Principles of mathematics

Teo Banica

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

2010 *Mathematics Subject Classification.* 01A67

Key words and phrases. Algebra, Geometry

Abstract. This is an introduction to mathematics, with emphasis on algebra and geometry. We first discuss numbers, fractions and percentages, and their basic applications, followed by real numbers, and with a look into philosophy and logic too. Then we get into plane geometry and trigonometry, and coordinates and some space geometry. We then go back to numbers, with more advanced theory, in relation with divisibility, prime numbers and related topics, and polynomials and their roots. Finally, we provide an introduction to functions and analysis, with the basics of the theory, followed by exponentials, logarithms and more trigonometry, and with the derivatives explained too.

Preface

Planet Earth, year 2090. A bit dark outside, for this time of the year, isn't it. Not many people around either, and for things like electricity, forget about it. But after all, it's not that bad, all this relaxation and silence. There is certainly food around, to be gathered, and wood for fire, and some folks left too, to hang out with, from time to time. And for electricity, civilization and stuff, do we really miss that, and we'll see later.

Congratulations, first, for having survived. I bet you don't even have an idea on what happened. Neither do I, writing from here, back in time in the 2020s, but I can only imagine that the marxist revolution has succeeded, sometimes around 2070, or at least, that was the plan. And then, what can I say, among these marxist folks the communists are usually reasonably peaceful, but not very sure about their various brothers and allies, these might have got into some form of disagreement, or something like that.

Anyway, life goes on, and time now for hunting, fishing, making fires, perhaps a bit of agriculture too, why not some metallurgy and medicine too. And good luck in learning all this, I have no idea where exactly from. In fact, I can only imagine that, with only the strong having survived, there is no college graduate left, on the whole planet.

This book will be here for teaching you some mathematics. Sure this is something a bit secondary, with respect to your technological needs at the present time, but the Winter nights are long and cold, and once done with repairing your gear, and doing other useful things, still plenty of time left, and have a look from time to time at this. Mathematics, and I'm telling you this, is something quite useful, not invented just for the sake of inventing things, and you will certainly learn some good tricks from here.

The book, which by the way is certified first-class mathematics, originally written for the mid-century marxist guerrilla, is organized in 4 parts, as follows:

Part I, with I actually standing for 1, and with this being a minor bug, deals with numbers. We will discuss here how to count things, in the best possible way.

Part II deals with angles, triangles and geometry. This knowledge, which is more advanced, is useful when building things, for craftsmanship, and sailing.

Part III goes back to numbers, which remain something extremely useful, and discusses more advanced aspects of them, sometimes in relation with geometry.

Part IV is an introduction to more advanced mathematics, namely functions and analysis, again with motivations coming from craftsmanship, and geometry.

In the hope that you will appreciate all this, and please, pass this knowledge to your friends, and children too. And do not do the same mistakes as your ancestors did, just live your life, and believe in the Sun, in Water, and in Fire, and things will be fine.

Many thanks to my various math school professors from the communist Romania, where I learned this stuff from, good and serious learning that was. Thanks as well to my colleagues and students here in France, every now and then I learn something new about basic mathematics, and good learning this is too. Finally, many thanks to my cats, for some help with trigonometry, that was the hardest part to write, dammit.

Cergy, January 2025
Teo Banica

Contents

Part I

Numbers

Oh, Shenandoah
I long to hear you
Look away, we're bound away
Across the wide Missouri

CHAPTER 1

Numbers

1a. Numbers

You certainly know a bit about numbers $1,2,3,4,\ldots$, and we will be here, with this book, for learning more about them. Many things can be said here, but instead of starting right away with some complicated mathematics, it is wiser to relax, and go back to these small numbers $1,2,3,4,\ldots$ that you know well, and have some more thinking at them. After all, these small numbers are something quite magic, worth some more thinking. And with the thinking work that we will be doing here being something useful.

So, reviewing the material from elementary school. Shall we start with 7×8, or perhaps with 6×7? I don't know about you, but personally I found these two computations both quite difficult, as a kid, these multiples of 7 are no joke, when learning arithmetic.

In answer, these are indeed tough computations, forget about them, and let us start with the very basics. Here will be our method, which is quite philosophical:

Method 1.1. *In order to better understand the small numbers* $1,2,3,4,\ldots$ *and their arithmetic, the best is to forget about these numbers, and reinvent them. With this being guaranteed to work, an inventor being not supposed to ever forget his invention.*

Ready for this? Hang on, and getting started now, here we are, in the dark. It is actually most convenient here to do assume that we are in the dark, say in a Stone Age cavern, lit only by a small fire, and with a pile of bloody ribs waiting to be counted, cooked, and eaten by our community. So, how to count these bloody ribs?

As a simple solution, we can invent some words for counting, ribs or any other type of objects. And going here with English, here is a proposal, for our first numbers:

$$\text{one, two, three, four, } \ldots$$

However, this method obviously has some limitations, because the more objects we want to count, the more words we will have to invent for them, and this is not very funny. In fact, we even risk, as leaders, to be killed and eaten by the tribe, on the grounds that our mathematics is too complicated and annoying. Well, this is how things were going during the Stone Age, people being honest and direct, nothing to do with the students nowadays, politely listening to whatever their math professor teaches them.

In short, we are in trouble here, and as problem to be solved, we have:

Problem 1.2. *Words are not very good for counting, we must invent something else, say some sort of bizarre signs.*

So, let us attempt to invent some suitable signs, doing the counting. The first thought here goes to the ribs themselves, that we want to count, which can be designated, pictorially, by vertical bars $|$. And with this, we certainly have our improved numeration system, which starts as follows, and can be continued indefinitely:

$$|\,,\ ||\,,\ |||\,,\ ||||\,,\ \ldots$$

However, there are still some bugs, with this new system, which remains not very practical for big numbers, say when counting small fruits. In addition, it is a bit of a pity to completely give up language, and to have no words for our signs, after all our one, two, three, four were not that bad, for the small numbers, and we are missing them.

A good solution to this, again by thinking at ribs, comes by thinking as well at the animals these ribs come from. Indeed, and by going now a bit abstract, we can group ribs into animals, and we can reach in this way to an even better numeration system. However, there are many ways of proceeding here, depending on how many ribs do we want our animals to have, on what signs we want to designate these animals, and also, on what words shall we use for designating the ribs inside such an animal.

Solving all these questions, in an ideal way for practice, does not look easy, so let us start with an attempt, and we will fine-tune later. Here is our definition:

Definition 1.3. *The numbers are signs of the following type,*

$$\bigcirc\ \bigcirc\ \bigcirc\,|\,|\,|\,|\,|$$

with each circle standing for an animal, itself standing for a number of ribs, according to:

$$\bigcirc=|\,|\,|\,|\,|\,|$$

Also, we agree to designate the number of ribs inside an animal by the words

one, two, three, four, five, six

and for counting animals, we can use these words too, followed by "ty".

Here "ty" is the name of a certain fatty and tasty animal, sort of a big and peaceful herbivore, which was wisepread during the Stone Age, and highly prized by our ancestors, but which unfortunately dissapeared in more modern times, due to overhunting.

So, very good, we have now our numbers, and even some nice words for designating them. As an example, here is a quite big number, that we can use whenever needed:

$$\bigcirc\ \bigcirc\ \bigcirc\,|\,|\,|\,|\ =\ \text{threety} - \text{four}$$

In practice now, we can do many things wich such numbers, but when it comes to counting seeds, or small fruits, we quite often reach to the limit of what we can do, with our numbering system, and more specifically, to the following number:

$$\bigcirc\ \bigcirc\ \bigcirc\ \bigcirc\ \bigcirc\ \bigcirc\quad |\,|\,|\,|\,|\,| = \text{sixty} - \text{six}$$

Of course, some tricks can be used here, but none is very good. For truly improving our numbering system, the best is to go back to Definition 1.3, and further recycle the idea there. Indeed, animals can be grouped into herds, and we are led in this way to:

Definition 1.4. *The numbers are signs of the following type,*

$$\bigstar\ \bigstar\ \bigstar\quad \bigcirc\ \bigcirc\quad |\,|\,|\,|$$

with the circles standing for animals, and the stars standing for herds, according to:

$$\bigcirc = |\,|\,|\,|\,|\,|\quad ,\quad \bigstar = \bigcirc\ \bigcirc\ \bigcirc$$

Also, we agree to designate the number of ribs inside an animal by the words

one, two, three, four, five, six

and for counting animals or herds, we can use these words, followed by "ty" and "gh".

Which looks very nice, because with this we can now count pretty much everything in this world, with our system being now bound by the following fairly large number:

$$\bigstar\ \bigstar\ \bigstar\ \bigstar\ \bigstar\ \bigstar\quad \bigcirc\ \bigcirc\quad |\,|\,|\,|\,|\,| = \text{sixgh} - \text{twoty} - \text{six}$$

This being said, there must be certainly room for better. Looking at the above big number, there is obviously something a bit wrong with it, and this leads us into:

Theorem 1.5. *For best results with our system, it is ideal to assume that the number of ribs of an animal equals the number of animals in a herd.*

Proof. This is somewhat obvious, because in the context of Definition 1.4, we can certainly improve everything there by assuming that herds consist of six animals. □

So, here we go again with improving our system, with our new definition being:

Definition 1.6. *The numbers are signs of the following type,*

$$\bigstar\ \bigstar\ \bigstar\quad \bigcirc\ \bigcirc\quad |\,|\,|\,|$$

with the circles standing for animals, and the stars standing for herds, according to:

$$\bigcirc = |\,|\,|\,|\,|\,|\quad ,\quad \bigstar = \bigcirc\ \bigcirc\ \bigcirc\ \bigcirc\ \bigcirc\ \bigcirc$$

Also, we agree to designate the number of ribs inside an animal by the words

one, two, three, four, five, six

and for counting animals or herds, we can use these words, followed by "ty" and "gh".

And with this, not only everything looks more logical and practical, but we can now count up to the following extremely large number:

★ ★ ★ ★ ★ ★ ○ ○ ○ ○ ○ ○ | | | | | | = sixgh – sixty – six

However, thinking some more, we can still improve this, simply by coming with some easy to draw symbols, representing one, two, three, four, five, six, as for instance:

$$1 = \text{one}$$

$$2 = \text{two}$$

$$3 = \text{three}$$

$$4 = \text{four}$$

$$5 = \text{five}$$

$$6 = \text{six}$$

Indeed, in the context of Definition 1.6, we can simply replace the rib, animal and herd symbols there by these new symbols, and things get easier. As an example here, the number given as example in Definition 1.6 take now the following simple form:

★ ★ ★ ○ ○ | | | | → 324

As for the biggest possible number, discussed above, this becomes:

★ ★ ★ ★ ★ ★ ○ ○ ○ ○ ○ ○ | | | | | | → 666

However, thinking some more, there is a bit of a bug with all this, because how to designate for instance the following number, with our new system:

★ ★ ★ | | | | → ?

In answer, we need a new symbol, for designating the lack of circles, or even better, the lack of anything, in general. Which looks like a quite tricky idea, so let us record this finding as a Theorem, with this meaning, as usual, thing found via hard work:

Theorem 1.7. *In order to improve our system, we need a new symbol, say*

$$0 = \text{zero}$$

standing for the lack of anything.

Proof. As already said, this is something that we came upon via some hard thinking. But now that we have it, the thing itself look quite trivial, so very good. □

Now armed with our new symbols $1, 2, 3, 4, 5, 6$, and with the above tricky symbol 0 too, we can substantially improve Definition 1.6, in the following way:

Definition 1.8. *The numbers are signs of the following type, with the components, called digits, standing for the number of herds, animals, and ribs*

$$253$$

and with the digits themselves designating the number of ribs inside an animal, from none up to all of them, according to the following system,

$$0=\text{zero},\ 1=\text{one},\ 2=\text{two},\ 3=\text{three},\ 4=\text{four},\ 5=\text{five},\ 6=\text{six}$$

telling us as well the words corresponding to these digits. For reading numbers, we agree as before to use these words, followed by "ty", "gh", and nothing at all.

Looks like we are now into quite serious mathematics, with our new system. However, there is still room for improvement, because we can forget if we want about ribs, animals and herds, and with this leading us into even bigger numbers, in the following way:

Definition 1.9. *The numbers are signs of the following type, of arbitrary length*

$$24015$$

with the components, called digits, and the words designating them being:

$$0=\text{zero},\ 1=\text{one},\ 2=\text{two},\ 3=\text{three},\ 4=\text{four},\ 5=\text{five},\ 6=\text{six}$$

For reading numbers, we can use these words, followed, in reverse order of appearance, by nothing at all, and then by "ty", "ry", "fy", "vy", "sy".

Here everything is quite self-explanatory, the idea being of course that we are expanding here our basic rib-animal-herd counting system with more and categories, of type "herds of herds" and so on, but with a problem coming from the fact that we are in the lack of a good system of words, for designating these new categories. However, in what regards reading the corresponding numbers, this is an easier problem, and we can use the system proposed as the end, which is something quite logical, coming from:

$$\begin{aligned}2=\text{two}&\quad\to\quad ty\\3=\text{three}&\quad\to\quad ry\\4=\text{four}&\quad\to\quad fy\\5=\text{five}&\quad\to\quad vy\\6=\text{six}&\quad\to\quad sy\end{aligned}$$

So, let us see how this latter system works. As a first example, we have:

$$23051=\text{twovy}-\text{threefy}-\text{fivety}-\text{one}$$

Which sound quite good, at least to my personal non-English native speaker ear. Let us record as well the biggest number that we can pronounce, with our system:

$$666666=\text{sixsy}-\text{sixvy}-\text{sixfy}-\text{sixry}-\text{sixty}-\text{six}$$

Which again, sounds quite good. It looks possible of course to work some more here, and come up with some further improvements to our system, and it is tempting to do indeed so. However, relaxing a bit, and looking at what we did so far, we are led into the following question, which perhaps is more fundamental, and comes first:

Question 1.10. *The number six plays a special role in the above, with* 6 *being the biggest digit. So, can we improve our system, by replacing six by other numbers?*

And tricky question this is, because thinking a bit at it, it is not even clear to which branch of science it belongs to. We will attempt to solve it, in the next section.

1b. Numeration bases

As explained above, the question is now, what should be the common number that we are using, of ribs of an animal, or of animals in a herd, and so on?

This is a quite subtle question, whose answer is not obvious, and this even if you know well math, as many of our ancestors did, over the centuries. So, let us work out some examples. As a first example here, which is something a bit formal, we have:

Example 1.11. *Numeration basis two.*

Many things can be said here, and we can even start, with this, to do some serious mathematics, with tables and rules for addition and multiplication, and for substraction and division too, and with many other interesting things that can be said, about this.

As a comment here, this system is not that unuseful or obsolote, because this is more or less what computer scientists are using, nowadays. But more on this later.

Next on our list, coming natually after numeration basis two, is of course:

Example 1.12. *Numeration basis three.*

As before, many things can be said here. Pros and cons. Note in passing that we are learning good mathematics here, with our numeration systems.

Coming next, we have:

Example 1.13. *Numeration basis four.*

This is somehow better than numeration basis two. Good to know.

Coming next, we have:

Example 1.14. *Numeration basis five.*

Quite interesting, and also nice pictorially, still used on prison walls, and in many other concrete situations. Pesronally, this is my favorite system, for counting things.

Coming next, we have:

Example 1.15. *Numeration basis six.*

Again, this is something quite interesting, nicely mixing two and three, and quite natural too, that we have already talked about in the previous section.

Coming next, we have:

Example 1.16. *Numeration basis seven.*

And with this, we are definitely into serious mathematics, and inventions, because that terrible 7×8 computation from school takes now a very simple form.

And we will stop here with our list of examples. But the question comes now, which system to use? And we have here several schools of thought:

(1) Numeration basis two, or better, four, or even better, eight, or perhaps even sixteen, or why not sixty-four, are something very natural and useful. In practice, and in view of what we can do, and what we can't, the choice is between eight and sixteen.

(2) Numeration basis three, or much better, because even, six, or why now twelve, or twenty-four are something natural and useful too. In practice now, again in view of what we can do, and what we can't, the choice here is between six and twelve.

(3) Finally, we have numeration basis five, or much better, because even, ten. Not very clear what the advantage of using ten would be, but at least, as an interesting observation, at least there is no dillema here, with fifty being barred, as being too big.

So, this was for the story of the bases of numeration, and in what follows we will use, as everyone or almost nowadays, basis ten, somehow for the reasons discussed above. As for the ten digits needed, my proposal would be to use the following signs:

$$0,1,2,3,4,5,6,7,8,9$$

And with this, we are ready to go, into some serious arithmetic.

1c. Sums and products

Sums and products.

1d. Basic arithmetic

Basic arithmetic.

1e. Exercises

Exercises:

Exercise 1.17.

Exercise 1.18.

Exercise 1.19.

Exercise 1.20.

Exercise 1.21.

Exercise 1.22.

Exercise 1.23.

Exercise 1.24.

Bonus exercise.

CHAPTER 2

Fractions

2a. Fractions

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 2.1. *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 never ever forget 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.

2b. Binomials, factorials

As a first theorem, solving a problem which often appears in real life, we have:

Theorem 2.2. *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 2.3. *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 2.2, we have:

CONVENTION 2.4. *By definition,* $0! = 1$.

This convention comes, and no surprise here, from Advice 2.3. Indeed, we obviously have $\binom{n}{n} = 1$, but if we want to recover this formula via Theorem 2.2 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 2.2 complemented by Convention 2.4 being in final form (trust me), we have:

THEOREM 2.5. *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 2.2, we have:

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

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

Theorem 2.5 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, what matters are the coefficients $\binom{n}{k}$, which form a triangle. So, it is enough to memorize this triangle, and this can be done by using:

THEOREM 2.6. *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, and we will be back to this, later in this book, on a regular basis, and with the idea being always the same, namely that in order to find such formulae you have a choice between algebra and combinatorics, a bit as in the above, and that when it comes to formal proofs, the brute-force computation method is something useful too.

In practice, the best is to master all 3 techniques. Among others, you will have in this way 3 different methods, for making sure that your formulae are correct indeed.

2c. Some probability

As an application to what we learned so far in this chapter, namely fractions, percentages, and rational numbers in general, let us do some probability. We first have:

Theorem 2.7. *The probabilities at poker are as follows:*

1. *One pair:* 0.533.
2. *Two pairs:* 0.120.
3. *Three of a kind:* 0.053.
4. *Full house:* 0.006.
5. *Straight:* 0.005.
6. *Four of a kind:* 0.001.
7. *Flush:* 0.000.
8. *Straight flush:* 0.000.

Proof. Let us consider indeed our deck of 32 cards, $7, 8, 9, 10, J, Q, K, A$. The total number of possibilities for a poker hand is:

$$\binom{32}{5} = \frac{32 \cdot 31 \cdot 30 \cdot 29 \cdot 28}{2 \cdot 3 \cdot 4 \cdot 5} = 32 \cdot 31 \cdot 29 \cdot 7$$

(1) For having a pair, the number of possibilities is:

$$N = \binom{8}{1}\binom{4}{2} \times \binom{7}{3}\binom{4}{1}^3 = 8 \cdot 6 \cdot 35 \cdot 64$$

Thus, the probability of having a pair is:

$$P = \frac{8 \cdot 6 \cdot 35 \cdot 64}{32 \cdot 31 \cdot 29 \cdot 7} = \frac{6 \cdot 5 \cdot 16}{31 \cdot 29} = \frac{480}{899} = 0.533$$

(2) For having two pairs, the number of possibilities is:

$$N = \binom{8}{2}\binom{4}{2}^2 \times \binom{24}{1} = 28 \cdot 36 \cdot 24$$

Thus, the probability of having two pairs is:

$$P = \frac{28 \cdot 36 \cdot 24}{32 \cdot 31 \cdot 29 \cdot 7} = \frac{36 \cdot 3}{31 \cdot 29} = \frac{108}{899} = 0.120$$

(3) For having three of a kind, the number of possibilities is:

$$N = \binom{8}{1}\binom{4}{3} \times \binom{7}{2}\binom{4}{1}^2 = 8 \cdot 4 \cdot 21 \cdot 16$$

Thus, the probability of having three of a kind is:

$$P = \frac{8 \cdot 4 \cdot 21 \cdot 16}{32 \cdot 31 \cdot 29 \cdot 7} = \frac{3 \cdot 16}{31 \cdot 29} = \frac{48}{899} = 0.053$$

(4) For having full house, the number of possibilities is:

$$N = \binom{8}{1}\binom{4}{3} \times \binom{7}{1}\binom{4}{2} = 8 \cdot 4 \cdot 7 \cdot 6$$

Thus, the probability of having full house is:

$$P=\frac{8\cdot 4\cdot 7\cdot 6}{32\cdot 31\cdot 29\cdot 7}=\frac{6}{31\cdot 29}=\frac{6}{899}=0.006$$

(5) For having a straight, the number of possibilities is:

$$N=4\left[\binom{4}{1}^4-4\right]=16\cdot 63$$

Thus, the probability of having a straight is:

$$P=\frac{16\cdot 63}{32\cdot 31\cdot 29\cdot 7}=\frac{9}{2\cdot 31\cdot 29}=\frac{9}{1798}=0.005$$

(6) For having four of a kind, the number of possibilities is:

$$N=\binom{8}{1}\binom{4}{4}\times\binom{7}{1}\binom{4}{1}=8\cdot 7\cdot 4$$

Thus, the probability of having four of a kind is:

$$P=\frac{8\cdot 7\cdot 4}{32\cdot 31\cdot 29\cdot 7}=\frac{1}{31\cdot 29}=\frac{1}{899}=0.001$$

(7) For having a flush, the number of possibilities is:

$$N=4\left[\binom{8}{4}-4\right]=4\cdot 66$$

Thus, the probability of having a flush is:

$$P=\frac{4\cdot 66}{32\cdot 31\cdot 29\cdot 7}=\frac{33}{4\cdot 31\cdot 29\cdot 7}=\frac{9}{25172}=0.000$$

(8) For having a straight flush, the number of possibilities is:

$$N=4\cdot 4$$

Thus, the probability of having a straight flush is:

$$P=\frac{4\cdot 4}{32\cdot 31\cdot 29\cdot 7}=\frac{1}{2\cdot 31\cdot 29\cdot 7}=\frac{1}{12586}=0.000$$

Thus, we have obtained the numbers in the statement. □

So far, so good, but you might argue, what if we model our problem as for our poker hand to be ordered, do we still get the same answer? In answer, sure yes, but let us check this. The probability for having four of a kind, computed in this new way, is then:

$$P(\text{four of a kind})=\frac{8\cdot 5\cdot 4\cdot 3\cdot 2\cdot 28}{32\cdot 31\cdot 30\cdot 29\cdot 28}=\frac{1}{31\cdot 29}=\frac{1}{899}$$

To be more precise, here on the bottom $32\cdot 31\cdot 30\cdot 29\cdot 28$ stands for the total number of possibilities for an ordered poker hand, 5 out of 32, and on top, exercise for you to figure out what the above numbers 8, 5, then $4\cdot 3\cdot 2$, and 28, stand for.

2d. Binomial laws

Here is now a theorem about flipping coins:

THEOREM 2.8. *When flipping a coin k times what you can win are quantities of type $\$k-2s$, with $s=0,1,\ldots,k$, with the probability for this to happen being:*

$$P(k-2s)=\frac{1}{2^k}\binom{k}{s}$$

Geometrically, your winning curve starts with probability $1/2^k$ of winning $-\$k$, then increases up to the tie situation, and then decreases, up to probability $1/2^k$ of winning $\$k$.

PROOF. All this is quite clear, the whole point being that, in order for you to win $k-s$ times and lose s times, over your k attempts, the number of possibilities is:

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

Thus, by dividing now by 2^k, which is the total number of possibilities, for the whole game, we are led to the probability in the statement, namely:

$$P(k-2s)=\frac{1}{2^k}\binom{k}{s}$$

Shall we doublecheck this? Sure yes, doublecheking is the first thing to be done, when you come across a theorem, in your mathematics. As a first check, the sum of probabilities that we found should be 1, which is intuitive, right, and 1 that is, as shown by:

$$\begin{aligned}\sum_{s=0}^kP(k-2s)&=&\sum_{s=0}^k\frac{1}{2^k}\binom{k}{s}\\&=&\frac{1}{2^k}\sum_{s=0}^k\binom{k}{s}\\&=&\frac{1}{2^k}\sum_{s=0}^k\binom{k}{s}1^s1^{k-s}\\&=&\frac{1}{2^k}(1+1)^k\\&=&\frac{1}{2^k}\times 2^k\\&=&1\end{aligned}$$

But shall we really trust this. Imagine for instance that you play your game for \$1000 instead of \$1 as basic gain, your life is obviously at stake, so all this is worth a second doublecheck, before being used in practice. So, as second doublecheck, let us verify that, on average, what you win is exactly \$0, which is something very intuitive, the game itself obviously not favoring you, nor your partner. But this can be checked as follows:

$$\begin{aligned}
\sum_{s=0}^k P(k-2s)\times(k-2s) &= \frac{1}{2^k}\sum_{s=0}^k\binom{k}{s}(k-2s)\\
&= \frac{1}{2^k}\sum_{s=0}^k\binom{k}{s}(k-s)-\frac{1}{2^k}\sum_{s=0}^k\binom{k}{s}s\\
&= \frac{1}{2^k}\sum_{s=0}^k\binom{k}{s}(k-s)-\frac{1}{2^k}\sum_{t=0}^k\binom{k}{k-t}(k-t)\\
&= \frac{1}{2^k}\sum_{s=0}^k\binom{k}{s}(k-s)-\frac{1}{2^k}\sum_{t=0}^k\binom{k}{t}(k-t)\\
&= 0
\end{aligned}$$

Here we have used a change of indices, namely $s=k-t$, along with the following formula, which is clear from the definition of binomial coefficients:

$$\binom{k}{t}=\binom{k}{k-t}$$

Summarizing, we have a good and valid theorem here, ready to be used in practice. □

More generally now, we have the following result:

THEOREM 2.9. *The following happen, in the context of a biased coin game:*

(1) *The Bernoulli laws μ_{ber} produce the binomial laws μ_{bin}, by iterating the game $k\in\mathbb{N}$ times, via the independence of the throws.*
(2) *We have in fact $\mu_{bin}=\mu_{ber}^{*k}$, with $*$ being the convolution operation for real probability measures, given by $\delta_x*\delta_y=\delta_{x+y}$, and linearity.*

PROOF. Obviously, this is something a bit informal, but let us prove this as stated, and we will come back later to it, with precise definitions, theorems and everything. In what regards the first assertion, nothing to be said there, this is what life teaches us. As for the second assertion, the formula $\mu_{bin}=\mu_{ber}^{*k}$ there certainly looks like mathematics, so job for us to figure out what this exactly means. And, this can be done as follows:

(1) The first idea is to encapsulate the data from the coin game into the probability measures associated to the Bernoulli and binomial laws. For the Bernoulli law, the corresponding measure is as follows, with the δ symbols standing for Dirac masses:

$$\mu_{ber}=(1-p)\delta_0+p\delta_1$$

As for the binomial law, here the measure is as follows, constructed in a similar way, you get the point I hope, again with the δ symbols standing for Dirac masses:

$$\mu_{bin}=\sum_{s=0}^k p^s(1-p)^{k-s}\binom{k}{s}\delta_s$$

(2) Getting now to independence, the point is that, as we will soon discover abstractly, the mathematics there is that of the following formula, with $*$ standing for the convolution operation for the real measures, which is given by $\delta_x*\delta_y=\delta_{x+y}$ and linearity:

$$\mu_{bin}=\underbrace{\mu_{ber}*\ldots*\mu_{ber}}_{k\ terms}$$

(3) To be more precise, this latter formula does hold indeed, as a straightforward application of the binomial formula, the formal proof being as follows:

$$\begin{aligned}\mu_{ber}^{*k}&=\big((1-p)\delta_0+p\delta_1\big)^{*k}\\&=\sum_{s=0}^k p^s(1-p)^{k-s}\binom{k}{s}\delta_0^{*(k-s)}*\delta_1^{*s}\\&=\sum_{s=0}^k p^s(1-p)^{k-s}\binom{k}{s}\delta_s\\&=\mu_{bin}\end{aligned}$$

(4) Summarizing, save for some uncertainties regarding what independence exactly means, mathematically speaking, and more on this in a moment, theorem proved. □

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

2e. Exercises

Exercises:

Exercise 2.10.

Exercise 2.11.

Exercise 2.12.

Exercise 2.13.

Exercise 2.14.

Exercise 2.15.

Exercise 2.16.

Exercise 2.17.

Bonus exercise.

CHAPTER 3

Real numbers

3a. Discussion

In more advanced mathematical terms, the operations on the rationals, namely sum, product and inversion, tell us that $\mathbb{Q}$ is a field, in the following sense:

Definition 3.1. *A field is a set F with a sum operation $+$ and a product operation $\times$, subject to the following conditions:*

(1) *$a+b=b+a$, $a+(b+c)=(a+b)+c$, there exists $0\in F$ such that $a+0=0$, and any $a\in F$ has an inverse $-a\in F$, satisfying $a+(-a)=0$.*
(2) *$ab=ba$, $a(bc)=(ab)c$, there exists $1\in F$ such that $a1=a$, and any $a\neq 0$ has a multiplicative inverse $a^{-1}\in F$, satisfying $aa^{-1}=1$.*
(3) *The sum and product are compatible via $a(b+c)=ab+ac$.*

The simplest possible field seems to be $\mathbb{Q}$. However, this is not exactly true, because, by a strange twist of fate, the numbers $0,1$, whose presence in a field is mandatory, $0,1\in F$, can form themselves a field, with addition as follows:

$$1+1=0$$

Let us summarize this finding, along with a bit more, obtained by suitably replacing our 2, used for addition, with an arbitrary prime number p, as follows:

Theorem 3.2. *The following happen:*

(1) *$\mathbb{Q}$ is the simplest field having the property $1+\ldots+1\neq 0$, in the sense that any field F having this property must contain it, $\mathbb{Q}\subset F$.*
(2) *The property $1+\ldots+1\neq 0$ can hold or not, and if not, the smallest number of terms needed for having $1+\ldots+1=0$ is a certain prime number p.*
(3) *$\mathbb{F}_p=\{0,1,\ldots,p-1\}$, with p prime, is the simplest field having the property $1+\ldots+1=0$, with p terms, in the sense that this implies $\mathbb{F}_p\subset F$.*

Proof. All this is basic number theory, the idea being as follows:

(1) This is clear, because $1+\ldots+1\neq 0$ tells us that we have an embedding $\mathbb{N}\subset F$, and then by taking inverses with respect to $+$ and $\times$ we obtain $\mathbb{Q}\subset F$.

(2) Again, this is clear, because assuming $1+\ldots+1=0$, with $p=ab$ terms, chosen minimal, we would have a formula as follows, which is a contradiction:

$$(\underbrace{1+\ldots+1}_{a\ terms})(\underbrace{1+\ldots+1}_{b\ terms})=0$$

(3) This follows a bit as in (1), with the copy $\mathbb{F}_p\subset F$ consisting by definition of the various sums of type $1+\ldots+1$, which must cycle modulo p, as shown by (2). □

Getting back now to our philosophical discussion regarding numbers, what we have in Theorem 3.2 is not exactly good news, suggesting that, on purely mathematical grounds, there is a certain rivalry between $\mathbb{Q}$ and $\mathbb{F}_p$, as being the simplest field. So, which of them shall we study, as being created first? Not an easy question, and as answer, we have:

Answer 3.3. *Ignoring what pure mathematics might say, and trusting instead physics and chemistry, we will choose to trust in $\mathbb{Q}$, as being the simplest field.*

In short, welcome to science, and with this being something quite natural for us, science being the topic of the present book. Moving ahead now, many things can be done with $\mathbb{Q}$, but getting straight to the point, one thing that fails is solving $x^2=2$:

Theorem 3.4. *The field $\mathbb{Q}$ does not contain a square root of 2:*

$$\sqrt{2}\notin\mathbb{Q}$$

In fact, among integers, only the squares, $n=m^2$ with $m\in\mathbb{N}$, have square roots in $\mathbb{Q}$.

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

(1) In what regards $\sqrt{2}$, assuming that $r=a/b$ with $a,b\in\mathbb{N}$ prime to each other satisfies $r^2=2$, we have $a^2=2b^2$, and so $a\in 2\mathbb{N}$. But then by using again $a^2=2b^2$ we obtain $b\in 2\mathbb{N}$ as well, which contradicts our assumption $(a,b)=1$.

(2) Along the same lines, any prime number $p\in\mathbb{N}$ has the property $\sqrt{p}\notin\mathbb{Q}$, with the proof here being as the above one for $p=2$, by congruence and contradiction.

(3) More generally, our claim is that any $n\in\mathbb{N}$ which is not a square has the property $\sqrt{n}\notin\mathbb{Q}$. Indeed, we can argue here that our number decomposes as $n=p_1^{a_1}\ldots p_k^{a_k}$, with $p_1,\ldots,p_k$ distinct primes, and our assumption that n is not a square tells us that one of the exponents $a_1,\ldots,a_k\in\mathbb{N}$ must be odd. Moreover, by extracting all the obvious squares from n, we can in fact assume $a_1=\ldots=a_k=1$. But with this done, we can set $p=p_1$, and the congruence argument from (2) applies, and gives $\sqrt{n}\notin\mathbb{Q}$, as desired. □

In short, in order to advance with our mathematics, we are in need to introduce the field of real numbers $\mathbb{R}$. You would probably say that this is very easy, via decimal writing, like everyone does, but before doing that, let me ask you a few questions:

(1) Honestly, do you really like the addition of real numbers, using the decimal form? Let us take, as example, the following computation:

$$12.456\,783\,872$$

$$+\ 27.536\,678\,377$$

This computation can surely be done, but, annoyingly, it must be done from right to left, instead of left to right, as we would prefer. I mean, personally I would be most interested in knowing first what happens at left, if the integer part is 39 or 40, but go do all the computation, starting from the right, in order to figure out that. In short, my feeling is that this addition algorithm, while certainly good, is a bit deceiving.

(2) What about multiplication. Here things become even more complicated, imagine for instance that Mars attacks, with δ-rays, which are something unknown to us, and $100,000$ stronger than γ-rays, and which have paralyzed all our electronics, and that in order to protect Planet Earth, you must do the following multiplication by hand:

$$12.456\,783\,872$$

$$\times\ 27.536\,678\,377$$

This does not look very inviting, doesn't it. In short, as before with the addition, there is a bit of a bug with all this, the algorithm being too complicated.

(3) Getting now to the problem that we were interested in, namely extracting the square root of 2, here the algorithm is as follows, not very inviting either:

$$1.4^2 < 2 < 1.5^2 \implies \sqrt{2} = 1.4\ldots$$

$$1.41^2 < 2 < 1.42^2 \implies \sqrt{2} = 1.41\ldots$$

$$1.414^2 < 2 < 1.415^2 \implies \sqrt{2} = 1.414\ldots$$

$$1.4142^2 < 2 < 1.4143^2 \implies \sqrt{2} = 1.4142\ldots$$

$$\ldots$$

In short, quite concerning all this, and don't count on such things, mathematics of the decimal form, if Mars attacks. Let us record these findings as follows:

Fact 3.5. *The real numbers $x \in \mathbb{R}$ can be certainly introduced via their decimal form, but with this, the field structure of $\mathbb{R}$ remains something quite unclear.*

And with this, it looks like we are a bit stuck, hope you agree with me.

3b. Real numbers

Getting now to the difficulties that we met, fortunately, there is a clever solution to this, due to Dedekind. His definition for the real numbers is as follows:

Definition 3.6. *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 3.7. *The equation $x^2=2$ has two solutions over the real numbers, namely the positive solution, denoted $\sqrt2$, and its negative counterpart, which is $-\sqrt2$.*

Proof. By using $x\to-x$, it is enough to prove that $x^2=2$ has exactly one positive solution $\sqrt2$. But this is clear, because $\sqrt2$ can only come from the following cut:

$$A_{\sqrt2}=\mathbb Q_-\bigsqcup\left\{p\in\mathbb Q_+\Big|p^2<2\right\}\quad,\quad B_{\sqrt2}=\left\{q\in\mathbb Q_+\Big|q^2>2\right\}$$

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

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

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

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

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

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

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

Theorem 3.9. *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:

$$A_x=\bigcup_{n\geq1}\left\{p\in\mathbb{Q}\Big|p\leq0.b_1\ldots b_n\right\}$$

(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 3.10. *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. □

3c. Some analysis

Getting back now to Theorem 3.9, that was definitely something quite difficult. Alternatively, we have the following definition for the real numbers:

Theorem 3.11. *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 chapter 2, we know from there what the rational numbers are. But, as a continuation of the material there, we can talk about the distance between such rational numbers, as being given by the formula in the statement, namely:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Theorem 3.12. *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 3.11. Thus, we get the result.

(3) Alternatively, by getting back to Theorem 3.9 and its proof, we can argue, based on that, that the $\mathbb{R}$ that we constructed coincides with the old $\mathbb{R}$ from Definition 3.6, the one constructed via Dedekind cuts, and this gives again all the assertions. □

3d. Irrational numbers

Many things can be said about rationals and irrationals, and we have:

THEOREM 3.13. *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. Following Fourier, we will do this by contradiction. So, assume $e = m/n$, with $m, n \in \mathbb{N}$, 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. □

We will be back to this in Part IV of the present book, when doing analysis.

3e. Exercises

Exercises:

Exercise 3.14.

Exercise 3.15.

Exercise 3.16.

Exercise 3.17.

Exercise 3.18.

Exercise 3.19.

Exercise 3.20.

Exercise 3.21.

Bonus exercise.

CHAPTER 4

Convergence

4a. Convergence

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.

Getting started, we already met, on several occasions, infinite sequences or sums, and their limits. Time now to clarify all this. We first have the following definition:

Definition 4.1. *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 4.2. *We have $1/n \to 0$.*

Proof. This is obvious, but let us prove it by using Definition 4.1. 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 4.1, 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 4.1 with:

Definition 4.3. *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 4.4. *We have $n^2 \to \infty$.*

Proof. As before, this is obvious, but let us prove it using Definition 4.3. We have:

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

Thus we can take $N = [\sqrt{K}] + 1$ in Definition 4.3, and we are done. □

We can unify and generalize Proposition 4.2 and Proposition 4.4, as follows:

Proposition 4.5. *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 4.2 and Proposition 4.4, first for a rational, and then for a real as well. □

There are many other interesting examples of explicit limits. Moving ahead, we have as well some general results about limits, summarized as follows:

Theorem 4.6. *The following happen:*

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

Proof. All this is elementary, coming from definitions:

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

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

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

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

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

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

(4) This is clear from definitions. □

Here are as well some general rules for computing limits:

Theorem 4.7. *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 4.8. *The $n \to \infty$ limits of quotients of polynomials are given by*

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

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

Proof. The first assertion comes from the following computation:

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

As for the second assertion, this comes from Proposition 4.5. □

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 4.9. *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 4.10. *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 4.11. *$\mathbb{R}$ 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 4.10 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 4.10. □

4b. Sums, series

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

Definition 4.12. *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 4.13. *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 4.14. *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 4.15. *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 4.14, 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 4.16. *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 4.14.

(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 4.15, 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 4.17. *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 4.15, we have:

Theorem 4.18. *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. □

4c. The number e

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

THEOREM 4.19. *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 4.20. *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 4.19 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 4.21. *We have the formula*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Observe that we used in the above a few things about functions, which are all intuitive, but not exactly trivial to prove. We will be back to this, with details, later on.

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 4.22. *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 used in this book, 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. □

4d. Poisson laws

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 4.23. *The probability for a random permutation $\sigma \in S_N$ to be a derangement, that is, to have no fixed points, is given by the following formula:*

$$P = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \ldots + (-1)^N \frac{1}{N!}$$

Thus we have the following asymptotic formula, in the $N \to \infty$ limit,

$$P \simeq \frac{1}{e}$$

with $e = 2.7182\ldots$ being the usual constant from analysis.

Proof. This is something very classical, which is best viewed by using the inclusion-exclusion principle. Consider indeed the following sets:

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

By inclusion-exclusion, the probability that we are interested in is given by:

$$\begin{aligned}P &= \frac{1}{N!}\left(|S_N|-\sum_i|S_N^i|+\sum_{i<j}|S_N^i\cap S_N^j|-\ldots+(-1)^N\sum_{i_1<\ldots<i_N}|S_N^{i_1}\cap\ldots\cap S_N^{i_N}|\right)\\ &= \frac{1}{N!}\sum_{k=0}^N(-1)^k\sum_{i_1<\ldots<i_k}(N-k)!\\ &= \frac{1}{N!}\sum_{k=0}^N(-1)^k\binom{N}{k}(N-k)!\\ &= \sum_{k=0}^N\frac{(-1)^k}{k!}\end{aligned}$$

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

More generally now, we have the following result:

THEOREM 4.24. *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. 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 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 (1), 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. Finally, the last assertion follows in a similar way, with all this being quite standard. □

4e. Exercises

Exercises:

Exercise 4.25.

Exercise 4.26.

Exercise 4.27.

Exercise 4.28.

Exercise 4.29.

Exercise 4.30.

Exercise 4.31.

Exercise 4.32.

Bonus exercise.

Part II

Geometry

But night is the cathedral
Where we recognized the sign
We strangers know each other now
As part of the whole design

CHAPTER 5

Triangles

5a. Parallel lines

Welcome to plane geometry. At the beginner level, which is ours for the moment, this is a story of points and lines. Here is a basic observation, to start with, and we will call this "axiom" instead of "theorem", as the statements which are true and useful are usually called, in mathematics, for reasons that will become clear in a moment:

Axiom 5.1. *Any two distinct points $P \neq Q$ determine a line, denoted PQ.*

Obviously, our axiom holds, and looks like something very useful. Need to draw anything, for various engineering purposes, at your job, or in your garage? The rule will be your main weapon, used exactly as in Axiom 5.1, that is, put the rule on the points $P \neq Q$ that your line must unite, and then draw that line PQ. Actually, in relation with this, we are rather used in practice to draw segments PQ. But in theory, meaning some sort of idealized practice, will having that segment extended to infinity hurt? Certainly not, so this is why our lines PQ in mathematics will be infinite, as above.

Getting now to point, as already announced, why is Axiom 5.1 an axiom, instead of being a theorem? You would probably argue here that this theorem can be proved by using a rule, as indicated above. However, and with my apologies for this, although rock-solid as a scientific proof, this rule thing does not stand as a mathematical proof. This is how things are, you will have to trust me here. And for further making my case, let me mention that my theoretical physics friends agree with me, on the grounds that, when looking with a good microscope at your rule, that rule is certainly bent.

Excuse me, but cat is here, meowing something. So, what is is, cat?

Cat 5.2. *In fact, spacetime itself is bent.*

Okay, thanks cat, so looks like we have multiple problems with the "rule proof" of Axiom 5.1, so that definitely does not qualify as a proof. And so Axiom 5.1 will be indeed an axiom, that is, a true and useful mathematical statement, coming without proof.

Getting now to more discussion, around Axiom 5.1, an interesting question appears in connection with our assumption there $P \neq Q$. Indeed, given a point Q in the plane, we can come up with a sequence of points $P_n \to Q$ vertically, and in this case the lines P_nQ

will all coincide with the vertical at Q. But we can then formally say that the $n \to \infty$ limit of these lines, which makes sense to be denoted QQ, is also the vertical at Q.

However, is this a good idea, or not. The point indeed is that, when doing exactly the same trick with a series of points $P_n \to Q$ horizontally, we will obtain in this way, as our limiting line QQ, the horizontal at Q. Which does not sound very good, but since we seem however to have some sort of valuable idea here, let us formulate:

Job 5.3. *Develop later some kind of analysis theory, generalizing plane geometry, where lines of type QQ make sense too, say as some sort of tangents.*

As a further comment now, still on Axiom 5.1, it is of course understood there that the points $P \neq Q$ appearing there, and the line PQ uniting them, lie in the given plane that we are interested in, in this Part I of the present book. However, Axiom 5.1 obviously holds too in space, and most likely, in higher dimensional spaces too.

So, the question which appears now is, on which type of spaces does Axiom 5.1 hold? And this is a quite interesting question, because if we take a sphere for instance, any two points $P \neq Q$ can be certainly united by a segment, which is by definition the shortest segment, on the sphere, uniting them. And, if we prolong this segment, in the obvious way, what we get is a circle uniting P, Q, that we can call line, and denote P, Q.

However, not so quick. There is in fact a bug with this, because if we take P to be the North Pole, and Q to be the South Pole, any meridian on the globe will do, as PQ. So, as a conclusion, Axiom 5.1 does not really hold on a sphere, but not by much.

Anyway, as before, we seem to have an idea here, so let us formulate:

Job 5.4. *Develop later some kind of advanced geometry theory, generalizing plane geometry, where certain lines PQ can take multiple values.*

And with this, done I guess with the discussion regarding Axiom 5.1, I can only presume that you got as tired of reading this, as I got tired of writing it. Well, this is how things are, geometry is no easy business, and there are certainly plenty of things to be done, and what we will be doing here, based on Axiom 5.1, will be just a beginning.

Excuse me, but cat is meowing again. So, what is it cat, and for God's sake, in the hope that this is not in connection with Axiom 5.1. Please have mercy.

Cat 5.5. *What about $PQ = \lambda P + (1 - \lambda)Q$ proving your axiom.*

Okay, thanks cat, but I was already having this in mind, for chapter 7 below. So, Axiom 5.1 remains an axiom, please everyone disagreeing with this get out of my math class, and enjoy the sunshine outside. And well, we will see later, in chapter 7 below, how cats and physicists can prove Axiom 5.1, or at least, what their claims are.

Moving ahead now, here is an interesting observation about lines and points in the plane, coming somehow as a complement to Axiom 5.1:

Observation 5.6. *Any two distinct lines $K \neq L$ determine a point, $P = K \cap L$, unless these two lines are parallel, $K||L$.*

So, what do we have here, axiom, theorem, or something else? Not very clear, but on the bottom line, this is something which is certainly true, useful, and provable as before, with a rule. Just carefully draw K, L, and you will certainly get upon $P = K \cap L$.

However, in contrast to Axiom 5.1, there is a bit of a bug with our statement, because we do not know yet, mathematically, what parallel lines means. So, let us formulate:

Definition 5.7. *We say that two lines are parallel, $K||L$, when they do not cross,*

$$K \cap L = \emptyset$$

or when they coincide, $K = L$. Otherwise, we say that K, L cross, and write $K\not|\not|L$.

Here we have tricked a bit, by agreeing to call parallel the pairs of identical lines too, and this for simplifying most of our mathematics, in what follows, trust me here.

As a first remark, with this definition in hand, Observation 5.6 makes now sense, as a formal mathematical statement, and skipping some discussion here, or rather leaving it as an exercise, for reasons which are somewhat clear, we will call this axiom:

Axiom 5.8. *Any two crossing lines $K\not|\not|L$ determine a point, $P = K \cap L$.*

Very good, and now with Axiom 5.1 and Axiom 5.8 in hand, we are potentially ready for doing some geometry. However, this is not exactly true, and we will need as well:

Axiom 5.9. *Given a point not lying on a line, $P \notin L$, we can draw through P a unique parallel to L. That is, we can find a line K satisfying $P \in K$, $K||L$.*

As before, we will leave as an exercise further meditating on all this.

Ready for some math? Here we go, and many things can be said here, especially about parallel lines, which are the main objects of basic geometry, as for instance:

Theorem 5.10 (Thales). *Proportions are kept, along parallel lines.*

Proof. This is indeed something very standard. □

Importantly, many things can be done with the parallel lines, with a suitably drawn such line hopefully solving, by some kind of miracle, your plane geometry problem.

We will see more illustrations for this general principle in the next section.

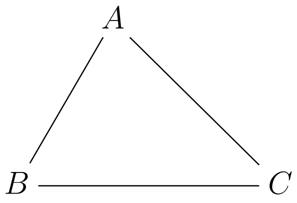
5b. Angles, triangles

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 5.11. *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:

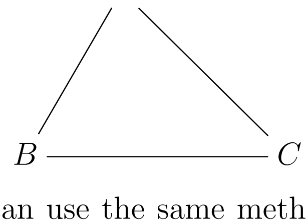
(1) Come with a small circle, inside ABC, and then inflate it, as to touch all 3 edges. The center of the circle will be then at equal distance from all 3 edges, so it will lie on all 3 angle bisectors. Thus, we have constructed the incenter, as required.

(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. □

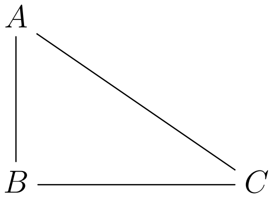
Many other things can be said about triangles, and we will be back to this. Importantly, we can now talk about angles, in the obvious way, by using triangles:

Fact 5.12. *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.

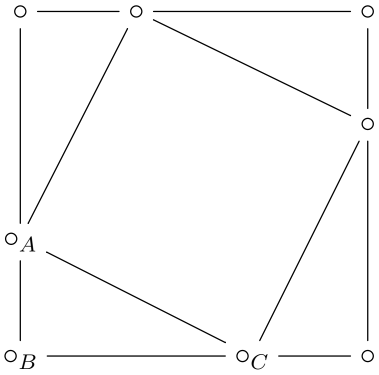
Getting started now with our study of angles, as a continuation of Fact 5.12, let us first talk about the simplest angle of them all, which is the right angle, denoted 90°. Many interesting things can be said about this right angle 90°, in particular with:

Theorem 5.13 (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. □

We will be back to Pythagoras in chapter 6, when doing trigonometry.

5c. Advanced results

Back now to the triangles, where everything comes from, along the same lines as Theorem 5.11, but at a more advanced level, we have the following result:

Fact 5.14. *Besides the above 4 centers, many more remarkable points can be associated to a triangle ABC, and most of these lie on a line, called Euler line of ABC.*

And exercise for you of course to remember or figure out how all this works, both statement and proof. As bonus exercise, learn about the nine-point circle too.

5d. Projective geometry

Switching topics, but still in relation with the parallel lines, that we constantly met in the above, you might have heard or not of projective geometry. In case you didn't yet, the general principle is that "this is the wonderland where parallel lines cross".

Which might sound a bit crazy, and not very realistic, but take a picture of some railroad tracks, and look at that picture. Do that parallel railroad tracks cross, on the picture? Sure they do. So, we are certainly not into abstractions here. QED.

Mathematically now, here are some axioms, to start with:

Definition 5.15. *A projective space is a space consisting of points and lines, subject to the following conditions:*

(1) *Each 2 points determine a line.*
(2) *Each 2 lines cross, on a point.*

As a basic example we have the usual projective plane $P^2_\mathbb{R}$, which is best seen as being the space of lines in $\mathbb{R}^3$ passing through the origin. To be more precise, let us call each of these lines in $\mathbb{R}^3$ passing through the origin a "point" of $P^2_\mathbb{R}$, and let us also call each plane in $\mathbb{R}^3$ passing through the origin a "line" of $P^2_\mathbb{R}$. Now observe the following:

(1) Each 2 points determine a line. Indeed, 2 points in our sense means 2 lines in $\mathbb{R}^3$ passing through the origin, and these 2 lines obviously determine a plane in $\mathbb{R}^3$ passing through the origin, namely the plane they belong to, which is a line in our sense.

(2) Each 2 lines cross, on a point. Indeed, 2 lines in our sense means 2 planes in $\mathbb{R}^3$ passing through the origin, and these 2 planes obviously determine a line in $\mathbb{R}^3$ passing through the origin, namely their intersection, which is a point in our sense.

Thus, what we have is a projective space in the sense of Definition 5.15. More generally now, we have the following construction, in arbitrary dimensions:

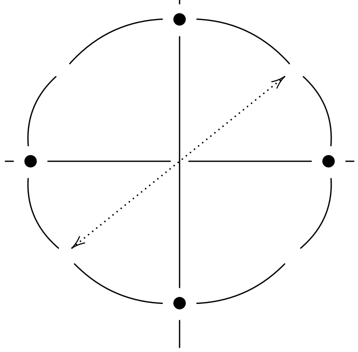
Theorem 5.16. *We can define the projective space $P^{N-1}_\mathbb{R}$ as being the space of lines in $\mathbb{R}^N$ passing through the origin, and in small dimensions:*

1. *$P^1_\mathbb{R}$ is the usual circle.*
2. *$P^2_\mathbb{R}$ is some sort of twisted sphere.*

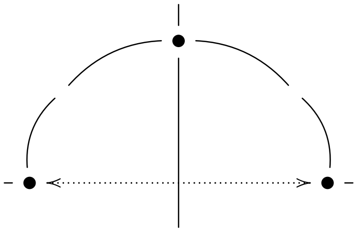
Proof. We have several assertions here, with all this being of course a bit informal, and self-explanatory, the idea and some further details being as follows:

(1) To start with, the fact that the space $P^{N-1}_\mathbb{R}$ constructed in the statement is indeed a projective space in the sense of Definition 5.15 follows from definitions, exactly as in the discussion preceding the statement, regarding the case $N = 3$.

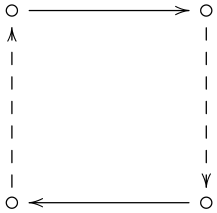
(2) At $N = 2$ now, a line in $\mathbb{R}^2$ passing through the origin corresponds to 2 opposite points on the unit circle $\mathbb{T} \subset \mathbb{R}^2$, according to the following scheme:



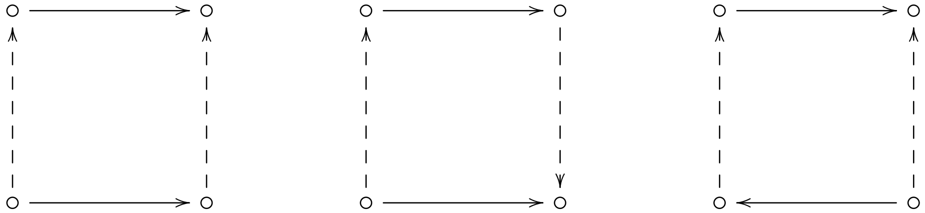
Thus, $P^1_\mathbb{R}$ corresponds to the upper semicircle of $\mathbb{T}$, with the endpoints identified, and so we obtain a circle, $P^1_\mathbb{R} = \mathbb{T}$, according to the following scheme:



(3) At $N = 3$, the space $P^2_\mathbb{R}$ corresponds to the upper hemisphere of the sphere $S^2_\mathbb{R} \subset \mathbb{R}^3$, with the points on the equator identified via $x = -x$. Topologically speaking, we can deform if we want the hemisphere into a square, with the equator becoming the boundary of this square, and in this picture, the $x = -x$ identification corresponds to a "identify opposite edges, with opposite orientations" folding method for the square:



(4) Thus, we have our space. In order to understand now what this beast is, let us look first at the other 3 possible methods of folding the square, which are as follows:



Regarding the first space, the one on the left, things here are quite simple. Indeed, when identifying the solid edges we get a cylinder, and then when further identifying the dotted edges, what we get is some sort of closed cylinder, which is a torus.

(5) Regarding the second space, the one in the middle, things here are more tricky. Indeed, when identifying the solid edges we get again a cylinder, but then when further identifying the dotted edges, we obtain some sort of "impossible" closed cylinder, called Klein bottle. This Klein bottle obviously cannot be drawn in 3 dimensions, but with a bit of imagination, you can see it, in its full splendor, in 4 dimensions.

(6) Finally, regarding the third space, the one on the right, we know by symmetry that this must be the Klein bottle too. But we can see this as well via our standard folding method, namely identifying solid edges first, and dotted edges afterwards. Indeed, we first obtain in this way a Möbius strip, and then, well, the Klein bottle.

(7) With these preliminaries made, and getting back now to the projective space $P^2_{\mathbb{R}}$, we can see that this is something more complicated, of the same type, reminding the torus and the Klein bottle. So, we will call it "sort of twisted sphere", as in the statement, and exercise for you to figure out how this beast looks like, in 4 dimensions. □

All this is quite exciting, and reminds childhood and primary school, but is however a bit tiring for our neurons, guess that is pure mathematics. It is possible to come up with some explicit formulae for the embedding $P^2_{\mathbb{R}} \subset \mathbb{R}^4$, which are useful in practice, allowing us to do some analysis over $P^2_{\mathbb{R}}$, and we will leave this as an instructive exercise.

All this is very interesting, but we will pause our study here, because we still have many other things to say. Getting now to finite fields, we have:

Theorem 5.17. *Given a field F, we can talk about the projective space P^{N-1}_F, as being the space of lines in F^N passing through the origin. At $N = 3$ we have*

$$|P^2_F| = q^2 + q + 1$$

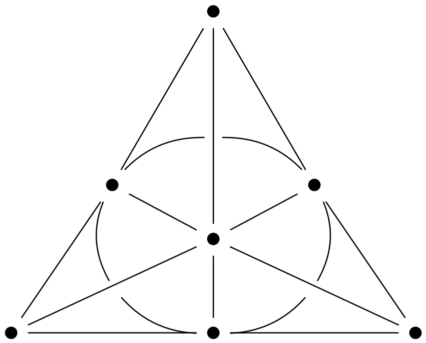
where $q = |F|$, in the case where our field F is finite.

Proof. This is indeed clear from definitions, with the cardinality coming from:

$$|P^2_F| = \frac{|F^3 - \{0\}|}{|F - \{0\}|} = \frac{q^3 - 1}{q - 1} = q^2 + q + 1$$

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

As an example, let us see what happens for the simplest finite field that we know, namely $F = \mathbb{Z}_2$. Here our projective plane, having $4 + 2 + 1 = 7$ points, and 7 lines, is a famous combinatorial object, called Fano plane, which is depicted as follows:



Here the circle in the middle is by definition a line, and with this convention, the basic axioms in Definition 5.15 are satisfied, in the sense that any two points determine a line, and any two lines determine a point. And isn't this beautiful.

5e. Exercises

Exercises:

Exercise 5.18.

Exercise 5.19.

Exercise 5.20.

Exercise 5.21.

Exercise 5.22.

Exercise 5.23.

Exercise 5.24.

Exercise 5.25.

Bonus exercise.

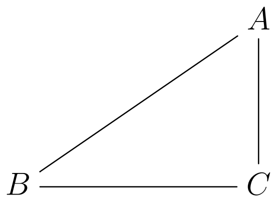
CHAPTER 6

Trigonometry

6a. Sine, cosine

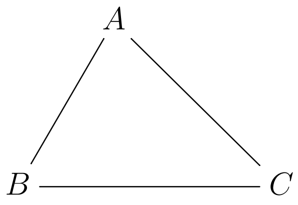
Now that we know about angles, and about Pythagoras' theorem too, it is tempting at this point to start talking about trigonometry. Let us begin with:

Definition 6.1. *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:



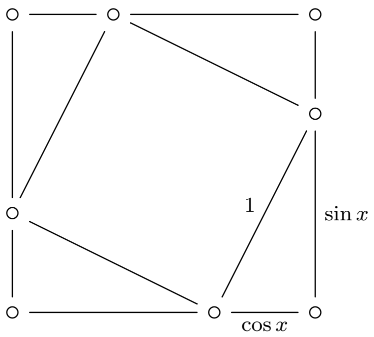
Getting now to more advanced theory, we first have:

Theorem 6.2. *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 6.3. *The sines and cosines of sums are given by*

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

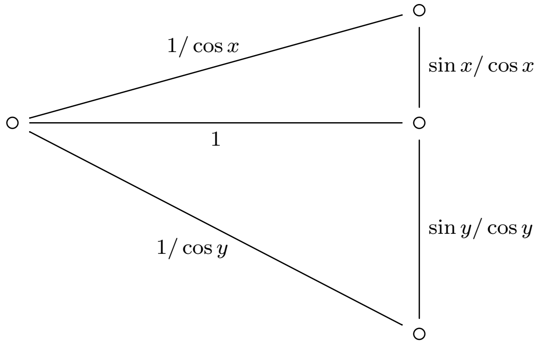
$$\cos(x + y) = \cos x \cos y - \sin x \sin y$$

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 6.3. 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.

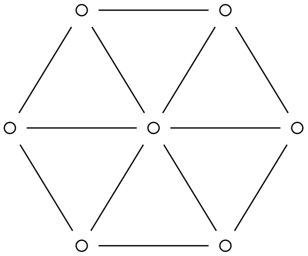
6b. The number pi

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 6.4. *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:

$$\pi = 3.14159\ldots$$

We will come back to such questions later, towards the end of the present book, once we will have appropriate tools for dealing with them.

It is also possible to prove that π is irrational, $\pi \notin \mathbb{Q}$, but this is not trivial either, with the idea being that of suitably adapting the proof for e:

THEOREM 6.5. *The number π is irrational.*

PROOF. This is indeed something quite routine, by using the same ideas as before for e, but with everything being now a bit more technical. □

Finally, again in analogy with our previous theory of e, we have the following well-known, and beautiful, probabilistic interpretation of π, due to Buffon:

Theorem 6.6. *The probability for a needle of length* 1*, when trown on a grid of parallel* 1*-spaced lines, to intersect one line, is* $P = 2/\pi$*.*

Proof. This is something quite tricky, and mandatory for learning well probability, because there are several possible modelings of the problem, leading, quite surprisingly, to different values of P. And, while a pure mathematician might find this a bit odd, and unfair, throwing a needle as in the statement is more than possible, in the real life, yielding to one such P, the correct one, and so with all mathematics leading to other P, be that very smart and formally correct, being therefore garbage. Welcome to probability. □

6c. Trigonometry

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

Theorem 6.7. *The following happen:*

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

Proof. The formula $L = 2\pi$ from Theorem 6.4 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 $\sin^2 x + \cos^2 x = 1$, 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 6.8. *The following happen, for small angles,* $x \simeq 0$*:*

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

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

6d. More trigonometry

More trigonometry. We will be back to this in chapter 7, with better tools.

6e. Exercises

Exercises:

Exercise 6.9.

Exercise 6.10.

Exercise 6.11.

Exercise 6.12.

Exercise 6.13.

Exercise 6.14.

Exercise 6.15.

Exercise 6.16.

Bonus exercise.

CHAPTER 7

Coordinates

7a. Real plane

Real plane.

7b. Complex plane

Let us discuss now the complex numbers. There is a lot of magic here, and we will carefully explain this material. Their definition is as follows:

Definition 7.1. *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 7.2. *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 7.3. *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:

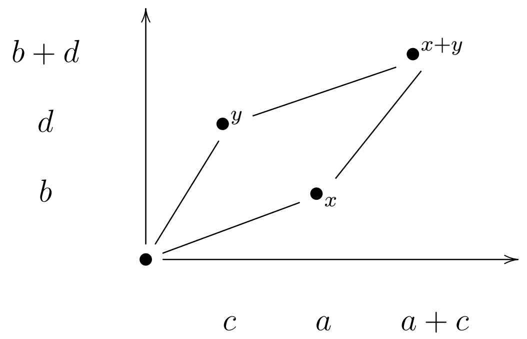
$$x=\begin{pmatrix}a\\ b\end{pmatrix}\quad,\quad y=\begin{pmatrix}c\\ d\end{pmatrix}$$

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. □

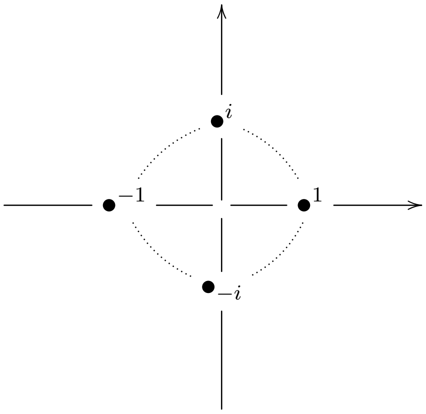
Here we have assumed that you are a bit familiar with vector calculus. If not, no problem, the idea is simply that vectors add by forming a parallelogram, as follows:



Observe that in our geometric picture from Proposition 7.3, the real numbers correspond to the numbers on the Ox axis. As for the purely imaginary numbers, these lie on the Oy axis, with the number i itself being given by the following formula:

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

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



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

ADVICE 7.4. *When drawing complex numbers, always begin with the coordinate axes Ox, Oy, and with a copy of the unit circle.*

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

DEFINITION 7.5. *The complex numbers $x = a + ib$ can be written in polar coordinates,*

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

with the connecting formulae being as follows,

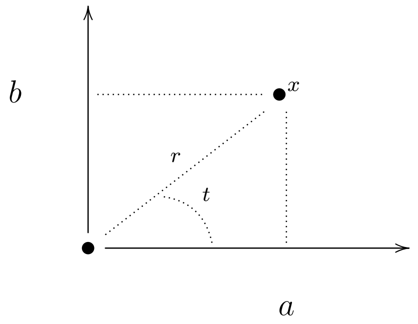
$$a = r \cos t \quad , \quad b = r \sin t$$

and in the other sense being as follows,

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

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

There is a clear relation here with the vector notation from Proposition 7.3, because r is the length of the vector, and t is the angle made by the vector with the Ox axis. To be more precise, the picture for what is going on in Definition 7.5 is as follows:



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

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

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

Theorem 7.6. *Two complex numbers written in polar coordinates,*

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

multiply according to the following formula:

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

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

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

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

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

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

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

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

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

Theorem 7.7. *The powers of a complex number, written in polar form,*

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

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

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

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

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

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

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 7.7, we have the following result:

Proposition 7.8. *Each complex number, written in polar form,*

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

has two square roots, given by the following formula:

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

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

Proof. The first assertion is clear indeed from the general formula in Theorem 7.7, at $k = 1/2$. As for its particular cases with $x \in \mathbb{R}$, these are clear from it. □

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

With the above results in hand, and notably with the square root formula from Proposition 7.8, we can now go back to the degree 2 equations, and we have:

Theorem 7.9. *The complex solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{C}$ are*

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

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

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

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

Now since we know from Proposition 7.8 that any complex number has a square root, we are led to the conclusion in the statement. □

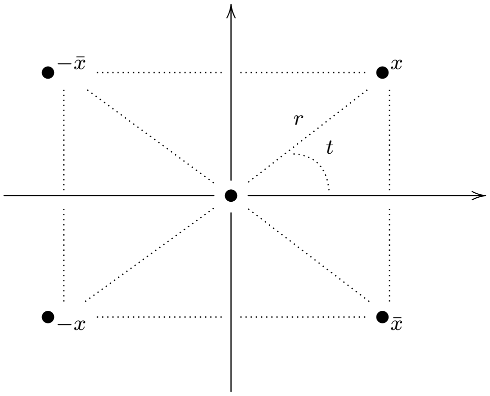
As a last general topic regarding the complex numbers, let us discuss conjugation. This is something quite tricky, complex number specific, as follows:

Definition 7.10. *The complex conjugate of $x = a + ib$ is the following number,*

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

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

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



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

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

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

Theorem 7.11. *The conjugation operation $x \to \bar{x}$ has the following properties:*

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

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

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

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

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

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

(4) This is clear indeed from the picture following Definition 7.10.

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

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

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

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

(6) This comes from the formula of the solutions, that we know from Theorem 7.2, 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. □

7c. Analysis tricks

Let us discuss now the final and most convenient writing of the complex numbers, $x = re^{it}$. The point with this formula comes from the following deep result:

Theorem 7.12. *We have the following formula,*

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

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

Proof. Our claim is that this follows from the formula of the complex exponential, and for the following formulae for the Taylor series of cos and sin, that we know well:

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

Indeed, let us first recall that we have the following formula, for the exponential of an arbitrary real number $x\in\mathbb R$, but which works in fact for any $x\in\mathbb C$:

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

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

$$\begin{aligned}e^{it}&=\sum_{k=0}^\infty\frac{(it)^k}{k!}\\&=\sum_{k=2l}\frac{(it)^k}{k!}+\sum_{k=2l+1}\frac{(it)^k}{k!}\\&=\sum_{l=0}^\infty(-1)^l\frac{t^{2l}}{(2l)!}+i\sum_{l=0}^\infty(-1)^l\frac{t^{2l+1}}{(2l+1)!}\\&=\cos t+i\sin t\end{aligned}$$

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

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

Theorem 7.13. *We have the following formula,*

$$e^{\pi i}=-1$$

and we have $E=mc^2$ as well.

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

(1) The first formula, $e^{\pi i}=-1$, which is actually the main formula in mathematics, comes from Theorem 7.12, by setting $t=\pi$. Indeed, we obtain:

$$\begin{aligned}e^{\pi i}&=\cos\pi+i\sin\pi\\&=-1+i\cdot 0\\&=-1\end{aligned}$$

(2) As for $E=mc^2$, which is the main formula in physics, this is something deep too. Although we will not really need it here, we recommend learning it as well, for symmetry reasons between math and physics, say from Feynman [33], [34], [35]. □

Now back to our $x = re^{it}$ objectives, with the above theory in hand we can indeed use from now on this notation, the complete statement being as follows:

THEOREM 7.14. *The complex numbers $x = a + ib$ can be written in polar coordinates,*

$$x = re^{it}$$

with the connecting formulae being

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

and in the other sense being

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

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

PROOF. This is a reformulation of our previous Definition 7.5, by using the formula $e^{it} = \cos t + i\sin t$ from Theorem 7.13, and multiplying everything by r. □

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

THEOREM 7.15. *In polar coordinates, the complex numbers multiply as*

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

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

PROOF. This is something that we already know, from Theorem 7.6, reformulated by using the notations from Theorem 7.14. Observe that this follows as well directly, from the fact that we have $e^{a+b} = e^a e^b$, that we know from analysis. □

The above formula is obviously very powerful. However, in polar coordinates we do not have a simple formula for the sum. Thus, this formalism has its limitations.

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

THEOREM 7.16. *We have the following operations on the complex numbers, written in polar form, as above:*

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

PROOF. This is something that we already know, from Theorem 7.7, but we can now discuss all this, from a more conceptual viewpoint, the idea being as follows:

(1) We have indeed the following computation, using Theorem 7.14:

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

(2) Once again by using Theorem 7.14, we have:

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

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

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

Due to Theorem 7.14, this operation $x\to x^a$ is indeed the correct one.

(4) This comes from the fact, that we know from Theorem 7.11, that the conjugation operation $x\to\bar{x}$ keeps the modulus, and switches the sign of the argument. □

Getting back to algebra, we know from Theorem 7.9 that any degree 2 equation has 2 complex roots. We can in fact prove that any polynomial equation, of arbitrary degree $N\in\mathbb{N}$, has exactly N complex solutions, counted with multiplicities:

Theorem 7.17. *Any polynomial $P\in\mathbb{C}[X]$ decomposes as*

$$P=c(X-a_1)\ldots(X-a_N)$$

with $c\in\mathbb{C}$ and with $a_1,\ldots,a_N\in\mathbb{C}$.

Proof. The problem is that of proving that our polynomial has at least one root, because afterwards we can proceed by recurrence. We prove this by contradiction. So, assume that P has no roots, and pick a number $z\in\mathbb{C}$ where $|P|$ attains its minimum:

$$|P(z)|=\min_{x\in\mathbb{C}}|P(x)|>0$$

Since $Q(t)=P(z+t)-P(z)$ is a polynomial which vanishes at $t=0$, this polynomial must be of the form ct^k + higher terms, with $c\neq 0$, and with $k\geq 1$ being an integer. We obtain from this that, with $t\in\mathbb{C}$ small, we have the following estimate:

$$P(z+t)\simeq P(z)+ct^k$$

Now let us write $t=rw$, with $r>0$ small, and with $|w|=1$. Our estimate becomes:

$$P(z+rw)\simeq P(z)+cr^kw^k$$

Now recall that we assumed $P(z)\neq 0$. We can therefore choose $w\in\mathbb{T}$ such that cw^k points in the opposite direction to that of $P(z)$, and we obtain in this way:

$$\begin{aligned}|P(z+rw)| &\simeq |P(z)+cr^kw^k|\\ &= |P(z)|(1-|c|r^k)\end{aligned}$$

Now by choosing $r > 0$ small enough, as for the error in the first estimate to be small, and overcame by the negative quantity $-|c|r^k$, we obtain from this:

$$|P(z + rw)| < |P(z)|$$

But this contradicts our definition of $z \in \mathbb{C}$, as a point where $|P|$ attains its minimum. Thus P has a root, and by recurrence it has N roots, as stated. $\square$

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

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

which are called roots of unity of order N.

Proof. This follows from the general multiplication formula for complex numbers from Theorem 7.15. Indeed, with $x = re^{it}$ our equation reads:

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

Thus $r = 1$, and $t \in [0, 2\pi)$ must be a multiple of $2\pi/N$, as stated. $\square$

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 7.7 and Theorem 7.16, the statement being as follows:

Theorem 7.19. *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 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. $\square$

7d. Ellipses, conics

Looking up, to the sky, the first thing that you see is the Sun, seemingly moving around the Earth on a circle, but a more careful study reveals that this circle is rather a deformed circle, called ellipsis. And good news, a full theory of ellipses is available, and this since the ancient Greeks, whose main findings were as follows:

Theorem 7.20. *The ellipses, taken centered at the origin* 0, *and squarely oriented with respect to* Oxy, *can be defined in* 4 *possible ways, as follows:*

(1) *As the curves given by an equation as follows, with* $a,b>0$*:*

$$\left(\frac{x}{a}\right)^2+\left(\frac{y}{b}\right)^2=1$$

(2) *Or given by an equation as follows, with* $q>0$, $p=-q$, *and* $l\in(0,2q)$*:*

$$d(z,p)+d(z,q)=l$$

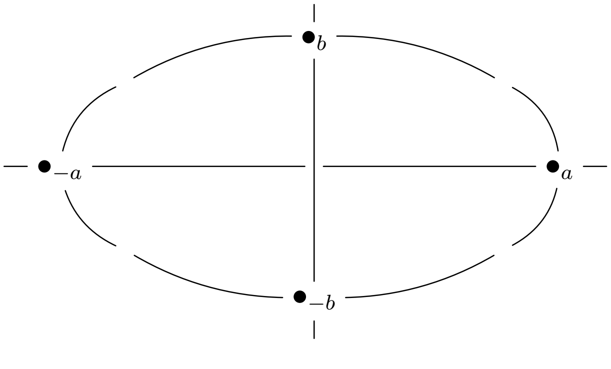
(3) *As the curves appearing when drawing a circle, from various perspectives:*

$$\bigcirc\quad\to\quad ?$$

(4) *As the closed non-degenerate curves appearing by cutting a cone with a plane.*

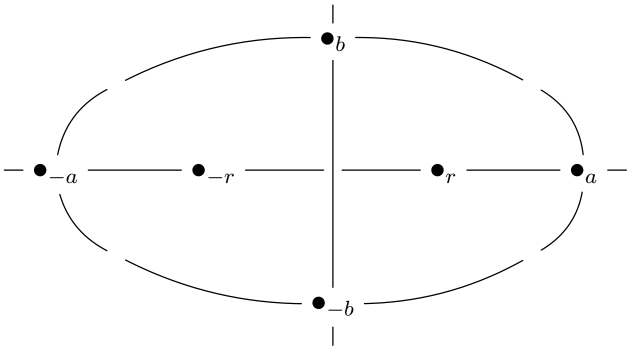
Proof. This might look a bit confusing, and you might say, what exactly is to be proved here. Good point, and in answer, what is to be proved is that the above constructions (1-4) give rise to the same class of curves. And this can be done as follows:

(1) To start with, let us draw a picture from what comes out of (1), which will be our main definition for the ellipses, in what follows. Here that is, making it clear what the parameters $a,b>0$ stand for, with $2a\times 2b$ being the gift box size for our ellipsis:



(2) Let us prove now that such an ellipsis has two focal points, as stated in (2). We must look for a number $r>0$, and a number $l>0$, such that our ellipsis appears as

$d(z,p)+d(z,q)=l$, with $p=(0,-r)$ and $q=(0,r)$, according to the following picture:



(3) Let us first compute these numbers $r,l>0$. Assuming that our result holds indeed as stated, by taking $z=(0,a)$, we see that the length l is:

$$l=(a-r)+(a+r)=2a$$

As for the parameter r, by taking $z=(b,0)$, we conclude that we must have:

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

(4) With these observations made, let us prove now the result. Given $l,r>0$, and setting $p=(0,-r)$ and $q=(0,r)$, we have the following computation, with $z=(x,y)$:

$$\begin{aligned}
& d(z,p)+d(z,q)=l\\
\iff\ & \sqrt{(x+r)^2+y^2}+\sqrt{(x-r)^2+y^2}=l\\
\iff\ & \sqrt{(x+r)^2+y^2}=l-\sqrt{(x-r)^2+y^2}\\
\iff\ & (x+r)^2+y^2=(x-r)^2+y^2+l^2-2l\sqrt{(x-r)^2+y^2}\\
\iff\ & 2l\sqrt{(x-r)^2+y^2}=l^2-4xr\\
\iff\ & 4l^2(x^2+r^2-2xr+y^2)=l^4+16x^2r^2-8l^2xr\\
\iff\ & 4l^2x^2+4l^2r^2+4l^2y^2=l^4+16x^2r^2\\
\iff\ & (4x^2-l^2)(4r^2-l^2)=4l^2y^2
\end{aligned}$$

(5) Now observe that we can further process the equation that we found as follows:

$$\begin{aligned}(4x^2-l^2)(4r^2-l^2)=4l^2y^2 \quad &\iff \quad \frac{4x^2-l^2}{l^2}=\frac{4y^2}{4r^2-l^2}\\ &\iff \quad \frac{4x^2-l^2}{l^2}=\frac{y^2}{r^2-l^2/4}\\ &\iff \quad \left(\frac{x}{2l}\right)^2-1=\left(\frac{y}{\sqrt{r^2-l^2/4}}\right)^2\\ &\iff \quad \left(\frac{x}{2l}\right)^2+\left(\frac{y}{\sqrt{r^2-l^2/4}}\right)^2=1\end{aligned}$$

(6) Thus, our result holds indeed, and with the numbers $l,r>0$ appearing, and no surprise here, via the formulae $l=2a$ and $r=\sqrt{a^2-b^2}$, found in (3) above.

(7) Getting back now to our theorem, we have two other assertions there at the end, labelled (3,4). But, thinking a bit, these assertions are in fact equivalent, and in what concerns us, we will rather focus on (4), which looks more mathematical. And in what regards this assertion (4), this can be established indeed, by doing some 3D computations, that we will leave here as an instructive exercise, for you. And with the promise that we will come back to this in a moment, with a full proof, in a more general setting. □

All this is very nice, but let us settle now as well the question of wandering asteroids. Observations show that these can travel on parabolas and hyperbolas, so what we need as mathematics is a unified theory of ellipses, parabolas and hyperbolas. And fortunately, this theory exists, also since the ancient Greeks, summarized as follows:

Theorem 7.21. *The conics, which are the algebraic curves of degree 2 in the plane,*

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

with $\deg P\leq 2$*, appear modulo degeneration by cutting a 2-sided cone with a plane, and can be classified into ellipses, parabolas and hyperbolas.*

Proof. This follows by further building on Theorem 7.20, as follows:

(1) Let us first classify the conics up to non-degenerate linear transformations of the plane, which are by definition transformations as follows, with $\det A\neq 0$:

$$\begin{pmatrix}x\\y\end{pmatrix}\to A\begin{pmatrix}x\\y\end{pmatrix}$$

Our claim is that as solutions we have the circles, parabolas, hyperbolas, along with some degenerate solutions, namely $\emptyset$, points, lines, pairs of lines, $\mathbb{R}^2$.

(2) As a first remark, it looks like we forgot precisely the ellipses, but via linear transformations these become circles, so things fine. As a second remark, all our claimed solutions can appear. Indeed, the circles, parabolas, hyperbolas can appear as follows:

$$x^2+y^2=1 \quad , \quad x^2=y \quad , \quad xy=1$$

As for $\emptyset$, points, lines, pairs of lines, $\mathbb{R}^2$, these can appear too, as follows, and with our polynomial P chosen, whenever possible, to be of degree exactly 2:

$$x^2=-1 \quad , \quad x^2+y^2=0 \quad , \quad x^2=0 \quad , \quad xy=0 \quad , \quad 0=0$$

Observe here that, when dealing with these degenerate cases, assuming $\deg P=2$ instead of $\deg P\leq 2$ would only rule out $\mathbb{R}^2$ itself, which is not worth it.

(3) Getting now to the proof of our claim in (1), classification up to linear transformations, consider an arbitrary conic, written as follows, with $a,b,c,d,e,f\in\mathbb{R}$:

$$ax^2+by^2+cxy+dx+ey+f=0$$

Assume first $a\neq 0$. By making a square out of ax^2, up to a linear transformation in (x,y), we can get rid of the term cxy, and we are left with:

$$ax^2+by^2+dx+ey+f=0$$

In the case $b\neq 0$ we can make two obvious squares, and again up to a linear transformation in (x,y), we are left with an equation as follows:

$$x^2\pm y^2=k$$

In the case of positive sign, $x^2+y^2=k$, the solutions are the circle, when $k\geq 0$, the point, when $k=0$, and $\emptyset$, when $k<0$. As for the case of negative sign, $x^2-y^2=k$, which reads $(x-y)(x+y)=k$, here once again by linearity our equation becomes $xy=l$, which is a hyperbola when $l\neq 0$, and two lines when $l=0$.

(4) In the case $b\neq 0$ the study is similar, with the same solutions, so we are left with the case $a=b=0$. Here our conic is as follows, with $c,d,e,f\in\mathbb{R}$:

$$cxy+dx+ey+f=0$$

If $c\neq 0$, by linearity our equation becomes $xy=l$, which produces a hyperbola or two lines, as explained before. As for the remaining case, $c=0$, here our equation is:

$$dx+ey+f=0$$

But this is generically the equation of a line, unless we are in the case $d=e=0$, where our equation is $f=0$, having as solutions $\emptyset$ when $f\neq 0$, and $\mathbb{R}^2$ when $f=0$.

(5) Thus, done with the classification, up to linear transformations as in (1). But this classification leads to the classification in general too, by applying now linear transformations to the solutions that we found. So, done with this, and very good.

(6) It remains to discuss the cone cutting. By suitably choosing our coordinate axes (x,y,z), we can assume that our cone is given by an equation as follows, with $k>0$:

$$x^2+y^2=kz^2$$

In order to prove the result, we must in principle intersect this cone with an arbitrary plane, which has an equation as follows, with $(a,b,c)\neq(0,0,0)$:

$$ax+by+cz=d$$

(7) However, before getting into computations, observe that what we want to find is a certain degree 2 equation in the above plane, for the intersection. Thus, it is convenient to change the coordinates, as for our plane to be given by the following equation:

$$z=0$$

(8) But with this done, what we have to do is to see how the cone equation $x^2+y^2=kz^2$ changes, under this change of coordinates, and then set $z=0$, as to get the (x,y) equation of the intersection. But this leads, via some thinking or computations, to the conclusion that the cone equation $x^2+y^2=kz^2$ becomes in this way a degree 2 equation in (x,y), which can be arbitrary, and so to the final conclusion in the statement. □

Ready for some physics? We have the following result:

Theorem 7.22. *Planets and other celestial bodies move around the Sun on conics,*

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

with $P\in\mathbb{R}[x,y]$ being of degree 2, which can be ellipses, parabolas or hyperbolas.

Proof. This is something quite long, due to Kepler and Newton. □

7e. Exercises

Exercises:

Exercise 7.23.

Exercise 7.24.

Exercise 7.25.

Exercise 7.26.

Exercise 7.27.

Exercise 7.28.

Exercise 7.29.

Exercise 7.30.

Bonus exercise.

CHAPTER 8

Space geometry

8a. Space geometry

Space geometry.

8b. Solid angles

Solid angles.

8c. Field lines

Field lines.

8d. Rotating bodies

Getting started with some applications, here is the notion what we will need:

Definition 8.1. *The vector product of two vectors in $\mathbb{R}^3$ is given by*

$$x \times y = ||x|| \cdot ||y|| \cdot \sin\theta \cdot n$$

where $n \in \mathbb{R}^3$ with $n \perp x, y$ and $||n|| = 1$ is constructed using the right-hand rule:

$$\begin{array}{l} \quad\uparrow_{x \times y} \\ \leftarrow_x \\ \swarrow_y \end{array}$$

Alternatively, in usual vertical linear algebra notation for all vectors,

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} x_2y_3 - x_3y_2 \\ x_3y_1 - x_1y_3 \\ x_1y_2 - x_2y_1 \end{pmatrix}$$

the rule being that of computing 2×2 determinants, and adding a middle sign.

Obviously, this definition is something quite subtle, and also something very annoying, because you always need this, and always forget the formula. Here are my personal methods. With the first definition, what I always remember is that:

$$||x \times y|| \sim ||x||, ||y|| \quad , \quad x \times x = 0 \quad , \quad e_1 \times e_2 = e_3$$

So, here's how it works. We are looking for a vector $x \times y$ whose length is proportional to those of x, y. But the second formula tells us that the angle θ between x, y must be involved via $0 \to 0$, and so the factor can only be $\sin\theta$. And with this we are almost there, it's just a matter of choosing the orientation, and this comes from $e_1 \times e_2 = e_3$.

As with the second definition, that I like the most, what I remember here is simply:

$$\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} = ?$$

In practice now, in order to get familiar with the vector products, nothing better than doing some classical mechanics. We have here the following key result:

Theorem 8.2. *In the gravitational* 2*-body problem, the angular momentum*

$$J = x \times p$$

with $p = mv$ being the usual momentum, is conserved.

Proof. There are several things to be said here, the idea being as follows:

(1) First of all the usual momentum, $p = mv$, is not conserved, because the simplest solution is the circular motion, where the moment gets turned around. But this suggests precisely that, in order to fix the lack of conservation of the momentum p, what we have to do is to make a vector product with the position x. Leading to J, as above.

(2) Regarding now the proof, consider indeed a particle m moving under the gravitational force of a particle M, assumed, as usual, to be fixed at 0. By using the fact that for two proportional vectors, $p \sim q$, we have $p \times q = 0$, we obtain:

$$\begin{aligned} \dot{J} &= \dot{x} \times p + x \times \dot{p} \\ &= v \times mv + x \times ma \\ &= m(v \times v + x \times a) \\ &= m(0 + 0) \\ &= 0 \end{aligned}$$

Now since the derivative of J vanishes, this quantity is constant, as stated. □

As another basic application of the vector products, still staying with classical mechanics, we have all sorts of useful formulae regarding rotating frames. We first have:

Theorem 8.3. *Assume that a* 3*D body rotates along an axis, with angular speed w. For a fixed point of the body, with position vector x, the usual* 3*D speed is*

$$v = \omega \times x$$

where $\omega = wn$, with n unit vector pointing North. When the point moves on the body

$$V = \dot{x} + \omega \times x$$

is its speed computed by an inertial observer O on the rotation axis.

Proof. We have two assertions here, both requiring some 3D thinking, as follows:

(1) Assuming that the point is fixed, the magnitude of $\omega \times x$ is the good one, due to the following computation, with r being the distance from the point to the axis:

$$||\omega \times x|| = w||x|| \sin t = wr = ||v||$$

As for the orientation of $\omega \times x$, this is the good one as well, because the North pole rule used above amounts in applying the right-hand rule for finding n, and so ω, and this right-hand rule was precisely the one used in defining the vector products $\times$.

(2) Next, when the point moves on the body, the inertial observer O can compute its speed by using a frame (u_1, u_2, u_3) which rotates with the body, as follows:

$$\begin{aligned} V &= \dot{x}_1 u_1 + \dot{x}_2 u_2 + \dot{x}_3 u_3 + x_1 \dot{u}_1 + x_2 \dot{u}_2 + x_3 \dot{u}_3 \\ &= \dot{x} + (x_1 \cdot \omega \times u_1 + x_2 \cdot \omega \times u_2 + x_3 \cdot \omega \times u_3) \\ &= \dot{x} + w \times (x_1 u_1 + x_2 u_2 + x_3 u_3) \\ &= \dot{x} + \omega \times x \end{aligned}$$

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

In what regards now the acceleration, the result, which is famous, is as follows:

Theorem 8.4. *Assuming as before that a 3D body rotates along an axis, the acceleration of a moving point on the body, computed by O as before, is given by*

$$A = a + 2\omega \times v + \omega \times (\omega \times x)$$

with $\omega = wn$ being as before. In this formula the second term is called Coriolis acceleration, and the third term is called centripetal acceleration.

Proof. This comes by using twice the formulae in Theorem 8.3, as follows:

$$\begin{aligned} A &= \dot{V} + \omega \times V \\ &= (\ddot{x} + \dot{\omega} \times x + \omega \times \dot{x}) + (\omega \times \dot{x} + \omega \times (\omega \times x)) \\ &= \ddot{x} + \omega \times \dot{x} + \omega \times \dot{x} + \omega \times (\omega \times x) \\ &= a + 2\omega \times v + \omega \times (\omega \times x) \end{aligned}$$

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

The truly famous result is actually the one regarding forces, obtained by multiplying everything by a mass m, and writing things the other way around, as follows:

$$ma = mA - 2m\omega \times v - m\omega \times (\omega \times x)$$

Here the second term is called Coriolis force, and the third term is called centrifugal force. These forces are both called apparent, or fictious, because they do not exist in the inertial frame, but they exist however in the non-inertial frame of reference, as explained above. And with of course the terms centrifugal and centripetal not to be messed up.

In fact, even more famous is the terrestrial application of all this, as follows:

THEOREM 8.5. *The acceleration of an object m subject to a force F is given by*

$$ma = F - mg - 2m\omega \times v - m\omega \times (\omega \times x)$$

with g pointing upwards, and with the last terms being the Coriolis and centrifugal forces.

PROOF. This follows indeed from the above discussion, by assuming that the acceleration A there comes from the combined effect of a force F, and of the usual g. □

We refer to any standard undergraduate mechanics book, such as Feynman [**33**], Kibble [**57**] or Taylor [**91**] for more on the above, including various numerics on what happens here on Earth, the Foucault pendulum, history of all this, and many other things. Let us just mention here, as a basic illustration for all this, that a rock dropped from 100m deviates about 1cm from its intended target, due to the formula in Theorem 8.5.

8e. Exercises

Exercises:

EXERCISE 8.6.

EXERCISE 8.7.

EXERCISE 8.8.

EXERCISE 8.9.

EXERCISE 8.10.

EXERCISE 8.11.

EXERCISE 8.12.

EXERCISE 8.13.

Bonus exercise.

Part III

Arithmetic

When it's summer in Siam
And the moon is full of rainbows
When it's summer in Siam
And we go through many changes

CHAPTER 9

Divisibility

9a. Divisibility

Divisibility basics.

9b. Factorization

Time now to get into prime numbers, which will be a main theme of discussion, for this Part II. How many primes do you know? The more the better, and those under 100 are mandatory, at the beginner level, here they are, in all their beauty:

$$2,3,5,7,11,13,17,19,23,29,31,37,41,43,47,53,59,61,67,71,73,79,83,89,97$$

We have already met prime numbers in the above, and even used some of their basic properties, that you were certainly very familiar with, but time now to review all this, on a more systematic basis. First, as definition for them, we have:

Definition 9.1. *The prime numbers are the integers $p > 1$ satisfying*

(1) *p does not decompose as $p = ab$, with $a, b > 1$.*
(2) *$p|ab$ implies $p|a$ or $p|b$.*
(3) *$a|p$ implies $a = 1, p$.*

with each of these properties uniquely determining them.

Here the equivalence between (1,2,3) comes from standard arithmetic, and you surely know this. Observe that we have ruled out $0, 1$ from being primes, and you may of course have a bit of thinking at this, and at $0, 1$ in general, but not too much, stay with us.

Still speaking things that you know, already used in the above, we have:

Theorem 9.2. *Any integer $n > 1$ decomposes uniquely as*

$$n = p_1^{a_1} \dots p_k^{a_k}$$

with $p_1 < \ldots < p_k$ primes, and with exponents $a_1, \ldots, a_k \geq 1$.

Proof. This is something that you certainly know, related to the equivalent conditions (1,2,3) in Definition 9.1, and exercise for you, to remember how all this works. Exercise as well, work out this for all integers $n \leq 100$, with no calculators allowed. □

As a first result about the prime numbers themselves, that you certainly know too, but this time coming with a full proof from me, I feel I can do that, we have:

Theorem 9.3. *There is an infinity of prime numbers.*

Proof. Indeed, assuming that we have finitely many prime numbers are $p_1,\ldots,p_k$, we can set $n=p_1\ldots p_k+1$, and this number n cannot factorize, contradiction. □

In practice, we can obtain the prime numbers as follows:

Theorem 9.4. *The set of prime numbers P can be obtained as follows:*

(1) *Start with* $2,3,4,5,6,7,8,9,10,11,12,\ldots$
(2) *Mark the first number, 2, as prime, and remove its multiples.*
(3) *Mark the new first number, 3, as prime, and remove its multiples.*
(4) *Mark the new first number, 5, as prime, and remove its multiples.*
(5) *And so on, with at each step a new prime number found.*

Proof. This algorithm for finding the primes, which is very old, and called "sieve method", is something obvious, with the first steps being as follows:

$$\begin{array}{ccccccccccccccccccc}
\underline{2} & 3 & \cancel{4} & 5 & \cancel{6} & 7 & \cancel{8} & 9 & \cancel{10} & 11 & \cancel{12} & 13 & \cancel{14} & 15 & \cancel{16} & 17 & \cancel{18} & 19 & \cancel{20}\\
 & \underline{3} & & 5 & & 7 & & \cancel{9} & & 11 & & 13 & & \cancel{15} & & 17 & & 19 & \\
 & & & \underline{5} & & 7 & & & & 11 & & 13 & & & & 17 & & 19 & \\
 & & & & & \underline{7} & & & & 11 & & 13 & & & & 17 & & 19 & \\
 & & & & & & & & & \underline{11} & & 13 & & & & 17 & & 19 & \\
 & & & & & & & & & & & \underline{13} & & & & 17 & & 19 & \\
 & & & & & & & & & & & \vdots & & & & & & &
\end{array}$$

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

We will be back to prime numbers on several occasions, in what follows, and notably in chapter 10 below, with a number of more advanced methods, in order to deal with them, and then in Part IV too, by using advanced analysis tools.

9c. Fermat theorem

Moving ahead, we will be mostly interested in congruence questions, with the aim of solving equations over the integers. Let us start with the following definition:

Definition 9.5. *We say that $a,b\in\mathbb Z$ are congruent modulo $c\in\mathbb Z$, and write*

$$a=b(c)$$

when c divides $b-a$.

A first interesting question concerns solving $a=0(n)$, with n fixed and small. By writing $n=p_1^{s_1}\ldots p_k^{s_k}$, the problem reduces to solving $a=0(q)$, with $q=p^s$ small prime power. And as you surely know, there are many tricks here, summarized as follows:

PROPOSITION 9.6. *Given a positive integer $a = a_1 \ldots a_r$, we have:*

(1) $2|a$ *when* $2|a_r$.
(2) $3|a$ *when* $3|\sum a_i$.
(3) $4|a$ *when* $4|a_{r-1}a_r$.
(4) $5|a$ *when* $5|a_r$.
(5) $8|a$ *when* $8|a_{r-2}a_{r-1}a_r$.
(6) $9|a$ *when* $9|\sum a_i$.
(7) $11|a$ *when* $11|\sum(-1)^i a_i$.
(8) $16|a$ *when* $16|a_{r-3}a_{r-2}a_{r-1}a_r$.

PROOF. Here the $q = 2^k, 5$ assertions follow from $10 = 2 \times 5$, the $q = 3, 9$ assertions follow from $10 = 9 + 1$, and the $q = 11$ assertion follows from $10 = 11 - 1$. □

All the above is certainly useful, in the daily life, but what is annoying is that for the missing values, $q = 7, 13$, nothing much intelligent, of the same level of simplicity, can be done. However, as mathematicians, we have solutions for everything, as shown by:

PROPOSITION 9.7. *Assuming that we have convinced mankind to change the numeration basis from* 10 *to* 14*, given a positive integer $a = a_1 \ldots a_r$, we have:*

(1) $2|a$ *when* $2|a_r$.
(2) $3|a$ *when* $3|\sum(-1)^i a_i$.
(3) $4|a$ *when* $4|a_{r-1}a_r$.
(4) $5|a$ *when* $5|\sum(-1)^i a_i$.
(5) $7|a$ *when* $7|a_r$.
(6) $8|a$ *when* $8|a_{r-2}a_{r-1}a_r$.
(7) $9|a$ *when* $9|\sum(-1)^i a_i$.
(8) $13|a$ *when* $13|\sum a_i$.
(9) $16|a$ *when* $16|a_{r-3}a_{r-2}a_{r-1}a_r$.

PROOF. Here the $q = 2^k, 7$ assertions follow from $14 = 2 \times 5$, the $q = 3, 5, 9$ assertions follow from $14 = 15 - 1$, and the $q = 13$ assertion follows from $14 = 13 + 1$. □

As a conclusion to all this, we have solved the $q = 7, 13$ problems mentioned above, but as a caveat, we have now $q = 11$ not working. And is this worth it or not, up to you to decide, and launch an online petition if enthusiastic about it.

Be said in passing, our Proposition 9.7 is a bit ill-formulated, mixing things written in basis 10 and basis 14, and we will leave fixing all this, with a fully correct mathematical statement, as another instructive exercise for you.

Moving ahead, congruences in general, but at a more advanced level, the mother of all results here is the following key theorem of Fermat:

Theorem 9.8. *We have the following congruence, for any prime p,*

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

called Fermat's little theorem.

Proof. The simplest way is to do this by recurrence on $a \in \mathbb{N}$, as follows:

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

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

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

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

9d. Finite fields

The Fermat theorem is particularly interesting when extended from the integers to the arbitrary field case, and can be used in order to elucidate the structure of finite fields. In order to discuss this question, let us start with some basic facts, as follows:

Definition 9.9. *A field is a set F with a sum operation $+$ and a product operation $\times$, subject to the following conditions:*

1. *$a + b = b + a$, $a + (b + c) = (a + b) + c$, there exists $0 \in F$ such that $a + 0 = 0$, and any $a \in F$ has an inverse $-a \in F$, satisfying $a + (-a) = 0$.*
2. *$ab = ba$, $a(bc) = (ab)c$, there exists $1 \in F$ such that $a1 = a$, and any $a \neq 0$ has a multiplicative inverse $a^{-1} \in F$, satisfying $aa^{-1} = 1$.*
3. *The sum and product are compatible via $a(b + c) = ab + ac$.*

You would say, obviously, the simplest possible field is $\mathbb{Q}$. However, this is not exactly true, because, by a strange twist of fate, the numbers $0, 1$, whose presence in a field is mandatory, $0, 1 \in F$, can form themselves a field, with structure as follows:

$$1 + 1 = 0$$

To be more precise, according to our field axioms, all operations of type $a * b$ with $a, b = 0, 1$ are uniquely determined, except for $1+1$. You would say that we must normally set $1 + 1 = 2$, with $2 \neq 0$ being a new field element, but the point is that $1 + 1 = 0$ is something natural too, this being the addition modulo 2. And, what we get is a field:

$$\mathbb{F}_2 = \{0, 1\}$$

Let us summarize this finding, along with a bit more, as follows:

Proposition 9.10. *$\mathbb{Q}$ is the simplest field having the property $1+\ldots+1\neq 0$, in the sense that any field F satisfying this condition must contain $\mathbb{Q}$:*

$$\mathbb{Q}\subset F$$

However, when dropping the assumption $1+\ldots+1\neq 0$, the above conclusion fails, for instance for the field $\mathbb{F}_2=\{0,1\}$, with addition $1+1=0$.

Proof. Here the first assertion is clear, because $1+\ldots+1\neq 0$ tells us that we have an embedding $\mathbb{N}\subset F$, and then by taking inverses with respect to $+$ and $\times$ we obtain $\mathbb{Q}\subset F$. As for the second assertion, this follows from the above discussion. □

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

Theorem 9.11. *Given a field F, define its characteristic $p=char(F)$ as being the smallest $p\in\mathbb{N}$ such that the following happens, and as $p=0$, if this never happens:*

$$\underbrace{1+\ldots+1}_{p\ times}=0$$

Then, assuming $p>0$, this characteristic p must be a prime number, we have a field embedding $\mathbb{F}_p\subset F$, and $q=|F|$ must be of the form $q=p^k$, with $k\in\mathbb{N}$.

Proof. Very crowded statement that we have here, the idea being as follows:

(1) The fact that $p>0$ must be prime comes by contradiction, by using:

$$(\underbrace{1+\ldots+1}_{a\ times})\times(\underbrace{1+\ldots+1}_{b\ times})=\underbrace{1+\ldots+1}_{ab\ times}$$

Indeed, assuming that we have $p=ab$ with $a,b>1$, the above formula corresponds to an equality of type $AB=0$ with $A,B\neq 0$ inside F, which is impossible.

(2) Back to the general case, F has a smallest subfield $E\subset F$, called prime field, consisting of the various sums $1+\ldots+1$, and their quotients. In the case $p=0$ we obviously have $E=\mathbb{Q}$. In the case $p>0$ now, the multiplication formula in (1) shows that the set $S=\{1+\ldots+1\}$ is stable under taking quotients, and so $E=S$.

(3) Now with $E=S$ in hand, we obviously have $(E,+)=\mathbb{Z}_p$, and since the multiplication is given by the formula in (1), we conclude that we have $E=\mathbb{F}_p$, as a field. Thus, in the case $p>0$, we have constructed an embedding $\mathbb{F}_p\subset F$, as claimed.

(4) In the context of the above embedding $\mathbb{F}_p\subset F$, we can say that F is a vector space over $\mathbb{F}_p$, and so we have $|F|=p^k$, with $k\in\mathbb{N}$ being the dimension of this space. □

In relation with Fermat, we can extend the trick in the proof there, as follows:

Proposition 9.12. *In a field F of characteristic $p>0$ we have*

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

for any two elements $a,b\in F$.

Proof. We have indeed the computation, exactly as in the proof of Fermat, by using the fact that the non-trivial binomial coefficients are all multiples of p:

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

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

Observe that we can iterate the Fermat formula, and we obtain $(a+b)^r=a^r+b^r$ for any power $r=p^s$. In particular we have, with $q=|F|$, the following formula:

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

But this is something quite interesting, showing that the following subset of F, which is closed under multiplication, is closed under addition too, and so is a subfield:

$$E=\left\{a\in F\Big|a^q=a\right\}$$

So, what is this subfield $E\subset F$? In the lack of examples, or general theory for subfields $E\subset F$, we are a bit in the dark here, but it seems quite reasonable to conjecture that we have $E=F$. Thus, our conjecture would be that we have the following formula, for any $a\in F$, and with this being the field extension of the Fermat theorem itself:

$$a^q=a$$

Now that we have our conjecture, let us think at a potential proof. And here, by looking at the proof of the Fermat theorem, the recurrence method from there, based on $a\to a+1$, cannot work as such, and must be suitably fine-tuned.

Thinking a bit, the recurrence from the proof of Fermat somehow rests on the fact that the additive group $\mathbb{Z}$ is singly generated, by $1\in\mathbb{Z}$. Thus, we need some sort of field extension of this single generation result, and in the lack of something additive here, the following theorem, which is something multiplicative, comes to the rescue:

Theorem 9.13. *Given a field F, any finite subgroup of its multiplicative group*

$$G\subset F-\{0\}$$

must be cyclic.

Proof. This can be done via some standard arithmetics, as follows:

(1) Let us pick an element $g\in G$ of highest order, $n=ord(g)$. Our claim, which will easily prove the result, is that the order $m=ord(h)$ of any $h\in G$ satisfies $m|n$.

(2) In order to prove this claim, let $d=(m,n)$, write $d=am+bn$ with $a,b\in\mathbb{Z}$, and set $k=g^ah^b$. We have then the following computations:

$$k^m=g^{am}h^{bm}=g^{am}=g^{d-bn}=g^d$$
$$k^n=g^{an}h^{bn}=h^{bn}=h^{d-am}=h^d$$

By using either of these formulae, say the first one, we obtain:

$$k^{[m,n]}=k^{mn/d}=(k^m)^{n/d}=(g^d)^{n/d}=g^n=1$$

Thus $ord(k)|[m,n]$, and our claim is that we have in fact $ord(k)=[m,n]$.

(3) In order to prove this latter claim, assume first that we are in the case $d=1$. But here the result is clear, because the formulae in (2) read $g=k^m,h=g^n$, and since $n=ord(g),m=ord(g)$ are prime to each other, we conclude that we have $ord(k)=mn$, as desired. As for the general case, where d is arbitrary, this follows from this.

(4) Summarizing, we have proved our claim in (2). Now since the order $n=ord(g)$ was assumed to be maximal, we must have $[m,n]|n$, and so $m|n$. Thus, we have proved our claim in (1), namely that the order $m=ord(h)$ of any $h\in G$ satisfies $m|n$.

(5) But with this claim in hand, the result follows. Indeed, since the polynomial x^n-1 has all the elements $h\in G$ as roots, its degree must satisfy $n\geq|G|$. On the other hand, from $n=ord(g)$ with $g\in G$, we have $n||G|$. We therefore conclude that we have $n=|G|$, which shows that G is indeed cyclic, generated by the element $g\in G$. □

We can now extend the Fermat theorem to the finite fields, as follows:

Theorem 9.14. *Given a finite field F, with $q=|F|$ we have*

$$a^q=a$$

for any $a\in F$.

Proof. According to Theorem 9.13 the multiplicative group $F-\{0\}$ is cyclic, of order $q-1$. Thus, the following formula is satisfied, for any $a\in F-\{0\}$:

$$a^{q-1}=1$$

Now by multiplying by a, we are led to the conclusion in the statement, with of course the remark that the formula there trivially holds for $a=0$. □

The Fermat polynomial X^p-X is something very useful, and its field generalization X^q-X, with $q=p^k$ prime power, can be used in order to elucidate the structure of finite fields. In order to discuss this question, let us start with a basic fact, as follows:

Proposition 9.15. *Given a finite field F, we have*

$$X^q-X=\prod_{a\in F}(X-a)$$

with $q=|F|$.

PROOF. We know from the Fermat theorem above that we have $a^q = a$, for any $a \in F$. We conclude from this that all the elements $a \in F$ are roots of the polynomial $X^q - X$, and so this polynomial must factorize as in the statement. □

The continuation of the story is more complicated, as follows:

THEOREM 9.16. *For any prime power $q = p^k$ there is a unique field $\mathbb{F}_q$ having q elements. At $k = 1$ this is the usual $\mathbb{F}_p$, and in general, this is the field making*

$$X^q - X = \prod_{a \in F} (X - a)$$

happen, in some abstract algebraic sense.

PROOF. We are punching here a bit above our weight, the idea being as follows:

(1) At $k = 1$ there is nothing much to be said, because the prime field embedding $\mathbb{F}_p \subset F$ found in Theorem 9.11 must be an isomorphism. Thus, done with this.

(2) At $k \geq 2$ however, both the construction and uniqueness of $\mathbb{F}_q$ are non-trivial. However, the idea is not that complicated. Indeed, instead of struggling first with finding a model for $\mathbb{F}_q$, and then struggling some more with proving the uniqueness, the point is that we can solve both these problems, at the same time, by looking at $X^q - X$.

(3) To be more precise, this polynomial $X^q - X$ must have some sort of abstract, minimal "splitting field", and this is how $\mathbb{F}_q$ comes, both existence and uniqueness. □

9e. Exercises

Exercises:

EXERCISE 9.17.

EXERCISE 9.18.

EXERCISE 9.19.

EXERCISE 9.20.

EXERCISE 9.21.

EXERCISE 9.22.

EXERCISE 9.23.

EXERCISE 9.24.

Bonus exercise.

CHAPTER 10

Prime numbers

10a. Euler formula

Many things can be said about the prime numbers, of analytic nature. At the beginning of everything here, we have the following famous formula, due to Euler:

THEOREM 10.1. *We have the following formula, implying $|P| = \infty$:*

$$\sum_{p\in P}\frac{1}{p}=\infty$$

Moreover, we have the following estimate for the partial sums of this series,

$$\sum_{p<N}\frac{1}{p}>\log\log N-\frac{1}{2}$$

valid for any integer $N \geq 2$.

PROOF. Here is the original proof, due to Euler. The idea is to use the factorization theorem, stating that we have $n = p_1^{a_1}\ldots p_k^{a_k}$, but written upside down, as follows:

$$\frac{1}{n}=\frac{1}{p_1^{a_1}}\cdots\frac{1}{p_k^{a_k}}$$

Indeed, summing now over $n \geq 1$ gives the following beautiful formula:

$$\sum_{n=1}^\infty\frac{1}{n}=\prod_{p\in P}\left(1+\frac{1}{p}+\frac{1}{p^2}+\frac{1}{p^3}+\ldots\right)=\prod_{p\in P}\left(1-\frac{1}{p}\right)^{-1}$$

In what concerns the sum on the left, this is well-known to be ∞. In what concerns now the product on the right, this can be estimated by using log, as follows:

$$\begin{aligned}
\log\left[\prod_{p\in P}\left(1-\frac{1}{p}\right)^{-1}\right] &= -\sum_{p\in P}\log\left(1-\frac{1}{p}\right)\\
&= \sum_{p\in P}\frac{1}{p}+\frac{1}{2p^2}+\frac{1}{3p^3}+\frac{1}{4p^4}+\ldots\\
&< \sum_{p\in P}\frac{1}{p}+\frac{1}{2p^2}+\frac{1}{2p^3}+\frac{1}{2p^4}+\ldots\\
&= \sum_{p\in P}\frac{1}{p}+\frac{1}{2}\sum_{p\in P}\frac{1}{p^2}\cdot\frac{1}{1-1/p}\\
&= \sum_{p\in P}\frac{1}{p}+\frac{1}{2}\sum_{p\in P}\frac{1}{p(p-1)}\\
&< \sum_{p\in P}\frac{1}{p}+\frac{1}{2}\sum_{n=2}^{\infty}\frac{1}{n(n-1)}\\
&= \sum_{p\in P}\frac{1}{p}+\frac{1}{2}
\end{aligned}$$

We therefore obtain the following estimate, which gives the first assertion:

$$\sum_{p\in P}\frac{1}{p}+\frac{1}{2}>\log\left(\sum_{n=1}^{\infty}\frac{1}{n}\right)=\infty$$

Regarding now the second assertion, the idea is to replace in the above computations the set P of all primes by the set of all primes $p<N$. We obtain in this way the following estimate, and with exercise for you, to work out the details:

$$\begin{aligned}
\sum_{p<N}\frac{1}{p}+\frac{1}{2} &> \log\left(\sum_{n=1}^{N}\frac{1}{n}\right)\\
&> \log\left(\int_1^N\frac{1}{x}\,dx\right)\\
&= \log\log N
\end{aligned}$$

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

The Euler formula and its proof are something of utter beauty, suggesting doing an enormous amount of things, and yes indeed, doing such things has been one of the favorite pastimes of mathematicians, since. Here is a brief account, of all this:

(1) The Euler formula $\sum_{p\in P}1/p=\infty$ basically tells us that there are "many primes", but what about the opposite, trying now to prove that there are "few primes"? Well, this comes too from the Euler formula, but in its refined version, with $\log\log N$:

$$\sum_{p<N}\frac{1}{p}\simeq\log\log N$$

Many things can be done here, one of the conclusions being that the N-th prime $\pi(N)$ satisfies $\pi(N)\sim N/\log N$. We will be back to this later in this book.

(2) Still talking analysis, an interesting observation, by Erdős, coming from his own proof of the Euler formula, regards the sets $S\subset\mathbb{N}$ satisfying the following condition:

$$\sum_{s\in S}\frac{1}{s}=\infty$$

Based on this, Erdős conjectured that such sets S contain arbitrarily long arithmetic progessions. And the point is that this is a very difficult and fascinating problem, with the case $S=P$ being settled only recently, by Green and Tao.

(3) Leaving aside now estimates and analysis, and going back to the beginning of Euler's proof, let us look more in detail at the formula there, namely:

$$\sum_{n=1}^\infty\frac{1}{n}=\prod_{p\in P}\left(1-\frac{1}{p}\right)^{-1}$$

This formula is something really beautiful, and the more you look at it, thinking at versions and so on, the more you are lost into the mysteries of number theory.

(4) To be more precise, the above formula suggests introducing the following function, depending on a parameter s, which can be integer, real, or even complex:

$$\zeta(s)=\sum_{n=1}^\infty\frac{1}{n^s}$$

And this is the famous Riemann zeta function, which obsesses all number theorists, be them algebraists, analysts, geometers, physicists, or amateurs. We will be talking about this magical function later in this book, in Part IV, after learning some analysis.

10b. Basic estimates

Let us start a more advanced study, by improving the Euler formula. We have:

Theorem 10.2. *We have the following formula, with sum over primes,*

$$\sum_{p<N}\frac{1}{p}>\log\log N-\frac{1}{2}$$

and the $1/2$ *constant on the right can be improved to* $\log(\pi^2/6)=0.49770..$

Proof. This is something quite straightforward, as follows:

(1) By using the unique factorization $n=p_1^{a_1}\ldots p_k^{a_k}$, we have:

$$\begin{aligned}\prod_{p<N}\left(1-\frac{1}{p}\right)^{-1}&=\prod_{p<N}\left(1+\frac{1}{p}+\frac{1}{p^2}+\frac{1}{p^3}+\ldots\right)\\&>\sum_{n=1}^{N-1}\frac{1}{n}\\&>\int_1^N\frac{1}{x}\,dx\\&=\log N\end{aligned}$$

(2) But the product on the left can be estimated by using log, as follows:

$$\begin{aligned}\log\left[\prod_{p<N}\left(1-\frac{1}{p}\right)^{-1}\right]&=-\sum_{p<N}\log\left(1-\frac{1}{p}\right)\\&=\sum_{p<N}\frac{1}{p}+\frac{1}{2p^2}+\frac{1}{3p^3}+\frac{1}{4p^4}+\ldots\\&<\sum_{p<N}\frac{1}{p}+\frac{1}{2p^2}+\frac{1}{2p^3}+\frac{1}{2p^4}+\ldots\\&=\sum_{p<N}\frac{1}{p}+\frac{1}{2}\sum_{p<N}\frac{1}{p^2}\cdot\frac{1}{1-1/p}\\&=\sum_{p<N}\frac{1}{p}+\frac{1}{2}\sum_{p<N}\frac{1}{p(p-1)}\\&<\sum_{p<N}\frac{1}{p}+\frac{1}{2}\sum_{n=2}^{\infty}\frac{1}{n(n-1)}\\&=\sum_{p<N}\frac{1}{p}+\frac{1}{2}\end{aligned}$$

(3) Thus, we are led to the estimate in the statement, namely:

$$\sum_{p<N}\frac{1}{p}>\log\log N-\frac{1}{2}$$

(4) In order now to improve this, a quick look at what we did in (1) and (2) reveals four $<$ signs, that we can all improve, if we want to. However, we will leave this for later, when talking about Mertens and his theorems. In the meantime, we would like to present a slight improvement, coming via a different technique, which is quite instructive.

(5) The point indeed is that we have a rival method, based by using the factorization $n=p_1\ldots p_km^2$, with p_i distinct primes. This factorization gives:

$$\begin{aligned}\sum_{n=1}^{N-1}\frac{1}{n} &< \prod_{p<N}\left(1+\frac{1}{p}\right)\sum_{m=1}^{N}\frac{1}{m^2}\\ &< \prod_{p<N}\exp\left(\frac{1}{p}\right)\sum_{m=1}^{\infty}\frac{1}{(m-1/2)(m+1/2)}\\ &= \exp\left(\sum_{p<N}\frac{1}{p}\right)\sum_{m=1}^{\infty}\frac{1}{m-1/2}-\frac{1}{m+1/2}\\ &= 2\exp\left(\sum_{p<N}\frac{1}{p}\right)\end{aligned}$$

We therefore obtain the following estimate, for our sum:

$$\sum_{p<N}\frac{1}{p}>\log\log N-\log 2$$

(6) However, $\log 2=0.69314..$ does not improve our $1/2$ constant, and we have to be more careful with our telescoping in (5). By separating the first term, we get closer:

$$\sum_{m=1}^{\infty}\frac{1}{m^2}<1+\frac{2}{3}=\frac{5}{3}\quad,\quad\log\left(\frac{5}{3}\right)=0.51082..$$

By separating the first two terms, we get even closer, but still not there:

$$\sum_{m=1}^{\infty}\frac{1}{m^2}<1+\frac{1}{4}+\frac{2}{5}=\frac{33}{20}\quad,\quad\log\left(\frac{33}{20}\right)=0.50077..$$

However, with the first three terms separated, what we get is a win:

$$\sum_{m=1}^{\infty}\frac{1}{m^2}<1+\frac{1}{4}+\frac{1}{9}+\frac{2}{7}=\frac{415}{252}\quad,\quad\log\left(\frac{415}{252}\right)=0.49884..$$

(7) In practice now, in order to finish this discussion, in a professional way, we can invoke the Basel formula, due to Euler, which is however something quite complicated:

$$\sum_{m=1}^{\infty}\frac{1}{m^2}=\frac{\pi^2}{6}$$

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

Although we will not need this here, with the above estimates to be soon improved by theorems of Mertens, let us prove however the formula that we used at the end:

Theorem 10.3. *We have the following formula, due to Euler,*

$$\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{6}$$

answering the Basel problem, asking for the computation of this sum.

Proof. This is something quite tricky. The original proof of Euler is as follows, making some manipulations on the Taylor series expansion of $\sin x/x$, based on the fact that the zeroes of this function appear at $x=k\pi$, with $k\in\mathbb{Z}$:

$$\begin{aligned}\frac{\sin x}{x} &= 1-\frac{x^2}{3!}+\frac{x^4}{5!}-\frac{x^6}{7!}+\ldots\\ &= \left(1-\frac{x}{\pi}\right)\left(1+\frac{x}{\pi}\right)\left(1-\frac{x}{2\pi}\right)\left(1+\frac{x}{2\pi}\right)\ldots\\ &= \left(1-\frac{x^2}{\pi^2}\right)\left(1-\frac{x^2}{4\pi^2}\right)\left(1-\frac{x^2}{9\pi^2}\right)\ldots\\ &= 1-\frac{1}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}x^2+\ldots\end{aligned}$$

In practice, all this needs a bit more justification, which can be obtained by taking the logarithm, or passing to complex numbers, of even passing to Fourier analysis, and getting the result from the Parseval formula. Exercise for you, to read all this. □

10c. Zeta function

Before moving ahead with the Mertens theorems, substantially improving the above, several comments are in order, with respect to the Euler method. Let us introduce:

Definition 10.4. *Associated to any $s>1$ is the function*

$$\zeta(s)=\sum_{n=1}^{\infty}\frac{1}{n^s}$$

called Riemann zeta function.

Observe that the above series converges indeed, as a Riemann sum approximation, by usual rectangles, of the following convergent integral:

$$\begin{aligned}\int_1^\infty\frac{1}{x^s}\,dx &= \left[\frac{x^{1-s}}{1-s}\right]_1^\infty\\ &= 0-\frac{1}{1-s}\\ &= \frac{1}{s-1}\\ &< \infty\end{aligned}$$

Based on this, we can further say that, more generally, the series converges for any $s\in\mathbb{C}$ satisfying $Re(s)>1$. And then, with a bit of complex analysis, we can have the zeta function working in the whole complex plane $\mathbb{C}$, as a meromorphic function there, by analytic continuation. But more on this, complex zeta, in Part IV below.

Here we will just use zeta at $s>1$, and why not its truncations too, at any $s\in\mathbb{R}$:

$$S_N=\sum_{n=1}^N\frac{1}{n^s}$$

As a first observation, the Basel formula, from Theorem 10.3, reformulates as:

Theorem 10.5. *We have the following formula, coming from the Basel problem:*

$$\zeta(2)=\frac{\pi^2}{6}$$

More generally, any value $\zeta(2k)$ with $k\in\mathbb{N}$ is a rational multiple of π^{2k}.

Proof. Here the formula of $\zeta(2)$ is what we have in Theorem 10.3, and the generalization to $\zeta(2k)$ with $k\in\mathbb{N}$ comes by further studying the Euler formula, namely:

$$\frac{\sin x}{x}=\left(1-\frac{x^2}{\pi^2}\right)\left(1-\frac{x^2}{4\pi^2}\right)\left(1-\frac{x^2}{9\pi^2}\right)\ldots$$

To be more precise, after some combinatorial work, that we will not get into here, we are led to the following formula, with B_n being the Bernoulli numbers:

$$\zeta(2k)=(-1)^{k+1}\frac{(2\pi)^{2k}B_{2k}}{2\cdot(2k)!}$$

In practice, this gives the following fomulae for the first few values $\zeta(2k)$:

$$\zeta(2)=\frac{\pi^2}{6}\quad,\quad\zeta(4)=\frac{\pi^4}{90}\quad,\quad\zeta(6)=\frac{\pi^6}{945}\quad,\quad\zeta(8)=\frac{\pi^8}{9450}$$

As usual, exercise for you to read more about this, as a continuation of the reading suggested in the proof of Theorem 10.3. All first-class mathematics, worth the effort. □

Many other things can be said about zeta, along the same lines, but it is not about this that we want to talk, in this chapter, with all this zeta material being deferred to Part IV below. What we want to discuss here is what happens to the Euler estimate from Theorem 10.2, when adding an exponent $s \in \mathbb{R}$ there. Let us start with:

Proposition 10.6. *The Euler estimate can be generalized into*

$$\sum_{p<N}\frac{1}{p^s} > \log\left(\int_1^N\frac{1}{x^s}\,dx\right) - \frac{1}{2}\sum_{n=2}^{N-1}\frac{1}{n^s(n^s-1)}$$

with the above integral given by the formula

$$\int_1^N\frac{1}{x^s}\,dx = \begin{cases}\frac{N^{1-s}-1}{1-s} & \text{if } s\neq 1\\ \log N & \text{if } s=1\end{cases}$$

involving now a real parameter $s \in \mathbb{R}$, with exactly the same proof.

Proof. By using the unique factorization $n = p_1^{a_1}\ldots p_k^{a_k}$, as before, we have:

$$\begin{aligned}\prod_{p<N}\left(1-\frac{1}{p^s}\right)^{-1} &= \prod_{p<N}\left(1+\frac{1}{p^s}+\frac{1}{p^{2s}}+\frac{1}{p^{3s}}+\ldots\right)\\ &> \sum_{n=1}^{N-1}\frac{1}{n^s}\\ &> \int_1^N\frac{1}{x^s}\,dx\end{aligned}$$

But the product on the left can be estimated by using log, as follows:

$$\begin{aligned}
\log\left[\prod_{p<N}\left(1-\frac{1}{p^s}\right)^{-1}\right] &= -\sum_{p<N}\log\left(1-\frac{1}{p^s}\right)\\
&= \sum_{p<N}\frac{1}{p^s}+\frac{1}{2p^{2s}}+\frac{1}{3p^{3s}}+\frac{1}{4p^{4s}}+\ldots\\
&< \sum_{p<N}\frac{1}{p^s}+\frac{1}{2p^{2s}}+\frac{1}{2p^{3s}}+\frac{1}{2p^{4s}}+\ldots\\
&= \sum_{p<N}\frac{1}{p^s}+\frac{1}{2}\sum_{p<N}\frac{1}{p^s}\cdot\frac{1}{1-1/p^s}\\
&= \sum_{p<N}\frac{1}{p^s}+\frac{1}{2}\sum_{p<N}\frac{1}{p^s(p^s-1)}\\
&< \sum_{p<N}\frac{1}{p^s}+\frac{1}{2}\sum_{n=2}^{N-1}\frac{1}{n^s(n^s-1)}
\end{aligned}$$

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

In the case $s>1$, which is the one of main interest, we obtain in this way:

Theorem 10.7. *We have the following Euler type estimate*

$$\sum_{p<N}\frac{1}{p^s}>\log\left(\frac{1-N^{1-s}}{s-1}\right)-\frac{\zeta(2s)}{2}$$

valid for any value of the parameter $s>1$.

Proof. In the case $s>1$ the estimate that we found in Proposition 10.6 gives:

$$\begin{aligned}
\sum_{p<N}\frac{1}{p^s} &> \log\left(\frac{1-N^{1-s}}{s-1}\right)-\frac{1}{2}\sum_{n=2}^{N-1}\frac{1}{n^s(n^s-1)}\\
&> \log\left(\frac{1-N^{1-s}}{s-1}\right)-\frac{1}{2}\sum_{n=2}^{\infty}\frac{1}{n^s(n^s-1)}\\
&> \log\left(\frac{1-N^{1-s}}{s-1}\right)-\frac{1}{2}\sum_{n=2}^{\infty}\frac{1}{(n-1)^{2s}}\\
&> \log\left(\frac{1-N^{1-s}}{s-1}\right)-\frac{\zeta(2s)}{2}
\end{aligned}$$

Here we have used the following inequality, with $\varepsilon = 1/n < 1$, which is true:

$$\begin{aligned}\frac{1}{n^s(n^s-1)} < \frac{1}{(n-1)^{2s}} &\iff (n-1)^{2s} < n^s(n^s-1)\\ &\iff \left(1-\frac{1}{n}\right)^{2s} < 1-\frac{1}{n^s}\\ &\iff (1-\varepsilon)^{2s} < 1-\varepsilon^s\\ &\iff (1-\varepsilon)^{2s-1} < \frac{1-\varepsilon^s}{1-\varepsilon}\end{aligned}$$

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

It is possible to futher build along the above lines, but we will leave this discussion for later, in Part IV, when talking more in detail about the Riemann zeta function.

10d. Mertens theorems

Moving ahead now, the continuation of the story involves the work of Mertens, that we would like to discuss now. Let us start with some analysis conventions:

Definition 10.8. *We use the following notations:*

(1) *We write $f \simeq g$ when $f - g \to 0$.*
(2) *We write $f \cong g$ when $f - g$ is bounded.*
(3) *We write $f \sim g$ when $f/g \to 1$.*
(4) *We write $f \approx g$ when f/g is bounded.*

Occasionaly, we will use as well the Landau $O(f), o(f)$ symbols, making it for 2 notations instead of 4. With these conventions, the formulae of Mertens are as follows:

Fact 10.9. *We have the following Mertens estimates, in the $N \to \infty$ limit,*

$$\sum_{p<N} \frac{\log p}{p} \cong \log N$$

$$\sum_{p<N} \frac{1}{p} \simeq \log\log N + M$$

$$\sum_{p<N} \log\left(1-\frac{1}{p}\right) \simeq -\log\log N - \gamma$$

$M = 0.26149..$ and $\gamma = 0.57721..$ being the Mertens and Euler-Mascheroni constants.

Obviously, these formulae are related, and there are many things that can be said here. We will do this slowly. To start with, we would like to talk about the second formula, which improves our Euler estimates before. The precise result here is as follows:

Theorem 10.10. *We have the following formula, with sum over primes,*

$$\sum_{p\leq N}\frac{1}{p}\simeq\log\log N+M$$

and with $M=0.26149..$ being a constant, called Mertens constant.

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

(1) As a first comment, observe that we have switched in the statement from sums over primes $p<N$, to sums over primes $p\leq N$. The point is that sums of type $p<N$ were best adapted to the Euler summation, which eventually leads to an integral of $1/x$, that we want to be $\log N$ instead of $\log(N+1)$. However, as we will see in a moment, the Mertens summation is best written with $p\leq N$. Of course, at the level of the final results, Theorem 10.2 and the present theorem, this does not matter, because:

$$\log\log N\simeq\log\log(N+1)$$

(2) Getting now to the proof, this is based on the following formula, which comes as usual from the unique factorization of integers, $n=p_1^{a_1}\ldots p_k^{a_k}$, with the sum being over prime powers p^k, and with the exponent $[N/p^k]$ being an integer part:

$$N!=\prod_{p^k\leq N}p^{[N/p^k]}$$

(3) By talking the logarithm, we obtain from this the following estimate:

$$\begin{aligned}\log N! &= \sum_{p^k\leq N}\left[\frac{N}{p^k}\right]\log p\\ &= \sum_{p^k\leq N}\left(\frac{N}{p^k}+o(1)\right)\log p\\ &= N\sum_{p^k\leq N}\frac{\log p}{p^k}+o(1)\sum_{p^k\leq N}\log p\end{aligned}$$

(4) By dividing by N and using $\log N!=N\log N+O(N)$, this gives:

$$\begin{aligned}\sum_{p^k\leq N}\frac{\log p}{p^k} &= \frac{\log N!}{N}+o\left(\frac{1}{N}\right)\sum_{p^k\leq N}\log p\\ &= \log N+o(1)+o\left(\frac{1}{N}\right)\sum_{p^k\leq N}\log p\end{aligned}$$

(5) Now let us analyze the sum on the right. We have:

$$\begin{aligned}\sum_{p^k\leq N}\log p &\leq \sum_{p\in(N,2N]}\log p\\ &\leq \log\binom{2N}{N}\\ &= O(N)\end{aligned}$$

(6) We conclude that the estimate in (4) can be written as follows:

$$\sum_{p^k\leq N}\frac{\log p}{p^k}=\log N+o(1)$$

(7) Now since the sum of reciprocals of squares is finite, $\sum_{k\geq 1}1/k^2<\infty$, we can remove all the squares from the sum on the left, and we are left with:

$$\sum_{p\leq N}\frac{\log p}{p}=\log N+o(1)$$

(8) But now by doing a partial summation, in the obvious way, this gives a formula as follows, with $M\in\mathbb{R}$ being a certain constant:

$$\sum_{p\leq N}\frac{1}{p}\simeq\log\log N+M+O\left(\frac{1}{\log N}\right)$$

Thus, we are led to the convergence conclusion in the statement, and of course with the precise numerics for the Mertens constant M remaining to be justified. □

Observe that the above proof crucially uses $\log N!=N\log N+O(N)$. Although we will not really need this, at this point, let us record the following famous result here:

Theorem 10.11. *We have the Stirling formula*

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

valid in the $N\to\infty$ limit.

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

(1) Let us first see what we can get with Riemann sums. We have:

$$\begin{aligned}\log(N!) &= \sum_{k=1}^N\log k\\ &\approx \int_1^N\log x\,dx\\ &= N\log N-N+1\end{aligned}$$

By exponentiating, this gives the following estimate, which is not bad:

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

(2) We can improve our estimate by replacing the rectangles from the Riemann sum approach to the integrals by trapezoids. In practice, this gives the following estimate:

$$\begin{aligned}\log(N!) &= \sum_{k=1}^N\log k\\ &\approx \int_1^N\log x\,dx+\frac{\log 1+\log N}{2}\\ &= N\log N-N+1+\frac{\log N}{2}\end{aligned}$$

By exponentiating, this gives the following estimate, which gets us closer:

$$N!\approx\left(\frac{N}{e}\right)^N\cdot e\cdot\sqrt{N}$$

(3) In order to conclude, we must take some kind of mathematical magnifier, and carefully estimate the error made in (2). Fortunately, this mathematical magnifier exists, called Euler-Maclaurin formula, and after some tough computations, we get to:

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

(4) However, all this remains a bit complicated, so we would like to present now an alternative approach to (3), which also misses some details, but better does the job, explaining where the $\sqrt{2\pi}$ factor comes from. First, by partial integration we have:

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

(5) 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:

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

(6) By exponentiating, we obtain from this the following estimate:

$$x^N e^{-x} \simeq \left(\frac{N}{e}\right)^N e^{-y^2/2N}$$

(7) Now by integrating, we obtain from this the following estimate:

$$\begin{aligned}
N! &= \int_0^\infty x^N e^{-x} dx \\
&\simeq \int_{-N}^N \left(\frac{N}{e}\right)^N e^{-y^2/2N}\, dy \\
&\simeq \left(\frac{N}{e}\right)^N \int_{\mathbb{R}} e^{-y^2/2N}\, dy \\
&= \left(\frac{N}{e}\right)^N \sqrt{2N} \int_{\mathbb{R}} e^{-z^2}\, dz \\
&= \left(\frac{N}{e}\right)^N \sqrt{2\pi N}
\end{aligned}$$

(8) Here we have used at the end the following key formula, due to Gauss:

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

Thus, we have proved the Stirling formula, as formulated in the statement. □

Now back to the Mertens second theorem, the continuation of the story, involving Mertens, Meissel and others, is quite long. The Mertens proof can be of course improved, with some technical bounds for M, and for the rate of convergence too.

However, skipping this discussion, which is quite technical, and getting to the point, the Mertens constant M itself, there are several interesting formulae for it. According to

Theorem 10.10, this constant appears by definition as follows:

$$M=\lim_{N\to\infty}\sum_{p<N}\frac{1}{p}-\log\log N$$

In order to further build on this, we will need the following standard result:

Theorem 10.12. *The following limit converges,*

$$\gamma=\lim_{N\to\infty}\sum_{n=1}^N\frac{1}{n}-\log N$$

the result being the Euler-Mascheroni constant $\gamma=0.57721..$

Proof. This is indeed something very standard, coming from basic calculus. In addition to the formula in the statement, there is a bewildering quantity of alternative formulae for γ, all being useful when doing number theory, which are as follows:

(1) First, we have the following alternative formula:

$$\gamma=-\int_0^\infty e^{-x}\log x\,dx$$

With a change of variables, this is equivalent to the following formula:

$$\gamma=-\int_0^1\log\left(\log\frac{1}{x}\right)dx$$

(2) We have as well the following formula, with $[.]$ being the integer part:

$$\gamma=\int_1^\infty\frac{1}{[x]}-\frac{1}{x}\,dx$$

Alternatively, in terms of the upper integer part $[[.]]$, we have:

$$\gamma=\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^n\left[\left[\frac{n}{k}\right]\right]-\frac{n}{k}$$

(3) In relation with the gamma function, we have the following formula:

$$\gamma=-\Gamma'(1)$$

Equivalently, still in terms of the gamma function, we have the following formula:

$$\gamma=\lim_{z\to0}\frac{1}{z}-\Gamma(z)$$

As a third formula for γ, still in terms of the gamma function, we have:

$$\gamma=\lim_{z\to0}\frac{1}{2z}\left(\frac{1}{\Gamma(1+z)}-\frac{1}{\Gamma(1-z)}\right)$$

(4) In relation now with the zeta function, we have the following formula:

$$\gamma=\sum_{n=2}^\infty(-1)^n\frac{\zeta(n)}{n}$$

Alternatively, still in terms of zeta, we have the following formula:

$$\gamma=\log\left(\frac{4}{\pi}\right)+\sum_{n=2}^\infty(-1)^n\frac{\zeta(n)}{2^{n-1}n}$$

(5) We have as well the following alternative formula:

$$\gamma=\lim_{s\to1^+}\sum_{n=1}^\infty\frac{1}{n^s}-\frac{1}{s^n}$$

In terms of the zeta function, this latter formula simply reads:

$$\gamma=\lim_{s\to1}\zeta(s)-\frac{1}{s-1}$$

Alternatively, still in terms of the zeta function around 1, this reads:

$$\gamma=\lim_{s\to0}\frac{\zeta(1+s)+\zeta(1-s)}{2}$$

(6) And as usual, exercise for you to do the calculus for all this, or of course look it up, in case the calculus turns too complicated. □

Now back to the Mertens constant, we have the following formula for it:

Theorem 10.13. *The Mertens constant is given by the formula*

$$M=\gamma+\sum_p\left(\log\left(1-\frac{1}{p}\right)+\frac{1}{p}\right)$$

with $\gamma=0.57721..$ *being the Euler-Mascheroni constant.*

Proof. We know that the Mertens constant appears by definition as follows:

$$\sum_{p<N}\frac{1}{p}\simeq\log\log N+M$$

But the Euler-Mascheroni constant is related as well to the primes, as follows:

$$\sum_{p<N}\log\left(1-\frac{1}{p}\right)\simeq-\log\log N-\gamma$$

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

Getting back now to the Mertens theorem, the above considerations eventually lead, via some more work, to the precise numeric figure from Theorem 10.10, namely:

$$M = 0.26149..$$

Changing topics now, as already mentioned in the above, Mertens proved in fact three theorems regarding the prime numbers, with Theorem 10.10, the most famous one, being his second theorem. His first theorem is a related formula, as follows:

Theorem 10.14. *We have the following formula,*

$$\sum_{p<N}\frac{\log p}{p}\cong\log N$$

with the sum being over primes.

Proof. This is indeed something quite standard, and with the precise upper bound obtained by Mertens being as follows:

$$\sum_{p<N}\frac{\log p}{p}<\log N+2$$

As usual, exercise for you, to read more about all this. □

As for the third theorem of Mertens, again related to all this, this is as follows:

Theorem 10.15. *We have the following formula,*

$$\prod_{p<N}\left(1-\frac{1}{p}\right)\approx\frac{e^{-\gamma}}{\log N}$$

with the product being over primes.

Proof. In order to establish the result, we can use the following formula:

$$\left(1-\frac{1}{p}\right)\left(1+\frac{1}{p}\right)=1-\frac{1}{p^2}$$

Indeed, this gives the following formula for the product in the statement:

$$\prod_{p<N}\left(1-\frac{1}{p}\right)=\prod_{p<N}\left(1-\frac{1}{p^2}\right)\prod_{p<N}\left(1-\frac{1}{p}\right)^{-1}$$

Now by inverting and applying the logarithm, we obtain:

$$\begin{aligned}\log\left[\prod_{p<N}\left(1-\frac{1}{p}\right)^{-1}\right] &= \log\left[\prod_{p<N}\left(1-\frac{1}{p^2}\right)^{-1}\right]+\log\left[\prod_{p<N}\left(1-\frac{1}{p}\right)\right]\\ &= \log\left[\prod_{p<N}\left(1+\frac{1}{p^2}+\frac{1}{p^4}+\ldots\right)\right]+\sum_{p<N}\log\left(1-\frac{1}{p}\right)\\ &\simeq \log\left[\sum_{n=1}^\infty\frac{1}{n^2}\right]+\sum_{p<N}\log\left(1-\frac{1}{p}\right)\\ &= \frac{\pi^2}{6}+\sum_{p<N}\log\left(1-\frac{1}{p}\right)\\ &\simeq \frac{\pi^2}{6}-\log\log N-\gamma\end{aligned}$$

Now by exponentiating, we are led to the conclusion in the statement: □

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

10e. Exercises

Exercises:

Exercise 10.16.

Exercise 10.17.

Exercise 10.18.

Exercise 10.19.

Exercise 10.20.

Exercise 10.21.

Exercise 10.22.

Exercise 10.23.

Bonus exercise.

CHAPTER 11

Squares, residues

11a. Squares, residues

Let us go back to what we did in chapter 9 with congruences. Our aim here will be that of further building on some of the theorems there. To be more precise, we will be interested in solving the following ubiquitous equation, over the integers:

$$a=b^2(c)$$

Many things can be said here, of various levels of difficulty.

11b. Legendre symbol

Inspired by the above, we have the following definition, which is something quite far-reaching, putting everything in relation with squares on a solid basis:

Definition 11.1. *The Legendre symbol is defined as follows,*

$$\left(\frac{a}{p}\right)=\begin{cases}1 & \text{if } \exists\, b\neq 0, a=b^2(p)\\ 0 & \text{if } a=0(p)\\ -1 & \text{if } \nexists\, b, a=b^2(p)\end{cases}$$

with $p\geq 3$ prime.

Now leaving aside all sorts of nice and amateurish things that can be said about $a=b^2(c)$, and going straight to the point, what we want to do is to compute this symbol. I mean, if we manage to have this symbol computed, that would be a big win.

As a first result on the subject, due to Euler, we have:

Theorem 11.2. *The Legendre symbol is given by the formula*

$$\left(\frac{a}{p}\right)=a^{\frac{p-1}{2}}(p)$$

called Euler formula for the Legendre symbol.

Proof. This is something not that complicated, the idea being as follows:

(1) We know from Fermat that we have $a^p = a(p)$, and leaving aside the case $a = 0(p)$, which is trivial, and therefore solved, this tells us that $a^{p-1} = 1(p)$. But since our prime p was assumed to be odd, $p \geq 3$, we can write this formula as follows:

$$\left(a^{\frac{p-1}{2}} - 1\right)\left(a^{\frac{p-1}{2}} + 1\right) = 0(p)$$

(2) Now let us think a bit at the elements of $\mathbb{F}_p - \{0\}$, which can be a quadratic residue, and which cannot. Since the squares b^2 with $b \neq 0$ are invariant under $b \to -b$, and give different b^2 values modulo p, up to this symmetry, we conclude that there are exactly $(p-1)/2$ quadratic residues, and with the remaining $(p-1)/2$ elements of $\mathbb{F}_p - \{0\}$ being non-quadratic residues. So, as a conclusion, $\mathbb{F}_p - \{0\}$ splits as follows:

$$\mathbb{F}_p - \{0\} = \left\{\frac{p-1}{2}\ squares\right\} \bigsqcup \left\{\frac{p-1}{2}\ non-squares\right\}$$

(3) Now by comparing what we have in (1) and in (2), the splits there must correspond to each other, so we are led to the following formula, valid for any $a \in \mathbb{F}_p - \{0\}$:

$$a^{\frac{p-1}{2}} = \begin{cases} 1 & \text{if } \exists\, b, a = b^2 \\ -1 & \text{if } \nexists\, b, a = b^2 \end{cases}$$

By comparing now with Definition 11.1, we obtain the formula in the statement. □

As a first consequence of the Euler formula, we have the following result:

Proposition 11.3. *We have the following formula, valid for any $a, b \in \mathbb{Z}$:*

$$\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right)$$

That is, the Legendre symbol is multiplicative in its upper variable.

Proof. This is clear indeed from the Euler formula, because $a^{\frac{p-1}{2}}(p)$ is obviously multiplicative in $a \in \mathbb{Z}$. Alternatively, this can be proved as well directly, by reasoning along the lines of (2) in the proof of Theorem 11.2. □

The above result looks quite conceptual, and as consequences, we have:

Proposition 11.4. *We have the following formula, telling us that modulo any prime number p, a product of non-squares is a square:*

$$\left(\frac{a}{p}\right) = -1\ ,\ \left(\frac{b}{p}\right) = -1 \implies \left(\frac{ab}{p}\right) = 1$$

Also, the Legendre symbol, regarded as a function

$$\chi : \mathbb{F}_p - \{0\} \to \{-1, 1\} \quad , \quad \chi(a) = \left(\frac{a}{p}\right)$$

is a character, in the sense that it is multiplicative.

Proof. The first asssertion is a consequence of Proposition 11.3, more or less equivalent to it, and with the remark that this formally holds at $p=2$ too, as $\emptyset \implies \emptyset$. As for the second assertion, this is just a fancy reformulation of Proposition 11.3. □

It is possible to say some further conceptual things, some sounding very fancy, in relation with Proposition 11.3 and Proposition 11.4. But remember that, according to the plan made in the beginning of this chapter, we are here for the kill, namely computing the Legendre symbol, no matter what, and with no prisoners taken.

11c. Quadratic reciprocity

So, computing the Legendre symbol. There are many things to be known here, and all must be known, for efficient application, to the real life. We have opted to present them all, of course with full proofs, when these proofs are easy, and leave the more complicated proofs for later. As a first and main result, which is something heavy, we have:

Theorem 11.5. *We have the quadratic reciprocity formula*

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right)=(-1)^{\frac{p-1}{2}\cdot\frac{q-1}{2}}$$

valid for any primes $p,q\geq 3$.

Proof. This is something quite tricky, one proof being as follows:

(1) First we have a combinatorial formula for the Legendre symbol, called Gauss lemma. Given a prime number $q\geq 3$, and $a\neq 0(q)$, consider the following sequence:

$$a\ ,\ 2a\ ,\ 3a\ ,\ \ldots\ ,\ \frac{q-1}{2}\,a$$

The Gauss lemma tells us that if we look at these numbers modulo q, and denote by n the number of residues modulo q which are greater than $q/2$, then:

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

(2) In order to prove this lemma, the idea is to look at the following product:

$$Z=a\times 2a\times 3a\times\ldots\times\frac{q-1}{2}\,a$$

Indeed, on one hand we have the following formula, with Euler used at the end:

$$Z=a^{\frac{q-1}{2}}\left(\frac{q-1}{2}\right)!=\left(\frac{a}{q}\right)\left(\frac{q-1}{2}\right)!$$

(3) On the other hand, we can compute Z in more complicated way, but leading to a simpler answer. Indeed, let us define the following function:

$$|x|=\begin{cases}x&\text{if }0<x<q/2\\ q-x&\text{if }q/2<x<q\end{cases}$$

With this convention, our product Z is given by the following formula, with n being as in (1), namely the number of residues modulo q which are greater than $q/2$:

$$Z=(-1)^n\times|a|\times|2a|\times|3a|\times\ldots\times\left|\frac{q-1}{2}\,a\right|$$

(4) But, the numbers $|ra|$ appearing in the above formula are all distinct, so up to a permutation, these must be exactly the numbers $1,2,\ldots,\frac{q-1}{2}$. That is, we have:

$$\left\{|a|\,,\,|2a|\,,\,|3a|\,,\,\ldots\,,\,\left|\frac{q-1}{2}\,a\right|\right\}=\left\{1,2,3,\ldots,\frac{q-1}{2}\right\}$$

Now by multiplying all these numbers, we obtain, via the formula in (3):

$$Z=(-1)^n\left(\frac{q-1}{2}\right)!$$

(5) But this is what we need, because when comparing with what we have in (2), we obtain the following formula, which is exactly the one claimed by the Gauss lemma:

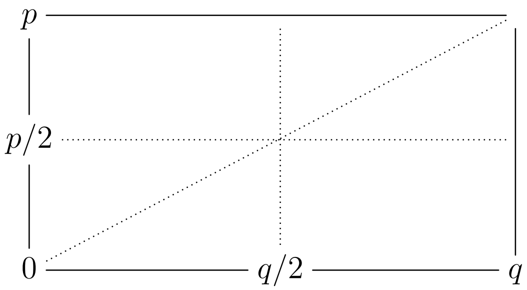
$$\left(\frac{a}{q}\right)=(-1)^n$$

(6) Next, we have a variation of this formula, due to Eisenstein. His formula for the Legendre symbol, this time involving a prime number numerator $p\geq 3$ in the symbol, is as follows, with the quantities on the right being integer parts, and with the proof being very similar to the proof of the Gauss lemma, that we will leave here as an exercise:

$$\left(\frac{p}{q}\right)=(-1)^n\quad,\quad n=\sum_{k=0}^{(q-1)/2}\left[\frac{2kp}{q}\right]$$

(7) The key point now is that, in this latter formula of Eisenstein, the number n itself counts the points of the lattice $\mathbb{Z}^2$ lying in the triangle $(0,0),(q,0),(q,p)$. So, based on

this observation, let us draw a picture, as follows:



(8) We must count the points of $\mathbb{Z}^2$ lying in the triangle $(0,0),(q,0),(q,p)$, modulo 2. This triangle has 3 components, when split by the dotted lines above. Since the points at right, in the small rectangle, and in the small triangle above it, will cancel modulo 2, we are left with the points at left, in the small triangle there, and the conclusion is that, if we denote by m the number of integer points there, we have the following formula:

$$\left(\frac{p}{q}\right)=(-1)^m$$

(9) Now by flipping the diagram, we have as well the following formula, with r being the number of integer points in the small triangle above the small triangle in (8):

$$\left(\frac{q}{p}\right)=(-1)^r$$

(10) But, since our two small triangles add up to a small rectangle, we have:

$$m+r=\frac{p-1}{2}\cdot\frac{q-1}{2}$$

Thus, by multiplying the formulae in (8) and (9), we are led to the result. □

We will see later in this chapter an alternative proof as well, for the above result.

As a comment now, the above result is extremely powerful, here being an illustration, computing the seemingly uncomputable number on the left in a matter of seconds:

$$\left(\frac{3}{173}\right)=(-1)^{\frac{3-1}{2}\cdot\frac{173-1}{2}}\left(\frac{173}{3}\right)=\left(\frac{173}{3}\right)=\left(\frac{2}{3}\right)=-1$$

In fact, when combining Theorem 11.5 with Proposition 11.3, it is quite clear that, no matter how big p is, if a has only small prime factors, we are saved.

Besides Proposition 11.3, the quadratic reciprocity formula comes accompanied by two other statements, which are very useful in practice. First, at $a=-1$, we have:

Proposition 11.6. *We have the following formula,*

$$\left(\frac{-1}{p}\right)=\begin{cases}1&\text{if }p=1(4)\\-1&\text{if }p=3(4)\end{cases}$$

solving in practice the equation $b^2=-1(p)$.

Proof. This follows from the Euler formula, which at $a=-1$ reads:

$$\left(\frac{-1}{p}\right)=(-1)^{\frac{p-1}{2}}(p)$$

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

As a second useful result, this time at $a=2$, we have:

Theorem 11.7. *We have the following formula,*

$$\left(\frac{2}{p}\right)=\begin{cases}1&\text{if }p=1,7(8)\\-1&\text{if }p=3,5(8)\end{cases}$$

solving in practice the equation $b^2=2(p)$.

Proof. This is actually a bit complicated. The Euler formula at $a=2$ gives:

$$\left(\frac{2}{p}\right)=2^{\frac{p-1}{2}}(p)$$

However, with more work, we have the following formula, which gives the result:

$$\left(\frac{2}{p}\right)=(-1)^{\frac{p^2-1}{8}}$$

We will be back to this later in this chapter, with a full proof for it. □

As a continuation of this, speaking Legendre symbol for small values of the upper variable, we can try to compute these for $a=\pm\,3,4,5,6,7,8,\ldots$ But by multiplicativity plus Proposition 11.6 plus Theorem 11.7 we are left with the case where $a=q$ is an odd prime, and we can solve the problem with quadratic reciprocity, so done.

Let us record however a few statements here, which can be useful in practice, and with this being mostly for illustration purposes, for Theorem 11.5. We first have:

Proposition 11.8. *We have the following formula,*

$$\left(\frac{3}{p}\right)=\begin{cases}1&\text{if }p=1,11(12)\\-1&\text{if }p=5,7(8)\end{cases}$$

valid for any prime $p\geq 5$.

Proof. By quadratic reciprocity, we have the following formula:

$$\left(\frac{3}{p}\right)=(-1)^{\frac{3-1}{2}\cdot\frac{p-1}{2}}\left(\frac{p}{3}\right)=(-1)^{\frac{p-1}{2}}\left(\frac{p}{3}\right)$$

Now since the sign depends on p modulo 4, and the symbol on the right depends on p modulo 3, we conclude that our symbol depends on p modulo 12, and the computation gives the formula in the statement. Finally, we have the following formula too:

$$\left(\frac{3}{p}\right)=(-1)^{\left[\frac{p+1}{6}\right]}$$

Indeed, the quantity on the right is something which depends on p modulo 12, and is in fact the simplest functional implementation of the formula in the statement. □

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

Proposition 11.9. *We have the following formula,*

$$\left(\frac{5}{p}\right)=\begin{cases}1 & \text{if } p=1,4(5)\\ -1 & \text{if } p=2,3(5)\end{cases}$$

valid for any odd prime $p\neq 5$.

Proof. By quadratic reciprocity, we have the following formula:

$$\left(\frac{5}{p}\right)=(-1)^{\frac{5-1}{2}\cdot\frac{p-1}{2}}\left(\frac{p}{5}\right)=\left(\frac{p}{5}\right)$$

Thus, we have the result. Alternatively, we have the following formula:

$$\left(\frac{5}{p}\right)=(-1)^{\left[\frac{2p+2}{5}\right]}$$

Indeed, this is the simplest implementation of the formula in the statement. □

Moving ahead now, we have the following interesting generalization of the Legendre symbol, to the case of denominators not necessarily prime, due to Jacobi:

Theorem 11.10. *The theory of Legendre symbols can be extended by multiplicativity into a theory of Jacobi symbols, according to the formula*

$$\left(\frac{a}{p_1^{s_1}\ldots p_k^{s_k}}\right)=\left(\frac{a}{p_1}\right)^{s_1}\ldots\left(\frac{a}{p_k}\right)^{s_k}$$

with the denominator being not necessarily prime, but just an arbitrary odd number, and this theory has as results those imported from the Legendre theory.

Proof. This is something self-explanatory, and we will leave listing the basic properties of the Jacobi symbols, based on the theory of Legendre symbols, as an exercise. □

The story is not over with Jacobi, because the denominator there is still odd, and positive. So, we have a problem to be solved, the solution to it being as follows:

THEOREM 11.11. *The theory of Jacobi symbols can be further extended into a theory of Kronecker symbols, according to the formula*

$$\left(\frac{a}{\pm p_1^{s_1}\ldots p_k^{s_k}}\right)=\left(\frac{a}{\pm 1}\right)\left(\frac{a}{p_1}\right)^{s_1}\ldots\left(\frac{a}{p_k}\right)^{s_k}$$

with the denominator being an arbitrary integer, via suitable values for

$$\left(\frac{a}{2}\right)\quad,\quad\left(\frac{a}{-1}\right)\quad,\quad\left(\frac{a}{0}\right)$$

and this theory has as results those imported from the Jacobi theory.

PROOF. Unlike the extension from Legendre to Jacobi, which was something straightforward, here we have some work to be done, in order to figure out the correct values of the 3 symbols in the statement. The answer for the first symbol is as follows:

$$\left(\frac{a}{2}\right)=\begin{cases}1&\text{if }a=\pm1(8)\\0&\text{if }a=0(2)\\-1&\text{if }a=\pm3(8)\end{cases}$$

The answer for the second symbol is as follows:

$$\left(\frac{a}{-1}\right)=\begin{cases}1&\text{if }a\geq0\\-1&\text{if }a<0\end{cases}$$

As for the answer for the third symbol, this is as follows:

$$\left(\frac{a}{0}\right)=\begin{cases}1&\text{if }a=\pm1\\0&\text{if }a\neq\pm1\end{cases}$$

And we will leave this as an instructive exercise, to figure out what the puzzle exactly is, and why these are the correct answers. And for an even better exercise, cover with a cloth the present proof, and try to figure out everything by yourself. □

11d. Gauss sums

Going back to what we learned so far about the Legendre symbols, there were several mysterious things there, that we will attempt to elucidate now.

Let us start with the $a=2$ case. The result here is as follows:

Theorem 11.12. *We have the following formula,*

$$\left(\frac{2}{p}\right)=\begin{cases}1&\text{if } p=1,7(8)\\-1&\text{if } p=3,5(8)\end{cases}$$

solving in practice the equation $b^2=2(p)$.

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

(1) As a first observation, the Euler formula at $a=2$ is as follows, obviously well below the quality of the very precise formula in the statement:

$$\left(\frac{2}{p}\right)=2^{\frac{p-1}{2}}(p)$$

As a second observation, the quadratic reciprocity formula, assuming that known, cannot help either, because in that formula $p,q\geq 3$ are odd primes.

(2) Thus, we must prove the result. As already mentioned before, the proof will come via the following formula, which is equivalent to the formula in the statement:

$$\left(\frac{2}{p}\right)=(-1)^{\frac{p^2-1}{8}}$$

Finally, let us mention too that, despite 2 being an even prime, the problematics here is a bit similar to the one of the quadratic reciprocity formula, and the proof below will contain many good ideas, that we will use later in the proof of quadratic reciprocity.

(3) Getting started now, let us set $w=e^{\pi i/4}$, so that $w^2=i$, do not ask me why, and then $t=w+w^{-1}$. We have of course $t=\sqrt{2}$, but it is better to forget this, and do formal arithmetics instead, with integers as scalars, based on the following computation:

$$\begin{aligned}t^2&=2+w^2+w^{-2}\\&=2+i-i\\&=2\end{aligned}$$

Now by using the Euler formula for the Legendre symbol, we have:

$$\begin{aligned}\left(\frac{2}{p}\right)&=2^{\frac{p-1}{2}}\ (p)\\&=(t^2)^{\frac{p-1}{2}}\ (p)\\&=t^{p-1}\ (p)\end{aligned}$$

(4) By multiplying now by t we obtain from this, in a formal sense, and I will leave it you to clarify all the details here, namely what this formal sense exactly means:

$$\left(\frac{2}{p}\right)t=t^p\ (p)$$

(5) On the other hand, by using the binomial formula, and the standard fact that all non-trivial binomial coefficients are multiples of p, we obtain, again formally:

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

(6) Now let us look at $w^p + w^{-p}$, as usual complex number. Since $w = e^{\pi i/4}$, this quantity will depend only on p modulo 8, and more precisely, we have:

$$w^p + w^{-p} = \begin{cases} w + w^{-1} & \text{if } p = \pm 1(8) \\ -w - w^{-1} & \text{if } p = \pm 3(8) \end{cases}$$

Thus $w^p + w^{-p} = \pm t$, with the sign depending on p modulo 8, and more specifically:

$$w^p + w^{-p} = (-1)^{\frac{p^2-1}{8}} t$$

(7) Time now to put everything together. By combining (4,5,6) we obtain:

$$\left(\frac{2}{p}\right) t = (-1)^{\frac{p^2-1}{8}} t \ (p)$$

By dividing by t, this gives the following formula:

$$\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}} \ (p)$$

But the mod p symbol can now be dropped, because our equality is between two ±1 quantities, and we obtain the formula in the statement. □

With the same idea, we can prove as well the quadratic reciprocity theorem:

THEOREM 11.13. *We have the quadratic reciprocity formula*

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\cdot\frac{q-1}{2}}$$

valid for any primes $p, q \geq 3$.

PROOF. This is something already advertised in the above, and we refer to the discussion there for the mighty power of this formula, and its enigmatic nature. However, thinking a bit, our $t = w + w^{-1}$ trick above can be adapted, as follows:

(1) To start with, we need an analogue of that $t = w + w^{-1}$ variable. For this purpose, let us set $w = e^{2\pi i/q}$, now that we have a prime $q \geq 3$ involved, and then:

$$t = \sum_{k=0}^{q-1} w^{k^2}$$

Observe that at $q = 2$, excluded by the statement, we have $w = -1$, and so $t = 1 + (-1) = 0$, instead of the $t = w + w^{-1}$ with $w = e^{\pi i/4}$ used before. However, believe me, this is due to some bizarre reasons, and the above t is the good variable, at $q \geq 3$.

(2) The above variable t is called Gauss sum, can be defined for any $q \in \mathbb{N}$, not necessarily prime, and can be explicitely computed, the formula being as follows:

$$t = \begin{cases} \sqrt{q} & \text{if } q = 1(4) \\ 0 & \text{if } q = 2(4) \\ \sqrt{q}\, i & \text{if } q = 3(4) \\ \sqrt{q}(1+i) & \text{if } q = 0(4) \end{cases}$$

In particular, assuming that q is odd, as is our $q \geq 3$ prime, we have:

$$t^2 = \begin{cases} q & \text{if } q = 1(4) \\ -q & \text{if } q = 3(4) \end{cases}$$

(3) In what follows we will only need this latter formula, for $q \geq 3$ prime, so let us prove this now, and with the comment that the proof of the first formula in (2) is something quite complicated, and better avoid that. We have, by definition of our variable t:

$$\begin{aligned} |t|^2 &= \sum_{kl} w^{k^2-l^2} \\ &= \sum_{kl} w^{(k+l)(k-l)} \\ &= \sum_{lr} w^{r(2l+r)} \\ &= \sum_r w^{r^2} \sum_l (w^{2r})^l \\ &= q \end{aligned}$$

(4) On the other hand, it is easy to see that t^2 is real, so $t^2 = \pm q$. With a bit more work it is possible to compute the sign too, $t^2 = (-1)^{\frac{q-1}{2}} q$, but we will not need this here, because the sign will come for free at the end of the proof, via a symmetry argument. So, as a conclusion, we have a formula as follows, for a certain $e_q \in \{0, 1\}$:

$$t^2 = (-1)^{e_q} q$$

(5) With this done, let us turn to the proof of our theorem, by using the variable t a bit as before, in the proof of Theorem 11.12. By using the Euler formula, we have:

$$\left(\frac{t^2}{p}\right) = (t^2)^{\frac{p-1}{2}} \ (p) = t^{p-1} \ (p)$$

By multiplying now by t we obtain from this, in a formal sense:

$$\left(\frac{t^2}{p}\right)t=t^p\ (p)$$

(6) In order to compute now t^p by other means, observe first that, if we denote by $\mathbb{Z}_q-\{0\}=S\sqcup N$ the partition into squares and non-squares, we have:

$$\begin{aligned}
t &= \sum_{k=0}^{q-1}w^{k^2}\\
&= 1+2\sum_{s\in S}w^s\\
&= \sum_{s\in S}w^s-\sum_{s\in N}w^s\\
&= \sum_{r=0}^{k-1}\left(\frac{r}{q}\right)w^r
\end{aligned}$$

(7) By using now the multinomial formula, with the observation that all the non-trivial multinomial coefficients are multiples of p, we obtain, in a formal sense:

$$\begin{aligned}
t^p &= \left(\sum_r\left(\frac{r}{q}\right)w^r\right)^p\\
&= \sum_r\left(\frac{r}{q}\right)w^{rp}\ (p)\\
&= \sum_s\left(\frac{p^{-1}s}{q}\right)w^s\ (p)\\
&= \left(\frac{p^{-1}}{q}\right)\sum_s\left(\frac{s}{q}\right)w^s\ (p)\\
&= \left(\frac{p}{q}\right)t\ (p)
\end{aligned}$$

(8) Time now to put everything together. By combining (5,7) we obtain:

$$\left(\frac{t^2}{p}\right)t=\left(\frac{p}{q}\right)t\ (p)$$

We can divide by t, and then drop the modulo p symbol, because our new equality, without t, is between two ±1 quantities, and we obtain:

$$\left(\frac{t^2}{p}\right)=\left(\frac{p}{q}\right)$$

Now by taking into account the formula found in (4), this reads:

$$\left(\frac{(-1)^{e_q}}{p}\right)\left(\frac{q}{p}\right)=\left(\frac{p}{q}\right)$$

By using the Euler formula for the symbol on the left, we obtain from this:

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right)=(-1)^{\frac{p-1}{2}\cdot e_q}$$

Now by symmetry we must have $e_q=\frac{q-1}{2}$, and this finishes the proof. □

We have seen in the above that the quadratic reciprocity theorem can be established via Gauss sums t, and this is certainly excellent news. However, we have mentioned in step (2) of our proof above a very nice, powerful and final formula for the Gauss sum t itself, and this even in the general case, where $q\in\mathbb N$ is not necessarily prime.

Time now to discuss all this. So, we want to solve the following question:

Question 11.14. *What is the value of the Gauss quadratic sum*

$$t=\sum_{k=0}^{q-1}w^{k^2}$$

where $w=e^{2\pi i/q}$, with $q\in\mathbb N$?

Let us begin with some experiments, at small values of q. We have here:

Proposition 11.15. *The first few Gauss sums are as follows:*

(1) *At $q=1$ we have $t=1$.*
(2) *At $q=2$ we have $t=0$.*
(3) *At $q=3$ we have $t=\sqrt{3}\,i$.*
(4) *At $q=4$ we have $t=2(1+i)$.*
(5) *At $q=5$ we have $t=\sqrt{5}$.*
(6) *At $q=6$ we have $t=0$.*
(7) *At $q=7$ we have $t=\sqrt{7}\,i$.*
(8) *At $q=8$ we have $t=2\sqrt{2}(1+i)$.*

Proof. The computations are as follows, with $w=e^{2\pi i/q}$:

(1) At $q=1$ we have $w=1$, and $t=1$.

(2) At $q=2$ we have $w=-1$, and $t=1+(-1)=0$

(3) At $q=3$ we have $w=e^{2\pi i/3}$, and the computation goes as follows:

$$\begin{aligned}
t&=1+w+w^4\\
&=1+2w\\
&=1+2\left(-\frac{1}{2}+\frac{\sqrt{3}}{2}\,i\right)\\
&=\sqrt{3}\,i
\end{aligned}$$

(4) At $q=4$ we have $w=i$, and the computation goes as follows:

$$\begin{aligned}
t&=1+i+i^4+i^9\\
&=1+i+1+i\\
&=2+2i\\
&=2(1+i)
\end{aligned}$$

(5) At $q=5$ we have $w=e^{2\pi i/5}$, and the computation goes as follows:

$$\begin{aligned}
t&=1+w+w^4+w^9+w^{16}\\
&=1+w+w^4+w^4+w\\
&=1+2(w+w^4)\\
&=1+4\cos\left(\frac{2\pi}{5}\right)\\
&=\sqrt{5}
\end{aligned}$$

Here we have used some crazy trigonometry at the end, which can be avoided, or rather proved, when thinking well, at where this trigonometry comes from, as follows:

$$\begin{aligned}
t^2&=(1+2w+2w^4)^2\\
&=1+4w^2+4w^3+4w+4w^4+8\\
&=5+4(1+w+w^2+w^3+w^4)\\
&=5
\end{aligned}$$

Observe that there is actually still some work to be done here, when extracting the square root of $t^2=5$. But the picture shows that the root is positive, $t=\sqrt{5}$.

(6) At $q=6$ it is most convenient to use $w=e^{2\pi i/3}$ as variable, as it is customary, and with this convention our root of unity is $e^{2\pi i/6}=-w^2$, and we have:

$$\begin{aligned}
t&=1-w^2+w^8-w^{18}+w^{32}-w^{50}\\
&=1-w^2+w^2-1+w^2-w^2\\
&=0
\end{aligned}$$

(7) At $q=7$ we have $w=e^{2\pi i/7}$, and the computation goes as follows:

$$\begin{aligned}
t&=1+w+w^4+w^9+w^{16}+w^{25}+w^{36}\\
&=1+w+w^4+w^2+w^2+w^4+w\\
&=1+2(w+w^2+w^4)\\
&=\sqrt{7}\,i
\end{aligned}$$

Here again we have used some crazy trigonometry, the justification being as follows, and with the correct root of $t^2=-7$, among $t=\pm\sqrt{7}\,i$, being $t=\sqrt{7}\,i$, as shown by the picture, with the components w,w^2,w^4 of our sum t tending to lie North-West:

$$\begin{aligned}
t^2&=(1+2w+2w^2+2w^4)^2\\
&=1+4w^2+4w^4+4w\\
&\quad+4w+4w^2+4w^4\\
&\quad+8w^3+8w^5+8w^6\\
&=1+8(w+w^2+w^3+w^4+w^5+w^6)\\
&=-7+8(1+w+w^2+w^3+w^4+w^5+w^6)\\
&=-7
\end{aligned}$$

(8) At $q=8$ we have $w=e^{\pi i/4}$, and the computation goes as follows:

$$\begin{aligned}
t&=1+w+w^4+w^9+w^{16}+w^{25}+w^{36}+w^{49}\\
&=1+w-1+w+1+w-1+w\\
&=4w\\
&=2\sqrt{2}(1+i)
\end{aligned}$$

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

All the above is quite interesting, and we can formulate our conclusion as follows:

Conclusion 11.16. *The first few quadratic Gauss sums are given by*

q	1	2	3	4	5	6	7	8
t	1	0	$\sqrt{3}\,i$	$2(1+i)$	$\sqrt{5}$	0	$\sqrt{7}\,i$	$2\sqrt{2}(1+i)$

with everything coming from easy algebra, except for the signs.

Moving ahead now with the general case, there is some obvious periodicity in the above table, of order 4, and with everything working fine, I mean with the dependence on q being clear in all cases modulo 4, we are led to the following statement:

Theorem 11.17. *We have the following formula for the Gauss sums,*

$$t=\begin{cases}\sqrt{q}&\text{if }q=1(4)\\0&\text{if }q=2(4)\\\sqrt{q}\,i&\text{if }q=3(4)\\\sqrt{q}(1+i)&\text{if }q=0(4)\end{cases}$$

valid for any $q\in\mathbb{N}$, not necessarily prime.

Proof. This is straightforward, except for that signs, the idea being as follows:

(1) To start with, let us compute $|t|^2$. This is something that we did in the proof of Theorem 11.13, for $q\geq 3$ prime, and the computation there can be recycled, as follows:

$$\begin{aligned}|t|^2&=\sum_{kl}w^{k^2-l^2}=\sum_{kl}w^{(k+l)(k-l)}\\&=\sum_{lr}w^{r(2l+r)}=\sum_r w^{r^2}\sum_l(w^{2r})^l\\&=\sum_r w^{r^2}\times\delta_{2|2r}\,q=q\sum_{q|2r}w^{r^2}\end{aligned}$$

(2) We have some cases here. For q odd we get 0, and for q even, we have:

$$\begin{aligned}|t|^2&=q(1+(w^{(q/2)^2})\\&=q(1+(w^{q/2})^{q/2}\\&=q(1+(-1)^{q/2})\end{aligned}$$

(3) We are therefore led to the following formula, for our variable $|t|^2$:

$$|t|^2=\begin{cases}q&\text{if }q=1(4)\\0&\text{if }q=2(4)\\q&\text{if }q=3(4)\\2q&\text{if }q=0(4)\end{cases}$$

(4) Now by extracting the square root, we have the following formula, for $|t|$:

$$|t|=\begin{cases}\sqrt{q}&\text{if }q=1(4)\\0&\text{if }q=2(4)\\\sqrt{q}&\text{if }q=3(4)\\\sqrt{2q}&\text{if }q=0(4)\end{cases}$$

(5) The question is now, shall we go ahead and compute t, or be less greedy, and compute t^2 first. And let us be modest, of course, and go with t^2 first. But here, it is pretty much clear, from the computations in the proof of Proposition 11.15, that we can

get away with some simple algebra, I mean with algebra a hair more complicated than that in (1,2) above. For this purpose, the best is to go with the following alternative definition of the Gauss sums, that we already met in the proof of Theorem 11.13:

$$t=\sum_{r=0}^{q-1}\left(\frac{r}{q}\right)w^r$$

(6) Now by taking the square of this quantity, and then working out what exactly happens at $q=1,2,3,0(4)$, exactly as in the proof of Proposition 11.15, and we will leave this as an instructive exercise, we are led to the following formula:

$$t^2=\begin{cases}q&\text{if }q=1(4)\\0&\text{if }q=2(4)\\-q&\text{if }q=3(4)\\2qi&\text{if }q=0(4)\end{cases}$$

(7) In what regards now t itself, by taking the square root, we must have:

$$t=\begin{cases}\pm\sqrt{q}&\text{if }q=1(4)\\0&\text{if }q=2(4)\\\pm\sqrt{q}\,i&\text{if }q=3(4)\\\pm\sqrt{q}(1+i)&\text{if }q=0(4)\end{cases}$$

(8) So, almost done, but thinking a bit, in fact we just got started. Indeed, remember from Proposition 11.15 that the computation of the signs is tricky business, done on pictures, more specifically at $q=5$ by arguing that the components of t tend to pull it East, and at $q=7$, by arguing that these components tend to pull it North-West.

(9) So, what kind of question is this, what we have left, geography or something? Well, in answer, such things are called mathematical analysis. Obviously, what we need are some estimates, with ε and everything, as to decide what is the approximate direction of the pull of the components of t, as to compute that missing sign.

(10) And we will stop here, with my apologies to you, and to mathematics well done, in general. For the story, Gauss himself struggled quite a bit with this question, and there have been countless other victims, afterwards. Incuding myself, once I got into this, in my research, not realized that this is the Gauss sign, that I'm looking for, and spent a few days with it, with the conclusion that the question is guaranteed undoable. □

So, this was for the story of Gauss sums. Still a gap left, but we will have the whole Part IV of this book for doing analysis, remind me at that time about that sign.

11e. Exercises

Exercises:

Exercise 11.18.

Exercise 11.19.

Exercise 11.20.

Exercise 11.21.

Exercise 11.22.

Exercise 11.23.

Exercise 11.24.

Exercise 11.25.

Bonus exercise.

CHAPTER 12

Polynomials, roots

12a. Polynomials, roots

We have seen that many number theory questions lead us into computing roots of polynomials $P \in \mathbb{Q}[X]$. We will investigate here such questions, with a detailed study of the arbitrary polynomials $P \in \mathbb{C}[X]$, and their roots, often by using analytic methods.

Let us start with something that we know well, but is always good to remember:

Theorem 12.1. *The solutions of $ax^2+bx+c=0$ with $a,b,c \in \mathbb{C}$ are*

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

with the square root of complex numbers being defined as $\sqrt{re^{it}}=\sqrt{r}e^{it/2}$.

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

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

Here we have used the fact, mentioned in the statement, that any complex number $z=re^{it}$ has indeed a square root, given by $\sqrt{z}=\sqrt{r}e^{it/2}$, plus in fact a second square root as well, namely $-\sqrt{z}$. Thus, we are led to the conclusion in the statement. □

Very nice all this, and you would probably say that the story is over here, with degree 2. However, not really. Here are a few tricks, in order to deal with degree 2 questions:

Tricks 12.2. *The following happen:*

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

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

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

In order to find the eigenvalues r, s, you are certainly very used to compute the characteristic polynomial, then apply Theorem 12.1. But my point is that this characteristic polynomial is of the form $x^2 - ax + b$, with $a = Tr(a)$ and $b = \det A$, so we can normally apply the trick in (1), provided of course that r, s are simple numbers, say integers.

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

Warning 12.3. *The above tricks work in pure mathematics, where the numbers r, s that we can meet are usually integers, or rationals. In applied mathematics, however, the numbers that we meet are integers or rationals with probability $P = 0$, so no tricks.*

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

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

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

Theorem 12.4. *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 12.5. *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 12.6. *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 12.4 and 12.5. □

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 will compute many other resultants, in what follows.

12b. The discriminant

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

Theorem 12.7. *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 12.5, 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 12.6:

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

Also, the fact that we have $\Delta(P) = 0$ precisely when P has a double root is clear from Theorem 12.5. 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 12.6, 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 12.6, and we obtain:

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

We will be back later to such formulae, in degree 3, and in degree 4 as well, with the comment however, coming in advance, that these formulae are not very beautiful.

At the theoretical level now, we have the following result, which is not trivial:

THEOREM 12.8. *The discriminant of a polynomial P is given by the formula*

$$\Delta(P) = a^{2N-2}\prod_{i<j}(r_i - r_j)^2$$

where a is the leading coefficient, and $r_1, \ldots, r_N$ are the roots.

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

(1) The first thought goes to the formula in Theorem 12.5, so let us see what that formula teaches us, in the case $Q = P'$. Let us write P, P' as follows:

$$P = a(x - r_1)\ldots(x - r_N)$$

$$P' = Na(x - p_1)\ldots(x - p_{N-1})$$

According to Theorem 12.5, the resultant of P,P' is then given by:

$$R(P,P')=a^{N-1}(Na)^N\prod_{ij}(r_i-p_j)$$

And bad news, this is not exactly what we wished for, namely the formula in the statement. That is, we are on the good way, but certainly have to work some more.

(2) Obviously, we must get rid of the roots $p_1,\ldots,p_{N-1}$ of the polynomial P'. In order to do this, let us rewrite the formula that we found in (1) in the following way:

$$\begin{aligned}
R(P,P') &= N^Na^{2N-1}\prod_i\left(\prod_j(r_i-p_j)\right)\\
&= N^Na^{2N-1}\prod_i\frac{P'(r_i)}{Na}\\
&= a^{N-1}\prod_i P'(r_i)
\end{aligned}$$

(3) In order to compute now P', and more specifically the values $P'(r_i)$ that we are interested in, we can use the Leibnitz rule. So, consider our polynomial:

$$P(x)=a(x-r_1)\ldots(x-r_N)$$

The Leibnitz rule for derivatives tells us that $(fg)'=f'g+fg'$, but then also that $(fgh)'=f'gh+fg'h+fgh'$, and so on. Thus, for our polynomial, we obtain:

$$P'(x)=a\sum_i(x-r_1)\ldots\underbrace{(x-r_i)}_{missing}\ldots(x-r_N)$$

Now when applying this formula to one of the roots r_i, we obtain:

$$P'(r_i)=a(r_i-r_1)\ldots\underbrace{(r_i-r_i)}_{missing}\ldots(r_i-r_N)$$

By making now the product over all indices i, this gives the following formula:

$$\prod_i P'(r_i)=a^N\prod_{i\neq j}(r_i-r_j)$$

(4) Time now to put everything together. By taking the formula in (2), making the normalizations in Theorem 12.7, and then using the formula found in (3), we obtain:

$$\begin{aligned}
\Delta(P) &= (-1)^{\binom{N}{2}}a^{N-2}\prod_i P'(r_i)\\
&= (-1)^{\binom{N}{2}}a^{2N-2}\prod_{i\neq j}(r_i-r_j)
\end{aligned}$$

(5) This is already a nice formula, which is very useful in practice, and that we can safely keep as a conclusion, to our computations. However, we can do slightly better, by grouping opposite terms. Indeed, this gives the following formula:

$$\begin{aligned}\Delta(P) &= (-1)^{\binom{N}{2}}a^{2N-2}\prod_{i\neq j}(r_i-r_j)\\ &= (-1)^{\binom{N}{2}}a^{2N-2}\prod_{i<j}(r_i-r_j)\cdot\prod_{i>j}(r_i-r_j)\\ &= (-1)^{\binom{N}{2}}a^{2N-2}\prod_{i<j}(r_i-r_j)\cdot(-1)^{\binom{N}{2}}\prod_{i<j}(r_i-r_j)\\ &= a^{2N-2}\prod_{i<j}(r_i-r_j)^2\end{aligned}$$

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

As applications now, the formula in Theorem 12.8 is quite useful for the real polynomials $P\in\mathbb{R}[X]$ in small degree, because it allows to say when the roots are real, or complex, or at least have some partial information about this. For instance, we have:

PROPOSITION 12.9. *Consider a polynomial with real coefficients, $P\in\mathbb{R}[X]$, assumed for simplicity to have nonzero discriminant, $\Delta\neq 0$.*

(1) *In degree 2, the roots are real when $\Delta>0$, and complex when $\Delta<0$.*
(2) *In degree 3, all roots are real precisely when $\Delta>0$.*

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

(1) The first assertion is something that you certainly know, coming from Theorem 12.1, but let us see how this comes via the formula in Theorem 12.8, namely:

$$\Delta(P)=a^{2N-2}\prod_{i<j}(r_i-r_j)^2$$

In degree $N=2$, this formula looks as follows, with r_1,r_2 being the roots:

$$\Delta(P)=a^2(r_1-r_2)^2$$

Thus $\Delta>0$ amounts in saying that we have $(r_1-r_2)^2>0$. Now since r_1,r_2 are conjugate, and with this being something trivial, meaning no need here for the computations in Theorem 12.1, we conclude that $\Delta>0$ means that r_1,r_2 are real, as stated.

(2) In degree $N=3$ now, we know from analysis that P has at least one real root, and the problem is whether the remaining 2 roots are real, or complex conjugate. For this purpose, we can use the formula in Theorem 12.8, which in degree 3 reads:

$$\Delta(P)=a^4(r_1-r_2)^2(r_1-r_3)^2(r_2-r_3)^2$$

We can see that in the case $r_1, r_2, r_3 \in \mathbb{R}$, we have $\Delta(P) > 0$. Conversely now, assume that $r_1 = r$ is the real root, coming from analysis, and that the other roots are $r_2 = z$ and $r_3 = \bar{z}$, with z being a complex number, which is not real. We have then:

$$\begin{aligned}\Delta(P) &= a^4(r-z)^2(r-\bar{z})^2(z-\bar{z})^2\\ &= a^4|r-z|^4(2iIm(z))^2\\ &= -4a^4|r-z|^4Im(z)^2\\ &< 0\end{aligned}$$

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

In relation with the above, for our result to be truly useful, we must of course compute the discriminant in degree 3. We will do this, along with applications, right next.

12c. Cardano formula

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

THEOREM 12.10. *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 12.5 and Theorem 12.6, 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 12.5. 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 12.6, 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 12.6 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 12.7, 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 12.11. *For a normalized degree* 3 *equation, namely*

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

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

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

is a real solution of our equation.

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

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

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

Regarding the other roots, we know from Proposition 12.9 that these are both real when $\Delta<0$, and complex conjugate when $\Delta<0$. Thus, in the context of Theorem 12.11, the other two roots are complex conjugate, the formula for them being as follows:

Proposition 12.12. *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 12.13. *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 12.11 and Proposition 12.12. Alternatively, we can redo the computation in their proof, which was nearly identical anyway, in the present setting, with x being given by the above formula, by using $w^3=1$. □

As a comment here, the formula in Theorem 12.13 holds of course in the case $\Delta>0$ too, and also when the coefficients are complex numbers, $p,q\in\mathbb{C}$, and this due to the fact that the proof rests on the nearly trivial computation from the proof of Theorem 12.11, or of Proposition 12.12. However, these extensions are quite often not very useful, because when it comes to extract all the above square and cubic roots, for complex numbers, you can well end up with the initial question, the one that you started with.

12d. Higher degree

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

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

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

Proof. The formula of Δ follows from the definition of the discriminant, from Theorem 12.7, with the resultant computed via Theorem 12.6, 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 12.14:

Proposition 12.15. *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 12.14, 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 12.14. □

We still have some work to do. Indeed, looking back at what we did in degree 3, the passage there from Theorem 12.10 to Theorem 12.11 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 12.15, 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 12.16. *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 12.15, 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 12.14. □

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

Theorem 12.17. *The roots of a normalized degree* 4 *equation, written as*

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

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

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

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

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

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

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

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

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

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

The idea will be that of adding a suitable common term, to both sides, as to make square on both sides, as to eventually end with a sort of double quadratic equation. For this purpose, our claim is that what we need is a number y satisfying:

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

Indeed, assuming that we have this number y, our equation becomes:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We still have to compute the number y appearing in the above via Cardano, and the result here, adding to what we already have in Theorem 12.17, is as follows:

Theorem 12.18 (continuation). *The value of y in the previous theorem is*

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

where the number t is given by the formula

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

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

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

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

With this in hand, we have the following computation:

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

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

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

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

In degree 5 and more, things become fairly complicated, and we have:

Theorem 12.19. *There is no general formula for the roots of polynomials of degree $N=5$ and higher, with the reason for this, coming from Galois theory, being that the group S_5 is not solvable. The simplest numeric example is $P=X^5-X-1$.*

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

(1) Given a field F, assume that the roots of $P\in F[X]$ can be computed by using iterated roots, a bit as for the degree 2 equation, or the degree 3 and 4 equations. Then, algebrically speaking, this gives rise to a tower of fields as folows, with $F_0=F$, and each F_{i+1} being obtained from F_i by adding a root, $F_{i+1}=F_i(x_i)$, with $x_i^{n_i}\in F_i$:

$$F_0\subset F_1\subset\ldots\subset F_k$$

(2) In order for Galois theory to apply to this situation, we must make all the extensions normal, which amounts in replacing each $F_{i+1}=F_i(x_i)$ by its extension $K_i(x_i)$, with K_i extending F_i by adding a n_i-th root of unity. Thus, with this replacement, we can assume that the tower in (1) in normal, meaning that all Galois groups are cyclic.

(3) Now by Galois theory, at the level of the corresponding Galois groups we obtain a tower of groups as follows as follows, which is a resolution of the last group G_k, the Galois group of P, in the sense of group theory, in the sense that all quotients are cyclic:

$$G_1\subset G_2\subset\ldots\subset G_k$$

As a conclusion, Galois theory tells us that if the roots of a polynomial $P\in F[X]$ can be computed by using iterated roots, then its Galois group $G=G_k$ must be solvable.

(4) In the generic case, the conclusion is that Galois theory tells us that, in order for all polynomials of degree 5 to be solvable, via square roots, the group S_5, which appears there as Galois group, must be solvable, in the sense of group theory. But this is wrong, because the alternating subgroup $A_5\subset S_5$ is simple, and therefore not solvable.

(5) Finally, regarding the polynomial $P=X^5-X-1$, some elementary computations here, based on arithmetic over $\mathbb{F}_2,\mathbb{F}_3$, and involving various cycles of length $2,3,5$, show that its Galois group is S_5. Thus, we have our counterexample.

(6) Finally, let us mention that all this shows as well that a random polynomial of degree 5 or higher is not solvable by square roots, and with this being an elementary consequence of the main result from (4), via some standard analysis arguments. □

There is a lot of further interesting theory that can be developed here, following Galois and others. For more on all this, we recommend any number theory book.

12e. Exercises

Exercises:

Exercise 12.20.

Exercise 12.21.

Exercise 12.22.

Exercise 12.23.

Exercise 12.24.

Exercise 12.25.

Exercise 12.26.

Exercise 12.27.

Bonus exercise.

Part IV

Functions

Dancing like there's no one there
Before she ever seemed to care
Now she wouldn't dare
It's so rock and roll to be alone

CHAPTER 13

Functions

13a. Functions, 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 13.1. *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 13.2. *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 13.1. So, let us pick $\varepsilon>0$. We want in the end to have something of type $|f(g(x))-f(g(y))|<\varepsilon$, so we must first use that ε,δ condition for the function f. So, let us start in this way. Since f is continuous at $g(x)$, we can find $\delta>0$ such that:

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

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

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

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

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

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

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

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

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

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

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

Theorem 13.3. *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 13.2 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 13.2.

(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:

$$\begin{aligned}\lim_{y\to0}\cos(x+y) &= \lim_{y\to0}(\cos x\cos y-\sin x\sin y)\\ &= \cos x\lim_{y\to0}\cos y-\sin x\lim_{y\to0}\sin y\\ &= \cos x\cdot 1-\sin x\cdot 0\\ &= \cos x\end{aligned}$$

(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 13.2. □

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

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

Proposition 13.4. *If a function $f:X\to\mathbb{R}$ has the Lipschitz property*

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

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

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

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

Definition 13.5. *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 13.6. *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. □

13b. Intermediate values

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

Definition 13.7. *The open and closed sets are defined as follows:*

(1) *Open means that there is a small interval around each point.*
(2) *Closed means that our set is closed under taking limits.*

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

Proposition 13.8. *A set $O\subset\mathbb{R}$ is open precisely when its complement $C\subset\mathbb{R}$ is closed, and vice versa.*

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

" $\implies$ " Assume that $O\subset\mathbb{R}$ is open, and let $C=\mathbb{R}-O$. In order to prove that C is closed, assume that $\{x_n\}_{n\in\mathbb{N}}\subset C$ converges to $x\in\mathbb{R}$. We must prove that $x\in C$, and 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. Getting now back to functions, we have:

Theorem 13.9. *A function is continuous precisely when $f^{-1}(O)$ is open, for any O open. Equivalently, $f^{-1}(C)$ must be closed, for any C closed.*

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

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

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

As a test for the above criterion, let us reprove the fact, that we know from Theorem 13.2, that if f,g are continuous, so is $f\circ g$. But this is clear, coming from:

$$(f\circ g)^{-1}(O)=g^{-1}(f^{-1}(O))$$

In short, not bad, because at least in relation with this specific problem, our proof using open sets is as simple as the simplest proof, namely the one using $f(x_n)\to f(x)$, and is simpler than the other proof that we know, namely the one with ε,δ.

In order to reach to true applications of Theorem 13.9, we will need to know more about the open and closed sets. Let us begin with a useful result, as follows:

Proposition 13.10. *The following happen:*

1. *Union of open sets is open.*
2. *Intersection of closed sets is closed.*
3. *Finite intersection of open sets is open.*
4. *Finite union of closed sets is closed.*

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

$$\left(\bigcup_iE_i\right)^c=\bigcap_iE_i^c\quad,\quad\left(\bigcap_iE_i\right)^c=\bigcup_iE_i^c$$

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 13.11. *The open and closed sets are as follows:*

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

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

(1) We know that the open intervals are those of type (a,b) with $a<b$, with the values $a,b=\pm\infty$ allowed, and by Proposition 13.10 a union of such intervals is open.

(2) Conversely, given $O\subset\mathbb R$ open, we can cover each point $x\in O$ with an open interval $I_x\subset O$, and we have $O=\cup_xI_x$, 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 $O=\cup_xI_x$ 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 13.8. □

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

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

Definition 13.12. *The compact and connected sets are defined as follows:*

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

As basic examples, the closed bounded intervals $[a,b]$ are compact, as we know from the proof of Theorem 13.6, and so are the finite unions of such intervals. As for connected sets, the basic examples here are the various types of intervals, namely (a,b), $(a,b]$, $[a,b)$, $[a,b]$, and it looks impossible to come up with more examples. In fact, we have:

Theorem 13.13. *The compact and connected sets are as follows:*

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

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

(1) The fact that compact implies both closed and bounded is clear from our definition of compactness, because assuming non-closedness or non-boundedness leads to an open cover having no finite subcover. As for the converse, we know from the proof of Theorem 13.6 that any closed bounded interval $[a,b]$ is compact, and it follows that any $K\subset\mathbb R$ closed and bounded is a closed subset of a compact set, which follows to be compact.

(2) This is something which is obvious, and this regardless of what "cannot be broken into parts" in Definition 13.12 exactly means, mathematically speaking, with several possible definitions being possible here, all being equivalent. Indeed, $E \subset \mathbb{R}$ having this property is equivalent to $a, b \in E \implies [a, b] \subset E$, and this gives the result. □

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

Theorem 13.14. *Assuming that f is continuous:*

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

Proof. These assertions both follow from our definition of compactness and connectedness, as formulated in Definition 13.12. To be more precise:

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

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

Let us record as well the following useful generalization of Theorem 13.6:

Theorem 13.15. *Any continuous function defined on a compact set*

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

is automatically uniformly continuous.

Proof. We can prove this as Theorem 13.6, by using the compactness of X. □

You might perhaps ask at this point, were Theorems 13.14 and 13.15 worth all this excursion into open and closed sets. Good point, and here is our answer, a beautiful and powerful theorem based on the above, which can be used for a wide range of purposes:

Theorem 13.16. *The following happen for a continuous function $f : [a, b] \to \mathbb{R}$:*

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

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

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

Theorem 13.17. *Assuming that a function f is continuous and invertible, this function must be monotone, and its inverse function f^{-1} must be monotone and continuous too. Moreover, this statement holds both locally, and globally.*

Proof. The fact that both f and f^{-1} are monotone follows from Theorem 13.16. Regarding now the continuity of f^{-1}, we want to prove that we have:

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

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

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

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

And with this, we have now all the needed generalities in our bag.

13c. Elementary functions

As a basic application of Theorem 13.17, we have:

Proposition 13.18. *The various usual inverse functions, such as the inverse trigonometric functions* arcsin*,* arccos*,* arctan*,* arccot*, are all continuous.*

Proof. This follows indeed from Theorem 13.17, with a course the full discussion needing some explanations on bijectivity and domains. But you surely know all that, and in what concerns us, our claim is simply that these beasts are all continuous, proved. □

As another basic application of this, we have:

Proposition 13.19. *The following happen:*

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

Proof. All these results come as applications of Theorem 13.16, as follows:

(1) This is clear from Theorem 13.16 (3), applied on $[-\infty,\infty]$.

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

(3) This follows as well by applying Theorem 13.16 (3) to the polynomial $P(z) = z^n - x$, but this time on $[0,\infty)$. □

There are many other things that can be said about roots of polyomials, and solutions of other equations of type $f(x) = 0$, by using Theorem 13.16. We will be back to this.

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

Theorem 13.20. *The function x^a is defined and continuous on $(0,\infty)$, for any $a\in\mathbb R$. Moreover, when trying to extend it to $\mathbb R$, we have 4 cases, as follows,*

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

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

Proof. The idea is that we know how to extract roots by using Proposition 13.19, and all the rest follows by continuity. To be more precise:

(1) Assume $a=p/q$, with $p,q\in\mathbb N$, $p\neq 0$ and q odd. Given a number $x\in\mathbb R$, we can construct the power x^a in the following way, by using Proposition 13.19:

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

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

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

(3) Assume first $a\in\mathbb Q_{even}$, $a>0$. This means $a=p/q$ 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 13.19. It is then straightforward to prove that x^a is indeed continuous on $[0,\infty)$, and not extendable either to the negatives. Thus, we are done with the case $a\in\mathbb Q_{even}$, $a>0$, and the case left, namely $a\in\mathbb R-\mathbb Q$, $a>0$, follows as well by continuity.

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

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

Theorem 13.21. *The function a^x is as follows:*

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

Proof. This is a sort of reformulation of Theorem 13.20, by exchanging the variables, $x\leftrightarrow a$. To be more precise, the situation is as follows:

(1) We know from Theorem 13.20 that things fine with x^a for $x>0$, no matter what $a\in\mathbb R$ is. But this means that things fine with a^x for $a>0$, no matter what $x\in\mathbb R$ is.

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

(3) Given $a<0$, we know from Theorem 13.20 that we cannot define a^x for $x\in\mathbb Q_{even}$. But since $\mathbb Q_{even}$ is dense in $\mathbb R$, this gives the result. □

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

Our goal now will be to extend the material from chapter 4 regarding the numeric sequences and series, to the case of the sequences and series of functions. To start with, we can talk about the convergence of sequences of functions, $f_n \to f$, as follows:

Definition 13.22. *We say that f_n converges pointwise to f, and write $f_n \to f$, if*

$$f_n(x) \to f(x)$$

for any x. Equivalently, $\forall x, \forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, |f_n(x) - f(x)| < \varepsilon$.

The question is now, assuming that f_n are continuous, does it follow that f is continuous? I am pretty much sure that you think that the answer is "yes", based on:

$$\begin{aligned}\lim_{y\to x} f(y) &= \lim_{y\to x}\lim_{n\to\infty} f_n(y)\\ &= \lim_{n\to\infty}\lim_{y\to x} f_n(y)\\ &= \lim_{n\to\infty} f_n(x)\\ &= f(x)\end{aligned}$$

However, this proof is wrong, because we know well from chapter 1 that we cannot intervert limits, 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 13.23. *The basic step function, namely the sign function*

$$sgn(x) = \begin{cases} -1 & , \quad x < 0 \\ 0 & , \quad x = 0 \\ 1 & , \quad x > 0 \end{cases}$$

can be approximated by suitable modifications of $\arctan(x)$. *Even worse, there are examples of $f_n \to f$ with each f_n continuous, and with f totally discontinuous.*

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

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

Definition 13.24. *We say that f_n converges uniformly to f, and write $f_n \to_u f$, if:*

$$\forall\varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, |f_n(x) - f(x)| < \varepsilon, \forall x$$

That is, the same condition as for $f_n \to f$ must be satisfied, but with the $\forall x$ at the end.

And it is this "$\forall x$ at the end" which makes the difference, and will make our theory work. In order to understand this, which is something quite subtle, let us compare Definition 5.22 and Definition 5.24. As a first observation, we have:

Proposition 13.25. *Uniform convergence implies pointwise convergence,*

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

but the converse is not true, in general.

Proof. Here the first assertion is clear from definitions, just by thinking at what is going on, with no computations needed. As for the second assertion, the simplest counterexamples here are the functions $f_n : [0, 1] \to \mathbb{R}$ given by $f_n(x) = x^n$, that we met before in Proposition 13.23. Indeed, uniform convergence on $[0, 1)$ would mean:

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

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

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

Theorem 13.26. *Assuming that f_n are continuous, and that*

$$f_n \to_u f$$

then f is continuous. That is, uniform limit of continuous functions is continuous.

Proof. As previously said, it is the "$\forall x$ at the end" in Definition 13.24 that will make this work. Indeed, let us try to prove that the limit f is continuous at some point x. For this, we pick a number $\varepsilon > 0$. Since $f_n \to_u f$, we can find $N \in \mathbb{N}$ such that:

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

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

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

But with this, we are done. Indeed, for $|x-y| < \delta$ we have:

$$\begin{aligned} |f(x)-f(y)| &\leq |f(x)-f_N(x)| + |f_N(x)-f_N(y)| + |f_N(y)-f(y)| \\ &\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \\ &= \varepsilon \end{aligned}$$

Thus, the limit function f is continuous at x, and we are done. □

Obviously, the notion of uniform convergence in Definition 13.24 is something quite interesting, worth some more study. As a first result, we have:

Proposition 13.27. *The following happen, regarding uniform limits:*

(1) $f_n \to_u f$, $g_n \to_u g$ *imply* $f_n + g_n \to_u f + g$.
(2) $f_n \to_u f$, $g_n \to_u g$ *imply* $f_n g_n \to_u fg$.
(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. □

Finally, there is some abstract mathematics to be done as well. Indeed, observe that the notion of uniform convergence, as formulated in Definition 13.24, means that:

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

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

Theorem 13.28. *The uniform convergence, $f_n \to_u f$, means that we have $f_n \to f$ with respect to the following distance,*

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

which is indeed a distance function.

Proof. Here the fact that d is indeed a distance, 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. □

13d. Binomial formula

With the above theory in hand, let us get now to interesting things, namely computations. Among others, because this is what a mathematician's job is, doing all sorts of 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 13.29. *We have the generalized binomial formula*

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

with the generalized binomial coefficients being given by

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

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

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

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

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

(2) For the exponent $a=-1$ this is something that we know from Part I 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. □

Let us discuss now some further generalizations of what we have.

Quite interestingly, we have as well the following result:

Theorem 13.30. *The generalized binomial formula, namely*

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

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

$$\sqrt{1+t}=1-2\sum_{k=1}^\infty C_{k-1}\left(\frac{-t}{4}\right)^k$$

with $C_k=\frac{1}{k+1}\binom{2k}{k}$ being the Catalan numbers, and at $a=-1/2$ we obtain

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

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

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

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

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

(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:

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

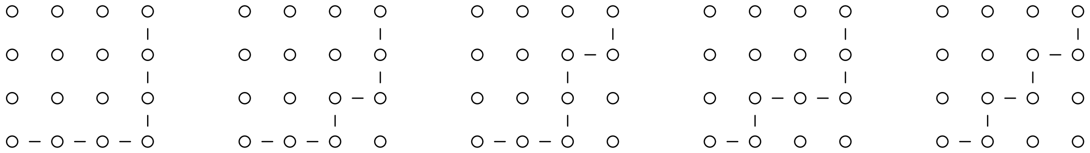
Thus, this leads to a sort of half-failure, the conclusion being that for proving by recurrence the second formula in (4), we need the third formula in (4).

(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:

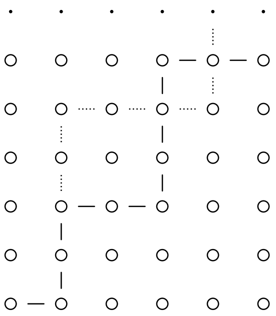
$$1, 1, 2, 5, 14, 42, 132, 429, \ldots$$

(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.

13e. Exercises

Exercises:

Exercise 13.31.

Exercise 13.32.

Exercise 13.33.

Exercise 13.34.

Exercise 13.35.

Exercise 13.36.

Exercise 13.37.

Exercise 13.38.

Bonus exercise.

CHAPTER 14

Derivatives

14a. Derivatives, rules

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 14.1. *A function $f:\mathbb{R}\to\mathbb{R}$ is called differentiable at x when*

$$f'(x)=\lim_{t\to0}\frac{f(x+t)-f(x)}{t}$$

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:

$$\lim_{t\to0}\left[f(x+t)-f(x)\right]=0$$

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 14.2. *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 14.1. We have:

$$\lim_{t\searrow0}\frac{|0+t|-|0|}{t}=\lim_{t\searrow0}\frac{t-0}{t}=1$$

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

$$\lim_{t\nearrow0}\frac{|0+t|-|0|}{t}=\lim_{t\nearrow0}\frac{-t-0}{t}=-1$$

Thus, the limit in Definition 14.1 does not converge, as desired. □

Generally speaking, the last assertion in Proposition 14.2 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 14.3. *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:

$$f'(x)=\lim_{t\to0}\frac{f(x+t)-f(x)}{t}$$

By multiplying by t, we obtain that we have, once again in the $t\to0$ limit:

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

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

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

Proposition 14.4. *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 14.5. *We have the following results:*

1. $(\sin x)' = \cos x$.
2. $(\cos x)' = -\sin x$.
3. $(e^x)' = e^x$.
4. $(\log x)' = x^{-1}$.

Proof. This is quite tricky, as always when computing derivatives, as follows:

(1) Regarding sin, the computation here goes as follows:

$$\begin{aligned}(\sin x)' &= \lim_{t\to 0}\frac{\sin(x+t)-\sin x}{t}\\ &= \lim_{t\to 0}\frac{\sin x\cos t+\cos x\sin t-\sin x}{t}\\ &= \lim_{t\to 0}\sin x\cdot\frac{\cos t-1}{t}+\cos x\cdot\frac{\sin t}{t}\\ &= \cos x\end{aligned}$$

Here we have used the fact, which is clear on pictures, by drawing the trigonometric circle, that we have $\sin t\simeq t$ for $t\simeq 0$, plus the fact, which follows from this and from Pythagoras, $\sin^2+\cos^2=1$, that we have as well $\cos t\simeq 1-t^2/2$, for $t\simeq 0$.

(2) The computation for cos is similar, as follows:

$$\begin{aligned}(\cos x)' &= \lim_{t\to 0}\frac{\cos(x+t)-\cos x}{t}\\ &= \lim_{t\to 0}\frac{\cos x\cos t-\sin x\sin t-\cos x}{t}\\ &= \lim_{t\to 0}\cos x\cdot\frac{\cos t-1}{t}-\sin x\cdot\frac{\sin t}{t}\\ &= -\sin x\end{aligned}$$

(3) For the exponential, the derivative can be computed as follows:

$$\begin{aligned}(e^x)' &= \left(\sum_{k=0}^\infty\frac{x^k}{k!}\right)'\\ &= \sum_{k=0}^\infty\frac{kx^{k-1}}{k!}\\ &= e^x\end{aligned}$$

(4) As for the logarithm, the computation here is as follows, using $\log(1+y)\simeq y$ for $y\simeq 0$, which follows from $e^y\simeq 1+y$ that we found in (3), by taking the logarithm:

$$\begin{aligned}(\log x)' &= \lim_{t\to 0}\frac{\log(x+t)-\log x}{t}\\ &= \lim_{t\to 0}\frac{\log(1+t/x)}{t}\\ &= \frac{1}{x}\end{aligned}$$

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

Speaking exponentials, we can now formulate a nice result about them:

Theorem 14.6. *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:

$$c_1=1\quad,\quad c_2=\frac{1}{2}\quad,\quad c_3=\frac{1}{2\times3}\quad,\quad c_4=\frac{1}{2\times3\times4}\quad,\quad\ldots$$

Thus we have $c_k=1/k!$, leading to the conclusion in the statement. □

Observe that the above result leads to a more conceptual explanation for the number e itself. To be more precise, $e\in\mathbb R$ is the unique number satisfying:

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

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

Theorem 14.7. *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 14.8. *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 14.7 (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 14.7 (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. □

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 14.9. *We have the following formulae,*

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

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

Proof. For tan, we have the following computation:

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

As for arctan, we can use here the following computation:

$$\begin{aligned}(\tan\circ\arctan)'(x) &= \tan'(\arctan x)\arctan'(x)\\ &= \frac{1}{\cos^2(\arctan x)}\arctan'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x'=1$, we obtain from this:

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

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

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

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

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

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

At the theoretical level now, further building on Theorem 14.3, we have:

Theorem 14.10. *The local minima and maxima of a differentiable function $f:\mathbb{R}\to\mathbb{R}$ appear at the points $x\in\mathbb{R}$ where:*

$$f'(x)=0$$

However, the converse of this fact is not true in general.

Proof. The first assertion follows from the formula in Theorem 14.3, namely:

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

Indeed, let us rewrite this formula, more conveniently, in the following way:

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

Now saying that our function f has a local maximum at $x\in\mathbb{R}$ means that there exists a number $\varepsilon>0$ such that the following happens:

$$f(x+t)\geq f(x)\quad,\quad\forall t\in[-\varepsilon,\varepsilon]$$

We conclude that we must have $f'(x)t\geq 0$ for sufficiently small t, and since this small t can be both positive or negative, this gives, as desired:

$$f'(x)=0$$

Similarly, saying that our function f has a local minimum at $x\in\mathbb{R}$ means that there exists a number $\varepsilon>0$ such that the following happens:

$$f(x+t)\leq f(x)\quad,\quad\forall t\in[-\varepsilon,\varepsilon]$$

Thus $f'(x)t\leq 0$ for small t, and this gives, as before, $f'(x)=0$. Finally, in what regards the converse, the simplest counterexample here is the following function:

$$f(x)=x^3$$

Indeed, we have $f'(x)=3x^2$, and in particular $f'(0)=0$. But our function being clearly increasing, $x=0$ is not a local maximum, nor a local minimum. □

As an important consequence of Theorem 14.10, we have:

THEOREM 14.11. *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 14.10. 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 14.10 can be used in order to find the maximum and minimum of any differentiable function, and this method is best recalled as follows:

ALGORITHM 14.12. *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.

14b. 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 14.13. *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 14.14. *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:

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

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.

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 14.15. *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:

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

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

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

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

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

Of course, this is not the end of the story, because these formulae remain quite opaque, and must be examined in view of Interpretation 14.14, in order to see what exactly is going on. Also, we have tan and the inverse trigonometric functions too. In short, plenty of good exercises here, for you, and the more you solve, the better your calculus will be.

(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. □

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

THEOREM 14.16. *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 Theorem 14.3 that $b = f'(x)$ is the correct choice for the coefficient of t. Thus, our approximation must be as follows:

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

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

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

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

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

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

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

Observe that this formula holds indeed, as an application of Theorem 14.3. Now by using this, if we denote by $\varphi(t)\simeq P(t)$ the formula to be proved, we have:

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

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

The above result substantially improves Theorem 14.3, and there are many applications of it. As a first such application, justifying Interpretation 14.14 (3), we have the following statement, which is a bit heuristic, but we will call it however Proposition:

Proposition 14.17. *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 14.17, as such, and with $t\to-t$ too:

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

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

By making the average, we obtain the following formula:

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

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 can improve as well Theorem 14.10, as follows:

THEOREM 14.18. *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 14.16, which in the case $f'(x) = 0$ reads:

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

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

As before with Theorem 14.10, the above result is not the end of the story with the mathematics of the local minima and maxima, because things are undetermined when:

$$f'(x) = f''(x) = 0$$

For instance the functions $\pm x^n$ with $n \in \mathbb{N}$ all satisfy this condition at $x = 0$, which is a minimum for the functions of type x^{2m}, a maximum for the functions of type $-x^{2m}$, and not a local minimum or local maximum for the functions of type $\pm x^{2m+1}$.

There are some comments to be made in relation with Algorithm 14.12 as well. Normally that algorithm stays strong, because Theorem 14.18 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 14.18 is good to know, when applying that algorithm.

14c. Convex functions

As a main concrete application now of the second derivative, which is something very useful in practice, and related to Interpretation 14.14 (4), we have the following result:

THEOREM 14.19. *Given a convex function $f : \mathbb{R} \to \mathbb{R}$, we have the following Jensen inequality, for any $x_1, \ldots, x_N \in \mathbb{R}$, and any $\lambda_1, \ldots, \lambda_N > 0$ summing up to 1,*

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

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

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

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

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

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

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

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

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

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

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

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

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

(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:

$$f''(x)\geq0$$

(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 14.20. *For any $p\in(1,\infty)$ we have the following inequality,*

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

and for any $p\in(0,1)$ we have the following inequality,

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

with in both cases equality precisely when $|x_1|=\ldots=|x_N|$.

Proof. This follows indeed from Theorem 14.19, because we have:

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

Thus x^p is convex for $p>1$ and concave for $p<1$, which gives the results. □

Observe that at $p=2$ we obtain as particular case of the above inequality the Cauchy-Schwarz inequality, or rather something equivalent to it, namely:

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

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 14.21. *We have the Young inequality,*

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

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

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

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

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

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

Now by exponentiating, we obtain the Young inequality. □

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.

14d. The 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 14.22. *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 of Theorem 14.16. To be more precise, if we denote by $\varphi(t)\simeq P(t)$ the approximation to be proved, we have:

$$\begin{aligned}\frac{\varphi(t)-P(t)}{t^n}&\simeq\frac{\varphi'(t)-P'(t)}{nt^{n-1}}\\&\simeq\frac{\varphi''(t)-P''(t)}{n(n-1)t^{n-2}}\\&\vdots\\&\simeq\frac{\varphi^{(n)}(t)-P^{(n)}(t)}{n!}\\&=\frac{f^{(n)}(x)-f^{(n)}(x)}{n!}\\&=0\end{aligned}$$

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

Here is a related interesting statement, inspired from the above proof:

Proposition 14.23. *For a polynomial of degree n, the Taylor approximation*

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

is an equality. The converse of this statement holds too.

PROOF. By linearity, it is enough to check the equality in question for the monomials $f(x)=x^p$, with $p\leq n$. But here, the formula to be proved is as follows:

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

We recognize the binomial formula, so our result holds indeed. As for the converse, this is clear, because the Taylor approximation is a polynomial of degree n. □

There are many other things that can be said about the Taylor formula, at the theoretical level, notably with a study of the remainder, when truncating this formula at a given order $n\in\mathbb{N}$. We will be back to this later, in chapter 16 below.

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

THEOREM 14.24. *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 13.

(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 13, when trying to directly establish particular cases of the generalized binomial formula. Gone all that.

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

PROPOSITION 14.25. *We have the following formula,*

$$\sqrt{1+t}=1-2\sum_{k=1}^\infty C_{k-1}\left(\frac{-t}{4}\right)^k$$

with $C_k=\frac{1}{k+1}\binom{2k}{k}$ being the Catalan numbers. Also, we have

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

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

PROOF. This is something that we already know from chapter 13, but time now to review all this. At $p=1/2$, the generalized binomial coefficients are:

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

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

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

Thus, Theorem 14.24 at $p=\pm1/2$ gives the formulae in the statement. □

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

THEOREM 14.26. *We have the following formulae,*

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

as well as the following formulae,

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

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

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

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

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

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

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

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

(3) Finally, the fact that the formulae in the statement extend beyond the small t setting, coming from Taylor series, is something standard too. We will leave this as an instructive exercise, and come back to it later, in chapter 16 below. □

14e. Exercises

Exercises:

Exercise 14.27.

Exercise 14.28.

Exercise 14.29.

Exercise 14.30.

Exercise 14.31.

Exercise 14.32.

Exercise 14.33.

Exercise 14.34.

Bonus exercise.

CHAPTER 15

Equations

15a. Differential equations

Differential equations.

15b. 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 15.1. *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 $a=v'(t)=x''(t)$.

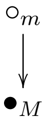
Summarizing, with Newton's theory of derivatives, as we learned it in this chapter, we can certainly do some mathematics for the motion of bodies. But, for these bodies to move, we need them to be acted upon by some forces, right? The simplest such force is gravity, and in our present, modest 1 dimensional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

Theorem 15.2. *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 extended by M, which is fixed in our formalism, on the object m which moves, is subject to the following equations:

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

To be more precise, in the first equation $G\simeq 6.674\times 10^{-11}$ is the gravitational constant, in usual SI units, and the sign is $-$ because F is attractive. The second equation is something standard and very intuitive, and the last two equations are those from Fact 3.28. Now observe that, with the above data for F, the equation $F=ma$ reads:

$$-G\cdot\frac{Mm}{x^2}=m\ddot{x}$$

Thus, by simplifying, we are led to the equation in the statement. □

15c. Wave equation

As more phsyics, we can talk as well about waves in 1 dimension, as follows:

Theorem 15.3. *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. □

15d. Heat equation

Along the same lines, we can talk as well about the heat equation in 1D, as follows:

THEOREM 15.4. *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 14, the heat equation as formulated above tells us that the rate of change $\dot{\varphi}$ of the temperature of the material at any given point must be proportional,

with proportionality factor $\alpha > 0$, to the average difference of temperature between that given point and the surrounding material. Which sounds reasonable.

(2) In practice now, we can use, a bit like before for the wave equation, a lattice model as follows, with distance $l > 0$ between the neighbors:

$$\text{——}\ \circ_{x-l} \overset{l}{\text{——}} \circ_x \overset{l}{\text{——}} \circ_{x+l}\ \text{——}$$

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

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

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

All this is very nice, so with the calculus that we know, we can certainly talk about physics. We will see later in this book how to deal with the above equations.

15e. Exercises

Exercises:

Exercise 15.5.

Exercise 15.6.

Exercise 15.7.

Exercise 15.8.

Exercise 15.9.

Exercise 15.10.

Exercise 15.11.

Exercise 15.12.

Bonus exercise.

CHAPTER 16

Integration

16a. 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 16.1. *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 16.2. *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 16.3. *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 16.1, 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 16.4. *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 16.2, 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 16.2. More on this later.

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

Proposition 16.5. *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 16.6. *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 16.1, because area means side times average height. Thus, we can formulate:

Theorem 16.7. *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 16.1, via some geometric thinking. Alternatively, this is something which certainly comes from Theorem 16.6. □

The point of view in Theorem 16.7 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 16.3, 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 16.7. The same goes for the results in Proposition 16.5, which become even more obvious with the viewpoint from Theorem 16.7. However, not everything trivializes in this way, and the result which is left, namely the formula $\int_{-1}^1\sqrt{1-x^2}\,dx=\pi/2$ from Proposition 16.3 (3), not only does not trivialize, but becomes quite opaque with our new philosophy.

In short, modesty. Integration is a quite delicate business, and we have several equivalent points of view on what an integral means, and all these points of view are useful, and must be learned, with none of them being clearly better than the others.

Going ahead with more interpretations of the integral, we have:

Theorem 16.8. *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 16.7 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 16.9. *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 16.9:

Conclusion 16.10. *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 16.11. *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 16.8, 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. □

16b. Riemann sums

Our purpose now will be to understand which functions $f:\mathbb{R}\to\mathbb{R}$ are integrable, and how to compute their integrals. For this purpose, the Riemann formula in Theorem 16.6 will be our favorite tool. Let us begin with some theory. We first have:

THEOREM 16.12. *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. □

Going ahead with more theory, let us establish some abstract properties of the integration operation. We already know from Proposition 16.5 that the integrals behave well with respect to sums and multiplication by scalars. Along the same lines, we have:

Theorem 16.13. *The integrals behave well with respect to taking limits,*

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

and with respect to taking infinite sums as well,

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

with both these formulae being valid, undwer mild assumptions.

Proof. This is something quite standard, by using the general theory developed in chapter 13 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 16.14. *Given a continuous function $f : [a,b] \to \mathbb{R}$, we have*

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

with this being called mean value property.

Proof. Our claim is that this follows from the following trivial estimate:

$$\min(f) \leq f \leq \max(f)$$

Indeed, by integrating this over $[a,b]$, we obtain the following estimate:

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

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

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

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

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

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

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 16.15. *We have the integration formula*

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

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

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

(1) By linearity we can assume that our interval $[a,b]$ is of the form $[0,c]$, and the formula that we want to establish is as follows:

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

(2) We can further assume $c = 1$, and by expressing the left term as a Riemann sum, we are in need of the following estimate, in the $N \to \infty$ limit:

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

(3) So, let us try to prove this. At $p = 0$, obviously nothing to do, because we have the following formula, which is exact, and which proves our estimate:

$$1^0 + 2^0 + \ldots + N^0 = N$$

(4) At $p = 1$ now, we are confronted with a well-known question, namely the computation of $1 + 2 + \ldots + N$. But this is simplest done by arguing that the average of the numbers $1, 2, \ldots, N$ being the number in the middle, we have:

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

Thus, we obtain the following formula, which again solves our question:

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

(5) At $p = 2$ now, go compute $1^2 + 2^2 + \ldots + N^2$. This is not obvious at all, so as a preliminary here, let us go back to the case $p = 1$, and try to find a new proof there, which might have some chances to extend at $p = 2$. The trick is to use 2D geometry. Indeed, consider the following picture, with stacks going from 1 to N:

$$\begin{array}{ccccc} & & & & \square\\ & & & & \vdots\\ & & \square & \ldots & \square\\ & \square & \square & \ldots & \square\\ \square & \square & \square & \ldots & \square \end{array}$$

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 16.16. *We have the following estimate,*

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

and so, by Riemann sums, we have the following integration formula,

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

valid for any exponent $p\in\mathbb{N}$, and perhaps for some other $p\in\mathbb{R}$.

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

THEOREM 16.17. *We have the following integration formula,*

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

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

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

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

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

16c. Advanced results

The problem is now, what to do with what we have, namely Conjecture 16.16 and Theorem 16.17. Not obvious, so stuck, and time to ask the cat. And cat says:

CAT 16.18. *Summing the infinitesimals of the rate of change of the function should give you the global change of the function. Obvious.*

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

This being said, wait. There is suggestion to connect integrals and derivatives, and this is in fact what we have, coming from Conjecture 16.16 and Theorem 16.17, due to:

$$\left(\frac{x^{p+1}}{p+1}\right)' = x^p \quad , \quad (e^x)' = e^x$$

So, eureka, we have our idea, thanks cat. Moving ahead now, following this idea, we first have the following result, called fundamental theorem of calculus:

THEOREM 16.19. *Given a continuous function $f : [a, b] \to \mathbb{R}$, if we set*

$$F(x) = \int_a^x f(s)ds$$

then $F' = f$. That is, the derivative of the integral is the function itself.

Proof. This follows from the Riemann integration picture, and more specifically, from the mean value property from Theorem 16.14. Indeed, we have:

$$\frac{F(x+t)-F(x)}{t}=\frac{1}{t}\int_x^{x+t}f(x)dx$$

On the other hand, our function f being continuous, by using the mean value property from Theorem 16.14, we can find a number $c\in[x,x+t]$ such that:

$$\frac{1}{t}\int_x^{x+t}f(x)dx=f(x)$$

Thus, putting our formulae together, we conclude that we have:

$$\frac{F(x+t)-F(x)}{t}=f(c)$$

Now with $t\to 0$, no matter how the number $c\in[x,x+t]$ varies, one thing that we can be sure about is that we have $c\to x$. Thus, by continuity of f, we obtain:

$$\lim_{t\to 0}\frac{F(x+t)-F(x)}{t}=f(x)$$

But this means exactly that we have $F'=f$, and we are done. □

We have as well the following result, which is something equivalent, and a hair more beautiful, also called fundamental theorem of calculus:

Theorem 16.20. *Given a function $F:\mathbb{R}\to\mathbb{R}$, we have*

$$\int_a^b F'(x)dx=F(b)-F(a)$$

for any interval $[a,b]$.

Proof. As already mentioned, this is something which follows from Theorem 16.19, and is in fact equivalent to it. Indeed, consider the following function:

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

By using Theorem 16.19 we have $G'=F'$, and so our functions F,G differ by a constant. But with $s=a$ we have $G(a)=0$, and so the constant is $F(a)$, and we get:

$$F(s)=G(s)+F(a)$$

Now with $s=b$ this gives $F(b)=G(b)+F(a)$, which reads:

$$F(b)=\int_a^b F'(x)dx+F(a)$$

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

As a first illustration for all this, solving our previous problems, we have:

Theorem 16.21. *We have the following integration formulae,*

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

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

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

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

Proof. We already know some of these formulae, but the best is to do everything, using the fundamental theorem of calculus. The computations go as follows:

(1) With $F(x) = x^{p+1}$ we have $F'(x) = px^p$, and we get, as desired:

$$\int_a^b px^p\,dx = b^{p+1} - a^{p+1}$$

(2) Observe first that the formula (1) does not work at $p = -1$. However, here we can use $F(x) = \log x$, having as derivative $F'(x) = 1/x$, which gives, as desired:

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

(3) With $F(x) = \cos x$ we have $F'(x) = -\sin x$, and we get, as desired:

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

(4) With $F(x) = \sin x$ we have $F'(x) = \cos x$, and we get, as desired:

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

(5) With $F(x) = e^x$ we have $F'(x) = e^x$, and we get, as desired:

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

(6) This is something more tricky. We are looking for a function satisfying:

$$F'(x) = \log x$$

This does not look doable, but fortunately the answer to such things can be found on the internet. But, what if the internet connection is down? So, let us think a bit, and try to solve our problem. Speaking logarithm and derivatives, what we know is:

$$(\log x)' = \frac{1}{x}$$

But then, in order to make appear log on the right, the idea is quite clear, namely multiplying on the left by x. We obtain in this way the following formula:

$$(x\log x)' = 1\cdot\log x + x\cdot\frac{1}{x} = \log x + 1$$

We are almost there, all we have to do now is to substract x from the left, as to get:

$$(x\log x - x)' = \log x$$

But this this formula in hand, we can go back to our problem, and we get the result. □

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

Definition 16.22. *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 14, 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 16.21 in a more digest form, as follows:

Theorem 16.23. *We have the following formulae for primitives,*

$$\int x^p = \frac{x^{p+1}}{p+1} \quad , \quad \int \frac{1}{x} = \log x$$

$$\int \sin x = -\cos x \quad , \quad \int \cos x = \sin x$$

$$\int e^x = e^x \quad , \quad \int \log x = x\log x - x$$

allowing us to compute the corresponding definite integrals too.

Proof. Here the various formulae in the statement follow from Theorem 16.21, or rather from the proof of Theorem 16.21, or even from chapter 14, for most of them, and the last assertion comes from the integration formula given after Definition 16.22. □

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

THEOREM 16.24. *We have the formula*

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

called integration by parts.

PROOF. This follows by integrating the Leibniz 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 16.25. *We have the change of variable formula*

$$\int_a^b f(x)dx=\int_c^d f(\varphi(t))\varphi'(t)dt$$

where $c=\varphi^{-1}(a)$ and $d=\varphi^{-1}(b)$.

PROOF. This follows with $f=F'$, from the following differentiation rule, that we know from chapter 14, 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 14, we have:

THEOREM 16.26. *Given a function $f:\mathbb{R}\to\mathbb{R}$, we have the formula*

$$f(x+t)=\sum_{k=0}^n\frac{f^{(k)}(x)}{k!}\,t^k+\int_x^{x+t}\frac{f^{(n+1)}(s)}{n!}(x+t-s)^n\,ds$$

called Taylor formula with integral formula for the remainder.

Proof. This is something which looks a bit complicated, so we will first do some verifications, and then we will go for the proof in general:

(1) At $n=0$ the formula in the statement is as follows, and certainly holds, due to the fundamental theorem of calculus, which gives $\int_x^{x+t} f'(s)ds = f(x+t)-f(x)$:

$$f(x+t)=f(x)+\int_x^{x+t} f'(s)ds$$

(2) At $n=1$, the formula in the statement becomes more complicated, as follows:

$$f(x+t)=f(x)+f'(x)t+\int_x^{x+t} f''(s)(x+t-s)ds$$

As a first observation, this formula holds indeed for the linear functions, where we have $f(x+t)=f(x)+f'(x)t$, and $f''=0$. So, let us try $f(x)=x^2$. Here we have:

$$f(x+t)-f(x)-f'(x)t=(x+t)^2-x^2-2xt=t^2$$

On the other hand, the integral remainder is given by the same formula, namely:

$$\begin{aligned}\int_x^{x+t} f''(s)(x+t-s)ds &= 2\int_x^{x+t}(x+t-s)ds\\ &= 2t(x+t)-2\int_x^{x+t} sds\\ &= 2t(x+t)-((x+t)^2-x^2)\\ &= 2tx+2t^2-2tx-t^2\\ &= t^2\end{aligned}$$

(3) Still at $n=1$, let us try now to prove the formula in the statement, in general. Since what we have to prove is an equality, this cannot be that hard, and the first thought goes towards differentiating. But this method works indeed, and we obtain the result.

(4) In general, the proof is similar, by differentiating, the computations being similar to those at $n=1$, and we will leave this as an instructive exercise. □

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 16.27. *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 16.28. *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. □

As another 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 16.29. *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 16.30. *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. □

16d. 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 16.31. *A probability density is a function $\varphi:\mathbb R\to\mathbb R$ satisfying*

$$\varphi\geq0\qquad,\qquad\int_{\mathbb R}\varphi(x)dx=1$$

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:

$$\psi\geq0\quad,\quad\alpha_i\geq0\quad,\quad\int_{\mathbb R}\psi(x)dx+\sum_i\alpha_i=1$$

Observe the obvious relation with intuitive probability theory, where the probability for something to happen is always positive, $P\geq0$, 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 16.31 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 16.31:

DEFINITION 16.32. *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 16.33. *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 16.32 and Definition 16.33, 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 16.34 (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

$$M_k=\int_\mathbb{R}x^kd\mu(x)$$

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 16.35. *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 16.36. *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.

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