. *The basic step function, namely the sign function*

$$sgn(x) = \begin{cases} -1 & , \quad x < 0 \\ 0 & , \quad x = 0 \\ 1 & , \quad x > 0 \end{cases}$$

*can be approximated by suitable modifications of* $\arctan(x)$*. Even worse, there are examples of* $f_n \to f$ *with each* $f_n$ *continuous, and with* $f$ *totally discontinuous.*

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

## 7b. Uniform convergence

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

DEFINITION 7.3. *We say that* $f_n$ *converges uniformly to* $f$*, and write* $f_n \to_u f$*, if:*

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, |f_n(x) - f(x)| < \varepsilon, \forall x$$

*That is, the same condition as for* $f_n \to f$ *must be satisfied, but with the* $\forall x$ *at the end.*

And it is this "$\forall x$ at the end" which makes the difference, and will make our theory work. In order to understand this, which is something quite subtle, let us compare Definition 7.1 and Definition 7.3. As a first observation, we have:

PROPOSITION 7.4. *Uniform convergence implies pointwise convergence,*

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

*but the converse is not true, in general.*

PROOF. Here the first assertion is clear from definitions, just by thinking at what is going on, with no computations needed. As for the second assertion, the simplest counterexamples here are the functions $f_n : [0, 1] \to \mathbb{R}$ given by $f_n(x) = x^n$, that we met before in Proposition 7.2. Indeed, uniform convergence on $[0, 1)$ would mean:

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

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

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

THEOREM 7.5. *Assuming that $f_n$ are continuous, and that*

$$f_n \to_u f$$

*then $f$ is continuous. That is, uniform limit of continuous functions is continuous.*

PROOF. As previously advertised, it is the "$\forall x$ at the end" in Definition 7.3 that will make this work. Indeed, let us try to prove that the limit $f$ is continuous at some point $x$. For this, we pick a number $\varepsilon > 0$. Since $f_n \to_u f$, we can find $N \in \mathbb{N}$ such that:

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

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

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

But with this, we are done. Indeed, for $|x - y| < \delta$ we have:

$$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f_N(x)| + |f_N(x) - f_N(y)| + |f_N(y) - f(y)| \\ &\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \\ &= \varepsilon \end{aligned}$$

Thus, the limit function $f$ is continuous at $x$, and we are done. □

Obviously, the notion of uniform convergence in Definition 7.3 is something quite interesting, worth some more study. As a first result, we have:

PROPOSITION 7.6. *The following happen, regarding uniform limits:*

(1) $f_n \to_u f$, $g_n \to_u g$ *imply* $f_n + g_n \to_u f + g$.
(2) $f_n \to_u f$, $g_n \to_u g$ *imply* $f_n g_n \to_u fg$.
(3) $f_n \to_u f$, $f \neq 0$ *imply* $1/f_n \to_u 1/f$.
(4) $f_n \to_u f$, $g$ *continuous imply* $f_n \circ g \to_u f \circ g$.
(5) $f_n \to_u f$, $g$ *continuous imply* $g \circ f_n \to_u g \circ f$.

PROOF. All this is routine, exactly as for the results for numeric sequences from chapter 4, that we know well, with no difficulties or tricks involved. □

## 7c. Spaces of functions

There is some abstract mathematics to be done as well. Indeed, observe that the notion of uniform convergence, as formulated in Definition 7.3, means that:

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

This suggests measuring the distance between functions via a supremum as above. In order to discuss this, we will need some abstract preliminaries. Let us start with:

DEFINITION 7.7. *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:

$$\sqrt{\sum_i (a_i + b_i)^2} \leq \sqrt{\sum_i a_i^2} + \sqrt{\sum_i b_i^2}$$

Now 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 calculus, and which follows by using the fact that $f(t) = \sum_i (a_i + tb_i)^2$ being positive, its discriminant must be negative.

As another example, we have $\mathbb{C}^N$, as well as any of its subsets $X \subset \mathbb{C}^N$. Indeed, this follows either from $\mathbb{C}^N \simeq \mathbb{R}^{2N}$, or directly, along the lines of the above proof for $\mathbb{R}^N$. To be more precise, after some algebra, we are left with proving the following inequality:

$$\left|\sum_i a_i \bar{b}_i\right|^2 \leq \left(\sum_i |a_i|^2\right)\left(\sum_i |b_i|^2\right)$$

But this is the complex version of the Cauchy-Schwarz inequality, that we know well from calculus, and which follows by using the fact that the function $f(t) = \sum_i |a_i + twb_i|^2$ with $t \in \mathbb{R}$ and $|w| = 1$ being positive, its discriminant must be negative.

Here is now another example, which at first looks new and interesting, but is in fact not new, because it appears as a subspace of a suitable $\mathbb{R}^N$:

PROPOSITION 7.8. *Given a finite set $X$, the following function is a metric on it, called discrete metric:*

$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}$$

*This metric space is in fact the $N$-simplex, which can be realized as a subspace of $\mathbb{R}^{N-1}$, or, more conveniently, as a subspace of $\mathbb{R}^N$.*

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

(1) First of all, the axioms from Definition 7.7 are trivially satisfied, and with the main axiom, namely the triangle inequality, basically coming from:

$$1 \leq 1 + 1$$

(2) At the level of examples, at $|X| = 1$ we obtain a point, at $|X| = 2$ we obtain a segment, at $|X| = 3$ we obtain an equilateral triangle, at $|X| = 4$ we obtain a regular tetrahedron, and so on. Thus, what we have in general, at $|X| = N$, is the arbitrary dimensional generalization of this series of geometric objects, called $N$-simplex.

(3) In what regards now the geometric generalization of the $N$-simplex, our above examples, namely segment, triangle, tetrahedron and so on, suggest to look for an embedding $X \subset \mathbb{R}^{N-1}$. Which is something which is certainly possible, but the computations here are quite complicated, involving a lot of trigonometry, as you can check yourself by studying the problem at $N = 4$, that is, parametrizing the regular tetrahedron in $\mathbb{R}^3$.

(4) However, mathematics, or perhaps physics come to the rescue, via the idea "add a dimension, for getting smarter". Indeed, when looking for an embedding $X \subset \mathbb{R}^N$ things drastically simplify, because we can simply take $X$ to be the standard basis of $\mathbb{R}^N$:

$$X = \{e_1, \ldots, e_N\}$$

Indeed, we have by definition $d(e_i, e_j) = 1$ for any $i \neq j$. So, we have solved our embedding problem, just like that, without doing any computations or trigonometry. □

Moving ahead now with some theory, and allowing us a bit of slopiness, since we are already quite familiar with such things, we have the following statement:

PROPOSITION 7.9. *We can talk about limits inside metric spaces $X$, by saying that*

$$x_n \to x \iff d(x_n, x) \to 0$$

*and we can talk as well about continuous functions $f : X \to Y$, by requiring that*

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

*and with these notions in hand, all the basic results from the cases $X = \mathbb{R}, \mathbb{C}$ extend.*

PROOF. All this is very standard, and we will leave this as an exercise, namely carefully checking what you learned in basic calculus, in relation with limits and continuity, in the cases $X = \mathbb{R}, \mathbb{C}$, and working out the metric space extensions of this. In fact, we have already talked in the above such things, when discussing the complex functions. □

More interestingly now, along the same lines, we can talk about open and closed sets inside metric spaces $X$, again in analogy with what we did for $X = \mathbb{R}, \mathbb{C}$, but with a whole lot of interesting new phenomena appearing, which are worth exploring.

So, we will do this in detail. Let us start with a basic result, as follows:

DEFINITION 7.10. *Let $X$ be a metric space.*

(1) *The open balls are the sets $B_x(r) = \{y \in X | d(x, y) < r\}$.*
(2) *The closed balls are the sets $\bar{B}_x(r) = \{y \in X | d(x, y) \leq r\}$.*
(3) *$U \subset X$ is called open if for any $x \in U$ we have a ball $B_x(r) \subset U$.*
(4) *$F \subset X$ is called closed if its complement $F^c \subset X$ is open.*

With this in hand, let us work out the basics. We first have the following result, clarifying some terminology issues from the above definition:

PROPOSITION 7.11. *The open balls are open, and the closed balls are closed.*

PROOF. This might sound like a joke, but it is not one, because this is the kind of thing that we have to check. Fortunately, all this is elementary, as follows:

(1) Given an open ball $B_x(r)$ and a point $y \in B_x(r)$, by using the triangle inequality we have $B_y(r') \subset B_x(r)$, with $r' = r - d(x, y)$. Thus, $B_x(r)$ is indeed open.

(2) Given a closed ball $\bar{B}_x(r)$ and a point $y \in B_x(r)^c$, by using the triangle inequality we have $B_y(r') \subset B_x(r)^c$, with $r' = d(x, y) - r$. Thus, $\bar{B}_x(r)$ is indeed closed. □

Here is now something more interesting, making the link with our intuitive understanding of the notion of closedness, coming from our experience so far with analysis:

THEOREM 7.12. *For a subset $F \subset X$, the following are equivalent:*

(1) *$F$ is closed in our sense, meaning that $F^c$ is open.*
(2) *We have $x_n \to x, x_n \in F \implies x \in F$.*

PROOF. We can prove this by double implication, as follows:

(1) $\implies$ (2) Assume by contradiction $x_n \to x, x_n \in F$ with $x \notin F$. Since we have $x \in F^c$, which is open, we can pick a ball $B_x(r) \subset F^c$. But this contradicts our convergence assumption $x_n \to x$, so we are done with this implication.

(2) $\implies$ (1) Assume by contradiction that $F$ is not closed in our sense, meaning that $F^c$ is not open. Thus, we can find $x \in F^c$ such that there is no ball $B_x(r) \subset F^c$. But with $r = 1/n$ this provides us with a point $x_n \in B_x(1/n) \cap F$, and since we have $x_n \to x$, this contradicts our assumption (2). Thus, we are done with this implication too. □

Here is another basic theorem about open and closed sets:

THEOREM 7.13. *Let $X$ be a metric space.*

(1) *If $U_i$ are open, then $\cup_i U_i$ is open.*
(2) *If $F_i$ are closed, then $\cap_i F_i$ is closed.*
(3) *If $U_1, \ldots, U_n$ are open, then $\cap_i U_i$ is open.*
(4) *If $F_1, \ldots, F_n$ are closed, then $\cup_i F_i$ is closed.*

*Moreover, both* (3) *and* (4) *can fail for infinite intersections and unions.*

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

(1) This is clear from definitions, because any point $x \in \cup_i U_i$ must satisfy $x \in U_i$ for some $i$, and so has a ball around it belonging to $U_i$, and so to $\cup_i U_i$.

(2) This follows from (1), by using the following well-known set theory formula:

$$\left(\bigcup_i U_i\right)^c = \bigcap_i U_i^c$$

(3) Given an arbitrary point $x \in \cap_i U_i$, we have $x \in U_i$ for any $i$, and so we have a ball $B_x(r_i) \subset U_i$ for any $i$. Now with this in hand, let us set:

$$B = B_x(r_1) \cap \ldots \cap B_x(r_n)$$

As a first observation, this is a ball around $x$, $B = B_x(r)$, of radius given by:

$$r = \min(r_1, \ldots, r_n)$$

But this ball belongs to all the $U_i$, and so belongs to their intersection $\cap_i U_i$. We conclude that the intersection $\cap_i U_i$ is open, as desired.

(4) This follows from (3), by using the following well-known set theory formula:

$$\left(\bigcap_i U_i\right)^c = \bigcup_i U_i^c$$

(5) Finally, in what regards the counterexamples at the end, these can be both found on $\mathbb{R}$, and we will leave this as an instructive exercise. □

Still in relation with the open and closed sets, let us formulate as well:

DEFINITION 7.14. *Let $X$ be a metric space, and $E \subset X$ be a subset.*

(1) *The interior $E^\circ \subset E$ is the set of points $x \in E$ which admit around them open balls $B_x(r) \subset E$.*
(2) *The closure $E \subset \bar{E}$ is the set of points $x \in X$ which appear as limits of sequences $x_n \to x$, with $x \in E$.*

These notions are quite interesting, because they make sense for any set $E$. That is, when $E$ is open, that is open and end of the story, and when $E$ is closed, that is closed and end of the story. In general, however, a set $E \subset X$ is not open or closed, and what we can best do to it, in order to study it with our tools, is to "squeeze" it, as follows:

$$E^\circ \subset E \subset \bar{E}$$

In practice now, in order to use the above notions, we need to know a number of things, including that fact that $E$ open implies $E^\circ = E$, the fact that $E$ closed implies $\bar{E} = E$, and many more such results. But all this can be done, and the useful statement here, summarizing all that we need to know about interiors and closures, is as follows:

THEOREM 7.15. *Let $X$ be a metric space, and $E \subset X$ be a subset.*

(1) *The interior $E^\circ \subset E$ is the biggest open set contained in $E$.*
(2) *The closure $E \subset \bar{E}$ is the smallest closed set containing $E$.*

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

(1) Let us first prove that the interior $E^\circ$ is open. For this purpose, pick $x \in E^\circ$. We know that we have a ball $B_x(r) \subset E$, and since this ball is open, it follows that we have $B_x(r) \subset E^\circ$. Thus, the interior $E^\circ$ is open, as claimed.

(2) Let us prove now that the closure $\bar{E}$ is closed. For this purpose, we will prove that the complement $\bar{E}^c$ is open. So, pick $x \in \bar{E}^c$. Then $x$ cannot appear as a limit of a sequence $x_n \to x$ with $x_n \in E$, so we have a ball $B_x(r) \subset \bar{E}^c$, as desired.

(3) Finally, the maximality and minimality assertions regarding $E^\circ$ and $\bar{E}$ are both routine too, coming from definitions, and we will leave them as exercises. □

As a continuation of this, we can talk as well about density, as follows:

DEFINITION 7.16. *We say that a subset $E \subset X$ is dense when:*

$$\bar{E} = X$$

*That is, any point of $X$ must appear as a limit of points of $E$.*

Obviously, this is something which is in tune with what we know so far from this book, and with the intuitive notion of density. As a basic example, we have $\bar{\mathbb{Q}} = \mathbb{R}$.

Again in analogy with what we know about $X = \mathbb{R}, \mathbb{C}$, we can talk about compact sets. However, things here are quite tricky, in the general metric space framework, substantially deviating from what we know, and we will do this in detail. Let us start with:

DEFINITION 7.17. *A set $K \subset X$ is called compact if any cover with open sets*

$$K \subset \bigcup_i U_i$$

*has a finite subcover, $K \subset (U_{i_1} \cup \ldots \cup U_{i_n})$.*

This definition, which is probably new to you, might seem overly abstract, but our claim is that this is the correct definition, and that there is no way of doing otherwise. Let us start with some examples, with $X = \mathbb{R}$. The situation here is as follows:

(1) A point is obviously compact, and we can choose that finite subcover with $n = 1$. Similarly, 2 points are compact, and we can choose the subcover with $n = 2$. More generally, $N$ points are compact, and we can choose the subcover with $n = N$.

(2) In contrast, the set $\mathbb{N} \subset \mathbb{R}$ is not compact, because we can cover it with of a suitable union of small open intervals around each point, and this open cover has no finite subcover. For the same reasons, the set $\{1/n | n \in \mathbb{N}\}$ is not compact either.

(3) However, and here comes an interesting point, the following set is compact:

$$K = \left\{ \frac{1}{n} \middle| n \in \mathbb{N} \right\} \cup \{0\}$$

Indeed, any open cover of it $\cup_i U_i$ has to cover 0, and by selecting an open set $U_i$ covering 0, this set $U_i$ will cover the whole $K$, except for finitely many points, due to $1/n \to 0$. But these finitely points left, say $N$ of them, can be covered by suitable sets $U_{i_1}, \ldots, U_{i_N}$, and by adding this family to the set $U_i$, we have our finite subcover.

As a conclusion to this, Definition 7.17 is in tune with what we know about the compact subsets $K \subset \mathbb{R}$, namely that these are the sets which are closed and bounded. However, and here comes our point, such things are wrong in general, due to:

PROPOSITION 7.18. *Given an infinite set $X$ with the discrete distance on it, namely $d(p,q) = 1 - \delta_{pq}$, which can be modeled as the basis of a suitable Hilbert space,*

$$X = \{e_x\}_{x \in X} \subset l^2(X)$$

*this set is closed and bounded, but not compact.*

PROOF. Here the first part, regarding the modeling of $X$, that we will actually not need here, is something that we already know. Regarding now the second part:

(1) $X$ being the total space, it is closed. In fact, since the points of $X$ are open, any subset $E \subset X$ is open, and by taking complements, any set $E \subset X$ is closed as well.

(2) $X$ is also bounded, because all distances are smaller than 1.

(3) However, our set $X$ is not compact, because its points being open, as noted above, $X = \cup_{x \in X} \{x\}$ is an open cover, having no finite subcover. □

Let us develop now the theory of compact sets, and see what we get. We first have:

PROPOSITION 7.19. *The following hold:*

(1) *Compact implies closed.*
(2) *Closed inside compact is compact.*
(3) *Compact intersected with closed is compact.*

PROOF. These assertions are all clear from definitions, as follows:

(1) Assume that $K \subset X$ is compact, and let us prove that $K$ is closed. For this purpose, we will prove that $K^c$ is open. So, pick $p \in K^c$. For any $q \in K$ we set $r = d(p,q)/3$, and we consider the following balls, separating $p$ and $q$:

$$U_q = B_p(r) \quad , \quad V_q = B_q(r)$$

We have then $K \subset \cup_{q \in K} V_q$, so we can pick a finite subcover, as follows:

$$K \subset (V_{q_1} \cup \ldots \cup V_{q_n})$$

With this done, consider the following intersection:

$$U = U_{q_1} \cap \ldots \cap U_{q_n}$$

This intersection is then a ball around $p$, and since this ball avoids $V_{q_1}, \ldots, V_{q_n}$, it avoids the whole $K$. Thus, we have proved that $K^c$ is open at $p$, as desired.

(2) Assume that $F \subset K$ is closed, with $K \subset X$ being compact. For proving our result, we can assume, by replacing $X$ with $K$, that we have $X = K$. In order to prove now that $F$ is compact, consider an open cover of it, as follows:

$$F \subset \bigcup_i U_i$$

By adding the set $F^c$, which is open, to this cover, we obtain a cover of $K$. Now since $K$ is compact, we can extract from this a finite subcover $\Omega$, and there are two cases:

– If $F^c \in \Omega$, by removing $F^c$ from $\Omega$ we obtain a finite cover of $F$, as desired.

– If $F^c \notin \Omega$, we are done too, because in this case $\Omega$ is a finite cover of $F$.

(3) This follows from (1) and (2), because if $K \subset X$ is compact, and $F \subset X$ is closed, then $K \cap F \subset K$ is closed inside a compact, so it is compact. □

As a second batch of results, which are useful as well, we have:

PROPOSITION 7.20. *The following hold:*

(1) *If $K_i \subset X$ are compact, satisfying $K_{i_1} \cap \ldots \cap K_{i_n} \neq \emptyset$, then $\cap_i K_i \neq \emptyset$.*
(2) *If $K_1 \supset K_2 \supset K_3 \supset \ldots$ are non-empty compacts, then $\cap_i K_i \neq \emptyset$.*
(3) *If $K$ is compact, and $E \subset K$ is infinite, then $E$ has a limit point in $K$.*
(4) *If $K$ is compact, any sequence $\{x_n\} \subset K$ has a limit point in $K$.*
(5) *If $K$ is compact, any $\{x_n\} \subset K$ has a subsequence which converges in $K$.*

PROOF. Again, these are elementary results, which can be proved as follows:

(1) Assume by contradiction $\cap_i K_i = \emptyset$, and let us pick $K_1 \in \{K_i\}$. Since any $x \in K_1$ is not in $\cap_i K_i$, there is an index $i$ such that $x \in K_i^c$, and we conclude that we have:

$$K_1 \subset \bigcup_{i \neq 1} K_i^c$$

But this can be regarded as being an open cover of $K_1$, that we know to be compact, so we can extract from it a finite subcover, as follows:

$$K_1 \subset \left(K_{i_1}^c \cup \ldots \cup K_{i_n}^c\right)$$

But this contradicts our non-empty intersection assumption, and we are done.

(2) This is a particular case of (1), proved above.

(3) We prove this by contradiction. So, assume that $E$ has no limit point in $K$. This means that any $p \in K$ can be isolated from the rest of $E$ by a certain open ball $V_p = B_p(r)$, and in both the cases that can appear, $p \in E$ or $p \notin E$, we have:

$$|V_p \cap E| = 0, 1$$

Now observe that these sets $V_p$ form an open cover of $K$, and so of $E$. But due to $|V_p \cap E| = 0, 1$ and to $|E| = \infty$, this open cover of $E$ has no finite subcover. Thus the same cover, regarded now as cover of $K$, has no finite subcover either, contradiction.

(4) This follows from (3) that we just proved, with $E = \{x_n\}$.

(5) This is a reformulation of (4), that we just proved. □

Getting now to more exciting theory, here is a key result about compactness:

THEOREM 7.21. *For a subset $K \subset \mathbb{R}^N$, the following are equivalent:*

(1) *$K$ is closed and bounded.*
(2) *$K$ is compact.*
(3) *Any infinite subset $E \subset K$ has a limiting point in $K$.*

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

(1) $\implies$ (2) As a first task, let us prove that any product of closed intervals is indeed compact. We can assume by linearity that we are dealing with the unit cube:

$$C_1 = \prod_{i=1}^{N} [0, 1] \subset \mathbb{R}^N$$

In order to prove that $C_1$ is compact, we proceed by contradiction. So, assume that we have an open cover as follows, having no finite subcover:

$$C_1 \subset \bigcup_i U_i$$

Now let us cut $C_1$ into $2^N$ small cubes, in the obvious way, over the $N$ coordinate axes. Then at least one of these small cubes, which are all covered by $\cup_i U_i$ too, has no finite subcover. So, let us call $C_2 \subset C_1$ one of these small cubes, having no finite subcover:

$$C_2 \subset \bigcup_i U_i$$

We can then cut $C_2$ into $2^N$ small cubes, and by the same reasoning, we obtain a smaller cube $C_3 \subset C_2$ having no finite subcover. And so on by recurrence, and we end up with a decreasing sequence of cubes, as follows, having no finite subcover:

$$C_1 \supset C_2 \supset C_3 \supset \ldots$$

Now since these decreasing cubes have edge size $1, 1/2, 1/4, \ldots$, their intersection must be a point. So, let us call $p$ this point, defined by the following formula:

$$\{p\} = \bigcap_k C_k$$

But this point $p$ must be covered by $\cup_i U_i$, so we can find an index $i$ such that $p \in U_i$. Now observe that $U_i$ must contain a whole ball around $p$, and so starting from a certain $K \in \mathbb{N}$, all the cubes $C_k$ will be contained in this ball, and so in $U_i$:

$$C_k \subset E_i \quad , \quad \forall k \geq K$$

But this is a contradiction, because $C_K$, and in fact the smaller cubes $C_k$ with $k > K$ as well, were assumed to have no finite subcover. Thus, we proved our claim. But with this claim in hand, the result is now clear. Indeed, assuming that $K \subset \mathbb{R}^N$ is closed and bounded, we can view it as a subset as a suitable big cube, of the following form:

$$K \subset \prod_{i=1}^{N} [-M, M] \subset \mathbb{R}^N$$

But, what we have here is a closed subset inside a compact set, which by Proposition 5.16 follows to be compact, as desired.

(2) $\implies$ (3) This is something that we already know, not needing $K \subset \mathbb{R}^N$.

(3) $\implies$ (1) We have to prove that $K$ as in the statement is both closed and bounded, and we can do both these things by contradiction, as follows:

– Assume first that $K$ is not closed. But this means that we can find a point $x \notin K$ which is a limiting point of $K$. Now let us pick $x_n \in K$, with $x_n \to x$, and consider the set $E = \{x_n\}$. According to our assumption, $E$ must have a limiting point in $K$. But this limiting point can only be $x$, which is not in $K$, contradiction.

– Assume now that $K$ is not bounded. But this means that we can find points $x_n \in K$ satisfying $||x_n|| \to \infty$, and if we consider the set $E = \{x_n\}$, then again this set must have a limiting point in $K$, which is impossible, so we have our contradiction, as desired. □

So long for compactness. As a last piece of general topology, in our metric space framework, we can talk as well about connectedness, as follows:

DEFINITION 7.22. *We can talk about connected sets $E \subset X$, as follows:*

(1) *We say that $E$ is connected if it cannot be separated as $E = E_1 \cup E_2$, with the components $E_1, E_2$ satisfying $E_1 \cap \bar{E}_2 = \bar{E}_1 \cap E_2 = \emptyset$.*
(2) *We say that $E$ is path connected if any two points $p, q \in E$ can be joined by a path, meaning a continuous $f : [0, 1] \to X$, with $f(0) = p$, $f(1) = q$.*

All this looks a bit technical, and indeed it is. To start with, (1) is something quite natural, but the separation condition there $E_1 \cap \bar{E}_2 = \bar{E}_1 \cap E_2 = \emptyset$ can be weakened into $E_1 \cap E_2 = \emptyset$, or strengthened into $\bar{E}_1 \cap \bar{E}_2 = \emptyset$, depeding on purposes, and with our (1) as formulated being the good compromise, for most purposes. As for (2), this condition is obviously something stronger, and we have in fact the following implications:

$$\text{convex} \implies \text{path connected} \implies \text{connected}$$

Anyway, leaving aside the discussion here, once all these questions clarified, the idea is that any set $E$ can be written as a disjoint union of connected components:

$$E = \bigsqcup_i E_i$$

Getting back now to more concrete things, remember that we are here in this book for studying functions, and doing calculus. And, regarding functions, we have:

THEOREM 7.23. *Assuming that $f : X \to Y$ is continuous, the following happen:*

(1) *If $O$ is open, then $f^{-1}(O)$ is open.*
(2) *If $C$ is closed, then $f^{-1}(C)$ is closed.*
(3) *If $K$ is compact, then $f(K)$ is compact.*
(4) *If $E$ is connected, then $f(E)$ is connected.*

PROOF. This is something fundamental, which can be proved as follows:

(1) This is clear from the definition of continuity, written with $\varepsilon, \delta$. In fact, the converse holds too, in the sense that if $f^{-1}(\text{open}) = \text{open}$, then $f$ must be continuous.

(2) This follows from (1), by taking complements. And again, the converse holds too, in the sense that if $f^{-1}(\text{closed}) = \text{closed}$, then $f$ must be continuous.

(3) Indeed, given an open cover $f(K) \subset \cup_i U_i$, we have an open cover $K \subset \cup_i f^{-1}(U_i)$, and so by compactness of $K$, a finite subcover $K \subset f^{-1}(U_{i_1}) \cup \ldots \cup f^{-1}(U_{i_n})$, and so finally a finite subcover $f(K) \subset U_{i_1} \cup \ldots \cup U_{i_n}$, as desired.

(4) This can be proved via the same trick as for (3). Indeed, any separation of $f(E)$ into two parts can be returned via $f^{-1}$ into a separation of $E$ into two parts, contradiction. □

Observe the power of (3,4) in the above result, which among others prove the mean value theorem, which is something non-trivial. Good mathematics that we have here.

In relation now with the notion of uniform continuity, we have the following result:

THEOREM 7.24. *The uniform convergence, $f_n \to_u f$, means that we have $f_n \to f$ with respect to the following distance,*

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

*which is indeed a distance function.*

PROOF. Here the fact that $d$ is indeed a distance, in the sense that it satisfies all the intuitive properties of a distance, including the triangle inequality, follows from definitions, and the fact that the uniform convergence can be interpreted as above is clear as well. □

## 7d. Power series

Power series.

## 7e. Exercises

Exercises:

EXERCISE 7.25.

EXERCISE 7.26.

EXERCISE 7.27.

EXERCISE 7.28.

EXERCISE 7.29.

EXERCISE 7.30.

EXERCISE 7.31.

EXERCISE 7.32.

Bonus exercise.

CHAPTER 8

# Elementary functions

## 8a. Binomial formula

With the above theory in hand, let us get now to some interesting things, namely 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 8.1. *We have the generalized binomial formula*

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

*with the generalized binomial coefficients being given by*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## 8b. Square roots

Quite interestingly, we have as well the following result:

THEOREM 8.2. *The generalized binomial formula, namely*

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

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

$$\sqrt{1+t} = 1 - 2\sum_{k=1}^{\infty} C_{k-1} \left(\frac{-t}{4}\right)^k$$

*with $C_k = \frac{1}{k+1}\binom{2k}{k}$ being the Catalan numbers, and at $a = -1/2$ we obtain*

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

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

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

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

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

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

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

(3) Summarizing, we have proved so far that the binomial formula at $a=\pm 1/2$ is equivalent to the explicit formulae in the statement, involving the Catalan numbers $C_k$, and the central binomial coefficients $D_k$. It remains now to prove that these two explicit formulae hold indeed. For this purpose, let us write these formulae as follows:

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

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

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

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

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

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

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

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

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

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

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

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

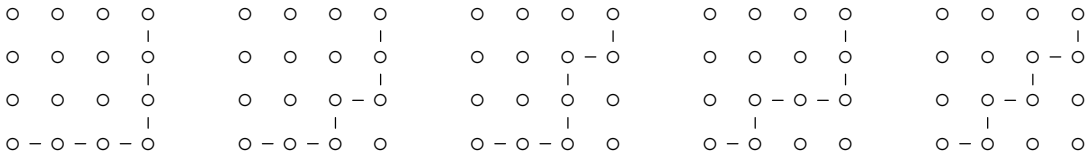
$$\begin{aligned}\sum_{k+l=n+1} D_k D_l &= D_0 D_{n+1} + \sum_{k+l=n} \left(4 - \frac{2}{k+1}\right) D_k D_l \\ &= D_{n+1} + 4 \sum_{k+l=n} D_k D_l - 2 \sum_{k+l=n} \frac{D_k D_l}{k+1} \\ &= D_{n+1} + 4^{n+1} - 2 \sum_{k+l=n} C_k D_l\end{aligned}$$

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

(6) All this suggests a systematic look at the three formulae in (4). According to our various observations above, these three formulae are equivalent, and so it is enough to prove one of them. We will chose here to prove the first one, namely:

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

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



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

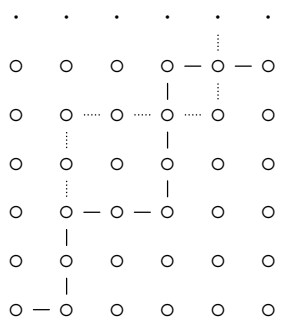
$$1, 1, 2, 5, 14, 42, 132, 429, \ldots$$

(8) We will prove that the numbers $C'_n$ satisfy the recurrence for the numbers $C_n$ that we want to prove, from (6), and on the other hand we will prove that we have $C'_n = C_n$. Getting to work, in what regards our first task, this is easy, because when looking at where our path last intersects the diagonal of the square, we obtain, as desired:

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

(9) In what regards now our second task, proving that we have $C'_n = C_n$, this is more tricky. If we ignore the assumption that our path must stay under the diagonal of the square, we have $\binom{2n}{n}$ such paths. And among these, we have the "good" ones, those that we want to count, and then the "bad" ones, those that we want to ignore.

(10) So, let us count the bad paths, those crossing the diagonal of the square, and reaching the higher diagonal next to it, the one joining $(0,1)$ and $(n,n+1)$. In order to count these, the trick is to "flip" their bad part over that higher diagonal, as follows:



(11) Now observe that, as it is obvious on the above picture, due to the flipping, the flipped bad path will no longer end in $(n,n)$, but rather in $(n-1,n+1)$. Moreover, more is true, in the sense that, by thinking a bit, we see that the flipped bad paths are precisely those ending in $(n-1,n+1)$. Thus, good news, we are done with the count.

(12) To finish now, by putting everything together, we have:

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

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

The generalized binomial formula holds in fact for any exponent $a \in \mathbb{Z}/2$, after some combinatorial pain, and even for any $a \in \mathbb{R}$, but this is non-trivial. More on this later.

## 8c. Catalan numbers

The Catalan numbers $C_k$ that we met above have many other interesting properties:

THEOREM 8.3. *The Catalan numbers $C_k$ count:*

(1) *The length $2k$ loops on $\mathbb{N}$, based at $0$.*
(2) *The noncrossing pairings of $1,\ldots,2k$.*
(3) *The noncrossing partitions of $1,\ldots,k$.*
(4) *The length $2k$ Dyck paths in the plane.*

PROOF. All this is standard combinatorics, the idea being as follows:

(1) To start with, in what regards the various objects involved, the length $2k$ loops on $\mathbb{N}$ are the length $2k$ loops on $\mathbb{N}$ that we know, and the same goes for the noncrossing pairings of $1,\ldots,2k$, and for the noncrossing partitions of $1,\ldots,k$, the idea here being that you must be able to draw the pairing or partition in a noncrossing way.

(2) Thus, we have definitions for all the objects involved, and in each case, if you start counting them, you always end up with the same sequence of numbers, namely:

$$1,2,5,14,42,132,429,1430,4862,16796,58786,\ldots$$

(3) In order to prove now that (1-4) produce indeed the same numbers, many things can be said. The idea is that, leaving aside mathematical brevity, and more specifically abstract reasonings of type $a=b,b=c\implies a=c$, what we have to do, in order to fully understand what is going on, is to etablish $\binom{4}{2}=6$ equalities, via bijective proofs.

(4) However, as a matter of having our theorem formally proved, the point is that, in each of the cases (1-4), the numbers $C_k$ that we get are easily seen to be subject to:

$$C_{k+1}=\sum_{a+b=k}C_aC_b$$

The initial data being the same, namely $C_1=1$ and $C_2=2$, in each of the cases (1-4) under consideration, we get indeed the same numbers. □

Here is as well a useful analytic result, regarding the Catalan numbers:

THEOREM 8.4. *The Catalan numbers have the following properties:*

(1) *They satisfy* $C_{k+1}=\sum_{a+b=k}C_aC_b$.
(2) *The series* $f(z)=\sum_{k\geq 0}C_kz^k$ *satisfies* $zf^2-f+1=0$.
(3) *This series is given by* $f(z)=\frac{1-\sqrt{1-4z}}{2z}$.
(4) *We have the formula* $C_k=\frac{1}{k+1}\binom{2k}{k}$.

PROOF. Consider indeed the generating series $f(z)=\sum_{k\geq 0}C_kz^k$ of the Catalan numbers. In terms of this series, the recurrence relation gives, as desired:

$$\begin{aligned}zf^2&=\sum_{a,b\geq 0}C_aC_bz^{a+b+1}\\&=\sum_{k\geq 1}\sum_{a+b=k-1}C_aC_bz^k\\&=\sum_{k\geq 1}C_kz^k\\&=f-1\end{aligned}$$

By solving the equation $zf^2-f+1=0$ found above, and choosing the solution which is bounded at $z=0$, we obtain the following formula, as claimed:

$$f(z)=\frac{1-\sqrt{1-4z}}{2z}$$

In order to compute $f$, we can use the generalized binomial formula, which gives:

$$\sqrt{1+t}=1-2\sum_{k=1}^{\infty}\frac{1}{k}\binom{2k-2}{k-1}\left(\frac{-t}{4}\right)^k$$

Thus, we obtain the following formula for our series $f$:

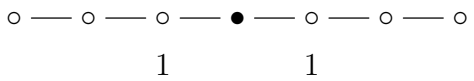
$$\begin{aligned}f(z)&=\frac{1-\sqrt{1-4z}}{2z}\\&=\sum_{k=1}^{\infty}\frac{1}{k}\binom{2k-2}{k-1}z^{k-1}\\&=\sum_{k=0}^{\infty}\frac{1}{k+1}\binom{2k}{k}z^k\end{aligned}$$

So, done. And exercise for you, to work out all the possible permutations of Theorem 8.2, Theorem 8.3 and Theorem 8.4, leading to various approaches to the numbers $C_k$. □

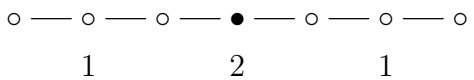
Many other things can be said about the Catalan numbers, mixing algebra, calculus, probability, and more. We will be back to this, on several occasions, in what follows.

## 8d. Special functions

As a continuation of the above material, let us try now to count the length $k$ paths on $\mathbb{Z}$, based at 0. At $k=1$ we have 2 such paths, ending at $-1$ and 1, and the count results can be pictured as follows, with everything being self-explanatory:




At $k=2$ now, we have 4 paths, one of which ends at $-2$, two of which end at 0, and one of which ends at 2. The results can be pictured as follows:




At $k=3$ now, we have 8 paths, the distribution of the endpoints being as follows:

As for $k = 4$, here we have 16 paths, the distribution of the endpoints being as follows:

$$\circ - \circ - \circ - \circ - \circ - \bullet - \circ - \circ - \circ - \circ - \circ$$
$$1 \qquad 4 \qquad 6 \qquad 4 \qquad 1$$

And good news, we can see in the above the Pascal triangle, namely:

$$\begin{array}{ccccccccccc}
 & & & & 1 & & & & & & \\
 & & & 1 & , & 1 & & & & & \\
 & & 1 & , & 2 & , & 1 & & & & \\
 & 1 & , & 3 & , & 3 & , & 1 & & & \\
1 & , & 4 & , & 6 & , & 4 & , & 1 & & \\
1 & , & 5 & , & 10 & , & 10 & , & 5 & , & 1 \\
 & & & & \vdots & & & & & &
\end{array}$$

Thus, eventually, we found the simplest graph ever, namely $\mathbb{Z}$, and we have the following result about it, coming as a complement to what we already know about $\mathbb{N}$:

THEOREM 8.5. *The paths on $\mathbb{Z}$ are counted by the binomial coefficients. In particular, the $2k$-paths based at $0$ are counted by the central binomial coefficients,*

$$\binom{2k}{k} \simeq \frac{4^k}{\sqrt{\pi k}}$$

*with the estimate, in the $k \to \infty$ limit, coming from the Stirling formula.*

PROOF. This basically follows from the above discussion, as follows:

(1) In what regards the count, we certainly have the Pascal triangle, as discovered above, and the rest is just a matter of finishing. There are many possible ways here, a straightforward one being that of arguing that the number $C_k^l$ of length $k$ loops $0 \to l$ is subject, due to the binary choice at the end, to the following recurrence relation:

$$C_k^l = C_{k-1}^{l-1} + C_{k-1}^{l+1}$$

But this is exactly the recurrence for the Pascal triangle, so done with the count.

(2) In what regards the estimate, this follows indeed from Stirling, as follows:

$$\begin{aligned}
\binom{2k}{k} &= \frac{(2k)!}{k!k!} \\
&\simeq \left(\frac{2k}{e}\right)^{2k} \sqrt{4\pi k} \times \left(\frac{e}{k}\right)^{2k} \frac{1}{2\pi k} \\
&= \frac{4^k}{\sqrt{\pi k}}
\end{aligned}$$

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

## 8e. Exercises

Exercises:

Exercise 8.6.

Exercise 8.7.

Exercise 8.8.

Exercise 8.9.

Exercise 8.10.

Exercise 8.11.

Exercise 8.12.

Exercise 8.13.

Bonus exercise.

# Part III

# Derivatives

*Put me up, put me down*
*Put my feet back on the ground*
*Put me up, take my heart*
*And make me happy*

CHAPTER 9

# Derivatives, rules

## 9a. Derivatives

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

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

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

DEFINITION 9.1. *A function $f : \mathbb{R} \to \mathbb{R}$ is called differentiable at $x$ when*

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

*called derivative of $f$ at that point $x$, exists.*

As a first remark, in order for $f$ to be differentiable at $x$, that is to say, in order for the above limit to converge, the numerator must go to 0, as the denominator $t$ does:

$$\lim_{t \to 0} [f(x+t) - f(x)] = 0$$

Thus, $f$ must be continuous at $x$. However, the converse is not true, a basic counterexample being $f(x) = |x|$ at $x = 0$. Let us summarize these findings as follows:

PROPOSITION 9.2. *If $f$ is differentiable at $x$, then $f$ must be continuous at $x$. However, the converse is not true, a basic counterexample being $f(x) = |x|$, at $x = 0$.*

PROOF. The first assertion is something that we already know, from the above. As for the second assertion, regarding $f(x) = |x|$, this is something quite clear on the picture of $f$, but let us prove this mathematically, based on Definition 9.1. We have:

$$\lim_{t \searrow 0} \frac{|0+t| - |0|}{t} = \lim_{t \searrow 0} \frac{t - 0}{t} = 1$$

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

$$\lim_{t \nearrow 0} \frac{|0+t| - |0|}{t} = \lim_{t \nearrow 0} \frac{-t - 0}{t} = -1$$

Thus, the limit in Definition 9.1 does not converge, as desired. □

Generally speaking, the last assertion in Proposition 9.2 should not bother us much, because most of the basic continuous functions are differentiable, and we will see examples in a moment. Before that, however, let us recall why we are here, namely improving the basic estimate $f(x+t)\simeq f(x)$. We can now do this, using the derivative, as follows:

THEOREM 9.3. *Assuming that $f$ is differentiable at $x$, we have:*

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

*In other words, $f$ is, approximately, locally affine at $x$.*

PROOF. Assume indeed that $f$ is differentiable at $x$, and let us set, as before:

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

By multiplying by $t$, we obtain that we have, once again in the $t\to 0$ limit:

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

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

## 9b. Basic examples

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

PROPOSITION 9.4. *We have the differentiation formula*

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

*valid for any exponent $p\in\mathbb{R}$.*

PROOF. We can do this in three steps, as follows:

(1) In the case $p\in\mathbb{N}$ we can use the binomial formula, which gives, as desired:

$$\begin{aligned}(x+t)^p&=\sum_{k=0}^{n}\binom{p}{k}x^{p-k}t^k\\&=x^p+px^{p-1}t+\ldots+t^p\\&\simeq x^p+px^{p-1}t\end{aligned}$$

(2) Let us discuss now the general case $p\in\mathbb{Q}$. We write $p=m/n$, with $m\in\mathbb{Z}$ and $n\in\mathbb{N}$. In order to do the computation, we use the following formula:

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

We set in this formula $a=(x+t)^{m/n}$ and $b=x^{m/n}$. We obtain, as desired:

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

(3) In the general case now, where $p\in\mathbb{R}$ is real, we can use a similar argument. Indeed, given any integer $n\in\mathbb{N}$, we have the following computation:

$$\begin{aligned}(x+t)^p-x^p &= \frac{(x+t)^{pn}-x^{pn}}{(x+t)^{p(n-1)}+\ldots+x^{p(n-1)}}\\ &\simeq \frac{(x+t)^{pn}-x^{pn}}{nx^{p(n-1)}}\end{aligned}$$

Now observe that we have the following estimate, with [.] being the integer part:

$$(x+t)^{[pn]}\leq(x+t)^{pn}\leq(x+t)^{[pn]+1}$$

By using the binomial formula on both sides, for the integer exponents $[pn]$ and $[pn]+1$ there, we deduce that with $n>>0$ we have the following estimate:

$$(x+t)^{pn}\simeq x^{pn}+pnx^{pn-1}t$$

Thus, we can finish our computation started above as follows:

$$(x+t)^p-x^p\simeq\frac{pnx^{pn-1}t}{nx^{pn-p}}=px^{p-1}t$$

But this gives $(x^p)'=px^{p-1}$, which finishes the proof. □

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

PROPOSITION 9.5. *We have the following results:*

(1) $(\sin x)'=\cos x$.
(2) $(\cos x)'=-\sin x$.
(3) $(e^x)'=e^x$.
(4) $(\log x)'=x^{-1}$.

PROOF. This is quite tricky, as always when computing derivatives, as follows:

(1) Regarding sin, the computation here goes as follows:

$$\begin{aligned}(\sin x)' &= \lim_{t\to 0}\frac{\sin(x+t)-\sin x}{t}\\ &= \lim_{t\to 0}\frac{\sin x\cos t+\cos x\sin t-\sin x}{t}\\ &= \lim_{t\to 0}\sin x\cdot\frac{\cos t-1}{t}+\cos x\cdot\frac{\sin t}{t}\\ &= \cos x\end{aligned}$$

Here we have used the fact, which is clear on pictures, by drawing the trigonometric circle, that we have $\sin t\simeq t$ for $t\simeq 0$, plus the fact, which follows from this and from Pythagoras, $\sin^2+\cos^2=1$, that we have as well $\cos t\simeq 1-t^2/2$, for $t\simeq 0$.

(2) The computation for cos is similar, as follows:

$$\begin{aligned}(\cos x)' &= \lim_{t\to 0}\frac{\cos(x+t)-\cos x}{t}\\ &= \lim_{t\to 0}\frac{\cos x\cos t-\sin x\sin t-\cos x}{t}\\ &= \lim_{t\to 0}\cos x\cdot\frac{\cos t-1}{t}-\sin x\cdot\frac{\sin t}{t}\\ &= -\sin x\end{aligned}$$

(3) For the exponential, the derivative can be computed as follows:

$$\begin{aligned}(e^x)' &= \left(\sum_{k=0}^{\infty}\frac{x^k}{k!}\right)'\\ &= \sum_{k=0}^{\infty}\frac{kx^{k-1}}{k!}\\ &= e^x\end{aligned}$$

(4) As for the logarithm, the computation here is as follows, using $\log(1+y)\simeq y$ for $y\simeq 0$, which follows from $e^y\simeq 1+y$ that we found in (3), by taking the logarithm:

$$\begin{aligned}(\log x)' &= \lim_{t\to 0}\frac{\log(x+t)-\log x}{t}\\ &= \lim_{t\to 0}\frac{\log(1+t/x)}{t}\\ &= \frac{1}{x}\end{aligned}$$

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

Speaking exponentials, we can now formulate a nice result about them:

THEOREM 9.6. *The exponential function, namely*

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

*is the unique power series satisfying $f' = f$ and $f(0) = 1$.*

PROOF. Consider indeed a power series satisfying $f' = f$ and $f(0) = 1$. Due to $f(0) = 1$, the first term must be 1, and so our function must look as follows:

$$f(x) = 1 + \sum_{k=1}^{\infty} c_k x^k$$

According to our differentiation rules, the derivative of this series is given by:

$$f(x) = \sum_{k=1}^{\infty} k c_k x^{k-1}$$

Thus, the equation $f' = f$ is equivalent to the following equalities:

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

But this system of equations can be solved by recurrence, as follows:

$$c_1 = 1 \quad , \quad c_2 = \frac{1}{2} \quad , \quad c_3 = \frac{1}{2 \times 3} \quad , \quad c_4 = \frac{1}{2 \times 3 \times 4} \quad , \quad \ldots$$

Thus we have $c_k = 1/k!$, leading to the conclusion in the statement. □

Observe that the above result leads to a more conceptual explanation for the number $e$ itself. To be more precise, $e \in \mathbb{R}$ is the unique number satisfying:

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

Which is very nice, at least we know one thing.

## 9c. Theorems, rules

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

THEOREM 9.7. *We have the following formulae:*

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

PROOF. All these formulae are elementary, the idea being as follows:

(1) This follows indeed from definitions, the computation being as follows:

$$\begin{aligned}(f+g)'(x) &= \lim_{t\to 0}\frac{(f+g)(x+t)-(f+g)(x)}{t}\\ &= \lim_{t\to 0}\left(\frac{f(x+t)-f(x)}{t}+\frac{g(x+t)-g(x)}{t}\right)\\ &= \lim_{t\to 0}\frac{f(x+t)-f(x)}{t}+\lim_{t\to 0}\frac{g(x+t)-g(x)}{t}\\ &= f'(x)+g'(x)\end{aligned}$$

(2) This follows from definitions too, the computation, by using the more convenient formula $f(x+t)\simeq f(x)+f'(x)t$ as a definition for the derivative, being as follows:

$$\begin{aligned}(fg)(x+t) &= f(x+t)g(x+t)\\ &\simeq (f(x)+f'(x)t)(g(x)+g'(x)t)\\ &\simeq f(x)g(x)+(f'(x)g(x)+f(x)g'(x))t\end{aligned}$$

Indeed, we obtain from this that the derivative is the coefficient of $t$, namely:

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

(3) Regarding compositions, the computation here is as follows, again by using the more convenient formula $f(x+t)\simeq f(x)+f'(x)t$ as a definition for the derivative:

$$\begin{aligned}(f\circ g)(x+t) &= f(g(x+t))\\ &\simeq f(g(x)+g'(x)t)\\ &\simeq f(g(x))+f'(g(x))g'(x)t\end{aligned}$$

Indeed, we obtain from this that the derivative is the coefficient of $t$, namely:

$$(f\circ g)'(x)=f'(g(x))g'(x)$$

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

We can of course combine the above formulae, and we obtain for instance:

PROPOSITION 9.8. *The derivatives of fractions are given by:*

$$\left(\frac{f}{g}\right)'=\frac{f'g-fg'}{g^2}$$

*In particular, we have the following formula, for the derivative of inverses:*

$$\left(\frac{1}{f}\right)'=-\frac{f'}{f^2}$$

*In fact, we have $(f^p)'=pf^{p-1}$, for any exponent $p\in\mathbb{R}$.*

PROOF. This statement is written a bit upside down, and for the proof it is better to proceed backwards. To be more precise, by using $(x^p)' = px^{p-1}$ and Theorem 9.7 (3), we obtain the third formula. Then, with $p = -1$, we obtain from this the second formula. And finally, by using this second formula and Theorem 9.7 (2), we obtain:

$$\begin{aligned}\left(\frac{f}{g}\right)' &= \left(f \cdot \frac{1}{g}\right)' \\ &= f' \cdot \frac{1}{g} + f\left(\frac{1}{g}\right)' \\ &= \frac{f'}{g} - \frac{fg'}{g^2} \\ &= \frac{f'g - fg'}{g^2}\end{aligned}$$

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

All the above might seem to start to be a bit too complex, with too many things to be memorized and so on, and as a piece of advice here, we have:

ADVICE 9.9. *Memorize and cherish the formula for fractions*

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$$

*along with the usual addition formula, that you know well*

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

*and generally speaking, never mess with fractions.*

With this coming from a lifelong calculus teacher and scientist, mathematics can be difficult, and many things can be pardoned, but not messing with fractions. And with this going beyond mathematics too, say if you want to make a living by selling apples or tomatoes at the market, fine, but you'll need to know well fractions, trust me.

Back to work now, with the above formulae in hand, we can do all sorts of computations for other basic functions that we know, including $\tan x$, or $\arctan x$:

PROPOSITION 9.10. *We have the following formulae,*

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

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

PROOF. For tan, we have the following computation:

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

As for arctan, we can use here the following computation:

$$\begin{aligned}(\tan\circ\arctan)'(x) &= \tan'(\arctan x)\arctan'(x) \\ &= \frac{1}{\cos^2(\arctan x)}\arctan'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x'=1$, we obtain from this:

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

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

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

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

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

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

## 9d. Local extrema

At the theoretical level now, further building on Theorem 9.3, we have:

THEOREM 9.11. *The local minima and maxima of a differentiable function $f:\mathbb{R}\to\mathbb{R}$ appear at the points $x\in\mathbb{R}$ where:*

$$f'(x)=0$$

*However, the converse of this fact is not true in general.*

PROOF. The first assertion follows from the formula $f(x+t)\simeq f(x)+f'(x)t$. Indeed, let us rewrite this formula, more conveniently, in the following way:

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

Now saying that our function $f$ has a local maximum at $x\in\mathbb{R}$ means that there exists a number $\varepsilon>0$ such that the following happens:

$$f(x+t)\geq f(x)\quad,\quad\forall t\in[-\varepsilon,\varepsilon]$$

We conclude that we must have $f'(x)t \geq 0$ for sufficiently small $t$, and since this small $t$ can be both positive or negative, this gives, as desired:

$$f'(x) = 0$$

Similarly, saying that our function $f$ has a local minimum at $x \in \mathbb{R}$ means that there exists a number $\varepsilon > 0$ such that the following happens:

$$f(x+t) \leq f(x) \quad , \quad \forall t \in [-\varepsilon, \varepsilon]$$

Thus $f'(x)t \leq 0$ for small $t$, and this gives, as before, $f'(x) = 0$. Finally, in what regards the converse, the simplest counterexample here is the following function:

$$f(x) = x^3$$

Indeed, we have $f'(x) = 3x^2$, and in particular $f'(0) = 0$. But our function being clearly increasing, $x = 0$ is not a local maximum, nor a local minimum. □

As an important consequence of Theorem 9.11, we have:

THEOREM 9.12. *Assuming that $f : [a, b] \to \mathbb{R}$ is differentiable, we have*

$$\frac{f(b) - f(a)}{b - a} = f'(c)$$

*for some $c \in (a, b)$, called mean value property of $f$.*

PROOF. In the case $f(a) = f(b)$, the result, called Rolle theorem, states that we have $f'(c) = 0$ for some $c \in (a, b)$, and follows from Theorem 9.11. Now in what regards our statement, due to Lagrange, this follows from Rolle, applied to the following function:

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

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

In practice, Theorem 9.11 can be used in order to find the maximum and minimum of any differentiable function, and this method is best recalled as follows:

ALGORITHM 9.13. *In order to find the minimum and maximum of $f : [a, b] \to \mathbb{R}$:*

(1) *Compute the derivative $f'$.*
(2) *Solve the equation $f'(x) = 0$.*
(3) *Add $a, b$ to your set of solutions.*
(4) *Compute $f(x)$, for all your solutions.*
(5) *Compute the min/max of all these $f(x)$ values.*
(6) *Then this is the min/max of your function.*

Needless to say, all this is very interesting, and powerful. The general problem in any type of applied mathematics is that of finding the minimum or maximum of some function, and we have now an algorithm for dealing with such questions. Very nice.

## 9e. Exercises

Exercises:

Exercise 9.14.

Exercise 9.15.

Exercise 9.16.

Exercise 9.17.

Exercise 9.18.

Exercise 9.19.

Exercise 9.20.

Exercise 9.21.

Bonus exercise.

CHAPTER 10

# Second derivatives

## 10a. Second derivatives

The derivative theory that we have is already quite powerful, and can be used in order to solve all sorts of interesting questions, but with a bit more effort, we can do better. Indeed, at a more advanced level, we can come up with the following notion:

DEFINITION 10.1. *We say that $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable if it is differentiable, and its derivative $f' : \mathbb{R} \to \mathbb{R}$ is differentiable too. The derivative of $f'$ is denoted*

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

*and is called second derivative of $f$.*

You might probably wonder why coming with this definition, which looks a bit abstract and complicated, instead of further developing the theory of the first derivative, which looks like something very reasonable and useful. Good point, and answer to this coming in a moment. But before that, let us get a bit familiar with $f''$. We first have:

INTERPRETATION 10.2. *The second derivative $f''(x) \in \mathbb{R}$ is the number which:*

(1) *Expresses the growth rate of the slope $f'(z)$ at the point $x$.*
(2) *Gives us the acceleration of the function $f$ at the point $x$.*
(3) *Computes how much different is $f(x)$, compared to $f(z)$ with $z \simeq x$.*
(4) *Tells us how much convex or concave is $f$, around the point $x$.*

So, this is the truth about the second derivative, making it clear that what we have here is a very interesting notion. In practice now, the situation is as follows:

(1) This is something very intuitive, which follows from the usual interpretation of the derivative, both as a growth rate, and a slope.

(2) This is some sort of reformulation of (1), using the intuitive meaning of the word "acceleration", with the relevant physics equations, due to Newton, being as follows:

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

To be more precise, here $x, v, a$ are the position, speed and acceleration, and the dot denotes the time derivative, and according to these equations, we have $a = \ddot{x}$, second derivative. We will be back to these equations later, in chapter 12 below.

(3) This is something more subtle, and very useful too, which is of statistical nature, and that we will clarify with some mathematics, in a moment.

(4) This is something quite subtle too, and again very useful in practice, that we will again clarify with some mathematics, later on this chapter.

All in all, what we have above is a mixture of trivial and non-trivial facts, and do not worry, we will get familiar with all this, in the next few pages.

## 10b. Basic examples

In practice now, in order to get familiar with the second derivatives, let us first compute the second derivatives of the functions that we are familiar with, see what we get.

The result here, which is perhaps not very enlightening at this stage of things, but which certainly looks technically useful, is as follows:

THEOREM 10.3. *The second derivatives of the basic functions are as follows:*

(1) $(x^p)'' = p(p-1)x^{p-2}$.
(2) $\sin'' = -\sin$.
(3) $\cos'' = -\cos$.
(4) $\exp'' = \exp$.
(5) $\log''(x) = -1/x^2$.

PROOF. The various formulae in the statement all follow from the various formulae for the derivatives established before, as follows:

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

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

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

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

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

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

At a more specialized level now, let us record as well the following result:

PROPOSITION 10.4. *We have the following second derivative formulae,*

$$(\tan x)'' = \frac{2\sin x}{\cos^3 x} \quad , \quad (\arctan x)'' = -\frac{2x}{(1+x^2)^2}$$

*and the second derivatives of the remaining trigonometric functions can be computed too.*

PROOF. Regarding the tangent, we have here the following computation:

$$\begin{aligned}(\tan x)'' &= \left(\frac{1}{\cos^2 x}\right)' \\ &= -\frac{(\cos^2 x)'}{\cos^4 x} \\ &= \frac{2\cos x\sin x}{\cos^4 x} \\ &= \frac{2\sin x}{\cos^3 x}\end{aligned}$$

As for the arctangent, the computation here is as follows:

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

Finally, we will leave the last assertion as an instructive exercise. □

At a theoretical level now, let us record the following result:

PROPOSITION 10.5. *There are functions which are differentiable, but not twice differentiable.*

PROOF. In order to construct a counterexample, recall first that the simplest example of a function which is continuous, but not differentiable, was $f(x) = |x|$, the idea behind this being to use a "piecewise linear function whose branches do not fit well". In connection now with our question, piecewise linear will not do, but we can use a similar idea, namely "piecewise quadratic function whose branches do not fit well". So, let us set:

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

This function is then differentiable, with its derivative being:

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

Now for getting our counterexample, we can set $a = -1, b = 1$, so that $f$ is:

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

Indeed, the derivative is $f'(x) = 2|x|$, which is not differentiable, as desired. □

## 10c. Taylor formula

Getting now to theory, our main purpose will be that of improving, with the help of the second derivative, the basic approximation formula for functions, namely:

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

In order to do so, things will be quite tricky, and a bit more geometric, and perhaps less intuitive, than before. We will be in need of the following standard result:

THEOREM 10.6. *The* $0/0$ *type limits can be computed according to the formula*

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

*called L'Hôpital's rule.*

PROOF. The above formula holds indeed, as an application of the general derivative theory from chapter 9, which gives, in the situation from the statement:

$$\begin{aligned}\frac{f(x+t)}{g(x+t)} &\simeq \frac{f(x)+f'(x)t}{g(x)+g'(x)t}\\ &= \frac{f'(x)t}{g'(x)t}\\ &= \frac{f'(x)}{g'(x)}\end{aligned}$$

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

We can now formulate the following key result:

THEOREM 10.7. *Any twice differentiable function* $f : \mathbb{R} \to \mathbb{R}$ *is locally quadratic,*

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

*with* $f''(x)$ *being as usual the derivative of the function* $f' : \mathbb{R} \to \mathbb{R}$ *at the point* $x$*.*

PROOF. 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 chapter 9 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 approximation 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 can use L'Hôpital's rule. Indeed, 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. □

As a first application of this, justifying Interpretation 10.2 (3), we have the following statement, which is a bit heuristic, but we will call it however Proposition:

PROPOSITION 10.8. *Intuitively speaking, the second derivative $f''(x) \in \mathbb{R}$ computes how much different is $f(x)$, compared to the average of $f(z)$, with $z \simeq x$.*

PROOF. As already mentioned, this is something a bit heuristic, but which is good to know. Let us write the formula in Theorem 10.7, as such, and with $t \to -t$ too:

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

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

By making the average, we obtain the following formula:

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

Now assume that we have found a way of averaging things over $t \in [-\varepsilon, \varepsilon]$, with the corresponding averages being denoted $I$. We obtain from the above:

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

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

Back now to rigorous mathematics, we have the following key result:

THEOREM 10.9. *The local minima and local maxima of a twice differentiable function $f : \mathbb{R} \to \mathbb{R}$ appear at the points $x \in \mathbb{R}$ where*

$$f'(x) = 0$$

*with the local minima corresponding to the case $f''(x) \geq 0$, and with the local maxima corresponding to the case $f''(x) \leq 0$.*

PROOF. The first assertion is something that we already know. As for the second assertion, we can use the formula in Theorem 10.7, which in the case $f'(x) = 0$ reads:

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

Indeed, assuming $f''(x) \neq 0$, it is clear that the condition $f''(x) > 0$ will produce a local minimum, and that the condition $f''(x) < 0$ will produce a local maximum. □

As before with derivatives, the above result is not the end of the story with the mathematics of the local minima and maxima, because things are undetermined when:

$$f'(x) = f''(x) = 0$$

For instance the functions $\pm x^n$ with $n \in \mathbb{N}$ all satisfy this condition at $x = 0$, which is a minimum for the functions of type $x^{2m}$, a maximum for the functions of type $-x^{2m}$, and not a local minimum or local maximum for the functions of type $\pm x^{2m+1}$.

We will be back to such questions in the next chapter, with a full, more advanced discussion about this, by using higher derivatives.

There are some comments to be made as well in relation with the algorithm presented at the end of the previous chapter, for finding the extrema of the function. Normally that algorithm stays strong, because Theorem 10.9 can only help in relation with the final steps, and is it worth it to compute the second derivative $f''$, just for getting rid of roughly $1/2$ of the $f(x)$ values to be compared. However, in certain cases, this method proves to be useful, so Theorem 10.9 is good to know, when applying that algorithm.

Again, we will be back to such questions in the next chapter, with a full, more advanced discussion about all this, by using higher derivatives.

## 10d. Convex functions

As a main concrete application now of the second derivative, which is something very useful in practice, and related to Interpretation 10.2 (4), we have the following result:

THEOREM 10.10. *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 $\lambda_i$ to be all equal, we obtain the following Jensen inequality, valid for any $x_1, \ldots, x_N \in \mathbb{R}$,*

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

*and once again with equality when $x_1 = \ldots = x_N$. A similar statement holds for the concave functions, with all the inequalities being reversed.*

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

(1) First, we can talk about convex functions in a usual, intuitive way, with this meaning by definition that the following inequality must be satisfied:

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x) + f(y)}{2}$$

(2) But this means, via a simple argument, by approximating numbers $t \in [0, 1]$ by sums of powers $2^{-k}$, that for any $t \in [0, 1]$ we must have:

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

Alternatively, via yet another simple argument, this time by doing some geometry with triangles, this means that we must have:

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

But then, again alternatively, by combining the above two simple arguments, the following must happen, for any $\lambda_1, \ldots, \lambda_N > 0$ summing up to 1:

$$f(\lambda_1 x_1 + \ldots + \lambda_N x_N) \leq \lambda_1 f(x_1) + \ldots + \lambda_N x_N$$

(3) Summarizing, all our Jensen inequalities, at $N = 2$ and at $N \in \mathbb{N}$ arbitrary, are equivalent. The point now is that, if we look at what the first Jensen inequality, that we took as definition for the convexity, exactly means, this is simply equivalent to:

$$f''(x) \geq 0$$

(4) Thus, we are led to the conclusions in the statement, regarding the convex functions. As for the concave functions, the proof here is similar. Alternatively, we can say that $f$ is concave precisely when $-f$ is convex, and get the results from what we have. □

As a basic application of the Jensen inequality, which is very classical, we have:

THEOREM 10.11. *For any $p \in (1,\infty)$ we have the following inequality,*

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

*and for any $p \in (0,1)$ we have the following inequality,*

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

*with in both cases equality precisely when $|x_1| = \ldots = |x_N|$.*

PROOF. This follows indeed from Theorem 10.10, because we have:

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

Thus $x^p$ is convex for $p > 1$ and concave for $p < 1$, which gives the results. □

Observe that at $p = 2$ we obtain as particular case of the above inequality the Cauchy-Schwarz inequality, or rather something equivalent to it, namely:

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

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

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

THEOREM 10.12. *We have the Young inequality,*

$$ab \leq \frac{a^p}{p}+\frac{b^q}{q}$$

*valid for any $a,b \geq 0$, and any exponents $p,q > 1$ satisfying $\frac{1}{p}+\frac{1}{q} = 1$.*

PROOF. We use the logarithm function, which is concave on $(0,\infty)$, due to:

$$(\log x)'' = \left(-\frac{1}{x}\right)' = -\frac{1}{x^2}$$

Thus we can apply the Jensen inequality, and we obtain in this way:

$$\begin{aligned}\log\left(\frac{a^p}{p}+\frac{b^q}{q}\right) &\geq \frac{\log(a^p)}{p}+\frac{\log(b^q)}{q}\\ &= \log(a)+\log(b)\\ &= \log(ab)\end{aligned}$$

Now by exponentiating, we obtain the Young inequality. □

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

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

In general, the Young inequality is something non-trivial, and the idea with it is that "when stuck with a problem, and with $ab \leq \frac{a^2+b^2}{2}$ not working, try Young".

Moving forward now, as a consequence of the Young inequality, we have:

THEOREM 10.13 (Hölder). *Assuming that $p, q \geq 1$ are conjugate, in the sense that*

$$\frac{1}{p} + \frac{1}{q} = 1$$

*we have the following inequality, valid for any two vectors $x, y \in \mathbb{C}^N$,*

$$\sum_i |x_i y_i| \leq \left( \sum_i |x_i|^p \right)^{1/p} \left( \sum_i |y_i|^q \right)^{1/q}$$

*with the convention that an $\infty$ exponent produces a $\max |x_i|$ quantity.*

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

(1) Assume first that we are dealing with finite exponents, $p, q \in (1, \infty)$. By linearity we can assume that $x, y$ are normalized, in the following way:

$$\sum_i |x_i|^p = \sum_i |y_i|^q = 1$$

In this case, we want to prove that the following inequality holds:

$$\sum_i |x_i y_i| \leq 1$$

For this purpose, we use the Young inequality, which gives, for any $i$:

$$|x_i y_i| \leq \frac{|x_i|^p}{p} + \frac{|y_i|^q}{q}$$

By summing now over $i = 1, \ldots, N$, we obtain from this, as desired:

$$\begin{aligned} \sum_i |x_i y_i| &\leq \sum_i \frac{|x_i|^p}{p} + \sum_i \frac{|y_i|^q}{q} \\ &= \frac{1}{p} + \frac{1}{q} \\ &= 1 \end{aligned}$$

(2) In the case $p = 1$ and $q = \infty$, or vice versa, the inequality holds too, trivially, with the convention that an $\infty$ exponent produces a max quantity, according to:

$$\lim_{p\to\infty}\left(\sum_i |x_i|^p\right)^{1/p} = \max|x_i|$$

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

As a consequence now of the Hölder inequality, we have:

THEOREM 10.14 (Minkowski). *Assuming $p \in [1,\infty]$, we have the inequality*

$$\left(\sum_i |x_i + y_i|^p\right)^{1/p} \leq \left(\sum_i |x_i|^p\right)^{1/p} + \left(\sum_i |y_i|^p\right)^{1/p}$$

*for any two vectors $x, y \in \mathbb{C}^N$, with our usual conventions at $p = \infty$.*

PROOF. We have indeed the following estimate, using the Hölder inequality, and the conjugate exponent $q \in [1,\infty]$, given by $1/p + 1/q = 1$:

$$\begin{aligned}
\sum_i |x_i + y_i|^p &= \sum_i |x_i + y_i| \cdot |x_i + y_i|^{p-1}\\
&\leq \sum_i |x_i| \cdot |x_i + y_i|^{p-1} + \sum_i |y_i| \cdot |x_i + y_i|^{p-1}\\
&\leq \left(\sum_i |x_i|^p\right)^{1/p}\left(\sum_i |x_i + y_i|^{(p-1)q}\right)^{1/q}\\
&+ \left(\sum_i |y_i|^p\right)^{1/p}\left(\sum_i |x_i + y_i|^{(p-1)q}\right)^{1/q}\\
&= \left[\left(\sum_i |x_i|^p\right)^{1/p} + \left(\sum_i |y_i|^p\right)^{1/p}\right]\left(\sum_i |x_i + y_i|^p\right)^{1-1/p}
\end{aligned}$$

Here we have used the following fact, at the end:

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

Now by dividing both sides by the last quantity at the end, we obtain:

$$\left(\sum_i |x_i + y_i|^p\right)^{1/p} \leq \left(\sum_i |x_i|^p\right)^{1/p} + \left(\sum_i |y_i|^p\right)^{1/p}$$

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

Good news, done with inequalities, and as a consequence of the above results, and more specifically of the Minkowski inequality obtained above, we can formulate:

THEOREM 10.15. *Given an exponent $p \in [1, \infty]$, the formula*

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

*with usual conventions at $p = \infty$, defines a norm on $\mathbb{C}^N$.*

PROOF. Man things can be said here, the idea being as follows:

(1) To start with, the normed space assertion follows from the Minkowski inequality, established above, which proves indeed the triangle inequality, for our norm.

(2) As a technical remark, the cases $p = 1, 2, \infty$, which are of particular interest, in practice, do not really need all the above, and exercise for you, to figure out all this.

(3) To be more precise, the cases $p = 1, \infty$ are elementary, and $p = 2$ only needs the knowledge of the Cauchy-Schwarz inequality, which is something elementary too.

(4) Finally, many interesting things can be said, about the spaces that we constructed. We will be back to this, on several occasions, in what follows. □

Very nice all this, but you might wonder at this point, what is the relation of all this with functions. In answer, Theorem 10.15 can be reformulated as follows:

THEOREM 10.16. *Given an exponent $p \in [1, \infty]$, the formula*

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

*defines a norm on the space of functions $f : \{1, \ldots, N\} \to \mathbb{C}$.*

PROOF. This is a just fancy reformulation of Theorem 10.15, by using the standard fact, that you might know from linear algebra, that the space formed by the functions $f : \{1, \ldots, N\} \to \mathbb{C}$ is canonically isomorphic to $\mathbb{C}^N$, in the obvious way, as follows:

$$f \to \begin{pmatrix} f(1) \\ \vdots \\ f(N) \end{pmatrix}$$

Moreover, in this picture, we must replace the sums from the $\mathbb{C}^N$ context, on the right, with integrals with respect to the counting measure on $\{1, \ldots, N\}$, in the function context, on the left. Thus, we are led to the conclusion in the statement. □

Finally, let us mention that there are functional versions of the above inequalities, meaning with numbers replaced by functions. We will discuss this, later in this book.

## 10e. Exercises

Exercises:

Exercise 10.17.

Exercise 10.18.

Exercise 10.19.

Exercise 10.20.

Exercise 10.21.

Exercise 10.22.

Exercise 10.23.

Exercise 10.24.

Bonus exercise.

CHAPTER 11

# Taylor formula

## 11a. Taylor formula

Back now to the general theory of the derivatives, and their theoretical applications, we can further develop our basic approximation method, at order 3, at order 4, and so on. Let us start with something nice and intuitive, as follows:

FACT 11.1. *Third derivatives, and the jerk.*

Here the terminology comes from real life and classical mechanics, where the jerk is by definition the derivative of the acceleration, and so is the second derivative of the speed, and so is the third derivative of the position, according to the following formulae:

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

As before with second derivatives, many other things can be said. Let us also record the formulae of the third derivatives of the basic functions, which are as follows:

THEOREM 11.2. *The third derivatives of the basic functions are as follows:*

(1) $(x^p)''' = p(p-1)(p-2)x^{p-3}$.
(2) $\sin''' = -\cos$.
(3) $\cos''' = \sin$.
(4) $\exp''' = \exp$.
(5) $\log'''(x) = 2/x^3$.

PROOF. The various formulae in the statement all follow from the various formulae for the second derivatives established before, as follows:

$$(x^p)''' = (p(p-1)x^{p-2})' = p(p-1)(p-2)x^{p-3}$$

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

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

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

$$(\log x)''' = (-1/x^2)' = 2/x^3$$

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

Getting now to the fourth derivatives, things are less intuitive here, in what regards the interpretation, but we can nevertheless do some computations, as follows:

THEOREM 11.3. *The fourth derivatives of the basic functions are as follows:*

(1) $(x^p)'''' = p(p-1)(p-2)(p-3)x^{p-4}$.
(2) $\sin'''' = \sin$.
(3) $\cos'''' = \cos$.
(4) $\exp'''' = \exp$.
(5) $\log''''(x) = -6/x^4$.

PROOF. The various formulae in the statement all follow from the various formulae for the third derivatives established before, as follows:

$$(x^p)'''' = (p(p-1)(p-2)x^{p-3})' = p(p-1)(p-2)(p-3)x^{p-4}$$

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

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

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

$$(\log x)'''' = (2/x^3)' = -6/x^4$$

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

Observe the magic brought by the fourth derivative at the level of basic trigonometric functions. This is perhaps something worth recording, as follows:

OBSERVATION 11.4. *The fourth derivative brings some periodicity magic at the level of basic trigonometric functions.*

With this discussed, and getting back now to our usual approximation business, the ultimate result on the subject, called Taylor formula, is as follows:

THEOREM 11.5. *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 $\varphi$ 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 can use L'Hôpital's rule, applied several times, as in the proof at order 2. To be more precise, if we denote by $\varphi(t) \simeq P(t)$ the approximation to be proved, we have:

$$\begin{aligned}
\frac{\varphi(t) - P(t)}{t^n} &\simeq \frac{\varphi'(t) - P'(t)}{nt^{n-1}} \\
&\simeq \frac{\varphi''(t) - P''(t)}{n(n-1)t^{n-2}} \\
&\vdots \\
&\simeq \frac{\varphi^{(n)}(t) - P^{(n)}(t)}{n!} \\
&= \frac{f^{(n)}(x) - f^{(n)}(x)}{n!} \\
&= 0
\end{aligned}$$

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

## 11b. The remainder

Here is a related interesting statement, inspired from the above proof:

PROPOSITION 11.6. *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 $n$. □

Many other things can be said, about the remainder. We will be back to this.

## 11c. Local extrema

In relation with the local extrema, we have the following result:

THEOREM 11.7. *The one-variable smooth functions are subject to the Taylor formula*

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

*which allows, via suitable truncations, to determine the local maxima and minima.*

PROOF. This is a compact summary of what we know from the above, with everything being in fact quite technical, and with the idea being as follows:

(1) In order to compute the local maxima and minima, a first method is by using the following formula, which comes straight from the definition of the derivative:

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

Indeed, this formula shows that when $f'(x)\neq 0$, the point $x$ cannot be a local minimum or maximum, due to the fact that $t\to -t$ will invert the growth.

(2) In relation with the problems left, the second derivative comes to the rescue. Indeed, we can use the following more advanced formula, coming via l'Hôpital's rule:

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

To be more precise, assume that we have $f'(x)=0$, as required by the study in (1). Then this second order formula simply reads:

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

But this is something very useful, telling us that when $f''(x)<0$, what we have is a local maximum, and when $f''(x)>0$, what we have is a local minimum. As for the remaining case, that when $f''(x)=0$, things here remain open.

(3) All this is very useful in practice, and with what we have in (1), complemented if needed with what we have in (2), we can in principle compute the local minima and maxima, without much troubles. However, if really needed, more tools are available. Indeed, we can use if we want the order 3 Taylor formula, which is as follows:

$$f(x+t)\simeq f(x)+f'(x)t+\frac{f''(x)}{2}\,t^2+\frac{f'''(x)}{6}\,t^3$$

To be more precise, assume that we are in the case $f'(x)=f''(x)=0$, which is where our joint algorithm coming from (1) and (2) fails. In this case, our formula becomes:

$$f(x+t)\simeq f(x)+\frac{f'''(x)}{6}\,t^3$$

But this solves the problem in the case $f'''(x) \neq 0$, because here we cannot have a local minimum or maximum, due to $t \to -t$ which switches growth. As for the remaining case, $f'''(x) = 0$, things here remain open, and we have to go at higher order.

(4) Summarizing, we have a recurrence method for solving our problem. In order to formulate now an abstract result about this, we can use the Taylor formula at order $n$:

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

Indeed, assume that we started to compute the derivatives $f'(x), f''(x), f''(x), \ldots$ of our function at the point $x$, with the goal of finding the first such derivative which does not vanish, and we found this derivative, as being the order $n$ one:

$$f'(x) = f''(x) = \ldots = f^{(n-1)}(x) = 0 \quad , \quad f^{(n)}(x) \neq 0$$

Then, the Taylor formula at $x$ at order $n$ takes the following form:

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

But this is exactly what we need, in order to fully solve our local extremum problem. Indeed, when $n$ is even, if $f^{(n)}(x) < 0$ what we have is a local maximum, and if $f^{(n)}(x) > 0$, what we have is a local minimum. As for the case where $n$ is odd, here we cannot have a local minimum or maximum, due to $t \to -t$ which switches growth. □

All the above, Theorem 11.7 and its proof, must be of course perfectly known, when looking for applications of such things. However, for theoretical purposes, let us record as well, in a very compact form, what is basically to be remembered:

THEOREM 11.8. *Given a differentiable function $f : \mathbb{R} \to \mathbb{R}$, we can always write*

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

*with $f^{(n)}(x) \neq 0$, and this tells us if $x$ is a local minimum, or maximum of $f$.*

PROOF. This was the conclusion of the proof of Theorem 11.7, and with the extra remark that local extremum means that $n$ is even, with in this case $f^{(n)}(x) < 0$ corresponding to local maximum, and $f^{(n)}(x) > 0$ corresponding to local minimum. □

There are of course many applications of the above.

## 11d. Basic applications

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

THEOREM 11.9. *We have the following formulae,*

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

*as well as the following formulae,*

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

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

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

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

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

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

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

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

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

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

(3) Finally, the fact that the formulae in the statement extend beyond the small $t$ setting, coming from Taylor series, is something standard too. We will leave this as an instructive exercise, and come back to it, later in this book. □

As another basic 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 11.10. *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, as we know from chapter 8.

(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:

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

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

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

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 $|t| < 1$, and we will leave this as an instructive exercise, and come back to it, later in this book. □

We can see from the above the power of the Taylor formula, saving us from quite complicated combinatorics. Remember indeed the mess from chapter 8, when trying to directly establish particular cases of the generalized binomial formula. Gone all that.

## 11e. Exercises

Exercises:

Exercise 11.11.

Exercise 11.12.

Exercise 11.13.

Exercise 11.14.

Exercise 11.15.

Exercise 11.16.

Exercise 11.17.

Exercise 11.18.

Bonus exercise.

CHAPTER 12

# Differential equations

## 12a. Newton, gravity

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

FACT 12.1. *Newton invented calculus for formulating the laws of motion as*

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

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

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

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

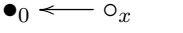
THEOREM 12.2. *The equation of a gravitational free fall, in 1 dimension, is*

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

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

PROOF. Assume indeed that we have a free falling object, in 1 dimension:


In order to reach to calculus as we know it, we must peform a rotation, as to have all this happening on the $Ox$ axis. By doing this, and assuming that $M$ is fixed at 0, our picture becomes as follows, with the attached numbers being now the coordinates:



Now comes the physics. The gravitational force exterted by $M$, which is fixed in our formalism, on the object $m$ which moves, is subject to the following equations:

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

To be more precise, in the first equation $G\simeq 6.674\times10^{-11}$ is the gravitational constant, in usual SI units, and the sign is $-$ because $F$ is attractive. The second equation is something standard and very intuitive, and the last two equations are those from Fact 12.1. Now observe that, with the above data for $F$, the equation $F=ma$ reads:

$$-G\cdot\frac{Mm}{x^2}=m\ddot{x}$$

Thus, by simplifying, we are led to the equation in the statement. □

## 12b. Wave equation

As more phsyics, we can talk as well about waves in 1 dimension, as follows:

THEOREM 12.3. *The wave equation in* 1 *dimension is*

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

*with the dot denoting time derivatives, and* $v>0$ *being the propagation speed.*

PROOF. In order to understand the propagation of the waves, let us model the space, which is $\mathbb{R}$ for us, as a network of balls, with springs between them, as follows:

$$\cdots\bowtie\bullet\bowtie\bullet\bowtie\bullet\bowtie\bullet\bowtie\bullet\bowtie\cdots$$

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:

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

We have two forces acting at $x$. First is the Newton motion force, mass times acceleration, which is as follows, with $m$ being the mass of each ball:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## 12c. Heat equation

Along the same lines, we can talk as well about the heat equation in 1D, as follows:

THEOREM 12.4. *The heat equation in* 1 *dimension is*

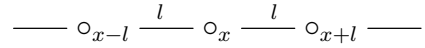
$$\dot{\varphi} = \alpha\varphi''$$

*where* $\alpha > 0$ *is the thermal diffusivity of the medium.*

PROOF. As before with the wave equation, this is not exactly a theorem, but rather what comes out of experiments, but we can justify this mathematically, as follows:

(1) As an intuitive explanation for this equation, since the second derivative $\varphi''$ computes the average value of a function $\varphi$ around a point, minus the value of $\varphi$ at that point, as we know from chapter 10, the heat equation as formulated above tells us that the rate of change $\dot{\varphi}$ of the temperature of the material at any given point must be proportional, with proportionality factor $\alpha > 0$, to the average difference of temperature between that given point and the surrounding material. Which sounds reasonable.

(2) In practice now, we can use, a bit like before for the wave equation, a lattice model as follows, with distance $l > 0$ between the neighbors:



In order to model now heat diffusion, we have to implement the intuitive mechanism explained above, and in practice, this leads to a condition as follows, expressing the change of the temperature $\varphi$, over a small period of time $\delta > 0$:

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

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

## 12d. Higher dimensions

In higher dimensions things are certainly more complicated. Let us start instead with some abstract mathematics. Here is a good, concrete question, which appears in mathematics, physics, and related disciplines, that we would like to solve:

QUESTION 12.5. *How to solve differential equations?*

Obviously, this question is quite broad, and as a first concrete example, let us examine the case of a falling object. If we denote by $x = x(t) : \mathbb{R} \to \mathbb{R}^3$ the position of our falling object, then its speed $v = v(t) : \mathbb{R} \to \mathbb{R}^3$ and acceleration $a = a(t) : \mathbb{R} \to \mathbb{R}^3$ are given by the following formulae, with the dots standing for derivatives with respect to time $t$:

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

Regarding now the equation of motion, this is as follows, coming from Newton, with $m$ being the mass of our object, and with $F$ being the gravitational force:

$$m \cdot a(t) = F(x(t))$$

Thus, in terms of derivatives as above, in order to have as only unknown the position vector $x = x(t) : \mathbb{R} \to \mathbb{R}^3$, the equation of motion is as follows:

$$m \cdot \ddot{x}(t) = F(x(t))$$

Which looks nice, but since what we have here is a degree 2 equation, instead of degree 1, which would be better, was it really a good idea to get rid of speed $v : \mathbb{R} \to \mathbb{R}^3$ and acceleration $a : \mathbb{R} \to \mathbb{R}^3$, and reformulate everything in terms of position $x : \mathbb{R} \to \mathbb{R}^3$.

Nevermind. So going all over again, with the aim this time of reaching to a degree 1 equation, let us replace our 3-dimensional unknown $x : \mathbb{R} \to \mathbb{R}^3$ with the 6-dimensional unknown $(x, v) : \mathbb{R} \to \mathbb{R}^6$. And with this done, surprise, we have our degree 1 system:

$$\begin{cases} \dot{x}(t) = v(t) \\ \dot{v}(t) = \frac{1}{m}F(x(t)) \end{cases}$$

Which was a nice trick, wasn't it. So, before going further, let us record the following conclusion, that we will come back to in a moment, after done with gravity:

CONCLUSION 12.6. *We can convert differential equations of higher order into differential equations of first order, by suitably enlarging the size of our unknown vectors.*

Now back to gravity and free falls, and to the degree 1 system found above, we will assume in what follows that our object is subject to a free fall under a uniform gravitational field. In practice, this means that the force $F$ is given by the following formula, with $m > 0$ being as usual the mass of our object, and with $g > 0$ being a certain constant:

$$F(x) = -mg \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

With this data, the system that we found takes the following form:

$$\begin{cases} \dot{x}(t) = v(t) \\ \\ \dot{v}(t) = -g \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \end{cases}$$

But this latter system is very easy to solve. Indeed, the second equation gives:

$$v(t) = v(0) - g \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} t$$

Now by integrating once again, we can recover as well the formula of $x$, as follows:

$$x(t) = x(0) + v(0)t - \frac{g}{2} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} t^2$$

Which is very nice, good work that we did here, so let us record our findings, along with a bit more, in the form of a complete statement, as follows:

THEOREM 12.7. *For a free fall in a uniform gravitational field, with gravitational acceleration constant $g > 0$, the equation of motion is*

$$x(t) = x(0) + v(0)t - \frac{g}{2} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} t^2$$

*and the trajectory is a parabola, unless in the case where the free fall is straight downwards, where the trajectory is a line.*

PROOF. This is a conclusion to what we found above, namely equation of motion, and its obvious implications, and the level of the corresponding trajectory. □

Now back to theory, let us go back to Conclusion 12.6, which was our main theoretical finding so far, and further comment on that. Of course in the case of extremely simple equations, like the above uniform gravity ones, there is no really need to use this trick, because you can directly integrate twice, and so on. However, in general, this remains a very useful trick, worth some discussion, and we will discuss this now.

Let us start with some generalities in one variable. We have here:

DEFINITION 12.8. *A general ordinary differential equation (ODE) is an equation as follows, with a function $x = x(t) : \mathbb{R} \to \mathbb{R}$ as unknown,*

$$F(t, x, \dot{x}, \ldots, x^{(k)}) = 0$$

*depending on a given function $F : U \to \mathbb{R}$, with $U \subset \mathbb{R}^{k+2}$ being an open set.*

As a first observation, under suitable assumptions on our function $F : U \to \mathbb{R}$, and more specifically non-vanishing of its partial derivatives, in all directions, we can use the implicit function theorem, in order to reformulate our equation as follows, for a certain function $f : V \to \mathbb{R}$, with $V \subset \mathbb{R}^{k+1}$ being a certain open set:

$$x^{(k)} = f(t, x, \dot{x}, \ldots, x^{(k-1)})$$

In practice, we will make this change, which often comes by default, when investigating questions coming from physics, and these will be the ODE that we will be interested in.

Now moving to several variables, more generally, let us formulate:

DEFINITION 12.9. *A standard system of ODE is a system as follows,*

$$x_1^{(k)} = f_1(t, x, \dot{x}, \ldots, x^{(k-1)})$$

$$\vdots$$

$$x_N^{(k)} = f_N(t, x, \dot{x}, \ldots, x^{(k-1)})$$

*with the unknown being a vector function $x = x(t) : \mathbb{R} \to \mathbb{R}^N$.*

Here the adjective "standard" refers to the implicit function theorem manipulation made in the above, which can be of course made in the context of several variables too.

Now with these abstract definitions in hand, we can go back to Conclusion 12.6, and formulate a more precise version of that observation, as follows:

THEOREM 12.10. *We can convert any standard system of ODE into a standard order 1 system of ODE, by suitably enlarging the size of the unknown vector.*

PROOF. This is indeed clear from definitions, because with $y = (x, \dot{x}, \ldots, x^{(k-1)})$, in the context of Definition 12.9, the system there takes the following form, as desired:

$$\dot{y}_1 = y_2$$

$$\dot{y}_2 = y_3$$

$$\vdots$$

$$\dot{y}_{k-1} = y_k$$

$$\dot{y}_k = f(t, y)$$

Thus, we are led to the conclusion in the statement. There are of course many explicit applications of this method, and further comments that can be made. □

Getting now to the point where we wanted to get, in order to get truly started with all this, with some mathematics going on, let us have a look at the systems of ODE which are linear. That is, we would like to solve equations as follows, with $f_i$ being linear:

$$x_1^{(k)} = f_1(t, x, \dot{x}, \ldots, x^{(k-1)})$$

$$\vdots$$

$$x_N^{(k)} = f_N(t, x, \dot{x}, \ldots, x^{(k-1)})$$

By doing the manipulation in Theorem 12.10, and assuming that we are in the "autonomous" case, where there is no time $t$ in our linear function which produces the system, we are led to a vector equation as follows, with $A \in M_N(\mathbb{R})$ being a certain matrix:

$$x' = Ax$$

But here, we are in familiar territory, namely very standard calculus, because in the 1D case, the solution simply appears by exponentiating, as follows:

$$x = e^{tA} x_0$$

Which is something very nice, and with this understood, we can now go back to our original Question 12.5, from the beginning of this section. As already mentioned, that question was something very broad, and as something more concrete now, we have:

QUESTION 12.11. *The solution of a system of linear differential equations,*

$$x' = Ax \quad , \quad x(0) = x_0$$

*with $A \in M_N(\mathbb{R})$, is normally given by $x = e^{tA} x_0$, and this because we should have:*

$$(e^{tA} x_0)' = A e^{tA} x_0$$

*But, what exactly is $e^{tA}$, and then, importantly, how to explicitly compute $e^{tA}$?*

To be more precise, again as with Question 12.5, this question appears indeed in a myriad contexts, all across physics and science, and with all this needing no further presentation. Observe also that, due to Theorem 12.10, this question allows us to deal with differential equations of higher order too, by enlarging the size of our vectors.

In answer now, the key to all this is advanced linear algebra theory. We will be back to all this at the end of the present book, where linear algebra will naturally appear in connection with the basic study of functions and their derivatives, in several variables.

## 12e. Exercises

Exercises:

Exercise 12.12.

Exercise 12.13.

Exercise 12.14.

Exercise 12.15.

Exercise 12.16.

Exercise 12.17.

Exercise 12.18.

Exercise 12.19.

Bonus exercise.

# Part IV

# Integrals

*You'll die as you lived, in a flash of the blade*
*In a corner forgotten by no one*
*You lived for the touch, for the feel of the steel*
*One man and his honor*

CHAPTER 13

# Integration theory

## 13a. Integration theory

We have seen so far the foundations of calculus, with lots of interesting results regarding the functions $f : \mathbb{R} \to \mathbb{R}$, and their derivatives $f' : \mathbb{R} \to \mathbb{R}$. The general idea was that in order to understand $f$, we first need to compute its derivative $f'$. The overall conclusion, coming from the Taylor formula, was that if we are able to compute $f'$, but then also $f''$, and $f'''$ and so on, we will have a good understanding of $f$ itself.

However, the story is not over here, and there is one more twist to the plot. Which will be a major twist, of similar magnitude to that of the Taylor formula. For reasons which are quite tricky, that will become clear later on, we will be interested in the integration of the functions $f : \mathbb{R} \to \mathbb{R}$. With the claim that this is related to calculus.

There are several possible viewpoints on the integral, which are all useful, and good to know. To start with, we have something very simple, as follows:

DEFINITION 13.1. *The integral of a continuous function $f : [a, b] \to \mathbb{R}$, denoted*

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

*is the area below the graph of $f$, signed $+$ where $f \geq 0$, and signed $-$ where $f \leq 0$.*

Here it is of course understood that the area in question can be computed, and with this being something quite subtle, that we will get into later. For the moment, let us just trust our intuition, our function $f$ being continuous, the area in question can "obviously" be computed. More on this later, but for being rigorous, however, let us formulate:

METHOD 13.2. *In practice, the integral of $f \geq 0$ can be computed as follows,*

(1) *Cut the graph of $f$ from 3mm plywood,*
(2) *Plunge that graph into a square container of water,*
(3) *Measure the water displacement, as to have the volume of the graph,*
(4) *Divide by $3 \times 10^{-3}$ that volume, as to have the area,*

*and for general $f$, we can use this plus $f = f_+ - f_-$, with $f_+, f_- \geq 0$.*

So far, so good, we have a rigorous definition, so let us do now some computations. In order to compute areas, and so integrals of functions, without wasting precious water, we can use our geometric knowledge. Here are some basic results of this type:

PROPOSITION 13.3. *We have the following results:*

(1) *When $f$ is linear, we have the following formula:*

$$\int_a^b f(x)dx = (b-a)\cdot\frac{f(a)+f(b)}{2}$$

(2) *In fact, when $f$ is piecewise linear on $[a=a_1,a_2,\ldots,a_n=b]$, we have:*

$$\int_a^b f(x)dx = \sum_{i=1}^{n-1}(a_{i+1}-a_i)\cdot\frac{f(a_i)+f(a_{i+1})}{2}$$

(3) *We have as well the formula $\int_{-1}^1\sqrt{1-x^2}\,dx=\pi/2$.*

PROOF. These results all follow from basic geometry, as follows:

(1) Assuming $f\geq 0$, we must compute the area of a trapezoid having sides $f(a),f(b)$, and height $b-a$. But this is the same as the area of a rectangle having side $(f(a)+f(b))/2$ and height $b-a$, and we obtain $(b-a)(f(a)+f(b))/2$, as claimed.

(2) This is clear indeed from the formula found in (1), by additivity.

(3) The integral in the statement is by definition the area of the upper unit half-disc. But since the area of the whole unit disc is $\pi$, this half-disc area is $\pi/2$. □

As an interesting observation, (2) in the above result makes it quite clear that $f$ does not necessarily need to be continuous, in order to talk about its integral. Indeed, assuming that $f$ is piecewise linear on $[a=a_1,a_2,\ldots,a_n=b]$, but not necessarily continuous, we can still talk about its integral, in the obvious way, exactly as in Definition 13.1, and we have an explicit formula for this integral, generalizing the one found in (2), namely:

$$\int_a^b f(x)dx = \sum_{i=1}^{n-1}(a_{i+1}-a_i)\cdot\frac{f(a_i^+)+f(a_{i+1}^-)}{2}$$

Based on this observation, let us upgrade our formalism, as follows:

DEFINITION 13.4. *We say that a function $f:[a,b]\to\mathbb{R}$ is integrable when the area below its graph is computable. In this case we denote by*

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

*this area, signed $+$ where $f\geq 0$, and signed $-$ where $f\leq 0$.*

As basic examples of integrable functions, we have the continuous ones, provided indeed that our intuition, or that Method 13.2, works indeed for any such function. We will soon see that this is indeed true, coming with mathematical proof. As further examples, we have the functions which are piecewise linear, or piecewise continuous. We will also see, later, as another class of examples, that the piecewise monotone functions are integrable. But more on this later, let us not bother for the moment with all this.

This being said, one more thing regarding theory, that you surely have in mind: is any function integrable? Not clear. I would say that if the Devil comes with some sort of nasty, totally discontinuous function $f : \mathbb{R} \to \mathbb{R}$, then you will have big troubles in cutting its graph from 3mm plywood, as required by Method 13.2. More on this later.

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

PROPOSITION 13.5. *We have the following formulae,*

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

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

*valid for any functions $f, g$ and any scalar $\lambda \in \mathbb{R}$.*

PROOF. Both these formulae are indeed clear from definitions. □

Moving ahead now, passed the above results, which are of purely algebraic and geometric nature, and perhaps a few more of the same type, which are all quite trivial and that we we will not get into here, we must do some analysis, in order to compute integrals. This is something quite tricky, and we have here the following result:

THEOREM 13.6. *We have the Riemann integration formula,*

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

*which can serve as a definition for the integral.*

PROOF. This is standard, by drawing rectangles. We have indeed the following formula, which can stand as a definition for the signed area below the graph of $f$:

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

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

THEOREM 13.7. *The integral of a function $f:[a,b]\to\mathbb{R}$ is given by*

$$\int_a^b f(x)dx=(b-a)\times A(f)$$

*where $A(f)$ is the average of $f$ over the interval $[a,b]$.*

PROOF. As explained above, this is clear from Definition 13.1, via some geometric thinking. Alternatively, this is something which certainly comes from Theorem 13.6. □

The point of view in Theorem 13.7 is something quite useful, and as an illustration for this, let us review the results that we already have, by using this interpretation. First, we have the formula for linear functions from Proposition 13.3, namely:

$$\int_a^b f(x)dx=(b-a)\cdot\frac{f(a)+f(b)}{2}$$

But this formula is totally obvious with our new viewpoint, from Theorem 13.7. The same goes for the results in Proposition 13.5, which become even more obvious with the viewpoint from Theorem 13.7. Thus, what we have in Theorem 13.7 is definitely useful.

However, not everything trivializes in this way, and the result which is left, from what we have so far, namely the formula $\int_{-1}^1\sqrt{1-x^2}\,dx=\pi/2$ from Proposition 13.3 (3), not only does not trivialize, but becomes quite opaque with our new philosophy.

In short, modesty. Integration is a quite delicate business, and we have several equivalent points of view on what an integral means, and all these points of view are useful, and must be learned, with none of them being clearly better than the others.

Going ahead with more interpretations of the integral, we have:

THEOREM 13.8. *We have the Monte Carlo integration formula,*

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

*with $x_1,\ldots,x_N\in[a,b]$ being random.*

PROOF. We recall from Theorem 13.7 that the idea is that we have a formula as follows, with the points $x_1, \ldots, x_N \in [a, b]$ being uniformly distributed:

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

But this works as well when the points $x_1, \ldots, x_N \in [a, b]$ are randomly distributed, for somewhat obvious reasons, and this gives the result. □

Observe that Monte Carlo integration works better than Riemann integration, for instance when trying to improve the estimate, via $N \to N + 1$. Indeed, in the context of Riemann integration, assume that we managed to find an estimate as follows, which in practice requires computing $N$ values of our function $f$, and making their average:

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

In order to improve this estimate, any extra computed value of our function $f(y)$ will be unuseful. For improving our formula, what we need are $N$ extra values of our function, $f(y_1), \ldots, f(y_N)$, with the points $y_1, \ldots, y_N$ being the midpoints of the previous division of $[a, b]$, so that we can write an improvement of our formula, as follows:

$$\int_a^b f(x)dx \simeq \frac{b-a}{2N} \sum_{k=1}^{2N} f\left(a + \frac{b-a}{2N} \cdot k\right)$$

With Monte Carlo, things are far more flexible. Assume indeed that we managed to find an estimate as follows, which again requires computing $N$ values of our function:

$$\int_a^b f(x)dx \simeq \frac{b-a}{N} \sum_{k=1}^{N} f(x_i)$$

Now if we want to improve this, any extra computed value of our function $f(y)$ will be helpful, because we can set $x_{n+1} = y$, and improve our estimate as follows:

$$\int_a^b f(x)dx \simeq \frac{b-a}{N+1} \sum_{k=1}^{N+1} f(x_i)$$

And isn't this potentially useful, and powerful, when thinking at practically computing integrals, either by hand, or by using a computer. Let us record this finding as follows:

CONCLUSION 13.9. *Monte Carlo integration works better than Riemann integration, when it comes to computing as usual, by estimating, and refining the estimate.*

As another interesting feature of Monte Carlo integration, this works better than Riemann integration, for functions having various symmetries, because Riemann integration can get "fooled" by these symmetries, while Monte Carlo remains strong.

As an example for this phenomeon, chosen to be quite drastic, let us attempt to integrate, via both Riemann and Monte Carlo, the following function $f : [0, \pi] \to \mathbb{R}$:

$$f(x) = \Big|\sin(120x)\Big|$$

The first few Riemann sums for this function are then as follows:

$$I_2(f) = \frac{\pi}{2}(|\sin 0| + |\sin 60\pi|) = 0$$

$$I_3(f) = \frac{\pi}{3}(|\sin 0| + |\sin 40\pi| + |\sin 80\pi|) = 0$$

$$I_4(f) = \frac{\pi}{4}(|\sin 0| + |\sin 30\pi| + |\sin 60\pi| + |\sin 90\pi|) = 0$$

$$I_5(f) = \frac{\pi}{5}(|\sin 0| + |\sin 24\pi| + |\sin 48\pi| + |\sin 72\pi| + |\sin 96\pi|) = 0$$

$$I_6(f) = \frac{\pi}{6}(|\sin 0| + |\sin 20\pi| + |\sin 40\pi| + |\sin 60\pi| + |\sin 80\pi| + |\sin 100\pi|) = 0$$

$$\vdots$$

Based on this evidence, we will conclude, obviously, that we have:

$$\int_0^\pi f(x)dx = 0$$

With Monte Carlo, however, such things cannot happen. Indeed, since there are finitely many points $x \in [0, \pi]$ having the property $\sin(120x) = 0$, a random point $x \in [0, \pi]$ will have the property $|\sin(120x)| > 0$, so Monte Carlo will give, at any $N \in \mathbb{N}$:

$$\int_0^\pi f(x)dx \simeq \frac{b-a}{N}\sum_{k=1}^{N} f(x_i) > 0$$

Again, this is something interesting, when practically computing integrals, either by hand, or by using a computer. So, let us record, as a complement to Conclusion 13.9:

Conclusion 13.10. *Monte Carlo integration is smarter than Riemann integration, because the symmetries of the function can fool Riemann, but not Monte Carlo.*

All this is good to know, when computing integrals in practice, especially with a computer. Finally, here is one more useful interpretation of the integral:

Theorem 13.11. *The integral of a function $f : [a, b] \to \mathbb{R}$ is given by*

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

*where $E(f)$ is the expectation of $f$, regarded as random variable.*

Proof. This is just some sort of fancy reformulation of Theorem 13.8, the idea being that what we can "expect" from a random variable is of course its average. We will be back to this later in this book, when systematically discussing probability theory. □

## 13b. Integrable functions

Our purpose now will be to understand which functions $f : \mathbb{R} \to \mathbb{R}$ are integrable, and how to compute their integrals. For this purpose, the Riemann formula in Theorem 13.6 will be our favorite tool. Let us begin with some theory. We first have:

THEOREM 13.12. *The following functions are integrable:*

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

PROOF. This is indeed something quite standard, as follows:

(1) It is enough to prove the first assertion for a function $f : [a, b] \to \mathbb{R}$ which is continuous, and our claim here is that this follows from the uniform continuity of $f$. To be more precise, given $\varepsilon > 0$, let us choose $\delta > 0$ such that the following happens:

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

In order to prove the result, let us pick two divisions of $[a, b]$, as follows:

$$I = [a = a_1 < a_2 < \ldots < a_n = b]$$

$$I' = [a = a'_1 < a'_2 < \ldots < a'_m = b]$$

Our claim, which will prove the result, is that if these divisions are sharp enough, of resolution $< \delta/2$, then the associated Riemann sums $\Sigma_I(f), \Sigma_{I'}(f)$ are close within $\varepsilon$:

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

(2) In order to prove this claim, let us denote by $l$ the length of the intervals on the real line. Our assumption is that the lengths of the divisions $I, I'$ satisfy:

$$l\big([a_i, a_{i+1}]\big) < \frac{\delta}{2} \quad , \quad l\big([a'_i, a'_{i+1}]\big) < \frac{\delta}{2}$$

Now let us intersect the intervals of our divisions $I, I'$, and set:

$$l_{ij} = l\big([a_i, a_{i+1}] \cap [a'_j, a'_{j+1}]\big)$$

The difference of Riemann sums that we are interested in is then given by:

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

(3) Now let us estimate $f(a_i) - f(a'_j)$. Since in the case $l_{ij} = 0$ we do not need this estimate, we can assume $l_{ij} > 0$. Now by remembering what the definition of the numbers $l_{ij}$ was, we conclude that we have at least one point $x \in \mathbb{R}$ satisfying:

$$x \in [a_i, a_{i+1}] \cap [a'_j, a'_{j+1}]$$

But then, by using this point $x$ and our assumption on $I, I'$ involving $\delta$, we get:

$$\begin{aligned} |a_i - a'_j| &\leq |a_i - x| + |x - a'_j| \\ &\leq \frac{\delta}{2} + \frac{\delta}{2} \\ &= \delta \end{aligned}$$

Thus, according to our definition of $\delta$ from (1), in relation to $\varepsilon$, we get:

$$|f(a_i) - f(a'_j)| < \varepsilon$$

(4) But this is what we need, in order to finish. Indeed, with the estimate that we found, we can finish the computation started in (2), as follows:

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

Thus our two Riemann sums are close enough, provided that they are both chosen to be fine enough, and this finishes the proof of the first assertion.

(5) Regarding now the second assertion, this is something more technical, that we will not really need in what follows. We will leave the proof here, which uses similar ideas to those in the proof of (1) above, namely subdivisions and estimates, as an exercise. □

We should mention that the above result is just a beginning, and many other things can be said about the integrable functions, and about the non-integrable functions too. For more on all this, we recommend any specialized measure theory book.

Going ahead with more theory, let us establish now some abstract properties of the integration operation. This will be actually quite technical, and we will be quite brief, and for more on this, we recommend as usual any specialized measure theory book.

We already know from Proposition 13.5 that the integrals behave well with respect to sums and multiplication by scalars. Along the same lines, we have:

THEOREM 13.13. *The integrals behave well with respect to taking limits,*

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

*and with respect to taking infinite sums as well,*

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

*with both these formulae being valid, undwer mild assumptions.*

PROOF. This is something quite standard, by using the general theory developed in chapter 7 for the sequences and series of functions. To be more precise, (1) follows by using the material there, via Riemann sums, and then (2) follows as a particular case of (1). We will leave for now the clarification of all this as an instructive exercise, and we will come back to it, with full details, at the end of the present chapter. □

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

THEOREM 13.14. *Given a continuous function $f : [a, b] \to \mathbb{R}$, we have*

$$\exists c \in [a, b] \quad , \quad \int_a^b f(x)dx = (b - a)f(c)$$

*with this being called mean value property.*

PROOF. Our claim is that this follows from the following trivial estimate:

$$\min(f) \leq f \leq \max(f)$$

Indeed, by integrating this over $[a, b]$, we obtain the following estimate:

$$(b - a)\min(f) \leq \int_a^b f(x)dx \leq (b - a)\max(f)$$

Now observe that this latter estimate can be written as follows:

$$\min(f) \leq \frac{\int_a^b f(x)dx}{b - a} \leq \max(f)$$

Since $f$ must takes all values on $[\min(f), \max(f)]$, we get a $c \in [a, b]$ such that:

$$\frac{\int_a^b f(x)dx}{b - a} = f(c)$$

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

## 13c. Abstract integration

Let us discuss now abstract measure theory. We first need measurable sets, and on the bottom line, we would like our open and closed sets, and their "combinations", to be measurable. But these "combinations" can be understood and axiomatized, and are called Borel sets. So, this will be the idea in what follows, talking about abstract measurable sets, then about Borel sets, and then coming with measures, and integration theory.

Let us start with the abstract measurable sets. We have here the following definition, which is something very general, not making reference to any metric on our space $X$, nor making reference to any measure on $X$, measuring these measurable sets:

DEFINITION 13.15. *An abstract measured space is a set $X$, given with a set of subsets $M \subset P(X)$, called measurable sets, which form an algebra, in the sense that:*

(1) $\emptyset, X \in M$.
(2) $E \in M \implies E^c \in M$.
(3) *$M$ is stable under countable unions and intersections.*

Obviously, this is something quite abstract. If the last axiom, (3), is only satisfied for the finite unions and intersections, we say that $M \subset P(X)$ is a finite algebra.

As a first observation, some of the axioms above are redundant. Indeed, assuming that (2) holds, we have the following equivalence, which can help in verifying (1):

$$\emptyset \in M \iff X \in M$$

The same goes for the axiom (3), with only one of the conditions there being in need to be verified, in practice, and this due to the following formula:

$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_c^i$$

However, it is most convenient to write Definition 13.15 as above, in a symmetric way with respect to the two operations involved, namely union and intersection.

At the level of examples of abstract measured spaces, there are many of them, and more on this in a moment, usually coming via the following result:

THEOREM 13.16. *Given a set of subsets $S \subset P(X)$, there is a smallest algebra*

$$M = \bar{S}$$

*containing it, called algebra generated by $X$.*

PROOF. This can be viewed in two possible ways, as follows:

(1) According to the axioms in Definition 13.15, what we have to do in order to construct $M = \bar{S}$ is to start with $S$, then add $\emptyset, X$ to it, along with the complements $E^c$

of all the sets $E \in S$, and then take countable unions and intersections of such sets. And, some elementary verifications show that what we get in this way is indeed an algebra.

(2) Alternatively, we can define $M = \bar{S}$ as being the intersection of all algebras containing $S$, and with the remark that we have at least one such algebra, namely $P(X)$ itself. It is then clear from definitions that $M$ is an algebra, as desired. □

Getting now to the concrete examples of abstract measured spaces, we have:

DEFINITION 13.17. *Any metric space $X$ is automatically an abstract measured space, with the algebra of measurable sets being*

$$B = \bar{O}$$

*that is, the smallest algebra containing the open sets, called Borel algebra of $X$.*

Observe that the Borel sets include all open sets, all closed sets, as well as all countable unions of closed sets, and all countable intersections of open sets. As an example here, in the case $X = \mathbb{R}$, with its usual topology, all kinds of intervals are Borel sets:

$$(a, b),\ [a, b],\ (a, b],\ [a, b) \in B$$

Indeed, the first interval is open, and the second one is closed, so these are certainly Borel sets. As for the third and fourth intervals, these appear as countable unions of closed intervals, or as countable intersections of open intervals, so they are Borel too.

Getting back now to the general case, following Definition 13.15, we have:

DEFINITION 13.18. *Given a measured space $(X, M)$, a measure on it is a function*

$$\mu : M \to [0, \infty]$$

*which is countably additive, in the sense that we have*

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$$

*for any countable family of disjoint measurable sets $E_i \in M$.*

This definition, which will play a key role in what follows, is something quite tricky, and there are several comments to be made about it, as follows:

(1) Obviously, what we axiomatized above are the positive measures, and for the moment this will do, and we will omit the term "positive". More on this later, when we will talk about differences $\mu - \eta$ of such measures, and about complex measures too.

(2) In contrast, we did not assume that our measures are finite, and this because many interesting spaces, such as $X = \mathbb{R}$ itself, are naturally of infinite measure, $\mu(X) = \infty$. In the case where the measure $\mu$ happens to be bounded, up to a rescaling we can assume $\mu(X) = 1$, and we say in this case that we have a probability measure on $X$.

(3) Yet another subtlety comes in relation with the countable additivity condition at the end, the point being that for many measured spaces $X$, the finite additivity condition is something strictly weaker, and leads to a wrong theory. More on this later.

Looking at what we have so far, Definition 13.17 and Definition 13.18, many natural questions appear, and leaving aside anything too specialized, we are led to:

QUESTION 13.19. *Given a metric space $X$, such as $X = \mathbb{R}$, or $X = \mathbb{R}^N$, how to construct a measure on it? Also, once we have such a measure, how to integrate the functions $f : X \to \mathbb{R}$, or $f : X \to \mathbb{C}$, with respect to this measure?*

As you can see, we have two questions here, and none is trivial:

(1) Indeed, regarding the first question, this is something that we do not know yet how to solve, even this even for very simple spaces like $X = \mathbb{R}$. Indeed, while we certainly know how to measure the real intervals, simply by setting $\mu(a, b) = b - a$, and then unions of such intervals too, by decomposing then into disjoint unions of intervals, and making sums, nothing guarantees that we can measure any Borel set $E \in B$, in this way.

(2) As for the second question, experience with the usual Riemann integral, that you know well from calculus, shows that such things can be quite tricky too. Observe also that this second question is more general than the first one, because we have $\mu(E) = \int \chi_E$ for any $E \in M$, so if we know how to integrate, we know how to measure.

In short, we are facing non-trivial questions here, and we need a plan. And, perhaps a bit surprisingly, the best plan in order to deal with Question 13.19 is as follows:

PLAN 13.20. *We will jointly develop measure and integration theory, as follows:*

(1) *We will first keep staying abstract, and understand how the functions $f : X \to \mathbb{R}$, or $f : X \to \mathbb{C}$, can be integrated, with respect to an abstract measure.*
(2) *With this understood, we will conclude that the integration over $X = \mathbb{R}$ can only be a straightforward extension of the usual Riemann integration.*
(3) *But, in the end, this will enable us to both measure all the Borel sets $E \subset \mathbb{R}$, and to integrate the measurable functions $f : \mathbb{R} \to \mathbb{C}$.*
(4) *Finally, we will discuss how to deal with $X = \mathbb{R}^N$ too, and with other product spaces $X = Y \times Z$, both measure theory and integration.*

As you can see, all this is quite tricky, the main idea behind this plan being functional analysis, that is, using functions and their integrals, instead of just fighting with abstract measure theory first, and looking at functions and their integrals afterwards.

Getting started now, let us first talk about measurable functions, in the general context of Definition 13.15, with no measure involved. We have here the following notion:

DEFINITION 13.21. *Given a measured space $X$, and a metric space $Y$, a function*

$$f : X \to Y$$

*is called measurable when it satisfies the following condition:*

$$U \in O \implies f^{-1}(U) \in M$$

*When $X$ comes with a measure, we also call such functions integrable.*

Obviously, this is something simplified, because for doing abstract measurability theory, our spaces should be abstract measured spaces in the sense of Definition 13.15, and so the functions that we should normally care about should be functions $f : X \to Y$, with both $X, Y$ being abstract measured spaces. However, in view of Definition 13.17, this is more or less that what we are doing here, by restricting the attention to the target spaces $Y$ which are metric, with this being the case that really matters.

Many things can be said about the measurable functions. We first have:

PROPOSITION 13.22. *The measurable functions have the following properties:*

(1) *If $f : X \to Y$ is measurable and $g : Y \to Z$ is continuous, $g \circ f$ is measurable.*
(2) *If $f, g : X \to \mathbb{R}$ are both measurable and $h : \mathbb{R}^2 \to Y$ is continuous, then the function $x \to h(f(x), g(x))$ is measurable.*
(3) *$f : X \to \mathbb{C}$ is measurable precisely when $Re(f), Im(f) : X \to \mathbb{R}$ are measurable. In this case, the function $|f| : X \to \mathbb{R}$ is measurable too.*
(4) *If $f, g : X \to \mathbb{C}$ are measurable, then so are $f + g, fg : X \to \mathbb{C}$.*

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

(1) This is clear from definitions, because we have:

$$U \in O \implies g^{-1}(U) \in O \implies f^{-1}(g^{-1}(U)) \in M$$

(2) By using (1), it is enough to check that the function $k(x) = (f(x), g(x))$ is measurable. But this follows by writing any open set $U \subset \mathbb{R}^2$ as a union of rectangles. Indeed, the preminage of each rectangle $R = I \times J$ is measurable, as shown by:

$$k^{-1}(R) = (f, g)^{-1}(I \times J) = f^{-1}(I) \cap g^{-1}(J) \in M$$

But then, with this in hand, since any open set $U \subset \mathbb{R}^2$ can be written as a union of rectangles, it follows that $k^{-1}(U)$ is measurable, as desired.

(3) This follows indeed by using (2), with the following functions:

$$h(z) = z, Re(z), Im(z), |z|$$

(4) In the real case, $f, g : X \to \mathbb{R}$, the result follows by using (2), with:

$$h(x, y) = x + y, xy$$

Then, the result can be extended to the complex case, $f, g : X \to \mathbb{C}$, by using (3). □

Next, we have the following useful characterization of the real measurable functions:

PROPOSITION 13.23. *A function $f : X \to [-\infty, \infty]$ is mesurable precisely when*

$$f^{-1}(I) \in M$$

*for any interval of type $I = (a, \infty]$, with $a \in \mathbb{R}$.*

PROOF. Consider the following set, which is easily seen to be an algebra:

$$\Omega = \left\{ E \subset [-\infty, \infty] \middle| f^{-1}(E) \in M \right\}$$

We want to prove that $\Omega$ contains all the open sets, and this can be done as follows:

(1) Pick $a \in \mathbb{R}$, and then pick an increasing sequence $a_n \to a$. We have then the following formula, which shows that we have $[-\infty, a) \in \Omega$:

$$[-\infty, a) = \bigcup_n [-\infty, a_n) = \bigcup_n (a_n, \infty]^c \in \Omega$$

(2) But with this in hand, we obtain, for any $a < b$, that we have:

$$(a, b) = [-\infty, b) \cap (a, \infty] \in \Omega$$

(3) Now since any open set $U \subset [-\infty, \infty]$ can be written as a union of open intervals, we conclude that we have $U \in \Omega$, as desired. □

Getting now to limits of measurable functions, we have the following result:

PROPOSITION 13.24. *Given an abstract measured space $X$, the measurable functions $f : X \to [-\infty, \infty]$ have the following properties:*

(1) *If $f_n$ are measurable, so are $g = \sup_n f_n$, and $h = \limsup_n f_n$.*
(2) *If $f_n$ are measurable, so are $k = \inf_n f_n$, and $l = \liminf_n f_n$.*
(3) *If $f_n \to f$ and $f_n$ are measurable, then $f$ is measurable.*
(4) *If $f, g$ are measurable, so are $h = \min(f, g)$ and $k = \max(f, g)$.*
(5) *If $f$ is measurable, so are $f^+ = \max(f, 0)$ and $f^- = -\min(f, 0)$.*

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

(1) For the function $g = \sup_n f_n$ we can use the measurability criterion from Proposition 13.23, along with the following fact, valid for any $a \in \mathbb{R}$:

$$g = \sup_n f_n \implies g^{-1}(a, \infty] = \bigcup_n f_n^{-1}(a, \infty]$$

By symmetry we obtain that the function $k = \inf_n f_n$ is measurable as well. But with these results in hand, the last assertion follows too, by using the following formula:

$$\limsup_n f_n = \inf_k \left( \sup_{n \geq k} f_n \right)$$

(2) Here the fact that $k = \inf_n f_n$ is measurable was already proved in the above, and for $l = \liminf_n f_n$ we can use the same argument, symmetry, or the following formula:

$$\liminf_n f_n = \sup_k \left( \inf_{n \geq k} f_n \right)$$

(3) This follows indeed from (1), or from (2), and from the following fact:

$$f_n \to f \implies f = \inf_n f_n = \sup_n f_n$$

(4) This is a trivial application of (1) and (2).

(5) This follows from (4), and from the fact that if $f$ is measurable, so is $-f$. □

As a main result now regarding the measurable functions, which will be of key importance in what follows, we have the following statement, with the convention that a step function means a function which takes finitely many values:

THEOREM 13.25. *Given a measurable function $f : X \to [0, \infty]$, we can write*

$$f(x) = \lim_{n \to \infty} \varphi_n(x)$$

*with $0 \leq \varphi_1 \leq \varphi_2 \leq \ldots \leq f$, and with each $\varphi_i$ being a measurable step function.*

PROOF. As a first observation, the converse holds too, thanks to the results in Proposition 13.24. Regarding now the proof, this goes as follows:

(1) First, it is clear by drawing a picture that we can approximate the identity of $[0, \infty]$ with step functions as in the statement. That is, we can obviously find an increasing sequence of measurable step functions $0 \leq \psi_1 \leq \psi_2 \leq \ldots \leq id$, satisfying:

$$\lim_{n \to \infty} \psi_n(x) = x$$

(2) Now let us set $\varphi_n = \psi_n \circ f$. According to the limiting formula above, we have:

$$\lim_{n \to \infty} \varphi_n(x) = f(x)$$

On the other hand, by using the measurability criterion from Proposition 13.23, it follows that our truncation functions $\varphi_n = \psi_n \circ f$ are measurable, as desired. □

Good news, with the above general theory understood, we can now integrate functions, by following the good old method of Riemann, that we know well from the beginning of this chapter. To be more precise, we have the following result, which is of course stated a bit informally, with some of the details being left to you, as an instructive exercise:

THEOREM 13.26. *We can integrate the measurable functions $f : X \to \mathbb{R}_+$ by setting*

$$\int_X f(x) d\mu(x) = \sup_{0 \leq \varphi \leq f} \int_X \varphi(x) d\mu(x)$$

*with sup over measurable step functions, then extend this by linearity.*

PROOF. This is something very standard, and we will leave the clarification of all this, both precise statement, and proof, as an instructive exercise. To be more precise:

(1) We can certainly integrate the step functions $\varphi : X \to \mathbb{R}_+$, by writing each such function as a linear combination of characteristic functions, as follows:

$$\varphi = \sum_i \lambda_i \chi_{E_i}$$

Indeed, with this formula in hand, we can integrate our function $\varphi$, as follows:

$$\int_X \varphi(x)\, d\mu(x) = \sum_i \lambda_i \mu(E_i)$$

The integral of step functions constructed in this way has then all the linearity and positivity properties that you might expect, and behaves well with respect to limits.

(2) Next, consider an arbitrary measurable function $f : X \to [0, \infty]$. We know from Theorem 13.25 that we can write this function as an increasing limit, as follows, with $0 \leq \varphi_1 \leq \varphi_2 \leq \ldots \leq f$, and with each $\varphi_i$ being a measurable step function:

$$f(x) = \lim_{n\to\infty} \varphi_n(x)$$

But this suggests to define the integral of $f$ by the formula in the statement, namely:

$$\int_X f(x) d\mu(x) = \sup_{0\leq\varphi\leq f} \int_X \varphi(x) d\mu(x)$$

Indeed, we can see that the integral constructed in this way has all the linearity and positivity properties that you might expect, and behaves well with respect to limits. □

More in detail now, here are some basic properties of the integrals, which are very similar to those of the Riemann integral, that we know from before:

PROPOSITION 13.27. *The integrals of measurable functions $f : X \to \mathbb{R}_+$ have the following properties:*

(1) $f \leq g$ *implies* $\int f \leq \int g$.
(2) $E \subset F$ *implies* $\int_E f \leq \int_F f$.
(3) $\int f + g = \int f + \int g$.
(4) $\int \lambda f = \lambda \int f$.

PROOF. All this is indeed very standard, all routine verifications. □

Summarizing, we know how to integrate the real positive functions, in our abstract measure theory setting. In general, we can use the following formula:

$$\int_X (f - g)(x)\, d\mu(x) = \int_X f(x) d\mu(x) - \int_X g(x) d\mu(x)$$

We can integrate as well the complex functions, by setting:

$$\int_X (f+ig)(x)\,d\mu(x)=\int_X f(x)d\mu(x)+i\int_X g(x)d\mu(x)$$

All this is, indeed, very standard, exactly as for the Riemann integral. Let us record these findings as an upgrade of Theorem 13.26, as follows, once again with the statement being a bit informal, and with some of the details being left as instructive exercises:

THEOREM 13.28. *We can integrate the measurable functions $f:X\to\mathbb{C}$ by setting*

$$\int_X f(x)d\mu(x)=\sup_{0\leq\varphi\leq f}\int_X\varphi(x)d\mu(x)$$

*with sup over measurable step functions, for $f:X\to\mathbb{R}_+$ then extend this by linearity.*

PROOF. This follows indeed from what we already know from Theorem 13.26, and from the above discussion, exactly as for the Riemann integral. □

## 13d. Lebesgue, Fatou

Many other things can be said here, following Lebesgue, Fatou and others. We first have the following result, regarding the monotone convergence, due to Lebesgue:

THEOREM 13.29 (Lebesgue). *Given an increasing sequence of measurable functions*

$$0\leq f_1\leq f_2\leq\ldots\leq\infty$$

*which converges pointwise, $f_n\to f$, their limit is measurable, and we have*

$$\int_X f_n(x)d\mu(x)\to\int_X f(x)d\mu(x)$$

*for any positive measure on $X$.*

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

(1) We first have the following obvious implication, showing that the sequence of integrals on the right converges, to a certain number in $[0,\infty]$:

$$f_1\leq f_2\leq\ldots\implies\int_X f_1(x)d\mu(x)\leq\int_X f_2(x)d\mu(x)\leq\ldots$$

Moreover, since we have $f_n\to f$, it follows that we have the following inequality:

$$\lim_{n\to\infty}\int f_n(x)d\mu(x)\leq\int_X f(x)d\mu(x)$$

(2) In order to prove now the reverse inequality, pick a measurable simple function $0\leq\varphi\leq f$, pick also a number $c\in(0,1)$, and consider the following sets:

$$E_n=\left\{x\in X\,\middle|\,f_n(x)\geq c\varphi(x)\right\}$$

These sets are then measurable, we have $E_1 \subset E_2 \subset \ldots$, and our claim is that:

$$X = \bigcup_n E_n$$

Indeed, given $x \in X$, if $f(x) = 0$ then $x \in E_1$ and things fine. Otherwise, we have $f(x) > 0$, and so $f(x) > c\varphi(x)$, since $c < 1$, and so $x \in E_n$ for some $n$, as desired.

(3) Now observe that, with $0 \leq \varphi \leq f$ and $c \in (0, 1)$ as above, we have:

$$\int_X f_n(x)d\mu(x) \geq \int_{E_n} f_n(x)d\mu(x) \geq c \int_{E_n} \varphi(x)d\mu(x)$$

By taking the limit of this estimate, with $n \to \infty$, we obtain:

$$\lim_{n\to\infty} \int f_n(x)d\mu(x) \geq c \int_X \varphi(x)d\mu(x)$$

Now with $c \to 1$, we obtain from this the following estimate:

$$\lim_{n\to\infty} \int f_n(x)d\mu(x) \geq \int_X \varphi(x)d\mu(x)$$

But this being true for any simple function $0 \leq \varphi \leq f$, we conclude that we have:

$$\lim_{n\to\infty} \int f_n(x)d\mu(x) \geq \int_X f(x)d\mu(x)$$

Thus we have the reverse of the estimate found in (1), which finishes the proof. □

Regarding now the series of functions, again following Lebesgue, we have:

THEOREM 13.30. *Given measurable functions $f_n : X \to [0, \infty]$, their sum*

$$f(x) = \sum_{n=1}^{\infty} f_n(x)$$

*is measurable, and we have the formula*

$$\int_X f(x)d\mu(x) = \sum_{n=1}^{\infty} \int_X f_n(x)d\mu(x)$$

*for any positive measure on $X$.*

PROOF. This follows indeed from Theorem 13.29, applied to the partial sums, and we will leave all the verifications here as an instructive exercise. □

Following now Fatou, we have as well the following key result:

THEOREM 13.31 (Fatou). *Given measurable functions $f_n : X \to [0, \infty]$, their limit*

$$f(x) = \liminf_n f_n(x)$$

*is measurable, and we have the formula*

$$\int_X f(x) d\mu(x) \leq \liminf_n \int_X f_n(x) d\mu(x)$$

*for any positive measure on $X$.*

PROOF. The first assertion is something that we know, from Proposition 13.24. As for the second assertion, this can be proved by using the same trick as there, namely:

$$\liminf_n f_n = \sup_k \left( \inf_{n \geq k} f_n \right)$$

Indeed, we can apply Theorem 13.29 to the following sequence of functions:

$$g_k(x) = \inf_{n \geq k} f_n(x)$$

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

Regarding the above result of Fatou, let us mention that there are examples where the inequality can be strict, with the standard example here being as follows:

$$f_n = \begin{cases} \chi_E & (n \text{ odd}) \\ 1 - \chi_E & (n \text{ even}) \end{cases}$$

Finally, as a last basic result, due to Lebesgue again, we have:

THEOREM 13.32 (Lebesgue). *Assuming $f_n \to f$ pointwise, and assuming too*

$$|f_n| \leq g$$

*with $\int g < \infty$, the following happen:*

(1) $\int |f| < \infty$.
(2) $\int |f_n - f| \to 0$.
(3) $\int f_n \to \int f$.

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

(1) This follows from $|f_n| \leq g$, which in the limit gives $|f| < g$, so $\int |f| < \infty$.

(2) Since $|f_n - f| \leq 2g$, we can apply Theorem 13.31 to the following functions:

$$h_n = 2g - |f_n - f|$$

We obtain in this way the following estimate:

$$\begin{aligned}\int_X 2g &\leq \liminf_n \int_X 2g - |f_n - f| \\ &= \int_X 2g + \liminf_n \left( - \int_X |f_n - f| \right) \\ &= \int_X 2g - \limsup_n \int_X |f_n - f|\end{aligned}$$

Now by substracting $\int_X 2g$, this estimate gives the following formula:

$$\limsup_n \int_X |f_n - f| \leq 0$$

We conclude that this limit must be 0, as claimed in (2).

(3) This follows indeed from (2). □

Many other things can be said, along these lines, which are more specialized, and such knowledge can be very useful when doing applied probability, and statistics. For our purposes here, which will be mostly pure mathematical, with a touch of theoretical physics, the above general theory, and results of Lebesgue and Fatou, will do.

## 13e. Exercises

Exercises:

Exercise 13.33.

Exercise 13.34.

Exercise 13.35.

Exercise 13.36.

Exercise 13.37.

Exercise 13.38.

Exercise 13.39.

Exercise 13.40.

Bonus exercise.

CHAPTER 14

# Main theorems

## 14a. Riemann sums

At the level of examples now, let us first look at the simplest functions that we know, namely the power functions $f(x) = x^p$. However, things here are tricky, as follows:

THEOREM 14.1. *We have the integration formula*

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

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

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

(1) By linearity we can assume that our interval $[a, b]$ is of the form $[0, c]$, and the formula that we want to establish is as follows:

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

(2) We can further assume $c = 1$, and by expressing the left term as a Riemann sum, we are in need of the following estimate, in the $N \to \infty$ limit:

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

(3) So, let us try to prove this. At $p = 0$, obviously nothing to do, because we have the following formula, which is exact, and which proves our estimate:

$$1^0 + 2^0 + \ldots + N^0 = N$$

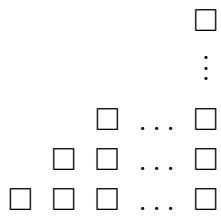
(4) At $p = 1$ now, we are confronted with a well-known question, namely the computation of $1 + 2 + \ldots + N$. But this is simplest done by arguing that the average of the numbers $1, 2, \ldots, N$ being the number in the middle, we have:

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

Thus, we obtain the following formula, which again solves our question:

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

(5) At $p = 2$ now, go compute $1^2 + 2^2 + \ldots + N^2$. This is not obvious at all, so as a preliminary here, let us go back to the case $p = 1$, and try to find a new proof there, which might have some chances to extend at $p = 2$. The trick is to use 2D geometry. Indeed, consider the following picture, with stacks going from 1 to $N$:



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

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

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

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

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

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

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

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

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

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

Alternatively, assuming that the gods of combinatorics are with us, we can see right away the following formula, which coupled with (4) gives the result:

$$1^3 + 2^3 + \ldots + N^3 = (1 + 2 + \ldots + N)^2$$

In any case, in practice, the above formula holds indeed, and you can check it for instance by recurrence, and this solves our problem:

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

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

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

CONJECTURE 14.2. *We have the following estimate,*

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

*and so, by Riemann sums, we have the following integration formula,*

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

*valid for any exponent $p\in\mathbb{N}$, and perhaps for some other $p\in\mathbb{R}$.*

We will see later that this conjecture is indeed true, and with the exact details regarding the exponents $p\in\mathbb{R}-\mathbb{N}$ too. However, all this is quite non-trivial.

Now, instead of struggling with the above conjecture, let us look at some other functions, which are not polynomial. And here, as good news, we have:

THEOREM 14.3. *We have the following integration formula,*

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

*valid for any two real numbers $a<b$.*

PROOF. This follows indeed from the Riemann integration formula, because:

$$\begin{aligned}
\int_a^b e^x dx &= \lim_{N\to\infty} \frac{e^a + e^{a+(b-a)/N} + e^{a+2(b-a)/N} + \ldots + e^{a+(N-1)(b-a)/N}}{N} \\
&= \lim_{N\to\infty} \frac{e^a}{N} \cdot \left(1 + e^{(b-a)/N} + e^{2(b-a)/N} + \ldots + e^{(N-1)(b-a)/N}\right) \\
&= \lim_{N\to\infty} \frac{e^a}{N} \cdot \frac{e^{b-a} - 1}{e^{(b-a)/N} - 1} \\
&= (e^b - e^a) \lim_{N\to\infty} \frac{1}{N(e^{(b-a)/N} - 1)} \\
&= e^b - e^a
\end{aligned}$$

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

Summarizing, we have some Riemann sum knowledge for the power functions, and the exponentials. For other functions, such as the trigonometric ones, such computations can be quite complicated, and so, with due apologies, we will have to stop our study here.

## 14b. Fundamental theorem

The problem is now, what to do with what we have, from the above. Not obvious, so stuck, and as always in such situations, time to ask the cat. And cat says:

CAT 14.4. *Summing the infinitesimals of the rate of change of the function should give you the global change of the function. Obvious.*

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

This being said, wait. There is suggestion to connect integrals and derivatives, and this is in fact what we have, coming from what we have from before, due to:

$$\left(\frac{x^{p+1}}{p+1}\right)' = x^p \quad , \quad (e^x)' = e^x$$

So, eureka, we have our idea, thanks cat. Moving ahead now, following this idea, we first have the following result, called fundamental theorem of calculus:

THEOREM 14.5. *Given a continuous function $f : [a, b] \to \mathbb{R}$, if we set*

$$F(x) = \int_a^x f(s)ds$$

*then $F' = f$. That is, the derivative of the integral is the function itself.*

PROOF. This follows from the Riemann integration picture, and more specifically, from the mean value property from chapter 13. Indeed, we have:

$$\frac{F(x+t)-F(x)}{t}=\frac{1}{t}\int_x^{x+t}f(x)dx$$

On the other hand, our function $f$ being continuous, by using the mean value property from chapter 13, we can find a number $c\in[x,x+t]$ such that:

$$\frac{1}{t}\int_x^{x+t}f(x)dx=f(x)$$

Thus, putting our formulae together, we conclude that we have:

$$\frac{F(x+t)-F(x)}{t}=f(c)$$

Now with $t\to 0$, no matter how the number $c\in[x,x+t]$ varies, one thing that we can be sure about is that we have $c\to x$. Thus, by continuity of $f$, we obtain:

$$\lim_{t\to 0}\frac{F(x+t)-F(x)}{t}=f(x)$$

But this means exactly that we have $F'=f$, and we are done. □

We have as well the following result, also called fundamental theorem of calculus:

THEOREM 14.6. *Given a function $F:\mathbb{R}\to\mathbb{R}$, we have*

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

*for any interval $[a,b]$.*

PROOF. As already mentioned, this is something which follows from Theorem 14.5, and is in fact equivalent to it. Indeed, consider the following function:

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

By using Theorem 14.5 we have $G'=F'$, and so our functions $F,G$ differ by a constant. But with $s=a$ we have $G(a)=0$, and so the constant is $F(a)$, and we get:

$$F(s)=G(s)+F(a)$$

Now with $s=b$ this gives $F(b)=G(b)+F(a)$, which reads:

$$F(b)=\int_a^b F'(x)dx+F(a)$$

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

As a first illustration for all this, solving our previous problems, we have:

THEOREM 14.7. *We have the following integration formulae,*

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

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

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

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

PROOF. We already know some of these formulae, but the best is to do everything, using the fundamental theorem of calculus. The computations go as follows:

(1) With $F(x) = x^{p+1}$ we have $F'(x) = px^p$, and we get, as desired:

$$\int_a^b px^p\, dx = b^{p+1} - a^{p+1}$$

(2) Observe first that the formula (1) does not work at $p = -1$. However, here we can use $F(x) = \log x$, having as derivative $F'(x) = 1/x$, which gives, as desired:

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

(3) With $F(x) = \cos x$ we have $F'(x) = -\sin x$, and we get, as desired:

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

(4) With $F(x) = \sin x$ we have $F'(x) = \cos x$, and we get, as desired:

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

(5) With $F(x) = e^x$ we have $F'(x) = e^x$, and we get, as desired:

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

(6) This is something more tricky. We are looking for a function satisfying:

$$F'(x) = \log x$$

This does not look doable, but fortunately the answer to such things can be found on the internet. But, what if the internet connection is down? So, let us think a bit, and try to solve our problem. Speaking logarithm and derivatives, what we know is:

$$(\log x)' = \frac{1}{x}$$

But then, in order to make appear log on the right, the idea is quite clear, namely multiplying on the left by $x$. We obtain in this way the following formula:

$$(x\log x)'=1\cdot\log x+x\cdot\frac{1}{x}=\log x+1$$

We are almost there, all we have to do now is to substract $x$ from the left, as to get:

$$(x\log x-x)'=\log x$$

But this this formula in hand, we can go back to our problem, and we get the result. □

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

DEFINITION 14.8. *Given $f$, we call primitive of $f$ any function $F$ satisfying:*

$$F'=f$$

*We denote such primitives by $\int f$, and also call them indefinite integrals.*

Observe that the primitives are unique up to an additive constant, in the sense that if $F$ is a primitive, then so is $F+c$, for any $c\in\mathbb{R}$, and conversely, if $F,G$ are two primitives, then we must have $G=F+c$, for some $c\in\mathbb{R}$, with this latter fact coming from a result from chapter 9, saying that the derivative vanishes when the function is constant.

As for the convention at the end, $F=\int f$, this comes from the fundamental theorem of calculus, which can be written as follows, by using this convention:

$$\int_a^b f(x)dx=\left(\int f\right)(b)-\left(\int f\right)(a)$$

By the way, observe that there is no contradiction here, coming from the indeterminacy of $\int f$. Indeed, when adding a constant $c\in\mathbb{R}$ to the chosen primitive $\int f$, when conputing the above difference the $c$ quantities will cancel, and we will obtain the same result.

We can now reformulate Theorem 14.7 in a more digest form, as follows:

THEOREM 14.9. *We have the following formulae for primitives,*

$$\int x^p=\frac{x^{p+1}}{p+1}\quad,\quad\int\frac{1}{x}=\log x$$

$$\int\sin x=-\cos x\quad,\quad\int\cos x=\sin x$$

$$\int e^x=e^x\quad,\quad\int\log x=x\log x-x$$

*allowing us to compute the corresponding definite integrals too.*

PROOF. Here the various formulae in the statement follow from Theorem 14.7, or rather from the proof of Theorem 14.7, or even from chapter 9, for most of them, and the last assertion comes from the integration formula given after Definition 14.8. □

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

THEOREM 14.10. *We have the formula*

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

*called integration by parts.*

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

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

Indeed, with our convention for primitives, this gives the formula in the statement. □

It is then possible to pass to usual integrals, and we obtain a formula here as well, as follows, also called integration by parts, with the convention $[\varphi]_a^b = \varphi(b) - \varphi(a)$:

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

In practice, the most interesting case is that when $fg$ vanishes on the boundary $\{a, b\}$ of our interval, leading to the following formula:

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

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

THEOREM 14.11. *We have the change of variable formula*

$$\int_a^b f(x)dx = \int_c^d f(\varphi(t))\varphi'(t)dt$$

*where $c = \varphi^{-1}(a)$ and $d = \varphi^{-1}(b)$.*

PROOF. This follows with $f = F'$, from the following differentiation rule, that we know from chapter 9, and whose proof is something elementary:

$$(F\varphi)'(t) = F'(\varphi(t))\varphi'(t)$$

Indeed, by integrating between $c$ and $d$, we obtain the result. □

As a main application now of our theory, in relation with advanced calculus, and more specifically with the Taylor formula from chapter 11, we have:

THEOREM 14.12. *Given a function $f : \mathbb{R} \to \mathbb{R}$, we have the formula*

$$f(x+t) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k + \int_x^{x+t} \frac{f^{(n+1)}(s)}{n!}(x+t-s)^n\, ds$$

*called Taylor formula with integral formula for the remainder.*

PROOF. This is something which looks a bit complicated, so we will first do some verifications, and then we will go for the proof in general:

(1) At $n = 0$ the formula in the statement is as follows, and certainly holds, due to the fundamental theorem of calculus, which gives $\int_x^{x+t} f'(s)ds = f(x+t) - f(x)$:

$$f(x+t) = f(x) + \int_x^{x+t} f'(s)ds$$

(2) At $n = 1$, the formula in the statement becomes more complicated, as follows:

$$f(x+t) = f(x) + f'(x)t + \int_x^{x+t} f''(s)(x+t-s)ds$$

As a first observation, this formula holds indeed for the linear functions, where we have $f(x+t) = f(x) + f'(x)t$, and $f'' = 0$. So, let us try $f(x) = x^2$. Here we have:

$$f(x+t) - f(x) - f'(x)t = (x+t)^2 - x^2 - 2xt = t^2$$

On the other hand, the integral remainder is given by the same formula, namely:

$$\begin{aligned}
\int_x^{x+t} f''(s)(x+t-s)ds &= 2\int_x^{x+t} (x+t-s)ds \\
&= 2t(x+t) - 2\int_x^{x+t} sds \\
&= 2t(x+t) - ((x+t)^2 - x^2) \\
&= 2tx + 2t^2 - 2tx - t^2 \\
&= t^2
\end{aligned}$$

(3) Still at $n = 1$, let us try now to prove the formula in the statement, in general. Since what we have to prove is an equality, this cannot be that hard, and the first thought goes towards differentiating. But this method works indeed, and we obtain the result.

(4) In general, the proof is similar, by differentiating, the computations being similar to those at $n = 1$, and we will leave this as an instructive exercise. □

## 14c. Basic applications

As a first application of our integration methods, we can now solve the 1D wave equation. In order to explain this, we will need a standard calculus result, as follows:

PROPOSITION 14.13. *The derivative of a function of type*

$$\varphi(x) = \int_{g(x)}^{h(x)} f(s)ds$$

*is given by the formula* $\varphi'(x) = f(h(x))h'(x) - f(g(x))g'(x)$.

PROOF. Consider a primitive of the function that we integrate, $F' = f$. We have:

$$\begin{aligned}\varphi(x) &= \int_{g(x)}^{h(x)} f(s)ds\\ &= \int_{g(x)}^{h(x)} F'(s)ds\\ &= F(h(x)) - F(g(x))\end{aligned}$$

By using now the chain rule for derivatives, we obtain from this:

$$\begin{aligned}\varphi'(x) &= F'(h(x))h'(x) - F'(g(x))g'(x)\\ &= f(h(x))h'(x) - f(g(x))g'(x)\end{aligned}$$

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

Now back to the 1D waves, the result here, due to d'Alembert, is as follows:

THEOREM 14.14. *The solution of the 1D wave equation $\ddot{\varphi} = v^2\varphi''$ with initial value conditions $\varphi(x,0) = f(x)$ and $\dot{\varphi}(x,0) = g(x)$ is given by the d'Alembert formula:*

$$\varphi(x,t) = \frac{f(x-vt) + f(x+vt)}{2} + \frac{1}{2v}\int_{x-vt}^{x+vt} g(s)ds$$

*Moreover, in the context of our previous lattice model discretizations, what happens is more or less that the above d'Alembert integral gets computed via Riemann sums.*

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

(1) Let us first check that the d'Alembert solution is indeed a solution of the wave equation $\ddot{\varphi} = v^2\varphi''$. The first time derivative is computed as follows:

$$\dot{\varphi}(x,t) = \frac{-vf'(x-vt) + vf'(x+vt)}{2} + \frac{1}{2v}(vg(x+vt) + vg(x-vt))$$

The second time derivative is computed as follows:

$$\ddot{\varphi}(x,t) = \frac{v^2f''(x-vt) + v^2f(x+vt)}{2} + \frac{vg'(x+vt) - vg'(x-vt)}{2}$$

Regarding now space derivatives, the first one is computed as follows:

$$\varphi'(x,t) = \frac{f'(x-vt) + f'(x+vt)}{2} + \frac{1}{2v}(g'(x+vt) - g'(x-vt))$$

As for the second space derivative, this is computed as follows:

$$\varphi''(x,t) = \frac{f''(x-vt) + f''(x+vt)}{2} + \frac{g''(x+vt) - g''(x-vt)}{2v}$$

Thus we have indeed $\ddot{\varphi} = v^2\varphi''$. As for the initial conditions, $\varphi(x,0) = f(x)$ is clear from our definition of $\varphi$, and $\dot{\varphi}(x,0) = g(x)$ is clear from our above formula of $\dot{\varphi}$.

(2) Conversely now, we can simply solve our equation, which among others will doublecheck the computations in (1). Let us make the following change of variables:

$$\xi = x - vt \quad , \quad \eta = x + vt$$

With this change of variables, which is quite tricky, mixing space and time variables, our wave equation $\ddot{\varphi} = v^2\varphi''$ reformulates in a very simple way, as follows:

$$\frac{d^2\varphi}{d\xi d\eta} = 0$$

But this latter equation tells us that our new $\xi, \eta$ variables get separated, and we conclude from this that the solution must be of the following special form:

$$\varphi(x,t) = F(\xi) + G(\eta) = F(x - vt) + G(x + vt)$$

Now by taking into account the intial conditions $\varphi(x,0) = f(x)$ and $\dot{\varphi}(x,0) = g(x)$, and then integrating, we are led to the d'Alembert formula. Finally, in what regards the last assertion, we will leave the study here as an instructive exercise. □

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

PROPOSITION 14.15. *The area of an ellipsis, given by the equation*

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$$

*with $a, b > 0$ being half the size of a box containing the ellipsis, is $A = \pi ab$.*

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

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

For any such $x$, the other coordinate $y$, satisfying $(x/a)^2 + (y/b)^2 = 1$, is given by:

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

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

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

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

$$\begin{aligned} A &= 2b\int_{-a}^{a}\sqrt{1-\frac{x^2}{a^2}}\,dx \\ &= \frac{4b}{a}\int_0^a\sqrt{a^2-x^2}\,dx \\ &= 4ab\int_0^1\sqrt{1-y^2}\,dy \\ &= 4ab\cdot\frac{\pi}{4} \\ &= \pi ab \end{aligned}$$

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

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

THEOREM 14.16. *The volume of the unit sphere is given by:*

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

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

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

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

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

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

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

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

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

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

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

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

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

## 14d. Some probability

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

DEFINITION 14.17. *A probability density is a function $\varphi:\mathbb{R}\to\mathbb{R}$ satisfying*

$$\varphi\geq 0\qquad,\qquad \int_{\mathbb{R}}\varphi(x)dx=1$$

*with the convention that we allow Dirac masses, $\delta_x$ with $x\in\mathbb{R}$, as components of $\varphi$.*

To be more precise, in what regards the convention at the end, which is something of physics flavor, this states that our density function $\varphi:\mathbb{R}\to\mathbb{R}$ must be a combination as follows, with $\psi:\mathbb{R}\to\mathbb{R}$ being a usual function, and with $\alpha_i,x_i\in\mathbb{R}$:

$$\varphi=\psi+\sum_i\alpha_i\delta_{x_i}$$

Assuming that $x_i$ are distinct, and with the usual convention that the Dirac masses integrate up to 1, the conditions on our density function $\varphi:\mathbb{R}\to\mathbb{R}$ are as follows:

$$\psi\geq 0\quad,\quad \alpha_i\geq 0\quad,\quad \int_{\mathbb{R}}\psi(x)dx+\sum_i\alpha_i=1$$

Observe the obvious relation with intuitive probability theory, where the probability for something to happen is always positive, $P\geq 0$, and where the overall probability for

something to happen, with this meaning for one of the possible events to happen, is of course $\Sigma P = 1$, and this because life goes on, and something must happen, right.

In short, what we are proposing with Definition 14.17 is some sort of continuous generalization of basic probability theory, coming from coins, dice and cards, that you know well. Moving now ahead, let us formulate, as a continuation of Definition 14.17:

DEFINITION 14.18. *We say that a random variable $f$ follows the density $\varphi$ if*

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

*holds, for any interval $[a,b] \subset \mathbb{R}$.*

With this, we are now one step closer to what we know from coins, dice, cards and so on. For instance when rolling a die, the corresponding density is as follows:

$$\varphi = \frac{1}{6}\left(\delta_1 + \delta_2 + \delta_3 + \delta_4 + \delta_5 + \delta_6\right)$$

In what regards now the random variables $f$, described as above by densities $\varphi$, the first questions regard their mean and variance, constructed as follows:

DEFINITION 14.19. *Given a random variable $f$, with probability density $\varphi$:*

(1) *Its mean is the quantity $M = \int_{\mathbb{R}} x\varphi(x)\,dx$.*
(2) *More generally, its $k$-th moment is $M_k = \int_{\mathbb{R}} x^k\varphi(x)\,dx$.*
(3) *Its variance is the quantity $V = M_2 - M_1^2$.*

Before going further, with more theory and examples, let us observe that, in both Definition 14.18 and Definition 14.19, what really matters is not the density $\varphi$ itself, but rather the related quantity $\mu = \varphi(x)dx$. So, let us upgrade our formalism, as follows:

DEFINITION 14.20 (upgrade). *A real probability measure is a quantity of the following type, with $\psi \geq 0$, $\alpha_i \geq 0$ and $x_i \in \mathbb{R}$, satisfying $\int_{\mathbb{R}} \psi(x)dx + \sum_i \alpha_i = 1$:*

$$\mu = \psi(x)dx + \sum_i \alpha_i \delta_{x_i}$$

*We say that a random variable $f$ follows $\mu$ when $P(f \in [a,b]) = \int_a^b d\mu(x)$. In this case*

$$M_k = \int_{\mathbb{R}} x^k d\mu(x)$$

*are called moments of $f$, and $M = M_1$ and $V = M_2 - M_1^2$ are called mean, and variance.*

In practice now, let us look for some illustrations for this. The simplest random variables are those following discrete laws, $\psi = 0$, and as a basic example here, when flipping a coin and being rewarded \$0 for heads, and \$1 for tails, the corresponding law is $\mu = \frac{1}{2}(\delta_0 + \delta_1)$. More generally, playing the same game with a biased coin, which lands on heads with probability $p \in (0, 1)$, leads to the following law, called Bernoulli law:

$$\mu = p\delta_0 + (1 - p)\delta_1$$

Many more things can be said here, notably with a study of what happens when you play the game $n$ times in a row, leading to some sort of powers of the Bernoulli laws, called binomial laws. In order to discuss this, let us first formulate:

DEFINITION 14.21. *Given $p \in [0, 1]$, the Bernoulli law of parameter $p$ is given by:*

$$P(\text{win}) = p \quad , \quad P(\text{lose}) = 1 - p$$

*More generally, the $k$-th binomial law of parameter $p$, with $k \in \mathbb{N}$, is given by*

$$P(s) = p^s(1 - p)^{k-s}\binom{k}{s}$$

*with the Bernoulli law appearing at $k = 1$, with $s = 1, 0$ here standing for win and lose.*

The point now is that the Bernoulli laws produce the binomial laws, simply by iterating the game, from 1 throw to $k \in \mathbb{N}$ throws. Obviously, what matters in all this is the "independence" of our coin throws, so let us record this finding, as follows:

THEOREM 14.22. *The following happen, in the context of the biased coin game:*

(1) *The Bernoulli laws $\mu_{ber}$ produce the binomial laws $\mu_{bin}$, by iterating the game $k \in \mathbb{N}$ times, via the independence of the throws.*
(2) *We have in fact $\mu_{bin} = \mu_{ber}^{*k}$, with $*$ being the convolution operation for real probability measures, given by $\delta_x * \delta_y = \delta_{x+y}$, and linearity.*

PROOF. This is something a bit informal, but let us prove this as stated, and we will come back later to it, with more details. In what regards the first assertion, this is what life teaches us. As for the second assertion, job for us to figure out what the formula there, $\mu_{bin} = \mu_{ber}^{*k}$, exactly means. And, this can be done as follows:

(1) The first idea is to encapsulate the data from Definition 14.21 into the probability measures associated to the Bernoulli and binomial laws. For the Bernoulli law, the corresponding measure is as follows, with the $\delta$ symbols standing for Dirac masses:

$$\mu_{ber} = (1 - p)\delta_0 + p\delta_1$$

As for the binomial law, here the measure is as follows, constructed in a similar way, you get the point I hope, again with the $\delta$ symbols standing for Dirac masses:

$$\mu_{bin} = \sum_{s=0}^{k} p^s(1 - p)^{k-s}\binom{k}{s}\delta_s$$

(2) Getting now to independence, the point is that, as we will soon discover abstractly, the mathematics there is that of the following formula, with $*$ standing for the convolution operation for the real measures, which is given by $\delta_x * \delta_y = \delta_{x+y}$ and linearity:

$$\mu_{bin} = \underbrace{\mu_{ber} * \ldots * \mu_{ber}}_{k\ terms}$$

(3) To be more precise, this latter formula does hold indeed, as a straightforward application of the binomial formula, the formal proof being as follows:

$$\begin{aligned}
\mu_{ber}^{*k} &= \big((1-p)\delta_0 + p\delta_1\big)^{*k} \\
&= \sum_{s=0}^{k} p^s(1-p)^{k-s}\binom{k}{s}\delta_0^{*(k-s)} * \delta_1^{*s} \\
&= \sum_{s=0}^{k} p^s(1-p)^{k-s}\binom{k}{s}\delta_s \\
&= \mu_{bin}
\end{aligned}$$

(4) Summarizing, save for some uncertainties regarding what independence exactly means, mathematically speaking, and more on this later, theorem proved. □

Skipping some further discussion here, and getting now straight to the point, the most important laws in discrete probability are the Poisson laws, constructed as follows:

DEFINITION 14.23. *The Poisson law of parameter* 1 *is the following measure,*

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

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

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

*with the letter "p" standing for Poisson.*

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

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

In general, the idea with the Poisson laws is that these appear a bit everywhere, in the real life, the reasons for this coming from the Poisson Limit Theorem (PLT). However, this theorem uses advanced calculus, and we will leave it for later. In the meantime, however, we can have some fun with moments, the result here being as follows:

THEOREM 14.24. *The moments of $p_1$ are the Bell numbers,*

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

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

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

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

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

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

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

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

Since the initial values coincide, $M_1 = B_1 = 1$ and $M_2 = B_2 = 2$, we obtain by recurrence $M_k = B_k$, as claimed. Regarding now the law $p_t$ with $t > 0$, we have here a

similar recurrence formula for the moments, as follows:

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

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

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

Now by using the above recurrence we obtain from this:

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

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

## 14e. Exercises

Exercises:

Exercise 14.25.

Exercise 14.26.

Exercise 14.27.

Exercise 14.28.

Exercise 14.29.

Exercise 14.30.

Exercise 14.31.

Exercise 14.32.

Bonus exercise.

CHAPTER 15

# Function spaces

## 15a. Normed spaces

Welcome to function space theory, also known as functional analysis. Although, for being fully honest, the basics here, which are quite algebraic, rather deserve the name "functional algebra". But do not worry, we will keep things as analytic as possible.

Let us start with something very general, as follows:

DEFINITION 15.1. *A normed space is a complex vector space $V$, which can be finite or infinite dimensional, together with a map*

$$||.|| : V \to \mathbb{R}_+$$

*called norm, subject to the following conditions:*

(1) $||x|| = 0$ *implies* $x = 0$.
(2) $||\lambda x|| = |\lambda| \cdot ||x||$, *for any* $x \in V$, *and* $\lambda \in \mathbb{C}$.
(3) $||x + y|| \leq ||x|| + ||y||$, *for any* $x, y \in V$.

As a basic example here, which is finite dimensional, we have the space $V = \mathbb{C}^N$, with the norm on it being the usual length of the vectors, namely:

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

Indeed, for this space (1) is clear, (2) is clear too, and (3) is something well-known, which is equivalent to the triangle inequality in $\mathbb{C}^N$, and which can be deduced from the Cauchy-Schwarz inequality. More on this, with some generalizations, in a moment.

Getting back now to the general case, we have the following result:

PROPOSITION 15.2. *Any normed vector space $V$ is a metric space, with*

$$d(x, y) = ||x - y||$$

*as distance. If this metric space is complete, we say that $V$ is a Banach space.*

PROOF. This follows from the definition of the metric spaces, as follows:

(1) The first distance axiom, $d(x, y) \geq 0$, and $d(x, y) = 0$ precisely when $x = y$, follows from the fact that the norm takes values in $\mathbb{R}_+$, and from $||x|| = 0 \implies x = 0$.

(2) The second distance axiom, which is the symmetry one, $d(x,y)=d(y,x)$, follows from our condition $||\lambda x||=|\lambda|\cdot||x||$, with $\lambda=-1$.

(3) As for the third distance axiom, which is the triangle inequality $d(x,y)\leq d(x,z)+d(y,z)$, this follows from our third norm axiom, namely $||x+y||\leq||x||+||y||$. □

Very nice all this, and it is possible to develop some general theory here, but before everything, however, we need more examples, besides $\mathbb{C}^N$ with its usual norm.

However, these further examples are actually quite tricky to construct, needing some inequality know-how. Let us start with a very basic result, as follows:

THEOREM 15.3 (Young). *We have the following inequality,*

$$ab\leq\frac{a^p}{p}+\frac{b^q}{q}$$

*valid for any $a,b\geq 0$, and any exponents $p,q>1$ satisfying $\frac{1}{p}+\frac{1}{q}=1$.*

PROOF. We use the logarithm function, which is concave on $(0,\infty)$, due to:

$$(\log x)''=\left(-\frac{1}{x}\right)'=-\frac{1}{x^2}$$

Thus we can apply the Jensen inequality, and we obtain in this way:

$$\begin{aligned}\log\left(\frac{a^p}{p}+\frac{b^q}{q}\right)&\geq\frac{\log(a^p)}{p}+\frac{\log(b^q)}{q}\\&=\log(a)+\log(b)\\&=\log(ab)\end{aligned}$$

Now by exponentiating, we obtain the Young inequality. □

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

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

In general, the Young inequality is something non-trivial, and the idea with it is that "when stuck with a problem, and with $ab\leq\frac{a^2+b^2}{2}$ not working, try Young".

We will be back to this general principle, in a moment, with some illustrations.

Moving forward now, as a consequence of the Young inequality, we have:

THEOREM 15.4 (Hölder). *Assuming that $p, q \geq 1$ are conjugate, in the sense that*

$$\frac{1}{p} + \frac{1}{q} = 1$$

*we have the following inequality, valid for any two vectors $x, y \in \mathbb{C}^N$,*

$$\sum_i |x_i y_i| \leq \left( \sum_i |x_i|^p \right)^{1/p} \left( \sum_i |y_i|^q \right)^{1/q}$$

*with the convention that an $\infty$ exponent produces a $\max |x_i|$ quantity.*

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

(1) Assume first that we are dealing with finite exponents, $p, q \in (1, \infty)$. By linearity we can assume that $x, y$ are normalized, in the following way:

$$\sum_i |x_i|^p = \sum_i |y_i|^q = 1$$

In this case, we want to prove that the following inequality holds:

$$\sum_i |x_i y_i| \leq 1$$

For this purpose, we use the Young inequality, which gives, for any $i$:

$$|x_i y_i| \leq \frac{|x_i|^p}{p} + \frac{|y_i|^q}{q}$$

By summing now over $i = 1, \ldots, N$, we obtain from this, as desired:

$$\begin{aligned} \sum_i |x_i y_i| &\leq \sum_i \frac{|x_i|^p}{p} + \sum_i \frac{|y_i|^q}{q} \\ &= \frac{1}{p} + \frac{1}{q} \\ &= 1 \end{aligned}$$

(2) In the case $p = 1$ and $q = \infty$, or vice versa, the inequality holds too, trivially, with the convention that an $\infty$ exponent produces a max quantity, according to:

$$\lim_{p \to \infty} \left( \sum_i |x_i|^p \right)^{1/p} = \max |x_i|$$

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

As a consequence now of the Hölder inequality, we have:

THEOREM 15.5 (Minkowski). *Assuming $p \in [1,\infty]$, we have the inequality*

$$\left(\sum_i |x_i+y_i|^p\right)^{1/p} \leq \left(\sum_i |x_i|^p\right)^{1/p} + \left(\sum_i |y_i|^p\right)^{1/p}$$

*for any two vectors $x, y \in \mathbb{C}^N$, with our usual conventions at $p = \infty$.*

PROOF. We have indeed the following estimate, using the Hölder inequality, and the conjugate exponent $q \in [1,\infty]$, given by $1/p + 1/q = 1$:

$$\begin{aligned}
\sum_i |x_i+y_i|^p &= \sum_i |x_i+y_i| \cdot |x_i+y_i|^{p-1} \\
&\leq \sum_i |x_i| \cdot |x_i+y_i|^{p-1} + \sum_i |y_i| \cdot |x_i+y_i|^{p-1} \\
&\leq \left(\sum_i |x_i|^p\right)^{1/p} \left(\sum_i |x_i+y_i|^{(p-1)q}\right)^{1/q} \\
&+ \left(\sum_i |y_i|^p\right)^{1/p} \left(\sum_i |x_i+y_i|^{(p-1)q}\right)^{1/q} \\
&= \left[\left(\sum_i |x_i|^p\right)^{1/p} + \left(\sum_i |y_i|^p\right)^{1/p}\right] \left(\sum_i |x_i+y_i|^p\right)^{1-1/p}
\end{aligned}$$

Here we have used the following fact, at the end:

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

Now by dividing both sides by the last quantity at the end, we obtain:

$$\left(\sum_i |x_i+y_i|^p\right)^{1/p} \leq \left(\sum_i |x_i|^p\right)^{1/p} + \left(\sum_i |y_i|^p\right)^{1/p}$$

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

Good news, done with inequalities, and as a consequence of the above results, and more specifically of the Minkowski inequality obtained above, we can formulate:

THEOREM 15.6. *Given an exponent $p \in [1,\infty]$, the formula*

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

*with usual conventions at $p = \infty$, defines a norm on $\mathbb{C}^N$, making it a Banach space.*

PROOF. Here the normed space assertion follows from the Minkowski inequality, established above, and the Banach space assertion is trivial, because our space being finite dimensional, by standard linear algebra all the Cauchy sequences converge. □

Very nice all this, but you might wonder at this point, what is the relation of all this with functions. In answer, Theorem 15.6 can be reformulated as follows:

THEOREM 15.7. *Given an exponent $p \in [1, \infty]$, the formula*

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

*with usual conventions at $p = \infty$, defines a norm on the space of functions*

$$f : \{1, \ldots, N\} \to \mathbb{C}$$

*making it a Banach space.*

PROOF. This is a just fancy reformulation of Theorem 15.6, by using the fact that the space formed by the functions $f : \{1, \ldots, N\} \to \mathbb{C}$ is canonically isomorphic to $\mathbb{C}^N$, in the obvious way, and by replacing the sums from the $\mathbb{C}^N$ context with integrals with respect to the counting measure on $\{1, \ldots, N\}$, in the function context. □

## 15b. Banach spaces

Moving now towards infinite dimensions and more standard analysis, the idea will be that of extending Theorem 15.7 to the arbitrary measured spaces. Let us start with:

THEOREM 15.8. *Given an exponent $p \in [1, \infty]$, the formula*

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

*with usual conventions at $p = \infty$, defines a norm on the space of sequences*

$$l^p = \left\{ (x_i)_{i \in \mathbb{N}} \Big| \sum_i |x_i|^p < \infty \right\}$$

*making it a Banach space.*

PROOF. As before with the finite sequences, the normed space assertion follows from the Minkowski inequality, established above, which extends without problems to the case of the infinite sequences, and with the Banach space assertion being clear too. □

We can unify and generalize what we have, in the following way:

THEOREM 15.9. *Given a discrete measured space $X$, and an exponent $p \in [1, \infty]$,*

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

*with usual conventions at $p = \infty$, defines a norm on the space of functions*

$$l^p(X) = \left\{ f : X \to \mathbb{C} \Big| \int_X |f(x)|^p < \infty \right\}$$

*making it a Banach space.*

PROOF. This is just a fancy reformulation of what we have:

(1) The case where $X$ is finite corresponds to Theorem 15.7.

(2) The case where $X$ is countable corresponds to Theorem 15.8.

(3) Finally, the case where $X$ is uncountable is easy to deal with too, by using the same arguments as in the countable case. □

In order to further extend the above result, to the case of the arbitrary measured spaces $X$, which are not necessarily discrete, let us start with:

THEOREM 15.10. *Given two functions $f, g : X \to \mathbb{C}$ and an exponent $p \geq 1$, we have*

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

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

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

*called Hölder inequality. These inequalities hold as well for $\infty$ values of the exponents.*

PROOF. This is very standard, exactly as in the case of sequences, finite or not, but since the above inequalities are really very general and final, here are the details:

(1) Let us first prove Hölder, in the case of finite exponents, $p, q \in (1, \infty)$. By linearity we can assume that $f, g$ are normalized, in the following way:

$$\int_X |f|^p = \int_X |g|^q = 1$$

In this case, we want to prove that the following inequality holds:

$$\int_X |fg| \leq 1$$

For this purpose, we use the Young inequality, which gives, for any $x \in X$:

$$|f(x)g(x)| \leq \frac{|f(x)|^p}{p} + \frac{|g(x)|^q}{q}$$

By integrating now over $x \in X$, we obtain from this, as desired:

$$\begin{aligned}\int_X |fg| &\leq \int_X \frac{|f(x)|^p}{p} + \int_X \frac{|g(x)|^q}{q}\\ &= \frac{1}{p} + \frac{1}{q}\\ &= 1\end{aligned}$$

(2) Let us prove now Minkowski, again in the finite exponent case, $p \in (1, \infty)$. We have the following estimate, using the Hölder inequality, and the conjugate exponent:

$$\begin{aligned}\int_X |f+g|^p &= \int_X |f+g| \cdot |f+g|^{p-1}\\ &\leq \int_X |f| \cdot |f+g|^{p-1} + \int_X |g| \cdot |f+g|^{p-1}\\ &\leq \left(\int_X |f|^p\right)^{1/p} \left(\int_X |f+g|^{(p-1)q}\right)^{1/q}\\ &+ \left(\int_X |g|^p\right)^{1/p} \left(\int_X |f+g|^{(p-1)q}\right)^{1/q}\\ &= \left[\left(\int |f|^p\right)^{1/p} + \left(\int_X |g|^p\right)^{1/p}\right] \left(\int_X |f+g|^p\right)^{1-1/p}\end{aligned}$$

Thus, we are led to the Minkowski inequality in the statement.

(3) Finally, in the infinite exponent cases we have similar results, which are trivial this time, with the convention that an $\infty$ exponent produces an essential supremum:

$$\lim_{p \to \infty} \left(\int_X |f|^p\right)^{1/p} = \text{ess sup}|f|$$

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

We can now extend Theorem 15.9, into something very general, as follows:

THEOREM 15.11. *Given a measured space $X$, and $p \in [1, \infty]$, the following space, with the convention that functions are identified up to equality almost everywhere,*

$$L^p(X) = \left\{f : X \to \mathbb{C} \Big| \int_I |f(x)|^p dx < \infty\right\}$$

*is a vector space, and the following quantity*

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

*is a norm on it, making it a Banach space.*

PROOF. This follows indeed from Theorem 15.10, with due attention to the null sets, and this because of the first normed space axiom, namely:

$$||x|| = 0 \implies x = 0$$

To be more precise, in order for this axiom to hold, we must identify the functions up to equality almost everywhere, as indicated in the statement. □

Very nice all this. So, we have our examples of Banach spaces, which look definitely interesting, and related to analysis. In the remainder of this chapter we will develop some general Banach space theory, and apply it to the above $L^p$ spaces.

Getting now to work, as a first result about the abstract normed spaces, we would like to talk about the linear maps $T : V \to W$, which can be thought of as being some kind of infinite matrices, when $V, W$ are infinite dimensional. We first have here:

PROPOSITION 15.12. *For a linear map $T : V \to W$, the following conditions are equivalent, and if they hold, we say that $T$ is bounded:*

(1) *$T$ is continuous.*
(2) *$T$ is continuous at $0$.*
(3) *$T$ maps the unit ball of $V$ into something bounded.*
(4) *$T$ is bounded, in the sense that $||T|| = \sup_{||x||=1} ||Tx||$ is finite.*

PROOF. Here the equivalences (1) $\iff$ (2) $\iff$ (3) $\iff$ (4) all follow from definitions, by using the linearity of $T$, and performing various rescalings, and with the number $||T||$ needed in (4) being the bound coming from (3). □

With the above result in hand, we can now formulate:

THEOREM 15.13. *Given two Banach spaces $V, W$, the bounded linear maps*

$$T : V \to W$$

*form a linear space $B(V, W)$, on which the following quantity is a norm,*

$$||T|| = \sup_{||x||=1} ||Tx||$$

*making $B(V, W)$ a Banach space. When $V = W$, we obtain a Banach algebra.*

PROOF. All this is very standard, and in the case $V = W$, for simplifying, which is the one that matters the most, the proof goes as follows:

(1) The fact that we have indeed an algebra, satisfying the product condition in the statement, follows from the following estimates, which are all elementary:

$$||S + T|| \leq ||S|| + ||T||$$
$$||\lambda T|| = |\lambda| \cdot ||T||$$
$$||ST|| \leq ||S|| \cdot ||T||$$

(2) Regarding now the last assertion, if $\{T_n\} \subset B(V)$ is Cauchy then $\{T_n x\}$ is Cauchy for any $x \in V$, so we can define the limit $T = \lim_{n\to\infty} T_n$ by setting:

$$Tx = \lim_{n\to\infty} T_n x$$

Let us first check that the application $x \to Tx$ is linear. We have:

$$\begin{aligned} T(x+y) &= \lim_{n\to\infty} T_n(x+y) \\ &= \lim_{n\to\infty} T_n(x) + T_n(y) \\ &= \lim_{n\to\infty} T_n(x) + \lim_{n\to\infty} T_n(y) \\ &= T(x) + T(y) \end{aligned}$$

Similarly, we have as well the following computation:

$$\begin{aligned} T(\lambda x) &= \lim_{n\to\infty} T_n(\lambda x) \\ &= \lambda \lim_{n\to\infty} T_n(x) \\ &= \lambda T(x) \end{aligned}$$

Thus we have a linear map $T : A \to A$. It remains to prove that we have $T \in B(V)$, and that we have $T_n \to T$ in norm. For this purpose, observe that we have:

$$\begin{aligned} & ||T_n - T_m|| \leq \varepsilon \ , \ \forall n, m \geq N \\ \implies & ||T_n x - T_m x|| \leq \varepsilon \ , \ \forall ||x|| = 1 \ , \ \forall n, m \geq N \\ \implies & ||T_n x - Tx|| \leq \varepsilon \ , \ \forall ||x|| = 1 \ , \ \forall n \geq N \\ \implies & ||T_N x - Tx|| \leq \varepsilon \ , \ \forall ||x|| = 1 \\ \implies & ||T_N - T|| \leq \varepsilon \end{aligned}$$

As a first consequence, we obtain $T \in B(V)$, because we have:

$$\begin{aligned} ||T|| &= ||T_N + (T - T_N)|| \\ &\leq ||T_N|| + ||T - T_N|| \\ &\leq ||T_N|| + \varepsilon \\ &< \infty \end{aligned}$$

As a second consequence, we obtain $T_N \to T$ in norm, and we are done. □

As a basic example for the above construction, in the case where both our spaces are finite dimensional, $V = \mathbb{C}^N$ and $W = \mathbb{C}^M$, with $N, M < \infty$, we obtain a matrix space:

$$B(\mathbb{C}^N, \mathbb{C}^M) = M_{M\times N}(\mathbb{C})$$

More on this later. On the other hand, of particular interest is as well the case $W = \mathbb{C}$ of the above construction, which leads to the following result:

THEOREM 15.14. *Given a Banach space $V$, its dual space, constructed as*

$$V^* = \left\{ f : V \to \mathbb{C},\ \text{linear and bounded} \right\}$$

*is a Banach space too, with norm given by:*

$$||f|| = \sup_{||x||=1} |f(x)|$$

*When $V$ is finite dimensional, we have $V \simeq V^*$.*

PROOF. This is clear indeed from Theorem 15.13, because we have:

$$V^* = B(V, \mathbb{C})$$

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

In order to better understand the linear forms, we will need:

THEOREM 15.15 (Hahn-Banach). *Given a Banach space $V$, the following happen:*

(1) *Given $x \in V - \{0\}$, there exists $f \in V^*$ with $f(x) \neq 0$.*

(2) *Given a subspace $W \subset V$, any $f \in W^*$ extends into a $\tilde{f} \in V^*$, of same norm.*

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

(1) As a first observation, (1) is weaker than (2).

(2) As a second observation, (2) can be proved in finite dimensions by using a direct sum decomposition $V = W \oplus U$, and setting $\tilde{f} \in V^*$ to be zero on $U$.

(3) In general, the proof is quite similar, by using the same ideas. To be more precise, we can first prove (1), and then by using using this, prove (2) as well. □

We can now formulate a key result, as follows:

THEOREM 15.16. *Given a Banach space $V$, we have an embedding as follows,*

$$V \subset V^{**}$$

*which is an isomorphism in finite dimensions, and for the $l^p$ and $L^p$ spaces too.*

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

(1) The fact that we have indeed a vector space embedding $V \subset V^{**}$ is clear from definitions, the formula of this embedding being as follows:

$$i(v)[f] = f(v)$$

(2) However, the fact that this embedding $V \subset V^{**}$ is isometric is something more subtle, which requires the use of the Hahn-Banach result from Theorem 15.15.

(3) Next, the fact that we have $V = V^{**}$ in finite dimensions is clear.

(4) Regarding now the formula $V = V^{**}$ for the various $l^p$ and $L^p$ spaces, this is something quite tricky. Let us start with the simplest case, that of the space $V = l^2$. We know that this space is given by definition by the following formula:

$$l^2 = \left\{ (x_i)_{i\in\mathbb{N}} \middle| \sum_i x_i^2 < \infty \right\}$$

Now let us look for linear forms $f : l^2 \to \mathbb{C}$. By linearity such a linear form must appear as follows, for certain scalars $a_i \in \mathbb{C}$, which must be such that $f$ is well-defined:

$$f\left((x_i)_i\right) = \sum_i a_i x_i$$

But, what does the fact that $f$ is well-defined mean? In answer, this means that the values of $f$ must all converge, which in practice means that we must have:

$$\sum_i x_i^2 < \infty \implies \left| \sum_i a_i x_i \right| < \infty$$

Moreover, we would like our linear form $f : l^2 \to \mathbb{C}$ to be bounded, and by denoting by $A = ||f|| < \infty$ the minimal bound, this means that we must have:

$$\left| \sum_i a_i x_i \right| \leq A \sqrt{\sum_i x_i^2}$$

Now recall that the Cauchy-Schwarz inequality tells us that we have:

$$\left| \sum_i a_i x_i \right| \leq \sqrt{\sum_i a_i^2} \cdot \sqrt{\sum_i x_i^2}$$

Thus, the linear form $f : l^2 \to \mathbb{C}$ associated to any $a = (a_i) \in l^2$ will do. Moreover, conversely, by examining the proof of Cauchy-Schwarz, we conclude that this condition $a = (a_i) \in l^2$ is in fact necessary. Thus, we have proved that we have:

$$(l^2)^* = l^2$$

But this gives the $V = V^{**}$ result in the statement for our space $V = l^2$, because by dualizing one more time we obtain, as desired:

$$(l^2)^{**} = (l^2)^* = l^2$$

(5) Getting now to more complicated spaces, let us look, more generally, at $L^2(X)$. We know that this space is given by definition by the following formula:

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

As before, when looking for linear forms $\varphi : L^2(X) \to \mathbb{C}$, by linearity, and with some mesure theory helping, our forms must appear via a formula as follows:

$$\varphi(f) = \int_X f(x)\varphi(x)dx$$

Now in order for this integral to converge, as for our map $\varphi : L^2(X) \to \mathbb{C}$ to be well-defined, and with the additional requirement that $\varphi$ must be actually bounded, we must have an inequality as follows, for a certain positive constant $A < \infty$:

$$\left|\int_X f(x)\varphi(x)dx\right| \leq A\sqrt{\int_X f(x)^2dx}$$

Now recall that the Cauchy-Schwarz inequality tells us that we have:

$$\left|\int_X f(x)\varphi(x)dx\right| \leq \sqrt{\int_X \varphi(x)^2dx} \cdot \sqrt{\int_X f(x)^2dx}$$

Thus, the linear form $\varphi : L^2(X) \to \mathbb{C}$ associated to any $\varphi \in L^2(X)$ will do. Moreover, conversely, by examining the proof of Cauchy-Schwarz, we conclude that this condition $\varphi \in L^2(X)$ is in fact necessary. Thus, we have proved that we have:

$$(L^2)^* = L^2$$

But this gives the $V = V^{**}$ result in the statement for our space $V = L^2$, because by dualizing one more time we obtain, as desired:

$$(L^2)^{**} = (L^2)^* = L^2$$

(6) Before getting further, let us mention that, more generally with respect to our $l^2, L^2$ computations, we have the following formula, valid for any Hilbert space $H$:

$$H^* \simeq \bar{H}$$

To be more precise, we can talk about Hilbert spaces, as being those Banach spaces whose norm comes from a scalar product, via $||x|| = \sqrt{< x, x >}$, and we will discuss this in the next chapter. And, the point is that, as we will see in the next chapter, any such Hilbert space has an orthogonal basis, which in practice means that we can write:

$$H = l^2(I)$$

Thus, we are apparently led to $H^* = H$, but this is not exactly true, because the correspondence $a \to f$ that we constructed in (4), and that we would like to rely upon, is antilinear, instead of being linear. Of course, this was not a problem in the context of (4), and nor is this a problem, for the same reasons, for a Hilbert space $H$ given with a basis, and so with an explicit isomorphism $H = l^2(I)$, as above. However, when talking about abstract Hilbert spaces $H$, coming without a basis, we must correct this, into:

$$H^* \simeq \bar{H}$$

But this gives the $V = V^{**}$ result in the statement for our Hilbert space $V = H$, because by dualizing one more time we obtain, as desired:

$$H^{**} = (\bar{H})^* = \bar{\bar{H}} = H$$

So long for $l^2, L^2$ spaces, and more general Hilbert spaces $H$. We will be back to this in the next chapter, when systematically discussing the Hilbert spaces.

(7) Moving ahead now, let us go back to the $l^p$ spaces, as in (4), but now with general exponents $p \in [1, \infty]$, instead of $p = 2$. The space $l^p$ is by definition given by:

$$l^p = \left\{ (x_i)_{i \in \mathbb{N}} \Big| \sum_i |x_i|^p < \infty \right\}$$

Now by arguing as in (4), a linear form $f : l^p \to \mathbb{C}$ must come as follows:

$$f\left((x_i)_i\right) = \sum_i a_i x_i$$

To be more precise, here $a_i \in \mathbb{C}$ are certain scalars, which are subject to an inequality as follows, for a certain constant $A < \infty$, making $f$ well-defined, and bounded:

$$\left| \sum_i a_i x_i \right| \leq A \left( \sum_i |x_i|^p \right)^{1/p}$$

Now recall that the Hölder inequality tells us that we have, with $\frac{1}{p} + \frac{1}{q} = 1$:

$$\left| \sum_i a_i x_i \right| \leq \left( \sum_i |x_i|^p \right)^{1/p} \left( \sum_i |a_i|^p \right)^{1/q}$$

Thus, the linear form $f : l^2 \to \mathbb{C}$ associated to any element $a = (a_i) \in l^q$ will do. Moreover, conversely, by examining the proof of Hölder, we conclude that this condition $a = (a_i) \in l^q$ is in fact necessary. Thus, we have proved that we have:

$$(l^p)^* = l^q$$

But this gives the $V = V^{**}$ result in the statement for our space $V = l^p$, because by dualizing one more time we obtain, as desired:

$$(l^p)^{**} = (l^q)^* = l^q$$

(8) All this is very nice, and time now to generalize everything that we know, by looking at the general spaces $L^p(X)$, with $p \in [1, \infty]$. These spaces are given by:

$$L^p(X) = \left\{ f : X \to \mathbb{C} \Big| \int_X |f(x)|^p dx < \infty \right\}$$

As before in (5), when looking for linear forms $\varphi : L^p(X) \to \mathbb{C}$, by linearity, and with some mesure theory helping, our forms must appear via a formula as follows:

$$\varphi(f) = \int_X f(x)\varphi(x)dx$$

Now in order for this integral to converge, as for our map $\varphi : L^p(X) \to \mathbb{C}$ to be well-defined, and with the additional requirement that $\varphi$ must be actually bounded, we must have an inequality as follows, for a certain positive constant $A < \infty$:

$$\left|\int_X f(x)\varphi(x)dx\right| \leq A\left(\int_X |f(x)|^p dx\right)^{1/p}$$

Now recall that the Hölder inequality tells us that we have, with $\frac{1}{p} + \frac{1}{q} = 1$:

$$\left|\int_X f(x)\varphi(x)dx\right| \leq \left(\int_X |f(x)|^p dx\right)^{1/p}\left(\int_X |\varphi(x)|^q dx\right)^{1/q}$$

Thus, the linear form $\varphi : L^p(X) \to \mathbb{C}$ associated to any function $\varphi \in L^q(X)$ will do. Moreover, conversely, by examining the proof of Hölder, we conclude that this condition $\varphi \in L^q(X)$ is in fact necessary. Thus, we have proved that we have:

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

But this gives the $V = V^{**}$ result in the statement for our space $V = L^p$, because by dualizing one more time we obtain, as desired:

$$(L^p)^{**} = (L^q)^* = L^p$$

(9) Finally, let us mention that not all Banach spaces satisfy $V = V^{**}$, with a basic counterexample here being the space $c_0$ of sequences $x_n \in \mathbb{C}$ satisfying $x_n \to 0$, with the sup norm. Indeed, computations show that we have the following formulae:

$$c_0^* = l^1 \quad , \quad (l^1)^* = l^\infty$$

Thus, in this case $V \subset V^{**}$ is the embedding $c_0 \subset l^\infty$, which is not an isomorphism. □

## 15c. Spectral theory

Many interesting things can be said about the Banach algebras:

DEFINITION 15.17. *A Banach algebra is a complex algebra with a norm satisfying*

$$||ab|| \leq ||a|| \cdot ||b||$$

*and which makes it a Banach space, in the sense that the Cauchy sequences converge.*

The basic examples of Banach algebras are the operator algebra $B(H)$, and its norm closed subalgebras $A \subset B(H)$, such as the algebras $A =< T >$ generated by a single operator $T \in B(H)$. There are many other examples, and more on this later.

Generally speaking, the elements $a \in A$ of a Banach algebra can be thought of as being bounded operators on some Hilbert space, which is not present. With this idea in mind, we can emulate spectral theory in our setting, the starting point being:

DEFINITION 15.18. *The spectrum of an element $a \in A$ is the set*

$$\sigma(a) = \left\{\lambda \in \mathbb{C}\middle| a - \lambda \notin A^{-1}\right\}$$

*where $A^{-1} \subset A$ is the set of invertible elements.*

As a basic example, the spectrum of a usual matrix $M \in M_N(\mathbb{C})$ is the collection of its eigenvalues, taken of course without multiplicities. In the case of the trivial algebra $A = \mathbb{C}$, appearing at $N = 1$, the spectrum of an element is the element itself.

Given an arbitrary Banach algebra element $a \in A$, and a rational function $f = P/Q$ having poles outside the spectrum $\sigma(a)$, we can construct the following element:

$$f(a) = P(a)Q(a)^{-1}$$

For simplicity, and due to the fact that the elements $P(a), Q(a)$ commute, so that the order is irrelevant, we write this element as a usual fraction, as follows:

$$f(a) = \frac{P(a)}{Q(a)}$$

With this convention, we have the following result:

THEOREM 15.19. *We have the "rational functional calculus" formula*

$$\sigma(f(a)) = f(\sigma(a))$$

*valid for any rational function $f \in \mathbb{C}(X)$ having poles outside $\sigma(a)$.*

PROOF. In order to prove this result, we can proceed in two steps, as follows:

(1) Assume first that we are in the polynomial function case, $f \in \mathbb{C}[X]$. We pick a scalar $\lambda \in \mathbb{C}$, and we decompose the polynomial $f - \lambda$ into factors:

$$f(X) - \lambda = c(X - r_1) \ldots (X - r_n)$$

By using this formula, we have then, as desired:

$$\begin{aligned}\lambda \notin \sigma(f(a)) &\iff f(a) - \lambda \in A^{-1}\\ &\iff c(a - r_1)\ldots(a - r_n) \in A^{-1}\\ &\iff a - r_1, \ldots, a - r_n \in A^{-1}\\ &\iff r_1, \ldots, r_n \notin \sigma(a)\\ &\iff \lambda \notin f(\sigma(a))\end{aligned}$$

(2) Assume now that we are in the general rational function case, $f \in \mathbb{C}(X)$. We pick a scalar $\lambda \in \mathbb{C}$, we write $f = P/Q$, and we set:

$$F = P - \lambda Q$$

By using now what we found in (1), for this polynomial, we obtain:

$$\begin{aligned}\lambda \in \sigma(f(a)) &\iff F(a) \notin A^{-1}\\ &\iff 0 \in \sigma(F(a))\\ &\iff 0 \in F(\sigma(a))\\ &\iff \exists \mu \in \sigma(a), F(\mu) = 0\\ &\iff \lambda \in f(\sigma(a))\end{aligned}$$

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

Summarizing, we have so far a beginning of theory. Let us prove now something that we do not know yet, namely that the spectra are non-empty:

$$\sigma(a) \neq \emptyset$$

In the present Banach algebra setting, this is definitely something non-trivial. In order to establish this result, we will need a number of analytic preliminaries, as follows:

PROPOSITION 15.20. *The following happen:*

(1) $||T|| < 1 \implies (1 - T)^{-1} = 1 + T + T^2 + \ldots$
(2) *The set* $B(V)^{-1}$ *is open.*
(3) *The map* $T \to T^{-1}$ *is differentiable.*

PROOF. All these assertions are elementary, as follows:

(1) This follows as in the scalar case, the computation being as follows, provided that everything converges under the norm, which amounts in saying that $||T|| < 1$:

$$\begin{aligned}(1 - T)(1 + T + T^2 + \ldots) &= 1 - T + T - T^2 + T^2 - T^3 + \ldots\\ &= 1\end{aligned}$$

(2) Assuming $T \in B(V)^{-1}$, let us pick $S \in B(V)$ such that:

$$||T - S|| < \frac{1}{||T^{-1}||}$$

We have then the following estimate:

$$\begin{aligned} ||1 - T^{-1}S|| &= ||T^{-1}(T - S)|| \\ &\leq ||T^{-1}|| \cdot ||T - S|| \\ &< 1 \end{aligned}$$

Thus we have $T^{-1}S \in B(V)^{-1}$, and so $S \in B(V)^{-1}$, as desired.

(3) In the scalar case, the derivative of $f(t) = t^{-1}$ is $f'(t) = -t^{-2}$. In the present normed space setting the derivative is no longer a number, but rather a linear transformation, which can be found by developing $f(T) = T^{-1}$ at order 1, as follows:

$$\begin{aligned} (T + S)^{-1} &= ((1 + ST^{-1})T)^{-1} \\ &= T^{-1}(1 + ST^{-1})^{-1} \\ &= T^{-1}(1 - ST^{-1} + (ST^{-1})^2 - \ldots) \\ &\simeq T^{-1}(1 - ST^{-1}) \\ &= T^{-1} - T^{-1}ST^{-1} \end{aligned}$$

Thus $f(T) = T^{-1}$ is indeed differentiable, with derivative $f'(T)S = -T^{-1}ST^{-1}$. □

We can now formulate a key result about spectra, as follows:

THEOREM 15.21. *The spectrum of a bounded operator $T \in B(V)$ is:*

(1) *Compact.*
(2) *Contained in the disc $D_0(||T||)$.*
(3) *Non-empty.*

PROOF. This can be proved by using Proposition 15.20, as follows:

(1) In view of (2) below, it is enough to prove that $\sigma(T)$ is closed. But this follows from the following computation, with $|\varepsilon|$ being small:

$$\begin{aligned} \lambda \notin \sigma(T) &\implies T - \lambda \in B(V)^{-1} \\ &\implies T - \lambda - \varepsilon \in B(V)^{-1} \\ &\implies \lambda + \varepsilon \notin \sigma(T) \end{aligned}$$

(2) This follows from the following computation:

$$\begin{aligned} \lambda > ||T|| &\implies \left|\left|\frac{T}{\lambda}\right|\right| < 1 \\ &\implies 1 - \frac{T}{\lambda} \in B(V)^{-1} \\ &\implies \lambda - T \in B(V)^{-1} \\ &\implies \lambda \notin \sigma(T) \end{aligned}$$

(3) Assume by contradiction $\sigma(T) = \emptyset$. Given a linear form $f \in B(V)^*$, consider the following map, which is well-defined, due to our assumption $\sigma(T) = \emptyset$:

$$\varphi : \mathbb{C} \to \mathbb{C} \quad , \quad \lambda \to f((T - \lambda)^{-1})$$

By using the fact that $T \to T^{-1}$ is differentiable, that we know from Proposition 15.20, we conclude that this map is differentiable, and so holomorphic. Also, we have:

$$\begin{aligned} \lambda \to \infty &\implies T - \lambda \to \infty \\ &\implies (T - \lambda)^{-1} \to 0 \\ &\implies f((T - \lambda))^{-1} \to 0 \end{aligned}$$

Thus by the Liouville theorem we obtain $\varphi = 0$. But, in view of the definition of $\varphi$, this gives $(T - \lambda)^{-1} = 0$, which is a contradiction, as desired. □

## 15d. Distributions

Distributions.

## 15e. Exercises

Exercises:

Exercise 15.22.

Exercise 15.23.

Exercise 15.24.

Exercise 15.25.

Exercise 15.26.

Exercise 15.27.

Exercise 15.28.

Exercise 15.29.

Bonus exercise.

CHAPTER 16

# Several variables

## 16a. Partial derivatives

Moving now to several variables, $N \geq 2$, as a first job, given a function $\varphi : \mathbb{R}^N \to \mathbb{R}$, we would like to find a quantity $\varphi'(x)$ making the following formula work:

$$\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h$$

But here, as in 1 variable, there are not so many choices, and the solution is that of defining $\varphi'(x)$ as being the row vector formed by the partial derivatives at $x$:

$$\varphi'(x) = \begin{pmatrix} \frac{d\varphi}{dx_1} & \ldots & \frac{d\varphi}{dx_N} \end{pmatrix}$$

To be more precise, with this value for $\varphi'(x)$, our approximation formula $\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h$ makes sense indeed, as an equality of real numbers, with $\varphi'(x)h \in \mathbb{R}$ being obtained as the matrix multiplication of the row vector $\varphi'(x)$, and the column vector $h$. As for the fact that our formula holds indeed, this follows by putting together the approximation properties of each of the partial derivatives $d\varphi/dx_i$, which give:

$$\varphi(x+h) \simeq \varphi(x) + \sum_{i=1}^{N} \frac{d\varphi}{dx_i} \cdot h_i = \varphi(x) + \varphi'(x)h$$

Before moving forward, you might say, why bothering with horizontal vectors, when it is so simple and convenient to have all vectors vertical, by definition. Good point, and in answer, we can indeed talk about the gradient of $\varphi$, constructed as follows:

$$\nabla\varphi = \begin{pmatrix} \frac{d\varphi}{dx_1} \\ \vdots \\ \frac{d\varphi}{dx_N} \end{pmatrix}$$

With this convention, $\nabla\varphi$ geometrically describes the slope of $\varphi$ at the point $x$, in the obvious way. However, the approximation formula must be rewritten as follows:

$$\varphi(x+h) \simeq \varphi(x) + < \nabla\varphi(x), h >$$

In what follows we will use both $\varphi'$ and $\nabla\varphi$, depending on the context. Moving now to second derivatives, the main result here is as follows:

THEOREM 16.1. *The second derivative of a function $\varphi : \mathbb{R}^N \to \mathbb{R}$, making the formula*

$$\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h + \frac{< \varphi''(x)h, h >}{2}$$

*work, is its Hessian matrix $\varphi''(x) \in M_N(\mathbb{R})$, given by the following formula:*

$$\varphi''(x) = \left( \frac{d^2\varphi}{dx_i dx_j} \right)_{ij}$$

*Moreover, this Hessian matrix is symmetric, $\varphi''(x)_{ij} = \varphi'(x)_{ji}$.*

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

(1) As a first observation, at $N = 1$ the Hessian matrix constructed above is simply the $1 \times 1$ matrix having as entry the second derivative $\varphi''(x)$, and the formula in the statement is something that we know well from chapter 10, namely:

$$\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h + \frac{\varphi''(x)h^2}{2}$$

(2) At $N = 2$ now, we obviously need to differentiate $\varphi$ twice, and the point is that we come in this way upon the following formula, called Clairaut formula:

$$\frac{d^2\varphi}{dxdy} = \frac{d^2\varphi}{dydx}$$

But, is this formula correct or not? As an intuitive justification for it, let us consider a product of power functions, $\varphi(z) = x^p y^q$. We have then our formula, due to:

$$\frac{d^2\varphi}{dxdy} = \frac{d}{dx}\left( \frac{dx^p y^q}{dy} \right) = \frac{d}{dx}\left( qx^p y^{q-1} \right) = pqx^{p-1}y^{q-1}$$

$$\frac{d^2\varphi}{dydx} = \frac{d}{dy}\left( \frac{dx^p y^q}{dx} \right) = \frac{d}{dy}\left( px^{p-1} y^q \right) = pqx^{p-1}y^{q-1}$$

Next, let us consider a linear combination of power functions, $\varphi(z) = \sum_{pq} c_{pq} x^p y^q$, which can be finite or not. We have then, by using the above computation:

$$\frac{d^2\varphi}{dxdy} = \frac{d^2\varphi}{dydx} = \sum_{pq} c_{pq} pq x^{p-1} y^{q-1}$$

Thus, we can see that our commutation formula for derivatives holds indeed, due to the fact that the functions in $x, y$ commute. Of course, all this does not fully prove our formula, in general. But exercise for you, to have this idea fully working, or to look up the standard proof of the Clairaut formula, using the mean value theorem.

(3) Moving now to $N = 3$ and higher, we can use here the Clairaut formula with respect to any pair of coordinates, which gives the Schwarz formula, namely:

$$\frac{d^2\varphi}{dx_i dx_j} = \frac{d^2\varphi}{dx_j dx_i}$$

Thus, the second derivative, or Hessian matrix, is symmetric, as claimed.

(4) Getting now to the main topic, namely approximation formula in the statement, in arbitrary $N$ dimensions, this is in fact something which does not need a new proof, because it follows from the one-variable formula in (1), applied to the restriction of $\varphi$ to the following segment in $\mathbb{R}^N$, which can be regarded as being a one-variable interval:

$$I = [x, x + h]$$

To be more precise, let $y \in \mathbb{R}^N$, and consider the following function, with $r \in \mathbb{R}$:

$$f(r) = \varphi(x + ry)$$

We know from (1) that the Taylor formula for $f$, at the point $r = 0$, reads:

$$f(r) \simeq f(0) + f'(0)r + \frac{f''(0)r^2}{2}$$

And our claim is that, with $h = ry$, this is precisely the formula in the statement.

(5) So, let us see if our claim is correct. By using the chain rule, we have the following formula, with on the right, as usual, a row vector multiplied by a column vector:

$$f'(r) = \varphi'(x + ry) \cdot y$$

By using again the chain rule, we can compute the second derivative as well:

$$\begin{aligned}
f''(r) &= (\varphi'(x + ry) \cdot y)' \\
&= \left( \sum_i \frac{d\varphi}{dx_i}(x + ry) \cdot y_i \right)' \\
&= \sum_i \sum_j \frac{d^2\varphi}{dx_i dx_j}(x + ry) \cdot \frac{d(x + ry)_j}{dr} \cdot y_i \\
&= \sum_i \sum_j \frac{d^2\varphi}{dx_i dx_j}(x + ry) \cdot y_i y_j \\
&= < \varphi''(x + ry)y, y >
\end{aligned}$$

(6) Time now to conclude. We know that we have $f(r) = \varphi(x + ry)$, and according to our various computations above, we have the following formulae:

$$f(0) = \varphi(x) \quad , \quad f'(0) = \varphi'(x) \quad , \quad f''(0) =< \varphi''(x)y, y >$$

Buit with this data in hand, the usual Taylor formula for our one variable function $f$, at order 2, at the point $r = 0$, takes the following form, with $h = ry$:

$$\begin{aligned}\varphi(x + ry) &\simeq \varphi(x) + \varphi'(x)ry + \frac{< \varphi''(x)y, y > r^2}{2}\\ &= \varphi(x) + \varphi'(x)t + \frac{< \varphi''(x)h, h >}{2}\end{aligned}$$

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

As before in the one variable case, many more things can be said, as a continuation of the above. For instance the local minima and maxima of $\varphi : \mathbb{R}^N \to \mathbb{R}$ appear at the points $x \in \mathbb{R}^N$ where the derivative vanishes, $\varphi'(x) = 0$, and where the second derivative $\varphi''(x) \in M_N(\mathbb{R})$ is positive, respectively negative. But, you surely know all this.

As a key observation now, generalizing what we know in 1 variable, we have:

PROPOSITION 16.2. *Intuitively, the following quantity, called Laplacian of $\varphi$,*

$$\Delta\varphi = \sum_{i=1}^{N} \frac{d^2\varphi}{dx_i^2}$$

*measures how much different is $\varphi(x)$, compared to the average of $\varphi(y)$, with $y \simeq x$.*

PROOF. As before with 1 variable, this is something a bit heuristic, but good to know. Let us write the formula in Theorem 16.1, as such, and with $h \to -h$ too:

$$\varphi(x + h) \simeq \varphi(x) + \varphi'(x)h + \frac{< \varphi''(x)h, h >}{2}$$

$$\varphi(x - h) \simeq \varphi(x) - \varphi'(x)h + \frac{< \varphi''(x)h, h >}{2}$$

By making the average, we obtain the following formula:

$$\frac{\varphi(x + h) + \varphi(x - h)}{2} = \varphi(x) + \frac{< \varphi''(x)h, h >}{2}$$

Thus, thinking a bit, we are led to the conclusion in the statement, modulo some discussion about integrating all this, that we will not really need, in what follows. □

With this understood, the problem is now, what can we say about the mathematics of $\Delta$? As a first observation, which is a bit speculative, the Laplace operator appears by

applying twice the gradient operator, in a somewhat formal sense, as follows:

$$\begin{aligned}\Delta\varphi &= \sum_{i=1}^{N}\frac{d^2\varphi}{dx_i^2}\\ &= \sum_{i=1}^{N}\frac{d}{dx_i}\cdot\frac{d\varphi}{dx_i}\\ &= \left\langle\begin{pmatrix}\frac{d}{dx_1}\\ \vdots\\ \frac{d}{dx_N}\end{pmatrix},\begin{pmatrix}\frac{d\varphi}{dx_1}\\ \vdots\\ \frac{d\varphi}{dx_N}\end{pmatrix}\right\rangle\\ &= <\nabla,\nabla\varphi>\end{aligned}$$

Thus, it is possible to write a formula of type $\Delta=\nabla^2$, with the convention that the square of the gradient $\nabla$ is taken in a scalar product sense, as above. However, this can be a bit confusing, and in what follows, we will not use this notation.

Instead of further thinking at this, and at double derivatives in general, let us formulate a more straightforward question, inspired by linear algebra, as follows:

QUESTION 16.3. *The Laplace operator being linear,*

$$\Delta(a\varphi+b\psi)=a\Delta\varphi+b\Delta\psi$$

*what can we say about it, inspired by usual linear algebra?*

In answer now, the space of functions $\varphi:\mathbb{R}^N\to\mathbb{R}$, on which $\Delta$ acts, being infinite dimensional, the usual tools from linear algebra do not apply as such, and we must be extremely careful. For instance, we cannot really expect to diagonalize $\Delta$, via some sort of explicit procedure, as we usually do in linear algebra, for the usual matrices.

Thinking some more, there is actually a real bug too with our problem, because at $N=1$ this problem becomes "what can we say about the second derivatives $\varphi'':\mathbb{R}\to\mathbb{R}$ of the functions $\varphi:\mathbb{R}\to\mathbb{R}$, inspired by linear algebra", with answer "not much".

And by thinking even more, still at $N=1$, there is a second bug too, because if $\varphi:\mathbb{R}\to\mathbb{R}$ is twice differentiable, nothing will guarantee that its second derivative $\varphi'':\mathbb{R}\to\mathbb{R}$ is twice differentiable too. Thus, we have some issues with the domain and range of $\Delta$, regarded as linear operator, and these problems will persist at higher $N$.

So, shall we trash Question 16.3? Not so quick, because, very remarkably, some magic comes at $N=2$ and higher in relation with complex analysis, according to:

PRINCIPLE 16.4. *The functions $\varphi : \mathbb{R}^N \to \mathbb{R}$ which are 0-eigenvectors of $\Delta$,*

$$\Delta\varphi = 0$$

*called harmonic functions, have the following properties:*

(1) *At $N = 1$, nothing spectacular, these are just the linear functions.*
(2) *At $N = 2$, these are, locally, the real parts of holomorphic functions.*
(3) *At $N \geq 3$, these still share many properties with the holomorphic functions.*

In order to understand this, or at least get introduced to it, let us first look at the case $N = 2$. Here, any function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ can be regarded as function $\varphi : \mathbb{C} \to \mathbb{R}$, depending on $z = x + iy$. But, in view of this, it is natural to enlarge the attention to the functions $\varphi : \mathbb{C} \to \mathbb{C}$, and ask which of these functions are harmonic, $\Delta\varphi = 0$. And here, we have the following remarkable result, making the link with complex analysis:

THEOREM 16.5. *Any holomorphic function $\varphi : \mathbb{C} \to \mathbb{C}$, when regarded as function*

$$\varphi : \mathbb{R}^2 \to \mathbb{C}$$

*is harmonic. Moreover, the conjugates $\bar{\varphi}$ of holomorphic functions are harmonic too.*

PROOF. The first assertion comes 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 $\Delta\bar{\varphi} = \overline{\Delta\varphi}$, which is clear from definitions, and which shows that if $\varphi$ is harmonic, then so is its conjugate $\bar{\varphi}$. □

Many more things can be said, along these lines.

## 16b. Multiple integrals

We can talk about multiple integrals, in the obvious way. Getting now to the general theory and rules, for computing such integrals, the key result here is the change of variable formula. In order to discuss this, let us start with something that we know well, in 1D:

PROPOSITION 16.6. *We have the change of variable formula*

$$\int_a^b f(x)dx = \int_c^d f(\varphi(t))\varphi'(t)dt$$

*where $c = \varphi^{-1}(a)$ and $d = \varphi^{-1}(b)$.*

PROOF. This follows with $f = F'$, via the following differentiation rule:

$$(F\varphi)'(t) = F'(\varphi(t))\varphi'(t)$$

Indeed, by integrating between $c$ and $d$, we obtain the result. □

In several variables now, we can only expect the above $\varphi'(t)$ factor to be replaced by something similar, a sort of "derivative of $\varphi$, arising as a real number". But this can only be the Jacobian $\det(\varphi'(t))$, and with this in mind, we are led to:

THEOREM 16.7. *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 $J_\varphi$ quantity, called Jacobian, being given by*

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

*and with this generalizing the formula from Proposition 16.6.*

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 16.6, the point here being that the absolute value on the derivative appears as to compensate for the lack of explicit bounds for the integral.

(2) In general now, we can first argue that, the formula in the statement being linear in $f$, we can assume $f = 1$. Thus we want to prove $vol(E) = \int_{\varphi^{-1}(E)} |J_\varphi(t)|dt$, and with $D = \varphi^{-1}(E)$, this amounts in proving $vol(\varphi(D)) = \int_D |J_\varphi(t)|dt$.

(3) Now since this latter formula is additive with respect to $D$, it is enough to prove that $vol(\varphi(D)) = \int_D J_\varphi(t)dt$, for small cubes $D$, and assuming $J_\varphi > 0$. But this follows by using the usual definition of the determinant, as a volume.

(4) The details and computations however are quite non-trivial, and can be found for instance in Rudin [79]. So, please read that. With this, reading the complete proof of the present theorem from Rudin, being part of the standard math experience. □

Many other things can be said, as a continuation of the above.

## 16c. Spherical coordinates

Time now do some exciting computations, with the technology that we have. In what regards the applications of Theorem 16.7, these often come via:

PROPOSITION 16.8. *We have polar coordinates in 2 dimensions,*

$$\begin{cases} x &= r\cos t \\ y &= r\sin t \end{cases}$$

*the corresponding Jacobian being $J = r$.*

PROOF. This is elementary, the Jacobian being:

$$\begin{aligned} J &= \begin{vmatrix} \frac{d(r\cos t)}{dr} & \frac{d(r\cos t)}{dt} \\ \frac{d(r\sin t)}{dr} & \frac{d(r\sin t)}{dt} \end{vmatrix} \\ &= \begin{vmatrix} \cos t & -r\sin t \\ \sin t & r\cos t \end{vmatrix} \\ &= r\cos^2 t + r\sin^2 t \\ &= r \end{aligned}$$

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

We can now compute the Gauss integral, which is the best calculus formula ever:

THEOREM 16.9. *We have the following formula,*

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

*called Gauss integral formula.*

PROOF. Let $I$ be the above integral. By using polar coordinates, we obtain:

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

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

Moving now to 3 dimensions, we have here the following result:

PROPOSITION 16.10. *We have spherical coordinates in* 3 *dimensions,*

$$\begin{cases} x &= r\cos s \\ y &= r\sin s\cos t \\ z &= r\sin s\sin t \end{cases}$$

*the corresponding Jacobian being* $J(r,s,t) = r^2\sin s$.

PROOF. The fact that we have indeed spherical coordinates is clear. Regarding now the Jacobian, this is given by the following formula:

$$\begin{aligned} & J(r,s,t) \\ = & \begin{vmatrix} \cos s & -r\sin s & 0 \\ \sin s\cos t & r\cos s\cos t & -r\sin s\sin t \\ \sin s\sin t & r\cos s\sin t & r\sin s\cos t \end{vmatrix} \\ = & r^2\sin s\sin t\begin{vmatrix} \cos s & -r\sin s \\ \sin s\sin t & r\cos s\sin t \end{vmatrix} + r\sin s\cos t\begin{vmatrix} \cos s & -r\sin s \\ \sin s\cos t & r\cos s\cos t \end{vmatrix} \\ = & r\sin s\sin^2 t\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} + r\sin s\cos^2 t\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} \\ = & r\sin s(\sin^2 t+\cos^2 t)\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} \\ = & r\sin s\times 1\times r \\ = & r^2\sin s \end{aligned}$$

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

Let us work out now the general spherical coordinate formula, in arbitrary $N$ dimensions. The formula here, which generalizes those at $N = 2, 3$, is as follows:

THEOREM 16.11. *We have spherical coordinates in* $N$ *dimensions,*

$$\begin{cases} x_1 &= r\cos t_1 \\ x_2 &= r\sin t_1\cos t_2 \\ \vdots & \\ x_{N-1} &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\cos t_{N-1} \\ x_N &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\sin t_{N-1} \end{cases}$$

*the corresponding Jacobian being given by the following formula,*

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

*and with this generalizing the known formulae at* $N = 2, 3$.

PROOF. As before, the fact that we have spherical coordinates is clear. Regarding now the Jacobian, also as before, by developing over the last column, we have:

$$\begin{aligned} J_N &= r\sin t_1\ldots\sin t_{N-2}\sin t_{N-1}\times\sin t_{N-1}J_{N-1}\\ &+ r\sin t_1\ldots\sin t_{N-2}\cos t_{N-1}\times\cos t_{N-1}J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}(\sin^2 t_{N-1}+\cos^2 t_{N-1})J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}J_{N-1}\end{aligned}$$

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

As a comment here, the above convention for spherical coordinates is one among many, designed to best work in arbitrary $N$ dimensions. Also, in what regards the precise range of the angles $t_1,\ldots,t_{N-1}$, we will leave this to you, as an instructive exercise.

As an application, let us compute the volumes of spheres. For this purpose, we must understand how the products of coordinates integrate over spheres. Let us start with the case $N=2$. Here the sphere is the unit circle $\mathbb{T}$, and with $z=e^{it}$ the coordinates are $\cos t,\sin t$. We can first integrate arbitrary powers of these coordinates, as follows:

PROPOSITION 16.12. *We have the following formulae,*

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

*where $\varepsilon(p)=1$ if $p$ is even, and $\varepsilon(p)=0$ if $p$ is odd, and where*

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

*with the product ending at 2 if $m$ is odd, and ending at 1 if $m$ is even.*

PROOF. Let us first compute the integral on the left in the statement:

$$I_p=\int_0^{\pi/2}\cos^p t\,dt$$

We do this by partial integration. We have the following formula:

$$\begin{aligned}(\cos^p t\sin t)' &= p\cos^{p-1}t(-\sin t)\sin t+\cos^p t\cos t\\ &= p\cos^{p+1}t-p\cos^{p-1}t+\cos^{p+1}t\\ &= (p+1)\cos^{p+1}t-p\cos^{p-1}t\end{aligned}$$

By integrating between 0 and $\pi/2$, we obtain the following formula:

$$(p+1)I_{p+1}=pI_{p-1}$$

Thus we can compute $I_p$ by recurrence, and we obtain:

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

But $I_0 = \frac{\pi}{2}$ and $I_1 = 1$, so we get the result. As for the second formula, this follows from the first one, with $t = \frac{\pi}{2} - s$. Thus, we have proved both formulae in the statement. □

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

THEOREM 16.13. *The volume of the unit sphere in $\mathbb{R}^N$ is given by*

$$V = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!}$$

*with our usual convention $N!! = (N-1)(N-3)(N-5)\ldots$*

PROOF. Let us denote by $B^+$ the positive part of the unit sphere, or rather unit ball $B$, obtained by cutting this unit ball in $2^N$ parts. At the level of volumes, we have:

$$V = 2^N V^+$$

We have the following computation, using spherical coordinates:

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

Here we have used the following formula, for computing the exponent of $\pi/2$:

$$\begin{aligned}\varepsilon(0)+\varepsilon(1)+\varepsilon(2)+\ldots+\varepsilon(N-2) &= 1+0+1+\ldots+\varepsilon(N-2)\\ &= \left[\frac{N-2}{2}\right]+1\\ &= \left[\frac{N}{2}\right]\end{aligned}$$

Thus, we obtain the formula in the statement. □

As main particular cases of the above formula, we have:

THEOREM 16.14. *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 16.13, 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 $V=\frac{\pi^2}{4}\cdot\frac{16}{8}=\frac{\pi^2}{2}$. □

The formula in Theorem 16.13 is certainly nice, but in practice, we would like to have estimates for that sphere volumes too. For this purpose, we will need:

THEOREM 16.15. *We have the Stirling formula*

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

*valid in the $N\to\infty$ limit.*

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

(1) Let us first see what we can get with Riemann sums. We have:

$$\begin{aligned}\log(N!) &= \sum_{k=1}^{N}\log k\\ &\approx \int_1^N\log x\,dx\\ &= N\log N-N+1\end{aligned}$$

By exponentiating, this gives the following estimate, which is not bad:

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

(2) We can improve our estimate by replacing the rectangles from the Riemann sum approach to the integrals by trapezoids. In practice, this gives the following estimate:

$$\begin{aligned}
\log(N!) &= \sum_{k=1}^{N} \log k \\
&\approx \int_1^N \log x\, dx + \frac{\log 1 + \log N}{2} \\
&= N \log N - N + 1 + \frac{\log N}{2}
\end{aligned}$$

By exponentiating, this gives the following estimate, which gets us closer:

$$N! \approx \left(\frac{N}{e}\right)^N \cdot e \cdot \sqrt{N}$$

(3) In order to conclude, we must take some kind of mathematical magnifier, and carefully estimate the error made in (2). Fortunately, this mathematical magnifier exists, called Euler-Maclaurin formula, and after some computations, this leads to:

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

(4) However, all this remains a bit complicated, so we would like to present now an alternative approach to (3), which also misses some details, but better does the job, explaining where the $\sqrt{2\pi}$ factor comes from. First, by partial integration we have:

$$N! = \int_0^\infty x^N e^{-x} dx$$

Since the integrand is sharply peaked at $x = N$, as you can see by computing the derivative of $\log(x^N e^{-x})$, this suggests writing $x = N + y$, and we obtain:

$$\begin{aligned}
\log(x^N e^{-x}) &= N \log x - x \\
&= N \log(N + y) - (N + y) \\
&= N \log N + N \log\left(1 + \frac{y}{N}\right) - (N + y) \\
&\simeq N \log N + N\left(\frac{y}{N} - \frac{y^2}{2N^2}\right) - (N + y) \\
&= N \log N - N - \frac{y^2}{2N}
\end{aligned}$$

By exponentiating, we obtain from this the following estimate:

$$x^N e^{-x} \simeq \left(\frac{N}{e}\right)^N e^{-y^2/2N}$$

Now by integrating, and using the Gauss formula, we obtain from this:

$$\begin{aligned} N! &= \int_0^\infty x^N e^{-x} dx \\ &\simeq \int_{-N}^N \left(\frac{N}{e}\right)^N e^{-y^2/2N}\, dy \\ &\simeq \left(\frac{N}{e}\right)^N \int_{\mathbb{R}} e^{-y^2/2N}\, dy \\ &= \left(\frac{N}{e}\right)^N \sqrt{2N} \int_{\mathbb{R}} e^{-z^2}\, dz \\ &= \left(\frac{N}{e}\right)^N \sqrt{2\pi N} \end{aligned}$$

Thus, we have proved the Stirling formula, as formulated in the statement. □

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

PROPOSITION 16.16. *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. □

We can now estimate the volumes of the spheres, as follows:

THEOREM 16.17. *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 16.13, 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 16.13, we can compute in the same way the area of the sphere, the result being as follows:

THEOREM 16.18. *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 16.13, as follows:

(1) We can either use the "pizza" argument from chapter 3, 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 16.13 for $V$, this gives the result.

(2) Or, we can start the computation in the same way as we started the proof of Theorem 16.13, the beginning of this computation being as follows:

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

Now by comparing with the beginning of the proof of Theorem 16.13, the only thing that changes is the following quantity, which now dissapears:

$$\int_0^1 r^{N-1}\, dr = \frac{1}{N}$$

Thus, we have $vol(S^+) = N \cdot vol(B^+)$, and so we obtain the following formula:

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

But this means $A = N \cdot V$, and together with the formula in Theorem 16.13 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 $N$. □

## 16d. Normal variables

We have kept the best for the end. As a starting point, we have:

DEFINITION 16.19. *Let $X$ be a probability space, that is, a space with a probability measure, and with the corresponding integration denoted $E$, and called expectation.*

(1) *The random variables are the real functions $f \in L^\infty(X)$.*
(2) *The moments of such a variable are the numbers $M_k(f) = E(f^k)$.*
(3) *The law of such a variable is the measure given by $M_k(f) = \int_{\mathbb{R}} x^k d\mu_f(x)$.*

Here the fact that a measure $\mu_f$ as above exists indeed is not exactly trivial. But we can do this by looking at formulae of the following type:

$$E(\varphi(f)) = \int_{\mathbb{R}} \varphi(x) d\mu_f(x)$$

Indeed, having this for monomials $\varphi(x) = x^n$, as above, is the same as having it for polynomials $\varphi \in \mathbb{R}[X]$, which in turn is the same as having it for the characteristic functions $\varphi = \chi_I$ of measurable sets $I \subset \mathbb{R}$. Thus, in the end, what we need is:

$$P(f \in I) = \mu_f(I)$$

But this formula can serve as a definition for $\mu_f$, and we are done.

Regarding now independence, we can formulate here the following definition:

DEFINITION 16.20. *Two variables $f, g \in L^\infty(X)$ are called independent when*

$$E(f^k g^l) = E(f^k)\,E(g^l)$$

*happens, for any $k, l \in \mathbb{N}$.*

Again, this definition hides some non-trivial things, the idea being a bit as before, namely that of looking at formulae of the following type:

$$E[\varphi(f)\psi(g)] = E[\varphi(f)]\,E[\psi(g)]$$

To be more precise, passing as before from monomials to polynomials, then to characteristic functions, we are led to the usual definition of independence, namely:

$$P(f \in I, g \in J) = P(f \in I)\,P(g \in J)$$

As a first result now, which is something very standard, we have:

THEOREM 16.21. *Assuming that $f, g \in L^\infty(X)$ are independent, we have*

$$\mu_{f+g} = \mu_f * \mu_g$$

*where $*$ is the convolution of real probability measures.*

PROOF. We have the following computation, using the independence of $f, g$:

$$\int_{\mathbb{R}} x^k d\mu_{f+g}(x) = E((f+g)^k) = \sum_r \binom{k}{r} M_r(f) M_{k-r}(g)$$

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

$$\begin{aligned}\int_{\mathbb{R}} x^k d(\mu_f * \mu_g)(x) &= \int_{\mathbb{R}\times\mathbb{R}} (x+y)^k d\mu_f(x) d\mu_g(y)\\ &= \sum_r \binom{k}{r} M_r(f) M_{k-r}(g)\end{aligned}$$

Thus $\mu_{f+g}$ and $\mu_f * \mu_g$ have the same moments, so they coincide, as claimed. □

As a second result on independence, which is more advanced, we have:

THEOREM 16.22. *Assuming that $f, g \in L^\infty(X)$ are independent, we have*

$$F_{f+g} = F_f F_g$$

*where $F_f(x) = E(e^{ixf})$ is the Fourier transform.*

PROOF. This is something which is very standard too, coming from:

$$\begin{aligned} F_{f+g}(x) &= \int_{\mathbb{R}} e^{ixz} d(\mu_f * \mu_g)(z) \\ &= \int_{\mathbb{R}\times\mathbb{R}} e^{ix(z+t)} d\mu_f(z) d\mu_g(t) \\ &= \int_{\mathbb{R}} e^{ixz} d\mu_f(z) \int_{\mathbb{R}} e^{ixt} d\mu_g(t) \\ &= F_f(x) F_g(x) \end{aligned}$$

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

Let us introduce now the normal laws. This can be done as follows:

DEFINITION 16.23. *The normal law of parameter* 1 *is the following measure:*

$$g_1 = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

*More generally, the normal law of parameter* $t > 0$ *is the following measure:*

$$g_t = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t} dx$$

*These are also called Gaussian distributions, with "g" standing for Gauss.*

Observe that the above laws have indeed mass 1, as they should. This follows indeed from the Gauss formula, which gives, with $x = \sqrt{2t}\, y$:

$$\begin{aligned} \int_{\mathbb{R}} e^{-x^2/2t} dx &= \int_{\mathbb{R}} e^{-y^2} \sqrt{2t}\, dy \\ &= \sqrt{2t} \int_{\mathbb{R}} e^{-y^2} dy \\ &= \sqrt{2t} \times \sqrt{\pi} \\ &= \sqrt{2\pi t} \end{aligned}$$

Generally speaking, the normal laws appear as bit everywhere, in real life. The reasons behind this phenomenon come from the Central Limit Theorem (CLT), that we will explain in a moment, after developing some general theory. As a first result, we have:

PROPOSITION 16.24. *We have the variance formula*

$$V(g_t) = t$$

*valid for any* $t > 0$.

PROOF. The first moment is 0, because our normal law $g_t$ is centered. As for the second moment, this can be computed as follows:

$$\begin{aligned} M_2 &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} x^2 e^{-x^2/2t}dx \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} (tx)\left(-e^{-x^2/2t}\right)' dx \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} te^{-x^2/2t}dx \\ &= t \end{aligned}$$

We conclude from this that the variance is $V = M_2 = t$. □

Here is another result, which is the key one for the study of the normal laws:

THEOREM 16.25. *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 16.26 (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:

$$F_f(x) = E(e^{ixf})$$

In terms of moments, we have the following formula:

$$\begin{aligned} F_f(x) &= E\left(\sum_{k=0}^{\infty}\frac{(ixf)^k}{k!}\right) \\ &= \sum_{k=0}^{\infty}\frac{(ix)^k E(f^k)}{k!} \\ &= \sum_{k=0}^{\infty}\frac{i^k M_k(f)}{k!}\,x^k \end{aligned}$$

Thus, the Fourier transform of the variable in the statement is:

$$\begin{aligned} F(x) &= \left[F_f\left(\frac{x}{\sqrt{n}}\right)\right]^n \\ &= \left[1-\frac{tx^2}{2n}+O(n^{-2})\right]^n \\ &\simeq \left[1-\frac{tx^2}{2n}\right]^n \\ &\simeq e^{-tx^2/2} \end{aligned}$$

But this latter function being the Fourier transform of $g_t$, we obtain the result. □

Let us discuss now some further properties of the normal law. We first have:

PROPOSITION 16.27. *The even moments of the normal law are the numbers*

$$M_k(g_t) = t^{k/2}\times k!!$$

*where* $k!! = (k-1)(k-3)(k-5)\ldots$, *and the odd moments vanish.*

PROOF. We have the following computation, valid for any integer $k\in\mathbb{N}$:

$$\begin{aligned} M_k &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} y^k e^{-y^2/2t}dy \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} (ty^{k-1})\left(-e^{-y^2/2t}\right)'dy \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} t(k-1)y^{k-2}e^{-y^2/2t}dy \\ &= t(k-1)\times\frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} y^{k-2}e^{-y^2/2t}dy \\ &= t(k-1)M_{k-2} \end{aligned}$$

Now recall from the proof of Proposition 16.24 that we have $M_0 = 1$, $M_1 = 0$. Thus by recurrence, we are led to the formula in the statement. □

We have the following alternative formulation of the above result:

PROPOSITION 16.28. *The moments of the normal law are the numbers*

$$M_k(g_t) = t^{k/2}|P_2(k)|$$

*where $P_2(k)$ is the set of pairings of $\{1, \ldots, k\}$.*

PROOF. Let us count the pairings of $\{1, \ldots, k\}$. In order to have such a pairing, we must pair 1 with one of the numbers $2, \ldots, k$, and then use a pairing of the remaining $k-2$ numbers. Thus, we have the following recurrence formula:

$$|P_2(k)| = (k-1)|P_2(k-2)|$$

As for the initial data, this is $P_1 = 0$, $P_2 = 1$. Thus, we are led to the result. □

We are not done yet, and here is one more improvement of the above:

THEOREM 16.29. *The moments of the normal law are the numbers*

$$M_k(g_t) = \sum_{\pi \in P_2(k)} t^{|\pi|}$$

*where $P_2(k)$ is the set of pairings of $\{1, \ldots, k\}$, and $|.|$ is the number of blocks.*

PROOF. This follows indeed from Proposition 16.28, because the number of blocks of a pairing of $\{1, \ldots, k\}$ is trivially $k/2$, independently of the pairing. □

Many other things can be said, as a continuation of the above.

## 16e. Exercises

Congratulations for having read this book, and no exercises for this final chapter.

# Bibliography


[1] A.A. Abrikosov, Fundamentals of the theory of metals, Dover (1988).

[2] V.I. Arnold, Ordinary differential equations, Springer (1973).

[3] V.I. Arnold, Lectures on partial differential equations, Springer (1997).

[4] V.I. Arnold, Catastrophe theory, Springer (1984).

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

[6] T. Banica, Calculus and applications (2024).

[7] T. Banica, Linear algebra and group theory (2024).

[8] T. Banica, Introduction to modern physics (2024).

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

[10] M.J. Benton, Vertebrate paleontology, Wiley (1990).

[11] M.J. Benton and D.A.T. Harper, Introduction to paleobiology and the fossil record, Wiley (2009).

[12] S.J. Blundell and K.M. Blundell, Concepts in thermal physics, Oxford Univ. Press (2006).

[13] B. Bollobás, Modern graph theory, Springer (1998).

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

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

[16] A.R. Choudhuri, Astrophysics for physicists, Cambridge Univ. Press (2012).

[17] J. Clayden, S. Warren and N. Greeves, Organic chemistry, Oxford Univ. Press (2012).

[18] D.D. Clayton, Principles of stellar evolution and nucleosynthesis, Univ. of Chicago Press (1968).

[19] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).

[20] A. Cottrell, An introduction to metallurgy, CRC Press (1997).

[21] C. Darwin, On the origin of species (1859).

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

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

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

[25] S.T. Dougherty, Combinatorics and finite geometry, Springer (2020).

[26] M. Dresher, The mathematics of games of strategy, Dover (1981).

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

[28] F. Dyson, Origins of life, Cambridge Univ. Press (1984).

[29] A. Einstein, Relativity: the special and the general theory, Dover (1916).

[30] L.C. Evans, Partial differential equations, AMS (1998).

[31] W. Feller, An introduction to probability theory and its applications, Wiley (1950).

[32] E. Fermi, Thermodynamics, Dover (1937).

[33] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics I: mainly mechanics, radiation and heat, Caltech (1963).

[34] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics II: mainly electromagnetism and matter, Caltech (1964).

[35] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics III: quantum mechanics, Caltech (1966).

[36] R.P. Feynman and A.R. Hibbs, Quantum mechanics and path integrals, Dover (1965).

[37] P. Flajolet and R. Sedgewick, Analytic combinatorics, Cambridge Univ. Press (2009).

[38] A.P. French, Special relativity, Taylor and Francis (1968).

[39] J.H. Gillespie, Population genetics, Johns Hopkins Univ. Press (1998).

[40] C. Godsil and G. Royle, Algebraic graph theory, Springer (2001).

[41] H. Goldstein, C. Safko and J. Poole, Classical mechanics, Addison-Wesley (1980).

[42] D.L. Goodstein, States of matter, Dover (1975).

[43] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).

[44] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).

[45] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).

[46] D.J. Griffiths, Revolutions in twentieth-century physics, Cambridge Univ. Press (2012).

[47] V.P. Gupta, Principles and applications of quantum chemistry, Elsevier (2016).

[48] W.A. Harrison, Solid state theory, Dover (1970).

[49] W.A. Harrison, Electronic structure and the properties of solids, Dover (1980).

[50] R.A. Horn and C.R. Johnson, Matrix analysis, Cambridge Univ. Press (1985).

[51] C.E. Housecroft and A.G. Sharpe, Inorganic chemistry, Pearson (2018).

[52] K. Huang, Introduction to statistical physics, CRC Press (2001).

[53] K. Huang, Fundamental forces of nature, World Scientific (2007).

[54] S. Huskey, The skeleton revealed, Johns Hopkins Univ. Press (2017).

[55] L. Hyman, Comparative vertebrate anatomy, Univ. of Chicago Press (1942).

[56] L.P. Kadanoff, Statistical physics: statics, dynamics and renormalization, World Scientific (2000).

[57] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).

[58] C. Kittel, Introduction to solid state physics, Wiley (1953).

[59] D.E. Knuth, The art of computer programming, Addison-Wesley (1968).

[60] M. Kumar, Quantum: Einstein, Bohr, and the great debate about the nature of reality, Norton (2009).

[61] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).

[62] L.D. Landau and E.M. Lifshitz, Mechanics, Pergamon Press (1960).

[63] L.D. Landau and E.M. Lifshitz, The classical theory of fields, Addison-Wesley (1951).

[64] L.D. Landau and E.M. Lifshitz, Quantum mechanics: non-relativistic theory, Pergamon Press (1959).

[65] S. Lang, Algebra, Addison-Wesley (1993).

[66] P. Lax, Linear algebra and its applications, Wiley (2007).

[67] P. Lax, Functional analysis, Wiley (2002).

[68] P. Lax and M.S. Terrell, Calculus with applications, Springer (2013).

[69] P. Lax and M.S. Terrell, Multivariable calculus with applications, Springer (2018).

[70] S. Ling and C. Xing, Coding theory: a first course, Cambridge Univ. Press (2004).

[71] J.P. Lowe and K. Peterson, Quantum chemistry, Elsevier (2005).

[72] S.J. Marshall, The story of the computer: a technical and business history, Create Space Publ. (2022).

[73] M.L. Mehta, Random matrices, Elsevier (2004).

[74] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge Univ. Press (2000).

[75] R.K. Pathria and and P.D. Beale, Statistical mechanics, Elsevier (1972).

[76] T.D. Pollard, W.C. Earnshaw, J. Lippincott-Schwartz and G. Johnson, Cell biology, Elsevier (2022).

[77] J. Preskill, Quantum information and computation, Caltech (1998).

[78] R. Rojas and U. Hashagen, The first computers: history and architectures, MIT Press (2000).

[79] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[80] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[81] W. Rudin, Functional analysis, McGraw-Hill (1973).

[82] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[83] B. Ryden and B.M. Peterson, Foundations of astrophysics, Cambridge Univ. Press (2010).

[84] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).

[85] R. Shankar, Fundamentals of physics I: mechanics, relativity, and thermodynamics, Yale Univ. Press (2014).

[86] R. Shankar, Fundamentals of physics II: electromagnetism, optics, and quantum mechanics, Yale Univ. Press (2016).

[87] N.J.A. Sloane and S. Plouffe, Encyclopedia of integer sequences, Academic Press (1995).

[88] A.M. Steane, Thermodynamics, Oxford Univ. Press (2016).

[89] S. Sternberg, Dynamical systems, Dover (2010).

[90] D.R. Stinson, Combinatorial designs: constructions and analysis, Springer (2006).

[91] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[92] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[93] J. von Neumann and O. Morgenstern, Theory of games and economic behavior, Princeton Univ. Press (1944).

[94] J. Watrous, The theory of quantum information, Cambridge Univ. Press (2018).

[95] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[96] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[97] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).

[98] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).

[99] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).

[100] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).

# Index