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, January 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

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

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

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

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

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

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

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

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

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

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

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

Theorem 1.2. *The number of possibilities of choosing k objects among n objects is*

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

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

Proof. Imagine a set consisting of n objects. We have n possibilities for choosing our 1st object, then $n-1$ possibilities for choosing our 2nd object, out of the $n-1$ objects left, and so on up to $n-k+1$ possibilities for choosing our k-th object, out of the $n-k+1$ objects left. Since the possibilities multiply, the total number of choices is:

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

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

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

So, where is the mistake? Thinking a bit, the number N that we computed is in fact the number of possibilities of choosing k ordered objects among n objects. Thus, we must divide everything by the number M of orderings of the k objects that we chose:

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

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

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

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

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

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

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

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*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\vdots$$

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

$$1$$

$$1\ ,\ 1$$

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

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

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

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

$$\vdots$$

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

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

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

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

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

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

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

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

By using the binomial formula, this formula becomes:

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

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

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

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:

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

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 $A_{xy}=A_xA_y$, 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:

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

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 $b^2-4ac\geq0$. 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:

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

(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:

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

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):

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

Let us also pick a number $c > 0$. Since the probability of having $x = r_1$ is certainly smaller than $c/2$, then the probability of having $x = r_2$ is certainly smaller than $c/4$, then the probability of having $x = r_3$ is certainly smaller than $c/8$ and so on, the probability for x to be rational satisfies the following inequality:

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

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

As a comment here, all the above 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. The condition $x_n \to x$ tells us that "when n is big, x_n is close to x", and to be more precise, it tells us that "when n is big enough, x_n gets arbitrarily close to x". But n big enough means $n \geq N$, for some $N \in \mathbb{N}$, and x_n arbitrarily close to x means $|x_n - x| < \varepsilon$, for some $\varepsilon > 0$. Thus, we are led to the above definition.

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:

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

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

There are many other examples, and more on this in a moment. Going ahead with more theory, let us complement Definition 1.17 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. As a basic illustration, 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:

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

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 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 ∞ and ∞. And so on. Of course, and generalizing the $x_n = n$ example, if $x_n \to x$ we obtain x and x.

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

THEOREM 1.26. *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:*

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

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

Proof. Let us denote the completion operation by $X \to \bar{X} = C_X/\sim$, where C_X is the space of Cauchy sequences over X, and $\sim$ is the above equivalence relation. Since by Theorem 1.26 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 $x_k=S_k-S_{k-1}\to 0$, and this gives the result. As for the simplest counterexample for the converse, this is $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots=\infty$, coming from Theorem 1.30.

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

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

As for the simplest counterexample for the converse, this is $1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\ldots<\infty$, coming from Theorem 1.31, coupled with $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots=\infty$ from (1).

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

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

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

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

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

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

THEOREM 1.33. *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. □

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

Functions. Examples, formulae and pictures.

Usually the graph of a function $f:\mathbb{R}\to\mathbb{R}$, which is something in 2D, is the best way to represent the function, but not always. For instance the 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.

The functions $f(x)=x^n$ and $f(x)=n^x$. We will be back to these.

2b. Polynomials, roots

Polynomials and their roots. Many things can be said here.

In order to solve $x^2=-1$, we must trick, in the following way:

Definition 2.1. *The complex numbers are variables of the form*

$$x=a+ib$$

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

$$i^2=-1$$

Each real number 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 2.2. *The complex solutions of $ax^2+bx+c=0$ with $a,b,c\in\mathbb{R}$ are*

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

with the square root of negative real numbers being defined as

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

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

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

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

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

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

Proposition 2.3. *The complex numbers, written as usual*

$$x = a + ib$$

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

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

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

Proof. Consider indeed two arbitrary complex numbers:

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

Their sum is then by definition the following complex number:

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

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

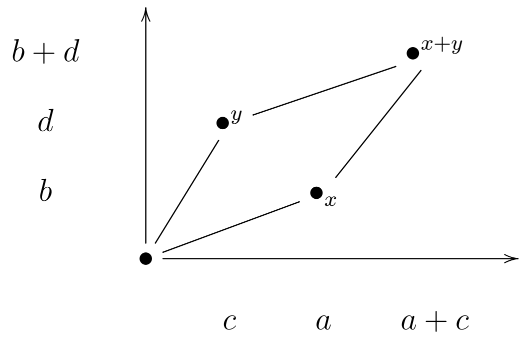
$$x = \begin{pmatrix} a \\ b \end{pmatrix} \quad , \quad y = \begin{pmatrix} c \\ d \end{pmatrix}$$

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

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

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

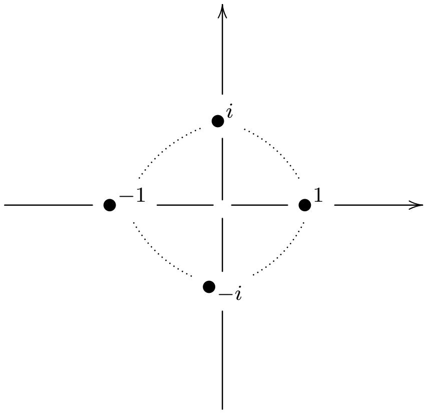
Here we have assumed that you are a bit familiar with vector calculus. If not, no problem, the idea is simply that vectors add by forming a parallelogram, as follows:



Observe that in our geometric picture from Proposition 2.3, the real numbers correspond to the numbers on the Ox axis. As for the purely imaginary numbers, these lie on the Oy axis, with the number i itself being given by the following formula:

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

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



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

Advice 2.4. *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.

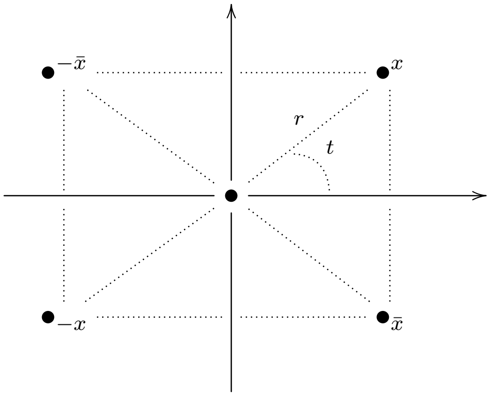
As a last general topic regarding the complex numbers, let us discuss conjugation. This is something quite tricky, complex number specific, as follows:

Definition 2.5. *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 2.6. *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 2.2, but we can deduce this as well directly, without computations. Indeed, by using our assumption that the coefficients are real, $a,b,c\in\mathbb{R}$, we have:

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

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

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, let us start with the following result:

Theorem 2.7. *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 standard, the idea being as follows:

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

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

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

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

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

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

$$S_r=\sum_i a_i^r$$

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

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

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

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

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

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

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

Theorem 2.8. *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,

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

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^ld^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^ld^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:

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

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 2.9. *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 $<k$:

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

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

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

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

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

(4) Finally, our claim is that we have $\det\Phi=R(P,Q)$. But this follows from the uniqueness of the resultant, up to a scalar, and with this uniqueness property being elementary to establish, along the lines of the proofs of Theorems 2.7 and 2.8. □

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 2.10. *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 2.8, 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 2.9:

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

Also, the fact that we have $\Delta(P) = 0$ precisely when P has a double root is clear from Theorem 2.8. 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 2.9, 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 2.9, and we obtain:

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

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

Theorem 2.11. *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 2.8 and Theorem 2.9, 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 2.8. 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 2.9, 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 2.9 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 2.10, 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 2.12. *For a normalized degree* 3 *equation, namely*

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

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

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

is a real solution of our equation.

Proof. The formula of Δ is clear from definitions, and with $108=4\times27$. Now with x as in the statement, by using $(a+b)^3=a^3+b^3+3ab(a+b)$, we have:

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

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

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 2.12, the other two roots are complex conjugate, the formula for them being as follows:

Proposition 2.13. *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 2.14. *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 2.12 and Proposition 2.13. 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 2.15. *The discriminant of* $P=ax^4+bx^3+cx^2+dx+e$ *is given by the following formula:*

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

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

PROOF. The formula of Δ follows from the definition of the discriminant, from Theorem 2.10, with the resultant computed via Theorem 2.9, 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 2.15:

Proposition 2.16. *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 2.15, with $a = 1, b = 0$, but we can deduce this as well directly. Indeed, the formula of Δ 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 2.15. □

We still have some work to do. Indeed, looking back at what we did in degree 3, the passage there from Theorem 2.11 to Theorem 2.12 was made of two operations, namely "depressing" the equation, that is, getting rid of the next-to-highest term, and then rescaling the coefficients, as for the formula of Δ to become as simple as possible.

In our present setting now, degree 4, with the depressing done as above, in Proposition 2.16, it remains to rescale the coefficients, as for the formula of Δ to become as simple as possible. And here, a bit of formula hunting, in relation with $2, 3$ powers, leads to:

Theorem 2.17. *The discriminant of a normalized degree 4 polynomial, written as*

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

is given by the following formula:

$$\Delta = 256 \times 27 \times \left(9p^4r - 2p^3q^2 - 6p^2r^2 + 6pq^2r - q^4 + r^3\right)$$

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

Proof. This follows from Proposition 2.16, with $c = 6p, d = 4q, e = 3r$, but we can deduce this as well directly. Indeed, the formula of Δ follows, quite easily, from:

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

As for the last assertion, this is something that we know from Theorem 2.15. □

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

Theorem 2.18. *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 2.18, is as follows:

THEOREM 2.19 (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 2.20.

Exercise 2.21.

Exercise 2.22.

Exercise 2.23.

Exercise 2.24.

Exercise 2.25.

Exercise 2.26.

Exercise 2.27.

Bonus exercise.

CHAPTER 3

Sin and cos

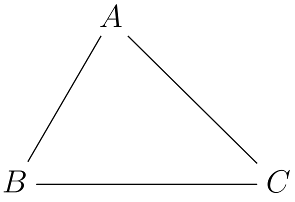
3a. Angles, triangles

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. *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:

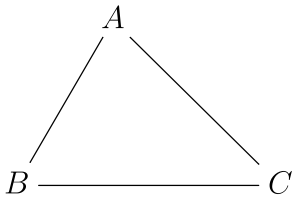
(1) Come with a small circle, inside ABC, and then inflate it, as to touch all 3 edges. The center of the circle will be then at equal distance from all 3 edges, so it will lie on all 3 angle bisectors. Thus, we have constructed the incenter, as required.

(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) Time to draw a new triangle, for clarity, since we are now on page two:



Regarding our problem, we can use 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, as desired. □

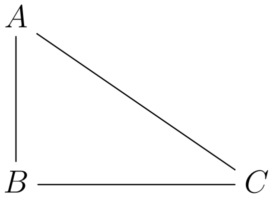
Many other things can be said about triangles, and we will be back to this. Importantly, we can now talk about angles, in the obvious way, by using triangles:

Fact 3.2. *We can talk about the angle between two crossing lines, and have some basic theory for the angles going, by using triangles.*

You might wonder of course what the values of these angles should be, say as real numbers. This is something quite tricky, that will take us some time to understand.

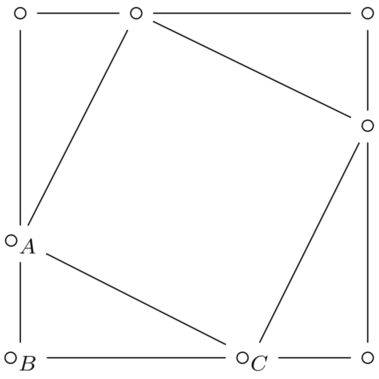
Getting started now with our study of angles, as a continuation of Fact 3.2, let us first talk about the simplest angle of them all, which is the right angle, denoted 90°. Many interesting things can be said about this right angle 90°, in particular with:

Theorem 3.3 (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:

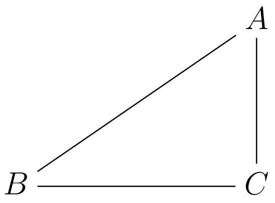
$$\begin{aligned} S &= AC^2+4\times\frac{AB\times BC}{2} \\ &= AC^2+2\times AB\times BC \end{aligned}$$

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

3b. Sine and cosine

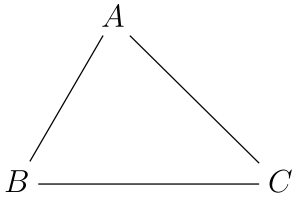
Now that we know about angles, and about Pythagoras' theorem too, it is tempting at this point to start talking about trigonometry. Let us begin with:

DEFINITION 3.4. *We can talk about sines and cosines, by using a right triangle*



in the obvious way, and ideally, by assuming $AB=1$.

Many interesting things can be said here, for instance regarding the sines and cosines of the angles of a triangle, which can be taken arbitrary, or of various special types:



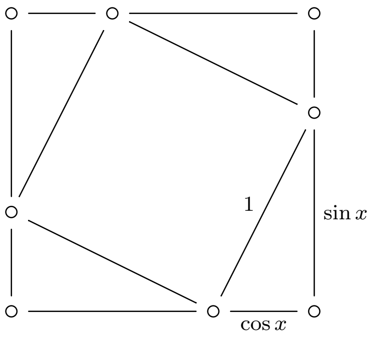
Getting now to more advanced theory, we first have:

THEOREM 3.5. *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 Phythagoras, 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 3.6. *The sines and cosines of sums are given by*

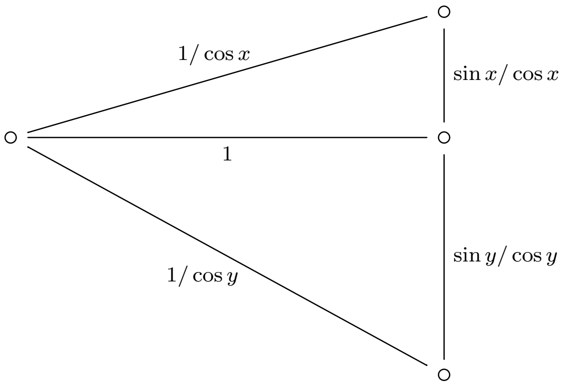
$$\sin(x+y)=\sin x\cos y+\cos x\sin y$$

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

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) Let us first establish the formula for the sines. In order to do so, consider the following picture, consisting of a length 1 line segment, with angles x,y drawn on each side, and with everything being completed, and lengths computed, as indicated:



Now let us compute the area of the big triangle, or rather the double of that area. We can do this in two ways, either directly, with a formula involving $\sin(x+y)$, or by using the two small triangles, involving functions of x,y. We obtain in this way:

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

But this gives the formula for $\sin(x+y)$ from the statement.

(2) Moving ahead, no need of new tricks for cosines, because by using the formula for $\sin(x+y)$ we can deduce a formula for $\cos(x+y)$, as follows:

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

(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 3.6. Observe in particular that with $x=y$ we obtain some interesting formulae for the duplication of angles, namely:

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

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

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.

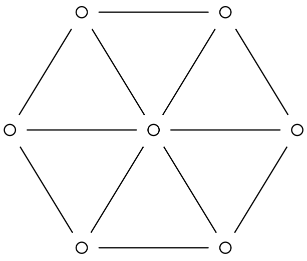
3c. Pi, trigonometry

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 π. And here, to start with, we have:

Theorem 3.7. *The following two definitions of π are equivalent:*

(1) *The length of the unit circle is $L=2\pi$.*
(2) *The area of the unit disk is $A=\pi$.*

Proof. In order to prove this theorem let us cut the unit disk as a pizza, into N slices, and forgetting about gastronomy, leave aside the rounded parts:



The area to be eaten can be then computed as follows, where H is the height of the slices, S is the length of their sides, and $P=NS$ is the total length of the sides:

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

Thus, with $N\to\infty$ we obtain that we have $A=L/2$, as desired. □

In what regards now the precise value of π, the above picture at $N=6$ shows that we have $\pi>3$, but not by much. The precise figure is as follows:

$$\pi=3.14159\ldots$$

Getting back now to trigonometry, the basics here are as follows:

THEOREM 3.8. *The following happen:*

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

PROOF. The formula $L=2\pi$ from Theorem 3.7 shows that the length of a quarter of the unit circle is $l=\pi/2$, and so the right angle has indeed this value, $\pi/2$. As for $\sin^2x+\cos^2x=1$, this is something that we know well, coming from Pythagoras. □

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 3.9. *The following happen, for small angles, $x\simeq0$:*

(1) $\sin x\simeq x$.
(2) $\cos x\simeq1-x^2/2$.
(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 3.10. *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,

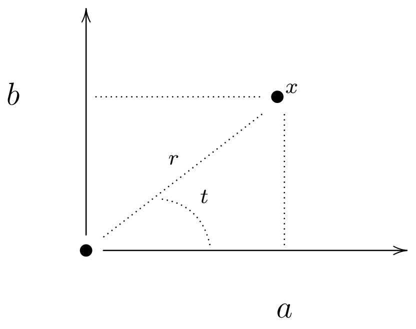
$$a=r\cos t\quad,\quad b=r\sin t$$

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 3.10 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 3.11. *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 3.12. *The powers of a complex number, written in polar form,*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

As a basic application of Theorem 3.12, we have the following result:

PROPOSITION 3.13. *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 3.12, at $k = 1/2$. As for its particular cases with $x \in \mathbb{R}$, these are clear from it. □

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

With the above results in hand, and notably with the square root formula from Proposition 3.13, we can now go back to the degree 2 equations, and we have:

THEOREM 3.14. *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 3.13 that any complex number has a square root, we are led to the conclusion in the statement. □

We will be back to this, roots of polynomials, later in this book.

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

THEOREM 3.15. *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 3.11 takes the following very simple form, with the arguments s, t being both taken modulo 2π:

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

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

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

which are called roots of unity of order N.

PROOF. This follows from the general multiplication formula for complex numbers from 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 3.15, the statement here being as follows:

Theorem 3.17. *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 3.18.

Exercise 3.19.

Exercise 3.20.

Exercise 3.21.

Exercise 3.22.

Exercise 3.23.

Exercise 3.24.

Exercise 3.25.

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.

Regarding e, we have the following 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:

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:

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

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

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

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

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

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

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

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

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

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

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

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

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

(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. □

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 I would like to have as well a combinatorial interpretation of it. And, here is a very nice result, of this type:

THEOREM 4.6. *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:

$$S_N^i=\left\{\sigma\in S_N\Big|\sigma(i)=i\right\}$$

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. □

More generally now, we have the following result:

Theorem 4.7. *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 4.6, 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 4.6, 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. □

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

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

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

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

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

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

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

Next, we have the following deep result, regarding the complex exponential:

Theorem 4.9. *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 4.8 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 4.10. *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 4.9, 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 4.11. *In polar coordinates, the complex numbers multiply as*

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

with the arguments s,t being taken modulo 2π.

Proof. This is something that we already know, from chapter 3, reformulated by using the notations from Theorem 4.10. Observe that this follows as well directly, from the fact that we have $e^{a+b}=e^ae^b$, that we know from analysis. □

As for the continuation of the story, involving square roots, and many other operations on the complex numbers, using the $x=re^{it}$ notation, we refer here to chapter 3.

As an interesting consequence of the above results, which is of great practical interest, we have the following useful method, for remembering the basic math formulae:

Method 4.12. *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:

$$\begin{aligned}e^{i(x+y)}=e^{ix}e^{iy}\quad&\implies\quad\cos(x+y)+i\sin(x+y)=(\cos x+i\sin x)(\cos y+i\sin y)\\&\implies\quad\begin{cases}\cos(x+y)=\cos x\cos y-\sin x\sin y\\\sin(x+y)=\sin x\cos y+\cos x\sin y\end{cases}\end{aligned}$$

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:

$$\begin{aligned}e^{ix}=\sum_k\frac{(ix)^k}{k!}\quad&\implies\quad\cos x+i\sin x=\sum_k\frac{(ix)^k}{k!}\\&\implies\quad\begin{cases}\cos x=\sum_{l=0}^\infty(-1)^l\frac{x^{2l}}{(2l)!}\\\sin x=\sum_{l=0}^\infty(-1)^l\frac{x^{2l+1}}{(2l+1)!}\end{cases}\end{aligned}$$

Finally, in what regards log, there is a trick here too, which is partial, namely:

$$\begin{aligned}\log(\exp x) = x \quad &\implies \quad \log\left(1 + x + \frac{x^2}{2} + \ldots\right) = x\\ &\implies \quad \log(1+y) = y - \frac{y^2}{2} + \ldots\end{aligned}$$

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

We have the following result, which is something of general interest:

Theorem 4.13. *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:

$$\frac{e^{x+y} - e^{-x-y}}{2} = \frac{e^x - e^{-x}}{2} \cdot \frac{e^y + e^{-y}}{2} + \frac{e^x + e^{-x}}{2} \cdot \frac{e^y - e^{-y}}{2}$$

$$\frac{e^{x+y} + e^{-x-y}}{2} = \frac{e^x + e^{-x}}{2} \cdot \frac{e^y + e^{-y}}{2} + \frac{e^x - e^{-x}}{2} \cdot \frac{e^y - e^{-y}}{2}$$

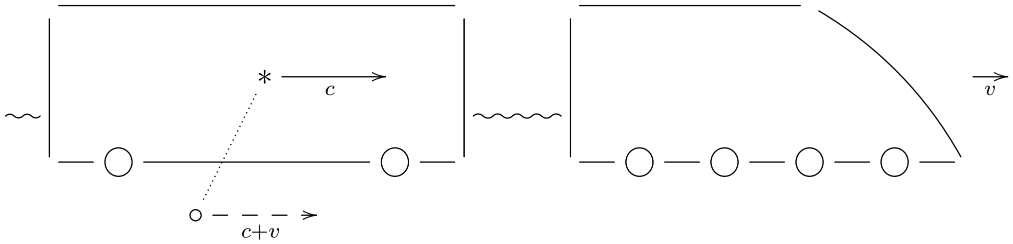
Finally, (5) is clear from the definition of sinh, cosh, and from $e^x = \sum_k \frac{x^k}{k!}$. □

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:

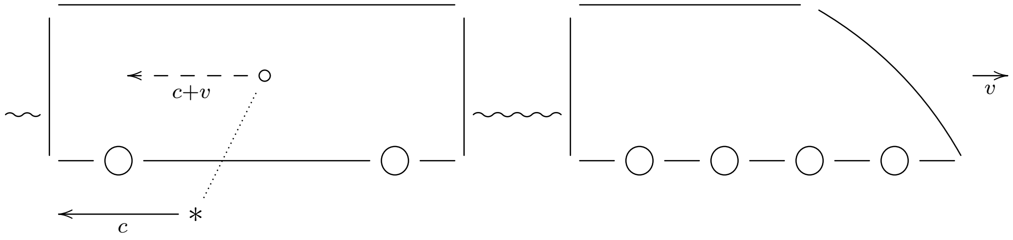
Fact 4.14 (Einstein principles). *The following happen:*

(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 4.14 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 4.15. *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 4.16. *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 $\pm 1\to i$, with negatives on the left, and positives on the right.

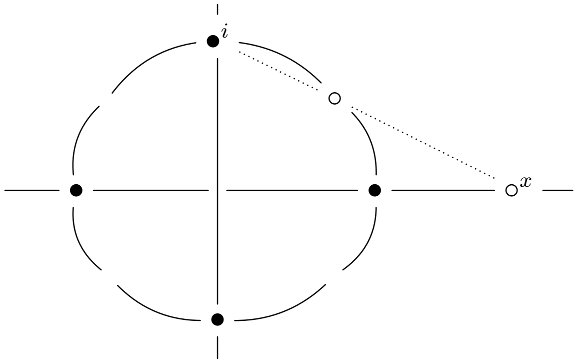
PROOF. Regarding the wrapping, as indicated, this is given by:

$$u\to e^{it}\quad,\quad t=\pi u-\frac{\pi}{2}$$

Indeed, this correspondence wraps $[-1,1]$ as above, the basic instances of our correspondence being as follows, and with everything being fine modulo 2π:

$$-1\to\frac{\pi}{2}\quad,\quad -\frac{1}{2}\to-\pi\quad,\quad 0\to-\frac{\pi}{2}\quad,\quad \frac{1}{2}\to 0\quad,\quad 1\to\frac{\pi}{2}$$

Regarding now the stereographic projection, the picture here is as follows:



Thus, by Thales, the formula of the stereographic projection is as follows:

$$\frac{\cos t}{x}=\frac{1-\sin t}{1}\implies x=\frac{\cos t}{1-\sin t}$$

Now if we compose our wrapping operation above with the stereographic projection, what we get is, via the above Thales formula, and some trigonometry:

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

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

The above result is very nice, but when it comes to physics, things do not work, for instance because of the wrong slope of the function $\varphi(u)=\tan\left(\frac{\pi u}{2}\right)$ at the origin, which makes our summing on $[-1,1]$ not compatible with the Galileo addition, at low speeds.

So, what to do? Obviously, trash Proposition 4.16, and start all over again. Getting back now to Puzzle 4.15, 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 4.17. *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 4.15, 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 4.17. So, we have our idea, and we are led to the following result:

Theorem 4.18. *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 4.19. *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 4.20.

Exercise 4.21.

Exercise 4.22.

Exercise 4.23.

Exercise 4.24.

Exercise 4.25.

Exercise 4.26.

Exercise 4.27.

Bonus exercise.

Part II

Continuity

Hold on tight, you know she's a little bit dangerous
She's got what it takes to make ends meet
The eyes of a lover that hit like heat
You know she's a little bit dangerous

CHAPTER 5

Continuous functions

5a. Continuity

We focus now our study on the functions $f:\mathbb R\to\mathbb R$ which are suitably regular. And, in what regards these regularity properties, the most basic of them 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 see examples in a moment. Still speaking theory, there are many equivalent formulations of the notion of continuity, with a well-known one, coming by reminding in the above definition what convergence of a sequence means, twice, for both the convergences $x_n\to x$ and $f(x_n)\to f(x)$, being as follows:

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

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

Theorem 5.2. *If f,g are continuous, then so are:*

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

Proof. Before anything, we should mention that the claim is that (1-4) hold indeed, provided that at the level of domains and ranges, the statement makes sense. For instance in (1,2,3) we are talking about functions having the same domain, and with $g(x)\neq0$ 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:

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

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

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

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

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

Alternatively, let us prove this as well by using that scary ε,δ condition given after Definition 5.1. So, let us pick $\varepsilon>0$. We want in the end to have something of type $|f(g(x))-f(g(y))|<\varepsilon$, so we must first use that ε,δ condition for the function f. So, let us start in this way. Since f is continuous at $g(x)$, we can find $\delta>0$ such that:

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

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

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

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

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

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

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

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

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

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

5b. Basic examples

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

Theorem 5.3. *The following functions are continuous:*

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

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

(1) Since $f(x)=x$ is continuous, by using Theorem 5.2 we obtain the result for exponents $n\in\mathbb N$, and then for general exponents $n\in\mathbb Z$ too.

(2) The statement here, which generalizes (1), follows exactly as (1), by using the various findings from Theorem 5.2.

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

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

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:

$$\begin{aligned}\lim_{y\to0}\sin(x+y)&=\lim_{y\to0}(\sin x\cos y+\cos x\sin y)\\&=\sin x\lim_{y\to0}\cos y+\cos x\lim_{y\to0}\sin y\\&=\sin x\cdot1+\cos x\cdot0\\&=\sin x\end{aligned}$$

(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) Finally, the fact that $\tan x$, $\cot x$ are continuous is clear from the fact that $\sin x$, $\cos x$ are continuous, by using the result regarding quotients from Theorem 5.2. □

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

5c. Further examples

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

PROPOSITION 5.4. *If a function $f:X\to\mathbb{R}$ has the Lipschitz property*

$$|f(x)-f(y)|\leq K|x-y|$$

for some $K>0$, then it is continuous.

PROOF. This is indeed clear from our definition of continuity. □

5d. Uniform continuity

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

DEFINITION 5.5. *A function $f:X\to\mathbb{R}$ is called uniformly continuous when:*

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

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

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

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

Indeed, it is clear by looking at the graph of f that, the further our point $x\in\mathbb{R}$ is from 0, the smaller our $\delta>0$ must be, compared to $\varepsilon>0$, in our ε,δ definition of continuity. Thus, given 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.

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

Theorem 5.6. *Any continuous function defined on a closed, bounded interval*

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

is automatically uniformly continuous.

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

(1) Given $\varepsilon>0$, for any $x\in[a,b]$ we know that we have a $\delta_x>0$ such that:

$$|x-y|<\delta_x\implies|f(x)-f(y)|<\frac{\varepsilon}{2}$$

So, consider the following open intervals, centered at the various points $x\in[a,b]$:

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

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

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

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

(2) So, let us prove first that the cover in (1) has a finite subcover. For this purpose, we proceed by contradiction. So, assume that $[a,b]$ has no finite subcover, and let us cut this interval in half. Then one of the halves must have no finite subcover either, and we can repeat the procedure, by cutting this smaller interval in half. And so on. But this leads to a contradiction, because the limiting point $x\in[a,b]$ that we obtain in this way, as the intersection of these smaller and smaller intervals, must be covered by something, and so one of these small intervals leading to it must be covered too, contradiction.

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

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

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

5e. Exercises

Exercises:

Exercise 5.7.

Exercise 5.8.

Exercise 5.9.

Exercise 5.10.

Exercise 5.11.

Exercise 5.12.

Exercise 5.13.

Exercise 5.14.

Bonus exercise.

CHAPTER 6

Intermediate values

6a. Open sets

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

Let us start with 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 examples here, the open intervals (a,b) are open, and the closed intervals $[a,b]$ are closed. Observe also that $\mathbb{R}$ itself is open and closed at the same time. Further examples, or rather results which are easy to establish, include the fact that the finite unions or intersections of open or closed sets are open or closed.

We will be back to all this later, with some precise results in this sense. For the moment, in order to get started, we will only need the following theoretical result:

Proposition 6.2. *A set $O\subset\mathbb{R}$ is open precisely when its complement $C\subset\mathbb{R}$ is closed, and vice versa.*

Proof. It is enough to prove the first assertion, since the "vice versa" part will follow from it, by taking complements. But this can be done as follows:

"$\implies$" Assume that $O\subset\mathbb{R}$ is open, and let $C=\mathbb{R}-O$. In order to prove that C is closed, assume that $\{x_n\}_{n\in\mathbb{N}}\subset C$ converges to $x\in\mathbb{R}$. We must prove that $x\in C$, and we will do this by contradiction. So, assume $x\notin C$. Thus $x\in O$, and since O is open we can find a small interval $(x-\varepsilon,x+\varepsilon)\subset O$. But since $x_n\to x$ this shows that $x_n\in O$ for n big enough, which contradicts $x_n\in C$ for all n, and we are done.

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

As basic illustrations for the above result, $\mathbb{R} - (a,b) = (-\infty, a] \cup [b, \infty)$ is closed, and $\mathbb{R} - [a,b] = (-\infty, a) \cup (b, \infty)$ is open. There are many other such illustrations.

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

Theorem 6.3. *A function is continuous precisely when $f^{-1}(O)$ is open, for any O open. Equivalently, $f^{-1}(C)$ must be closed, for any C closed.*

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

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

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

As a test for the above criterion, let us reprove the fact, that we know from 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 ε, δ.

In order to reach now to true applications of Theorem 6.3, we will need to know more about the open and closed sets. Let us begin with a useful result, as follows:

Proposition 6.4. *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 6.5. *The open and closed sets are as follows:*

1. *The open sets are the disjoint unions of open intervals.*
2. *The closed sets are the complements of these unions.*

Proof. We have two assertions to be proved, the idea being as follows:

(1) We know that the open intervals are those of type (a,b) with $a<b$, with the values $a,b=\pm\infty$ allowed, and by Proposition 6.4 a union of such intervals is open.

(2) Conversely, given $O\subset\mathbb{R}$ open, we can cover each point $x\in O$ with an open interval $I_x\subset O$, and we have $O=\cup_x I_x$, so O is a union of open intervals.

(3) In order to finish the proof of the first assertion, it remains to prove that the union $O=\cup_x I_x$ in (2) can be taken to be disjoint. For this purpose, our first observation is that, by approximating points $x\in O$ by rationals $y\in\mathbb{Q}\cap O$, we can make our union to be countable. But once our union is countable, we can start merging intervals, whenever they meet, and we are left in the end with a countable, disjoint union, as desired.

(4) Finally, the second assertion comes from Proposition 6.2. □

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 $O\subset\mathbb{R}$.

6b. Compact sets

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

Definition 6.6. *The compact and connected sets are defined as follows:*

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

As basic examples, the closed bounded intervals $[a,b]$ are compact, as we know from chapter 5, and so are the finite unions of such intervals. As for connected sets, the basic examples here are the various types of intervals, namely (a,b), $(a,b]$, $[a,b)$, $[a,b]$, and it looks impossible to come up with more examples. In fact, we have:

Theorem 6.7. *The compact and connected sets are as follows:*

(1) *The compact sets are those which are closed and bounded.*
(2) *The connected sets are the various types of intervals.*

Proof. This is something quite intuitive, the idea being as follows:

(1) The fact that compact implies both closed and bounded is clear from our definition of compactness, because assuming non-closedness or non-boundedness leads to an open cover having no finite subcover. As for the converse, we know from 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.

(2) This is something which is obvious, and this regardless of what "cannot be broken into parts" in Definition 6.6 exactly means, mathematically speaking, with several possible definitions being possible here, all being equivalent. Indeed, $E\subset\mathbb{R}$ having this property is equivalent to $a,b\in E\implies[a,b]\subset E$, and this gives the result. □

We will be back to all this later, when looking at open, closed, compact and connected sets in $\mathbb{R}^N$, or more general spaces, where things are more complicated than in $\mathbb{R}$.

Now with this discussed, let us go back to the continuous functions. We have:

Theorem 6.8. *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 6.6. 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 6.9. *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. □

6c. Intermediate values

Very nice all this, but you might perhaps ask at this point, were Theorems 6.8 and 6.9 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 6.10. *The following happen for a continuous function $f : [a,b] \to \mathbb{R}$:*

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

Proof. All these statements are related, and are called altogether "intermediate value theorem". Regarding now the proof, one way of viewing things is that since $[a,b]$ is compact and connected, the set $f([a,b])$ is compact and connected too, and so it is a certain closed bounded interval $[c,d]$, and this gives all the results. However, this is based on rather advanced technology, as developed above, and it is possible to prove (1-3) directly as well, and we will leave this as an instructive exercise for you, reader. □

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

Theorem 6.11. *Assuming that a function f is continuous and invertible, this function must be monotone, and its inverse function f^{-1} must be monotone and continuous too. Moreover, this statement holds both locally, and globally.*

Proof. The fact that both f and f^{-1} are monotone follows from Theorem 6.10. Regarding now the continuity of f^{-1}, we want to prove that we have:

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

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

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

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

As a basic application of Theorem 6.11, we have:

Proposition 6.12. *The various usual inverse functions, such as the inverse trigonometric functions* arcsin, arccos, arctan, arccot*, are all continuous.*

Proof. This follows indeed from Theorem 6.11, with a course the full discussion needing some explanations on bijectivity and domains. But you surely know all that, and in what concerns us, our claim is simply that these beasts are all continuous, proved. □

As another basic application of this, we have:

Proposition 6.13. *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 6.10, as follows:

(1) This is clear from Theorem 6.10 (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 6.10 (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 6.10. We will be back to this.

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

Theorem 6.14. *The function x^a is defined and continuous on $(0,\infty)$, for any $a\in\mathbb R$. Moreover, when trying to extend it to $\mathbb R$, we have 4 cases, as follows,*

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

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

Proof. The idea is that we know how to extract roots by using Proposition 6.13, and all the rest follows by continuity. To be more precise:

(1) Assume $a=p/q$, with $p,q\in\mathbb N$, $p\neq 0$ and q odd. Given a number $x\in\mathbb R$, we can construct the power x^a in the following way, by using Proposition 6.13:

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

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

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

(3) Assume first $a\in\mathbb Q_{even}$, $a>0$. This means $a=p/q$ with $p,q\in\mathbb N$, $p\neq 0$ and q even, and as before in (1), we can set $x^a=\sqrt[q]{x^p}$ for $x\geq 0$, by using Proposition 6.13. 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 6.15. *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 6.14, by exchanging the variables, $x\leftrightarrow a$. To be more precise, the situation is as follows:

(1) We know from Theorem 6.14 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 6.14 that we cannot define a^x for $x\in\mathbb Q_{even}$. But since $\mathbb Q_{even}$ is dense in $\mathbb R$, this gives the result. □

Summarizing, we have been quite successful with our theory of continuous functions, having how full results, regarding the definition and continuity property, for all basic functions from mathematics. All this is of course just a beginning, and we will be back to these functions on regular occasions, in what follows.

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 we have the following result:

Theorem 6.16. *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 6.17.

Exercise 6.18.

Exercise 6.19.

Exercise 6.20.

Exercise 6.21.

Exercise 6.22.

Exercise 6.23.

Exercise 6.24.

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 $\lim_{x\to1}x^N=1$, 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

Finally, 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, and in relation with this, we have the following result:

THEOREM 7.7. *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. Functional analysis

Functional analysis.

7e. Exercises

Exercises:

EXERCISE 7.8.

EXERCISE 7.9.

EXERCISE 7.10.

EXERCISE 7.11.

EXERCISE 7.12.

EXERCISE 7.13.

EXERCISE 7.14.

EXERCISE 7.15.

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:

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

(3) Summarizing, we have proved so far that the binomial formula at $a=\pm 1/2$ is equivalent to the explicit formulae in the statement, involving the Catalan numbers C_k, and the central binomial coefficients D_k. It remains now to prove that these two explicit formulae hold indeed. For this purpose, let us write these formulae as follows:

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

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

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

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

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

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

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

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

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

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

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

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

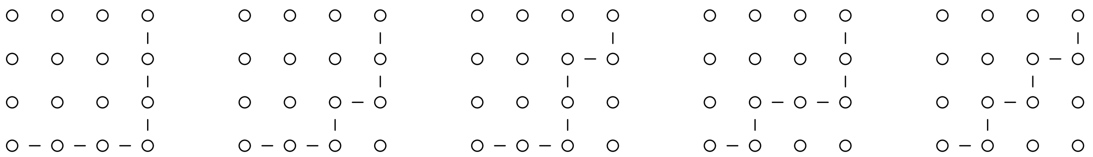
$$\begin{aligned}\sum_{k+l=n+1}D_kD_l&=D_0D_{n+1}+\sum_{k+l=n}\left(4-\frac{2}{k+1}\right)D_kD_l\\&=D_{n+1}+4\sum_{k+l=n}D_kD_l-2\sum_{k+l=n}\frac{D_kD_l}{k+1}\\&=D_{n+1}+4^{n+1}-2\sum_{k+l=n}C_kD_l\end{aligned}$$

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

(6) All this suggests a systematic look at the three formulae in (4). According to our various observations above, these three formulae are equivalent, and so it is enough to prove one of them. We will chose here to prove the first one, namely:

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

(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:

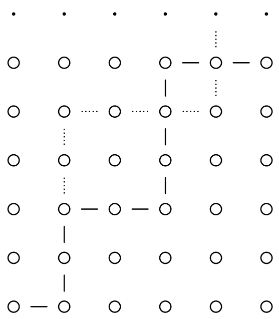
$$1,1,2,5,14,42,132,429,\ldots$$

(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 $f(z)=\sum_{k\geq0}C_kz^k$ 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 $f(z)=\sum_{k\geq0}C_kz^k$ of the Catalan numbers. In terms of this series, the recurrence relation gives, as desired:

$$\begin{aligned}zf^2&=\sum_{a,b\geq0}C_aC_bz^{a+b+1}\\&=\sum_{k\geq1}\sum_{a+b=k-1}C_aC_bz^k\\&=\sum_{k\geq1}C_kz^k\\&=f-1\end{aligned}$$

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. □

8d. Special functions

Special functions.

8e. Exercises

Exercises:

Exercise 8.5.

Exercise 8.6.

Exercise 8.7.

Exercise 8.8.

Exercise 8.9.

Exercise 8.10.

Exercise 8.11.

Exercise 8.12.

Bonus exercise.

Part III

Derivatives

Up ahead in the distance
I saw a shimmering light
My head grew heavy and my sight grew dim
I had to stop for the night

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:

$$\lim_{t\to 0}\left[f(x+t)-f(x)\right]=0$$

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:

$$\begin{aligned}(\sin x)' &= \lim_{t\to0}\frac{\sin(x+t)-\sin x}{t}\\ &= \lim_{t\to0}\frac{\sin x\cos t+\cos x\sin t-\sin x}{t}\\ &= \lim_{t\to0}\sin x\cdot\frac{\cos t-1}{t}+\cos x\cdot\frac{\sin t}{t}\\ &= \cos x\end{aligned}$$

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:

$$\begin{aligned}(\cos x)' &= \lim_{t\to0}\frac{\cos(x+t)-\cos x}{t}\\ &= \lim_{t\to0}\frac{\cos x\cos t-\sin x\sin t-\cos x}{t}\\ &= \lim_{t\to0}\cos x\cdot\frac{\cos t-1}{t}-\sin x\cdot\frac{\sin t}{t}\\ &= -\sin x\end{aligned}$$

(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:

$$\begin{aligned}(\log x)' &= \lim_{t\to0}\frac{\log(x+t)-\log x}{t}\\ &= \lim_{t\to0}\frac{\log(1+t/x)}{t}\\ &= \frac{1}{x}\end{aligned}$$

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:

$$f(x)=1+\sum_{k=1}^\infty c_kx^k$$

According to our differentiation rules, the derivative of this series is given by:

$$f(x)=\sum_{k=1}^\infty kc_kx^{k-1}$$

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:

$$\begin{aligned}(f+g)'(x) &= \lim_{t\to0}\frac{(f+g)(x+t)-(f+g)(x)}{t}\\ &= \lim_{t\to0}\left(\frac{f(x+t)-f(x)}{t}+\frac{g(x+t)-g(x)}{t}\right)\\ &= \lim_{t\to0}\frac{f(x+t)-f(x)}{t}+\lim_{t\to0}\frac{g(x+t)-g(x)}{t}\\ &= f'(x)+g'(x)\end{aligned}$$

(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, (1) follows from the usual interpretation of the derivative, as both a growth rate, and a slope. Regarding (2), this is some sort of reformulation of (1), using the intuitive meaning of the word “acceleration”, with the relevant physics equations, due to Newton, being as follows:

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

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 at the end of the present chapter.

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

All in all, what we have above is a mixture of trivial and non-trivial facts, and do not worry, we will get familiar with all this, in the next few pages.

10b. Basic examples

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

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

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

Proof. We have several assertions here, the idea being as follows:

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

$$(x^p)'' = (px^{p-1})' = p(p-1)x^{p-2}$$
$$(\sin x)'' = (\cos x)' = -\sin x$$
$$(\cos x)'' = (-\sin x)' = -\cos x$$
$$(e^x)'' = (e^x)' = e^x$$
$$(\log x)'' = (-1/x)' = -1/x^2$$

(2) Regarding now the counterexample, recall first that the simplest example of a function which is continuous, but not differentiable, was $f(x) = |x|$, the idea behind this being to use a "piecewise linear function whose branches do not fit well". In connection now with our question, piecewise linear will not do, but we can use a similar idea, namely "piecewise quadratic function whose branches do not fit well". So, let us set:

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

This function is then differentiable, with its derivative being:

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

Now for getting our counterexample, we can set $a=-1,b=1$, so that f is:

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

Indeed, the derivative is $f'(x)=2|x|$, which is not differentiable, as desired. □

10c. Taylor formula

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

THEOREM 10.4. *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 formula in the statement, namely:

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

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

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

Observe that this formula holds indeed, as an application of the theory in chapter 9. Now by using this, if we denote by $\varphi(t) \simeq P(t)$ the formula to be proved, we have:

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

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

As a first application, justifying Interpretation 10.2 (3), we have the following statement, which is a bit heuristic, but we will call it however Proposition:

Proposition 10.5. *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 10.4, as such, and with $t \to -t$ too:

$$\begin{aligned}f(x+t) &\simeq f(x) + f'(x)t + \frac{f''(x)}{2}\,t^2\\ f(x-t) &\simeq f(x) - f'(x)t + \frac{f''(x)}{2}\,t^2\end{aligned}$$

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 to rigorous mathematics, we have as well the following result:

Theorem 10.6. *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 10.4, which in the case $f'(x) = 0$ reads:

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

Indeed, assuming $f''(x) \neq 0$, it is clear that the condition $f''(x) > 0$ will produce a local minimum, and that the condition $f''(x) < 0$ will produce a local maximum. □

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 $\pm x^{2m+1}$.

There are some comments to be made 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 10.6 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 10.6 is good to know, when applying that algorithm.

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 10.7. *Given a convex function $f : \mathbb{R} \to \mathbb{R}$, we have the following Jensen inequality, for any $x_1, \ldots, x_N \in \mathbb{R}$, and any $\lambda_1, \ldots, \lambda_N > 0$ summing up to 1,*

$$f(\lambda_1 x_1 + \ldots + \lambda_N x_N) \leq \lambda_1 f(x_1) + \ldots + \lambda_N x_N$$

with equality when $x_1 = \ldots = x_N$. In particular, by taking the weights λ_i to be all equal, we obtain the following Jensen inequality, valid for any $x_1, \ldots, x_N \in \mathbb{R}$,

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

and once again with equality when $x_1 = \ldots = x_N$. A similar statement holds for the concave functions, with all the inequalities being reversed.

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

(1) First, we can talk about convex functions in a usual, intuitive way, with this meaning by definition that the following inequality must be satisfied:

$$f\left(\frac{x+y}{2}\right)\leq\frac{f(x)+f(y)}{2}$$

(2) But this means, via a simple argument, by approximating numbers $t\in[0,1]$ by sums of powers 2^{-k}, that for any $t\in[0,1]$ we must have:

$$f(tx+(1-t)y)\leq tf(x)+(1-t)f(y)$$

Alternatively, via yet another simple argument, this time by doing some geometry with triangles, this means that we must have:

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

But then, again alternatively, by combining the above two simple arguments, the following must happen, for any $\lambda_1,\ldots,\lambda_N>0$ summing up to 1:

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

(3) Summarizing, all our Jensen inequalities, at $N=2$ and at $N\in\mathbb{N}$ arbitrary, are equivalent. The point now is that, if we look at what the first Jensen inequality, that we took as definition for the convexity, 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 10.8. *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 10.7, 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.

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

Theorem 10.9. *We have the Young inequality,*

$$ab\leq\frac{a^p}{p}+\frac{b^q}{q}$$

valid for any $a,b\geq 0$, and any exponents $p,q>1$ satisfying $\frac{1}{p}+\frac{1}{q}=1$.

Proof. We use the logarithm function, which is concave on $(0,\infty)$, due to:

$$(\log x)''=\left(-\frac{1}{x}\right)'=-\frac{1}{x^2}$$

Thus we can apply the Jensen inequality, and we obtain in this way:

$$\begin{aligned}\log\left(\frac{a^p}{p}+\frac{b^q}{q}\right)&\geq\frac{\log(a^p)}{p}+\frac{\log(b^q)}{q}\\&=\log(a)+\log(b)\\&=\log(ab)\end{aligned}$$

Now by exponentiating, we obtain the Young inequality. □

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

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

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

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 10.10.

Exercise 10.11.

Exercise 10.12.

Exercise 10.13.

Exercise 10.14.

Exercise 10.15.

Exercise 10.16.

Exercise 10.17.

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, the ultimate result on the subject, called Taylor formula, being as follows:

THEOREM 11.1. *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:

$$P(t)=a_0+a_1t+\ldots+a_nt^n$$

The natural conditions to be imposed are those stating that P and φ should match at $t=0$, at the level of the actual value, of the derivative, second derivative, and so on up the n-th derivative. Thus, we are led to the approximation in the statement:

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

In order to prove now that this approximation holds indeed, we 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 11.2. *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. □

11c. Local extrema

In relation with the local extrema, we have the following result:

Theorem 11.3. *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 11.3 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 11.4. *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 11.3, 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. □

11d. Basic applications

As a 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 11.5. *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:

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

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

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

But this is exactly our generalized binomial formula, so we are done with the case where t is small. With a bit more care, we obtain that this holds for any $|t| < 1$, 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 11.6.

Exercise 11.7.

Exercise 11.8.

Exercise 11.9.

Exercise 11.10.

Exercise 11.11.

Exercise 11.12.

Exercise 11.13.

Bonus exercise.

CHAPTER 12

Differential equations

12a. Differential equations

Differential equations.

12b. 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 $a = v'(t) = x''(t)$.

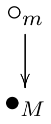
Summarizing, with Newton's theory of derivatives, as we learned it in the previous chapters, we can certainly do some mathematics for the motion of bodies. But, for these bodies to move, we need them to be acted upon by some forces, right? The simplest such force is gravity, and in our present, modest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

$$\bullet_0 \longleftarrow \circ_x$$

Now comes the physics. The gravitational force extented by M, which is fixed in our formalism, on the object m which moves, is subject to the following equations:

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

To be more precise, in the first equation $G \simeq 6.674 \times 10^{-11}$ is the gravitational constant, in usual SI units, and the sign is $-$ because F is attractive. The second equation is something standard and very intuitive, and the last two equations are those from Fact 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. □

12c. 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:

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

Now let us send an impulse, and see how balls will be moving. For this purpose, we zoom on one ball. The situation here is as follows, l being the spring length:

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

We have two forces acting at x. First is the Newton motion force, mass times acceleration, which is as follows, with m being the mass of each ball:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

12d. 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 φ'' computes the average value of a function φ around a point, minus the value of φ 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:

$$-\!\!- \circ_{x-l} \overset{l}{-\!\!-} \circ_x \overset{l}{-\!\!-} \circ_{x+l} -\!\!-$$

In order to model now heat diffusion, we have to implement the intuitive mechanism explained above, and in practice, this leads to a condition as follows, expressing the change of the temperature φ, over a small period of time $\delta > 0$:

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

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

12e. Exercises

Exercises:

Exercise 12.5.

Exercise 12.6.

Exercise 12.7.

Exercise 12.8.

Exercise 12.9.

Exercise 12.10.

Exercise 12.11.

Exercise 12.12.

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 3×10^{-3} that volume, as to have the area,*

and for general f, we can use this plus $f=f_+-f_-$, with $f_+,f_-\geq 0$.

So far, so good, we have a rigorous definition, so let us do now some computations. In order to compute areas, and so integrals of functions, without wasting precious water, we can use our geometric knowledge. Here are some basic results of this type:

Proposition 13.3. *We have the following results:*

(1) *When f is linear, we have the following formula:*

$$\int_a^b f(x)dx=(b-a)\cdot\frac{f(a)+f(b)}{2}$$

(2) *In fact, when f is piecewise linear on $[a=a_1,a_2,\ldots,a_n=b]$, we have:*

$$\int_a^b f(x)dx=\sum_{i=1}^{n-1}(a_{i+1}-a_i)\cdot\frac{f(a_i)+f(a_{i+1})}{2}$$

(3) *We have as well the formula $\int_{-1}^1\sqrt{1-x^2}\,dx=\pi/2$.*

Proof. These results all follow from basic geometry, as follows:

(1) Assuming $f\geq 0$, we must compute the area of a trapezoid having sides $f(a),f(b)$, and height $b-a$. But this is the same as the area of a rectangle having side $(f(a)+f(b))/2$ and height $b-a$, and we obtain $(b-a)(f(a)+f(b))/2$, as claimed.

(2) This is clear indeed from the formula found in (1), by additivity.

(3) The integral in the statement is by definition the area of the upper unit half-disc. But since the area of the whole unit disc is π, this half-disc area is $\pi/2$. □

As an interesting observation, (2) in the above result makes it quite clear that f does not necessarily need to be continuous, in order to talk about its integral. Indeed, assuming that f is piecewise linear on $[a=a_1,a_2,\ldots,a_n=b]$, but not necessarily continuous, we can still talk about its integral, in the obvious way, exactly as in Definition 13.1, and we have an explicit formula for this integral, generalizing the one found in (2), namely:

$$\int_a^b f(x)dx=\sum_{i=1}^{n-1}(a_{i+1}-a_i)\cdot\frac{f(a_i^+)+f(a_{i+1}^-)}{2}$$

Based on this observation, let us upgrade our formalism, as follows:

Definition 13.4. *We say that a function $f:[a,b]\to\mathbb{R}$ is integrable when the area below its graph is computable. In this case we denote by*

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

this area, signed $+$ where $f\geq 0$, and signed $-$ where $f\leq 0$.

As basic examples of integrable functions, we have the continuous ones, provided indeed that our intuition, or that Method 13.2, works indeed for any such function. We will soon see that this is indeed true, coming with mathematical proof. As further examples, we have the functions which are piecewise linear, or piecewise continuous. We will also see, later, as another class of examples, that the piecewise monotone functions are integrable. But more on this later, let us not bother for the moment with all this.

This being said, one more thing regarding theory, that you surely have in mind: is any function integrable? Not clear. I would say that if the Devil comes with some sort of nasty, totally discontinuous function $f:\mathbb{R}\to\mathbb{R}$, then you will have big troubles in cutting its graph from 3mm plywood, as required by Method 13.2. More on this later.

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

Proposition 13.5. *We have the following formulae,*

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

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

valid for any functions f,g and any scalar $\lambda\in\mathbb{R}$.

Proof. Both these formulae are indeed clear from definitions. □

Moving ahead now, passed the above results, which are of purely algebraic and geometric nature, and perhaps a few more of the same type, which are all quite trivial and that we we will not get into here, we must do some analysis, in order to compute integrals. This is something quite tricky, and we have here the following result:

Theorem 13.6. *We have the Riemann integration formula,*

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

which can serve as a definition for the integral.

Proof. This is standard, by drawing rectangles. We have indeed the following formula, which can stand as a definition for the signed area below the graph of f:

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

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.

13b. Monte Carlo

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. □

13c. 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 ε:

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

(2) In order to prove this claim, let us denote by l the length of the intervals on the real line. Our assumption is that the lengths of the divisions I,I' satisfy:

$$l\big([a_i,a_{i+1}]\big)<\frac{\delta}{2}\quad,\quad l\big([a_i',a_{i+1}']\big)<\frac{\delta}{2}$$

Now let us intersect the intervals of our divisions I,I', and set:

$$l_{ij}=l\big([a_i,a_{i+1}]\cap[a_j',a_{j+1}']\big)$$

The difference of Riemann sums that we are interested in is then given by:

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

(3) Now let us estimate $f(a_i)-f(a_j')$. Since in the case $l_{ij}=0$ we do not need this estimate, we can assume $l_{ij}>0$. Now by remembering what the definition of the numbers l_{ij} was, we conclude that we have at least one point $x\in\mathbb R$ satisfying:

$$x\in[a_i,a_{i+1}]\cap[a_j',a_{j+1}']$$

But then, by using this point x and our assumption on I,I' involving δ, we get:

$$\begin{aligned}|a_i-a_j'|&\leq|a_i-x|+|x-a_j'|\\&\leq\frac{\delta}{2}+\frac{\delta}{2}\\&=\delta\end{aligned}$$

Thus, according to our definition of δ from (1), in relation to ε, we get:

$$|f(a_i)-f(a_j')|<\varepsilon$$

(4) But this is what we need, in order to finish. Indeed, with the estimate that we found, we can finish the computation started in (2), as follows:

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

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.

13d. General results

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,*

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

and with respect to taking infinite sums as well,

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

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 the clarification of all this as an instructive exercise. □

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*

$$\exists c\in[a,b]\quad,\quad\int_a^bf(x)dx=(b-a)f(c)$$

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:

$$(b-a)\min(f)\leq\int_a^bf(x)dx\leq(b-a)\max(f)$$

Now observe that this latter estimate can be written as follows:

$$\min(f)\leq\frac{\int_a^bf(x)dx}{b-a}\leq\max(f)$$

Since f must takes all values on $[\min(f),\max(f)]$, we get a $c\in[a,b]$ such that:

$$\frac{\int_a^bf(x)dx}{b-a}=f(c)$$

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

13e. Exercises

Exercises:

Exercise 13.15.

Exercise 13.16.

Exercise 13.17.

Exercise 13.18.

Exercise 13.19.

Exercise 13.20.

Exercise 13.21.

Exercise 13.22.

Bonus exercise.

CHAPTER 14

Riemann sums

14a. Power functions

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*

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

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:

$$\int_0^c x^pdx=\frac{c^{p+1}}{p+1}$$

(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:

$$\begin{matrix} & & & & \square \\ & & & & \vdots \\ & & \square & \ldots & \square \\ & \square & \square & \ldots & \square \\ \square & \square & \square & \ldots & \square \end{matrix}$$

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

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

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

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

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

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

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

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

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

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

Alternatively, assuming that the gods of combinatorics are with us, we can see right away the following formula, which coupled with (4) gives the result:

$$1^3 + 2^3 + \ldots + N^3 = (1 + 2 + \ldots + N)^2$$

In any case, in practice, the above formula holds indeed, and you can check it for instance by recurrence, and this solves our problem:

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

(8) Thus, 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,

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

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.

14b. Exponentials

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,*

$$\int_a^b e^xdx=e^b-e^a$$

valid for any two real numbers $a<b$.

PROOF. This follows indeed from the Riemann integration formula, because:

$$\begin{aligned}
\int_a^b e^x dx &= \lim_{N\to\infty}\frac{e^a+e^{a+(b-a)/N}+e^{a+2(b-a)/N}+\ldots+e^{a+(N-1)(b-a)/N}}{N}\\
&= \lim_{N\to\infty}\frac{e^a}{N}\cdot\left(1+e^{(b-a)/N}+e^{2(b-a)/N}+\ldots+e^{(N-1)(b-a)/N}\right)\\
&= \lim_{N\to\infty}\frac{e^a}{N}\cdot\frac{e^{b-a}-1}{e^{(b-a)/N}-1}\\
&= (e^b-e^a)\lim_{N\to\infty}\frac{1}{N\left(e^{(b-a)/N}-1\right)}\\
&= e^b-e^a
\end{aligned}$$

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

14c. Basic estimates

Basic estimates.

14d. Further results

Further results.

14e. Exercises

Exercises:

EXERCISE 14.4.

EXERCISE 14.5.

EXERCISE 14.6.

EXERCISE 14.7.

EXERCISE 14.8.

EXERCISE 14.9.

EXERCISE 14.10.

EXERCISE 14.11.

Bonus exercise.

CHAPTER 15

Main theorems

15a. Fundamental theorem

The problem is now, what to do with what we have, from the previous chapter. Not obvious, so stuck, and time to ask the cat. And cat says:

CAT 15.1. *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 chapter 14, 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 15.2. *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 15.3. *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 15.2, and is in fact equivalent to it. Indeed, consider the following function:

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

By using Theorem 15.2 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 15.4. *We have the following integration formulae,*

$$\int_a^b x^pdx=\frac{b^{p+1}-a^{p+1}}{p+1}\quad,\quad\int_a^b\frac{1}{x}\,dx=\log\left(\frac{b}{a}\right)$$

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

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

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 15.5. *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 15.4 in a more digest form, as follows:

Theorem 15.6. *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 15.4, or rather from the proof of Theorem 15.4, or even from chapter 9, for most of them, and the last assertion comes from the integration formula given after Definition 15.5. □

15b. Further theorems

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

Theorem 15.7. *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 15.8. *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 15.9. *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. □

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 15.10. *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 15.11. *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 φ, 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 ξ, η 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. □

15c. Areas, volumes

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

Proposition 15.12. *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 15.13. *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. □

15d. Some probability

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

Definition 15.14. *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, δ_x with $x\in\mathbb R$, as components of φ.

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 15.14 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 15.14:

DEFINITION 15.15. *We say that a random variable f follows the density φ 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 φ, the first questions regard their mean and variance, constructed as follows:

DEFINITION 15.16. *Given a random variable f, with probability density φ:*

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 15.15 and Definition 15.16, what really matters is not the density φ itself, but rather the related quantity $\mu = \varphi(x)dx$. So, let us upgrade our formalism, as follows:

DEFINITION 15.17 (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 μ 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. Skipping some discussion here, and getting straight to the point, the most important laws in discrete probability are the Poisson laws, constructed as follows:

Definition 15.18. *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 15.19. *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. □

Many other things can be said, as a continuation of the above.

15e. Exercises

Exercises:

Exercise 15.20.

Exercise 15.21.

Exercise 15.22.

Exercise 15.23.

Exercise 15.24.

Exercise 15.25.

Exercise 15.26.

Exercise 15.27.

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 φ, 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 φ 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 φ' 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:

$$\varphi''(x)=\left(\frac{d^2\varphi}{dx_idx_j}\right)_{ij}$$

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 1×1 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 φ 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, $\varphi(z)=x^py^q$. We have then our formula, due to:

$$\frac{d^2\varphi}{dxdy}=\frac{d}{dx}\left(\frac{dx^py^q}{dy}\right)=\frac{d}{dx}\left(qx^py^{q-1}\right)=pqx^{p-1}y^{q-1}$$

$$\frac{d^2\varphi}{dydx}=\frac{d}{dy}\left(\frac{dx^py^q}{dx}\right)=\frac{d}{dy}\left(px^{p-1}y^q\right)=pqx^{p-1}y^{q-1}$$

Next, let us consider a linear combination of power functions, $\varphi(z)=\sum_{pq}c_{pq}x^py^q$, which can be finite or not. We have then, by using the above computation:

$$\frac{d^2\varphi}{dxdy}=\frac{d^2\varphi}{dydx}=\sum_{pq}c_{pq}pqx^{p-1}y^{q-1}$$

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 φ 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 φ,*

$$\Delta\varphi=\sum_{i=1}^N\frac{d^2\varphi}{dx_i^2}$$

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 Δ? 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 ∇ 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 Δ 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 Δ, 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 Δ, 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 Δ,*

$$\Delta\varphi=0$$

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 $z=x+iy$:

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

As for the second assertion, this follows from $\Delta\bar{\varphi}=\overline{\Delta\varphi}$, which is clear from definitions, and which shows that if φ 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 φ, arising as a real number". But this can only be the Jacobian $\det(\varphi'(t))$, and with this in mind, we are led to:

Theorem 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_φ quantity, called Jacobian, being given by

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

and with this generalizing the formula from Proposition 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:

$$V=2^NV^+$$

We have the following computation, using spherical coordinates:

$$\begin{aligned}V^+ &= \int_{B^+}1\\ &= \int_0^1\int_0^{\pi/2}\ldots\int_0^{\pi/2} r^{N-1}\sin^{N-2}t_1\ldots\sin t_{N-2}\,drdt_1\ldots dt_{N-1}\\ &= \int_0^1 r^{N-1}\,dr\int_0^{\pi/2}\sin^{N-2}t_1\,dt_1\ldots\int_0^{\pi/2}\sin t_{N-2}dt_{N-2}\int_0^{\pi/2}1dt_{N-1}\\ &= \frac{1}{N}\times\left(\frac{\pi}{2}\right)^{[N/2]}\times\frac{(N-2)!!}{(N-1)!!}\cdot\frac{(N-3)!!}{(N-2)!!}\ldots\frac{2!!}{3!!}\cdot\frac{1!!}{2!!}\cdot 1\\ &= \frac{1}{N}\times\left(\frac{\pi}{2}\right)^{[N/2]}\times\frac{1}{(N-1)!!}\\ &= \left(\frac{\pi}{2}\right)^{[N/2]}\frac{1}{(N+1)!!}\end{aligned}$$

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:

$$N!=\int_0^\infty x^Ne^{-x}dx$$

Since the integrand is sharply peaked at $x=N$, as you can see by computing the derivative of $\log(x^Ne^{-x})$, this suggests writing $x=N+y$, and we obtain:

$$\begin{aligned}\log(x^Ne^{-x}) &= N\log x-x\\ &= N\log(N+y)-(N+y)\\ &= N\log N+N\log\left(1+\frac{y}{N}\right)-(N+y)\\ &\simeq N\log N+N\left(\frac{y}{N}-\frac{y^2}{2N^2}\right)-(N+y)\\ &= N\log N-N-\frac{y^2}{2N}\end{aligned}$$

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 $M_k(f)=\int_\mathbb{R}x^kd\mu_f(x)$.*

Here the fact that a measure μ_f as above exists indeed is not exactly trivial. But we can do this by looking at formulae of the following type:

$$E(\varphi(f))=\int_\mathbb{R}\varphi(x)d\mu_f(x)$$

Indeed, having this for monomials $\varphi(x)=x^n$, as above, is the same as having it for polynomials $\varphi\in\mathbb{R}[X]$, which in turn is the same as having it for the characteristic functions $\varphi=\chi_I$ of measurable sets $I\subset\mathbb{R}$. Thus, in the end, what we need is:

$$P(f\in I)=\mu_f(I)$$

But this formula can serve as a definition for μ_f, and we are done.

Regarding now independence, we can formulate here the following definition:

Definition 16.20. *Two variables $f,g\in L^\infty(X)$ are called independent when*

$$E(f^kg^l)=E(f^k)\,E(g^l)$$

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:

$$\int_{\mathbb R}x^kd\mu_{f+g}(x)=E((f+g)^k)=\sum_r\binom{k}{r}M_r(f)M_{k-r}(g)$$

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

$$\begin{aligned}\int_{\mathbb R}x^kd(\mu_f*\mu_g)(x)&=\int_{\mathbb R\times\mathbb R}(x+y)^kd\mu_f(x)d\mu_g(y)\\&=\sum_r\binom{k}{r}M_r(f)M_{k-r}(g)\end{aligned}$$

Thus μ_{f+g} and $\mu_f*\mu_g$ have the same moments, so they coincide, as claimed. □

As 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*

$$F_{f+g}=F_fF_g$$

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:

$$\begin{aligned}M_2&=\frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}}x^2e^{-x^2/2t}dx\\&=\frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}}(tx)\left(-e^{-x^2/2t}\right)'dx\\&=\frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}}te^{-x^2/2t}dx\\&=t\end{aligned}$$

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:

$$\begin{aligned}F_f(x)&=&E\left(\sum_{k=0}^\infty\frac{(ixf)^k}{k!}\right)\\&=&\sum_{k=0}^\infty\frac{(ix)^kE(f^k)}{k!}\\&=&\sum_{k=0}^\infty\frac{i^kM_k(f)}{k!}\,x^k\end{aligned}$$

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 $k\in\mathbb{N}$:

$$\begin{aligned}M_k&=&\frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}}y^ke^{-y^2/2t}dy\\&=&\frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}}(ty^{k-1})\left(-e^{-y^2/2t}\right)'dy\\&=&\frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}}t(k-1)y^{k-2}e^{-y^2/2t}dy\\&=&t(k-1)\times\frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}}y^{k-2}e^{-y^2/2t}dy\\&=&t(k-1)M_{k-2}\end{aligned}$$

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