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. Numbers, Functions

Abstract. This is an introduction to mathematics, with emphasis on geometric aspects. We first discuss numbers, counting, fractions and percentages, and their basic applications. Then we get into plane geometry, with a study of triangles and trigonometry, followed by coordinates and complex numbers. We then go into functions and analysis, with the basics of the theory explained, followed by exponentials, logarithms and more trigonometry, and with the derivatives and integrals discussed too. Finally, we provide an introduction to vector calculus, space geometry and basic mechanics.

Preface

The foundations of modern mathematics, as they were developed a few centuries ago, and even before, are quite vast, and take some time to be explored. Normally the learning is done in two steps, following the story of mathematics itself, as follows:

(1) At the beginning we certainly have basic algebra and geometry. With basic algebra dealing with the numbers and their properties, such as divisibility, equations, prime numbers and related topics. And with basic geometry dealing with triangles, circles and more complicated curves in the plane, and with a look into trigonometry too.

(2) And then, we have basic analysis, and vector calculus. With basic analysis dealing with the functions, polynomial or more general, their properties, such as continuity, and tools for dealing with them, such as differentiation and integration. And with vector calculus being something more advanced, mixing algebra, geometry and analysis.

The present book is an introduction to this, foundational mathematics, by using the above traditional learning scheme, namely numbers and basic algebra, to start with, followed by basic geometry and trigonometry, followed by functions and basic analysis, and with some vector calculus material, which is more advanced, at the end.

The book, meant to be accessible to anyone knowing and loving basic mathematics from school, and by this I mean we will start from 0 or almost, but we will go quite fast, with the goal of learning a maximum of things, is organized in 4 parts, as follows:

Part I - Numbers. We discuss here numbers, first with a philosophical discussion regarding numeration bases and notations, then with a look into basic arithmetic, and basic counting too, followed by real numbers and what can be done with them, including sequences and series, and with an introduction to more advanced aspects at the end.

Part II - Geometry. Here we discuss geometry, first with the basic study of triangles, their various centers, and other things that can be said about them, then with a look into angles and trigonometry, and then with more advanced aspects, namely real coordinates, and complex numbers too, and finally conics and other basic algebraic curves.

Part III - Functions. Here we discuss functions, first with a classical study of the polynomials and their roots, in general and in low degree too, followed by the standard modern theory of continuous functions, meaning basics, intermediate value theorem, derivatives of first, second and higher order, and then integration and basic applications.

Part IV - Vectors. We discuss here vectors, first with a general introduction to space geometry, featuring tetrahedra, polyhedra and related topics, then with a discussion of basic linear algebra and matrix theory, followed by a discussion of multivariable functions and their analysis, and with an introduction to basic mechanics at the end.

In the hope that you will find this book useful. Personally, I collected here what I usually have to say to my 1st year students, and sometimes 2nd year and higher too, and by including also some material which, although being basic, beautiful, and desirable to learn, is no longer deemed fashionable in the present times. In a word, I collected here what I think first year mathematics should mean, in a future, better world.

As for the second and third year mathematics, these should be normally dedicated, a bit like the learning in physics, chemistry, biology, engineering, computer science and economics goes, to taking students to truly modern science, as we presently know it. Many things to learned here, and we will provide some references at the end.

It is a pleasure to thanks to my cats, for some help with the organization of the book. Also, some of the tricks using complex numbers come from them.

Cergy, October 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,*

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

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

$$\text{one},\ \text{two},\ \text{three},\ \text{four},\ \text{five},\ \text{six}$$

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:

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

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,*

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

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

$$\text{one},\ \text{two},\ \text{three},\ \text{four},\ \text{five},\ \text{six}$$

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:

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

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:

$$\bigstar\bigstar\bigstar\quad\bigcirc\bigcirc\quad ||||\quad\to\quad 324$$

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

$$\bigstar\bigstar\bigstar\bigstar\bigstar\bigstar\quad\bigcirc\bigcirc\bigcirc\bigcirc\bigcirc\bigcirc\quad ||||||\quad\to\quad 666$$

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:

$$\bigstar\bigstar\bigstar\quad ||||\quad\to\quad ?$$

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:

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

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

Question 1.10 is something quite subtle, 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. Here the numbers are sequences of type*

$$n = a_1a_2\ldots a_k$$

with $a_i \in \{0,1\}$, and $a_1 \neq 0$, and with the counting going as follows:

(1) *If a set has $a_1 = 1$ objects, the set count is $n = a_1$,*
(2) *If a set consists of $a_1 = 1$ pairs, followed by $a_2 \in \{0,1\}$ objects, the set count is $n = a_1a_2$,*
(3) *If a set consists of $a_1 = 1$ quadruplets, followed by $a_2 \in \{0,1\}$ pairs, and then by $a_3 \in \{0,1\}$ objects, the count is $n = a_1a_2a_3$,*

.. and so on, the idea being that we can count any set, no matter how big, in this way.

Which sounds quite exciting, doesn't it. More in detail now, here is how the counting in basis two goes, and with this looking like something quite simple:

$$|\circ| = 1$$

$$|\circ\circ| = 10$$

$$|\circ\circ\circ| = 11$$

$$|\circ\circ\circ\circ| = 100$$

$$|\circ\circ\circ\circ\circ| = 101$$

$$|\circ\circ\circ\circ\circ\circ| = 110$$

$$|\circ\circ\circ\circ\circ\circ\circ| = 111$$

$$|\circ\circ\circ\circ\circ\circ\circ\circ| = 1000$$

$$|\circ\circ\circ\circ\circ\circ\circ\circ\circ| = 1001$$

$$\ldots$$

Regarding the addition table, this is something ridiculously simple, as follows:

$$\begin{array}{cc} + & 1 \\ 1 & 10 \end{array}$$

As for the multiplication table, this is ridiculously simple too, as follows:

$$\begin{array}{cc} \times & 1 \\ 1 & 1 \end{array}$$

So, shall we use this new system? I would rather say no, on the grounds that what we have in the above seems to require only two neurons for understanding, and we certainly have more neurons than that. So, our old numeration system, using the digits $0, 1, 2, 3, 4, 5, 6$ and their magic, looks like something more advanced.

Before leaving numeration basis two, however, let us mention that this system is used, successfully, by our friends the computers. But we are smarter than them.

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

Example 1.12. *Numeration basis three. Here the numbers are sequences of type*

$$n = a_1 a_2 \ldots a_k$$

with $a_i \in \{0, 1, 2\}$, and $a_1 \neq 0$, and with the counting going as follows:

(1) *If a set has $a_1 \in \{1, 2\}$ objects, the set count is $n = a_1$,*

(2) *If a set consists of $a_1 \in \{1, 2\}$ triples, followed by $a_2 \in \{0, 1, 2\}$ objects, the set count is $n = a_1 a_2$,*

(3) *If a set consists of $a_1 \in \{1, 2\}$ triples of triples, followed by $a_2 \in \{0, 1, 2\}$ triples, and then by $a_3 \in \{0, 1, 2\}$ objects, the count is $n = a_1 a_2 a_3$,*

.. and so on, the idea being that we can count any set, no matter how big, in this way.

As before, many things can be said here. Here is how the set counting goes:

$$|\circ| = 1$$

$$|\circ\circ| = 2$$

$$|\circ\circ\circ| = 10$$

$$|\circ\circ\circ\circ| = 11$$

$$|\circ\circ\circ\circ\circ| = 12$$

$$|\circ\circ\circ\circ\circ\circ| = 20$$

$$|\circ\circ\circ\circ\circ\circ\circ| = 21$$

$$|\circ\circ\circ\circ\circ\circ\circ\circ| = 22$$

$$|\circ\circ\circ\circ\circ\circ\circ\circ\circ| = 200$$

$$|\circ\circ\circ\circ\circ\circ\circ\circ\circ\circ| = 201$$

$$\ldots$$

Regarding now the addition table, this is something quite fun too, as follows:

$+$	1	2
1	2	10
2	10	11

As for the multiplication table, this is again something quite exciting, as follows:

$\times$	1	2
1	1	2
2	2	11

Time now to draw some conclusions, so, shall we use this new system? I would again say no, again on the grounds that what we have in the above seems to require only few neurons for understanding, and we certainly have more neurons than that.

Coming next, we have numeration basis four, whose theory is as follows:

Example 1.13. *Numeration basis four. Here the numbers are sequences of type*

$$n=a_1a_2\ldots a_k$$

with $a_i\in\{0,1,2,3\}$, and $a_1\neq 0$, counting in the obvious way, the addition table is

$+$	1	2	3
1	2	3	10
2	3	10	11
3	10	11	12

the multiplication table is

$\times$	1	2	3
1	1	2	3
2	2	10	12
3	3	12	21

and in practice, this is a sort of a better version of numeration basis two.

To be more precise here, in what regards the last assertion, it is quite clear that everything that we can do, as tricks, in basis two, can be seen as well in basis four. And so, that basis four is more advanced than basis two, due to the more symbols used.

In any case, as before with basis two, all this rather belongs to computer science. So, we will not use this numeration basis, let the computers use it, if they want to.

Coming next, we have something quite interesting, as follows:

Example 1.14. *Numeration basis five. Here the numbers are sequences of type*

$$n = a_1a_2\ldots a_k$$

with $a_i \in \{0,1,2,3,4\}$, *and* $a_1 \neq 0$, *counting in the obvious way, the addition table is*

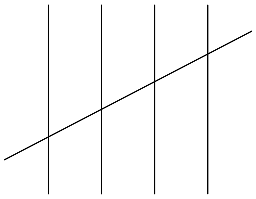
+	1	2	3	4
1	2	3	4	10
2	3	4	10	11
3	4	10	11	12
4	10	11	12	13

the multiplication table is

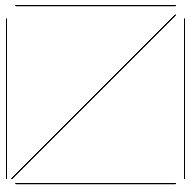
×	1	2	3	4
1	1	2	3	4
2	2	4	11	13
3	3	11	14	22
4	4	13	22	31

and in practice, this is something quite efficient, for counting.

To be more precise here, in what regards the last assertion, there is certainly some truth there, that you might be aware of, because the chunks of five objects are very easy to represent, with a well-known convention for this being as follows:



An alternative convention here, which is widely used as well, is as follows:



Quite interesting all this, and still used on prison walls, and in many other concrete situations. Personally, this is my favorite system, for counting things.

Coming next, we have numeration basis six, which again is something interesting:

Example 1.15. *Numeration basis six. Here the numbers are sequences of type*

$$n = a_1a_2\ldots a_k$$

with $a_i \in \{0,1,2,3,4,5\}$*, and* $a_1 \neq 0$*, counting in the obvious way, the addition table is*

+	1	2	3	4	5
1	2	3	4	5	10
2	3	4	5	10	11
3	4	5	10	11	12
4	5	10	11	12	13
5	10	11	12	13	14

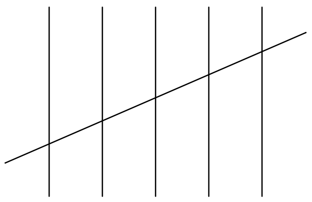
the multiplication table is

×	1	2	3	4	5
1	1	2	3	4	5
2	2	4	10	12	14
3	3	10	13	20	23
4	4	12	20	24	32
5	5	14	23	32	41

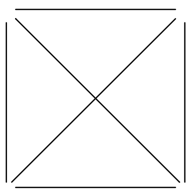
and in practice, this beats both basis two, and basis three.

To be more precise here, in what regards the last assertion, it is pretty much clear that all sorts of tricks from basis two and basis three can be done in basis six too.

In what regards graphics, the chunks of six objects are quite easy to represent too, with a well-known convention for this being as follows:



An alternative convention here, which is widely used as well, is as follows:



Summarizing, quite interesting numeration basis that we have here, nicely mixing two and three, and that can be successfully used, for various purposes.

Coming next, we have numeration basis seven, which is something fun too:

Example 1.16. *Numeration basis seven. Here the numbers are sequences of type*

$$n = a_1a_2\ldots a_k$$

with $a_i \in \{0,1,2,3,4,5,6\}$, and $a_1 \neq 0$, counting as usual, the addition table is

+	1	2	3	4	5	6
1	2	3	4	5	6	10
2	3	4	5	6	10	11
3	4	5	6	10	11	12
4	5	6	10	11	12	13
5	6	10	11	12	13	14
6	10	11	12	13	14	15

the multiplication table is

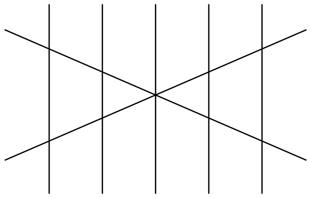
×	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	11	13	15
3	3	6	12	15	21	24
4	4	11	15	22	26	33
5	5	13	21	26	34	42
6	6	15	24	33	42	51

and in practice, this solves some of our school 7-related nightmares.

To be more precise here, in what regards the last assertion, remember that damn 6×7 and 7×8 computations from school, that we all had big troubles with. Well, in basis seven these two computations take a very simple form, as follows:

$$6\times 10 = 60 \quad , \quad 10\times 11 = 110$$

In what regards the graphics, however, not very good news here, because with the heptagon being hard to draw, we are basically left with ugly pictures, as follows:



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. The basics here are as follows:

Definition 1.17. *In numeration basis ten the numbers are sequences of type*

$$n = a_1a_2\ldots a_k$$

with $a_i \in \{0,1,2,3,4,5,6,7,8,9\}$, *and* $a_1 \neq 0$, *counting as usual, the addition table is*

+	1	2	3	4	5	6	7	8	9
1	2	3	4	5	6	7	8	9	10
2	3	4	5	6	7	8	9	10	11
3	4	5	6	7	8	9	10	11	12
4	5	6	7	8	9	10	11	12	13
5	6	7	8	9	10	11	12	13	14
6	7	8	9	10	11	12	13	14	15
7	8	9	10	11	12	13	14	15	16
8	9	10	11	12	13	14	15	16	17
9	10	11	12	13	14	15	16	17	18

the multiplication table is

×	1	2	3	4	5	6	7	8	9
1	1	2	3	4	5	6	7	8	9
2	2	4	6	8	10	12	14	16	18
3	3	6	9	12	15	18	21	24	27
4	4	8	12	16	20	24	28	32	36
5	5	10	15	20	25	30	35	40	45
6	6	12	18	24	30	36	42	48	54
7	7	14	21	28	35	42	49	56	63
8	8	16	24	32	40	48	56	64	72
9	9	18	27	36	45	54	63	72	81

and in practice, this is the numeration basis that we will be using.

Now that we have our numeration basis, let us develop some theory for it. To start with, in mathematical notation, our usual counting rules can be summarized as follows, with obvious meanings for the sum and product operations + and ×:

$$a_1 = a_1$$
$$a_1a_2 = 10 \times a_1 + a_2$$
$$a_1a_2a_3 = 100 \times a_1 + 10 \times a_2 + a_3$$
$$a_1a_2a_3a_4 = 1000 \times a_1 + 100 \times a_2 + 10 \times a_3 + a_4$$
$$\ldots$$

We conclude that, again in standard mathematical notation, we have the following formula, for an arbitrary number $n = a_1a_2\ldots a_k$, as in Definition 1.17:

$$a_1a_2\ldots a_k = 10^{k-1} \times a_1 + 10^{k-2} \times a_2 + \ldots + 10 \times a_{k-1} + a_k$$

Regarding now the addition of our numbers, in order to add two arbitrary numbers, $n = a_1a_2\ldots a_k$ and $m = b_1b_2\ldots b_s$, we can do this in the following way:

$$\begin{aligned}
& a_1a_2\ldots a_k + b_1b_2\ldots b_s \\
= \ & (10^{k-1} \times a_1 + 10^{k-2} \times a_2 + \ldots + 10 \times a_{k-1} + a_k) \\
& + (10^{s-1} \times b_1 + 10^{s-2} \times b_2 + \ldots + 10 \times b_{s-1} + b_s) \\
= \ & 10^{k-1} \times a_1 + 10^{s-1} \times b_1 + \ldots\ldots + 10 \times a_{k-1} + 10 \times b_{s-1} + a_k + b_s \\
= \ & 10(10^{k-2} \times a_1 + 10^{s-2} \times b_1 + \ldots\ldots + a_{k-1} + b_{s-1}) + a_k + b_s
\end{aligned}$$

Thus, proceeding from right to left, the last digit will obviously be $a_k + b_s$, or rather the last digit of $a_k + b_s$, in case $a_k + b_s \geq 10$, and so on, up to the first digit.

Equivalently, we have here the basic algorithm for addition, obtained by putting $n = a_1a_2\ldots a_k$ on top of $m = b_1b_2\ldots b_s$, and summing as above, that you know well.

Getting now to multiplication, in order to multiply two arbitrary numbers, $n = a_1a_2\ldots a_k$ and $m = b_1b_2\ldots b_s$, we can do this in the following way:

$$\begin{aligned}
& a_1a_2\ldots a_k \times b_1b_2\ldots b_s \\
= \ & (10^{k-1} \times a_1 + 10^{k-2} \times a_2 + \ldots + 10 \times a_{k-1} + a_k) \\
& \times (10^{s-1} \times b_1 + 10^{s-2} \times b_2 + \ldots + 10 \times b_{s-1} + b_s) \\
= \ & 10^{k+s-2} \times a_1b_1 + \ldots\ldots + 10 \times a_{k-1}b_s + 10 \times a_kb_{s-1} + a_kb_s \\
= \ & 10(10^{k+s-3} \times a_1b_1 + \ldots\ldots + a_{k-1}b_s + a_kb_{s-1}) + a_kb_s
\end{aligned}$$

Thus, when proceeding from right to left, the last digit will obviously be a_kb_s, or rather the last digit of a_kb_s, in case $a_kb_s \geq 10$, and so on, up to the first digit.

Equivalently, we have the algorithm for multiplication, obtained by putting $n = a_1a_2\ldots a_k$ on top of $m = b_1b_2\ldots b_s$, and multiplying as above, that you know well.

1c. Basic arithmetic

With counting and numbers understood, let us develop now some basic arithmetic. We have here the following key notion, which often appears in the real life:

Definition 1.18. *We say that b divides a, and we write $b|a$, when*

$$a=bc$$

for some number c. In this case we also use the following notation,

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

with this being called fraction, for designating this quotient number c.

All this is quite intuitive, and the fractions are subject to a number of simple formulae, which are all useful, in the real life, which can be summarized as follows:

Theorem 1.19. *The fractions add and substract according to the formulae*

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

and they multiply and divide according to the formulae

$$\frac{a}{b}\cdot\frac{c}{d}=\frac{ac}{bd}\quad,\quad\frac{a}{b}:\frac{c}{d}=\frac{ad}{bc}$$

provided, at the end, that the quotient a/b is a multiple of the quotient c/d.

Proof. Let us see indeed how the mathematical proof goes. According to Definition 1.18 we have the following equality, that we will use many times, in what follows:

$$\frac{a}{b}=\frac{ad}{bd}\quad,\quad\forall d$$

(1) In what regards the addition formula, this can be established as follows:

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

(2) The proof of the substraction formula is similar, as follows:

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

(3) In what regards now the multiplication formula, this comes from:

$$\left(\frac{a}{b}\cdot\frac{c}{d}\right)bd=\left(\frac{a}{b}\cdot b\right)\left(\frac{c}{d}\cdot d\right)=ac$$

(4) As for the division formula, this can be proved as follows:

$$\left(\frac{a}{b}:\frac{c}{d}\right)bc=\frac{abc}{b}:\frac{c}{d}=ac:\frac{c}{d}=ad$$

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

Moving ahead with more arithmetic, we have the following result:

Theorem 1.20. *Given two numbers a, b, we can talk about their greatest common divisor (a, b), and their least common multiple $[a, b]$. We can write*

$$a = da' \quad , \quad b = db'$$

with a', b' being numbers having no common divisor, and we have:

$$(a, b) = d \quad , \quad [a, b] = da'b'$$

Also, (a, b) and $[a, b]$ are subject to the formula $(a, b)[a, b] = ab$.

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

(1) To start with, we can talk indeed about the greatest common divisor $d = (a, b)$ of any two numbers $a, b \in \mathbb{N}$, in several equivalent ways, as follows:

– The simplest way is to argue that since all common divisors $e|a, b$ satisfy $e \leq a, b$, we can pick the greatest such common divisor, $d = (a, b)$.

– Alternatively, we can say that $d = (a, b)$ is the smallest number having the property that any common divisor $e|a, b$ must divide it, $e|d$.

– Yet another approach is by recurrence on $a + b$. Indeed, assuming $a > b$, we can perform the division $a = bn + c$, and then set $(a, b) = (b, c)$, by recurrence.

(2) In practice now, there is some discussion needed here, in order to prove that the above 3 methods yield indeed the same number $d = (a, b)$. But this is best seen via the third method, which produces the same numbers as the first and second methods.

(3) So, this was for the story of the greatest common divisor (a, b), and in what regards the least common multiple $[a, b]$, the story here is similar, and we will leave the details as an exercise. Alternatively, if looking for a quick formal proof here, we can define $[a, b]$ by starting with (a, b), and using the formula $(a, b)[a, b] = ab$, discussed below.

(4) Next, with $d = (a, b)$, we can decompose our two numbers as follows:

$$a = da' \quad , \quad b = db'$$

We have then the following implications, coming from definitions:

$$(a, b) = d \implies (da', db') = d \implies (a', b') = 1$$

Thus, we managed to write a, b as in the statement, and with this done, it is clear that we must have $[a, b] = da'b'$, so we have both formulae in the statement, namely:

$$(a, b) = d \quad , \quad [a, b] = da'b'$$

(5) Finally, observe that we have the following formula:

$$ab = d^2a'b' = d \times da'b' = (a, b)[a, b]$$

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

We can basically do the same with three numbers, as follows:

THEOREM 1.21. *Given three numbers a,b,c, we can talk about their greatest common divisor (a,b,c), and their least common multiple $[a,b,c]$, and if we write*

$$a=da'\quad,\quad b=db'\quad,\quad c=dc'$$

with $(a',b',c')=1$, and then further decompose each pair (a',b'), (a',c'), (b',c'), by using their respective greatest common divisors, we are led to a decomposition as follows,

$$a=dpqx\quad,\quad b=dpry\quad,\quad c=dqrz$$

and in terms of this decomposition, we have the following formulae:

$$(a,b,c)=d\quad,\quad [a,b,c]=dpqrxyz$$

Also, $(a,b,c)^2[a,b,c]$ divides abc, but with these numbers being different, in general.

PROOF. As before, we can talk about (a,b,c) and $[a,b,c]$, in the obvious way. Now if we set $d=(a,b,c)$, we can decompose our three numbers as follows:

$$a=da'\quad,\quad b=db'\quad,\quad c=dc'$$

We have then the following implications, coming from definitions:

$$(a,b,c)=d\implies(da',db',dc')=d\implies(a',b',c')=1$$

Thus, we have managed to write a,b,c as in the statement, and we have:

$$(a,b,c)=d$$

In order to compute now $[a',b',c']$, we can look at the pairs (a',b'), (a',c'), (b',c'), and apply to them the theory that we learned in Theorem 1.20. Indeed, let us set:

$$p=(a',b')\quad,\quad q=(a',c')\quad,\quad r=(b',c')$$

We are led in this way to decompositions as follows, for the numbers a',b',c':

$$a'=pqx\quad,\quad b'=pry\quad,\quad c'=qrz$$

As a conclusion, our original numbers a,b,c decompose as follows:

$$a=dpqx\quad,\quad b=dpry\quad,\quad c=dqrz$$

But with these formulae in hand, the numbers that we were looking for are:

$$(a,b,c)=d\quad,\quad [a,b,c]=dpqrxyz$$

Now when multiplying our numbers a,b,c, we have the following formula:

$$\begin{aligned}abc&=dpqx\cdot dpry\cdot dqrz\\&=d^2\cdot dpqrxyz\cdot pqr\\&=(a,b,c)^2\cdot[a,b,c]\cdot pqr\end{aligned}$$

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

In the general case now, that of k numbers, we have the following result:

Theorem 1.22. *Given numbers $a_1,\ldots,a_k$, we can talk about their greatest common divisor $(a_1,\ldots,a_k)$, and their least common multiple $[a_1,\ldots,a_k]$, and we can write*

$$a_1=da_1'\quad,\quad\ldots\quad,\quad a_k=da_k'$$

with $d=(a_1,\ldots,a_k)$, and with $(a_1',\ldots,a_k')=1$. Also, the product

$$(a_1,\ldots,a_k)^{k-1}[a_1,\ldots,a_k]$$

divides the product $a_1\ldots a_k$, but these numbers are different, in general.

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

(1) As before, the fact that we can talk indeed about greatest common divisors, and about least common multiples, is something which is clear from definitions.

(2) Also as before, we can divide our numbers $a_1,\ldots,a_k$ by their common divisor $d=(a_1,\ldots,a_k)$, and we reach to a decomposition as follows, with $(a_1',\ldots,a_k')=1$:

$$a_1=da_1'\quad,\quad\ldots\quad,\quad a_k=da_k'$$

(3) However, and here comes the point, when it comes to suitably decomposing our numbers $a_1,\ldots,a_k$, or rather their reduced versions $a_1',\ldots,a_k'$, by using their various common divisors, as we did in Theorem 1.20 at $k=2$, and in Theorem 1.21 at $k=3$, things become considerably more complicated at $k=4$ and higher. We can only recommend here doing some computations at $k=4$, in order to understand what the difficulty is.

(4) Summarizing, we cannot say much, as a continuation of this, along the lines of what we did before at $k=2,3$, and the only obvious thing that can be said, completing our theorem, is the last assertion, which is something that we know, from Theorem 1.21. □

Moving on, still talking divisibility, here is a key result:

Theorem 1.23. *Given two numbers satisfying $(a,b)=1$, we can write*

$$ae+bf=1$$

for certain integers $e,f\in\mathbb{Z}$.

Proof. This might sound a bit surprising, but give me any two numbers which are prime to each other, say 7 and 10, and after thinking a bit, here is what I find:

$$7\times 3-10\times 2=1$$

So, let us try to prove now the result, in general. For this purpose, let us look at:

$$b,2b,3b,\ldots,ab$$

This is a certain collection of a numbers, and our claim is that the remainders of these numbers, modulo a, are different. Indeed, this is something coming from:

$$\begin{aligned} cb=db(a) &\iff a\mid(c-d)b\\ &\iff a\mid c-d\\ &\iff c=d \end{aligned}$$

Thus, our a numbers above are distinct modulo a, and so we have, still modulo a:

$$\left\{b,2b,3b,\ldots,ab\right\}=\left\{1,2,3,\ldots,a\right\}$$

But this does the job, because we get a certain $f\in\{1,\ldots,a\}$ such that:

$$bf=1(a)$$

Thus we must have $ae+bf=1$, for a certain $e\in\mathbb{Z}$, as desired. □

Along the same lines, here is a useful generalization of the above result:

Theorem 1.24. *Given numbers satisfying* $(a_1,\ldots,a_k)=1$*, we can write*

$$a_1e_1+a_2e_2+\ldots+a_ke_k=1$$

for certain integers $e_1,\ldots,e_k\in\mathbb{Z}$*.*

Proof. We already know from Theorem 1.23 that this holds at $k=2$, and in general, the result will follow from this, the idea being as follows:

(1) Let us first see how the proof goes at $k=3$. Given three numbers a,b,c having no common divisor, $(a,b,c)=1$, let us write them as in Theorem 1.21, as follows:

$$a=pqx\quad,\quad b=pry\quad,\quad c=qrz$$

Now let us look at the sums of the following type, with $e,f,g\in\mathbb{Z}$:

$$ae+bf+cg=pqxe+pryf+qrzg$$

– As a first observation, since we have $(qx,ry)=1$, we know from Theorem 1.23 that the quantities of type $qxe+ryf$ will range over the whole $\mathbb{Z}$.

– Next, and getting now towards what we want to prove, we conclude from this that the quantities of type $pqxe+pryf$ range over the whole $p\mathbb{Z}$.

– But then, since we have $(p,qrz)=1$, we conclude, again by using Theorem 1.23, that the above sums $pqxe+pryf+qrzg$ range over the whole $\mathbb{Z}$.

(2) Thus, theorem proved at $k=3$, and the proof in general is similar, by recurrence on k. To be more precise, it is technically convenient to look at a slightly more general statement, saying that given $a_1,\ldots,a_k$, we can always find $e_1,\ldots,e_k\in\mathbb{Z}$ such that:

$$a_1e_1+a_2e_2+\ldots+a_ke_k=(a_1,\ldots,a_k)$$

But with this picture in hand, it is quite clear that the above arguments used at $k=3$ will apply in general, and will give the result, by recurrence on $k\in\mathbb{N}$. □

Moving ahead, we will be mostly interested in congruence questions, based on:

Definition 1.25. *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. There is a bit of recursivity that can be used, in relation with this, as shown by:

$$6|a\iff 2|a\text{ and }3|a$$

$$\begin{aligned}10|a&\iff 2|a\text{ and }5|a\\ 12|a&\iff 3|a\text{ and }4|a\\ 14|a&\iff 2|a\text{ and }7|a\\ 15|a&\iff 3|a\text{ and }5|a\\ 18|a&\iff 2|a\text{ and }9|a\\ 20|a&\iff 4|a\text{ and }5|a\\ 21|a&\iff 3|a\text{ and }7|a\end{aligned}$$

$$22|a\iff 2|a\text{ and }11|a$$

In general, based on these observations, the idea is that by writing $n=n_1\ldots n_k$ with the factors n_i having no common divisor, we just have to solve this question for certain special values of n, excluding $n=6,10,12,14,15,18,20,21,22,\ldots$

These special values of n are called "powers of primes", and many things can be said about them. More on them later in this chapter, and then later in this book.

In practice, the first such numbers, powers of primes, are as follows:

$$q=2,3,4,5,7,8,9,11,13,16,17,19,23,\ldots$$

And in what regards solving $a=0(q)$, with respect to these powers of primes, there are many useful tricks here, which can be summarized as follows:

Theorem 1.26. *Given a positive integer $a=a_1\ldots a_k$, we have:*

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

Proof. Here the $q = 2^r, 5$ 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:

Theorem 1.27. *Assuming that we have convinced mankind to change the numeration basis from* 10 *to* 14*, given a positive integer $a = a_1 \ldots a_k$, we have:*

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

Proof. Here the $q = 2^r, 7$ 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 $14 = 13 + 1$. □

In short, we have solved indeed the $q = 7, 13$ problems, but as a caveat, we have now $q = 11$ not working. And is this worth it or not, up to you to decide here.

1d. Prime numbers

Time now to get into prime numbers, which will be a main theme of discussion, in this book. 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, but time now to review all this, on a more systematic basis. To start with, as a definition for them, we have:

Definition 1.28. *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 the conditions (1,2,3) is something intuitive and standard, which can be deduced by using our common divisor technology developed above. Observe also 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 we know, already used in the above, we have:

THEOREM 1.29. *Any integer $n > 1$ decomposes uniquely as*

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

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

PROOF. This is something very standard, related to the equivalent conditions (1,2,3) in Definition 1.28, which formally comes by recurrence on n. As an interesting exercise here, work out this for all the integers $n \leq 100$, with no calculators allowed. □

As a first result now about the prime numbers themselves, we have:

THEOREM 1.30. *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 1.31. *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. □

The above algorithm, while mathematically rather trivial, is something quite fascinating, because it suggests all sorts of mechanical ways of dealing with the primes, via analysis and physics and engineering. Let us record this as a conjecture:

Conjecture 1.32. *A good analyst, physicist and engineer would probably have no troubles in elucidating everything about primes, using the sieve method.*

And we will end the present opening chapter with this. Mystery.

1e. Exercises

This was a quite basic chapter, about numbers and their main properties, save perhaps for the discussion regarding the numeration bases. As exercises on this, we have:

Exercise 1.33. *Do all your computations, for a full day in a row, in basis* 2.

Exercise 1.34. *Then the next day, do all your computations in basis* 3.

Exercise 1.35. *And the day after, do all your computations in basis* 6.

Exercise 1.36. *Review the algorithms of addition and multiplication, in basis* 10.

Exercise 1.37. *Fill in all the details, for* $abc=(a,b,c)^2\cdot[a,b,c]\cdot pqr$.

Exercise 1.38. *Study too, with some counterexamples, what happens for* a,b,c,d.

Exercise 1.39. *Invent some clever criteria for the divisibility with* 7 *and* 13.

Exercise 1.40. *Fill in all the details, in the proof of the unique factorization.*

As bonus exercise, learn a bit more about the prime numbers. We will be back to them, later in this book, but the more you know in advance, the better that will be.

CHAPTER 2

Fractions

2a. Fractions

Time now for some more complicated mathematics, going beyond what we know about integers. We recall from chapter 1 that given an integer dividing another integer, $b|a$, we can talk about the corresponding quotient c, given by $a = bc$, and denoted as follows:

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

The fractions are subject to a number of formulae, that we explained in chapter 1, which are all useful, in the real life. For addition and substraction, the formulae are:

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

As for multiplication and division, here the formulae are as follows:

$$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd} \quad , \quad \frac{a}{b} : \frac{c}{d} = \frac{ad}{bc}$$

The point now is that we can talk about fractions even when $b|a$ fails, in the obvious way. And, with this convention, the above formulae still hold. Let us start with:

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".

So, this is the definition of the rational numbers, which will take us some time, in order to properly understand. Generally speaking, the idea will be as follows:

(1) We will usually treat, based on a number of abstract results that we will prove, the rational numbers as usual numbers, that is, as integers.

(2) With the remark of course that the rational numbers are not necessarily integers. It is only that their arithmetic is quite similar to that of the integers.

Getting to work now, we must first talk about addition and substraction, and then about multiplication and division. As a first result, regarding the addition, we have:

Theorem 2.2. *We can add the rational numbers $r = a/b$ according to the rule*

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

and with this convention, we have the following formulae,

$$(r+s)+t=r+(s+t)\quad,\quad r+s=s+r$$

called associativity and commutativity of the addition operation.

Proof. We can certainly define an operation as in the statement, and with this done, our operation is indeed associative, as shown by the following computation:

$$\begin{aligned}\left(\frac{a}{b}+\frac{c}{d}\right)+\frac{e}{f} &= \frac{ad+bc}{bd}+\frac{e}{f}\\ &= \frac{(ad+bc)f+bde}{bdf}\\ &= \frac{adf+bcf+bde}{bdf}\\ &= \frac{adf+b(cf+de)}{bdf}\\ &= \frac{a}{b}+\frac{cf+de}{df}\\ &= \frac{a}{b}+\left(\frac{c}{d}+\frac{e}{f}\right)\end{aligned}$$

As for the commutativity of the sum, this is clear from definitions, as shown by:

$$\begin{aligned}\frac{a}{b}+\frac{c}{d} &= \frac{ad+bc}{bd}\\ &= \frac{cb+da}{db}\\ &= \frac{c}{d}+\frac{a}{b}\end{aligned}$$

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

Next, we must talk about substraction. The result here is similar, as follows:

Proposition 2.3. *We can substract the rationals according to the rule*

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

and with this, all basic formulae that we know for integer fractions still hold.

PROOF. This is something quite self-explanatory, and we will leave doing some computations here, in order to make sure that everything works fine, as an exercise. □

Summarizing, everything fine with the addition of rationals. Next, we must talk about multiplication. The result here is quite similar to Theorem 2.2, as follows:

THEOREM 2.4. *We can multiply the rational numbers $r = a/b$ according to the rule*

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

and with this convention, we have the following formulae,

$$(rs)t = r(st) \quad , \quad rs = sr$$

called associativity and commutativity of the multiplication operation.

PROOF. We can certainly define an operation as in the statement, and with this done, our operation is indeed associative, as shown by the following computation:

$$\begin{aligned}\left(\frac{a}{b}\cdot\frac{c}{d}\right)\frac{e}{f} &= \frac{ac}{bd}\cdot\frac{e}{f}\\ &= \frac{ace}{bdf}\\ &= \frac{a}{b}\cdot\frac{ce}{df}\\ &= \frac{a}{b}\left(\frac{c}{d}\cdot\frac{e}{f}\right)\end{aligned}$$

As for the commutativity property, this is clear from definitions, as shown by:

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

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

Next, we must talk about division. The result here is again routine, as follows:

PROPOSITION 2.5. *We can divide the rationals according to the rule*

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

and with this, all basic formulae that we know for integer fractions still hold.

PROOF. This is indeed quite self-explanatory, and we will leave doing some computations here, in order to make sure that everything works fine, as an exercise. □

As a conclusion now, things fine with both the addition and the multiplication, and you might probably think that we are done with all this algebra. Well, you must be kidding. Many other basic algebraic things remain to be done, such as:

Theorem 2.6. *The addition and multiplication of rationals are subject to*

$$r(s+t)=rs+rt\quad,\quad(r+s)t=rt+st$$

called distributivity formulae.

Proof. The verification of the first formula goes as follows:

$$\begin{aligned}\frac{a}{b}\left(\frac{c}{d}+\frac{e}{f}\right)&=\frac{a}{b}\cdot\frac{cf+de}{df}\\&=\frac{acf+ade}{bdf}\\&=\frac{acf}{bdf}+\frac{ade}{bdf}\\&=\frac{ac}{bd}+\frac{ae}{bf}\\&=\frac{a}{b}\cdot\frac{c}{d}+\frac{a}{b}\cdot\frac{e}{f}\end{aligned}$$

As for the second formula, this follows from the first one, and commutativity:

$$\begin{aligned}(r+s)t&=t(r+s)\\&=tr+ts\\&=rt+st\end{aligned}$$

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

Summarizing, many operations that we have here, and job for us to get familiar with all this, via practice, tricks and so on. Here is my favorite trick for fractions, which is something quite trivial, but I will call this Theorem, because this is perhaps the matematical formula that I use the most, in my complicated, daily quantum physics work:

Theorem 2.7. *We have the following substraction formula,*

$$\frac{1}{n}-\frac{1}{n+1}=\frac{1}{n(n+1)}$$

valid for any $n\in\mathbb{N}$. As illustrations for this, we have

$$1-\frac{1}{2}=\frac{1}{2}\quad,\quad\frac{1}{2}-\frac{1}{3}=\frac{1}{6}\quad,\quad\frac{1}{3}-\frac{1}{4}=\frac{1}{12}\quad,\quad\frac{1}{4}-\frac{1}{5}=\frac{1}{20}\quad\ldots$$

and with the knowledge of these latter formulae being mandatory too.

Proof. This is something trivial, but since we called our result Theorem, as mathematicians do, let us pull out now a complete proof, also as mathematicians do. We have the following computation, based on our general formula for substracting fractions:

$$\begin{aligned}\frac{1}{n}-\frac{1}{n+1}&=&\frac{1\times(n+1)-1\times n}{n(n+1)}\\&=&\frac{(n+1)-n}{n(n+1)}\\&=&\frac{1}{n(n+1)}\end{aligned}$$

Thus, theorem proved, and for the particular cases at the end, I will leave it to you. The more such particular cases you know well, the better your mathematics will be. □

At a more abstract level now, we have the following result, regarding the sums:

THEOREM 2.8. *The sum of two fractions is always of the following form,*

$$\frac{a}{b}+\frac{c}{d}=\frac{e}{[b,d]}$$

with $e\in\mathbb{Z}$ being a certain number. More generally, the sum of n fractions is of the form

$$\frac{a_1}{b_1}+\ldots+\frac{a_n}{b_n}=\frac{e}{[b_1,\ldots,b_n]}$$

with $e\in\mathbb{Z}$ being a certain number.

PROOF. In what regards the first assertion, we know from chapter 1 that the least common multiple $[b,d]$ appears as follows, for certain integers p,q:

$$[b,d]=bp=dq$$

But with this, we have the following computation, proving the first assertion:

$$\begin{aligned}\frac{a}{b}+\frac{c}{d}&=&\frac{ap}{bp}+\frac{cq}{dq}\\&=&\frac{ap}{[b,d]}+\frac{cq}{[b,d]}\\&=&\frac{ap+cq}{[b,d]}\end{aligned}$$

As for the second assertion, its proof is similar. We know that the least common multiple $[b_1,\ldots,b_n]$ appears as follows, for certain integers $p_1,\ldots,p_n$:

$$[b_1,\ldots,b_n]=b_1p_1=\ldots=b_np_n$$

But with these formulae in hand, we have the following computation:

$$\begin{aligned}\frac{a_1}{b_1}+\ldots+\frac{a_n}{b_n} &= \frac{a_1p_1}{b_1p_1}+\ldots+\frac{a_np_n}{b_np_n}\\ &= \frac{a_1p_1}{[b_1,\ldots,b_n]}+\ldots+\frac{a_np_n}{[b_1,\ldots,b_n]}\\ &= \frac{a_1p_1+\ldots+a_np_n}{[b_1,\ldots,b_n]}\end{aligned}$$

Thus, theorem proved, but as before with many other such things, a lot of practice is needed, meaning working out a lot of exercises, in order to master well this method. □

Summarizing, we have a nice theory of rational numbers, extending well what we knew from before, from chapter 1, regarding the fractions $r=a/b$ with $b|a$.

There is actually one more thing to be talked about, in relation with this, namely the ordering of the fractions. The result here, formulated a bit informally, is as follows:

Theorem 2.9. *We can order the positive fractions as follows,*

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

and with this, all the basic results about the ordering of numbers extend.

Proof. This is obviously something a bit informal, but there is a reason for this, there are in fact countless things to be checked here, and my concern as math professor is you getting a bit bored by all this, and starting to attend a chemistry class instead. This being said, as a sample verification, let us attempt to prove the following fact:

$$\frac{a}{b}<\frac{c}{d}\,,\ \frac{e}{f}<\frac{g}{h}\implies\frac{a}{b}+\frac{e}{f}<\frac{c}{d}+\frac{g}{h}$$

But this can be proved indeed, with some patience, as follows:

$$\begin{aligned}\frac{a}{b}+\frac{e}{f}<\frac{c}{d}+\frac{g}{h} &\iff \frac{af+be}{bf}<\frac{ch+dg}{dh}\\ &\iff (af+be)dh<bf(ch+dg)\\ &\iff adfh+bdeh<bcfh+bdgf\\ &\iff adfh-bcfh<bdgf-bdeh\\ &\iff (ad-bc)fh<bd(gf-eh)\end{aligned}$$

Indeed, what we have at the end does hold, because our assumptions give:

$$(ad-bc)fh<0<bd(gf-eh)$$

So, verification done, and for the rest, exercise of course for you to formulate a more precise version of the statement, and perform the other verifications that are needed. □

Getting now a bit abstract, the basic operations on the rational numbers, namely sum, product and inversion, tell us that $\mathbb{Q}$ is a field, in the following sense:

Definition 2.10. *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 $a(b+c)=ab+ac$.*

So, this is the much feared definition of the fields, and more on this later in this book. In the meantime, let us record the following result, coming from the above:

Theorem 2.11. *The rational numbers $\mathbb{Q}$ form a field, with operations given by:*

$$\frac{a}{b}+\frac{c}{d}=\frac{ad+bc}{bd}\quad,\quad\frac{a}{b}\cdot\frac{c}{d}=\frac{ac}{bd}$$

In fact, $\mathbb{Q}$ is the smallest field containing $\mathbb{Z}$, or even the smallest field containing $\mathbb{N}$.

Proof. This is more or less clear from the above, as follows:

(1) We have indeed a field, with the operations in the statement, with the verification of the various field axioms being something clear, coming from the above results.

(2) As a comment here, we have opted to include in the statement only the basic operations for fractions, that of the sum, and product. The other operations, regarding substraction, division, inverses, can be all deduced easily from these two operations.

(3) Regarding now the last assertion, when searching for a field containing $\mathbb{N}$, by looking at the equation $c=a-b$ we are led into $\mathbb{Z}$. But then, by looking at the equation $c=a/b$, we are led into $\mathbb{Q}$. Thus, the field that we were looking for is $\mathbb{Q}$. □

Still staying a bit abstract, as a further result about rationals, in relation with what we like to do the most, since the beginning of this book, namely counting, we have:

Theorem 2.12. *The field of rational numbers $\mathbb{Q}$ is countable.*

Proof. We can count indeed the positive rationals, with some redundancies, by arranging them in a table, and snaking our way inside this table, as follows:

$$\begin{tikzcd}
1/1 \arrow[r] & 1/2 \arrow[dl] & 1/3 \arrow[r] & 1/4 \arrow[dl] & 1/5 \arrow[r] & 1/6 \arrow[dl] & \cdots \\
2/1 \arrow[d] & 2/2 \arrow[ur] & 2/3 \arrow[dl] & 2/4 \arrow[ur] & 2/5 \arrow[dl] & 2/6 & \cdots \\
3/1 \arrow[ur] & 3/2 \arrow[dl] & 3/3 \arrow[ur] & 3/4 \arrow[dl] & 3/5 & 3/6 & \cdots \\
4/1 \arrow[d] & 4/2 \arrow[ur] & 4/3 \arrow[dl] & 4/4 & 4/5 & 4/6 & \cdots \\
5/1 \arrow[ur] & 5/2 \arrow[dl] & 5/3 & 5/4 & 5/5 & 5/6 & \cdots \\
6/1 & 6/2 & 6/3 & 6/4 & 6/5 & 6/6 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{tikzcd}$$

Thus, after eliminating the redundancies, and then adding the negatives, which must be countable too, say via an alternating $+/-$ scheme, theorem proved. $\square$

Many other things can be said, as a continuation of the above, notably with the question of explicitly listing the elements $\mathbb{Q}$, if possible in some sort of increasing order. However, as we will soon discover, by doing some analysis, this is not really possible.

2b. Binomials, factorials

Time now to get into some interesting mathematics, by using our knowledge of numbers. As a first theorem, solving a problem which often appears in real life, we have:

Theorem 2.13. *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:

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

However, when thinking well, 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:

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

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

The binomial numbers, as constructed above, are quite fascinating objects, and the more you know about them, the better your mathematics will be. Trust me here.

To start with, here are some basic formulae for binomial coefficients that you should definitely memorize, and pull right away, when needed in your computations:

$$\binom{n}{1}=n$$

$$\binom{n}{2}=\frac{n(n-1)}{2}$$

$$\binom{n}{3}=\frac{n(n-1)(n-2)}{6}$$

$$\binom{n}{4}=\frac{n(n-1)(n-2)(n-3)}{24}$$

Here are as well some numerics, with $n=k,k+1,k+2,\ldots$ in each case, that you should know well too, and pull out instantly, when needed in your computations:

$$\binom{n}{2}=1,3,6,10,15,21,28,\ldots$$

$$\binom{n}{3}=1,4,10,20,35,56,\ldots$$

$$\binom{n}{4}=1,5,15,35,70,\ldots$$

Finally, still talking numerics, as an important adding to Theorem 2.13, we have:

Convention 2.14. *By definition we have the formula*

$$0!=1$$

as for the following binomial coefficient computation to work,

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

in agreement with what Theorem 2.13 says, requiring $\binom{n}{n}=1$.

Going ahead now with more mathematics and less philosophy, with Theorem 2.13 complemented by this convention being in final form, we have:

Theorem 2.15. *We have the binomial formula*

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

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

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

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

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

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

In order to finish, it remains to compute the coefficients C_k. But, according to our product formula, C_k is the number of choices for the k needed a variables among the n available a variables. Thus, according to Theorem 2.13, we have:

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

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

Theorem 2.14 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:

$$\begin{gathered}(a+b)^0=1\\(a+b)^1=a+b\\(a+b)^2=a^2+2ab+b^2\\(a+b)^3=a^3+3a^2b+3ab^2+b^3\\(a+b)^4=a^4+4a^3b+6a^2b^2+4ab^3+b^4\\(a+b)^5=a^5+5a^4b+10a^3b^2+10a^2b^3+5ab^4+b^5\\\vdots\end{gathered}$$

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 2.16. *The Pascal triangle, formed by the binomial coefficients* $\binom{n}{k}$,

$$\begin{gathered}1\\1\ ,\ 1\\1\ ,\ 2\ ,\ 1\\1\ ,\ 3\ ,\ 3\ ,\ 1\\1\ ,\ 4\ ,\ 6\ ,\ 4\ ,\ 1\\1\ ,\ 5\ ,\ 10\ ,\ 10\ ,\ 5\ ,\ 1\\\vdots\end{gathered}$$

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 the idea being always be 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.

Getting now to more advanced things regarding the binomial coefficients, let us formulate, as a complement to the various particular cases discussed before:

Definition 2.17. *The central binomial coefficients are the following numbers,*

$$D_n=\binom{2n}{n}$$

which are not to be confused with the middle binomial coefficients,

$$E_n=\binom{n}{[n/2]}$$

with [.] *standing as usual for the integer part.*

Observe that we can recover the central binomial coefficients as particular cases of the middle binomial coefficients, due to the following trivial formula:

$$D_n=E_{2n}$$

However, in practice, the central binomial coefficients D_n are the truly interesting quantities, and the middle binomial coefficients E_n remain something quite secondary.

Regarding the numerics for the central binomial coefficients, these are as follows:

$$D_n = 1, 2, 6, 20, 70, 252, 924, 3432, 12870, 48620, \ldots$$

This sequence is actually something quite fascinating, and if you are a number theory nerd, and hope so are you, one of the first things that you will discover, by playing with it, is that these central binomial coefficients factorize as follows:

$$\begin{aligned} D_n &= 1\times 1, 2\times 1, 3\times 2, 4\times 5, 5\times 14, 6\times 42,\\ &\quad 7\times 132, 8\times 429, 9\times 1430, 10\times 4862, \ldots \end{aligned}$$

Thus, we are led in this way to the following conjecture:

Conjecture 2.18. *The central binomial coefficients factorize as*

$$D_n = (n+1)C_n$$

with $C_n = 1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, \ldots$ *being certain integers.*

However, this is something which is not trivial to prove, with bare hands, and we will leave it for later in this chapter, once we will know more things.

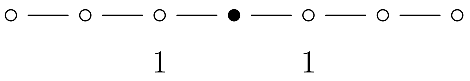
2c. Catalan numbers

We would like to count now loops on graphs, with this being a quite interesting question. Think for instance percolation, when making coffee, each droplet of water will have to make its way through the coffee particles, and this is how making coffee works.

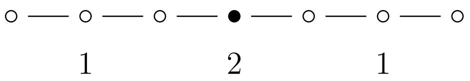
In practice now, let us start with the following question:

Question 2.19. *What is the number of length k paths on $\mathbb{Z}$, based at 0?*

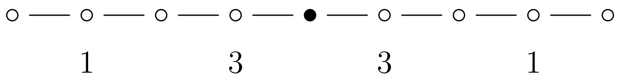
In answer, at $k = 1$ we have 2 such paths, ending at -1 and 1, and the count results can be pictured as follows, with everything being self-explanatory:



At $k = 2$ now, we have 4 paths, one of which ends at -2, two of which end at 0, and one of which ends at 2. The results can be pictured as follows:



At $k = 3$ now, we have 8 paths, the distribution of the endpoints being as follows:



As for $k = 4$, here we have 16 paths, the distribution of the endpoints being as follows:

$$\begin{array}{ccccccccccccccccccccc}
\circ & - & \circ & - & \circ & - & \circ & - & \circ & - & \bullet & - & \circ & - & \circ & - & \circ & - & \circ & - & \circ \\
 & & 1 & & & & 4 & & & & 6 & & & & 4 & & & & 1 & &
\end{array}$$

And good news, we can see in the above the Pascal triangle, namely:

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

Thus, we can answer Question 2.19, in the following way:

Theorem 2.20. *The paths on $\mathbb{Z}$ are counted by the binomial coefficients. In particular, the $2k$-paths based at 0 are counted by the numbers*

$$D_k = \binom{2k}{k}$$

called central binomial coefficients.

Proof. This follows from the above discussion. Indeed, we certainly have the Pascal triangle, and the rest is just a matter of finishing. There are many possible ways here, a straightforward one being that of arguing that the number E_k^l of length k loops $0 \to l$ is subject, due to the binary choice at the end, to the following recurrence relation:

$$E_k^l = E_{k-1}^{l-1} + E_{k-1}^{l+1}$$

But this is exactly the recurrence for the Pascal triangle, as desired. □

As a second example, let us try now to count the loops of $\mathbb{N}$, based at 0. This is something less obvious, and at the experimental level, the result is as follows:

Proposition 2.21. *The Catalan numbers C_k, counting the loops on $\mathbb{N}$ based at 0,*

$$C_k = \#\left\{0 - i_1 - \ldots - i_{2k-1} - 0\right\}$$

are numerically $1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, \ldots$

Proof. To start with, we have indeed $C_1 = 1$, the only loop here being $0 - 1 - 0$. Then we have $C_2 = 2$, due to two possible loops, namely:

$$0 - 1 - 0 - 1 - 0$$

$$0 - 1 - 2 - 1 - 0$$

Then we have $C_3 = 5$, the possible loops here being as follows:

$$0 - 1 - 0 - 1 - 0 - 1 - 0$$

$$0 - 1 - 0 - 1 - 2 - 1 - 0$$

$$0 - 1 - 2 - 1 - 0 - 1 - 0$$

$$0 - 1 - 2 - 1 - 2 - 1 - 0$$

$$0 - 1 - 2 - 3 - 2 - 1 - 0$$

In general, the same method works, with $C_4 = 14$ being left to you, as an exercise, and with C_5 and higher to me, and I will be back with the solution, in due time. □

Obviously, computing the numbers C_k is no easy task, and finding the formula of C_k, out of the data that we have, does not look as an easy task either. So, let us look for other objects counted by the same numbers C_k. With a bit of luck, among these objects some will be easier to count than the others, and this will eventually compute C_k.

This was for the strategy. In practice now, we first have the following result:

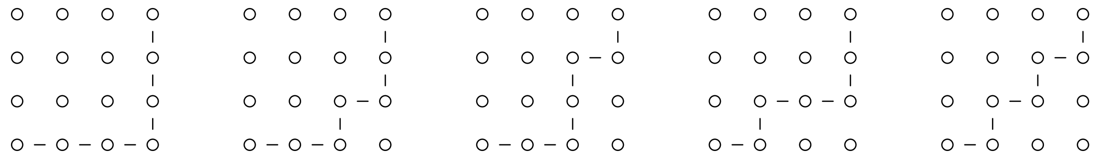
Theorem 2.22. *The Catalan numbers C_k count:*

1. *The length $2k$ loops on $\mathbb{N}$, based at 0.*
2. *The noncrossing pairings of $1, \ldots, 2k$.*
3. *The noncrossing partitions of $1, \ldots, k$.*
4. *The length $2k$ Dyck paths in the plane.*

Proof. All this is standard combinatorics, the idea being as follows:

(1) To start with, in what regards the various objects involved, the length $2k$ loops on $\mathbb{N}$ are the length $2k$ loops on $\mathbb{N}$ that we know, and the same goes for the noncrossing pairings of $1, \ldots, 2k$, and for the noncrossing partitions of $1, \ldots, k$, the idea here being that you must be able to draw the pairing or partition in a noncrossing way.

(2) Regarding now the length $2k$ Dyck paths in the plane, these are by definition the paths from $(0, 0)$ to (k, k), marching North-East over the integer lattice $\mathbb{Z}^2 \subset \mathbb{R}^2$, by staying inside the square $[0, k] \times [0, k]$, and staying as well under the diagonal of this square. As an example, here are the 5 possible Dyck paths at $n = 3$:



(3) Thus, we have definitions for all the objects involved, and in each case, if you start counting them, as we did in Proposition 2.21 with the loops on $\mathbb{N}$, you always end up with the same sequence of numbers, namely those found in Proposition 2.21:

$$1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, \ldots$$

(4) In order to prove now that (1-4) produce indeed the same numbers, many things can be said. The idea is that, leaving aside mathematical brevity, and more specifically abstract reasonings of type $a = b, b = c \implies a = c$, what we have to do, in order to fully understand what is going on, is to etablish $\binom{4}{2} = 6$ equalities, via bijective proofs.

(5) But this can be done, indeed. As an example here, the noncrossing pairings of $1, \ldots, 2k$ from (2) are in bijection with the noncrossing partitions of $1, \ldots, k$ from (3), via fattening the pairings and shrinking the partitions. We will leave the details here as an instructive exercise, and exercise as well, to add (1) and (4) to the picture.

(6) However, matter of having our theorem formally proved, I mean by me professor and not by you student, here is a less elegant argument, which is however very quick, and does the job. The point is that, in each of the cases (1-4) under consideration, the numbers C_k that we get are easily seen to be subject to the following recurrence:

$$C_{k+1} = \sum_{a+b=k} C_a C_b$$

The initial data being the same, namely $C_1 = 1$ and $C_2 = 2$, in each of the cases (1-4) under consideration, we get indeed the same numbers. □

Now we can pass to the second step, namely selecting in the above list the objects that we find the most convenient to count, and count them. This leads to:

Theorem 2.23. *The Catalan numbers are given by the formula*

$$C_k = \frac{1}{k+1}\binom{2k}{k}$$

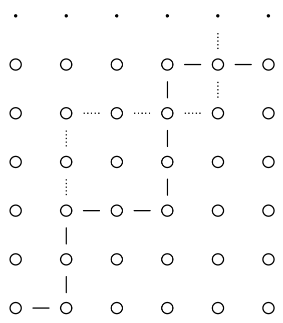
with this being best seen by counting the length $2k$ Dyck paths in the plane.

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

(1) Let us count indeed the Dyck paths in the plane. For this purpose, we use a trick. Indeed, if we ignore the assumption that our path must stay under the diagonal of the square, we have $\binom{2k}{k}$ 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.

(2) 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 $(k, k + 1)$. In order to

count these, the trick is to "flip" their bad part over that higher diagonal, as follows:



(3) Now observe that, as it is obvious on the above picture, due to the flipping, the flipped bad path will no longer end in (k,k), but rather in $(k-1,k+1)$. Moreover, more is true, in the sense that, by thinking a bit, we see that the flipped bad paths are precisely those ending in $(k-1,k+1)$. Thus, we can count these flipped bad paths, and so the bad paths, and so the good paths too, and so good news, we are done.

(4) To finish now, by putting everything together, we have:

$$\begin{aligned}
C_k &= \binom{2k}{k} - \binom{2k}{k-1} \\
&= \binom{2k}{k} - \frac{k}{k+1}\binom{2k}{k} \\
&= \frac{1}{k+1}\binom{2k}{k}
\end{aligned}$$

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

Good work that we did here, and among others, Conjecture 2.18 is now proved. Many other things can be said about the Catalan numbers. We will be back to this.

2d. Binomial laws

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. Let us start with:

Definition 2.24. *A discrete probability space is a set X, usually finite or countable, whose elements $x \in X$ are called events, together with a function*

$$P : X \to [0,\infty)$$

called probability function, which is subject to the condition

$$\sum_{x\in X} P(x) = 1$$

telling us that the overall probability for something to happen is 1.

As a comment here, our condition $\sum_{x\in X}P(x)=1$ perfectly makes sense, and this even if X is uncountable, because the sum of positive numbers is always defined, as a number in $[0,\infty]$, and this no matter how many positive numbers we have.

As a second comment, once we have a probability function $P:X\to[0,\infty)$ as above, with $P(x)\in[0,1]$ telling us what the probability for an event $x\in X$ to happen is, we can compute what the probability for a set of events $Y\subset X$ to happen is, by setting:

$$P(Y)=\sum_{y\in Y}P(y)$$

With this discussed, let us explore now the basic examples, coming from the real life. And here, there are many things to be learned. As a first example, we have:

Example 2.25. *Flipping coins.*

Here things are simple and clear, because when you flip a coin the corresponding discrete probability space, together with its probability measure, is as follows:

$$X=\big\{\text{heads, tails}\big\}\quad,\quad P(\text{heads})=P(\text{tails})=\frac{1}{2}$$

In the case where the coin is biased, as to land on heads with probability $2/3$, and on tails with probability $1/3$, the corresponding probability space is as follows:

$$X=\big\{\text{heads, tails}\big\}\quad,\quad P(\text{heads})=\frac{2}{3}\quad,\quad P(\text{tails})=\frac{1}{3}$$

More generally, given any number $p\in[0,1]$, we have an abstract probability space as follows, where we have replaced heads and tails by win and lose:

$$X=\big\{\text{win, lose}\big\}\quad,\quad P(\text{win})=p\quad,\quad P(\text{lose})=1-p$$

Finally, things become more interesting when flipping a coin, biased or not, several times in a row. We will be back to this in a moment, with details.

Example 2.26. *Rolling dice.*

Again, things here are simple and clear, because when you throw a die the corresponding probability space, together with its probability measure, is as follows:

$$X=\big\{1,\ldots,6\big\}\quad,\quad P(i)=\frac{1}{6}\ ,\ \forall i$$

As before with coins, we can further complicate this by assuming that the die is biased, say landing on face i with probability $p_i\in[0,1]$. In this case the corresponding probability space, together with its probability measure, is as follows:

$$X=\big\{1,\ldots,6\big\}\quad,\quad P(i)=p_i\quad,\quad p_i\geq 0\ ,\ \sum_i p_i=1$$

Also as before with coins, things become more interesting when throwing a die several times in a row, or equivalently, when throwing several identical dice at the same time. In this latter case, with n identically biased dice, the probability space is as follows:

$$X=\Big\{1,\ldots,6\Big\}^n\quad,\quad P(i_1\ldots i_n)=p_{i_1}\ldots p_{i_n}\quad,\quad p_i\geq 0\ ,\ \sum_ip_i=1$$

Observe that the sum 1 condition in Definition 2.24 is indeed satisfied, and with this proving that our dice modeling is bug-free, due to the following computation:

$$\begin{aligned}\sum_{i\in X}P(i)&=\sum_{i_1,\ldots,i_n}P(i_1\ldots i_n)\\&=\sum_{i_1,\ldots,i_n}p_{i_1}\ldots p_{i_n}\\&=\sum_{i_1}p_{i_1}\ldots\sum_{i_n}p_{i_n}\\&=1\end{aligned}$$

Getting back now to theory, in the general context of Definition 2.24, we can see that what we have there is very close to the biased die, from Example 2.26. Indeed, in the general context of Definition 2.24, we can say that what happens is that we have a die with $|X|$ faces, which is biased such that it lands on face i with probability $P(i)$.

Which is something quite interesting, allowing us to have some intuition on what is going on, in discrete probability. So, let us record this finding, as follows:

Conclusion 2.27. *Discrete probability can be understood as being about throwing a general die, having an arbitrary number of faces, and which is arbitrarily biased too.*

Moving ahead now, as usual in probability and statistics, we are mainly interested in winning. But, winning what? In case we are dealing with a usual die, what we win is what the die says, and on average, what we win is the following quantity:

$$E=\frac{1+2+3+4+5+6}{6}=3.5$$

In case we are dealing with the biased die in Example 2.26, again what we win is what the die says, and on average, what we win is the following quantity:

$$E=\sum_ii\times p_i$$

With this understood, what about coins? Here, before doing any computation, we have to assign some numbers to our events, and a standard choice here is as follows:

$$f:\Big\{\text{heads, tails}\Big\}\to\mathbb R\quad,\quad f(\text{heads})=1\quad,\quad f(\text{tails})=0$$

With this choice made, what we can expect to win is the following quantity:

$$\begin{aligned}E(f) &= f(\text{heads}) \times P(\text{heads}) + f(\text{tails}) \times P(\text{tails})\\ &= 1 \times \frac{1}{2} + 0 \times \frac{1}{2}\\ &= \frac{1}{2}\end{aligned}$$

In short, you get the point. In order to do some math, in the context of Definition 2.24, we need a random variable $f : X \to \mathbb{Q}$, and the math will consist in computing the expectation of this variable, $E(f) \in \mathbb{Q}$. Alternatively, in order to do some business in the context of Definition 2.24, we need some form of "money", and our random variable $f : X \to \mathbb{Q}$ will stand for that money, and then $E(f) \in \mathbb{Q}$, for the average gain.

Let us axiomatize this situation as follows:

Definition 2.28. *A random variable on a probability space X is a function*

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

and the expectation of such a random variable is the quantity

$$E(f) = \sum_{x \in X} f(x)P(x)$$

which is best thought as being the average gain, when the game is played.

Here the word "game" refers to the probability space interpretation from Conclusion 2.27. Indeed, in that context, with our discrete set of events X being thought of as corresponding to a generalized die, and by thinking of f as representing some sort of money, the above quantity $E(f)$ is what we win, on average, when playing the game.

As a further piece of basic probability, coming this time as a theorem, we have:

Theorem 2.29. *Given a random variable $f : X \to \mathbb{Q}$, if we define its law as being*

$$\mu = \sum_{x \in X} P(x)\delta_{f(x)}$$

regarded as probability measure on $\mathbb{Q}$, then the expectation is given by the formula

$$E(f) = \int_{\mathbb{Q}} y \, d\mu(y)$$

with the usual convention that each Dirac mass integrates up to 1.

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

(1) To start with, given a random variable $f : X \to \mathbb{Q}$, we can certainly talk about its law μ, as being the formal linear combination of Dirac masses in the statement. Our

claim is that this is a probability measure on $\mathbb{Q}$, in the sense of Definition 2.24. Indeed, the weight of each point $y\in\mathbb{Q}$ is the following quantity, which is positive, as it should:

$$d\mu(y)=\sum_{f(x)=y}P(x)$$

Moreover, the total mass of this measure is 1, as it should, due to:

$$\begin{aligned}\sum_{y\in\mathbb{Q}}d\mu(y)&=\sum_{y\in\mathbb{Q}}\sum_{f(x)=y}P(x)\\&=\sum_{x\in X}P(x)\\&=1\end{aligned}$$

(2) Still talking basics, let us record as well the following alternative formula for the law, which is clear from definitions, and that we will often use, in what follows:

$$\mu=\sum_{y\in\mathbb{Q}}P(f=y)\delta_y$$

(3) Now let us compute the expectation of f. With the usual convention that each Dirac mass integrates up to 1, as mentioned in the statement, we have:

$$\begin{aligned}E(f)&=\sum_{x\in X}P(x)f(x)\\&=\sum_{y\in\mathbb{Q}}y\sum_{f(x)=y}P(x)\\&=\int_{\mathbb{Q}}y\,d\mu(y)\end{aligned}$$

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

Let us talk now about the key notion in probability, which is independence. Motivated by what happens when flipping a biased coin several times in a row, we have:

Definition 2.30. *Given $p\in[0,1]$, the Bernoulli law of parameter p is given by:*

$$P(\text{win})=p\quad,\quad P(\text{lose})=1-p$$

More generally, the k-th binomial law of parameter p, with $k\in\mathbb{N}$, is given by

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

with the Bernoulli law appearing at $k=1$, with $s=1,0$ here standing for win and lose.

Let us try now to understand the relation between the Bernoulli and binomial laws. Indeed, we know that the Bernoulli laws produce the binomial laws, simply by iterating the game, from 1 throw to $k\in\mathbb N$ throws. Obviously, what matters in all this is the "independence" of our coin throws, so let us record this finding, as follows:

Theorem 2.31. *The following happen, in the context of the 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 discuss this, and we will come back later to it, with precise definitions, general theorems and everything:

(1) The idea is to encapsulate the data from Definition 2.30 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. □

Getting to formal mathematical work now, let us start with the following straightforward definition, inspired by what happens for coins and dice:

Definition 2.32. *We say that two variables $f,g:X\to\mathbb{Q}$ are independent when*

$$P(f=x,g=y)=P(f=x)P(g=y)$$

happens, for any $x,y\in\mathbb{Q}$.

As already mentioned, this is something very intuitive, inspired by what happens for coins, dice and cards. As a first result now regarding independence, we have:

Theorem 2.33. *Assuming that $f,g:X\to\mathbb{Q}$ are independent, we have:*

$$E(fg)=E(f)E(g)$$

More generally, we have in fact the following formula, for any $k,l\in\mathbb{N}$,

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

and the converse holds, in the sense that this formula implies the independence of f,g.

Proof. We have indeed the following computation, using the independence of f,g:

$$\begin{aligned}E(f^kg^l)&=\sum_{xy}x^ky^lP(f=x,g=y)\\&=\sum_{xy}x^ky^lP(f=x)P(g=y)\\&=\sum_xx^kP(f=x)\sum_yy^lP(g=y)\\&=E(f^k)E(g^l)\end{aligned}$$

As for the last assertion, this is clear too, because having the above computation work, for any $k,l\in\mathbb{N}$, amounts in saying that the independence formula for f,g holds. □

Regarding now the convolution operation, motivated by what we found before, in Theorem 2.31, let us start with the following abstract definition:

Definition 2.34. *Given a space X with a sum operation $+$, we can define the convolution of any two discrete probability measures on it,*

$$\mu=\sum_ia_i\delta_{x_i}\quad,\quad\nu=\sum_jb_j\delta_{y_j}$$

as being the discrete probability measure given by the following formula:

$$\mu*\nu=\sum_{ij}a_ib_j\delta_{x_i+y_j}$$

That is, the convolution operation $*$ is defined by $\delta_x*\delta_y=\delta_{x+y}$, and linearity.

Observe that this agrees with what we did before with coins, and Bernoulli and binomial laws. We have in fact the following general result, clarifying Theorem 2.31:

THEOREM 2.35. *Assuming that $f,g:X\to\mathbb{Q}$ are independent, we have*

$$\mu_{f+g}=\mu_f*\mu_g$$

where $*$ is the convolution of real probability measures.

PROOF. We have indeed the following straightforward verification:

$$\begin{aligned}\mu_{f+g}&=\sum_{x\in\mathbb{Q}}P(f+g=x)\delta_x\\&=\sum_{y,z\in\mathbb{Q}}P(f=y,g=z)\delta_{y+z}\\&=\sum_{y,z\in\mathbb{Q}}P(f=y)P(g=z)\delta_y*\delta_z\\&=\left(\sum_{y\in\mathbb{Q}}P(f=y)\delta_y\right)*\left(\sum_{z\in\mathbb{Q}}P(g=z)\delta_z\right)\\&=\mu_f*\mu_g\end{aligned}$$

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

2e. Exercises

This was supposed to be a routine chapter on fractions, but we have seen that all this quickly gets us into combinatorics and probability. As exercises on this, we have:

EXERCISE 2.36. *Review if needed the proof of $\frac{a}{b}-\frac{c}{d}=\frac{ad-bc}{bd}$.*

EXERCISE 2.37. *Review also, if needed, the proof of $\frac{a}{b}:\frac{c}{d}=\frac{ad}{bc}$.*

EXERCISE 2.38. *Fill in all the details, for $\frac{a}{b}<\frac{c}{d}\iff ad<bc$.*

EXERCISE 2.39. *Learn about the integer and fractional part of fractions.*

EXERCISE 2.40. *Can we avoid redundancies in our counting method for $\mathbb{Q}_+$?*

EXERCISE 2.41. *Compute some more binomials $\binom{n}{k}$, at small values of k.*

EXERCISE 2.42. *Learn more about the Catalan numbers C_k, as much as you can.*

EXERCISE 2.43. *Find more examples of independent variables, in the real life.*

As bonus exercise, in case you are reading this book with the idea in mind of setting up some business, you can start reading some more systematic probability.

CHAPTER 3

Real numbers

3a. Real numbers

Many things can be done with the rational numbers $\mathbb{Q}$, as we have seen in the above. However, getting straight to the point, one thing that fails is solving $x^2 = 2$:

THEOREM 3.1. *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:

$$\begin{array}{r} 12.456\,783\,872 \\ +\ 27.536\,678\,377 \end{array}$$

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 3.2. *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. Fortunately, there is a clever solution to this, due to Dedekind. His definition is as follows:

Definition 3.3. *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 3.4. *The equation $x^2=2$ has two solutions over the real numbers, namely the positive solution, denoted $\sqrt{2}$, and its negative counterpart, which is $-\sqrt{2}$.*

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

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

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

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

Theorem 3.5. *The solutions of $ax^2+bx+c=0$ with $a,b,c\in\mathbb{R}$ are*

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

provided that $b^2-4ac\geq0$. In the case $b^2-4ac<0$, there are no solutions.

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

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

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

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

THEOREM 3.6. *The real numbers $x \in \mathbb{R}$ can be written in decimal form,*

$$x = \pm a_1 \ldots a_n . b_1 b_2 b_3 \ldots\ldots$$

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 $x = \pm a_1 \ldots a_n . b_1 b_2 b_3 \ldots\ldots$ 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 $[x]$:

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

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.

3b. Limits, series

Time now to get into calculus. Here is what you need to know:

Definition 3.7. *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 3.8. *We have $1/n \to 0$.*

Proof. This is obvious, but let us prove it by using Definition 3.7. 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 3.7, 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 3.7 with:

Definition 3.9. *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 3.10. *We have $n^2 \to \infty$.*

Proof. As before, this is obvious, but let us prove it using Definition 3.9. We have:

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

Thus we can take $N = [\sqrt{K}] + 1$ in Definition 3.9, and we are done. □

We can unify and generalize Proposition 3.8 and Proposition 3.9, as follows:

Proposition 3.11. *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 3.8 and Proposition 3.9, first for a rational, and then for a real as well. □

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

Theorem 3.12. *The following happen:*

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

Proof. All this is elementary, coming from definitions:

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

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

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

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

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

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

(4) This is clear from definitions. □

Here are as well some general rules for computing limits:

Theorem 3.13. *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 3.14. *The $n\to\infty$ limits of quotients of polynomials are given by*

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

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

Proof. The first assertion comes from the following computation:

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

As for the second assertion, this comes from Proposition 3.11. □

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 3.15. *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 3.16. *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. □

Good news, with our current knowledge of the reals, we are now ready to get into some truly interesting mathematics. Let us start with the following definition:

Definition 3.17. *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 a first, basic example of series, which can converge or diverge, we have:

Theorem 3.18. *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 3.19. *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 3.20. *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 3.19, 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 3.21. *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 3.19.

(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 3.20, 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 3.22. *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 3.21, we have:

Theorem 3.23. *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. □

And good news, what we learned in the above will do, as general theory regarding the series. Of course, we will be back from time to time to theory, whenever needed.

3c. The number e

All the above was a bit theoretical, and as something more concrete now, which is at the origins of all modern mathematics, we have the following key result:

THEOREM 3.24. *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 3.25. *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 3.24 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 3.26. *We have the formula*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

3d. 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. In order to discuss this, we need to know more about counting. We first have the following well-known, and useful formula:

Theorem 3.27. *We have the following formula,*

$$\left|\left(\bigcup_i A_i\right)^c\right|=|A|-\sum_i|A_i|+\sum_{i<j}|A_i\cap A_j|-\sum_{i<j<k}|A_i\cap A_j\cap A_k|+\ldots$$

called inclusion-exclusion principle.

Proof. This is indeed quite clear, by thinking a bit, as follows:

(1) In order to count $(\cup_iA_i)^c$, we certainly have to start with $|A|$.

(2) Then, we obviously have to remove each $|A_i|$, and so remove $\sum_i|A_i|$.

(3) But then, we have to put back each $|A_i\cap A_j|$, and so put back $\sum_{i<j}|A_i\cap A_j|$.

(4) Afterwards, we must remove each $|A_i\cap A_j\cap A_k|$, so remove $\sum_{i<j<k}|A_i\cap A_j\cap A_k|$.

$\vdots$

(5) And so on, which leads to the formula in the statement. □

Getting now towards what we wanted to do, in relation with e, let us start with the following definition, which is something very standard:

Definition 3.28. *A permutation of $\{1,\ldots,N\}$ is a bijection, as follows:*

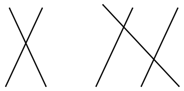
$$\sigma:\{1,\ldots,N\}\to\{1,\ldots,N\}$$

The set of such permutations is denoted S_N.

There are many possible notations for the permutations, the basic one consisting in writing the numbers $1,\ldots,N$, and below them, their permuted versions:

$$\sigma=\begin{pmatrix}1&2&3&4&5\\2&1&4&5&3\end{pmatrix}$$

Another method, which is certainly faster, and which is actually my personal favorite, is by denoting the permutations as diagrams, acting from top to bottom:

$\sigma =$ 

Here are some basic properties of the permutations:

Theorem 3.29. *The permutations have the following properties:*

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

Proof. In order to construct a permutation $\sigma\in S_N$, we have:
– N choices for the value of $\sigma(N)$.
– $(N-1)$ choices for the value of $\sigma(N-1)$.
– $(N-2)$ choices for the value of $\sigma(N-2)$.
$\vdots$
– and so on, up to 1 choice for the value of $\sigma(1)$.
Thus, we have $N!$ choices, as claimed. As for the second assertion, this is clear. □

With this discussed, here is now the application of the inclusion-exclusion principle that we were having in mind, making appear e, in a nice combinatorial way:

Theorem 3.30. *The probability for a random permutation $\sigma\in S_N$ to be a derangement, that is, to have no fixed points, is given by the following formula:*

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

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

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

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

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

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

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. □

In order to further build on this, let us formulate the following key definition:

Definition 3.31. *The Poisson law of parameter* 1 *is the following measure,*

$$p_1=\frac{1}{e}\sum_{k\geq0}\frac{\delta_k}{k!}$$

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

$$p_t=e^{-t}\sum_{k\geq0}\frac{t^k}{k!}\,\delta_k$$

with the letter "p" standing for Poisson.

We are using here, as usual, some simplified notations for these laws. Observe that our laws have indeed mass 1, as they should, due to the following key formula:

$$e^t=\sum_{k\geq0}\frac{t^k}{k!}$$

We will see in the moment why these measures appear a bit everywhere, the reasons for this coming from the Poisson Limit Theorem (PLT), which is closely related to our previous investigations regarding the Bernoulli and binomial laws.

For the moment, let us first develop some general theory, for these Poisson laws. We first have the following theoretical result, regarding them:

Theorem 3.32. *We have the following formula, for any* $s,t>0$,

$$p_s*p_t=p_{s+t}$$

so the Poisson laws form a convolution semigroup.

Proof. By using $\delta_k * \delta_l = \delta_{k+l}$ and the binomial formula, we obtain:

$$\begin{aligned}
p_s * p_t &= e^{-s}\sum_k \frac{s^k}{k!}\,\delta_k * e^{-t}\sum_l \frac{t^l}{l!}\,\delta_l\\
&= e^{-s-t}\sum_n \delta_n \sum_{k+l=n}\frac{s^k t^l}{k!l!}\\
&= e^{-s-t}\sum_n \frac{(s+t)^n}{n!}\,\delta_n\\
&= p_{s+t}
\end{aligned}$$

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

Next in line, we have the following result, which is fundamental as well:

Theorem 3.33. *The Poisson laws appear as formal exponentials*

$$p_t = \sum_k \frac{t^k(\delta_1-\delta_0)^{*k}}{k!}$$

with respect to the convolution of measures $*$.

Proof. By using the binomial formula, the measure on the right is:

$$\begin{aligned}
\mu &= \sum_k \frac{t^k}{k!}\sum_{r+s=k}(-1)^s\frac{k!}{r!s!}\delta_r\\
&= \sum_k t^k \sum_{r+s=k}(-1)^s\frac{\delta_r}{r!s!}\\
&= \sum_r \frac{t^r\delta_r}{r!}\sum_s \frac{(-1)^s t^s}{s!}\\
&= \frac{1}{e^t}\sum_r \frac{t^r\delta_r}{r!}\\
&= p_t
\end{aligned}$$

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

As a third and main result, we have the Poisson Limit Theorem, as follows:

Theorem 3.34 (PLT). *We have the following convergence, in moments,*

$$\left(\left(1-\frac{t}{n}\right)\delta_0 + \frac{t}{n}\,\delta_1\right)^{*n} \to p_t$$

for any $t > 0$.

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

(1) Given a discrete random variable $f : X \to \mathbb{R}$, let us define its formal Fourier transform as being $F(x) = E(e^{ifx})$, with i being an arbitrary number as we know them, real, or even a formal variable. Assuming now that f, g are independent, we have:

$$\begin{aligned}
F_{f+g}(x) &= \int_{\mathbb{R}} e^{ixz} d\mu_{f+g}(z)\\
&= \int_{\mathbb{R}} e^{ixz} d(\mu_f * \mu_g)(z)\\
&= \int_{\mathbb{R}\times\mathbb{R}} e^{ix(y+t)} d\mu_f(y) d\mu_g(t)\\
&= \int_{\mathbb{R}} e^{ixy} d\mu_f(y) \int_{\mathbb{R}} e^{ixt} d\mu_g(t)\\
&= F_f(x) F_g(x)
\end{aligned}$$

(2) Now let us denote by ν_n the measure in the statement, under the convolution sign. We have the following computation, for the formal Fourier transform of the limit:

$$\begin{aligned}
F_{\delta_r}(x) = e^{irx} \quad &\implies \quad F_{\nu_n}(x) = \left(1 - \frac{t}{n}\right) + \frac{t}{n} e^{ix}\\
&\implies \quad F_{\nu_n^{*n}}(x) = \left(\left(1 - \frac{t}{n}\right) + \frac{t}{n} e^{ix}\right)^n\\
&\implies \quad F_{\nu_n^{*n}}(x) = \left(1 + \frac{(e^{ix} - 1)t}{n}\right)^n\\
&\implies \quad F(y) = \exp\left((e^{ix} - 1)t\right)
\end{aligned}$$

(3) On the other hand, the formal Fourier transform of p_t is given by:

$$\begin{aligned}
F_{p_t}(x) &= e^{-t} \sum_k \frac{t^k}{k!} F_{\delta_k}(x)\\
&= e^{-t} \sum_k \frac{t^k}{k!} e^{ikx}\\
&= e^{-t} \sum_k \frac{(e^{ix} t)^k}{k!}\\
&= \exp(-t) \exp(e^{ix} t)\\
&= \exp\left((e^{ix} - 1)t\right)
\end{aligned}$$

(4) Now by comparing (2) and (3), we are led to the conclusion in the statement. □

Many other things can be said here, mixing Bernoulli laws and variables, binomial laws and variables, and Poisson laws and variables, in particular further clarifying the material from chapter 2. For all this, we recommend any specialized probability book.

Getting back now to permutations, we have the following result:

THEOREM 3.35. *The main character of S_N, which counts the fixed points,*

$$\chi(\sigma)=\#\left\{i\in\{1,\ldots,N\}\Big|\sigma(i)=i\right\}$$

follows the Poisson law p_1, in the $N\to\infty$ limit. More generally, the variable

$$\chi_t(\sigma)=\#\left\{i\in\{1,\ldots,[tN]\}\Big|\sigma(i)=i\right\}$$

with $t\in(0,1]$ follows the Poisson law p_t, in the $N\to\infty$ limit.

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

(1) In order to establish the first result in the statement, regarding the main character, we must prove the following formula, for any $r\in\mathbb{N}$, in the $N\to\infty$ limit:

$$P(\chi=r)\simeq\frac{1}{r!e}$$

We already know, from Theorem 3.30, that this formula holds at $r=0$:

$$P(\chi=0)\simeq\frac{1}{e}$$

In the general case, we have to count the permutations $\sigma\in S_N$ having exactly r points. Now since having such a permutation amounts in choosing r points among $1,\ldots,N$, and then permuting the $N-r$ points left, without fixed points allowed, we have:

$$\begin{aligned}\#\left\{\sigma\in S_N\Big|\chi(\sigma)=r\right\}&=\binom{N}{r}\#\left\{\sigma\in S_{N-r}\Big|\chi(\sigma)=0\right\}\\&=\frac{N!}{r!(N-r)!}\#\left\{\sigma\in S_{N-r}\Big|\chi(\sigma)=0\right\}\\&=N!\times\frac{1}{r!}\times\frac{\#\left\{\sigma\in S_{N-r}\Big|\chi(\sigma)=0\right\}}{(N-r)!}\end{aligned}$$

By dividing everything by $N!$, we obtain from this the following formula:

$$\frac{\#\left\{\sigma\in S_N\Big|\chi(\sigma)=r\right\}}{N!}=\frac{1}{r!}\times\frac{\#\left\{\sigma\in S_{N-r}\Big|\chi(\sigma)=0\right\}}{(N-r)!}$$

Now by using the computation at $r = 0$, that we already have, from Theorem 3.30, it follows that with $N \to \infty$ we have the following estimate:

$$\begin{aligned} P(\chi = r) &\simeq \frac{1}{r!} \cdot P(\chi = 0) \\ &\simeq \frac{1}{r!} \cdot \frac{1}{e} \end{aligned}$$

Thus, we obtain as limiting measure the Poisson law of parameter 1, as stated.

(2) Regarding now the second assertion, involving an arbitrary parameter $t \in (0, 1]$, the proof here is similar. To be more precise, by using the inclusion-exclusion principle, as in the proof of Theorem 3.30, we first have the following formula:

$$P(\chi_t = 0) \simeq \frac{1}{e^t}$$

But then, we can generalize this formula, by proceeding as in (1) above, into:

$$P(\chi_t = r) \simeq \frac{t^r}{r! e^t}$$

Thus, we obtain as limiting measure the Poisson law of parameter t, as stated. □

3e. Exercises

As it became customary with this book, this was supposed to be a quite basic chapter, which however ended up getting a bit out of control. As exercises on this, we have:

Exercise 3.36. *Find, or look up, the best algorithm for extracting square roots.*

Exercise 3.37. *Learn about computers, how they avoid, or not, the* $999\ldots$ *issue.*

Exercise 3.38. *Further meditate on the reals, and their various possible definitions.*

Exercise 3.39. *Meditate as well about the various conventions involving* $0, 1, \infty$.

Exercise 3.40. *Learn more about Cauchy sequences, and about* $\bar{\mathbb{Q}} = \mathbb{R}$ *too.*

Exercise 3.41. *Can you sum, pictorially, the series* $\sum_n x^n$*, when* x *is rational?*

Exercise 3.42. *Learn, with full details, the arithmetic-geometric inequality.*

Exercise 3.43. *Learn more about the Poisson laws, and their various properties.*

As bonus exercise, learn more about e, for instance by consulting an alternative book, doing things in a different way from the one here. All approaches are good to know.

CHAPTER 4

Number theory

4a. Decimal writing

Time now for some more advanced number theory. As a first job, let us review the definition of the real numbers. By using the Cauchy criterion for sequences, we have:

Theorem 4.1. *$\mathbb{R}$ is the completion of $\mathbb{Q}$, in the sense that it is the space of Cauchy sequences over $\mathbb{Q}$, identified when the virtual limit is the same, in the sense that:*

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

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

Proof. 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:

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

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

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

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

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

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

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

Very nice all this, so have have two equivalent definitions for the real numbers. Getting back now 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 4.2. *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 4.1. Thus, we get the result.

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

Let us record as well the following result, coming as a useful complement to the above:

THEOREM 4.3. *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. □

As a concrete result now, regarding e, which is more advanced, we have:

THEOREM 4.4. *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. □

4b. Group theory

Getting now to algebraic aspects of the numbers, we would like to talk about groups, fields and other algebraic beasts, which are all intimately related to numbers. Let us first talk about groups. Their definition is something very simple, as follows:

Definition 4.5. *A group is a set G endowed with a multiplication operation*

$$(g,h) \to gh$$

which must satisfy the following conditions:

1. *Associativity: we have, $(gh)k = g(hk)$, for any $g,h,k \in G$.*
2. *Unit: there is an element $1 \in G$ such that $g1 = 1g = g$, for any $g \in G$.*
3. *Inverses: for any $g \in G$ there is $g^{-1} \in G$ such that $gg^{-1} = g^{-1}g = 1$.*

The multiplication law is not necessarily commutative. In the case where it is, in the sense that $gh=hg$, for any $g,h\in G$, we call G abelian, en hommage to Abel, and we usually denote its multiplication, unit and inverse operation as follows:

$$(g,h)\to g+h\quad,\quad 0\in G\quad,\quad g\to -g$$

However, this is not a general rule, and rather the converse is true, in the sense that if a group is denoted as above, this means that the group must be abelian.

There are many examples of groups, with typically all the basic systems of numbers that we know being groups. Here are some standard illustrations, for this fact:

Theorem 4.6. *We have the following groups, and non-groups:*

1. $(\mathbb{Z},+)$ *is a group.*
2. $(\mathbb{Q},+)$, $(\mathbb{R},+)$, *are groups as well.*
3. $(\mathbb{N},+)$ *is not a group.*
4. $(\mathbb{Q}^*,\cdot\,)$ *is a group.*
5. $(\mathbb{R}^*,\cdot\,)$ *is a group as well.*
6. $(\mathbb{N}^*,\cdot\,)$, $(\mathbb{Z}^*,\cdot\,)$ *are not groups.*

Proof. All this is clear from the definition of the groups, as follows:

(1) The group axioms are indeed satisfied for $(\mathbb{Z},+)$, with the group operation being the sum $(g,h)\to g+h$, the unit element being 0, and the inverse map being $g\to -g$. Indeed, the axioms correspond to the following formulae, which are all trivial:

$$(g+h)+k=g+(h+k)$$

$$g+0=0+g=g$$

$$g+(-g)=(-g)+g=0$$

(2) Once again, the group axioms are satisfied for $(\mathbb{Q},+)$ and $(\mathbb{R},+)$, for the same reasons as for $(\mathbb{Z},+)$, and with the remark that for $\mathbb{Q}$ we are using here the fact that the sum of any two rational numbers is a rational number, coming from:

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

(3) In $(\mathbb{N},+)$ the first two group axioms are satisfied, for the same reasons as for $(\mathbb{Z},+)$. However, we do not have inverses, so we do not have a group:

$$-1\notin\mathbb{N}$$

As a comment here, such beasts, which satisfy the first two group axioms, but not the third one, are called semigroups. Thus, $(\mathbb{N},+)$ is a semigroup, which is not a group.

(4) The group axioms are indeed satisfied for $(\mathbb{Q}^*,\cdot\,)$, with the product gh being the usual product, 1 being the usual 1, and g^{-1} being the usual g^{-1}. Observe that we must remove indeed the element $0\in\mathbb{Q}$, because in a group, any element must be invertible:

$$0\notin\mathbb{Q}^*$$

(5) The group axioms are indeed satisfied for $(\mathbb{R}^*,\cdot\,)$ too, with the product gh being again the usual product, 1 being the usual 1, and g^{-1} being the usual g^{-1}.

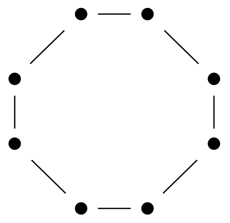
(6) In what regards $(\mathbb{N}^*,\cdot\,),(\mathbb{Z}^*,\cdot\,)$, here the first two group axioms are satisfied, but not the third one, for instance due to the fact that the element 2 has no inverse:

$$\frac{1}{2}\notin\mathbb{Z}^*$$

Thus, both $(\mathbb{N}^*,\cdot\,),(\mathbb{Z}^*,\cdot\,)$ are semigroups, which are not groups. □

Getting now to the finite group case, which is of particular interest in relation with numbers, as a basic example here we have the cyclic group $\mathbb{Z}_N$, constructed as follows:

Definition 4.7. *The cyclic group $\mathbb{Z}_N$ is the group formed by the N rotations of the regular N-gon, with the group operation being the composition of these rotations:*



Alternatively, $\mathbb{Z}_N=\{0,1,2,\ldots,N-1\}$ is the group of remainders modulo N, with the usual addition operation for such remainders.

Here the fact that the above two definitions of $\mathbb{Z}_N$ are indeed equivalent comes from the fact that, with the first approach, if we set $\mathbb{Z}_N=\{R_0,R_1,R_2,\ldots,R_{N-1}\}$, with $R_0=id$, and with $R_1,R_2,\ldots$ being the other rotations, listed in increasing counterclockwise order, the group law is given by $R_kR_l=R_{k+l}$, with $k+l$ taken modulo N. Thus, we have $\mathbb{Z}_N=\{0,1,2,\ldots,N-1\}$, with the group operation being $(k,l)\to k+l$, modulo N.

As a first observation, the above cyclic groups $\mathbb{Z}_N$ are all abelian. We can construct further abelian groups by taking products of such cyclic groups, as follows:

Proposition 4.8. *The following groups are all finite, and abelian,*

$$G=\mathbb{Z}_{N_1}\times\ldots\times\mathbb{Z}_{N_k}$$

for any choice of the numbers $N_1,\ldots,N_k\in\mathbb{N}$.

Proof. This is something trivial, with the obvious definition for ×, coming from the fact that a product of abelian groups must be abelian too. We will see later in this chapter that any finite abelian group must appear as above, as a product of cyclic groups. □

As just mentioned, we will talk more about this later. In the meantime, however, let us examine a bit the groups appearing in Proposition 4.8. In the simplest case, that of a product of two cyclic groups, we have the following useful result:

Theorem 4.9. *Given two integers satisfying $(M,N)=1$, we have:*

$$\mathbb{Z}_M\times\mathbb{Z}_N=\mathbb{Z}_{MN}$$

In the case $(M,N)>1$ this fails, and the group on the left is not cyclic.

Proof. This follows from some basic arithmetic, the idea being as follows:

(1) In order to establish the identification in the statement, consider the following map, which is obviously well-defined, for any two positive integers $M,N\in\mathbb{N}$:

$$f:\mathbb{Z}_M\times\mathbb{Z}_N\to\mathbb{Z}_{MN}\quad,\quad f(a,b)=Na+Mb$$

To be more precise, the fact that f is well-defined comes from the following facts:

$$a=a'(M)\implies Na=Na'(MN)$$

$$b=b'(N)\implies Mb=Mb'(MN)$$

In order to prove now that f is bijective, when $(M,N)=1$, we can invoke a standard result from basic arithmetic, coming as a consequence of the division algorithm for the integers, stating that for $(M,N)=1$, we can always find $p,q\in\mathbb{Z}$ such that:

$$Mp+Nq=1$$

Indeed, this shows that we have $f(q,p)=1$, and by further multiplying everything by a given $c\in\mathbb{Z}_{MN}$, taken arbitrary, we can have $c\in Im(f)$, as follows:

$$\begin{aligned}f(cq,cp)&=Ncq+Mcp\\&=(Nq+Mp)c\\&=c\end{aligned}$$

Thus f is surjective, and since its domain and range has the same cardinality, it is bijective. Regarding now the group theory conditions, we first have:

$$\begin{aligned}f(a+a',b+b')&=N(a+a')+M(b+b')\\&=(Na+Mb)+(Na'+Mb')\\&=f(a,b)+f(a',b')\end{aligned}$$

The condition regarding the units is also satisfied, coming from:

$$f(0,0)=N0+M0=0$$

Finally, the condition regarding the inverses holds too, as shown by:

$$f(-a,-b)=-Na-Mb=-f(a,b)$$

Thus our map $f:\mathbb{Z}_M\times\mathbb{Z}_N\to\mathbb{Z}_{MN}$ is indeed a group identification, as desired.

(2) With this discussed, let us prove now the second assertion, stating that when assuming $(M,N)>1$, we have a non-identification, as follows:

$$\mathbb{Z}_M\times\mathbb{Z}_N\neq\mathbb{Z}_{MN}$$

But, how to prove such things? Well, we must trick. Assume by contradiction that we have identification, and with the identification, from right to left, being as follows:

$$f:\mathbb{Z}_{MN}\to\mathbb{Z}_M\times\mathbb{Z}_N\quad,\quad f(1)=(a,b)$$

Now if we set $K=[M,N]<MN$, then we have, with K terms in the sum:

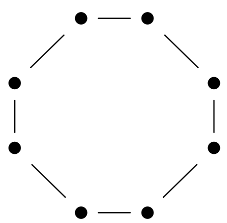
$$\begin{aligned}f(K)&=(a,b)+\ldots+(a,b)\\&=(Ka,Kb)\\&=(0,0)\end{aligned}$$

But this is a contradiction, because the elements $K\neq 0$ and 0 have the same image under f, and so f cannot be injective. Thus, non-identification claim proved, and the very last assertion follows too, because $\mathbb{Z}_M\times\mathbb{Z}_N$ having MN elements, and not being equal to the only cyclic group having MN elements, namely $\mathbb{Z}_{MN}$, it cannot be cyclic. □

As a second basic example of a finite group, we have the symmetric group S_N. We already met and studied this group in chapter 3, and for the moment, this will do.

Moving on, as a third basic example of finite group, lying in complexity somewhere between $\mathbb{Z}_N$ and S_N, we have the dihedral group D_N, which appears as follows:

DEFINITION 4.10. *The dihedral group D_N is the symmetry group of*



that is, of the regular polygon having N vertices.

Many interesting things can be said about D_N. To start with, we have $D_2=\mathbb{Z}_2$, and $D_3=S_3$. In general, the dihedral group D_N has $2N$ elements, as follows:

– First we have the N rotations $R_1,\ldots,R_N\in\mathbb{Z}_N$, with R_k being the rotation of angle $360°k/N$. When labeling the vertices of the N-gon $1,\ldots,N$ we have $R_k:i\to k+i$.

– Then we have N symmetries $S_1, \ldots, S_N$, with S_k being the symmetry with respect to the Ox axis rotated by $180° k/N$. The symmetry formula is $S_k : i \to k - i$.

Now let us see how these rotations and symmetries multiply. We have:

$$\begin{aligned} R_k R_l &: i \to l + i \to k + l + i \\ R_k S_l &: i \to l - i \to k + l - i \\ S_k R_l &: i \to l + i \to k - l - i \\ S_k S_l &: i \to l - i \to k - l + i \end{aligned}$$

We conclude that we can talk about D_N abstractly, if we want to, as follows:

THEOREM 4.11. *D_N is the group formed by $R_1, \ldots, R_N$ and $S_1, \ldots, S_N$, with*

$$R_k R_l = R_{k+l} \quad , \quad R_k S_l = S_{k+l}$$

$$S_k R_l = S_{k-l} \quad , \quad S_k S_l = R_{k-l}$$

being the multiplication formulae for these group elements.

PROOF. This follows indeed from the above discussion. □

As a continuation of this, observe that $\mathbb{Z}_N \times \mathbb{Z}_2$ is the group having $2N$ elements, $r_1, \ldots, r_N$ and $s_1, \ldots, s_N$, which multiply according to the following rules:

$$r_k r_l = r_{k+l} \quad , \quad r_k s_l = s_{k+l}$$

$$s_k r_l = s_{k+l} \quad , \quad s_k s_l = r_{k+l}$$

We conclude that D_N must appear as some sort of "twisted version" of $\mathbb{Z}_N \times \mathbb{Z}_2$, and with a bit of algebraic know-how, we can even formulate a nice finding, as follows:

FACT 4.12. *The dihedral group D_N can be decomposed as*

$$D_N = \mathbb{Z}_N \rtimes \mathbb{Z}_2$$

with $\rtimes$ being a sort of "twisted version" of the usual $\times$ operation.

And I will leave it to you to learn more about this, from any standard algebra book, such as Lang [60]. Getting now to some general theory, we first have:

THEOREM 4.13. *Given a finite group G and a subgroup $H \subset G$, the sets*

$$G/H = \left\{ gH \Big| g \in G \right\} \quad , \quad H \backslash G = \left\{ Hg \Big| g \in G \right\}$$

consist of partitions of G into subsets of size H, and we have the following formula:

$$|G/H| = |H \backslash G| = \frac{|G|}{|H|}$$

In particular, the order of the subgroup divides the order of the group, $|H| \Big| |G|$.

Proof. The partition claim for the set G/H constructed in the statement can be deduced as follows, and the proof for $H\backslash G$ is similar:

$$gH\cap kH\neq\emptyset\iff g^{-1}k\in H\iff gH=kH$$

But with this in hand, the cardinality formulae are all clear. □

As a continuation of the above, which is something fundamental, we have:

Theorem 4.14. *Given a subgroup $H\subset G$ which is normal, in the sense that*

$$gH=Hg\quad,\quad\forall g\in G$$

the space $G/H=H\backslash G$ is a group, with multiplication $(gH)(sH)=gsH$.

Proof. Assume indeed that $H\subset G$ is normal, and that g,k,s,t are such that:

$$gH=kH\quad,\quad sH=tH$$

We have then the following computation, by using the normality condition:

$$gsH=gtH=gHt=kHt=ktH$$

Thus $G/H=H\backslash G$ is a indeed group, with multiplication $(gH)(sH)=gsH$. □

As another continuation of Theorem 4.13, which is something fundamental too, we can talk about the order of group elements, inside any finite group, as follows:

Theorem 4.15. *Given a finite group G, any $g\in G$ generates a cyclic subgroup*

$$<g>=\{1,g,g^2,\ldots,g^{k-1}\}$$

with $k=ord(g)$ being the smallest number $k\in\mathbb N$ satisfying $g^k=1$. Also, we have

$$ord(g)\Big|\,|G|$$

that is, the order of any group element divides the order of the group.

Proof. In order to prove the first assertion, let $g\in G$, and consider the semigroup $<g>\subset G$ formed by the sequence of powers of g:

$$<g>=\{1,g,g^2,g^3,\ldots\}\subset G$$

Since G was assumed to be finite, the sequence of powers must cycle, $g^n=g^m$ for some $n<m$. But this shows that $g^k=1$, with $k=m-n$, which gives:

$$<g>=\{1,g,g^2,\ldots,g^{k-1}\}$$

Moreover, we can choose the number $k\in\mathbb N$ to be minimal with this property, and with this choice, we have a set without repetitions. Thus $<g>\subset G$ is indeed a group, and more specifically a cyclic group, whose order is as follows:

$$|<g>|=k=ord(g)$$

Thus, we proved the first assertion, and with this in hand, the second assertion, namely $ord(g)\,|\,|G|$, follows from Theorem 4.13, applied to the subgroup $<g>\subset G$. □

As a main result now about groups, dealing with the abelian case, we have:

Theorem 4.16. *The finite abelian groups are the following groups:*

$$G=\mathbb{Z}_{N_1}\times\ldots\times\mathbb{Z}_{N_k}$$

That is, the finite abelian groups are the products of cyclic groups.

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

(1) In order to prove our result, assume that G is finite and abelian. For any prime number $p\in\mathbb{N}$, let us define $G_p\subset G$ to be the subset of elements having as order a power of p. Equivalently, this subset $G_p\subset G$ can be defined as follows:

$$G_p=\left\{g\in G\Big|\exists k\in\mathbb{N},g^{p^k}=1\right\}$$

(2) It is then routine to check, based on definitions, that each G_p is a subgroup. Our claim now is that we have a direct product decomposition as follows:

$$G=\prod_p G_p$$

(3) Indeed, by using the fact that our group G is abelian, we have a group map as follows, with the order of the factors when computing $\prod_p g_p$ being irrelevant:

$$\prod_p G_p\to G\quad,\quad (g_p)\to\prod_p g_p$$

Moreover, it is routine to check that this group map is both injective and surjective, via some simple manipulations, so we have our group decomposition, as in (2).

(4) Thus, we are left with proving that each component G_p decomposes as a product of cyclic groups, having as orders powers of p, as follows:

$$G_p=\mathbb{Z}_{p^{r_1}}\times\ldots\times\mathbb{Z}_{p^{r_s}}$$

But this is something that can be checked by recurrence on $|G_p|$, via some routine computations, and so we are led to the conclusion in the statement. See Lang [60]. □

4c. Finite fields

Still in relation with numbers, we would like to talk now about fields. Let us start with the following key theorem of Fermat, for the usual integers:

Theorem 4.17. *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. □

The Fermat theorem is particularly interesting when extended from the integers to the arbitrary field case. In order to discuss this question, let us start with:

Theorem 4.18. *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 4.19. *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, in the lack of anything obvious, we have the following theorem, which comes to the rescue:

Theorem 4.20. *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 arithmetic, 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 4.21. *Given a finite field F, with $q=|F|$ we have*

$$a^q=a$$

for any $a\in F$.

Proof. According to Theorem 4.20 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 4.22. *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 4.23. *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 4.18 must be an equality. 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. For details here, we recommend any solid abstract algebra book, such as Lang [60]. □

4d. Legendre symbol

We would like to end this chapter on number theory with a discussion about squares. Of particular interest is the equation $a = b^2(c)$, and in relation with this, we have the following definition, putting everything on a solid basis:

DEFINITION 4.24. *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. And here, as a first result on the subject, due to Euler, we have:

THEOREM 4.25. *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 4.24, we obtain the formula in the statement. □

As a first consequence of the Euler formula, we have the following result:

Proposition 4.26. *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 multiplicative in $a\in\mathbb{Z}$. Alternatively, this can be proved as well directly, exercise for you. □

The above result looks quite conceptual, and as consequences, we have:

Proposition 4.27. *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 4.26, 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 4.26. □

Getting now to the explicit computation of the Legendre symbol, many things can be said here, with the central result, which is something quite heavy, being as follows:

Theorem 4.28. *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):

$$Z = (-1)^n \left(\frac{q-1}{2}\right)!$$

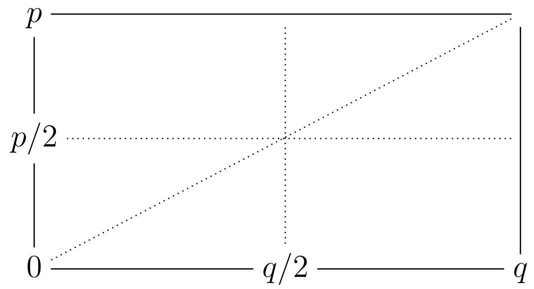
(5) But this is what we need, because when comparing with what we have in (2), we obtain the following formula, which is exactly the one claimed by the Gauss lemma:

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

(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. □

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:

$$\begin{aligned}\left(\frac{3}{173}\right)&=(-1)^{\frac{3-1}{2}\cdot\frac{173-1}{2}}\left(\frac{173}{3}\right)\\&=\left(\frac{173}{3}\right)\\&=\left(\frac{2}{3}\right)\\&=-1\end{aligned}$$

Besides Proposition 4.26, the quadratic reciprocity formula comes accompanied by two other statements, which are very useful in practice. First, at $a=-1$, we have:

PROPOSITION 4.29. *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 4.30. *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 improvise, and prove the result. The proof will come via the following formula, which is equivalent to the formula in the statement:

$$\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}$$

(3) Getting started now, let us be crazy, and introduce a formal number i satisfying $i^2 = -1$, then a formal number w satisfying $w^2 = i$, and then set $t = w + w^{-1}$:

$$i^2 = -1 \quad , \quad w^2 = i \quad , \quad t = w + w^{-1}$$

(4) As a remark that you might have, in case you know a bit about complex numbers, you might say that the above is not that crazy, but rather stupid, because $t = \sqrt{2}$. In answer, yes I know, but it is better to forget this, and do formal arithmetic instead, with integers as scalars, based on our rules above, and the following computation:

$$\begin{aligned} t^2 &= 2 + w^2 + w^{-2} \\ &= 2 + i - i \\ &= 2 \end{aligned}$$

(5) Now by using the Euler formula for the Legendre symbol, we have:

$$\begin{aligned} \left(\frac{2}{p}\right) &= 2^{\frac{p-1}{2}} \ (p) \\ &= (t^2)^{\frac{p-1}{2}} \ (p) \\ &= t^{p-1} \ (p) \end{aligned}$$

(6) 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:

$$\left(\frac{2}{p}\right) t = t^p \ (p)$$

(7) 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:

$$\begin{aligned}t^p &= (w+w^{-1})^p\\ &= \sum_{k=0}^p\binom{k}{p}w^kw^{k-p}\\ &= w^p+w^{-p}\ (p)\end{aligned}$$

(8) Now let us look at the quantity w^p+w^{-p}. Since we have $w^4=-1$, this quantity will depend only on p modulo 8, and more precisely, we have:

$$w^p+w^{-p}=\begin{cases}w+w^{-1}&\text{if } p=\pm1(8)\\ -w-w^{-1}&\text{if } p=\pm3(8)\end{cases}$$

Thus $w^p+w^{-p}=\pm t$, with the sign depending on p modulo 8, and more specifically:

$$w^p+w^{-p}=(-1)^{\frac{p^2-1}{8}}t$$

(9) Time now to put everything together. By combining (6,7,8) 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. □

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 4.29 plus Theorem 4.30 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 4.28. We first have:

PROPOSITION 4.31. *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 $p\geq5$.

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 4.32. *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, we have the following generalization of the Legendre symbol:

Theorem 4.33. *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 4.34. *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. In practice, 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:

$$\left(\frac{a}{-1}\right)=\begin{cases}1&\text{if }a\geq 0\\-1&\text{if }a<0\end{cases}$$

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 solving the rest of the puzzle as an instructive exercise. □

4e. Exercises

This was our first truly advanced chapter, and as exercises on this, we have:

Exercise 4.35. *Write a short essay on the reals, introduced via Cauchy sequences.*

Exercise 4.36. *Write as well an essay, with the reals introduced via decimal form.*

Exercise 4.37. *Compute decimals of e, no calculators allowed, as many as you can.*

Exercise 4.38. *Write a formula of type $D_N=\mathbb{Z}_N\rtimes\mathbb{Z}_2$, based on the above.*

Exercise 4.39. *Learn more about finite fields, history, and some details too.*

Exercise 4.40. *Look up and learn other proofs of quadratic reciprocity.*

Exercise 4.41. *Fill in the details, for the theory of the Jacobi symbols.*

Exercise 4.42. *Fill in the details, for the theory of Kronecker symbols.*

As bonus exercise, and no surprise here, start reading a number theory book.

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 will be 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 it, 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.

Moving ahead now, as a natural question, do any two lines $K \neq L$ determine a point? Normally yes, because assuming $P, Q \in K \cap L$ we would have $K = L = PQ$, contradiction.

However, it might happen that these distinct lines $K\neq L$ are parallel, $K||L$, in which case we have $K\cap L=\emptyset$. In order to further discuss this, let us formulate:

Definition 5.3. *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 $K\not\parallel L$.

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.

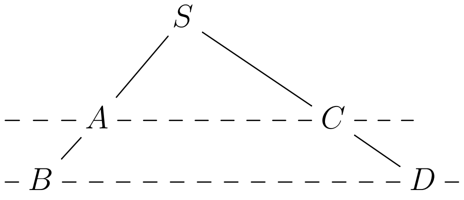
Very good, and now with Axiom 5.1 and Definition 5.3, we are potentially ready for doing some geometry. However, this is not exactly true, and we will need as well:

Axiom 5.4. *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 $P\in K$, $K||L$.*

To be more precise, this is again something which obviously holds, but cannot be established, as a theorem. I mean just try, and you will see that you will fail. As before with Axiom 5.1, we will leave as an exercise some 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. We first have:

Theorem 5.5 (Thales). *Proportions are kept, along parallel lines. That is, given a configuration as follows, consisting of two parallel lines, and of two extra lines,*



the following equality holds:

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

Moreover, the converse holds too, in the sense that this implies $AC||BD$.

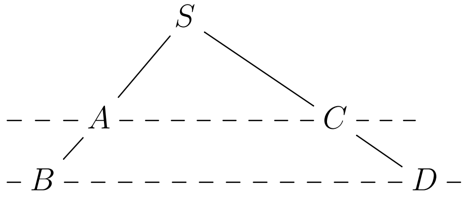
Proof. We have indeed the following computation, based on the usual area formula for the triangles, that is, half of side times height, used multiple times:

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

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

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

Theorem 5.6 (Thales 2). *In the context of the Thales theorem configuration,*

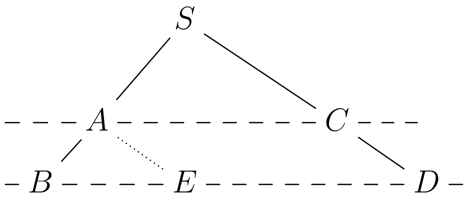


the following equality, involving the same number, holds as well:

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

However, the converse of this does not necessarily hold.

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



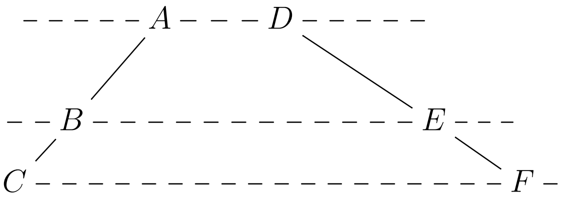
By using Theorem 5.5, we have then the following computation, as desired:

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

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

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

THEOREM 5.7 (Thales 3). *Given a configuration as follows, consisting of three parallel lines, and of two extra lines, which can cross or not,*



the following equality holds:

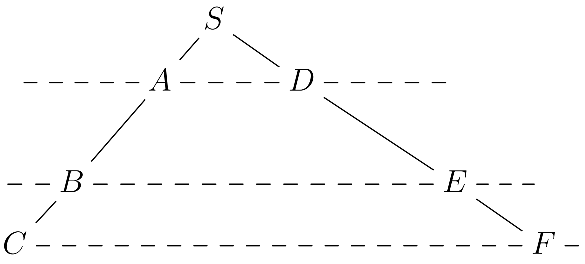
$$\frac{AB}{BC}=\frac{DE}{EF}$$

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

PROOF. We have two cases here, as follows:

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

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



Now by using Theorem 5.5 several times, we obtain:

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

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

Summarizing, 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 many more illustrations for this general principle in this chapter.

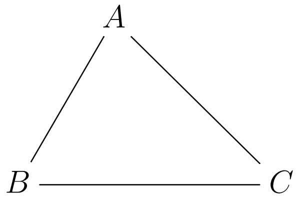
5b. Angles, triangles

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

THEOREM 5.8. *Given a triangle ABC, the following happen:*

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

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



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

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

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

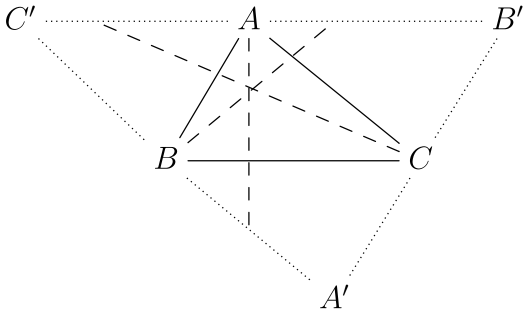
$$P=\frac{1}{3}\cdot A+\frac{2}{3}\cdot\frac{B+C}{2}$$

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

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

(4) This is tougher, and I must admit that, when writing this book, I first struggled a bit with this, then ended looking it up on the internet. So, here is the trick. Draw a

parallel to BC at A, and similarly, parallels to AB and AC at C and B. You will get in this way a bigger triangle, upside-down, $A'B'C'$. But then, the circumcenter of $A'B'C'$, that we know to exist from (3), will be the orthocenter of ABC:

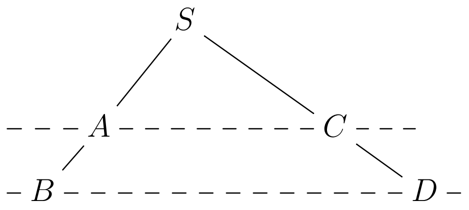


Thus, we are led to the conclusions in the statement. □

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

Fact 5.9. *We can talk about the angle between two crossing lines, and have some basic theory for the angles going, by using triangles, and Thales, in the obvious way.*

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



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

Definition 5.10. *We say that two triangles are similar, and we write*

$$SAC \sim SBD$$

when their respective angles are equal.

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

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

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

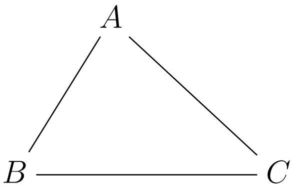
Definition 5.11. *We can talk about the numeric value of angles, as follows:*

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

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

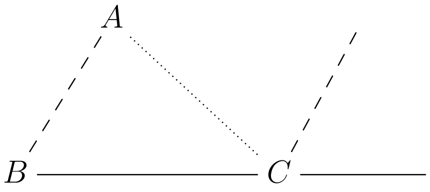
Getting back to work now, theorems and proofs, in relation with the above, here is a key result, which will be our main tool for the study of the angles:

Theorem 5.12. *In an arbitrary triangle*



the sum of all three angles is 180°.

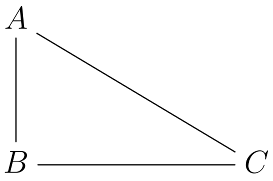
Proof. This does not seem obvious to prove, with bare hands, but as usual, in such situations, some tricky parallels can come to the rescue. Let us prolong indeed the segment BC a bit, on the C side, and then draw a parallel at C, to the line AB, as follows:



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

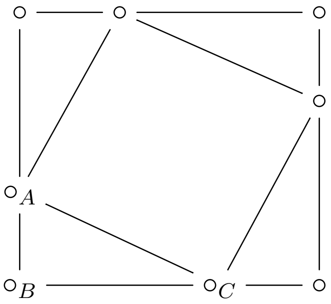
Going ahead now with our study of angles, as a continuation of the above, let us first talk about the simplest angle of them all, which is the right angle, denoted 90°. A triangle having one of the angles equal to 90° is called right triangle, and we have:

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:

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

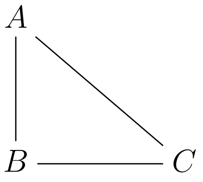
On the other hand, the outer square is made of the smaller square, having side AC, and of four identical right triangles, having sizes AB, BC. Thus:

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

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

As an interesting consequence, making a link with $\sqrt{2}$ and numbers, we have:

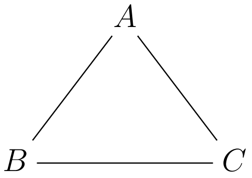
PROPOSITION 5.14. *In a right triangle ABC, with $AB = BC$,*



the small angles are 45°*, and if* $AB = BC = 1$ *then* $AC = \sqrt{2}$.

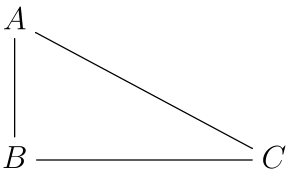
PROOF. The first assertion is clear indeed from the fact that the sum of all angles is 180°. As for the second assertion, this comes from Pythagoras. □

As a next interesting angle, we have the 60° angle. This usually appears in the context of the equilateral triangles, that is, of those triangles having all sides equal:



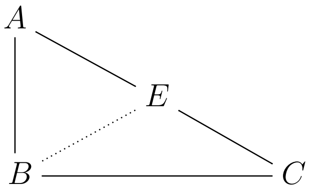
Indeed, the angles being equal, and summing up to 180°, they must be all equal to 60°. Many things can be said about 60°, and about $90° - 60° = 30°$ too, including:

PROPOSITION 5.15. *In a right triangle having small angles* $30°, 60°$*,*



we have $AC = 2AB$*. Also, we have* $BC = \sqrt{3}AB$*.*

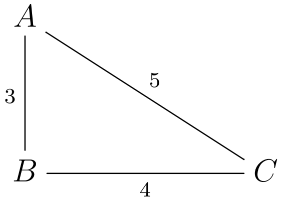
PROOF. The first assertion comes from an equilateral triangle, as follows:

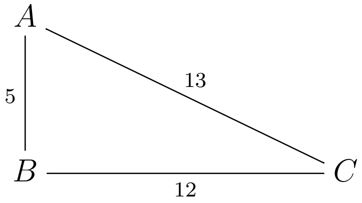


As for the second assertion, this comes from Pythagoras, via $1 + 3 = 4$. □

Still talking Pythagoras, as a concrete and useful application, we have:

THEOREM 5.16. *A triangle having sides* $3, 4, 5$*, or having sides* $5, 12, 13$*, must be a right triangle:*





Thus, for drawing right angles, you only need a loop, with 12 *or* 30 *knots on it.*

PROOF. These assertions both come from the Pythagoras theorem, or rather from its converse, which is clear from it, and from the following equalities:

$$9+16=25\quad,\quad 25+144=169$$

As for the second assertion, this is a standard application to engineering. □

Along the same lines, at a more advanced level, we have the following result, which fully closes the discussion, regarding the Pythagoras equation over the integers:

THEOREM 5.17. *The Pythagoras equation, namely*

$$a^2+b^2=c^2$$

can be fully solved over the integers, the solutions being

$$a=d(m^2-n^2)\quad,\quad b=2dmn\quad,\quad c=d(m^2+n^2)$$

with $(m,n)=1$, up to exchanging a,b.

PROOF. This is something standard, due to Euclid, the idea being as follows:

(1) Let us try to solve $a^2+b^2=c^2$. If we divide a,b,c by their greatest common divisor $d=(a,b,c)$, the equation is still satisfied. Thus, we can assume $(a,b,c)=1$, and we want to prove that the solutions are as follows, up to exchanging a,b:

$$a=m^2-n^2\quad,\quad b=2mn\quad,\quad c=m^2+n^2$$

(2) To start with, in one sense our result is clear, because given any two numbers m,n, the above formulae produce a solution to our equation, as shown by:

$$\begin{aligned}(m^2-n^2)^2+(2mn)^2 &= m^4+n^4-2m^2n^2+4m^2n^2\\ &= m^4+n^4+2m^2n^2\\ &= (m^2+n^2)^2\end{aligned}$$

(3) So, we must prove now the converse, stating that if a,b,c satisfying $(a,b,c)=1$ are solutions of $a^2+b^2=c^2$, then we can write them as in (1). For this purpose, the first observation is that, due to $a^2+b^2=c^2$, our assumption $(a,b,c)=1$ implies:

$$(a,b)=(a,c)=(b,c)=1$$

(4) Let us study now the parity of a,b,c. Since $(a,b)=1$, one of these two numbers, say a, is odd. Now assuming that b is odd too, we would get $a^2+b^2=2(4)$, which is impossible, due to $a^2+b^2=c^2$. Thus b must be even, and as a conclusion to this study, up to exchanging a,b, we can assume that the parity of our numbers is as follows:

$$a=\text{odd}\quad,\quad b=\text{even}\quad,\quad c=\text{odd}$$

(5) Now comes the trick. We can rewrite our equation in the following way:

$$\begin{aligned} a^2+b^2=c^2 &\iff b^2=c^2-a^2 \\ &\iff b^2-(c-a)(c+a) \\ &\iff \frac{c+a}{b}=\frac{b}{c-a} \end{aligned}$$

(6) With this done, let us look at the fraction on the left. This is a rational number, so we can write it in reduced form, as follows, with $(m,n)=1$:

$$\frac{c+a}{b}=\frac{m}{n}$$

Now observe that our equation, as reformulated in (5), takes the following form:

$$\frac{c+a}{b}=\frac{m}{n}\quad,\quad\frac{c-a}{b}=\frac{n}{m}$$

Equivalently, our equation, as reformulated in (5), takes the following form:

$$\frac{c}{b}+\frac{a}{b}=\frac{m}{n}\quad,\quad\frac{c}{b}-\frac{a}{b}=\frac{n}{m}$$

But this latter system is equivalent to the following two formulae:

$$\frac{a}{b}=\frac{1}{2}\left(\frac{m}{n}-\frac{m}{n}\right)=\frac{m^2-n^2}{2mn}$$

$$\frac{c}{b}=\frac{1}{2}\left(\frac{m}{n}+\frac{m}{n}\right)=\frac{m^2+n^2}{2mn}$$

(7) Good work that we did, and time to breathe, and see what we have. We have proved so far that if a,b,c satisfying $(a,b,c)=1$ are solutions of $a^2+b^2=c^2$, then up to exchanging a,b, we can find numbers m,n satisfying $(m,n)=1$, such that:

$$\frac{a}{b}=\frac{m^2-n^2}{2mn}\quad,\quad\frac{c}{b}=\frac{m^2+n^2}{2mn}$$

Which sounds nice, because due to $(a,b)=(b,c)=1$, as noted in (3), the two fractions on the left are in reduced form. So, if we manage to prove that the two fractions on the right are in reduced form too, this would finish the proof, because we would get:

$$a=m^2-n^2\quad,\quad b=2mn\quad,\quad c=m^2+n^2$$

(8) So, let us look now at the two fractions on the right, appearing above. As a first observation, due to $(m,n)=1$, the following two fractions are in reduced form:

$$\frac{m^2-n^2}{mn}\quad,\quad\frac{m^2+n^2}{mn}$$

The problem, however, is that the fractions in (7) are the halves of these quantities. So, all we need is a study modulo 2, and with this, normally done.

(9) Getting now to the endgame, from $(m,n)=1$, the case where both m,n are even is excluded. But the case where both m,n are odd is excluded too, due to:

$$\frac{a}{b}=\frac{m^2-n^2}{2mn}$$

Indeed, if m,n were both to be odd, we would have $m^2-n^2=0(4)$ and $2mn=2(4)$, so the fraction on the right, when reduced, would have an even denominator. But this would tell us that b must be even, which contradicts our b odd choice from (4).

(10) Summarizing, one of the numbers m,n must be even, and the other must be odd. But this does the job, because it shows that m^2-n^2 and m^2+n^2 are both odd, so when dividing the reduced fractions from (7) by 2, these fractions remain still reduced. Thus, as a conclusion to our study, the following two fractions are reduced:

$$\frac{m^2-n^2}{2mn}\quad,\quad\frac{m^2+n^2}{2mn}$$

(11) So, theorem proved. Indeed, as indicated in (7), let us look now at:

$$\frac{a}{b}=\frac{m^2-n^2}{2mn}\quad,\quad\frac{c}{b}=\frac{m^2+n^2}{2mn}$$

Since all fractions appearing here are in reduced form, we obtain from this:

$$a=m^2-n^2\quad,\quad b=2mn\quad,\quad c=m^2+n^2$$

And finally, as indicated in (1), by multiplying a,b,c by an arbitrary number d, we obtain the general solutions from the statement, namely:

$$a=d(m^2-n^2)\quad,\quad b=2dmn\quad,\quad c=d(m^2+n^2)$$

(12) At the level of the interesting examples now, there are of course many of them, and we have for instance a solution coming as follows:

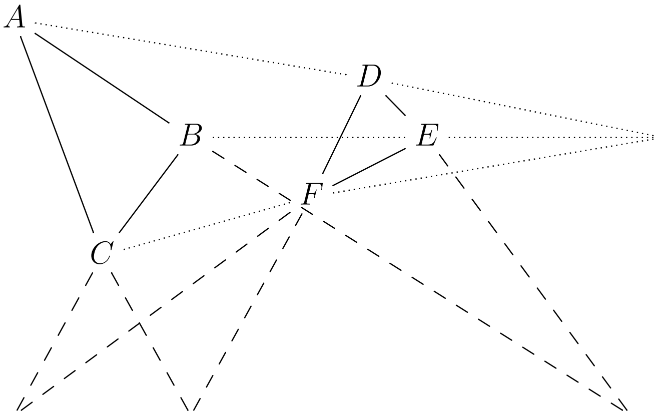
$$9^2+40^2=1681=41^2$$

Which is quite interesting for engineering purposes, in view of Theorem 5.16. Indeed, if our 12-knot device is not accurate enough for our problem, and the 30-knot device is not accurate either, we can come up with a 90-knot device, based on the above solution. □

5c. Desargues, Pappus

Back now to points and lines, as a basic statement, which is something quite subtle, due to Desargues, we have the following fact, that we will prove in what follows:

Fact 5.18 (Desargues). *Two triangles are in perspective centrally if and only if they are in perspective axially. That is, in the context of a configuration of type*



the lines AD, BE, CF cross, so that ABC, DEF are in central perspective, if and only if $AB \cap DE, AC \cap DF, BC \cap EF$ are collinear, so that ABC, DEF are in axial perspective.

Obviously, this is something that can be very useful for various technical computations and drawings, and more on this later. Getting now to the proof of the result, this is something quite tricky. So, with a bit of imagination, we first have:

Theorem 5.19. *The Desargues claim holds in one sense: central perspectivity implies axial perspectivity.*

Proof. The trick here is to pass in 3D, as follows:

(1) Assume first that we are in 3D, with our triangles ABC and DEF lying in distinct planes, say $ABC \subset P$ and $DEF \subset Q$. Assuming central perspectivity, the lines AD, BE cross, so the points A, B, D, E are coplanar. But this tells us that the lines AB, DE cross, and that, in addition, their crossing point lies on the intersection of the planes P, Q:

$$(AB \cap DE) \in P \cap Q$$

But a similar argument, again using central perspectivity, shows that we have also:

$$(AC \cap DF) \in P \cap Q \quad , \quad (BC \cap EF) \in P \cap Q$$

Now since the intersection $P \cap Q$ is a certain line in space, we obtain the result.

(2) Thus, almost there, with the theorem proved when the triangles ABC and DEF are both in 3D, in generic position, and the rest is just a matter of finishing. Indeed, when ABC and DEF are still in 3D, but this time lying in the same plane, the result follows too, by perturbing a bit our configuration, as to make it generic. And with this we are done indeed, because we are now in 2D, exactly as in the setting of the theorem. □

In order to prove now to converse, there are several methods and tricks available. We can use for instance the following result, which is something having its own interest:

Theorem 5.20. *We have a duality between points and lines, obtained by fixing a circle in the plane, say of center O and radius $r > 0$, and doing the following,*

(1) *Given a point P, construct Q on the line OP, as to have $OP \cdot OQ = r^2$,*

(2) *Draw the perpendicular at Q on the line OQ. This is the dual line p,*

and this duality $P \leftrightarrow p$ transforms collinear points into concurrent lines.

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

(1) In order to establish the result, the idea will be that of proving that we have the following implication, with $P_n \leftrightarrow p$ and $L \leftrightarrow l$ being instances of our duality:

$$P_1, \ldots, P_n \in l \implies L \in p_1, \ldots, p_n$$

But here, we can assume $n = 1$. Thus, we must prove that the following happens:

$$P \in l \implies L \in p$$

(2) In order to prove now this latter fact, given a point P, construct its dual line $P \leftrightarrow p$ via a point Q as in the statement, satisfying the following formula:

$$OP \cdot OQ = r^2$$

Now assuming $P \in l$, as above, let us construct the dual point $L \leftrightarrow l$, by projecting O on the line l, into a point $R \in l$, and then requiring that $L \in OR$ must satisfy:

$$OL \cdot OR = r^2$$

With these constructions made, we want to prove that the following happens:

$$L \in p$$

(3) But this is best seen by considering the following intersection point:

$$S = p \cap OR$$

Indeed, we can see that we have two similar triangles appearing, as follows:

$$OPR \sim OSQ$$

Now by using the Thales theorem, we obtain the following formula:

$$\frac{OP}{OR} = \frac{OS}{OQ}$$

(4) But this formula can be written as follows, using $OP \cdot OQ = r^2$:

$$OS \cdot OR = OP \cdot OQ = r^2$$

Now by comparing with $OL \cdot OR = r^2$, we conclude that we have:

$$L = S$$

Now since $S \in p$ by definition, we have $L \in p$, which proves our claim in (2). □

The point now is that the Desargues configuration is self-dual, so we obtain:

Theorem 5.21. *The Desargues claim holds in the other sense too: axial perspectivity implies central perspectivity.*

Proof. Let us look at the Desargues configuration, involving triangles ABC and DEF, and then at the dual Desargues configuration, involving triangles abc and def. We have then the following things happening, both coming from Theorem 5.20:

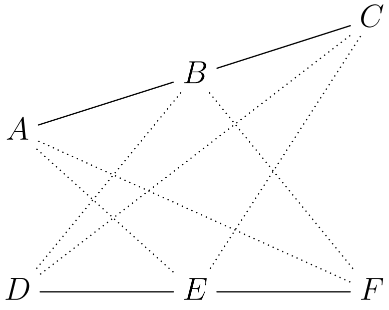
(1) The original triangles ABC, DEF are in central perspective precisely when the dual triangles abc, def are in axial perspective.

(2) The original triangles ABC, DEF are in axial perspective precisely when the dual triangles abc, def are in central perspective.

But with this, we are done, because Theorem 5.19 applied to the dual triangles abc, def gives the present result, for the original triangles ABC, DEF. □

Summarizing, done with Desargues, and we have learned many interesting things, on this occasion. Next, we have the following fact, going back in time, to Pappus:

Fact 5.22 (Pappus). *Given a configuration as follows,*



the three middle points are collinear.

As before with Desargues, or rather with the tricky implication of Desargues, proving such things will need some preparations. So, temporarily forgetting about Pappus, we have the following result, which is something having its own interest:

Theorem 5.23. *We can talk about the cross ratio of four collinear points A, B, C, D, as being the following quantity, signed according to our usual sign conventions,*

$$(A, B, C, D) = \frac{AC \cdot BD}{BC \cdot AD}$$

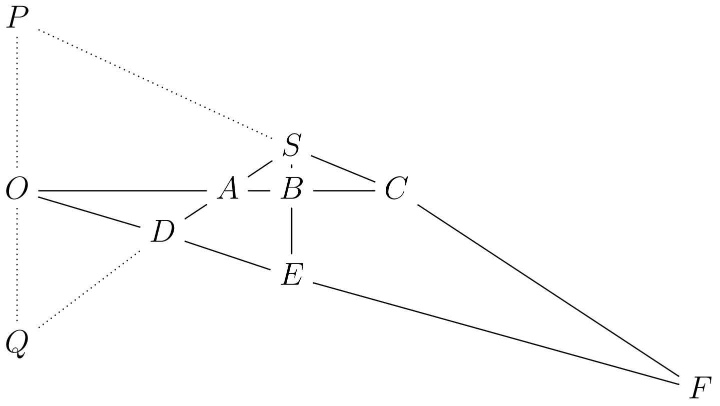
and with this notion in hand, points in central perspective have the same cross ratio:

$$(A, B, C, D) = (A', B', C', D')$$

Moreover, the converse of this fact holds too.

Proof. As before with Theorem 5.20, there is a lot of mathematics hidden here, and with the formula in the statement coming by drawing a suitable parallel line, and computing both $(A,B,C,D),(A',B',C',D')$ in terms of the new points which appear:

(1) Consider first the following picture, with the points A,B,C,D,E,F and S,O being as indicated, and with a parallel line to SE drawn on the left, as indicated:



(2) We have then the following equality, obtained by using the Thales theorem:

$$\begin{aligned}(O,B,C,A) &= \frac{OC}{BC}\cdot\frac{BA}{OA}\\ &= \frac{PO}{SB}\cdot\frac{SB}{OQ}\\ &= \frac{PO}{OQ}\end{aligned}$$

On the other hand, again by using the Thales theorem, we have as well:

$$\begin{aligned}(O,E,F,D) &= \frac{OF}{EF}\cdot\frac{ED}{OD}\\ &= \frac{PO}{SE}\cdot\frac{SE}{OQ}\\ &= \frac{PO}{OQ}\end{aligned}$$

We conclude that in the context of the above configuration, we have:

$$(O,B,C,A)=(O,E,F,D)$$

(3) But this gives the equality in statement, by suitably generalizing what we found, somewhat by "blowing up" the point O on the left into a pair of distinct points. To be more precise, let us turn now to the precise equality to be proved, namely:

$$(A,B,C,D)=(A',B',C',D')$$

Here the points A,B,C,D and A',B',C',D' are assumed to be in perspectivity, say with respect to a center of perspectivity S. Consider as well the following intersection:

$$O=ABCD\cap A'B'C'D'$$

(4) We have the following formula, coming from the definition of the cross ratio:

$$\begin{aligned}(A,B,C,D)&=\frac{AC\cdot BD}{BC\cdot AD}\\&=\frac{AC\cdot OD}{OC\cdot AD}\cdot\frac{OC\cdot BD}{BC\cdot OD}\\&=\frac{OC\cdot BD}{BC\cdot OD}\Big/\frac{OC\cdot AD}{AC\cdot OD}\\&=\frac{(O,B,C,D)}{(O,A,C,D)}\end{aligned}$$

On the other hand, a similar computation shows that we have as well:

$$(A',B',C',D')=\frac{(O',B',C',D')}{(O',A',C',D')}$$

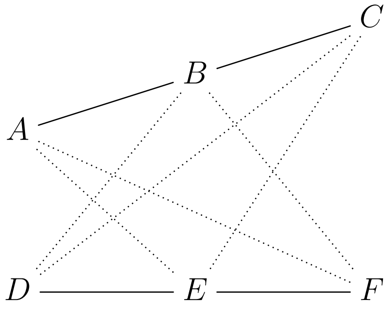
(5) But with these formulae in hand, by using (2) twice, we obtain:

$$\begin{aligned}(A,B,C,D)&=\frac{(O,B,C,D)}{(O,A,C,D)}\\&=\frac{(O',B',C',D')}{(O',A',C',D')}\\&=(A',B',C',D')\end{aligned}$$

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

Good news, we can now prove the Pappus theorem, as follows:

Theorem 5.24 (Pappus). *Given a configuration as follows,*



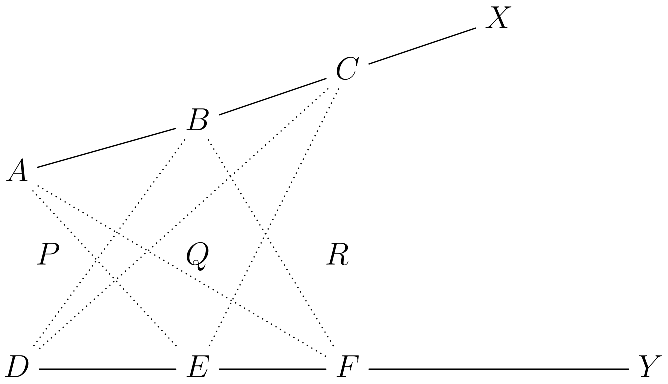
the three middle points are collinear.

Proof. We can use the cross ratio technology from Theorem 5.23, as follows:

(1) Consider indeed the Pappus configuration in the statement, then let us call P,Q,R the middle points appearing there, and construct points X,Y as follows:

$$X=AC\cap DR\quad,\quad Y=AR\cap DF$$

We obtain in this way an enlarged configuration, which looks as follows:



(2) We have then the following equalities, with the first one coming from Theorem 5.23, via the central perspective coming from the point R, and with the second one being something trivial, valid for any cross ratio, coming from definitions:

$$(A,C,B,X)=(Y,E,F,D)=(D,F,E,Y)$$

(3) But with this equality, we can conclude. Consider indeed the following point, appearing on the left in the picture, that we will need too, in what follows:

$$K=AD\cap PQ$$

Now let us see what happens to the configurations $ACBX$ and $DFEY$, when projected respectively from the points D,A, on the line PQ. Via these projections, we have:

$$ACB\to KQP\quad,\quad DFE\to KQP$$

(4) Now remember the cross ratio formula found in (2), namely:

$$(A,C,B,X)=(D,F,E,Y)$$

In view of this, and by applying again Theorem 5.23, this time in reverse form, we conclude that the images of X,Y via the above projections must coincide:

$$(DX\cap AY)\in PQ$$

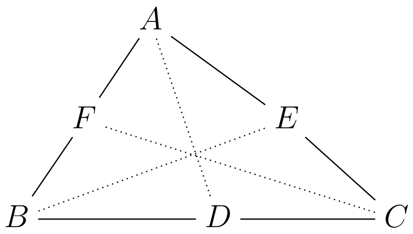
But, according to our conventions above, $DX\cap AY=R$, so we obtain, as desired:

$$R\in PQ$$

(5) Thus, result proved. As a further comment, observe that there is a relation with Desargues too. Finally, note that the Pappus configuration is self-dual. □

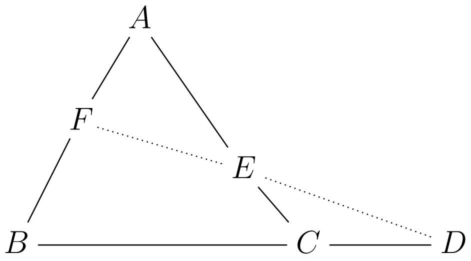
5d. Menelaus and Ceva

Let us go back now to basic triangle geometry and centers, as developed before in this chapter. In order to further build on that material, and systematically look at triangle centers, we would like to have general crossing results, of the following type:



We will discusss this slowly, with several results on this subject, and on related topics. First on our list we have the following key result, due to Menelaus:

Theorem 5.25 (Menelaus). *In a configuration of the following type, with a triangle ABC cut by a line FED,*



we have the following formula, with all segments being taken oriented:

$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=-1$$

Moreover, the converse holds, with this formula guaranteeing that F,E,D are colinear.

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

(1) Let us first try to prove the following equality, which is a bit weaker than what the theorem says, with all segments being by definition taken oriented:

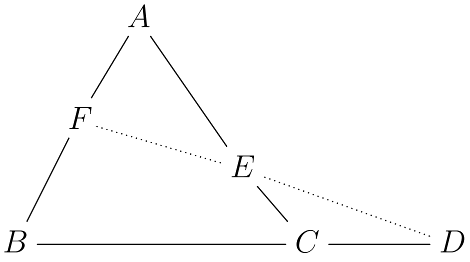
$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=1$$

But this is something clear, because by projecting the vertices A,B,C on the line DEF, into points A',B',C', we have the following computation:

$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=\frac{AA'}{BB'}\cdot\frac{BB'}{CC'}\cdot\frac{CC'}{AA'}=1$$

(2) Next, we must see what happens to the above equality, when allowing the segments to be oriented. But here, there are several cases to be considered, depending on whether

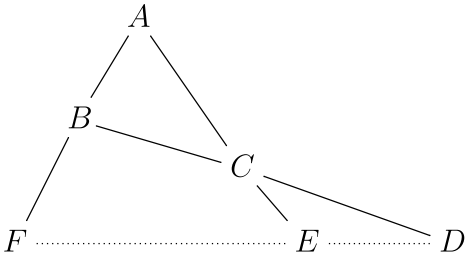
the line DEF intersects the triangle ABC, a bit as in the picture in the statement, or not. Let us first examine the crossing configuration, as in the statement, namely:



In this case, with all the segments being by definition taken oriented, we are led indeed to the formula in the statement, as follows:

$$\begin{aligned}\frac{AF}{FB}\cdot\frac{BD}{DC}\frac{CE}{EA} &= \frac{|AF|}{|FB|}\left(-\frac{|BD|}{|DC|}\right)\cdot\frac{CE}{EA}\\ &= -\frac{|AF|}{|FB|}\cdot\frac{|BD|}{|DC|}\cdot\frac{|CE|}{|EA|}\\ &= -1\end{aligned}$$

(3) Let us examine now the non-crossing configuration, which is as follows:



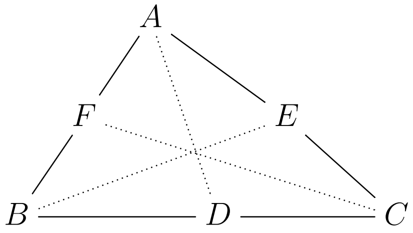
In this case, again with all the segments being by definition taken oriented, we are again led to the formula in the statement, as follows:

$$\begin{aligned}\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA} &= \left(-\frac{|AF|}{|FB|}\right)\left(-\frac{|BD|}{|DC|}\right)\left(-\frac{CE}{EA}\right)\\ &= -\frac{|AF|}{|FB|}\cdot\frac{|BD|}{|DC|}\cdot\frac{|CE|}{|EA|}\\ &= -1\end{aligned}$$

(4) Thus, we have proved the formula in the statement. As for the converse, this follows from the main result, in the obvious way, and as usual with converses of such statements, we will leave the discussion here as an instructive exercise for you. □

We can now answer our original question about crossing lines inside a triangle, drawn from the vertices, with the following remarkable result, due to Ceva:

Theorem 5.26 (Ceva). *In a configuration of the following type, with a triangle ABC containing inner lines AD, BE, CF which cross,*



we have the following formula:

$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=1$$

Moreover, the converse holds, with this formula guaranteeing that AD, BE, CF cross.

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

(1) A first way of proving this result is by using the Menelaus theorem, applied twice. Indeed, if we denote by O the point in the middle in the above picture, we have the following formula, coming from the line COF cutting the triangle ABD:

$$\frac{AF}{FB}\cdot\frac{BC}{CD}\cdot\frac{DO}{OA}=-1$$

On the other hand, again by using the Menelaus theorem, we have as well the following formula, coming this time from the line BEO cutting the triangle ADC:

$$\frac{AO}{OD}\cdot\frac{DB}{BC}\cdot\frac{CE}{EA}=-1$$

By multiplying now the above two formulae, we obtain, as desired:

$$\begin{aligned}
1 &= \frac{AF}{FB}\cdot\frac{BC}{CD}\cdot\frac{DO}{OA}\times\frac{AO}{OD}\cdot\frac{DB}{BC}\cdot\frac{CE}{EA}\\
&= \frac{AF}{FB}\cdot\frac{BC}{CD}\times\frac{DB}{BC}\cdot\frac{CE}{EA}\\
&= \frac{AF}{FB}\cdot\frac{BC}{DC}\times\frac{BD}{BC}\cdot\frac{CE}{EA}\\
&= \frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}
\end{aligned}$$

(2) An alternative proof, which is more elegant, is by using the same idea as for Menelaus, namely some fractions which cancel. Again by denoting by O the point in the middle, we have the following formulae for the quotient AF/FB, in terms of areas:

$$\frac{AF}{FB}=\frac{AFO}{FBO}=\frac{AFC}{FBC}$$

We deduce from this that we have the following extra formula for AF/FB:

$$\frac{AF}{FB}=\frac{AFC-AFO}{FBC-FBO}=\frac{AOC}{BOC}$$

Similarly, we have the following formulae for BD/DC, and for CE/EA:

$$\frac{BD}{DC}=\frac{AOB}{AOC}\quad,\quad\frac{CE}{EA}=\frac{BOC}{AOB}$$

Now by multiplying all these formulae we obtain, as desired:

$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=\frac{AOC}{BOC}\cdot\frac{AOB}{AOC}\cdot\frac{BOC}{AOB}=1$$

(3) As for the converse, this follows from the main result, in the obvious way, and as usual with such converses, we will leave the discussion here as an exercise. □

As a basic application of the Ceva theorem, we have now a new point of view on the barycenter. Indeed, the fact that the medians of a triangle cross can be seen as coming from the Ceva theorem, via the following trivial computation:

$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=1\times1\times1=1$$

As further applications of the Ceva theorem, we can try to reprove the incenter and orthocenter theorems too. However, this is something more tricky, involving a bit of trigonometry, and we will defer the discussion here, for the next chapter.

5e. Exercises

Welcome to plane geometry exercises, all beautiful, and here are some:

Exercise 5.27. *Learn more about the duality between points and lines.*

Exercise 5.28. *Fill in all the details in the proof of Desargues.*

Exercise 5.29. *Learn more about the cross ratio of collinear points.*

Exercise 5.30. *Fill in all the details in the proof of Pappus.*

Exercise 5.31. *Clarify the relation between Menelaus and Ceva.*

Exercise 5.32. *Learn about the Euler line, and the nine-point circle too.*

Exercise 5.33. *Learn about the theorems of Pascal, and Brianchon.*

Exercise 5.34. *Learn also a bit about projective geometry.*

As bonus exercise, and no surprise here, read a plane geometry book.

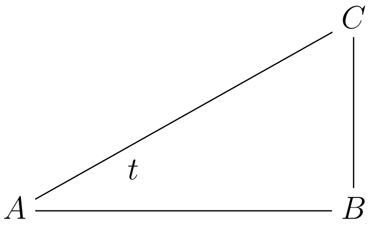
CHAPTER 6

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. *Given a right triangle ABC,*



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

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

We call the sine and cosine basic trigonometric functions.

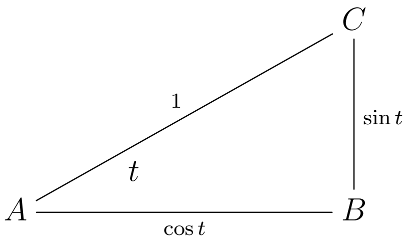
As a first observation, the sine and cosine do not depend on the choice of the given right triangle ABC having an angle t at A, and this due to the Thales theorem. In view of this, whenever possible, we will choose the right triangle ABC as to have:

$$AC=1$$

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

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

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



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

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

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

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

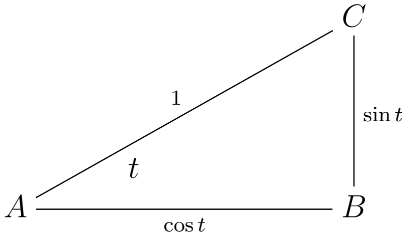
Observe that the numbers in the above two lists are the same, but written backwards in the second list. In fact, we have the following result, regarding this:

Theorem 6.2. *The sines and cosines are subject to the formulae*

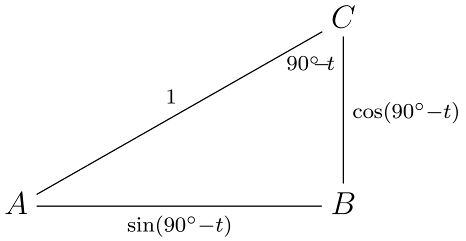
$$\sin(90^\circ - t) = \cos t \quad , \quad \cos(90^\circ - t) = \sin t$$

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

Proof. In order to understand this, the best is to choose our right triangle ABC with $AC = 1$. In this case, the picture coming from Definition 6.1 is as follows:



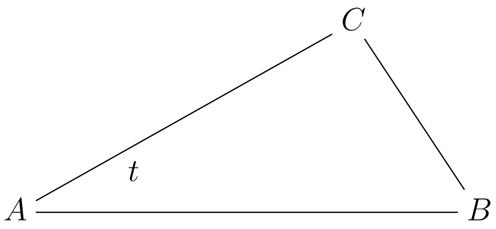
On the other hand, by focusing now at the angle at C, and perhaps twisting a bit our minds too, we have as well the following picture, for the same triangle:



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

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

Cat 6.3. *The area of an arbitrary triangle, having an angle t at A,*

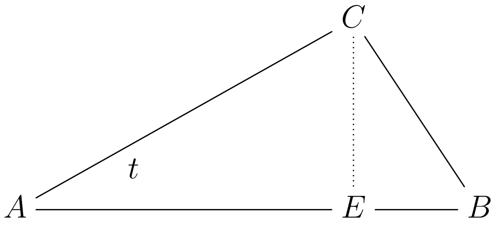


is given by the following formula, making appear the sine:

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

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

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



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

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

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

Moving forward now, still in relation with Cat 6.3, we have the following question:

Question 6.4. *What happens to the cat formula,*

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

when the angle at A is obtuse, $t>90°$?

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

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

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

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

Thus, Question 6.4 answered, with our conclusions being as follows:

Theorem 6.5. *We can talk about the sine of any angle $t \in [0^\circ, 180^\circ]$, according to*

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

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

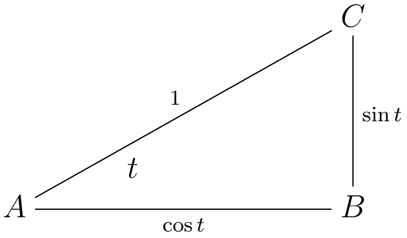
$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

holds for any triangle, without any assumption on it.

Proof. This follows indeed from the above discussion. □

Moving ahead now, defining sines as in Definition 6.1 for $t \in [0^\circ, 90^\circ]$, and as above for $t \in [90^\circ, 180^\circ]$ certainly does the job, as explained above, but is not very elegant. So, let us try to improve this. We have here the following obvious speculation:

Speculation 6.6. *The sine of any angle $t \in [0^\circ, 180^\circ]$ can be defined geometrically, according to the usual picture*



with the convention that for $t > 90^\circ$, the triangle is drawn at the left of A.

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

$$AB \to BA$$

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

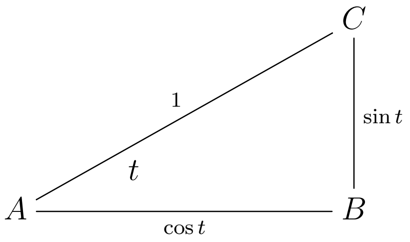
$$AB \to -AB$$

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

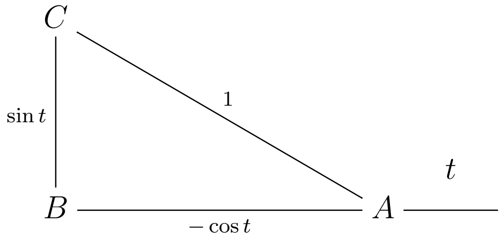
$$\cos t = -\cos(180^\circ - t)$$

Very good all this, so let us update now Theorem 6.5, and by incorporating as well Speculation 6.6, in the form of a grand result, in the following way:

Theorem 6.7 (update). *We can talk about the sine and cosine of any angle $t \in [0^\circ, 180^\circ]$, according to the following picture,*



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



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

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

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

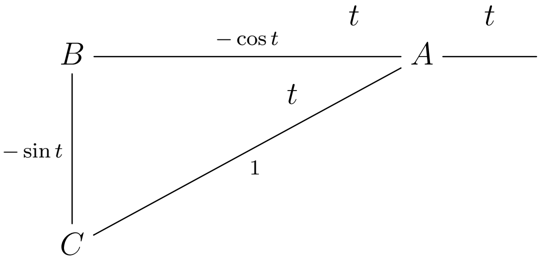
$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

holds for any triangle, without any assumption on it.

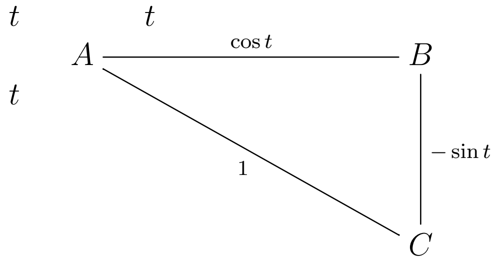
Proof. This follows indeed by putting together all the above. □

Which sounds quite good, and normally end of the story, but let us be crazy now, and try to talk as well about the sine or cosine of angles $t < 0^\circ$, or $t > 180^\circ$.

Indeed, we know the recipe, namely suitably drawing our right triangle, with attention to positive and negatives. Thus, for $t \in [180^\circ, 270^\circ]$, our picture should be as follows:



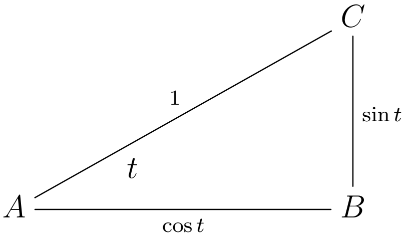
As for the next case, $t \in [270^\circ, 360^\circ]$, here our picture should be as follows:



But with this, we are done, because adding or substracting 360° to our angles won't change the corresponding right triangle, and so won't change the sine and cosine.

Hope you're still with me, after all these speculations. Good work that we did, and time now to further improve Theorem 6.7, into something really final, as follows:

THEOREM 6.8 (final update). *We can talk about the sine and cosine of any angle* $t \in \mathbb{R}$*, according to the following picture,*



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

PROOF. This follows by putting together all the above:

(1) The first assertion follows indeed from the above discussion.

(2) As for the basic trigonometry formulae mentioned at the end of the statement, these are as follows, and in the hope of course that I forgot none:

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

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

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

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

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

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

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

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

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

In order to study now sin and cos, let us first update the numerics that we have, for very simple angles in $[0^\circ, 90^\circ]$, to more angles, in $[0^\circ, 360^\circ]$. We have here:

Theorem 6.9. *The sines of the basic angles are as follows,*

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

$$\sin 120^\circ = \frac{\sqrt{3}}{2} \quad , \quad \sin 135^\circ = \frac{1}{\sqrt{2}} \quad , \quad \sin 150^\circ = \frac{1}{2} \quad , \quad \sin 180^\circ = 0$$

$$\sin 210^\circ = -\frac{1}{2} \quad , \quad \sin 225^\circ = -\frac{1}{\sqrt{2}} \quad , \quad \sin 240^\circ = -\frac{\sqrt{3}}{2} \quad , \quad \sin 270^\circ = -1$$

$$\sin 300^\circ = -\frac{\sqrt{3}}{2} \quad , \quad \sin 315^\circ = -\frac{1}{\sqrt{2}} \quad , \quad \sin 330^\circ = -\frac{1}{2} \quad , \quad \sin 360^\circ = 0$$

and the cosines and tangents are given by similar formulae.

Proof. This is indeed self-explanatory, with input coming from the above. □

6b. Trigonometry

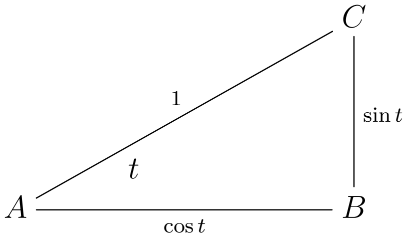
The problem is now, how to get beyond the above formulae? Not an easy question, but do not worry, we will be back to this, in due time. For the moment, as a complement to the above, let us record the following key formula, coming from Pythagoras:

Theorem 6.10. *The sines and cosines are subject to the formula*

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

coming from Pythagoras' theorem.

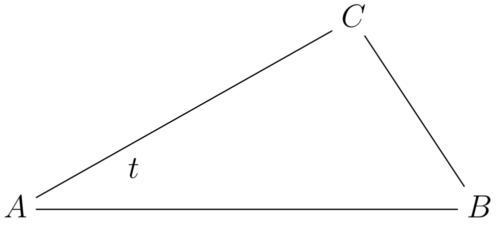
Proof. Consider indeed the defining picture for sin and cos, namely:



By applying now Pythagoras, we are led to the formula in the statement. □

In relation with this, with our knowledge of the sine and cosine, we can now formulate a technical generalization of the Pythagoras theorem, in the following way:

Theorem 6.11. *Given an arbitrary triangle, as follows,*

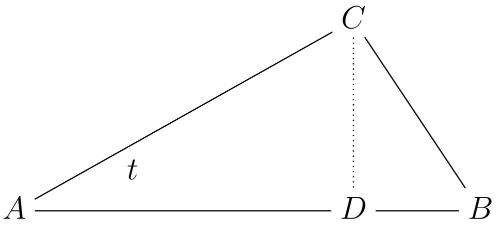


the length of the side which is away from the vertex A is given by the formula

$$BC^2 = AB^2 + AC^2 - 2AB \cdot AC \cdot \cos t$$

called law of cosines, and with this generalizing Pythagoras.

Proof. Let us draw indeed an altitude of our triangle, as follows:



We have then the following computation, coming from Pythagoras, applied twice:

$$\begin{aligned} BC^2 &= CD^2 + BD^2 \\ &= CD^2 + (AB - AD)^2 \\ &= CD^2 + AB^2 + AD^2 - 2AB \cdot AD \\ &= AB^2 + AC^2 - 2AB \cdot AD \\ &= AB^2 + AC^2 - 2AB \cdot AC \cdot \cos t \end{aligned}$$

Finally, the last assertion is clear, because with $\cos t = 0$ we obtain Pythagoras. □

The above result looks quite interesting, for engineering purposes, and we have:

Conclusion 6.12. *The law of cosines found above can be effectively used for making money, by computing distances BC over wild land, for various interested customers.*

Which might sound quite interesting, for us humans, but my cat, who is not into making money, seems unfazed. In fact, here is what he has to say, about this:

Cat 6.13. *That law of cosines is too complicated, no match for my law of sines:*

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

I would suggest you humans to look into the quantity

$$< AB, AC >= AB \cdot AC \cdot \cos t$$

in order to understand what the cosines are good for. And change your diet, too.

Quite interesting all this, but in practice, this $< AB, AC >$ quantity does not seem to be something very intuitive, at least to my human brain. We will leave this for later.

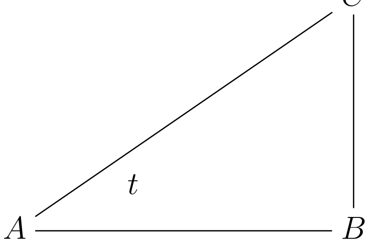
Back now to the basics, it is possible to say many more things about angles and $\sin t$, $\cos t$, and also talk about some supplementary quantities, such as the tangent:

Definition 6.14. *We can talk about the tangent of angles $t \in \mathbb{R}$, as being given by*

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

with $\sin t, \cos t$ being defined as before.

In more geometric terms, consider an arbitrary right triangle, as follows:



We have then the following computation, for the tangent of t:

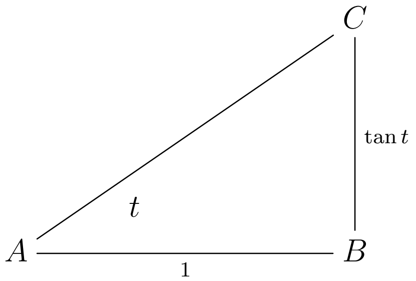
$$\tan t=\frac{\sin t}{\cos t}=\frac{BC}{AC}\Big/\frac{AB}{AC}=\frac{BC}{AB}$$

Thus, the tangent defined above complements the sine and cosine, because we have:

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

A similar interpretation works for obtuse right triangles, and even for right triangles with an arbitrary angle $t\in\mathbb{R}$, and we can formulate, in the spirit of Theorem 6.8:

Theorem 6.15. *We can talk, geometrically, about the tangent of any angle $t\in\mathbb{R}$, according to the following picture,*



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

Proof. Here the first assertion follows by reasoning as in the proof of Theorem 6.8, or simply follows from Theorem 6.8 itself. As for the second assertion, the basic formulae for the tangent, all coming from what we know, are as follows:

$$\tan(-t)=-\tan t$$

$$\tan(90°-t)=\frac{1}{\tan t}\quad,\quad\cos(90°+t)=-\frac{1}{\tan t}$$

$$\tan(180°-t)=-\tan t\quad,\quad\tan(180°+t)=\tan t$$

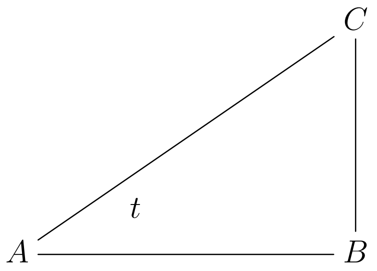
Let us record as well the formulae for the basic angles. These are as follows:

$$\tan 0°=0\quad,\quad\tan 30°=\frac{1}{\sqrt{3}}\quad,\quad\sin 45°=1\quad,\quad\sin 60°=\sqrt{3}$$

$$\tan 120°=-\sqrt{3}\quad,\quad\tan 135°=-1\quad,\quad\tan 150°=-\frac{1}{\sqrt{3}}\quad,\quad\tan 180°=0$$

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

Very nice all this, but are we really done with generalities and definitions? Not yet, because, let us go back to our basic right triangle, with an angle t, as follows:



We know from the above that we have the following formulae:

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

However, there are still 3 fractions left, in need of a name, so let us formulate:

Definition 6.16. *We can talk about the secant, cosecant and cotangent, as being*

$$\sec t=\frac{AC}{BC}\quad,\quad\csc t=\frac{AC}{AB}\quad,\quad\cot t=\frac{BC}{AB}$$

in the context of a right triangle, as above, or equivalently, as being

$$\sec t=\frac{1}{\sin t}\quad,\quad\csc t=\frac{1}{\cos t}\quad,\quad\cot t=\frac{1}{\tan t}$$

in terms of the standard trigonometric functions $\sin,\cos,\tan$.

As an application now, remember the discussion following the Ceva theorem, from the previous chapter? We had some unfinished business there, in what regards the applications, and we promised to get back to this, once we know some trigonometry. So, time to do this, and as good news, we get into something involving secants and cotangents:

Theorem 6.17. *The barycenter, incenter and orthocenter theorems can be all deduced from the Ceva theorem, with the computations being as follows,*

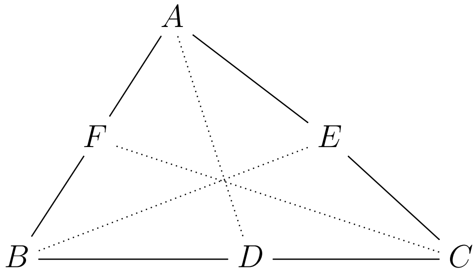
$$1\times 1\times 1=1$$

$$\frac{\sec A}{\sec B}\cdot\frac{\sec B}{\sec C}\cdot\frac{\sec C}{\sec A}=1$$

$$\frac{\cot A}{\cot B}\cdot\frac{\cot B}{\cot C}\cdot\frac{\cot C}{\cot A}=1$$

with A,B,C *being the angles of our triangle.*

Proof. Let us first recall from chapter 5 that the Ceva theorem concerns a configuration as follows, with a triangle ABC containing inner lines AD, BE, CF:



The theorem states that AD, BE, CF cross precisely when the following happens:

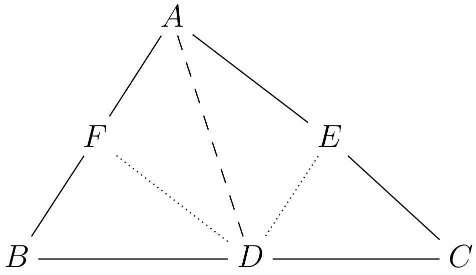
$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=1$$

Regarding now the barycenter, incenter and orthocenter, the situation is as follows:

(1) In what regards the barycenter, the computation is trivial, as follows:

$$1\times 1\times 1=1$$

(2) In order to deal now with the incenter, consider indeed a triangle, with an angle bisector drawn, and with two perpendiculars drawn as well, as indicated:



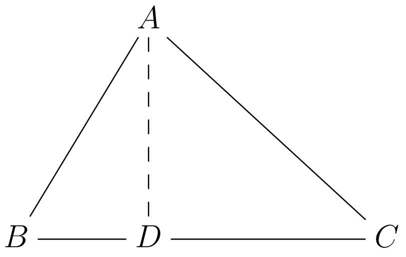
We have then the following computation, using $FD=DE$:

$$\frac{BD}{DC}=\frac{FD\sec B}{DE\sec C}=\frac{\sec B}{\sec C}$$

We conclude that Ceva gives indeed the incenter, with the computation being:

$$\frac{\sec A}{\sec B}\cdot\frac{\sec B}{\sec C}\cdot\frac{\sec C}{\sec A}=1$$

(3) Finally, in order to deal now with the orthocenter, a bit in a similar way, consider indeed a triangle, with an altitude drawn, as follows:



We have then the following computation, coming from definitions:

$$\frac{BD}{DC}=\frac{BD}{AD}\Big/\frac{DC}{AD}=\frac{\cot B}{\cot C}$$

Thus Ceva gives as well the orthocenter, with the computation being as follows:

$$\frac{\cot A}{\cot B}\cdot\frac{\cot B}{\cot C}\cdot\frac{\cot C}{\cot A}=1$$

And with this being something nice, remember the mess with the orthocenter when first proving the theorem, with that trick involved. Gone all that. □

6c. Sums, duplication

Getting back now to the basics, namely sine, cosine and tangent, how these can be computed, and what can be done with them, we have the following key result:

Theorem 6.18. *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 the tangent too, namely

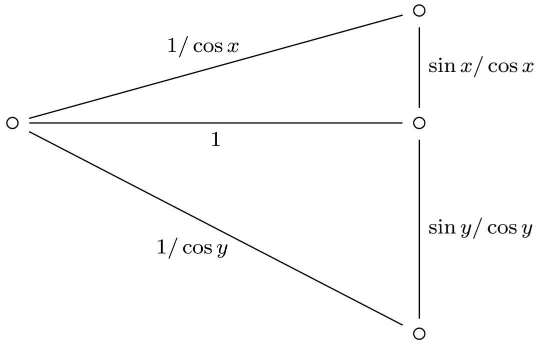
$$\tan(x+y)=\frac{\tan x+\tan y}{1-\tan x\tan y}$$

provided of course that the denominator is nonzero.

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, namely:

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

(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:

$$\begin{aligned}\tan(x+y)&=\frac{\sin x\cos y+\cos x\sin y}{\cos x\cos y-\sin x\sin y}\\&=\frac{\sin x\cos y/\cos x\cos y+\cos x\sin y/\cos x\cos y}{1-\sin x\sin y/\cos x\cos y}\\&=\frac{\tan x+\tan y}{1-\tan x\tan y}\end{aligned}$$

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

The above theorem is something very useful, in practice, so let us record as well what happens when replacing sums by substractions. The formulae here are as follows:

Theorem 6.19. *The sines and cosines of differences 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 the tangent too, namely

$$\tan(x-y)=\frac{\tan x-\tan y}{1+\tan x\tan y}$$

provided of course that the denominator is nonzero.

Proof. These are all consequences of what we have in Theorem 6.18, as follows:

(1) Regarding the sine, we have here the following computation:

$$\begin{aligned}\sin(x-y)&=&\sin x\cos(-y)+\cos x\sin(-y)\\&=&\sin x\cos y-\cos x\sin y\end{aligned}$$

(2) Regarding the cosine, the computation here is similar, as follows:

$$\begin{aligned}\cos(x-y)&=&\cos x\cos(-y)-\sin x\sin(-y)\\&=&\cos x\cos y+\sin x\sin y\end{aligned}$$

(3) Finally, in what regards the tangent, I would not mess with it, and say instead:

$$\begin{aligned}\tan(x-y)&=&\frac{\sin x\cos y-\cos x\sin y}{\cos x\cos y+\sin x\sin y}\\&=&\frac{\sin x\cos y/\cos x\cos y-\cos x\sin y/\cos x\cos y}{1+\sin x\sin y/\cos x\cos y}\\&=&\frac{\tan x-\tan y}{1+\tan x\tan y}\end{aligned}$$

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

As an application, we can now deal with all multiples of 15°, as follows:

Theorem 6.20. *The sine, cosine and tangent of multiples of* 15° *are given by*

$$\sin 15^\circ=\frac{\sqrt{3}-1}{2\sqrt{2}}\ ,\ \sin 30^\circ=\frac{1}{2}\ ,\ \sin 45^\circ=\frac{1}{\sqrt{2}}\ ,\ \sin 60^\circ=\frac{\sqrt{3}}{2}\ ,\ \sin 75^\circ=\frac{\sqrt{3}+1}{2\sqrt{2}}$$

$$\cos 15^\circ=\frac{\sqrt{3}+1}{2\sqrt{2}}\ ,\ \cos 30^\circ=\frac{\sqrt{3}}{2}\ ,\ \cos 45^\circ=\frac{1}{\sqrt{2}}\ ,\ \cos 60^\circ=\frac{1}{2}\ ,\ \cos 75^\circ=\frac{\sqrt{3}-1}{2\sqrt{2}}$$

$$\tan 15^\circ=\frac{\sqrt{3}-1}{\sqrt{3}+1}\ ,\ \tan 30^\circ=\frac{1}{\sqrt{3}}\ ,\ \tan 45^\circ=1\ ,\ \tan 60^\circ=\sqrt{3}\ ,\ \tan 75^\circ=\frac{\sqrt{3}+1}{\sqrt{3}-1}$$

plus $\sin 0^\circ=0$, *and various periodicity formulae.*

PROOF. For the quantity $\sin 15^\circ = \cos 75^\circ$, we have the following computation:

$$\begin{aligned}\sin 15^\circ &= \sin(45^\circ - 30^\circ)\\ &= \sin 45^\circ \cos 30^\circ - \cos 45^\circ \sin 30^\circ\\ &= \frac{1}{\sqrt{2}}\cdot\frac{\sqrt{3}}{2} - \frac{1}{\sqrt{2}}\cdot\frac{1}{2}\\ &= \frac{\sqrt{3}-1}{2\sqrt{2}}\end{aligned}$$

Also, for the quantity $\cos 15^\circ = \sin 75^\circ$, we have the following computation:

$$\begin{aligned}\cos 15^\circ &= \cos(45^\circ - 30^\circ)\\ &= \cos 45^\circ \cos 30^\circ + \sin 45^\circ \sin 30^\circ\\ &= \frac{1}{\sqrt{2}}\cdot\frac{\sqrt{3}}{2} + \frac{1}{\sqrt{2}}\cdot\frac{1}{2}\\ &= \frac{\sqrt{3}+1}{2\sqrt{2}}\end{aligned}$$

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

Next, with $x = y$ in Theorem 6.18 we obtain some interesting formulae, as follows:

THEOREM 6.21. *The sines of the doubles of angles are given by*

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

and the corresponding cosines are given by the following equivalent formulae,

$$\begin{aligned}\cos(2t) &= \cos^2 t - \sin^2 t\\ &= 2\cos^2 t - 1\\ &= 1 - 2\sin^2 t\end{aligned}$$

with all these three formulae being useful, in practice.

PROOF. By taking $x = y = t$ in the formulae from Theorem 6.18, we obtain:

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

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

As for the extra formulae for $\cos(2t)$, these follow by using $\cos^2 + \sin^2 = 1$. □

Let us record as well the formula for the tangents, which is as follows:

THEOREM 6.22. *The tangents of the doubles of angles are given by*

$$\tan(2t) = \frac{2\tan t}{1 - \tan^2 t}$$

provided as usual that the denominator is nonzero.

Proof. This follows indeed by taking $x = y = t$ in the formula for tangents from Theorem 6.18. Equivalently, you can check, as an easy, instructive exercise, that this is indeed what we get, by dividing the sine and cosine computed in Theorem 6.21. □

As a first application of the above duplication results, we have:

Theorem 6.23. *The sine, cosine and tangent of* 22.5° *are given by*

$$\sin 22.5^\circ = \frac{\sqrt{2-\sqrt{2}}}{2} \quad , \quad \cos 22.5^\circ = \frac{\sqrt{2+\sqrt{2}}}{2} \quad , \quad \tan 22.5^\circ = \sqrt{2}-1$$

and for the odd multiples of 22.5°*, we have similar formulae.*

Proof. For the cosine we can use $\cos(2t) = 2\cos^2 t - 1$, and we obtain:

$$\begin{aligned}\cos 22.5^\circ &= \sqrt{\frac{1+\frac{1}{\sqrt{2}}}{2}}\\ &= \sqrt{\frac{\sqrt{2}+1}{2\sqrt{2}}}\\ &= \frac{\sqrt{2+\sqrt{2}}}{2}\end{aligned}$$

For the sine we can use Pythagoras, $\sin^2 + \cos^2 = 1$, and we obtain:

$$\begin{aligned}\sin 22.5^\circ &= \sqrt{1-\cos^2 22.5^\circ}\\ &= \sqrt{1-\frac{2+\sqrt{2}}{4}}\\ &= \frac{\sqrt{2-\sqrt{2}}}{2}\end{aligned}$$

Finally, by taking the quotient we obtain a formula for the tangent, as follows:

$$\begin{aligned}\tan 22.5^\circ &= \sqrt{\frac{2-\sqrt{2}}{2+\sqrt{2}}}\\ &= \sqrt{\frac{(2-\sqrt{2})^2}{(2+\sqrt{2})(2-\sqrt{2})}}\\ &= \frac{2-\sqrt{2}}{\sqrt{2}}\\ &= \sqrt{2}-1\end{aligned}$$

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

Along the same lines, at a more advanced level, we have as well:

THEOREM 6.24. *The sine, cosine and tangent of* 7.5° *are given by*

$$\sin 7.5^\circ=\sqrt{\frac{4-\sqrt{2}-\sqrt{6}}{8}}\ ,\ \cos 7.5^\circ=\sqrt{\frac{4+\sqrt{2}+\sqrt{6}}{8}}\ ,\ \tan 7.5^\circ=\sqrt{\frac{4-\sqrt{2}-\sqrt{6}}{4+\sqrt{2}+\sqrt{6}}}$$

and for the odd multiples of 7.5°, *we have similar formulae.*

PROOF. For the cosine we can use $\cos(2t)=2\cos^2 t-1$, and we obtain:

$$\begin{aligned}\cos 7.5^\circ &= \sqrt{\frac{1+\frac{1+\sqrt{3}}{2\sqrt{2}}}{2}}\\ &= \sqrt{\frac{2\sqrt{2}+1+\sqrt{3}}{4\sqrt{2}}}\\ &= \sqrt{\frac{4+\sqrt{2}+\sqrt{6}}{8}}\end{aligned}$$

For the sine we can use Pythagoras, $\sin^2+\cos^2=1$, and we obtain:

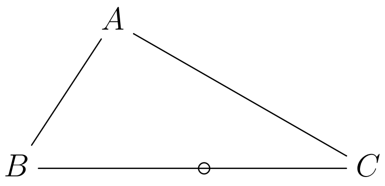
$$\begin{aligned}\sin 7.5^\circ &= \sqrt{1-\cos^2 7.5^\circ}\\ &= \sqrt{1-\frac{4+\sqrt{2}+\sqrt{6}}{8}}\\ &= \sqrt{\frac{4-\sqrt{2}-\sqrt{6}}{8}}\end{aligned}$$

Finally, by taking the quotient we obtain the formula for the tangent. □

6d. Circles, pi

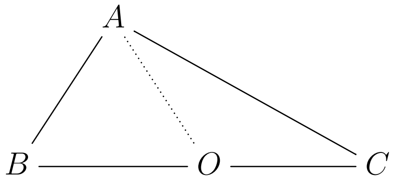
Let us get now into a more advanced study of the angles, by using circles, which are more advanced technology. We have here the following key result, to start with:

THEOREM 6.25. *Any triangle lying on a circle, with two vertices on a diameter,*



is a right triangle.

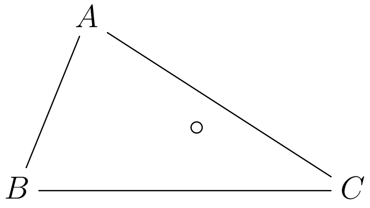
Proof. This is clear, because we have on the picture of our triangle, with the center of the circle marked, two isosceles triangles appearing, as follows:



Thus, at the level of the corresponding angles, the 180° equation for our triangle reads $b+(b+c)+c=180^\circ$, so the angle at A is indeed $b+c=90^\circ$, as claimed. □

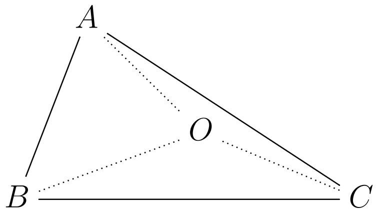
More generally now, we have the following result:

Theorem 6.26. *Given a triangle ABC lying on a circle,*



the angle at A does not depend on the exact position of A on the circle. In fact this angle equals half the angle BOC, with O being the middle of the circle.

Proof. This follows a bit as before, by drawing the center of the circle and the corresponding 3 rays, which make appear 3 isosceles triangles, as follows:



Thus, the angles of our triangle ABC are as follows, with p,q,r being the smaller angles which appear on the above picture, from left to right:

$$a=p+q \quad , \quad b=p+r \quad , \quad c=r+q$$

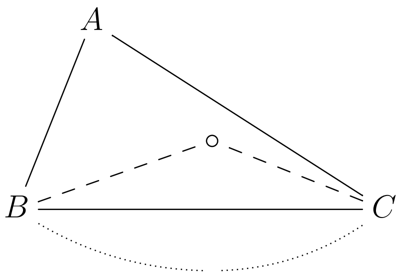
Now let us look at the angle BOC. This is given by the following formula:

$$\begin{aligned} BOC &= 180^\circ-2r\\ &= (a+b+c)-2r\\ &= (2p+2q+2r)-2r\\ &= 2p+2q \end{aligned}$$

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

The above results are quite interesting. Indeed, based on them, we can formulate:

Conclusion 6.27. *We can measure angles A by putting them on a circle, as follows, and assigning to A the length of the arc BC:*



Alternatively, we can put our angle in the middle of the circle, and assign to it the corresponding arc length, with this yielding half of the quantity defined before.

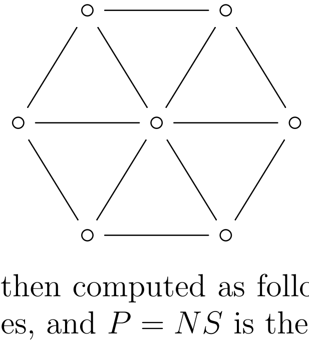
Which sounds very nice and useful, but in order to further build on this, we have two pressing issues to be clarified, first being the practical computation of arc lengths, and second being to decide which of the above two conventions is the best.

In order to discuss these questions, we first need to talk about π. And here, we have the following result, which can be regarded as being something axiomatic:

Theorem 6.28. *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:

$$A = N \times \frac{HS}{2} = \frac{HP}{2} \simeq \frac{1 \times L}{2}$$

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. More can be said by using results like Theorems 6.23 and 6.24, by replacing the hexagon used in the above with higher polygons, which gives:

$$\pi = 3.14159\ldots$$

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

Definition 6.29. *The value of an angle is obtained by putting that angle on the center of a circle of radius* 1*, and measuring the corresponding arc length.*

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

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

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

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

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

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

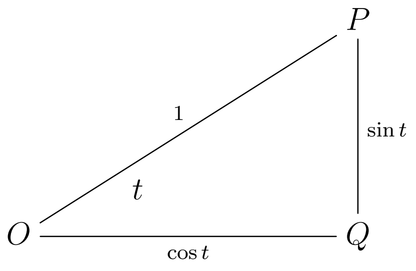
$$210^\circ = \frac{7\pi}{6} \quad , \quad 240^\circ = \frac{4\pi}{3} \quad , \quad 300^\circ = \frac{5\pi}{3} \quad , \quad 330^\circ = \frac{11\pi}{6}$$

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

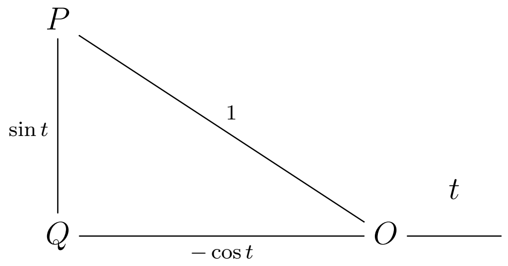
Theorem 6.30. *The sine and cosine of an angle are obtained by putting the angle on the unit circle, as above, then projecting on the vertical and the horizontal, and then measuring the oriented segments that we get, on the vertical and horizontal.*

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

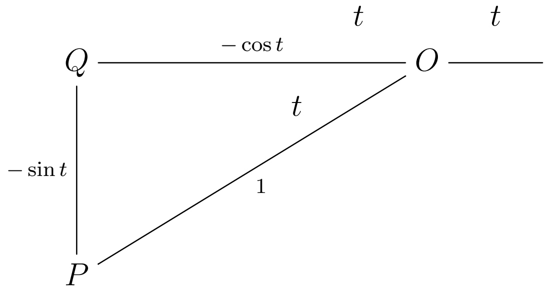
(1) In the simplest case, namely $t \in [0, \pi/2]$, the sine and cosine are indeed computed according to the following picture, which is the one in the statement:



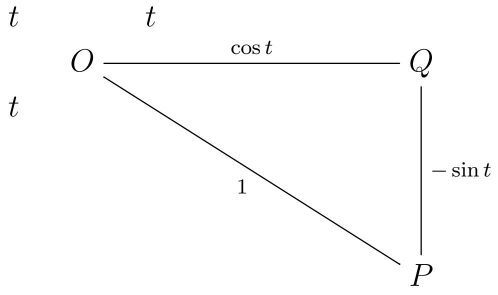
(2) In the case of obtuse angles, $t \in [\pi/2, \pi]$, the picture becomes as follows:



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



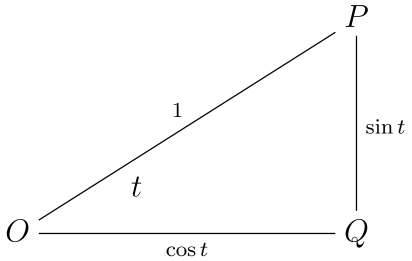
(4) As for the last case, namely $t \in [3\pi/2, 2\pi]$, here our picture is as follows:



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

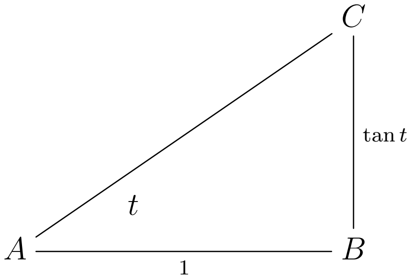
As an interesting fact, we can complement Theorem 6.30 with a statement regarding the tangent, the other basic trigonometric function, as follows:

Theorem 6.31. *The tangent of an angle can be obtained by putting the angle on the unit circle, as before,*



and then measuring the oriented segment that we get, on the vertical, outside the circle, on the vertical tangent at right.

Proof. This is, again, something quite self-explanatory, with the picture here being something that we already know from before, namely:



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

Let us get now into an interesting question, namely estimating $\sin, \cos, \tan$ and the other trigonometric functions. For this purpose, let us first recall the basic formulae for the sums of angles, that we established before, which were as follows:

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

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

$$\tan(x+y) = \frac{\tan x + \tan y}{1 - \tan x \tan y}$$

Obviously, these formulae allow us to transport our approximation questions around $t = 0$, so with this understood, let us get now to what happens around 0.

And here, to start with, we have the following basic estimates:

Theorem 6.32. *We have the following estimates,*

$$\sin t \leq t \leq \tan t$$

valid for small sngles.

PROOF. The above two estimates are indeed both clear from our circle picture for the angles, and trigonometric functions. One interesting question concerns the exact range of the above estimates, and we will leave the discussion here as an interesting exercise. □

In fact, by using our circle technology, we are led to the following result:

THEOREM 6.33. *The following happen, for small angles:*

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

PROOF. This can be indeed established as follows:

(1) This is clear indeed on the circle.

(2) This comes from (1), and from Pythagoras. Indeed, knowing $\sin t \simeq t$, when looking for a quantity $\cos t$ making the Pythagoras formula $\cos^2 t + \sin^2 t = 1$ hold, we are led, via some quick thinking, to the formula $\cos t \simeq 1 - t^2/2$, as stated. Here is the verification, and with the result itself coming via some reverse engineering, from this:

$$\begin{aligned}\left(1-\frac{t^2}{2}\right)^2+t^2 &= \left(1-t^2+\frac{t^4}{4}\right)+t^2\\ &\simeq 1-t^2+t^2\\ &= 1\end{aligned}$$

(3) This is again clear on the circle. □

6e. Exercises

Welcome to trigonometry exercises, such a pleasure, and here are some:

EXERCISE 6.34. *Further meditate on the need for the sine, and the cosine.*

EXERCISE 6.35. *Learn more on the various secondary trigonometric functions.*

EXERCISE 6.36. *Compute the trigonometric functions of all multiples of* $7.5°$.

EXERCISE 6.37. *Compute the trigonometric functions of all multiples of* $3.75°$.

EXERCISE 6.38. *Compute* $\tan(kt)$ *as function of* $\tan(t)$, *for* $k \in \mathbb{N}$ *small.*

EXERCISE 6.39. *Learn about the Chebycheff polynomials, of first and second kind.*

EXERCISE 6.40. *Further build on the above, with some better estimates for* π.

As bonus exercise, find and read an old-style, dusty trigonometry book.

CHAPTER 7

Coordinates

7a. Real plane

Welcome to plane geometry, take two. What we have been doing so far was certainly great work, certainly needed for understanding what is going on, no question about this, but that material was a bit old, essentially going back to the old Greeks. Time now for some true modern things, from a few hundred centuries ago, no longer than that.

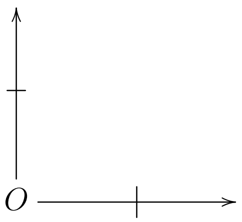
The general principle of modern geometry, coming from the work of Descartes and others, is something very simple and bright, as follows:

Principle 7.1. *Everything that we know about plane geometry, including angles and trigonometry, can be better understood, and substantially generalized, by using vectors,*

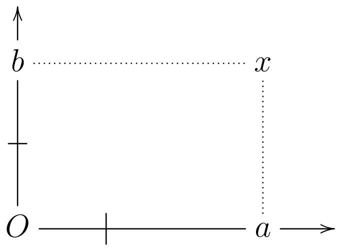
$$x=\begin{pmatrix}a\\ b\end{pmatrix}$$

with $a,b\in\mathbb R$, which such a vector describing the position of a point x in the plane with respect to a given system of coordinates, with $a,b\in\mathbb R$ being the coordinates of x.

To be more precise here, let us fix a system of coordinates in the plane, with this meaning fixing a point O, called the origin, and then a pair of orthogonal lines passing through O. We will assume in addition that these two orthogonal lines are oriented, by marking arrows on them, and also we will specify the unit length on each of them, with the complete picture of our coordinate system being as follows:



Now given a point x in the plane, we can project it onto the coordinate axes, and call the numbers $a,b\in\mathbb R$ describing the positions of these projections, with respect to the origin O, the coordinates of x, with the picture for this being as follows:



Observe now that, conversely, given two real numbers $a, b \in \mathbb{R}$, these will uniquely determine a certain point x is the plane, constructed according to the above picture. That is, we draw a on the horizontal axis, b on the vertical axis, than we draw perpendiculars as above, and x will be then the intersection of these two perpendiculars.

Summarizing, a point x in the plane and a pair of real numbers $a, b \in \mathbb{R}$ is the same thing. In view of this, we agree to use the following notation, for this correspondence, and also make the convention that, with x viewed in this way, it will be called vector:

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

In practice now, with all this digested, it is actually convenient to forget about the plane, coordinates and projections, and summarize this discussion as follows:

Definition 7.2. *A vector is a pair of real numbers, written vertically:*

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

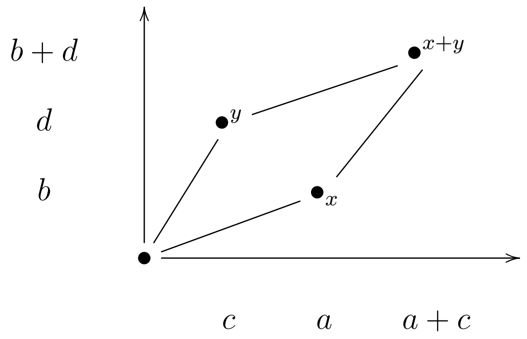
We identify the vectors with the points in the plane, in the obvious way.

Many interesting things can be done with vectors, and of particular interest is the summing operation for such vectors, given by the following formula:

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

Geometrically, and coming as a simple application of the Thales theorem, the idea with this operation is that the vectors add by forming a parallelogram, as shown by:

Theorem 7.3. *The vector addition can be understood geometrically,*

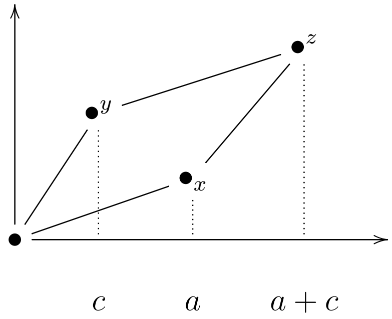


with $x+y$ completing the parallelogram based at O,x,y.

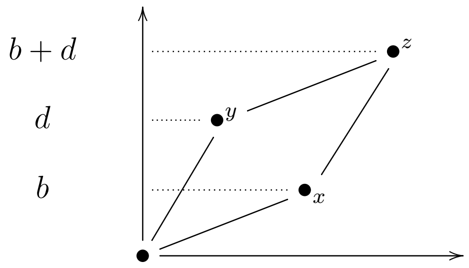
Proof. This is something quite self-explanatory. Consider indeed a parallelogram in the plane, with three of its vertices being as follows:

$$O=\begin{pmatrix}0\\0\end{pmatrix}\quad,\quad x=\begin{pmatrix}a\\b\end{pmatrix}\quad,\quad y=\begin{pmatrix}c\\d\end{pmatrix}$$

Now let us draw verticals from x,y, and from the fourth vertex z too. From Thales we obtain that the first coordinate of z is $a+c$, according to the following picture:



Similarly, if we draw horizontals from x,y, and from z too, from Thales we obtain that the second coordinate of z is $b+d$, according to the following picture:



Thus we are led to the picture in the statement, and with the final conclusion being that the coordinates of the fourth vertex z can be computed according to:

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

But this is exactly the summing formula for the vectors, as desired. □

In practice, the summing operation is usefully complemented by the multiplication by scalars operation, which is given by the following very intuitive formula:

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

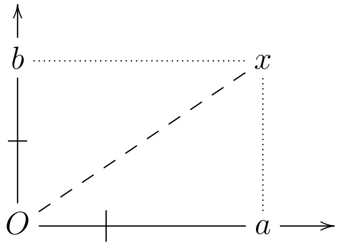
Finally, of particular interest too, in relation with the computation of the lengths, is the following result, allowing us to compute the length of any vector:

Theorem 7.4. *The length of a vector is given by the following formula:*

$$x=\begin{pmatrix}a\\ b\end{pmatrix}\implies ||x||=\sqrt{a^2+b^2}$$

Also, the vector lengths satisfy $||\lambda x||=|\lambda|\cdot||x||$, *and* $||x+y||\leq||x|+||y||$.

Proof. In what regards the first assertion, this follows as a basic application of the theorem of Pythagoras, according to the following picture:



Regarding now the second assertion, with $x=\binom{a}{b}$, we have indeed:

$$\begin{aligned}||\lambda x|| &= \sqrt{(\lambda a)^2+(\lambda b)^2}\\ &= |\lambda|\sqrt{a^2+b^2}\\ &= |\lambda|\cdot||x||\end{aligned}$$

Finally, in what regards the last assertion, this is something clear geometrically, expressing the fact that the side of a triangle, and more specifically of the triangle having vertices $O,x,x+y$, is smaller than the sum of the other two sides. However, we can prove this algebrically as well. Indeed, with $x=\binom{a}{b}$ and $y=\binom{c}{d}$, we must prove:

$$\sqrt{(a+c)^2+(b+d)^2}\leq\sqrt{a^2+b^2}+\sqrt{c^2+d^2}$$

And I will leave finishing this to you, by raising to the square and so on. □

And with this, good news, we have all the needed vector calculus tools, in our bag, and we can now start exploring what we can do, with this new formalism.

As a first good surprise, in what regards the axiomatics from chapter 5, that is literally nuked by coordinates. We first have, indeed, regarding the first axiom of geometry, that we started chapter 5 with, the following theorem, coming along with a trivial proof:

THEOREM 7.5. *Any two distinct points $P\neq Q$ determine a line, given by*

$$L=\left\{\lambda P+(1-\lambda)Q\Big|\lambda\in\mathbb R\right\}$$

in affine coordinates.

PROOF. This is somewhat clear, but let us do this in detail. We can say that the line L determined by P,Q consists of the set of vectors R such that we have:

$$QR\sim QP$$

That is, L consists of the set of vectors R such that we have, for a certain $\lambda\in\mathbb R$:

$$QR=\lambda QP$$

By using now the standard rules of vector calculus, this equation reads:

$$\begin{aligned}QR=\lambda QP&\iff R-Q=\lambda(P-Q)\\&\iff R=Q+\lambda(P-Q)\\&\iff R=\lambda P+(1-\lambda)Q\end{aligned}$$

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

Thus, very good news, axiom becoming theorem, what more can we wish for. Still speaking lines, let us have some further look at them. We have the following result:

THEOREM 7.6. *The lines in the plane are the solutions of equations of type*

$$ax+by+c=0$$

with $(a,b)\neq(0,0)$, and in addition, the following happen:

1. *Two such lines coincide when their triples (a,b,c) are proportional.*
2. *Two such lines are parallel or coincide when their pairs (a,b) are proportional.*

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

(1) As explained and Theorem 7.5 and its proof, with the convention that a line appears by uniting two points, the equations of these lines are as follows, with $P\neq Q$:

$$L=\left\{\lambda P+(1-\lambda)Q\Big|\lambda\in\mathbb R\right\}$$

Thus, in terms of coordinates, the lines are given by equations of the following type, with $(p,r)\neq(q,s)$, and with $\lambda\in\mathbb{R}$ being a parameter which varies:

$$\begin{cases}x=\lambda p+(1-\lambda)q\\ y=\lambda r+(1-\lambda)s\end{cases}$$

Equivalently, we can say that the lines are given by equations of the following type, with $(p,r)\neq(q,s)$, and with $\lambda\in\mathbb{R}$ being a parameter which varies:

$$\begin{cases}x=q+\lambda(p-q)\\ y=s+\lambda(r-s)\end{cases}$$

But now, by eliminating λ, in the obvious way, we are led to the conclusion that the lines are given by equations of the following type, with $(a,b)\neq(0,0)$:

$$ax+by+c=0$$

(2) In what regards now the second assertion, stating that two such lines coincide when their triples (a,b,c) are proportional, this is something clear.

(3) As for the last assertion, stating that two such lines are parallel or coincide when their pairs (a,b) are proportional, this is something clear too. □

In what follows we will often use the formula in Theorem 7.6, which is more convenient than the one in Theorem 7.5, for various algebraic computations. However, one problem with this sometimes comes from our lack of intuition regarding the parameters a,b,c. We will be back to this issue at the end of the present section, with an answer to it.

Moving on, and still following the material from the beginning of chapter 5, as a second piece of good news, our second geometry axiom becomes a theorem too:

THEOREM 7.7. *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 $P\in K$, $K||L$.*

PROOF. According to Theorem 7.6, we can assume that our line L is given by an equation of the following type, with $(a,b)\neq(0,0)$:

$$ax+by+c=0$$

As for the point P, with the notation $P=(x,y)$, the condition in the statement, namely $P\notin L$, tells us that the following must happen:

$$ax+by+c\neq0$$

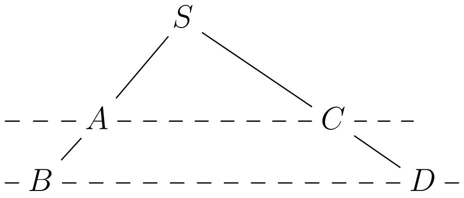
In view of this, let us pick now $\gamma\in\mathbb{R}$ such that the following equality holds:

$$ax+by+\gamma=0$$

But this formula, with x,y being now variables again, defines a certain line K, which certainly passes through P, and which is parallel to L too, as desired. □

Getting now to the next thing that we did in chapter 5, namely the Thales theorem, and as further good news, that can be proved too with coordinates, as follows:

Theorem 7.8 (Thales). *Proportions are kept, along parallel lines. That is, given a configuration as follows, consisting of two parallel lines, and of two extra lines,*



the following equality holds:

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

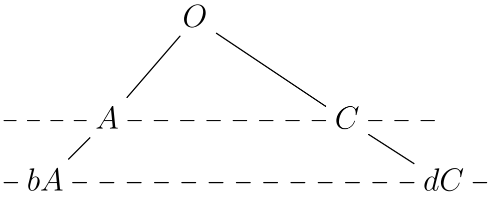
Moreover, the converse of this holds too, in the sense that, in the context of a picture as above, if this equality is satisfied, then the lines AC and BD must be parallel.

Proof. Many things can be said here, the idea being as follows:

(1) In what regards the main assertion, we can assume if we want, by translation, that the point S is the origin, $S=O$. Now with this assumption made, since O,A,B are collinear, and since O,C,D are collinear too, we must have, for certain $b,d\in\mathbb{R}$:

$$B=bA\quad,\quad D=dC$$

Thus, the picture of the Thales configuration becomes as follows, with $b,d\in\mathbb{R}$:



(2) Now let us prove the main assertion. We have the following equivalences:

$$\begin{aligned}AC||BD&\iff D-B=\lambda(C-A)\\&\iff dC-bA=\lambda(C-A)\\&\iff d=b\end{aligned}$$

But with this in hand, $d=b$, we obtain indeed the Thales formula, as follows:

$$\frac{OA}{OB}=\frac{1}{b}=\frac{1}{d}=\frac{OC}{OD}$$

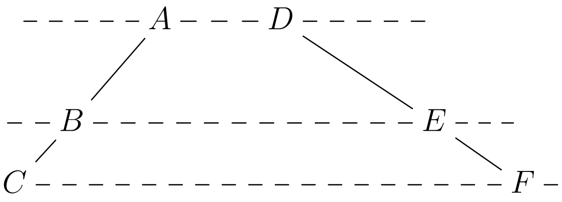
(3) Conversely now, we can still use the convention $S=O$ and the equalities $B=bA$ and $D=dC$ found in (1), and the picture there too, and we have, as claimed:

$$\begin{aligned}\frac{OA}{OB}=\frac{OC}{OD}&\implies\frac{1}{b}=\frac{1}{d}\\&\implies b=d\\&\implies AC||BD\end{aligned}$$

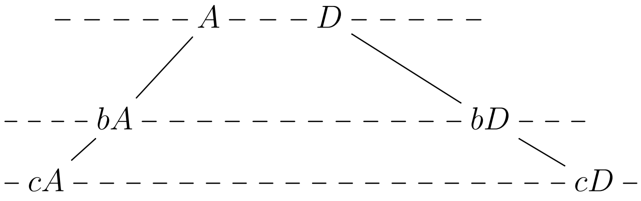
(4) Finally, let us mention that the other formulations of the Thales theorem, also from chapter 5, are also clear with coordinates. Indeed, for the above configuration, with the convention $S=O$, the improved conclusion, Thales 2, is as follows:

$$\frac{OA}{OB}=\frac{OC}{OD}=\frac{AC}{BD}$$

(5) Getting now to the Thales 3 configuration, also by following the material from chapter 5, this was as follows, with two lines meeting two parallel lines:



But here, save for a discussion of the case $AC||DF$, where the Thales 3 formula is clear, we can assume that $AC\cap DF$ is the origin O. And then, by proceeding as in (1), our picture becomes as follows, with $b,c\in\mathbb{R}$ being certain parameters:



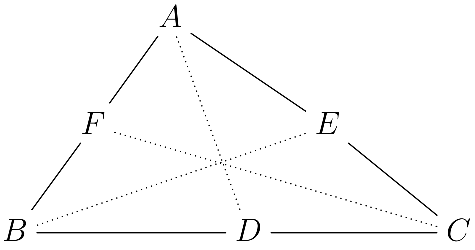
We conclude that the Thales 3 formula holds indeed, as follows:

$$\frac{AB}{BC}=\frac{b-1}{c-1}=\frac{DE}{EF}$$

Thus, fully done with Thales, in all its formulations, using coordinates. □

Getting now to triangles, the barycenter theorem drastically simplifies, as follows:

Theorem 7.9 (Barycenter). *Given a triangle ABC, its medians cross,*



at a point called barycenter, lying at $1/3-2/3$ on each median.

Proof. Let us call $A,B,C\in\mathbb R^2$ the coordinates of the vertices 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. □

Regarding now more advanced plane geometry results, these can be often investigated by using scalar products, whose theory can be summarized as follows:

Theorem 7.10. *If we define the scalar product of two vectors by*

$$\left\langle\begin{pmatrix}a\\b\end{pmatrix},\begin{pmatrix}c\\d\end{pmatrix}\right\rangle=ac+bd$$

then the following happen:

(1) $<A+B,C>=<A,C>+<B,C>$.
(2) $<A,B+C>=<A,B>+<A,C>$.
(3) $<\lambda A,B>=<A,\lambda B>=\lambda<A,B>$.
(4) $||A||=\sqrt{<A,A>}$.
(5) $A\perp B\iff<A,B>=0$.
(6) $<A,B>=||A||\cdot||B||\cdot\cos t$, *with t being the angle between A,B.*
(7) $<A,B>=<A',B>=<A,B'>$, *prime meaning projection on the other vector.*

In addition, the line equation $ax+by+c=0$ can be written as $<\binom{a}{b},\binom{x}{y}>=-c$.

Proof. Many things going on here, the idea being as follows:

(1-3) These formulae, very useful in practice, are all clear from definitions.

(4-7) To start with, the formula in (4) is clear, coming from:

$$\left\langle\begin{pmatrix}a\\b\end{pmatrix},\begin{pmatrix}a\\b\end{pmatrix}\right\rangle=a^2+b^2=\left|\left|\begin{pmatrix}a\\b\end{pmatrix}\right|\right|^2$$

Observe that this formula agrees with what (6) says. In fact, more generally, the scalar product of two proportional vectors is as follows, again in agreement with (6):

$$\left\langle\begin{pmatrix}a\\ b\end{pmatrix},\begin{pmatrix}\lambda a\\ \lambda b\end{pmatrix}\right\rangle=\lambda a^2+\lambda b^2=\pm\left|\left|\begin{pmatrix}a\\ b\end{pmatrix}\right|\right|\cdot\left|\left|\begin{pmatrix}\lambda a\\ \lambda b\end{pmatrix}\right|\right|$$

In order to prove now (5), we can assume using (3) that we have $||A||=||B||=1$. But here, assuming $A\perp B$, if s is the angle formed by A with the Ox axis, we have:

$$<A,B>=\left\langle\begin{pmatrix}\cos s\\ \sin s\end{pmatrix},\pm\begin{pmatrix}-\sin s\\ \cos s\end{pmatrix}\right\rangle=0$$

Getting now to (6), which will prove as well the converse of this, again we can assume $||A||=||B||=1$, and if s is the angle formed by A with the Ox axis, we have:

$$\begin{aligned}<A,B> &= \left\langle\begin{pmatrix}\cos s\\ \sin s\end{pmatrix},\begin{pmatrix}\cos(s+t)\\ \sin(s+t)\end{pmatrix}\right\rangle\\ &= \cos s\cos(s+t)+\sin s\sin(s+t)\\ &= \cos((s+t)-s)\\ &= \cos t\end{aligned}$$

As for (7), this is a reformulation of (6), using the above formula of $<A,\lambda A>$.

(8) Finally, the last assertion is clear, and with this answering a question raised after Theorem 7.6. By the way, talking answers to previous questions, observe that (6) provides an answer to our philosophical questions regarding the cosine, from chapter 6. □

As a conclusion, the coordinates perform quite well. There are of course many exercises that can be worked out here, and we will formulate some, at the end of this chapter.

7b. Complex plane

Let us discuss now the complex numbers, which are an even stronger tool. There is a lot of magic here, and we will carefully explain this material. We first have:

Definition 7.11. *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:

Theorem 7.12. *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 7.13. *The complex numbers, written as usual*

$$x = a + ib$$

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

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

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

Proof. Consider indeed two arbitrary complex numbers:

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

Their sum is then by definition the following complex number:

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

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

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

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

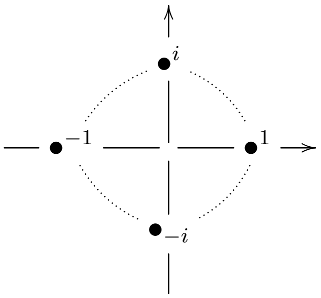
$$x + y = \begin{pmatrix} a + c \\ b + d \end{pmatrix}$$

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

Observe that in our geometric picture from Proposition 7.13, 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:



Summarizing, 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 7.14. *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,

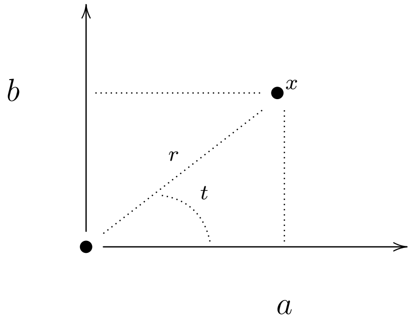
$$a = r\cos t \quad , \quad b = r\sin t$$

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 7.13, 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 7.14 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 7.15. *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 7.16. *The powers of a complex number, written in polar form,*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

As a basic application of Theorem 7.16, we have the following result:

Proposition 7.17. *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 7.16, at $k = 1/2$. As for its particular cases with $x \in \mathbb{R}$, these are clear from it. □

With the above results in hand, and notably with the square root formula from Proposition 7.17, we can go back now to the degree 2 equations, and we have:

Theorem 7.18. *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 7.17 that any complex number has a square root, we are led to the conclusion in the statement. □

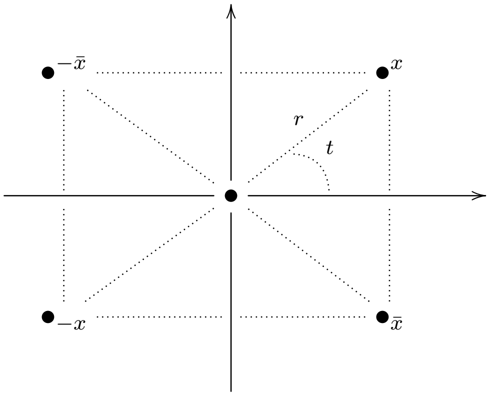
As a last general topic regarding the complex numbers, let us discuss conjugation. This is something quite tricky, complex number specific, as follows:

Definition 7.19. *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 7.20. *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 7.19.

(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 7.12, 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. Euler formula

We would like to discuss now the final and most convenient writing of the complex numbers, which is a variation on the polar writing, as follows:

$$x=re^{it}$$

For this purpose, let us start with the following basic result:

THEOREM 7.21. *We can exponentiate the complex numbers, according to the formula*

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

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

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

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

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

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

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

As a consequence of the above formula $e^{x+y}=e^xe^y$, we have the following result:

PROPOSITION 7.22. *The exponential of complex numbers is given by*

$$e^{s+it}=e^se^{it}$$

with e^s being a usual real exponential, and with e^{it}, in need to be computed.

PROOF. This is indeed something self-explanatory, coming from $e^{x+y}=e^xe^y$, and with the somewhat non-standard notation $x=s+it$ being something needed later. □

Now let us get to the remaining problem, computation of e^{it} with $t\in\mathbb{R}$. Here are a few elementary observations, regarding the operation $t\to e^{it}$:

Proposition 7.23. *For $t \in \mathbb{R}$ the number e^{it} belongs to the unit circle,*

$$e^{it} \in \mathbb{T}$$

and the operation $t \to e^{it}$ is subject to the following formulae,

$$e^{i(s+t)} = e^{is}e^{it} \quad , \quad e^{i0} = 1 \quad , \quad (e^{it})^{-1} = e^{it}$$

telling us $t \to e^{it}$ is a group morphism $\mathbb{R} \to \mathbb{T}$.

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

(1) To start with, we have the following formula, valid for any $x \in \mathbb{C}$:

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

We have as well the following computation, again valid for any $x \in \mathbb{C}$:

$$e^x e^{-x} = e^{x-x} = e^0 = 1 \implies (e^x)^{-1} = e^{-x}$$

(2) But with these two formulae in hand, we can prove the first assertion. Indeed, the first formula, applied with $x = it$, with $t \in \mathbb{R}$, gives the following equality:

$$e^{-it} = \overline{e^{it}}$$

As for the second formula above, again applied with $x = it$, this gives:

$$(e^{it})^{-1} = e^{it}$$

We conclude that the complex number $z = e^{it}$ has the following property:

$$z^{-1} = \bar{z}$$

But this is exactly the equation of the unit circle $\mathbb{T}$, as desired.

(3) Regarding now the various formulae in the statement, for the operation $t \to e^{it}$, these are all trivial, coming from the multiplicativity formula $e^{x+y} = e^x e^y$.

(4) As for the final conclusion, this is something quite intuitive, telling us that $t \to e^{it}$ transforms the additive structure of $\mathbb{R}$ into the multiplicative structure of $\mathbb{T}$. □

What is next? Well, we will have to improvise a bit, and we are led in this way to the following fundamental result of Euler, regarding the complex exponential:

Theorem 7.24. *We have the following formula,*

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

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

Proof. There are several possible proofs of this, the idea being as follows:

(1) Intuitive proof. We know from Proposition 7.23 that $t \to e^{it}$ is a group morphism $\mathbb{R} \to \mathbb{T}$. But in view of this, barring any pathologies, this operation can only appear by "wrapping". That is, we must have a formula as follows, for a certain $\alpha \in \mathbb{R}$:

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

In order now to find the parameter $\alpha \in \mathbb{R}$, let us look at what happens around $t = 0$. As a first observation, at $t = 0$ precisely, our formula is as follows, true:

$$e^0 = \cos 0 + i \sin 0$$

The point now is that, around $t = 0$, we have the following elementary estimate, simply obtained by truncating the series defining the exponential:

$$e^{it} \simeq 1 + it$$

On the other hand, we know from chapter 6 that we have $\sin t \simeq t$ and $\cos t \simeq 1 - t^2/2$, for $t \simeq 0$. We conclude that we have the following estimate, for $t \simeq 0$:

$$\cos(\alpha t) + i\sin(\alpha t) \simeq 1 + i\alpha t$$

Thus we must have $\alpha = 1$, and we are led to the Euler formula in the statement.

(2) Calculus proof. This is something more solid, obtained by differentiating the following function, using the various available calculus rules, and getting 0:

$$f(t) = e^{-it}(\cos t + i \sin t)$$

Indeed, this shows that our function f must be constant, equal to $f(0) = 1$, as desired. We will discuss this in detail in chapter 11, when doing calculus. □

As a well-known application of the Euler formula, we have:

Theorem 7.25. *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 7.24, 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 [31], [32], [33]. □

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

THEOREM 7.26. *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 7.14, by using the formula $e^{it} = \cos t + i \sin t$ from Theorem 7.24, 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 7.27. *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 7.15, reformulated by using the notations from Theorem 7.26. Observe that this follows as well directly, from the fact that we have $e^{x+y} = e^x e^y$, that we know from Theorem 7.21. □

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 7.28. *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 7.16, 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 7.27:

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

(2) Once again by using Theorem 7.27, 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 7.27, this operation $x\to x^a$ is indeed the correct one.

(4) This comes from the fact, that we know from Theorem 7.20, 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 7.18 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 7.29. *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. As before with the Euler formula, we are punching here a bit above our weight, because investigating such things normally needs some training in analysis, which will be job for us in Part III. This being said, here is how the proof goes:

(1) 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:

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

(2) 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:

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

(3) If we write $t=rw$, with $r>0$ small, and with $|w|=1$, our estimate becomes:

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

(4) 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:

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

(5) 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:

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

(6) 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. □

7d. Roots of unity

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

Theorem 7.30. *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 7.27. Indeed, with $x=re^{it}$ our equation reads:

$$r^Ne^{itN}=1$$

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

As an illustration here, the roots of unity of small order 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 $\{\pm1,\pm w,\pm w^2\}$, 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 7.16 and Theorem 7.28, the statement being as follows:

Theorem 7.31. *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. □

In relation now with geometry, the roots of unity are something very useful. For instance in order to draw equilateral triangles, what we need to do is to multiply by $w = e^{2\pi i/3}$, or by $w^2 = e^{4\pi i/3}$. We will leave some study here as an exercise.

7e. Exercises

There is nothing more pleasant and relaxing than coordinates, gone all that plane geometry tricks, with coordinates everything works. As exercises on all this, we have:

Exercise 7.32. *Prove the Desargues theorem, using coordinates.*

Exercise 7.33. *Prove the Pappus theorem, using coordinates.*

Exercise 7.34. *Prove the Menelaus theorem, using coordinates.*

Exercise 7.35. *Prove the Ceva theorem, using coordinates.*

Exercise 7.36. *Discuss the Euler line, using coordinates.*

Exercise 7.37. *Discuss the nine-point circle, using coordinates.*

Exercise 7.38. *Clarify what we said above, in relation with scalar products.*

Exercise 7.39. *Solve the above exercises, and more, by using complex coordinates.*

As bonus exercise, read a bit of history of mathematics, and in particular, learn about the history of coordinates, in plane and in space, and in mathematics and physics.

CHAPTER 8

Plane curves

8a. Ellipses, conics

Time to end the present Part II of this book, on geometry, with some tough and beautiful results, coming as a continuation of the above. And there are so many things to be discussed here, namely the conics, which are the core of the whole modern mathematics and physics, then more general plane curves, and then with a look into $\mathbb{R}^3$ too.

We will be quite brief, with some proofs missing, and with some other based on material that we have not studied yet, coming in Parts III and IV. So, take what we will be talking about here as a nice story, or as a physics class if you prefer, and for more, come back here after the whole book read, and you will certainly understand better.

Let us start with some astronomy. Looking up, to the sky, the first thing that you see is the Sun, seemingly moving around the Earth on a circle. However, a more careful study reveals that this circle is rather a deformed circle, called ellipse. And good news, a full theory of ellipses is available, and this since the ancient Greeks, as follows:

Theorem 8.1. *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* $l\in(0,2q)$*:*

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

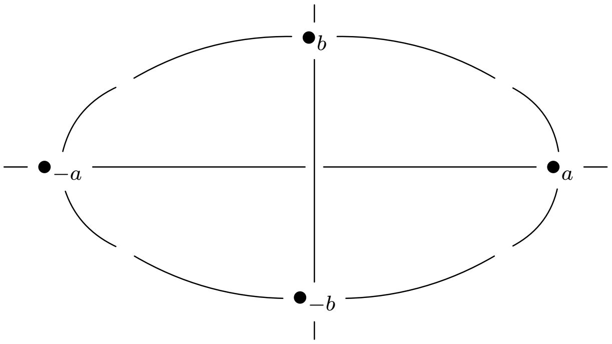
(3) *As the curves appearing when drawing a circle, from various perspectives:*

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

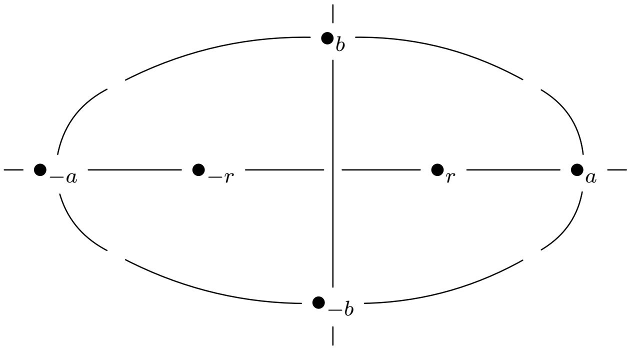
(4) *As the closed non-degenerate curves appearing by cutting a cone with a plane.*

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 ellipse:



(2) Let us prove now that such an ellipse has two focal points, as stated in (2). We must look for a number $r > 0$, and a number $l > 0$, such that our ellipse 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, labeled (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 8.2. *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 8.1, as follows:

(1) We first need to talk about linear transformations of the plane. Normally this is the business of linear algebra, to be discussed in Part IV, but for our purposes here, we will only need some basics, which are not hard to explain. So, recall that the main operations on vectors are the sum, and the multiplication by scalars:

$$\begin{pmatrix}x\\y\end{pmatrix}+\begin{pmatrix}z\\t\end{pmatrix}=\begin{pmatrix}x+z\\y+t\end{pmatrix}\quad,\quad\lambda\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}\lambda x\\\lambda y\end{pmatrix}$$

Now since these two operations are what produces our vector mathematics, it makes sense to look at the transformations of the plane $f:\mathbb{R}^2\to\mathbb{R}^2$ preserving them:

$$f(p+q)=f(p)+f(q)\quad,\quad f(\lambda p)=\lambda f(p)$$

Such maps $f:\mathbb{R}^2\to\mathbb{R}^2$ are called linear, and in practice, there are plenty of them, including all the rotations around 0, all the symmetries with respect to lines passing through 0, and all the projections on these same lines passing through 0, too.

(2) Getting now to the mathematics of the linear maps $f:\mathbb{R}^2\to\mathbb{R}^2$, we have:

$$\begin{aligned}f\begin{pmatrix}x\\y\end{pmatrix}&=f\left(\begin{pmatrix}x\\0\end{pmatrix}+\begin{pmatrix}0\\y\end{pmatrix}\right)\\&=f\left(x\begin{pmatrix}1\\0\end{pmatrix}+y\begin{pmatrix}0\\1\end{pmatrix}\right)\\&=xf\begin{pmatrix}1\\0\end{pmatrix}+yf\begin{pmatrix}0\\1\end{pmatrix}\end{aligned}$$

Thus, if we set $\binom{a}{c}=f\binom{1}{0}$ and $\binom{b}{d}=f\binom{0}{1}$, we have the following formula:

$$f\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}ax+by\\cx+dy\end{pmatrix}$$

And the point is that, conversely, any such formula defines a linear map $f:\mathbb{R}^2\to\mathbb{R}^2$, with both the formulae $f(p+q)=f(p)+f(q)$ and $f(\lambda p)=\lambda f(p)$ being obvious.

(3) As a continuation of this, at a more advanced level, we can say that such maps $f:\mathbb{R}^2\to\mathbb{R}^2$ come from the 2×2 matrices, according to the following formula:

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

Indeed, in order to do so, we can define the multiplication on the right as follows:

$$\begin{pmatrix}a&b\\c&d\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}ax+by\\cx+dy\end{pmatrix}$$

(4) Finally, as a last piece of theory, some of our linear maps $f:\mathbb{R}^2\to\mathbb{R}^2$ are "degenerate", with this happening when the vectors $\binom{a}{c}=f\binom{1}{0}$ and $\binom{b}{d}=f\binom{0}{1}$ are proportional. But this latter proportionality means $ad=bc$, so if we want to restrict the attention to the non-degenerate maps, we must impose the following condition:

$$ad\neq bc$$

(5) Getting to work now, we would first like to classify the conics up to non-degenerate linear transformations, which these being the transformations as follows:

$$f\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}ax+by\\cx+dy\end{pmatrix}\quad,\quad ad\neq bc$$

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

(6) 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.

(7) Getting now to the proof of our claim in (5), 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 $l=0$.

(8) 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 $f=0$.

(9) Thus, done with the classification, up to linear transformations as in (5). 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.

(10) 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 $(a,b,c)\neq(0,0,0)$:

$$ax+by+cz=d$$

(11) 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:

$$z=0$$

(12) 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. □

In relation now with physics, following Kepler and Newton, we have:

Theorem 8.3. *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 very standard, which needs however some advanced calculus, that we will learn later in this book. So, patience, we will learn the needed calculus soon, and come back to this, with a proof. In the meantime, here is the idea:

(1) According to observations and calculations performed over the centuries, and first formalized by Newton, following some groundbreaking work of Kepler, the force of attraction between two bodies of masses M,m is given by the following formula, with d being the distance between the two bodies, and $G\simeq 6.674\times 10^{-11}$ being a constant:

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

(2) Now assuming that M is fixed at $0\in\mathbb R^2$, the force exterted on m positioned at $x\in\mathbb R^2$, regarded as vector $F\in\mathbb R^2$, is given by the following formula, with $K=GM$:

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

(3) On the other hand, again following Newton, we have the following sequence of general equalities, with x,v,a being the position, speed and acceleration, and with the dot, called derivative, standing for the rate of change of the function in question:

$$F=ma=m\dot v=m\ddot x$$

(4) Now by putting everything together, we conclude that the equation of motion of m, assuming that M is fixed at 0, is something quite simple, as follows:

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

(5) Let us first study a simple particular case, that of the circular solutions. To be more precise, we are interested in solutions of the following type:

$$x=(r\cos\alpha t,r\sin\alpha t)$$

In this case we have $||x||=r$, so our equation of motion becomes:

$$\ddot x=-\frac{Kx}{r^3}$$

On the other hand, differentiating x twice leads to the following formula:

$$\ddot x=-\alpha^2x$$

Thus, we have a circular solution when the parameters r,α satisfy:

$$r^3\alpha^2=K$$

(6) In general now, when looking for arbitrary solutions, things are certainly more complicated, but the idea remains the same, namely polar coordinates, and calculus. And this leads to conics, as stated. We will discuss this, with details, in chapter 16. □

8b. Algebraic curves

As a conclusion to what we did so far, conics are at the core of everything, mathematics, physics, life. But, what is next? A natural answer to this question comes from:

Definition 8.4. *An algebraic curve in $\mathbb{R}^2$ is the vanishing set*

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

of a polynomial $P\in\mathbb{R}[X,Y]$ of arbitrary degree.

We already know well the algebraic curves in degree 2, which are the conics, and a first problem is, what results from what we learned about conics have a chance to be relevant to the arbitrary algebraic curves. And normally none, because the ellipses, parabolas and hyperbolas are obviously very particular curves, having very particular properties.

Let us record however a useful statement here, as follows:

Proposition 8.5. *The conics can be written in cartesian, polar, parametric or complex coordinates, with the equations for the unit circle being*

$$x^2+y^2=1\quad,\quad r=1\quad,\quad x=\cos t\,,\,y=\sin t\quad,\quad |z|=1$$

and with the equations for ellipses, parabolas and hyperbolas being similar.

Proof. The equations for the circle are clear, those for ellipses can be found in the above, and we will leave as an exercise those for parabolas and hyperbolas. □

As a true answer to our question now, coming this time from a very modest conic, namely $xy=0$, that we dismissed in the above as being "degenerate", we have:

Theorem 8.6. *The following happen, for curves C defined by polynomials P:*

1. *In degree $d=2$, curves can have singularities, such as $xy=0$ at $(0,0)$.*
2. *In general, assuming $P=P_1\ldots P_k$, we have $C=C_1\cup\ldots\ldots\cup C_k$.*
3. *A union of curves $C_i\cup C_j$ is generically non-smooth, unless disjoint.*
4. *Due to this, we say that C is non-degenerate when P is irreducible.*

Proof. All this is self-explanatory, the details being as follows:

(1) This is something obvious, just the story of two lines crossing.

(2) This comes from the following trivial fact, with the notation $z=(x,y)$:

$$P_1\ldots P_k(z)=0\iff P_1(z)=0,\text{ or }P_2(z)=0,\ \ldots\ ,\text{ or }P_k(z)=0$$

(3) This is something very intuitive, and it actually takes a bit of time to imagine a situation where $C_1\cap C_2\neq\emptyset$, $C_1\not\subset C_2$, $C_2\not\subset C_1$, but $C_1\cup C_2$ is smooth. In practice now, "generically" has of course a mathematical meaning, in relation with probability, and our assertion does say something mathematical, that we are supposed to prove. But,

we will not insist on this, and leave this as an instructive exercise, precise formulation of the claim, and its proof, in the case you are familiar with probability theory.

(4) This is just a definition, based on the above, that we will use in what follows. □

With degree 1 and 2 investigated, and our conclusions recorded, let us get now to degree 3, see what new phenomena appear here. And here, to start with, we have the following remarkable curve, well-known from calculus, because 0 is not a maximum or minimum of the function $x \to y$, despite the derivative vanishing there:

$$x^3 = y$$

Also, in relation with set theory and logic, and with the foundations of mathematics in general, we have the following curve, which looks like the empyset $\emptyset$:

$$(x - y)(x^2 + y^2 - 1) = 0$$

But, it is not about counterexamples to calculus, or about logic, that we want to talk about here. As a first truly remarkable degree 3 curve, or cubic, we have the cusp:

Proposition 8.7. *The standard cusp, which is the cubic given by*

$$x^3 = y^2$$

has a singularity at $(0, 0)$*, with only* 1 *tangent line at that singularity.*

Proof. The two branches of the cusp are indeed both tangent to Ox, because:

$$y' = \pm\frac{3}{2}\sqrt{x} \implies y'(0) = 0$$

Observe also that what happens for the cusp is different from what happens for $xy = 0$, precisely because we have 1 line tangent at the singularity, instead of 2. □

As a second remarkable cubic, which gets the crown, and the right to have a Theorem about it, we have the Tschirnhausen curve, which is as follows:

Theorem 8.8. *The Tschirnhausen cubic, given by the following equation,*

$$x^3 = x^2 - 3y^2$$

makes the dream of $xy = 0$ *come true, by self-intersecting, and being non-degenerate.*

Proof. This is something self-explanatory, by drawing a picture, but there are several other interesting things that can be said about this curve, and the family of curves containing it, depending on a parameter, and up to basic transformations, as follows:

(1) Let us start with the curve written in polar coordinates as follows:

$$r\cos^3\left(\frac{\theta}{3}\right) = a$$

With $t=\tan(\theta/3)$, the equations of the coordinates are as follows:

$$x=a(1-3t^2)\quad,\quad y=at(3-t^2)$$

Now by eliminating t, we reach to the following equation:

$$(a-x)(8a+x)^2=27ay^2$$

(2) By translating horizontally by $8a$, and changing signs of variables, we have:

$$x=3a(3-t^2)\quad,\quad y=at(3-t^2)$$

Now by eliminating t, we reach to the following equation:

$$x^3=9a(x^2-3y^2)$$

But with $a=1/9$ this is precisely the equation in the statement. □

In degree 4 now, quartics, we have enough dimensions for "improving" the cusp and the Tschirnhausen curve. First we have the cardioid, which is as follows:

Proposition 8.9. *The cardioid, which is a quartic, given in polar coordinates by*

$$2r=a(1-\cos\theta)$$

makes the dream of $x^3=y^2$ come true, by being a closed curve, with a cusp.

Proof. As before with the Tschirnhausen curve, this is something self-explanatory, by drawing a picture, but there are several things that must be said, as follows:

(1) The cardioid appears by definition by rolling a circle of radius $c>0$ around another circle of same radius $c>0$. With θ being the rolling angle, we have:

$$x=2c(1-\cos\theta)\cos\theta$$

$$y=2c(1-\cos\theta)\sin\theta$$

(2) Thus, in polar coordinates we get the equation in the statement, with $a=4c$:

$$r=2c(1-\cos\theta)$$

(3) Finally, in cartesian coordinates, the equation is as follows:

$$(x^2+y^2)^2+4cx(x^2+y^2)=4c^2y^2$$

Thus, what we have is indeed a degree 4 curve, as claimed. □

Still in degree 4, the crown gets to the Bernoulli lemniscate, which is as follows:

Theorem 8.10. *The Bernoulli lemniscate, a quartic, which is given by*

$$r^2=a^2\cos 2\theta$$

makes the dream of $x^3=x^2-3y^2$ come true, by being closed, and self-intersecting.

Proof. As usual, this is something self-explanatory, by drawing a picture, which looks like ∞, but there are several other things that must be said, as follows:

(1) In cartesian coordinates, the equation is as follows, with $a^2 = 2c^2$:

$$(x^2 + y^2)^2 = c^2(x^2 - y^2)$$

(2) Also, we have the following nice complex reformulation of this equation:

$$|z + c| \cdot |z - c| = c^2$$

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

In degree 5, in the lack of any spectacular quintic, let us record:

Theorem 8.11. *Unlike in degree* $3, 4$*, where equations can be solved, by the Cardano formula, in degree* 5 *this generically does not happen, an example being*

$$x^5 - x - 1 = 0$$

having Galois group S_5*, not solvable. Geometrically, this tells us that the intersection of the quintic* $y = x^5 - x - 1$ *with the line* $y = 0$ *cannot be computed.*

Proof. Obviously off-topic, but with no good quintic available, and still a few more minutes before the bell ringing, I had to improvise a bit, and tell you about this:

(1) As indicated, the degree 3 equations can be solved a bit like the degree 2 ones, but with the formula, due to Cardano, being more complicated. With some square making tricks, which are non-trivial either, the Cardano formula applies to degree 4 as well.

(2) In degree 5 or higher, none of this is possible. Long story here, the idea being that in order for $P = 0$ to be solvable, the group $Gal(P)$ must be solvable, in the sense of group theory. But, unlike S_3, S_4 which are solvable, S_5 and higher are not solvable. □

Back now to our usual business, in degree 6, sextics, we first have here:

Proposition 8.12. *The trefoil sextic, or Kiepert curve, which is given by*

$$r^3 = a^3 \cos 3\theta$$

looks like a trefoil, closed curve, with a triple self-intersection.

PROOF. As before, drawing a picture is mandatory. With $z=re^{i\theta}$ we have:

$$\begin{aligned}
r^3=a^3\cos 3\theta \quad &\iff \quad r^3\cos 3\theta=\left(\frac{r^2}{a}\right)^3\\
&\iff \quad z^3+\bar{z}^3=2\left(\frac{z\bar{z}}{a}\right)^3\\
&\iff \quad (x+iy)^3+(x-iy)^3=2\left(\frac{x^2+y^2}{a}\right)^3\\
&\iff \quad x^3-3xy^2=\left(\frac{x^2+y^2}{a}\right)^3\\
&\iff \quad (x^2+y^2)^3=a^3(x^3-3xy^2)
\end{aligned}$$

Thus, we have indeed a sextic, as claimed. □

We also have in degree 6 the most beautiful of curves them all, the Cayley sextic:

THEOREM 8.13. *The Cayley sextic, given in polar coordinates by*

$$r=a\cos^3\left(\frac{\theta}{3}\right)$$

makes the dream of everyone come true, by looking like a self-intersecting heart.

PROOF. As before, picture mandatory. With $z=re^{i\theta}$ and $u=z^{1/3}$ we have:

$$\begin{aligned}
r=a\cos^3\left(\frac{\theta}{3}\right) \quad &\iff \quad ar\cos^3\left(\frac{\theta}{3}\right)=r^2\\
&\iff \quad a\left(\frac{u+\bar{u}}{2}\right)^3=r^2\\
&\iff \quad a(u^3+\bar{u}^3+3u\bar{u}(u+\bar{u}))=8r^2\\
&\iff \quad 3au\bar{u}\cdot\frac{u+\bar{u}}{2}=4r^2-ax\\
&\iff \quad 27a^3r^6\cdot\frac{r^2}{a}=(4r^2-ax)^3\\
&\iff \quad 27a^2(x^2+y^2)^2=(4x^2+4y^2-ax)^3
\end{aligned}$$

Thus, we have indeed a sextic, as claimed. □

8c. Spirals, lemniscates

Quite remarkably, most of the above curves are sinusoidal spirals, in the following sense, and with actually the term "sinusoidal spiral" being a bit unfortunate:

THEOREM 8.14. *The sinusoidal spirals, which are as follows,*

$$r^n = a^n \cos n\theta$$

with $a \neq 0$ and $n \in \mathbb{Q} - \{0\}$, include the following curves:

(1) *$n = -1$ line.*
(2) *$n = 1$ circle, $n = -1/2$ parabola, $n = -2$ hyperbola.*
(3) *$n = -3$ Humbert cubic, $n = -1/3$ Tschirnhausen curve.*
(4) *$n = 1/2$ cardioid, $n = 2$ Bernoulli lemniscate.*
(5) *$n = 3$ Kiepert trefoil, $n = 1/3$ Cayley sextic.*

PROOF. We first have to prove that the sinusoidal spirals are indeed algebraic curves. But this is best done by using the complex coordinate $z = re^{i\theta}$, as follows:

$$\begin{aligned} r^n = a^n \cos n\theta \quad &\iff \quad r^n \cos n\theta = \left(\frac{r^2}{a}\right)^n \\ &\iff \quad z^n + \bar{z}^n = 2\left(\frac{z\bar{z}}{a}\right)^n \\ &\iff \quad (x+iy)^n + (x-iy)^n = 2\left(\frac{x^2+y^2}{a}\right)^n \end{aligned}$$

As a first observation now, in the case $n \in \mathbb{N}$ we can simply use the binomial formula, and we get an algebraic equation of degree $2n$, as follows:

$$\sum_{k=0}^{[n/2]} (-1)^k \binom{n}{2k} x^{n-2k} y^{2k} = \left(\frac{x^2+y^2}{a}\right)^n$$

In general, things are a bit more complicated, as shown for instance by our computation for the Cayley sextic. However, the same idea as there applies, and we are led in this way to the equation of an algebraic curve, as claimed. Regarding now the examples:

(1) At $n = -1$ the equation is as follows, producing a line:

$$r \cos \theta = a \iff x = a$$

(2) At $n = 1$ the equation is as follows, producing a circle:

$$r = a \cos \theta \iff r^2 = ax \iff x^2 + y^2 = ax$$

(3) At $n = -1/2$ the equation is as follows, producing a parabola:

$$a = r \cos^2(\theta/2) \iff r + x = 2a \iff y^2 = 4a(a - x)$$

(4) At $n = -2$ the equation is as follows, producing a hyperbola:

$$a^2 = r \cos^2 2\theta \iff a^2 = 2x^2 - r^2 \iff (x+y)(x-y) = a^2$$

(5) At $n=-3$ the equation is as follows, producing a curve with 3 components, which looks like some sort of "trivalent hyperbola", called Humbert cubic:

$$r^3\cos 3\theta=a^3\iff z^3+\bar{z}^3=2a^3\iff x^3-3xy^2=a^3$$

(6) As for the other curves, this follows from our various formulae above. □

Let us study now more in detail the sinusoidal spirals. We first have:

Proposition 8.15. *The sinusoidal spirals, which with $z=x+iy$ are*

$$z^n+\bar{z}^n=2\left(\frac{z\bar{z}}{a}\right)^n$$

with $a\neq 0$ and $n\in\mathbb{Q}-\{0\}$, are as follows:

1. *With $n=-m$, $m\in\mathbb{N}$, the equation is $z^m+\bar{z}^m=2a^m$, degree m.*
2. *With $n=m$, $m\in\mathbb{N}$, the equation is $z^m+\bar{z}^m=2(z\bar{z}/a)^m$, degree $2m$.*
3. *With $n=-1/m$, $m\in\mathbb{N}$, the equation is $(z^{1/m}+\bar{z}^{1/m})^m=2^ma$.*
4. *With $n=1/m$, $m\in\mathbb{N}$, the equation is $(z^{1/m}+\bar{z}^{1/m})^m=2^mz\bar{z}/a$.*

Proof. This is something self-explanatory, the details being as follows:

(1) With $n=-m$ and $m\in\mathbb{N}$ as in the statement, the equation is, as claimed:

$$z^{-m}+\bar{z}^{-m}=2\left(\frac{z\bar{z}}{a}\right)^{-m}\iff z^m+\bar{z}^m=2a^m$$

(2) This is an empty statement, just a matter of using the new variable $m=n$.

(3) With $n=-1/m$ and $m\in\mathbb{N}$ as in the statement, the equation is, as claimed:

$$\begin{aligned}z^{-1/m}+\bar{z}^{-1/m}=2\left(\frac{z\bar{z}}{a}\right)^{-1/m}&\iff z^{1/m}+\bar{z}^{1/m}=2a^{1/m}\\&\iff (z^{1/m}+\bar{z}^{1/m})^m=2^ma\end{aligned}$$

(4) With $n=1/m$ and $m\in\mathbb{N}$ as in the statement, the equation is, as claimed:

$$z^{1/m}+\bar{z}^{1/m}=2\left(\frac{z\bar{z}}{a}\right)^{1/m}\iff (z^{1/m}+\bar{z}^{1/m})^m=2^m\cdot\frac{z\bar{z}}{a}$$

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

Observe that in the fractionary cases, $n=\pm 1/m$, the equations in the above statement are not polynomial in x,y, unless at very small values of m. To be more precise:

(1) In the case $n=-1/m$, we certainly have at $m=1,2,3$ the $d=1$ line, $d=2$ parabola, and $d=3$ Tschirnhausen curve, but at $m=4$ things change, with the equation $(z^{1/4}+\bar{z}^{1/4})^4=16a$ being no longer polynomial in x,y, and requiring a further square operation to make it polynomial, and therefore leading to a curve of degree $d=8$.

(2) As for the case $n=1/m$, this is more complicated, with the data that we have at $m=1,2,3$, namely the $d=2$ circle, $d=3$ cardioid, and $d=6$ Cayley sextic, being not very good, and with things getting even more complicated at $m=4$ and higher.

In short, things quite complicated, and the general case, $n=\pm p/q$ with $p,q\in\mathbb N$, is certainly even more complicated. Instead of insisting on this, let us focus now on the simplest sinusoidal spirals that we have, namely those with $n=\pm m$, with $m\in\mathbb N$.

The point indeed is that the sinusoidal spirals with $n\in\mathbb N$ are also part of another remarkable family of plane algebraic curves, going back to Cassini, as follows:

Theorem 8.16. *The polynomial lemniscates, which are as follows,*

$$|P(z)|=b^n$$

with $P\in\mathbb C[X]$ having n distinct roots, and $b>0$, include the following curves:

1. *The sinusoidal spirals with $n\in\mathbb N$, including the $n=1$ circle, $n=2$ Bernoulli lemniscate, and $n=3$ Kiepert trefoil.*
2. *The Cassini ovals, which are the quartics given by $|z+c|\cdot|z-c|=b^2$, covering too the Bernoulli lemniscate, appearing at $b=c$.*

Proof. This is something quite self-explanatory, the details being as follows:

(1) Regarding the sinusoidal spirals with $n\in\mathbb N$, their equation is, with $a^n=2c^n$:

$$\begin{aligned}z^n+\bar z^n=2\left(\frac{z\bar z}{a}\right)^n &\iff c^n(z^n+\bar z^n)=(z\bar z)^n\\ &\iff (z^n-c^n)(\bar z^n-c^n)=c^{2n}\\ &\iff |z^n-c^n|=c^n\end{aligned}$$

(2) Regarding the Cassini ovals, these correspond to the case where the polynomial $P\in\mathbb C[X]$ has degree 2, and we already know from the above that these cover the Bernoulli lemniscate. In general, the equation for the Cassini ovals is:

$$\begin{aligned}|z+c|\cdot|z-c|=b^2 &\iff |z^2-c^2|=b^2\\ &\iff (z^2-c^2)(\bar z^2-c^2)=b^4\\ &\iff (z\bar z)^2-c^2(z^2+\bar z^2)+c^4=b^4\\ &\iff (x^2+y^2)^2-c^2(x^2-y^2)+c^4=b^4\\ &\iff (x^2+y^2)^2=c^2(x^2-y^2)+b^4-c^4\end{aligned}$$

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

The polynomial lemniscates can be geometrically understood as follows:

Theorem 8.17. *The equation $|P(z)| = b$ defining the polynomial lemniscates can be written as follows, in terms of the roots $c_1, \ldots, c_n$ of the polynomial P,*

$$\sqrt[n]{\prod_{k=1}^{n} |z - c_i|} = b$$

telling us that the geometric mean of the distances from z to the vertices of the polygon formed by $c_1, \ldots, c_n$ must be the constant $b > 0$.

Proof. This is something self-explanatory, and as an illustration, let us work out the case of sinusoidal spirals with $n \in \mathbb{N}$. Here with $w = e^{2\pi i/n}$ we have:

$$z^n - c^n = \prod_{k=1}^{n} (z - cw^k)$$

Thus, the sinusoidal spiral equation reformulates as follows:

$$\begin{aligned} |z^n - c^n| = c^n \quad & \iff \quad \prod_{k=1}^{n} |z - cw^k| = c^n \\ & \iff \quad \sqrt[n]{\prod_{k=1}^{n} |z - cw^k|} = c \end{aligned}$$

Thus, for a sinusoidal spiral with positive integer parameter, the geometric mean of the distances to the vertices of a regular polygon must equal the radius of the polygon. □

Regarding now the sinusoidal spirals with $n \in -\mathbb{N}$, these are too part of another remarkable family of plane algebraic curves, constructed as follows:

Theorem 8.18. *Given points in the plane $c_1, \ldots, c_n \in \mathbb{C}$ and a number $d \in \mathbb{R}$, construct the associated stelloid as being the set of points $z \in \mathbb{C}$ verifying*

$$\frac{1}{n} \sum_{k=1}^{n} \alpha_v (z - c_i) = d$$

with α_v denoting the angle with respect to a direction v. Then the stelloid is an algebraic curve, not depending on v, and at the level of examples we have the sinusoidal spirals with $n \in -\mathbb{N}$, including the $n = -1$ line, $n = -2$ hyperbola, and $n = -3$ Humbert cubic.

Proof. All this is quite self-explanatory, and we will leave the verification of the various generalities regarding the stelloids, as well as the verification of the relation with the sinusoidal spirals with $n \in -\mathbb{N}$, as an instructive exercise. As a bonus exercise, try understanding the precise relation between stelloids, and polynomial lemniscates. □

So long for plane algebraic curves. Needless to say, all the above is old-style, first class mathematics, having countless applications. For instance when doing classical mechanics or electrodynamics, you will certainly meet polynomial lemniscates and stelloids, when looking at the field lines. Also, the image of any circle passing though 0 by $z \to z^2$ is a cardioid, and the famous Mandelbrot set is organized around such a cardioid.

8d. Algebraic manifolds

We would like to end the present chapter and Part II with a discussion on what happens in higher dimensions, with an introduction to modern algebraic geometry. As before with other things in this chapter, we will be quite quick, and advanced.

Let us first get to $\mathbb{R}^3$. Here we are right away into a dillema, because the plane curves have two possible generalizations. First we have the algebraic curves in $\mathbb{R}^3$:

Definition 8.19. *An algebraic curve in $\mathbb{R}^3$ is a curve as follows,*

$$C = \left\{(x,y,z) \in \mathbb{R}^3 \Big| P(x,y,z) = 0,\ Q(x,y,z) = 0\right\}$$

appearing as the joint zeroes of two polynomials P, Q.

These curves look of course like the usual plane curves, and at the level of the phenomena that can appear, these are similar to those in the plane, involving singularities and so on, but also knotting, which is a new phenomenon. However, it is hard to say something with bare hands about knots, and we will not get into this, in this book.

On the other hand, as another natural generalization of the plane curves, and this might sound a bit surprising, we have the surfaces in $\mathbb{R}^3$, constructed as follows:

Definition 8.20. *An algebraic surface in $\mathbb{R}^3$ is a surface as follows,*

$$S = \left\{(x,y,z) \in \mathbb{R}^3 \Big| P(x,y,z) = 0\right\}$$

appearing as the zeroes of a polynomial P.

The point indeed is that, as it was the case with the plane curves, what we have here is something defined by a single equation. And with respect to many questions, having a single equation matters a lot, and this is why surfaces in $\mathbb{R}^3$ are "simpler" than curves in $\mathbb{R}^3$. In fact, believe me, they are even the correct generalization of the curves in $\mathbb{R}^2$.

As an example of what can be done with surfaces, which is very similar to what we did with the conics $C \subset \mathbb{R}^2$ before, we have the following result:

Theorem 8.21. *The degree 2 surfaces $S \subset \mathbb{R}^3$, called quadrics, are the ellipsoid*

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

which is the only compact one, plus 16 more, which can be explicitly listed.

Proof. We will be quite brief here, because we intend to rediscuss all this in a moment, with more details, in arbitrary N dimensions, the idea being as follows:

(1) The equations for a quadric $S \subset \mathbb{R}^2$ are best written as follows, with $A \in M_3(\mathbb{R})$ being a matrix, $B \in M_{1\times 3}(\mathbb{R})$ being a row vector, and $C \in \mathbb{R}$ being a constant:

$$< Au, u > +Bu + C = 0$$

(2) By doing now the linear algebra, and we will come back to this in a moment, with details, or by invoking the theorem of Sylvester on quadratic forms, we are left, modulo degeneracy and linear transformations, with signed sums of squares, as follows:

$$\pm x^2 \pm y^2 \pm z^2 = 0, 1$$

(3) Thus the sphere is the only compact quadric, up to linear transformations, and by applying now linear transformations to it, we are led to the ellipsoids in the statement.

(4) As for the other quadrics, there are many of them, a bit similar to the parabolas and hyperbolas in 2 dimensions, and some work here leads to a 16 item list. □

With this done, instead of further insisting on the surfaces $S \subset \mathbb{R}^3$, or getting into their rivals, the curves $C \subset \mathbb{R}^3$, which appear as intersections of such surfaces, $C = S \cap S'$, let us get instead to arbitrary N dimensions, see what the axiomatics looks like there, with the hope that this will clarify our dimensionality dillema, curves vs surfaces.

So, moving to N dimensions, we have here the following definition, to start with:

Definition 8.22. *An algebraic hypersurface in $\mathbb{R}^N$ is a space of the form*

$$S = \left\{(x_1, \ldots, x_N) \in \mathbb{R}^N \Big| P(x_1, \ldots, x_N) = 0, \forall i\right\}$$

appearing as the zeroes of a polynomial $P \in \mathbb{R}[x_1, \ldots, x_N]$.

Again, this is a quite general definition, covering both the plane curves $C \subset \mathbb{R}$ and the surfaces $S \subset \mathbb{R}^2$, which is certainly worth a systematic exploration. But, no hurry with this, for the moment we are here for talking definitons and axiomatics.

In order to have now a full collection of beasts, in all possible dimensions $N \in \mathbb{N}$, and of all possible dimensions $k \in \mathbb{N}$, we must intersect such algebraic hypersurfaces. We are led in this way to the zeroes of families of polynomials, as follows:

Definition 8.23. *An algebraic manifold in $\mathbb{R}^N$ is a space of the form*

$$X = \left\{(x_1, \ldots, x_N) \in \mathbb{R}^N \Big| P_i(x_1, \ldots, x_N) = 0, \forall i\right\}$$

with $P_i \in \mathbb{R}[x_1, \ldots, x_N]$ being a family of polynomials.

As a first observation, as already mentioned, such a manifold appears as an intersection of hypersurfaces S_i, those associated to the various polynomials P_i:

$$X = S_1 \cap \ldots \cap S_k$$

There is actually a bit of a discussion needed here, regarding the parameter $k \in \mathbb{N}$, shall we allow this parameter to be $k = \infty$ too, or not. We will discuss this later.

Let us first look more in detail at the hypersurfaces. We have here:

THEOREM 8.24. *The degree 2 hypersurfaces $S \subset \mathbb{R}^N$, called quadrics, are up to degeneracy and to linear transformations the hypersurfaces of the following form,*

$$\pm x_1^2 \pm \ldots \pm x_N^2 = 0, 1$$

and with the sphere being the only compact one.

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

(1) The equations for a quadric $S \subset \mathbb{R}^N$ are best written as follows, with $A \in M_N(\mathbb{R})$ being a matrix, $B \in M_{1\times N}(\mathbb{R})$ being a row vector, and $C \in \mathbb{R}$ being a constant:

$$< Ax, x > + Bx + C = 0$$

(2) By doing the linear algebra, or by invoking the theorem of Sylvester on quadratic forms, we are left, modulo linear transformations, with signed sums of squares:

$$\pm x_1^2 \pm \ldots \pm x_N^2 = 0, 1$$

(3) To be more precise, with linear algebra, by evenly distributing the terms x_ix_j above and below the diagonal, we can assume that our matrix $A \in M_N(\mathbb{R})$ is symmetric. Thus A must be diagonalizable, and by changing the basis of $\mathbb{R}^N$, as to have it diagonal, our equation becomes as follows, with $D \in M_N(\mathbb{R})$ being now diagonal:

$$< Dx, x > + Ex + F = 0$$

(4) But now, by making squares in the obvious way, which amounts in applying yet another linear transformation to our quadric, the equation takes the following form, with $G \in M_N(-1, 0, 1)$ being diagonal, and with $H \in \{0, 1\}$ being a constant:

$$< Gx, x >= H$$

(5) Now barring the degenerate cases, we can further assume $G \in M_N(-1, 1)$, and we are led in this way to the equation claimed in (2) above, namely:

$$\pm x_1^2 \pm \ldots \pm x_N^2 = 0, 1$$

(6) In particular we see that, up to some degenerate cases, namely emptyset and point, the only compact quadric, up to linear transformations, is the one given by:

$$x_1^2 + \ldots + x_N^2 = 1$$

(7) But this is the unit sphere, so are led to the conclusions in the statement. □

Getting now to the case of arbitrary general manifolds, as in Definition 8.23, the key to their study is abstract algebra, guided by the following fundamental question:

Question 8.25. *Given an algebraic manifold in $\mathbb{R}^N$, appearing as*

$$X=\left\{(x_1,\ldots,x_N)\in\mathbb{R}^N\Big|P_i(x_1,\ldots,x_N)=0,\forall i\right\}$$

what are the polynomials $P\in\mathbb{R}[x_1,\ldots,x_N]$ vanishing on X? Conversely, given a set

$$I\subset\mathbb{R}[x_1,\ldots,x_N]$$

what is the manifold X where all the polynomials $P\in I$ vanish?

Obviously, this is something important, because assuming that we managed to find an answer, we will have a useful "algebraic geometry" correspondence, as follows:

$$\left(X\subset\mathbb{R}^N\right)\quad\longleftrightarrow\quad\left(I\subset\mathbb{R}[x_1,\ldots,x_N]\right)$$

In practice now, we already know a bit about the beasts on the left X, so let us study the beasts on the right I. Here are a few basic observations, about them:

(1) To start with, assuming that $X\subset\mathbb{R}^N$ comes from polynomials $\{P_i\}$, the set $I\subset\mathbb{R}[x_1,\ldots,x_N]$ of polynomials vanishing on X obviously contains $\{P_i\}$.

(2) However, much more is true. Indeed, if we come with any family of polynomials $\{Q_i\}\subset\mathbb{R}[x_1,\ldots,x_N]$, it is then clear that we must have $\sum_i P_iQ_i\in I$.

(3) Getting now a bit abstract, we can see that, more generally, $I\subset\mathbb{R}[x_1,\ldots,x_N]$ must be stable under sums, and must satisfy $P\in I\implies PQ\in I$, $\forall Q$.

And so, question now, in view of all this, what are the beasts $I\subset\mathbb{R}[x_1,\ldots,x_N]$ that we are looking for? In answer, these must be ideals, in the following sense:

Definition 8.26. *We have notions of rings, modules and ideals, as follows:*

1. *A ring R is a set with operations $+$ and $\times$, satisfying the usual conditions for such operations, except for $ab=ba$, and for $a\neq 0\implies\exists a^{-1}$.*
2. *A module V over a ring R is a vector space, but we will call it ring, and keep the name vector spaces for the modules over fields, $R=F$.*
3. *An ideal $I\subset R$ is a subgroup with the left ideal property $i\in I,r\in R\implies ir\in I$, or the right ideal property $i\in I,r\in R\implies ri\in I$, or both.*

In what follows we will be mainly interested in the ring $R=\mathbb{R}[x_1,\ldots,x_N]$, which is commutative, $ab=ba$. For such rings, the 3 notions of ideals in (3) coincide.

In relation now with our algebraic geometry questions, we can reformulate our notion of algebraic manifold, in commutative algebra terms, as follows:

Theorem 8.27. *The algebraic manifolds are precisely the sets of the form*

$$X=\left\{x\in\mathbb{R}^N\Big|P(x)=0,\forall P\in I\right\}$$

with $I\subset\mathbb{R}[x_1,\ldots,x_N]$ being a certain ideal.

Proof. In one sense, this comes from the discussion after Question 8.25, and in the other sense this is trivial, because we can write $I=\{P_i|i\in I\}$, with $P_i=i$. □

In order to further discuss now the correspondence $X\leftrightarrow I$, we need to know more algebra. Let us start with the following basic fact, in the context of Definition 8.26:

Theorem 8.28. *Let R be a ring, and $I\subset R$ be an additive subgroup.*

(1) *I is a two-sided ideal precisely when $F=R/I$ is a ring.*
(2) *If R is commutative, $I\subset R$ is a maximal ideal precisely when F is a field.*

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

(1) Since the additive group $(R,+)$ is abelian, given an additive subgroup $I\subset R$ we can form the quotient group $F=R/I$, which is abelian too, with addition as follows:

$$(a+I)+(b+I)=(a+b+I)$$

The question is now, can we turn this abelian group F into a ring? Normally the multiplication can only be as follows, and with this clarifying our statement:

$$(a+I)(b+I)=(ab+I)$$

But, will this work. In practice, the following condition must be satisfied:

$$(a+I)=(a'+I)\ ,\ (b+I)=(b'+I)\implies(ab+I)=(a'b'+I)$$

But this amounts in the following condition to be satisfied:

$$a-a'\in I\ ,\ b-b'\in I\implies ab-a'b'\in I$$

Now comes the math. We have the following identity, which shows that if $I\subset R$ is a two-sided ideal, then the above condition is satisfied, and so done:

$$ab-a'b'=a(b-b')+(a-a')b'$$

Conversely, if the above condition is satisfied, we have in particular:

$$i-0\in I\ ,\ r-r\in I\implies ir-0r\in I$$

$$r-r\in I\ ,\ i-0\in I\implies ri-r0\in I$$

Thus $I\subset R$ must be a two-sided ideal, and this finishes the proof of (1).

(2) Assume first that $F=R/I$ is a field. This means that any nonzero element of F is invertible, and with our usual conventions for F, this reads:

$$\forall a\notin I\ ,\ \exists b\in R\ ,\ (ab+I)=(1+I)$$

Now assume by contradiction that $I\subset R$ is not maximal, so that we have a bigger ideal $I\subset J\subset R$. If we pick $a\in J-I$, we obtain, by the above, the following:

$$a\in J-I\ ,\ b\in R\ ,\ ab=1+i\ ,\ i\in I$$

But this is contradictory, because since J is an ideal, containing I, we must have $ab,i\in J$, and so we conclude that we have $1\in J$, which in turn gives:

$$J=R$$

Conversely, assume now that I is maximal, and assume too, by contradiction, that $F=R/I$ is not a field. Then we can find a zero divisor in F, which reads:

$$(a+I)(b+I)=(I)\ ,\ a,b\notin I$$

In other words, we can find $ab\in I$ with $a,b\notin I$. But then, let us look at:

$$I\subset I+aR\subset R$$

What we have in the middle is an ideal, and it is also clear, from $a\notin I$, that the inclusion on the left is proper. As for the inclusion on the right, our claim is that this is proper too. Indeed, assuming otherwise, we would have a formula as follows:

$$i+ac=1\ ,\ i\in I$$

Now by multiplying everything by b, we obtain from this:

$$ib+acb=b\ ,\ i\in I$$

But this is contradictory, because on the left we have $ib\in I$ and $acb=(ab)c\in I$, which gives $b\in I$, contradicting $b\notin I$. Thus, claim proved, which gives the result. □

Getting back now to algebraic geometry, we first have the following result:

Theorem 8.29 (Hilbert basis theorem). *Any ideal of polynomials*

$$I\subset\mathbb{R}[x_1,\ldots,x_N]$$

is finitely generated, $I=(P_1,\ldots,P_k)$, for some $P_i\in\mathbb{R}[x_1,\ldots,x_N]$.

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

(1) Following Emmy Noether, let us call a ring R Noetherian when any ideal $I\subset R$ is finitely generated. Equivalently, any increasing sequence of ideals $I_1\subset I_2\subset\ldots$ must stabilize, in the sense that we must have $I_n=I_{n+1}=\ldots$, for some $n\in\mathbb{N}$.

(2) We want to prove that $\mathbb{R}[x_1,\ldots,x_N]$ is Noetherian, and we will do this by recurrence on N. Since $R=\mathbb{R}$ is clearly Noetherian, as being a field, we are left with proving the recurrence step. And, for this purpose, we will prove something which is a bit more general, namely that if a ring R is Noetherian, then so is the ring $R[X]$.

(3) We do this by contradiction. So, assume that R is Noetherian, and that $R[X]$ is not Noetherian, so that we have an ideal $I\subset R[X]$ which is not finitely generated.

(4) In order to find a contradiction, let us pick $P_1 \in I$ of minimial degree $d_1 \in \mathbb{N}$, then $P_2 \in I/(P_1)$ of minimal degree $d_2 \in \mathbb{N}$, then $P_3 \in I/(P_1, P_2)$ of minimal degree $d_3 \in \mathbb{N}$, and so on. Since our ideal $I \subset R[X]$ was assumed to be not finitely generated, this procedure will not stop, and we obtain an increasing sequence, as follows:

$$d_1 \leq d_2 \leq d_3 \leq \ldots$$

(5) Now let $a_i \in R$ be the leading coefficient of each P_i, and set $J = (a_1, a_2, \ldots) \subset R$. Since R was assumed to be Noetherian, we can find $n \in \mathbb{N}$ such that $J = (a_1, \ldots, a_n)$. Thus, we have a formula as follows, for certain scalars $\lambda_i \in R$:

$$a_{n+1} = \sum_{i=1}^{n} \lambda_i a_i$$

(6) With this done, consider the following polynomial, with $\lambda_i \in R$ as above:

$$Q = \sum_{i=1}^{n} \lambda_i X^{d_{n+1} - d_i} P_i$$

This polyomial satisfies then $Q \in (P_1, \ldots, P_n)$, and has the same leading coefficient as $P_{n+1} \notin (P_1, \ldots, P_n)$. Thus, the following polynomial has degree $< d_{n+1}$:

$$P_{n+1} - Q \in I/(P_1, \ldots, P_n)$$

But this is a contradiction, as desired, and this finishes the proof. □

In practice, Theorem 8.29 is best remembered geometrically, as follows:

Theorem 8.30. *The algebraic manifolds $X \subset \mathbb{R}^N$ are precisely the intersections*

$$X = S_1 \cap \ldots \cap S_k$$

with $S_i \subset \mathbb{R}^N$ being hypersurfaces.

Proof. Indeed, given an algebraic manifold $X \subset \mathbb{R}^N$, we can consider the ideal $I \subset \mathbb{R}[x_1, \ldots, x_N]$ of polynomials vanishing on X, then write $I = (P_1, \ldots, P_k)$ with $k < \infty$, as in Theorem 8.29, and then set $S_i \subset \mathbb{R}^N$ to be the set of zeroes of P_i. □

Moving ahead now, let us further investigate the correspondence $X \leftrightarrow I$. We would like this to be bijective, but there are at least 2 obstructions to this, as follows:

(1) To start with, assuming $P^k = 0$ on X, we have $P = 0$ on X. In view of this, we must restrict the attention to the ideals I which are "radical", $P^k \in I \implies P \in I$.

(2) Also, at $N = 1$, the ideal $I = (x^2 + 1) \subset \mathbb{R}[x]$ produces the manifold $X = \emptyset$. In view of this, we must trade $\mathbb{R}$ for $\mathbb{C}$, where arbitrary polynomials have roots.

So, these are two obvious obstructions, with respective solutions to them, and coming now as good news, there is no third obstruction, as shown by the following result:

Theorem 8.31 (Nullstellensatz). *We have a correspondence*

$$\Big(X\subset\mathbb C^N\Big)\quad\longleftrightarrow\quad\Big(I\subset\mathbb C[x_1,\ldots,x_N]\Big)$$

between algebraic manifolds in $\mathbb C^N$, and radical ideals of $\mathbb C[x_1,\ldots,x_N]$.

Proof. This is something quite tricky, due to Hilbert, as follows:

(1) To start with, we have traded $\mathbb R$ for $\mathbb C$, but this will not affect much what we know, notably with Theorem 8.29 still holding in this setting, with the same proof.

(2) We know that at $N=1$ polynomials have roots, so here $I=(P)\implies X_I\neq\emptyset$. The point now is that, by doing some algebra, in the spirit of what we did in the proof of Theorem 8.29, something similar happens in arbitrary N dimensions, in the sense that any proper ideal $I\subset\mathbb C[x_1,\ldots,x_N]$ produces a non-empty manifold, $X_I\neq\emptyset$.

(3) Next, what we want to prove is that given an ideal $I\subset\mathbb C[x_1,\ldots,x_N]$, any polynomial $P\in\mathbb C[x_1,\ldots,x_N]$ vanishing on X_I has the property $P^k\in I$, for some $k\in\mathbb N$. For this purpose, we can add 1 dimension, and consider the following ideal:

$$J=<I,x_{N+1}P(x_1,\ldots,x_N)-1>$$

(4) Now since we have $X_J=\emptyset$, by (2) we conclude that J is trivial. In order now to best interpret this finding, consider the following algebra:

$$\mathbb C[x_1,\ldots,x_N][P^{-1}]=\mathbb C[x_1,\ldots,x_{N+1}]/(x_{N+1}P-1)$$

The triviality of J gives then a formula of the following type, with $f_i\in I$:

$$1=f_0+f_1x_{N+1}+\ldots+f_kx_{N+1}^k$$

Now by multiplying by P^k, we obtain from this $P^k\in I$, as desired. □

8e. Exercises

This was a particularly pleasant chapter, and as exercises on this, we have:

Exercise 8.32. *Fill in all the details for the conics appearing via cone cuts.*

Exercise 8.33. *Learn more about the focal points of ellipses, and other conics.*

Exercise 8.34. *Learn more about gravity, and Kepler and Newton.*

Exercise 8.35. *Work out equations for the conics, in polar coordinates.*

Exercise 8.36. *Learn more about quintics, and about Galois theory too.*

Exercise 8.37. *Learn more about sinusoidal spirals, and their properties.*

Exercise 8.38. *Learm as well more about polynomial lemniscates, and stelloids.*

Exercise 8.39. *Learn more about the quadrics, and their classification.*

As bonus exercise, and no surprise here, start reading some algebraic geometry.

Part III

Functions

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

Polynomials

9a. Polynomials, roots

Welcome to functions, which are the topic of the present Part III of this book. As a goal, we would like to understand the functions $f:\mathbb R\to\mathbb R$, or perhaps $f:X\to\mathbb R$ with $X\subset\mathbb R$ being a suitable subset, and with this meaning things like solving $f(x)=0$, or, importantly in relation with applications, finding the minimum or maximum of f.

There are many things to be learned here, and we will go slowly. The simplest possible functions, that we would like to investigate in this chapter, are the polynomial ones:

$$P\in\mathbb R[X]\quad\rightsquigarrow\quad P:\mathbb R\to\mathbb R$$

Getting started, let us first have a look at the simplest polynomials that we know, namely the degree 2 ones. You certainly know that these are suitably represented by their graphs, which are parabolas, and with these parabolas being drawn as follows:

Method 9.1. *In order to draw the graph of $P(x)=ax^2+bx+c$:*

(1) *We must first compute the discriminant, $\Delta=b^2-4ac$.*
(2) *Which leads to 4 cases, depending on whether a,Δ are positive or not.*
(3) *And so to 4 cases, regarding the position and orientation of the parabola.*
(4) *Next, we must compute $x=-b/2a$, where the symmetry axis is.*
(5) *So, we must first draw $(x,P(x))$, and then the parabola, according to (3),*
(6) *With the zeroes $(z,0)$ with $z=(-b\pm\sqrt{\Delta})/2a$ represented too, when $\Delta\geq 0$.*

Which sounds quite simple, but in practice, in what regards the computation of the zeroes, which is the hardest part, there is a trick that you must know, as follows:

Theorem 9.2. *The roots r,s of a degree 2 equation, written as*

$$x^2-ax+b=0$$

can be computed by using $r+s=a$, $rs=b$.

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

(1) To start with, given an arbitrary equation $Ax^2+Bx+C=0$, we can always divide by A, then switch the sign of B, as to reach to the above form, $x^2-ax+b=0$.

(2) Next, let us look for the roots r,s. These must satisfy the following equations:

$$r^2-ar+b=0$$

$$s^2-as+b=0$$

By making the difference and the sum, these equations are equivalent to:

$$(r-s)(r+s)-a(r-s)=0$$

$$(r+s)^2-2rs-a(r+s)+2b=0$$

But, assuming that the roots are distinct, $r\neq s$, the first equation gives $r+s=a$, and with this in hand, the second equation becomes $rs=b$, as desired.

(3) Thus, result proved, modulo a discussion regarding the case $r=s$. But this case appears when $\Delta=a^2-4b$ vanishes, and with the common root here being $r=s=a/2$, and this fits with our equations, which are in this case $r+s=a$, $rs=a^2/4$.

(4) As a comment now, the above manipulations might seem quite wizarding, but we have as well the following more advanced proof, to be fully justified later:

$$\begin{aligned}x^2-ax+b=(x-r)(x-s)\quad&\iff\quad x^2-ax+b=x^2-(r+s)x+rs\\&\iff\quad r+s=a,\ rs=b\end{aligned}$$

To be more precise, as we will see later in this chapter, the roots r,s must be given by the condition on the left, and so the above equivalences give the result. □

Here is an illustration for this. With the help of the general formula, we find:

$$x^2-8x+15=0\iff x=\frac{8\pm\sqrt{64-60}}{2}=\frac{8\pm 2}{2}=3,5$$

With our trick, however, the computation is almost instant, as follows:

$$x^2-8x+15=0\iff r+s=8,\ rs=15\iff r,s=3,5$$

Which is not bad, hope you agree with me here. Finally, for this discussion to be complete, let us mention too the following important fact:

Warning 9.3. *The above trick works in pure mathematics, where the numbers r,s that we 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 $P=0$.

In degree 3 now, things get more complicated. As a first observation, we have:

Fact 9.4. *Any degree* 3 *polynomial, say taken with leading coefficient* 1,

$$P = x^3 + ax^2 + bx + c$$

must have at least one root, on the grounds that P must travel as follows:

$$P(-\infty) = -\infty \qquad \rightsquigarrow \qquad P(\infty) = \infty$$

Moreover, the same argument applies to any $P \in \mathbb{R}[X]$ of odd degree.

In order to exploit this fact, we need to know more about polynomials, and their roots. As a starting point here, we have Theorem 9.2, telling us that the roots r, s of a degree 2 polynomial $x^2 - ax + b$ can be computed by using the following formulae:

$$r + s = a \quad , \quad rs = b$$

Moreover, as explained in the proof of Theorem 9.2, these formulae come in fact from the following remarkable identity, assuming as above that r, s are the roots:

$$x^2 - ax + b = (x - r)(x - s)$$

In order to discuss such things for arbitrary polynomials, let us start with:

Proposition 9.5. *For a polynomial $P \in \mathbb{R}[X]$ and a number $r \in \mathbb{R}$, the following conditions are equivalent:*

1. *$P(r) = 0$.*
2. *$P(x) = (x - r)Q$, with $Q \in \mathbb{R}[X]$.*

Proof. The point here is that we can divide the polynomials, a bit as we divide the integers, and by dividing P by $x - r$ we are led to a formula as follows, with the quotient being a certain polynomial $Q \in \mathbb{R}[X]$, and the remainder being a constant $c \in \mathbb{R}$:

$$P(x) = (x - r)Q + c$$

But with this, the equivalence in the statement is clear, by taking $x = r$. □

Now by applying this iteratively, we are led to the following key result:

Theorem 9.6. *Any polynomial $P \in \mathbb{R}[X]$ can be written as*

$$P(x) = (x - r_1)^{n_1} \ldots (x - r_k)^{n_k} Q$$

with $r_1, \ldots, r_k \in \mathbb{R}$ being the roots, $n_1, \ldots, n_k \in \mathbb{N}$, and $Q \in \mathbb{R}[X]$ having no roots.

Proof. This follows indeed by applying Proposition 9.5 iteratively, with the term $\prod_i (x - r_i)^{n_i}$ growing over the time, until it has to stop, due to the fact that the remainder $Q \in \mathbb{R}[X]$ becomes a constant, or more generally, a polynomial having no roots. □

As an illustration here, consider a degree 3 polynomial, chosen for simplifying to be of the following special form, with leading coefficient 1:

$$P = x^3 + ax^2 + bx + c$$

Since we have $P(-\infty) = -\infty$ and $P(\infty) = \infty$, we have at least one root $r \in \mathbb{R}$. Thus $P = (x-r)Q$ with Q being of degree 2, which leads to the following 3 possible situations, with r, s, t being distinct real numbers, and Q being of degree 2, having no roots:

$$P = (x-r)(x-s)(x-t)$$

$$P = (x-r)^2(x-s)$$

$$P = (x-r)Q$$

As a useful complement now to Theorem 9.6, which generalizes and further clarifies what we said in Theorem 9.2 and its proof, in degree 2, we have:

Theorem 9.7. *Given a polynomial $P \in \mathbb{R}[X]$, with leading coefficient* 1*,*

$$P(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_1x + a_0$$

assuming that P has the maximum of n roots, when counted with multiplicities, so that

$$P(x) = (x-r_1)\ldots(x-r_n)$$

these roots, taken with multiplicities, satisfy $\sum_i r_i = -a_{n-1}$ and $\prod_i r_i = (-1)^n a_0$.

Proof. This is clear indeed from the formula $P(x) = (x-r_1)\ldots(x-r_n)$, by expanding the product, and identifying the terms of degree $n-1$, and of degree 0. □

As yet another basic thing about roots, that you should know, we have:

Theorem 9.8. *Given a polynomial $P \in \mathbb{Z}[X]$, with leading coefficient* 1*,*

$$P(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_1x + a_0$$

any integer root $r \in \mathbb{Z}$ must satisfy $r|a_0$. A similar result holds for $P \in \mathbb{Q}[X]$.

Proof. This is clear indeed from $P(r) = r^n + a_{n-1}r^{n-1} + \ldots + a_1r + a_0$, because assuming $P(r) = 0$, this formula can be written in the following way:

$$r(r^{n-1} + a_{n-1}r^{n-2} + \ldots + a_1) = -a_0$$

Thus we have $r|a_0$, as claimed. As for the extension to the case $P \in \mathbb{Q}[X]$, this is something straightforward, that we will leave here as an instructive exercise. □

Many other things can be said, along these lines, notably with the Eisenstein criterion, making use of a prime number p, which is something very useful. Exercise for you, to learn more about all this, various algebraic tricks for dealing with polynomials.

Getting now to heavier tools, we can differentiate the polynomials, as follows:

Theorem 9.9. *We can formally differentiate the polynomials, according to*

$$(x^n)'=nx^{n-1}$$

and to the following linearity rules, allowing to pass to linear combinations:

$$(P+Q)'=P'+Q'\quad,\quad(\lambda P)'=\lambda P'$$

This differentiation operation satisfies the following rules,

$$(PQ)'=P'Q+PQ'\quad,\quad(P\circ Q)'=P'(Q)Q'$$

called Leibnitz rule for products, and chain derivative rule.

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

(1) To start with, we can certainly differentiate the polynomials according to the recipe in the statement, with the precise general formula being as follows:

$$\begin{aligned}&P=a_nx^n+a_{n-1}x^{n-1}+\ldots+a_1x+a_0\\ \implies\ &P'=na_nx^{n-1}+(n-1)a_{n-1}x^{n-2}+\ldots+a_1\end{aligned}$$

(2) In what regards the Leibnitz rule, by linearity we can assume that we are dealing with monomials, $P=x^m$ and $Q=x^n$. But here, the Leibnitz rule comes from:

$$\begin{aligned}(x^{m+n})'&=(m+n)x^{m+n-1}\\&=mx^{m-1}x^n+nx^mx^{n-1}\\&=(x^m)'x^n+x^m(x^n)'\end{aligned}$$

(3) Finally, in what regards the chain rule, again by linearity we can assume that we have $P=x^m$ and $Q=x^n$. And here, the result comes via the following computation:

$$\begin{aligned}(x^{mn})'&=mnx^{mn-1}\\&=mx^{mn-n}\cdot nx^{n-1}\\&=[(mx^{m-1})\circ x^n]\cdot nx^{n-1}\\&=[(x^m)'\circ x^n]\cdot(x^n)'\end{aligned}$$

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

As a comment here, you might wonder what the quantity $P'(x)\in\mathbb{R}$ exactly stands for, intuitively speaking, when a particular $x\in\mathbb{R}$ is given. Good question, and in answer, the derivatives as introduced in Theorem 9.9 are some sort of black magic, used since long by mathematicians, and it took mankind a long time, culminating with the work of Netwon, in order to truly understand what is really going on, with all this.

In short, trust me here, with Theorem 9.9, which sounds a bit medieval, we are on the good way towards modernity. And more about modernity later, no worries for that.

As an application now of our derivatives, as introduced above, we have:

THEOREM 9.10. *Given a polynomial $P \in \mathbb{R}[X]$, the following happen:*

(1) *Any multiple root of P must be a root of P'.*
(2) *In fact, the multiple roots of P are the common roots of P, P'.*
(3) *If $P(r) = 0$ with multiplicity k, then $P'(r) = 0$ with multiplicity $k - 1$.*

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

(1) We have indeed the following computation, based on the general differentiation rules from Theorem 9.9, and more specifially, on the Leibnitz rule there:

$$\begin{aligned}[(x-r)^2Q]' &= [(x-r)^2]'Q + (x-r)^2Q' \\ &= 2(x-r)Q + (x-r)^2Q' \\ &= (x-r)(2Q + (x-r)Q')\end{aligned}$$

Here we have used the following formula, which is something trivial:

$$[(x-r)^2]' = 2(x-r)$$

But with $P = (x-r)^2Q$, this leads to the conclusion in the statement.

(2) We know this in one sense from (1). In the other sense, assume that:

$$P(r) = P'(r) = 0$$

Now let us divide P by $(x-r)^2$. This must give a formula as follows:

$$P = (x-r)^2Q + c(x-r)$$

By using now the computation in (1), we can see that $P'(r) = 0$ amounts in saying that $(c(x-r))'$ vanishes at r, so that $c = 0$. Thus, $P = (x-r)^2Q$, as desired.

(3) We have indeed the following computation, generalizing the one in (1):

$$\begin{aligned}[(x-r)^kQ]' &= [(x-r)^k]'Q + (x-r)^kQ' \\ &= k(x-r)^{k-1}Q + (x-r)^kQ' \\ &= (x-r)^{k-1}(kQ + (x-r)Q')\end{aligned}$$

Here we have used the following formula, coming from the chain rule:

$$[(x-r)^k]' = k(x-r)^{k-1}$$

Thus, with $P = (x-r)^kQ$, we are led to the conclusion in the statement. □

The above result is something quite amazing, raising the possibility of deciding if P has multiple roots, without computing the roots in question. Indeed, for this purpose we can simply compute P', and then successively perform the division algorithm for P, P', a bit like for the usual integers, as to compute the greatest common divisor $D = (P, P')$. And then, if D has degree ≥ 1, our original polynomial P must have a double root.

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

Theorem 9.11. *Given a polynomial $P \in \mathbb{R}[X]$, compute P', and perform the division algorithm for P, P', as to get to the greatest common divisor $D = (P, P')$.*

1. *P has multiple roots precisely when $\deg D \geq 1$.*
2. *In fact, the multiple roots of P are precisely the roots of D.*
3. *Moreover, via $P \to D$, all root multiplicities get lowered by 1.*

Proof. This follows indeed as indicated above. To be more precise, assume that P factorizes as follows, with r_i being its multiple roots, with multiplicities $n_i \geq 2$:

$$P(x) = (x - r_1)^{n_1} \ldots (x - r_k)^{n_k} Q$$

According to Theorem 9.10, the polynomial P' is then of the following form:

$$P'(x) = (x - r_1)^{n_1 - 1} \ldots (x - r_k)^{n_k - 1} R$$

Thus, the common divisor $D = (P, P')$ is given by the following formula:

$$D(x) = (x - r_1)^{n_1 - 1} \ldots (x - r_k)^{n_k - 1}$$

But this leads to the various conclusions in the statement. □

9b. The resultant

Time now to get more systematically into the mathematics of polynomials of small degree. To start with, it is convenient to upgrade our formalism, and work over $\mathbb{C}$, where all polynomials have roots. Indeed, let us recall from chapter 7 that we have:

Theorem 9.12. *Any polynomial $P \in \mathbb{C}[X]$ decomposes as*

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

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

Proof. This is something from chapter 7, the idea being as follows:

(1) 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:

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

Here we have assumed that this minimum is attained, and with this happening indeed, coming from the continuity of P. More on this in chapter 10, when doing analysis.

(2) Now 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:

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

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

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

(4) 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. $\square$

Getting now to small degree polynomials, let us start with something that we know well, but is always good to remember, such a pleasure to do this computation again:

Theorem 9.13. *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:

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

Thus, we are led to the conclusion in the statement. $\square$

Moving now to degree 3 and higher, things here are 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 9.14. *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) In order to understand what is going on, let us begin with some examples, in small degree. Consider first a monic polynomial of degree 2, factorized as follows:

$$x^2 + ax + b = (x - r)(x - s)$$

We have then the following equations, showing that the first assertion holds:

$$r+s=-a\quad,\quad rs=b$$

Regarding now the second assertion, there are many symmetric functions in r,s, and all can be expressed as polynomials in a,b, as shown by the following computations:

$$r^2+s^2=(r+s)^2-2rs=a^2-2b$$

$$r^2s+rs^2=(r+s)rs=-ab$$

$$r^3+s^3=(r+s)^3-3(r+s)rs=-a^3+3ab$$

$$r^2s^2=(rs)^2=b^2$$

$$r^3s+rs^3=(r^2+s^2)rs=a^2b-2b^2$$

$$r^4+s^4=(r+s)(r^3+s^3)-(r^3s+rs^3)=a^4-4a^2b+2b^2$$

$$\vdots$$

(2) Let us see as well what happens in degree 3. Consider a polynomial as follows:

$$x^3+ax^2+bx+c=(x-r)(x-s)(x-t)$$

We have then the following equations, showing that the first assertion holds:

$$r+s+t=-a\quad,\quad rs+rt+st=b\quad,\quad rst=-c$$

As for the second assertion, there are many symmetric functions in r,s,t, and again these are polynomials in a,b,c. Indeed, we first have the following computations:

$$r^2+s^2+t^2=(r+s+t)^2-2(rs+rt+st)=a^2-2b$$

$$r^2s+rs^2+r^2t+rt^2+s^2t+st^2=(r+s+t)(rs+rt+st)-3rst=-ab+3c$$

Next, for the missing degree 3 symmetric function in r,s,t, which is the sum of cubes, we have the following computation, based on our various formulae above:

$$\begin{aligned}&r^3+s^3+t^3\\=\;&(r+s+t)(r^2+s^2+t^2)-(r^2s+rs^2+r^2t+rt^2+s^2t+st^2)\\=\;&-a(a^2-2b)-(-ab+3c)\\=\;&-a^3+3ab-3c\end{aligned}$$

Getting started now with degree 4, we first have here the following function:

$$\begin{aligned}r^2s^2+r^2t^2+s^2t^2&=(rs+rt+st)^2-2rst(r+s+t)\\&=b^2-2ac\end{aligned}$$

And so on, you get the point, the idea being that, a bit as for the polynomials of degree 2, the second assertion holds as well for polynomials of degree 3.

(3) In general now, 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.

(4) 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:

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

(5) 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 (3), 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.

(6) So, 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 9.15. *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 let us work out some illustrations, both for the computation of $R(P,Q)$, and for what happens when $R(P,Q)=0$:

EXAMPLE 9.16. *Degree 2 and degree 1.*

Consider a polynomial of degree 2, and a polynomial of degree 1:

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

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

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

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

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

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

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

Thus, Theorem 9.15 holds indeed in this case, degree 2 and degree 1.

Example 9.17. *Degree* 3 *and degree* 1.

Consider a polynomial of degree 3, and a polynomial of degree 1:

$$P=ax^3+bx^2+cx+d\quad,\quad Q=ex+f$$

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

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

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

$$\begin{aligned}R(P,Q) &= ae^3(p-s)(q-s)(r-s)\\ &= ae^3(pqr-(pq+pr+qr)s+(p+q+r)s^2-s^3)\\ &= -de^3-ce^3s-be^3s^2-ae^3s^3\\ &= -de^3+ce^2f-bef^2+af^3\end{aligned}$$

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 $s=-f/e$, and we have:

$$P(s)=-\frac{af^3}{e^3}+\frac{bf^2}{e^2}-\frac{cf}{e}+d=-\frac{R(P,Q)}{e^3}$$

Thus, Theorem 9.15 holds indeed in this case, degree 3 and degree 1.

Example 9.18. *Degree* 2 *and degree* 2.

Consider indeed two polynomials of degree 2, written as follows:

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

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

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

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

$$\begin{aligned}
R(P,Q) &= a^2d^2(p-r)(q-r)(p-s)(q-s)\\
&= a^2d^2(pq-(p+q)r+r^2)(pq-(p+q)s+s^2)\\
&= d^2(c+br+ar^2)(c+bs+as^2)\\
&= d^2(c^2+bc(r+s)+ac(r^2+s^2)+b^2rs+ab(r^2s+s^2r)+a^2r^2s^2)\\
&= c^2d^2-bcde+ace^2-2acdf+b^2df-abef+a^2f^2
\end{aligned}$$

As for the fact that $R(P,Q)=0$ corresponds indeed to the fact that P,Q have a common root, I will leave some verifications here to you, as an instructive exercise.

9c. The discriminant

We can go back now to our original question, finding the analogue of the discriminant $\Delta = b^2-4ac$ for higher degree polynomials, with the following result about this:

THEOREM 9.19. *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. This comes from the various results that we have, as follows:

(1) The fact that the discriminant $\Delta(P)$ constructed above is a polynomial in the coefficients of P, with integer coefficients, comes from Theorem 9.15, coupled with the fact that the division by the leading coefficient a is indeed possible, under $\mathbb{Z}$.

(2) Also, the fact that we have $\Delta(P)=0$ precisely when P has a double root is clear from Theorem 9.15. 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 an illustration, let us see what happens in degree 2. Here we have:

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

Thus, the resultant is given by the following formula:

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

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

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

At the theoretical level now, we have the following result, which is not trivial:

THEOREM 9.20. *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 9.15, 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 9.15, 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:

$$\begin{aligned} R(P, P') &= N^N a^{2N-1} \prod_i \left(\prod_j (r_i - p_j) \right) \\ &= N^N a^{2N-1} \prod_i \frac{P'(r_i)}{Na} \\ &= a^{N-1} \prod_i P'(r_i) \end{aligned}$$

(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 9.19, 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 9.20 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:

THEOREM 9.21. *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 9.13, but let us see how this comes via the formula in Theorem 9.20, 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 9.13, 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 9.20, 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.

9d. Cardano formulae

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

Theorem 9.22. *The discriminant of a degree 3 polynomial,*

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

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

Proof. We can use the same method as in Examples 9.16, 9.17 and 9.18. 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, as desired. □

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:

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

And, regarding such equations, we have the following famous result of Cardano:

Theorem 9.23. *For a normalized degree 3 equation, namely*

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

the discriminant is $\Delta=-108(p^3+q^2)$. Assuming $p,q\in\mathbb{R}$ and $\Delta<0$, the 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. The formula of Δ comes from Theorem 9.22, 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:

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

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

As a comment here, the formula for roots in Theorem 9.23 holds in the case $\Delta>0$ too, and also when the coefficients are complex numbers, $p,q\in\mathbb{C}$. However, these extensions are usually not very useful, because when it comes to extract the above square and cubic roots, you can well end up with the initial question, the one that you started with.

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 9.24. *The discriminant of $P = ax^4 + bx^3 + cx^2 + dx + e$ is given by:*

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

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

Proof. The formula of Δ follows from a routine computation, say exercise for you, and we will be back to this in chapter 14, with a more clever method for dealing with such questions. 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 some manipulations on our polynomials, as to have them in simpler form, and we have the following version of Theorem 9.24:

Theorem 9.25. *The discriminant of a normalized degree 4 polynomial,*

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

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 indeed from the general formula in Theorem 9.24. □

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

Theorem 9.26. *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 9.26, is as follows:

Theorem 9.27 (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:

Fact 9.28. *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 $P=X^5-X-1$.*

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 abstract algebra book.

9e. Exercises

We have seen in this chapter that polynomials are a non-trivial business. As exercises on this, polynomials, and chosen of course non-trivial, we have:

Exercise 9.29. *Clarify what we said, in relation with symmetric functions.*

Exercise 9.30. *Compute some further resultants, for low degree polynomials.*

Exercise 9.31. *Clarify all details in relation with the meaning of Δ, in degree 3.*

Exercise 9.32. *Clarify also details in relation with the meaning of Δ, in degree 4.*

Exercise 9.33. *Establish the explicit formula of Δ given above, in degree 4.*

Exercise 9.34. *Read more about Cardano, including trigonometric aspects.*

Exercise 9.35. *Read as well about various trigonometric aspects, in degree 4.*

Exercise 9.36. *Learn various algebraic and analytic tricks, in degree 5 and higher.*

As bonus exercise, which might however take some time, read some Galois theory, with all the needed algebra preliminaries. That would be an excellent investment.

CHAPTER 10

Functions

10a. Functions, continuity

Welcome to functions, take two. We have tried in chapter 9 to get introduced to them, via polynomials $P \in \mathbb{R}[X]$, which are the simplest functions, no question about this. However, precisely because the polynomials $P \in \mathbb{R}[X]$ are something so simple and classical, we ended up doing some complicated mathematics with them. So, time to change our approach, radically, by looking instead at the arbitrary functions $f : \mathbb{R} \to \mathbb{R}$, with the hope of course that this can lead to some interesting mathematics.

In short, let us forget everything that we know, polynomials and other mathematics, and start our study abstractly, with the following definition:

Definition 10.1. *A real function is a correspondence as follows:*

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

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

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

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

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

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

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

Definition 10.2 (update). *More generally, we call function any correspondence*

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

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

In practice, however, this will not change much to what we already had, from Definition 10.1. Indeed, any function $f : X \to Y$ with $Y \subset \mathbb{R}$ can be regarded as a function $f : X \to \mathbb{R}$ in the obvious way, by composing it with the inclusion $Y \subset \mathbb{R}$, as follows:

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

However, Definition 10.2 can be something useful, in relation with the notions of injectivity, or surjectivity. Consider for instance the usual square function:

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

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

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

Which is good, but this latter function is still not surjective. However, we can make it surjective, by using the framework of Definition 10.2, as follows:

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

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

Proposition 10.3. *Any function $f : X \to \mathbb{R}$ can be made into a function*

$$f : X \to Y$$

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

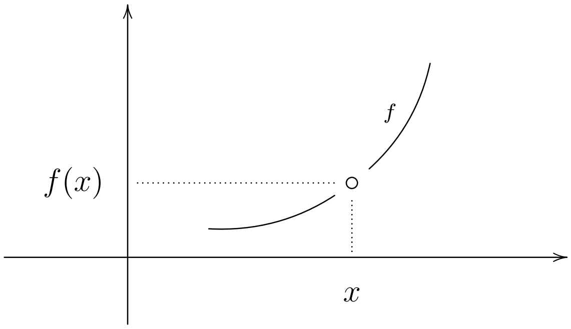
Proof. This is indeed something clear from definitions, as explained above. □

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

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

Question 10.4. *How to suitably represent our functions $f : \mathbb{R} \to \mathbb{R}$?*

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



You are certainly familiar with this, drawing such graphs, so let us record here:

ANSWER 10.5. *The functions $f:\mathbb{R}\to\mathbb{R}$ are usually well represented by their graphs, drawn as usual in 2D, with the convention $y=f(x)$.*

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

THEOREM 10.6. *Given a bijective function $f:\mathbb{R}\to\mathbb{R}$, its inverse function*

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

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

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

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

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

In what regards now the more general functions, $f:X\to Y$ with $X,Y\subset\mathbb{R}$, as in Definition 10.2, pretty much the same can be said here, and we have:

THEOREM 10.7 (update). *Given a bijective function $f:X\to Y$, its inverse function*

$$f^{-1}:Y\to X$$

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

PROOF. This is indeed a straightforward generalization of Theorem 10.6. □

We will see in what follows many other applications of the graphs of functions, for countless other questions that we can have, about them. However, as a word of warning, the graph of a function is not everything. For instance the very basic function $f(x)=2x$

remains best thought of as it comes, in 1D, as being the function which elongates all the distances by 2, and with this property being harder to see on its graph:

Warning 10.8. *The graph is not everything, with for instance the function*

$$f(x)=2x$$

being best thought of as it comes, as the function elongating all distances by 2*.*

With this discussed, let us focus now our study on the functions $f:\mathbb{R}\to\mathbb{R}$ which are suitably regular, with the hope as usual of getting into interesting mathematics. And, in what regards these regularity properties, the most basic of them is continuity:

Definition 10.9. *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.

Regarding now the basic examples of countinous functions, there are many of them, and we will discuss them in a moment, once we will have some basic tools, in order to prove that this or that function is continuous or not, without much pain. As a matter, however, of having a first illustration for Definition 10.9, let us record here:

Theorem 10.10. *The basic power functions, namely*

$$f(x)=x^k$$

with $k\in\mathbb{N}$, are all continuous.

Proof. According to Definition 10.9, we want to prove that we have:

$$x_n\to x\implies x_n^k\to x^k$$

(1) A first method is by using the results from chapter 3 regarding the sequences. To be more precise, we know from there that the following formula holds:

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

But with $x_n=y_n$, this leads to the following formula:

$$\lim_{n\to\infty}x_n^2=\left(\lim_{n\to\infty}x_n\right)^2$$

Obviously, we can iterate this method, and so for any $k\in\mathbb{N}$, we have:

$$\lim_{n\to\infty}x_n^k=\left(\lim_{n\to\infty}x_n\right)^k$$

But now, assuming $x_n\to x$ as above, this formula gives, as desired:

$$\lim_{n\to\infty}x_n^k=x^k$$

(2) As a second method, more direct, we must estimate quantities of type $(x+t)^k-x^k$, with t small. But we can do this with the binomial formula, which gives, for $|t|\leq 1$:

$$\begin{aligned}
|(x+t)^k-x^k| &= \left|\sum_{s=0}^k\binom{k}{s}x^{k-s}t^s-x^k\right|\\
&= \left|\sum_{s=1}^k\binom{k}{s}x^{k-s}t^s\right|\\
&\leq \sum_{s=1}^k\binom{k}{s}|x|^{k-s}|t|^s\\
&\leq |t|\sum_{s=1}^k\binom{k}{s}|x|^{k-s}\\
&\leq |t|\sum_{s=0}^k\binom{k}{s}|x|^{k-s}\\
&= |t|(1+|x|)^k
\end{aligned}$$

Now assume $x_n\to x$. We can then write $x_n=x+t_n$, and by choosing our $n>>0$ as to have $|t_n|\leq 1$, we can use the above estimate, which gives:

$$|x_n^k-x^k|\leq|t_n|(1+|x|^k)$$

Now since we have $t_n\to 0$, we obtain from this $x_n^k\to x^k$, as desired. □

Getting back now to the general theory, and to Definition 10.9 as stated, many things can be said. To start with, there are many other equivalent formulations of the notion of continuity, with a well-known, useful, and much feared one, being as follows:

THEOREM 10.11. *A function $f:X\to\mathbb{R}$ is continuous when*

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

holds.

PROOF. Let us prove this, with no fear. According to Definition 10.9, in order for our function f to be continuous, the following must happen, for any $x\in X$:

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

Now when reminding what convergence of a sequence exactly means, for both the convergences $x_n\to x$ and $f(x_n)\to f(x)$, we are led to the conclusion in the statement. □

So long for the continuity basics. As a last piece of general theory, there are also many interesting functions which have discontinuities, and for these, we can use:

Definition 10.12. *Given a function $f : X \to \mathbb{R}$ and $x \in X$, we set*

$$f(x_-) = \lim_{y \nearrow x} f(y) \quad , \quad f(x_+) = \lim_{y \searrow x} f(y)$$

provided that these two limits exist indeed, and we call the quantity

$$J_f(x) = f(x_+) - f(x_-)$$

which does not depend on $f(x)$, the jump of f at the given point $x \in X$.

As a first observation, assuming that a function $f : X \to \mathbb{R}$ is continuous at $x \in X$, its jump there is zero, so that we have the following implications:

$$f \text{ continuous at } x \implies J_f(x) = 0$$

$$f \text{ continuous} \implies J_f(x) = 0, \ \forall x \in X$$

Observe also that the converses of these implications do not necessarily hold, because the jump $J_f(x)$ as constructed above does not depend on $f(x)$, so we can easily construct counterexamples, just by modifying the value $f(x)$. As illustrations now, we have:

Proposition 10.13. *For the basic step function, given by*

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

we have $J_f(0) = 1$, as we should. Also, for the inverse function

$$g(x) = \frac{1}{x}$$

say defined with $g(0) = \alpha$, we have $J_g(0) = \infty$, again as we should.

Proof. Both formulae are clear from definitions. Indeed, we have:

$$\begin{aligned} J_f(0) &= f(0_+) - f(0_-) \\ &= \lim_{y \searrow 0} 1 - \lim_{y \nearrow 0} 0 \\ &= 1 - 0 \\ &= 1 \end{aligned}$$

As for the second formula, the computation here is similar, as follows:

$$\begin{aligned} J_g(0) &= g(0_+) - g(0_-) \\ &= \lim_{y \searrow 0} \frac{1}{y} - \lim_{y \nearrow 0} \frac{1}{y} \\ &= \infty - (-\infty) \\ &= \infty \end{aligned}$$

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

Getting back now to generalities, we can use the jump in the following way:

THEOREM 10.14. *Assuming that $f : X \to \mathbb{R}$ is discontinuous at $x \in X$, we can make it continuous there, by suitably changing $f(x)$, precisely when $J_f(x) = 0$.*

PROOF. Indeed, assuming $J_f(x) = 0$, we can make f continuous at x by setting:

$$f(x) = f(x_+) = f(x_-)$$

As for the converse, this is clear too, as already observed after Definition 10.12. □

Back now to the continuous functions, in order to get towards examples, let us start with the following theoretical result, regarding the various operations on functions:

THEOREM 10.15. *If f, g are continuous, then so are:*

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

PROOF. Before anything, we should mention that the claim is that (1-4) hold indeed, provided that at the level of domains and ranges, the statement makes sense. For instance in (1,2,3) we are talking about functions having the same domain, and with $g(x) \neq 0$ for the needs of (3), and there is a similar discussion regarding (4).

(1) The claim here is that if both f, g are continuous at a point x, then so is the sum $f + g$. But this is clear from the similar result for sequences, namely:

$$\lim_{n\to\infty}(x_n + y_n) = \lim_{n\to\infty} x_n + \lim_{n\to\infty} y_n$$

(2) Again, the statement here is similar, and the result follows from:

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

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

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

(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, using that scary ε,δ condition from Theorem 10.11, let us pick $\varepsilon > 0$. 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. □

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

THEOREM 10.16. *The following functions are continuous:*

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

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

(1) Since $f(x) = x$ is continuous, by using Theorem 10.15 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 10.15.

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

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

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

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

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:

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

Alternatively, we can use the same method as for sin, and we get, as desired:

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

(4) The fact that the functions $\tan x$ and $\sec x$, $\csc x$, $\cot x$ are continuous too is clear from the fact that $\sin x$, $\cos x$ are continuous, by using Theorem 10.15 (3).

(5) The continuity of $x\to e^x$ comes at $x=0$ from the following computation:

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

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

$$\lim_{t\to 0}e^{x+t}=\lim_{t\to 0}e^xe^t=e^x\lim_{t\to 0}e^t=e^x\cdot 1=e^x$$

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

10b. Intermediate values

We would like to explain now an alternative formulation of the notion of continuity, which is quite abstract, but is definitely worth learning, because it is quite powerful, solving some of the questions that we have left. Let us start with:

Definition 10.17. *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 10.18. *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 10.19. *For a function $f : \mathbb{R} \to \mathbb{R}$, the following are equivalent:*

(1) *f is continuous.*
(2) *$f^{-1}(O)$ is open, for any O open.*
(3) *$f^{-1}(C)$ is closed, for any C closed.*

Proof. This is something which follows from definitions, as follows:

(1) $\iff$ (2) This equivalence, which is the main one, is best viewed by using the ε, δ definition of continuity, from Theorem 10.11, 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 from the definition of the open sets that if O is open, then $f^{-1}(O)$ is open, and that the converse holds too.

(1) $\iff$ (3) This follows either directly, by using the $f(x_n) \to f(x)$ definition of continuity, or indirectly, by combining (1) $\iff$ (2) above with (2) $\iff$ (3) below.

(2) $\iff$ (3) This is clear indeed by taking complements, and using Proposition 10.18, both for the sets in question, and for their preimages via f. □

As a test for the above criterion, let us reprove the fact, that we know from Theorem 10.15, that if f, g are continuous, so is $f \circ g$. But this is clear, coming from:

$$(f \circ g)^{-1}(O) = g^{-1}(f^{-1}(O))$$

In short, not bad, because at least in relation with this specific problem, our proof using open sets is as simple as the simplest proof, namely the one using $f(x_n)\to f(x)$, and is simpler than the other proof that we know, namely the one with ε,δ.

In order to reach now to true applications of Theorem 10.19, we will need to know more about the open and closed sets. Let us begin with a useful result, as follows:

Proposition 10.20. *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 10.21. *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 10.20 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 $O = \cup_x I_x$ in (2) can be taken to be disjoint. For this purpose, our first observation is that, by approximating points $x \in O$ by rationals $y \in \mathbb{Q} \cap O$, we can make our union to be countable. But once our union is countable, we can start merging intervals, whenever they meet, and we are left in the end with a countable, disjoint union, as desired.

(4) Finally, the second assertion comes from Proposition 10.18. □

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 10.22. *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 will soon see, 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 the following result:

Theorem 10.23. *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) Let us first prove that any compact set K must be closed. But this is clear by contradiction, because assuming that that $\{x_n\} \subset K$ has as limit $x \notin K$, we can find an open cover of K, as follows, obviously having no open subcover:

$$K \subset \bigcup_{n\in\mathbb{N}} \left[x - \frac{1}{n}\,,\, x + \frac{1}{n}\right]^c$$

Similarly, any compact set K must be bounded, by using the following cover:

$$K \subset \bigcup_{N\in\mathbb{N}} (-N, N)$$

(2) It remains now to prove the converse, stating that a closed and bounded set K must be compact. This is something more tricky, and we will do this first for the unit interval $K = [0,1]$. So, consider an arbitrary open cover of $[0,1]$, as follows:

$$[0,1] \subset \bigcup_i U_i$$

Assume now by contradiction that this cover has no finite subcover. Then one of the corresponding covers of $[0,1/2]$ and $[1/2,1]$ must have the same property, say the corresponding cover of $[0,1/2]$ has the same property. But then one of the corresponding covers of $[0,1/4]$ and $[1/4,1/2]$ must have the same property. And so on, and in the end, what we get is a descreasing sequence of intervals having this property:

$$[0,1]\supset[a_1,b_1]\supset[a_2,b_2]\supset\ldots$$

But this is contradictory, because the limiting point $x\in[0,1]$ of this descreasing sequence of intervals must be covered by one of the sets U_i, say via $(x-\varepsilon,x+\varepsilon)\subset U_i$, and if we choose $n\in\mathbb N$ such that $1/2^n<\varepsilon$, then we will have an inclusion as follows, which can be regarded as being a finite subcover, with just 1 open set involved:

$$[a_n,b_n]\subset(x-\varepsilon,x+\varepsilon)\subset U_i$$

(3) Summarizing, we have proved that the unit interval $K=[0,1]$ is compact, and by using now translations and dilations, in the obvious way, and we will leave this as an exercise, we conclude that any closed bounded interval $K=[a,b]$ is compact.

(4) Time now to finish the proof started in (2). Assuming that K is closed and bounded, by boundedness we have an inclusion as follows, with a,b being finite:

$$K\subset[a,b]$$

Now consider an open cover of K. By adding the set $[a,b]-K$ to this cover, we obtain an open cover of $[a,b]$, which by (3) must have a finite subcover. Now by removing the set $[a,b]-K$ from this subcover, in case it is present, we obtain a finite subcover of K. We conclude that our closed and bounded set K must be compact, as claimed.

(5) Finally, the second assertion in the statement is something quite obvious, and this regardless of what “cannot be broken into parts” in Definition 10.22 exactly means, 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. □

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

Theorem 10.24. *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 10.22. To be more precise:

(1) This comes indeed from the fact, based on Theorem 10.19, 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 which is clear as well, because if $f(E)$ can be split into two parts, then by applying f^{-1} we can split as well E into two parts. □

You might perhaps ask at this point, was Theorem 10.24 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 10.25. *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, the best 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. Just like that. □

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

THEOREM 10.26. *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 10.25. 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. □

10c. Elementary functions

Time now for some concrete applications, of our intermediate value technology learned above. As a first such application, which is something fundamental, we have:

THEOREM 10.27. *We can talk about the following inverse functions, which are all well-defined, continuous and monotone, exactly as the original functions are:*

(1) $\arcsin : (-1, 1) \to (-\pi/2, \pi/2)$, *the inverse of* $\sin : (-\pi/2, \pi/2) \to (-1, 1)$.
(2) $\arccos : (-1, 1) \to (0, \pi)$, *the inverse of* $\cos : (0, \pi) \to (-1, 1)$.
(3) $\arctan : \mathbb{R} \to (-\pi/2, \pi/2)$, *the inverse of* $\tan : (-\pi/2, \pi/2) \to \mathbb{R}$.
(4) $\log : (0, \infty) \to \mathbb{R}$, *the inverse of* $\exp : \mathbb{R} \to (0, \infty)$.

PROOF. This follows indeed from Theorem 10.26, with the discussion regarding continuity and monotony being something that we already know, from earlier in this book. □

As another basic application of our intermediate value technology, we have:

Theorem 10.28. *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 10.25, as follows:

(1) This is clear from Theorem 10.25 (3), applied on $[-\infty,\infty]$.

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

(3) This follows as well from Theorem 10.25 (3), applied to the same polynomial $P(z) = z^n - x$, but this time on the interval $[0,\infty)$. □

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

Theorem 10.29. *The function x^a is defined and continuous on $(0,\infty)$, for any $a \in \mathbb{R}$. Moreover, when trying to extend it to $\mathbb{R}$, we have 4 cases, as follows,*

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

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

Proof. The idea is that we know how to extract roots by using Theorem 10.28, and all the rest follows by continuity. To be more precise:

(1) Assume $a = p/q$, with $p,q \in \mathbb{N}$, $p \neq 0$ and q odd. Given a number $x \in \mathbb{R}$, we can construct the power x^a in the following way, by using Theorem 10.28:

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

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

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

(3) Assume first $a \in \mathbb{Q}_{even}$, $a > 0$. This means $a = p/q$ with $p,q \in \mathbb{N}$, $p \neq 0$ and q even, and as before in (1), we can set $x^a = \sqrt[q]{x^p}$ for $x \geq 0$, by using Theorem 10.28. 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 10.30. *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 10.29, by exchanging the variables, $x \leftrightarrow a$. To be more precise, the situation is as follows:

(1) We know from Theorem 10.29 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 10.29 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. □

Getting back to theory, a closer look at the wealth of the functions that we have reveals that "some functions are more continuous than some other". To be more precise, here is a theoretical result, making the difference between the various types of functions:

THEOREM 10.31. *Consider the following properties, regarding $f : X \to \mathbb{R}$ with $X \subset \mathbb{R}$:*

(1) *f has the following property, for some $K > 0$, called Lipschitz property:*

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

(2) *f is uniformly continuous, in the sense that the following happens:*

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

(3) *f is continuous in the usual sense, namely:*

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

We have then $(1) \implies (2) \implies (3)$. *Also, the converse implications do not hold.*

PROOF. This is something quite self-explanatory, with $(1) \implies (2)$ being clear, coming by talking $\delta = \varepsilon/K$, and with $(2) \implies (3)$ being plainly trivial. As for the counterexamples, x^2 is continuous but not uniformly continuous, for quite obvious reasons, while $\sqrt{|x|}$ is uniformly continuous but not Lipschitz, again for obvious reasons. □

Quite interesting all this, and in practice, all this seems to have something to do with the slope of the graph of f, computed at various points. We will be back to this later, in chapter 11, when talking slopes of graphs, or derivatives, which can help with this.

In the meantime, let us record the following key result, due to Heine and Cantor:

Theorem 10.32. *Any continuous function defined on a compact set*

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

and in particular, defined on $X=[a,b]$, is automatically uniformly continuous.

Proof. Given $\varepsilon>0$, for any $x\in X$ we know that we have a $\delta_x>0$ such that:

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

So, consider the following open intervals, centered at the various points $x\in X$:

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

These cover our compact set X, so consider a finite subcover of this cover:

$$X\subset\bigcup_i U_{x_i}$$

With this done, consider as well the following number, which is strictly positive:

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

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. □

10d. Weierstrass theorem

Our goal now will be to extend the material from chapter 3 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 10.33. *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 3 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:

Theorem 10.34. *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 $\arctan(x)$.

Proof. We know that $\arctan(x)$ 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 $\pi/2$, which amounts in considering the following function:

$$f(x) = \frac{2}{\pi}\arctan(x)$$

Now with this done, we must stretch the variable x, as to get our function closer and closer to $sgn(x)$. This can be done in several ways, a standard one being as follows:

$$g_n(x) = \frac{2}{\pi}\arctan(nx)$$

So, let us see if this works. First, we have the following computation, for $x > 0$:

$$\begin{aligned} \lim_{n\to\infty} g_n(x) &= \frac{2}{\pi}\lim_{n\to\infty}\arctan(nx) \\ &= \frac{2}{\pi}\arctan(\infty) \\ &= \frac{2}{\pi}\cdot\frac{\pi}{2} \\ &= 1 \end{aligned}$$

Similarly, we have the following computation, this time for $x < 0$:

$$\begin{aligned}\lim_{n\to\infty} g_n(x) &= \frac{2}{\pi}\lim_{n\to\infty}\arctan(nx)\\ &= \frac{2}{\pi}\arctan(-\infty)\\ &= \frac{2}{\pi}\left(-\frac{\pi}{2}\right)\\ &= -1\end{aligned}$$

Finally, for $x = 0$ the limit is that of the constant 0 sequence, as follows:

$$\lim_{n\to\infty} g_n(0) = \lim_{n\to\infty} 0 = 0$$

We conclude from this that we have the following pointwise convergence:

$$\lim_{n\to\infty} g_n(x) = \begin{cases} -1 & , \quad x < 0\\ 0 & , \quad x = 0\\ 1 & , \quad x > 0\end{cases}$$

In other words, we have proved that we have the following approximation:

$$\lim_{n\to\infty}\frac{2}{\pi}\arctan(nx) = sgn(x)$$

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

So, this is the situation with pointwise convergence, and not very good all this, hope you agree with me. In fact, even worse, we have truly scary things, as follows:

FACT 10.35. *There are examples of pointwise convergence of functions*

$$f_n \to f$$

with each f_n being continuous, and with f totally discontinuous.

To be more precise, this is something a bit more technical, that we will not really need in what follows, and that we will leave 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 10.36. *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 10.33 and Definition 10.36. As a first observation, we have:

PROPOSITION 10.37. *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 that we met before Theorem 10.34, namely:

$$f_n : [0,1] \to \mathbb{R} \quad , \quad f_n(x) = x^n$$

Indeed, uniform convergence of these functions 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 10.38. *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 10.36 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. □

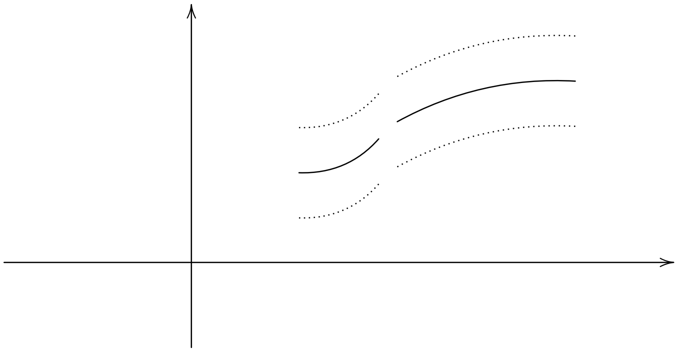
Obviously, the notion of uniform convergence in Definition 10.36 is something quite interesting, worth some more study. As a next result about it, we have:

Proposition 10.39. *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 3, that we know well, with no difficulties or tricks involved. □

Still talking generalities, it is also possible to draw pictures, and more specifically, to use the "strip" interpretation of the notion of uniform continuity, which is as follows:



To be more precise, assuming that the curve in the middle represents f, the uniform convergence $f_n \to_u f$ means that, whenever $\varepsilon > 0$ is given, if we draw the $(-\varepsilon, \varepsilon)$ strip around f, as above, the graph of f_n with $n >> 0$ must lie in the strip.

There is some abstract mathematics to be done as well, as follows:

Proposition 10.40. *The uniform convergence condition* $f_n \to_u f$ *is equivalent to*

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

and with the sup and the limit being, as usual, not to be interverted.

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

(1) To start with, what we say above is indeed clear from definitions. And as a comment, all this is even more clear when using the "strip" interpretation of the notion of uniform continuity, given just before the statement.

(2) In what regards now the last assertion, this is our usual word of warning, regarding such things, but out of curiosity, let us see what happens, when doing that bad thing. To

start with, the formula in the statement can be written as follows:

$$\lim_{n\to\infty}\sup_x\left|f_n(x)-f(x)\right|=0$$

Now let us do the bad thing, namely interverting the limit and the supremum:

$$\sup_x\lim_{n\to\infty}\left|f_n(x)-f(x)\right|=0$$

So, what does this latter condition mean? Since a supremum of positive numbers vanishes precisely when all the positive numbers vanish, this is equivalent to:

$$\lim_{n\to\infty}\left|f_n(x)-f(x)\right|=0\quad,\quad\forall x$$

But, what we have here is the old notion of convergence, the pointwise one. □

Getting now to more concrete things, we have the following fundamental result, due to Weierstrass, regarding the approximation of functions by polynomials:

Theorem 10.41 (Weierstrass). *Any continuous function on a closed interval*

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

can be uniformly approximated by polynomials.

Proof. This is indeed something very classical, with a well-known, constructive proof, being by using an approximation by suitable Bernstein polynomials, namely:

$$f_n(x)=\sum_{k=0}^n f\left(\frac{k}{n}\right)b_{kn}(x)$$

To be more precise, we assume here that $[a,b]=[0,1]$, and we set:

$$b_{kn}(x)=\binom{n}{k}x^k(1-x)^{n-k}$$

As for the proof of this, this is something well-known, which goes as follows:

(1) Consider indeed the basic Bernstein polynomials b_{kn}, as constructed above. These remind the binomial laws, so it is with some probability that we will start. We have the following formulae, which are all elementary to establish, and which in probabilistic terms are dealing with the moments of order $0,1,2$ of the binomial laws:

$$\sum_k\binom{n}{k}x^k(1-x)^{n-k}=1$$

$$\sum_k\frac{k}{n}\binom{n}{k}x^k(1-x)^{n-k}=x$$

$$\sum_k\left(x-\frac{k}{n}\right)^2\binom{n}{k}x^k(1-x)^{n-k}=\frac{x(1-x)}{n}$$

(2) In terms of the basic Bernstein polynomials b_{kn}, the above formulae read:

$$\sum_k b_{kn}(x)=1$$

$$\sum_k \frac{k}{n}\cdot b_{kn}(x)=x$$

$$\sum_k \left(x-\frac{k}{n}\right)^2 b_{kn}(x)=\frac{x(1-x)}{n}$$

(3) Now consider our arbitrary continuous function $f:[0,1]\to\mathbb{R}$, and construct for any $n\in\mathbb{N}$ the approximation indicated above, namely:

$$f_n(x)=\sum_{k=0}^n f\left(\frac{k}{n}\right)b_{kn}(x)$$

In order to estimate the error $|f_n-f|$, we use the uniform continuity property of f. So, pick $\varepsilon>0$, and then $\delta>0$ such that the following happens:

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

Now with this done, we have the following estimate, using the first formula in (2) at the first step, the uniform continuity at the last step, and with $M=\sup|f|$:

$$\begin{aligned}
&|f_n(x)-f(x)|\\
=&\left|\sum_k\left(f\left(\frac{k}{n}\right)-f(x)\right)b_{kn}(x)\right|\\
\leq&\sum_k\left|f\left(\frac{k}{n}\right)-f(x)\right|b_{kn}(x)\\
=&\sum_{\left|x-\frac{k}{n}<\delta\right|}\left|f\left(\frac{k}{n}\right)-f(x)\right|b_{kn}(x)+\sum_{\left|x-\frac{k}{n}\geq\delta\right|}\left|f\left(\frac{k}{n}\right)-f(x)\right|b_{kn}(x)\\
\leq&\ \varepsilon+M\sum_{\left|x-\frac{k}{n}\geq\delta\right|}b_{kn}(x)
\end{aligned}$$

(4) The point now is that the last sum on the right can be estimated by using the Chebycheff inequality, based on the third formula from (2), and we obtain:

$$\begin{aligned}\sum_{\left|x-\frac{k}{n}\geq\delta\right|} b_{kn}(x) &\leq \sum_k \delta^{-2}\left(x-\frac{k}{n}\right)^2 b_{kn}(x)\\ &= \delta^{-2}\,\frac{x(1-x)}{n}\\ &\leq \frac{\delta^{-2}}{4n}\end{aligned}$$

(5) Now by putting everything together, we obtain the following estimate:

$$|f_n(x)-f(x)|\leq\varepsilon+\frac{\delta^{-2}M}{4n}$$

Thus we have indeed $|f_n - f| \to 0$, uniform convergence, as desired.

(6) Summarizing, present theorem proved, modulo some learning in relation with the Chebycheff inequality that we used above, that we will leave as an exercise. □

10e. Exercises

There is no serious mathematics without functions, and as exercises here, we have:

Exercise 10.42. *Imagine various mechanical devices, that can represent functions.*

Exercise 10.43. *Learn more about discontinuous functions, and their jumps.*

Exercise 10.44. *Learn more about compact sets, and their applications.*

Exercise 10.45. *Learn as well about the various types of connected sets.*

Exercise 10.46. *Learn some other proofs for the intermediate value theorem.*

Exercise 10.47. *Clarify what we said, about inverse trigonometric functions.*

Exercise 10.48. *Clarify what we said about about the functions a^x and x^a.*

Exercise 10.49. *Compute Lipschitz constants for all the basic functions.*

As bonus exercise, read more about the Weierstrass theorem, and its applications.

CHAPTER 11

Derivatives

11a. 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? To be more precise, the above approximation means that we have a formula as follows, with $\varepsilon(t)\to0$ for $t\to0$:

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

Thus, what we are looking for is a better approximation, of the following type, with the function $\rho(t)$ being some sort of simple approximation of the error term $\varepsilon(t)$:

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

And a bit of thinking at all this, or just drawing a picture, suggests to look at the slope of f at the point x. Which leads us into the following notion:

Definition 11.1. *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.

We will see in a moment that this definition provides the key to the solution of our approximation problem. Before that, however, let us comment a bit on this notion.

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:

$$\lim_{t\to0}[f(x+t)-f(x)]=0$$

Thus, f must be continuous at x. However, the converse is not true, a basic counterexample being $f(x)=|x|$ at $x=0$. Let us summarize these findings as follows:

Proposition 11.2. *If f is differentiable at x, then f must be continuous at x. However, the converse is not true, with the modulus function*

$$f(x)=|x|$$

being a basic counterexample for this, 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 11.1. We have:

$$\lim_{t\searrow 0}\frac{|0+t|-|0|}{t}=\lim_{t\searrow 0}\frac{t-0}{t}=1$$

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

$$\lim_{t\nearrow 0}\frac{|0+t|-|0|}{t}=\lim_{t\nearrow 0}\frac{-t-0}{t}=-1$$

Thus, the limit in Definition 11.1 does not converge, as desired. □

Generally speaking, the last assertion in Proposition 11.2 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 11.3. *Assuming that f is differentiable at x, we have:*

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

In other words, f is, approximately, locally affine at x.

Proof. Assume indeed that f is differentiable at x, and let us set, as before:

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

By multiplying by t, we obtain that we have, once again in the $t\to 0$ limit:

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

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

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

Theorem 11.4. *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:

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

With $p = m/n$ with $m \in \mathbb{Z}$ and $n \in \mathbb{N}$, as above, we set in this formula:

$$a = (x+t)^{m/n} \quad , \quad b = x^{m/n}$$

We obtain in this way, as desired, the following approximation:

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

(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:

Theorem 11.5. *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:

$$\begin{aligned}(\sin x)' &= \lim_{t\to 0}\frac{\sin(x+t)-\sin x}{t}\\ &= \lim_{t\to 0}\frac{\sin x\cos t+\cos x\sin t-\sin x}{t}\\ &= \lim_{t\to 0}\sin x\cdot\frac{\cos t-1}{t}+\cos x\cdot\frac{\sin t}{t}\\ &= \cos x\end{aligned}$$

(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 11.6. *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 the following conditions:

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

Due to $f(0)=1$, the first term must be 1, so our function must look as follows:

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

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

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

Thus, the equation $f'=f$ is equivalent to the following equalities:

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

But this system of equations can be solved by recurrence, as follows:

$$c_1=1\quad,\quad c_2=\frac{1}{2}\quad,\quad c_3=\frac{1}{2\times 3}\quad,\quad c_4=\frac{1}{2\times 3\times 4}\quad,\quad\ldots$$

Thus we have $c_k=1/k!$, leading to the conclusion in the statement. □

Observe that the above result leads to a more conceptual explanation for the number e itself. To be more precise, $e\in\mathbb R$ is the unique number satisfying:

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

Which is something good to know, you can even attend Bourbaki seminars now.

Let us work out now some general results, for the computation of the derivatives. We have here the following statement, which is the key to everything computations:

THEOREM 11.7. *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:

Theorem 11.8. *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 11.7 (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 11.7 (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 11.9. *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 the tangent, 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 11.3, we have:

Theorem 11.10. *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 11.3, 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 11.10, we have:

Theorem 11.11. *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 11.10. 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 11.10 can be used in order to find the maximum and minimum of any differentiable function, and this method is best recalled as follows:

Algorithm 11.12. *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.*

To be more precise, we are using here Theorem 11.10, or rather the obvious extension of this result to the case of the functions $f:[a,b]\to\mathbb{R}$. This tells us that the local minima and maxima of our function f, and in particular the global minima and maxima, can be found among the zeroes of the first derivative f', with the endpoints a,b added. Thus, what we have to do is to compute these "candidates", as explained in steps (1-2-3), and then see what each candidate is exactly worth, as explained in steps (4-5-6).

Needless to say, all this is very interesting, and powerful. The general problem in any type of applied mathematics is that of finding the minimum or maximum of some function, and we have now an algorithm for dealing with such questions. Very nice.

11b. 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 11.13. *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.

But you might probably wonder why coming with this new 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 the second derivatives f''. Regarding them, we first have:

Interpretation 11.14. *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 indeed a very interesting notion. In practice now, the situation is as follows:

(1) This is something which is clear, and very intuitive, coming from the usual interpretation of the derivative, as both a growth rate, and a slope.

(2) This is some sort of reformulation of (1), using the intuitive meaning of the word "acceleration", with the relevant physics equations, due to Newton, being as follows:

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

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

(3) This is something more subtle, of statistical nature, and which is very useful for applications, that we will clarify with some mathematics, in a moment.

(4) This is something quite subtle too, which is again very useful for applications, and that we will clarify as well with some mathematics, in a moment.

All in all, what we have in Interpretation 11.14 is a mixture of trivial and non-trivial facts, and do not worry, we will get familiar with all this, in the next few pages.

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

Theorem 11.15. *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 11.14, 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:

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

This function is then differentiable, with its derivative being:

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

Thus, 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 11.16. *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 11.3 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 11.3. 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 11.3, and there are many applications of it. As a first such application, justifying Interpretation 11.14 (3), we have the following statement, which is a bit heuristic, but we will call it however Proposition:

Proposition 11.17. *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 11.16, 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:

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

But this is what our statement says, save for some uncertainties regarding the averaging method I, 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 11.10, as follows:

Theorem 11.18. *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 11.16, 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 11.10, 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 11.12 as well. Normally that algorithm stays strong, because Theorem 11.18 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 11.18 is good to know, when applying that algorithm.

11c. Convex functions

As a main concrete application now of the second derivative, which is something very useful in practice, and related to Interpretation 11.14 (4), we have the following result:

Theorem 11.19. *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,*

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

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:

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

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

$$f''(x) \geq 0$$

(4) Thus, we are led to the conclusions in the statement, regarding the convex functions. As for the concave functions, the proof here is similar. Alternatively, we can say that f is concave precisely when $-f$ is convex, and get the results from what we have. □

As a basic application of the Jensen inequality, which is very classical, we have:

THEOREM 11.20. *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 11.19, 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:

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

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

Theorem 11.21 (Young). *We have the following inequality,*

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

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

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

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

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

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

Now by exponentiating, we obtain the Young inequality. □

Moving forward now, as a consequence of the Young inequality, we have:

Theorem 11.22 (Hölder). *Assuming that $p,q\geq 1$ are conjugate, in the sense that*

$$\frac{1}{p}+\frac{1}{q}=1$$

we have the following inequality, valid for any two vectors $x,y\in\mathbb{R}^N$,

$$\sum_i|x_iy_i|\leq\left(\sum_i|x_i|^p\right)^{1/p}\left(\sum_i|y_i|^q\right)^{1/q}$$

with the convention that an ∞ exponent produces a $\max|x_i|$ quantity.

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

(1) Assume first that we are dealing with finite exponents, $p,q\in(1,\infty)$. By linearity we can assume that x,y are normalized, in the following way:

$$\sum_i|x_i|^p=\sum_i|y_i|^q=1$$

But in this case, we use the Young inequality, which gives, as desired:

$$\begin{aligned}\sum_i|x_iy_i|&\leq\sum_i\frac{|x_i|^p}{p}+\sum_i\frac{|y_i|^q}{q}\\&=\frac{1}{p}+\frac{1}{q}\\&=1\end{aligned}$$

(2) In the case $p=1$ and $q=\infty$, or vice versa, the inequality holds too, trivially, with the convention that an ∞ exponent produces a max quantity, according to:

$$\lim_{p\to\infty}\left(\sum_i|x_i|^p\right)^{1/p}=\max|x_i|$$

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

As a consequence now of the Hölder inequality, we have:

Theorem 11.23 (Minkowski). *Assuming $p\in[1,\infty]$, we have the inequality*

$$\left(\sum_i|x_i+y_i|^p\right)^{1/p}\leq\left(\sum_i|x_i|^p\right)^{1/p}+\left(\sum_i|y_i|^p\right)^{1/p}$$

for any two vectors $x,y\in\mathbb R^N$, with our usual conventions at $p=\infty$.

Proof. We have indeed the following estimate, using the Hölder inequality, and the conjugate exponent $q\in[1,\infty]$, given by $1/p+1/q=1$:

$$\begin{aligned}\sum_i|x_i+y_i|^p&=\sum_i|x_i+y_i|\cdot|x_i+y_i|^{p-1}\\&\leq\sum_i|x_i|\cdot|x_i+y_i|^{p-1}+\sum_i|y_i|\cdot|x_i+y_i|^{p-1}\\&\leq\left(\sum_i|x_i|^p\right)^{1/p}\left(\sum_i|x_i+y_i|^{(p-1)q}\right)^{1/q}\\&+\left(\sum_i|y_i|^p\right)^{1/p}\left(\sum_i|x_i+y_i|^{(p-1)q}\right)^{1/q}\\&=\left[\left(\sum_i|x_i|^p\right)^{1/p}+\left(\sum_i|y_i|^p\right)^{1/p}\right]\left(\sum_i|x_i+y_i|^p\right)^{1-1/p}\end{aligned}$$

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

The Minkowski theorem is quite interesting, allowing us to redefine the length of vectors in $\mathbb R^N$, in case the usual length does not perform well, in the following way:

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

And there are many applications of this trick, all across advanced mathematics.

11d. Taylor formula

Back now to the general theory of the derivatives, and their theoretical applications, we can further develop our basic approximation method, at order 3, at order 4, and so on. Let us start with something nice and intuitive, coming from physics, as follows:

FACT 11.24. *In analogy with the fact that the second derivative measures the acceleration of the slope, the third derivative measures the jerk of the slope.*

Here the terminology comes from real life and classical mechanics, where the jerk is by definition the derivative of the acceleration, and so is the second derivative of the speed, or third derivative of the position, according to the following formulae:

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

As before with second derivatives, many other things can be said. Let us also record the formulae of the third derivatives of the basic functions, which are as follows:

THEOREM 11.25. *The third derivatives of the basic functions are as follows:*

(1) $(x^p)''' = p(p-1)(p-2)x^{p-3}$.
(2) $\sin''' = -\cos$.
(3) $\cos''' = \sin$.
(4) $\exp''' = \exp$.
(5) $\log'''(x) = 2/x^3$.

PROOF. The various formulae in the statement all follow from the various formulae for the second derivatives established before, as follows:

$$(x^p)''' = (p(p-1)x^{p-2})' = p(p-1)(p-2)x^{p-3}$$

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

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

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

$$(\log x)''' = (-1/x^2)' = 2/x^3$$

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

Getting now to the fourth derivatives, things are less intuitive here, in what regards the interpretation, but we can nevertheless do some computations, as follows:

THEOREM 11.26. *The fourth derivatives of the basic functions are as follows:*

(1) $(x^p)'''' = p(p-1)(p-2)(p-3)x^{p-4}$.
(2) $\sin'''' = \sin$.
(3) $\cos'''' = \cos$.
(4) $\exp'''' = \exp$.
(5) $\log''''(x) = -6/x^4$.

Proof. The various formulae in the statement all follow from the various formulae for the third derivatives established before, as follows:

$$(x^p)'''' = (p(p-1)(p-2)x^{p-3})' = p(p-1)(p-2)(p-3)x^{p-4}$$

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

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

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

$$(\log x)'''' = (2/x^3)' = -6/x^4$$

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

With this discussed, and getting back now to our usual approximation business, the ultimate result on the subject, called Taylor formula, is as follows:

Theorem 11.27. *Any function $f:\mathbb{R}\to\mathbb{R}$ can be locally approximated as*

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

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

Proof. Consider the function to be approximated, namely:

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

Let us try to best approximate this function at a given order $n\in\mathbb{N}$. We are therefore looking for a certain polynomial in t, of the following type:

$$P(t) = a_0 + a_1 t + \ldots + a_n t^n$$

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

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

In order to prove now that this approximation holds indeed, we can use L'Hôpital's rule, applied several times, as in the proof of Theorem 11.16. 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 11.28. *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 the next chapter.

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

Theorem 11.29. *Given a differentiable function $f : \mathbb{R} \to \mathbb{R}$, we can always write*

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

with $f^{(n)}(x) \neq 0$, and this tells us if x is a local minimum, or maximum of f.

Proof. This is indeed something self-explanatory, coming from Theorem 11.27, with the number $n\in\mathbb N$ in question being the smallest one such that $f^{(n)}(x)\neq 0$. □

As a concrete application now of the Taylor formula, we have:

Theorem 11.30. *We have 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 well as 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 Taylor series, and in general as well, with $x\in(-1,1]$ needed for log.

Proof. There are several assertions here, the proofs being as follows:

(1) Regarding the Taylor series statements, we can use here the following formulae:

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

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

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

(2) Regarding now the convergence away from 0, we already know that this happens for e^x. In order to discuss sin and cos, we will need the Euler formula, namely:

$$e^{ix}=\cos x+i\sin x$$

To be more precise, this is a formula that we more or less established in chapter 7, with the promise to come back later to it, after learning calculus. So, let us set:

$$f(x)=\frac{\cos x+i\sin x}{e^{ix}}$$

The point is that we can compute the derivative of f, and we obtain:

$$\begin{aligned}f'(x)&=(e^{-ix}(\cos x+i\sin x))'\\&=-ie^{-ix}(\cos x+i\sin x)+e^{-ix}(-\sin x+i\cos x)\\&=e^{-ix}(-i\cos x+\sin x)+e^{-ix}(-\sin x+i\cos x)\\&=0\end{aligned}$$

We conclude from this that f is constant, equal to $f(0)=1$, as desired.

(3) Getting back now to our questions regarding sin and cos, we have the following computation, valid for any $x \in \mathbb{R}$, based on the usual formula of the exponential:

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

Now by comparing this with $e^{ix} = \cos x + i \sin x$, we obtain, for any $x \in \mathbb{R}$:

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

(4) Finally, in what regards the logarithm, we know that we have, as Taylor series:

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

By using the general theory of series from chapter 3, we can see that this series does not converge for $|x| > 1$, nor at $x = -1$. Thus, we are left with the question whether the above formula holds or not, at $x \in (-1, 1]$. And in order to prove that it is so, we must check one of the following formulae, with $\log(1+x)$ standing for the above series:

$$\exp(\log(1+x)) = 1 + x \quad , \quad \log(\exp(x)) = x$$

But this can be done indeed, with some patience, and we will leave the computations here, based on the binomial formula, as an instructive exercise. □

As another application of our Taylor formula technology, we have:

THEOREM 11.31. *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 $|t|<1$:

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

(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) General case, $p\in\mathbb{R}$. Many things can be said here, for instance with direct combinatorial proofs at $p\in\mathbb{Z}$, or even at $p\in\mathbb{Z}/2$, and skipping the discussion here, let us investigate right away the general 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. As for the general case, which reads $|t|<1$ with our normalization $x=1$ above, this follows from $(1+t)^p=\exp(p\log(1+t))$, using Theorem 11.30. □

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

Theorem 11.32. *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. 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 11.31 at $p=\pm1/2$ gives the formulae in the statement. □

11e. Exercises

Welcome to calculus, eventually, such a joy, and as calculus exercises, we have:

EXERCISE 11.33. *Clarify all the details in the proof of $(x^p)'=px^{p-1}$.*

EXERCISE 11.34. *Compute the derivatives of the remaining trigomometric functions.*

EXERCISE 11.35. *Further meditate on what we said, regarding the meaning of f''.*

EXERCISE 11.36. *Compute the second derivatives of all trigomometric functions.*

EXERCISE 11.37. *Clarify everything that we said, in relation with convex functions.*

EXERCISE 11.38. *Compute the third derivatives of all trigomometric functions.*

EXERCISE 11.39. *Learn more about the Taylor formula, and the remainder.*

EXERCISE 11.40. *Work out the Taylor formula for all trigomometric functions.*

As bonus exercise, set up an optimization business, with what you learned from here.

CHAPTER 12

Integrals

12a. 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 12.1. *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 12.2. *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 12.3. *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 12.1, 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 12.4. *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 12.2, 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 12.2. More on this later.

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

Proposition 12.5. *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 12.6. *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 12.1, because area means side times average height. Thus, we can formulate:

Theorem 12.7. *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 12.1, via some geometric thinking. Alternatively, this is something which certainly comes from Theorem 12.6. □

The point of view in Theorem 12.7 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 12.3, 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 12.7. The same goes for the results in Proposition 12.5, which become even more obvious with the viewpoint from Theorem 12.7. 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 12.3 (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 12.8. *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 12.7 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 12.9. *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 12.9:

Conclusion 12.10. *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 12.11. *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 12.8, 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 chapter, when systematically discussing probability theory. □

12b. 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 12.6 will be our favorite tool. Let us begin with some theory. We first have:

Theorem 12.12. *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 12.5 that the integrals behave well with respect to sums and multiplication by scalars. Along the same lines, we have:

Theorem 12.13. *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 10 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 12.14. *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 12.15. *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 has it 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 12.16. *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,

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

valid for any exponent $p\in\mathbb N$, and perhaps for some other $p\in\mathbb R$.

We will see later that this conjecture is indeed true, and with the exact details regarding the exponents $p\in\mathbb R-\mathbb N$ too. 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 12.17. *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. □

12c. Main theorems

The problem is now, what to do with what we have, namely Conjecture 12.16 and Theorem 12.17. Not obvious, so stuck, and time to ask the cat. And cat says:

CAT 12.18. *Summing the infinitesimals of the rate of change of the function should give you the global change of the function. Obvious.*

Which sounds quite odd, guess cat must be a reincarnation of Newton or Leibnitz, these gentlemen used to talk like that. And shall we trust such things, in the present modern age, where we have nuclear technology, internet and TikTok, and many more.

This being said, wait. There is suggestion to connect integrals and derivatives, and this is in fact what we have, coming from Conjecture 12.16 and Theorem 12.17, 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 12.19. *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 12.14. 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 12.14, 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 12.20. *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 12.19, and is in fact equivalent to it. Indeed, consider the following function:

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

By using Theorem 12.19 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 12.21. *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 12.22. *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 11, 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 12.21 in a more digest form, as follows:

Theorem 12.23. *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 12.21, or rather from the proof of Theorem 12.21, or even from chapter 11, for most of them, and the last assertion comes from the integration formula given after Definition 12.22. □

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

THEOREM 12.24. *We have the formula*

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

called integration by parts.

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

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

Indeed, with our convention for primitives, this gives the formula in the statement. □

It is then possible to pass to usual integrals, and we obtain a formula here as well, as follows, also called integration by parts, with the convention $[\varphi]_a^b=\varphi(b)-\varphi(a)$:

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

In practice, the most interesting case is that when fg vanishes on the boundary $\{a,b\}$ of our interval $[a,b]$, 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 12.25. *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 11, 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. □

Finally, as yet another interesting consequence of our technology, we have:

THEOREM 12.26. *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. □

And with this, good news, we have all the needed tools in our bag, for dealing with all sorts of integration problems. So, hang on, tough computations to come.

As a first application, we can compute all sorts of areas and volumes. Normally such things 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. We first have:

THEOREM 12.27. *The area of an ellipse, 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 ellipse, is $A=\pi ab$.

PROOF. The idea is that of cutting the ellipse 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 ellipse 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 ellipse 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. □

Still talking ellipses, in what regards the length things are quite tricky, as follows:

THEOREM 12.28. *The length of an ellipse, given by* $(x/a)^2+(y/b)^2=1$*, is*

$$L=4\int_0^{\pi/2}\sqrt{a^2\sin^2t+b^2\cos^2t}\,dt$$

and with this integral being generically not computable.

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

(1) To start with, in the case where our ellipse is a circle, say of radius R, the area and length are related by the formula $A=LR/2$, as we know well from chapter 6, from the "pizza" argument there, and in this case $A=\pi R^2$ coming from Theorem 12.27 translates into $L=2\pi R$, as it should. The problem, however, is that the pizza argument from chapter 6 obviously does not work for general ellipses, so we must find something else.

(2) So, what is the length of a curve $\gamma:[a,b]\to\mathbb{R}^N$? Good question, and in answer, a physicist would say that this is the quantity obtained by integrating the magnitude of the velocity vector over the curve, with respect to time. But this velocity vector is $\gamma'(t)$, having magnitude $||\gamma'(t)||$, so we are led to the following formula:

$$L(\gamma)=\int_a^b||\gamma'(t)||dt$$

(3) Regarding now mathematicians, these would say that the length of a curve is the following quantity, with $(t_1=a,t_2,\ldots,t_{n-1},t_n=b)$ being a uniform division of (a,b):

$$L(\gamma)=\lim_{n\to\infty}\sum_{i=1}^n||\gamma(t_i)-\gamma(t_{i-1})||$$

But, by using the fundamental theorem of calculus, we can write this as follows:

$$L(\gamma)=\lim_{n\to\infty}\sum_{i=1}^n\left|\left|\int_{t_{i-1}}^{t_i}\gamma'(t)dt\right|\right|$$

And the point now is that, by doing some standard analysis, that we will leave here as an instructive exercise, we are led to the formula in (2).

(4) Getting back now to the ellipses, we can compute their length, as follows:

$$\begin{aligned}L&=4\int_0^{\pi/2}\sqrt{\left(\frac{dx}{dt}\right)^2+\left(\frac{dy}{dt}\right)^2}\,dt\\&=4\int_0^{\pi/2}\sqrt{\left(\frac{da\cos t}{dt}\right)^2+\left(\frac{db\sin t}{dt}\right)^2}\,dt\\&=4\int_0^{\pi/2}\sqrt{a^2\sin^2t+b^2\cos^2t}\,dt\end{aligned}$$

(5) As for the last assertion, when $a=b=R$ we get of course $L=2\pi R$, as we should, but in general, when $a\neq b$, there is no trick for computing the above integral. □

Moving now to 3D, as an obvious challenge here, we can try to compute the area and volume of the sphere, and more generally of the ellipsoids. We have here:

Theorem 12.29. *The volume of the unit sphere in $\mathbb{R}^3$ is given by:*

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

More generally, the volume of an ellipsoid, $(x/a)^2+(y/b)^2+(z/c)^2=1$, is:

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

The area of the unit sphere is $A=4\pi$. For ellipsoids, this is generically not computable.

Proof. There are several things going on here, as follows:

(1) Let us first compute the volume of the ellipsoid, which at $a=b=c=1$ will give the volume of the unit sphere. The range of the first coordinate x is as follows:

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

Now when the first coordinate x is fixed, the other coordinates y,z vary on an ellipse, given by the equation $(y/b)^2+(z/c)^2=1-(x/a)^2$, which can be written as follows:

$$\left(\frac{y}{\beta}\right)^2+\left(\frac{z}{\gamma}\right)^2=1\quad:\quad\beta=b\sqrt{1-\left(\frac{x}{a}\right)^2}\ ,\ \gamma=c\sqrt{1-\left(\frac{x}{a}\right)^2}$$

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

$$A(x)=\pi\beta\gamma=\pi bc\left[1-\left(\frac{x}{a}\right)^2\right]$$

We conclude that the volume of the ellipsoid is given, as claimed, by:

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

(2) At $a=b=c=1$ we get $V=4\pi/3$, and this gives the area of the unit sphere too, because the "pizza" method from chapter 6 obviously applies, and gives:

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

(3) Finally, the last assertion, regarding the area of ellipsoids, is something quite informal, coming from the last assertion in Theorem 12.28, which was informal too. □

There are of course many other computations that can be done, along the same lines, and we will be back to this in Part IV, when discussing space geometry and calculus.

As yet another application of our integration theory, this time in relation with advanced calculus, and more specifically with the Taylor formula, we have:

Theorem 12.30. *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. □

12d. Some probability

As yet another application of the integration theory developed above, let us develop now some theoretical probability theory. We already know, from chapters 2-3, what discrete probability is, so time now to discuss the continuous and general case too.

In practice, when trying to axiomatize probability, in mathematical terms, things can be quite tricky. So, here comes our point, the definition saving us is as follows:

Definition 12.31. *A probability density is a function $\varphi : \mathbb{R} \to \mathbb{R}$ satisfying*

$$\varphi \geq 0 \qquad , \qquad \int_{\mathbb{R}} \varphi(x)dx = 1$$

with the convention that we allow Dirac masses, δ_x with $x \in \mathbb{R}$, as components of φ.

To be more precise, in what regards the convention at the end, which is something of physics flavor, this states that our density function $\varphi:\mathbb{R}\to\mathbb{R}$ must be a combination as follows, with $\psi:\mathbb{R}\to\mathbb{R}$ being a usual function, and with $\alpha_i, x_i\in\mathbb{R}$:

$$\varphi=\psi+\sum_i\alpha_i\delta_{x_i}$$

Assuming that x_i are distinct, and with the usual convention that the Dirac masses integrate up to 1, the conditions on our density function $\varphi:\mathbb{R}\to\mathbb{R}$ are as follows:

$$\psi\geq 0\quad,\quad\alpha_i\geq 0\quad,\quad\int_\mathbb{R}\psi(x)dx+\sum_i\alpha_i=1$$

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

In short, what we are proposing with Definition 12.31 is some sort of continuous generalization of basic probability theory, coming from coins, dice and cards, that we know well. Moving now ahead, let us formulate, as a continuation of Definition 12.31:

Definition 12.32. *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 about 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 12.33. *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 12.32 and Definition 12.33, 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 12.34 (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 $P(f\in[a,b])=\int_a^bd\mu(x)$. 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. Skipping some discussion here about Bernoulli and binomial laws, that we know from chapter 2, and getting to the Poisson laws from chapter 3, which are the main laws in discrete probability, we have:

Theorem 12.35. *The moments of the Poisson law 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 the Poisson law p_t of parameter $t>0$, 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. As for the general $t>0$ formula in the statement, its proof is similar, again by recurrence, and we will leave this as an instructive exercise. □

Regarding now continuous probability, the main laws here are the normal ones, but the problem is that their introduction and study is something quite advanced, requiring multivariable calculus and integration. So patience, we will be back to this later.

12e. Exercises

There is no true mathematics without integration, and as exercises, we have:

Exercise 12.36. *Further meditate on the integrable and non-integrable functions.*

Exercise 12.37. *Meditate at the random numbers, used for Monte Carlo.*

Exercise 12.38. *Design some good algorithms for producing random numbers.*

Exercise 12.39. *Learn when exactly integrals commute with limits, or sums.*

Exercise 12.40. *Fully compute $1^p+2^p+\ldots+N^p$, at small values of p.*

Exercise 12.41. *Do some computations of your own, for the length of the ellipses.*

Exercise 12.42. *Clarify what we said in relation with the Taylor formula remainder.*

Exercise 12.43. *Learn more about the Poisson laws, and their various properties.*

As bonus exercise, in relation with this, learn some systematic probability theory.

Part IV

Vectors

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

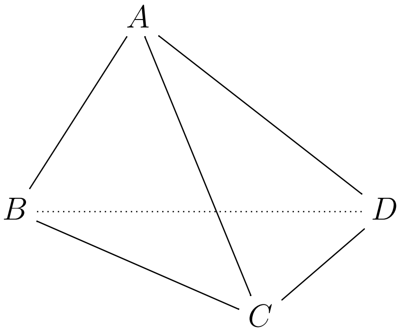
CHAPTER 13

Space geometry

13a. Space geometry

Welcome to space geometry, in the usual 3 dimensions that we live in, and in higher dimensions too. Many interesting things can be said here, in analogy with what we know from Part II about triangles. Let us start with something very basic, as follows:

THEOREM 13.1. *Any tetrahedron in three-dimensional space*



has a barycenter, lying $1/4-3/4$ *on the medians, uniting vertices to opposite barycenters.*

PROOF. This is something self-explanatory, which is best seen by using coordinates. Indeed, the barycenter of our tetrahedron can only be given by the following formula:

$$P=\frac{A+B+C+D}{4}$$

Now observe that this formula can be written in the following way:

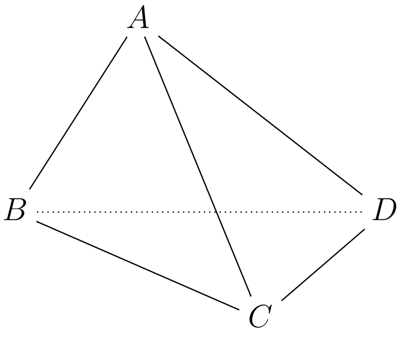
$$P=\frac{1}{4}\cdot A+\frac{3}{4}\cdot\frac{B+C+D}{3}$$

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

As in the case of the triangles, there is some further discussion here in relation with physical barycenters, when considering that the vertices, or edges, or faces, or the whole solid body itself, have mass. We will leave the study here as an interesting exercise.

Moving on, as a second basic result, again as for the triangles, we have:

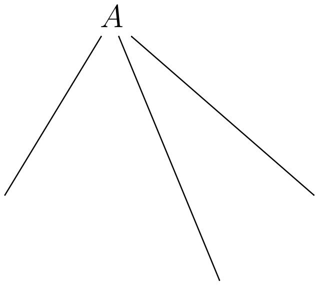
Theorem 13.2. *Any tetrahedron in three-dimensional space*



has an incenter, where the solid angle bisectors cross.

Proof. Again, this is something quite self-explanatory, and as in the case of the triangles, there are several ways of precisely stating and proving this, as follows:

(1) As a first approach, which is straightforward, we can base our study on the notion of solid angle bisector, as stated. Consider indeed a solid angle, as follows:



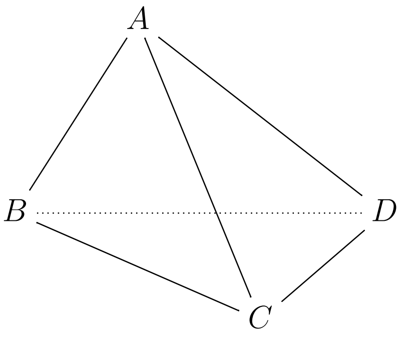
This solid angle has then a bisector, and with this best seen by fitting a sphere into our angle. Indeed, if O is the center of the sphere, AO is the angle bisector.

(2) Now the point is that the 4 angle bisectors cross indeed, and this can be seen for instance by interpreting each angle bisector as being an intersection of 3 planes, in the obvious way. Indeed, the total of 12 planes that we have must intersect.

(3) But the simplest is to argue that the incenter appears by fitting, or rather by inflating, a sphere inside our tetrahedron. Indeed, once our sphere is duly inflated, as to touch the faces, its center will be the incenter of our tetrahedron. $\square$

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

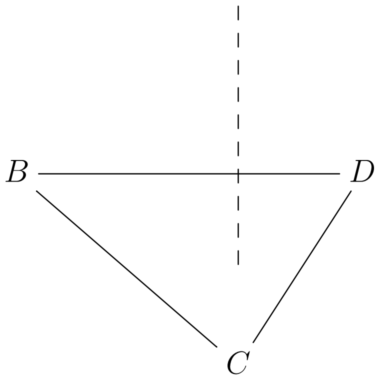
THEOREM 13.3. *Any tetrahedron in three-dimensional space*



has a circumcenter, where the perpendicular bisectors cross.

PROOF. Again, this is something quite self-explanatory, and as in the case of the triangles, there are several ways of precisely stating and proving this, as follows:

(1) As a first approach, which is straightforward, we can base our study on the notion of perpendicular bisector, as stated. Consider indeed a triangle in space:



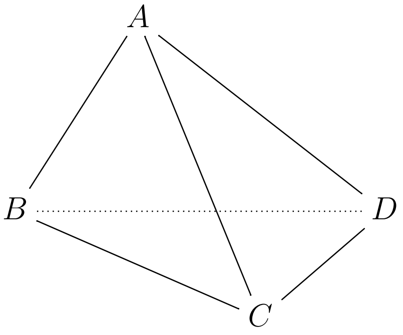
This triangle has then a perpendicular bisector, as indicated on the above picture, and with this best seen by fitting our triangle into a sphere. Indeed, if O is the center of the sphere, AO is the perpendicular bisector.

(2) Now the point is that the 4 perpendicular bisectors cross indeed, and this can be seen for instance by interpreting each perpendicular bisector as being an intersection of 3 planes, in the obvious way. Indeed, the total of 12 planes that we have must intersect.

(3) But the simplest is to argue that the circumcenter appears by fitting, or rather by deflating, a sphere outside our tetrahedron. Indeed, once our sphere is duly deflated, as to touch the vertices, its center will be the circumcenter of our tetrahedron. □

Regarding now the orthocenter, things here are quite complicated, as follows:

Theorem 13.4. *Under suitable assumptions, the tetrahedra in 3D space*

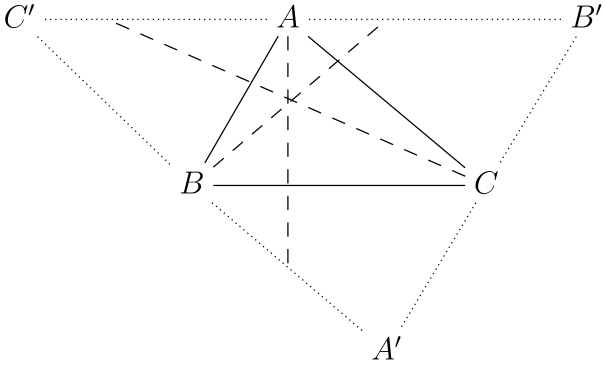


have an orthocenter, where the altitudes cross.

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

(1) To start with, the altitudes of a tetrahedron do not cross, in general. We will leave some thinking here, and the construction of counterexamples, as an exercise.

(2) Along the same lines, but a bit more philosophically, let us look at the 2-dimensional proof, of the existence of the orthocenter. The trick and picture were as follows:



But such things won't work in three dimensions, somehow for obvious reasons, and again, we will leave some thinking here as an instructive exercise.

(3) Getting now to what can be done, as to have some theory and results going on, following Monge and others, the idea is that we can talk about orthocentric tetrahedra. Consider indeed a tetrahedron whose opposite edges are orthogonal:

$$AB \perp CD \quad , \quad AC \perp BD \quad , \quad AD \perp BC$$

In this situation the altitudes cross, and their intersection, the orthocenter, coincides with the Monge point, appearing as the intersection of the 6 midplanes, which pass through the middle of each of the 6 edges, and are orthogonal to the opposite edge. □

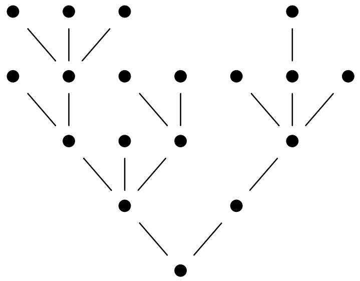
13b. Graphs, polyhedra

Switching a bit topics, let us discuss now, still in relation with space geometry questions, the graphs. Here is a general principle, regarding graphs and their geometry:

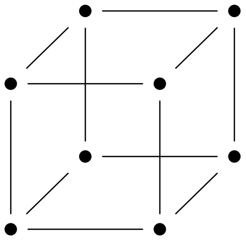
Principle 13.5. *Graphs fall into three classes:*

1. *Trees and other graphs which can be drawn without crossings are good.*
2. *If we can still do this, but on a torus, the graph is bad.*
3. *And the rest is evil.*

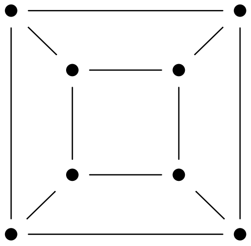
Here the fact that trees are indeed planar is obvious, and as an illustration, here is some sort of "random" tree, which is clearly planar, no question about this:



Of course, there are many other interesting examples of planar graphs. Consider for instance the cube graph, which is one of the most important graphs, as follows:

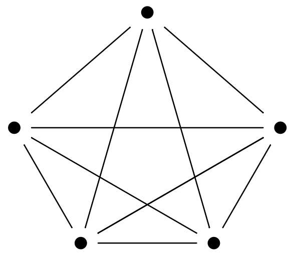


At the first glance, this graph does not look very planar. However, after thinking a bit, we can draw it as follows, making it clear that this graph is planar:



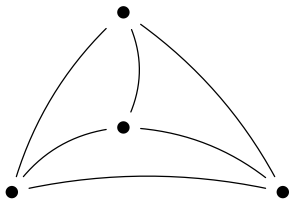
In order to find now some basic examples of non-planar graphs, the simplest method is by looking at the complete graphs, or simplices. This leads to the following result:

Proposition 13.6. *When looking at simplices, the segment K_2, the triangle K_3 and the tetrahedron K_4 are planar. However, the next simplex K_5, namely*

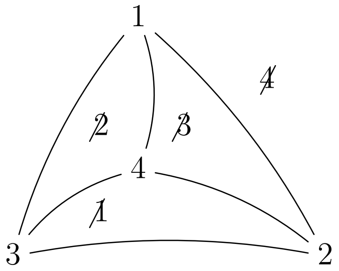


is not planar. Nor are the higher simplices, K_N with $N \geq 6$, planar.

Proof. To start with, the graphs K_2, K_3, K_4 are obviously planar. Regarding now the non-planarity of K_5, let us first have to draw its subgraph K_4 in a planar way:



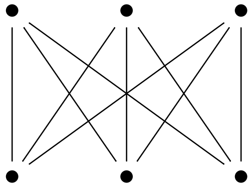
But with this in hand, it is clear that there is no room in the plane for our 5th vertex, as to avoid crossings. Indeed, we have 4 possible regions in the plane for this 5th vertex, and each of them is forbidden by the edge towards a certain vertex, as follows:



Finally, the fact that the higher simplices, K_N with $N \geq 6$, are not planar either follows from the fact that their subgraphs K_5 are not planar. □

In order to find some further examples of non-planar graphs, we can look as well at the bipartite simplices, and we are led to the following result:

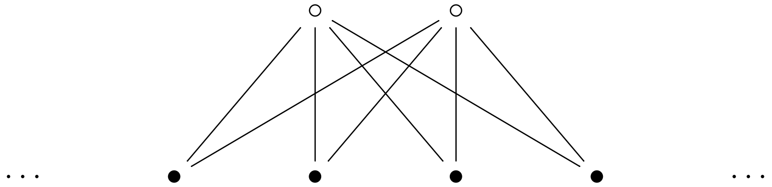
Proposition 13.7. *When looking at bipartite simplices, the square $K_{2,2}$ is planar, and so are all the graphs $K_{2,N}$. However, the next such graph, namely $K_{3,3}$,*



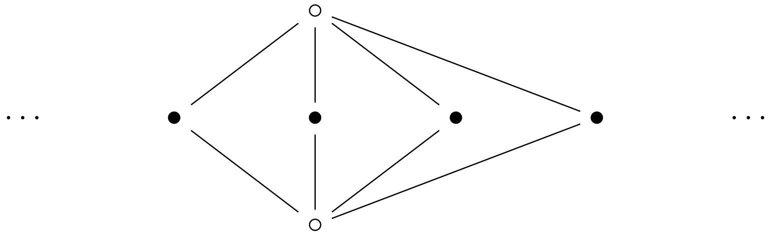
called "utility graph" is not planar. Nor are planar the graphs $K_{M,N}$, for any $M,N\geq 3$.

Proof. Again, this is something elementary and intuitive, as follows:

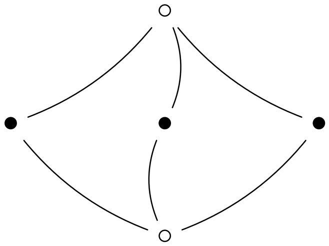
(1) The square $K_{2,2}$ is obviously planar. Regarding now the bipartite simplex $K_{2,N}$ with $N\geq 2$ arbitrary, this graph looks at follows, with N vertices in the lower row:



But this graph is planar too, because we can draw it in the following way:



(2) Regarding now $K_{3,3}$, as before with the simplex K_5, the result here is quite clear by thinking a bit, and drawing pictures. To be more precise, reasoning by contradiction, we first have to draw its subgraph $K_{2,3}$ in a planar way, and this is done as follows:



But now, it is clear that there is no room in the plane for our 6th vertex, as to avoid crossings. Finally, $K_{M,N}$ with $M,N\geq 3$ is not planar either, because it contains $K_{3,3}$. □

In general now, as a first main result about the planar graphs, we have:

Theorem 13.8. *The fact that a graph X is non-planar can be checked as follows:*

(1) *Kuratowski criterion: X contains a subdivision of K_5 or $K_{3,3}$.*
(2) *Wagner criterion: X has a minor of type K_5 or $K_{3,3}$.*

Proof. This is obviously something quite powerful, when thinking at the potential applications, and non-trivial to prove as well, the idea being as follows:

(1) Regarding the Kuratowski criterion, the convention is that "subdivision" means graph obtained by inserting vertices into edges, e.g. replacing $\bullet - \bullet$ with $\bullet - \bullet - \bullet$.

(2) Regarding the Wagner criterion, the convention there is that "minor" means graph obtained by contracting certain edges into vertices.

(3) Regarding now the proofs, the Kuratowski and Wagner criteria are more or less equivalent, and their proof is via standard, although long, recurrence methods. □

As a second fundamental result now about the planar graphs, we have:

Theorem 13.9. *For a connected planar graph we have the Euler formula*

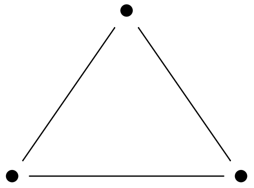
$$v - e + f = 2$$

with v, e, f being the number of vertices, edges and faces.

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

(1) Regarding the precise statement, given a connected planar graph, drawn in a planar way, without crossings, we can certainly talk about the numbers v and e, as for any graph, and also about f, as being the number of faces that our graph has, in our picture, with these including by definition the outer face too, the one going to ∞. With these conventions, the claim is that the Euler formula $v - e + f = 2$ holds indeed.

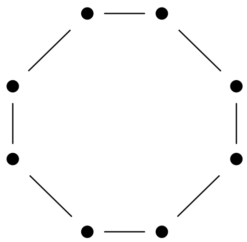
(2) As a first illustration for how this formula works, consider a triangle:



Here we have $v = e = 3$, and $f = 2$, with this accounting for the interior and exterior, and we conclude that the Euler formula holds indeed in this case, as follows:

$$3 - 3 + 2 = 2$$

(3) More generally now, let us look at an arbitrary N-gon graph:



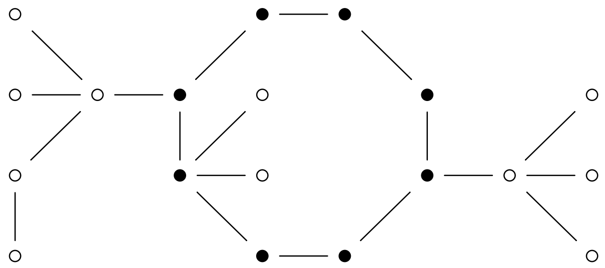
Then, for this graph, the Euler formula holds indeed, as follows:

$$N - N + 2 = 2$$

(4) With these examples discussed, let us look now for a proof. The idea will be to proceed by recurrence on the number of faces f. And here, as a first observation, the result holds at $f = 1$, where our graph must be planar and without cycles, and so must be a tree. Indeed, with N being the number of vertices, the Euler formula holds, as:

$$N - (N - 1) + 1 = 2$$

(5) At $f = 2$ now, our graph must be an N-gon as above, but with some trees allowed to grow from the vertices, with an illustrating example here being as follows:

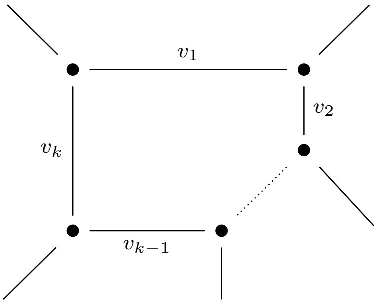


But here we can argue, again based on the fact that for a rooted tree, the non-root vertices are in obvious bijection with the edges, that removing all these trees won't change the problem. So, we are left with the problem for the N-gon, already solved in (3).

(6) And so on, the idea being that we can first remove all the trees, by using the argument in (5), and then we are left with some sort of agglomeration of N-gons, for which we can check the Euler formula directly, a bit as in (3), or by recurrence.

(7) To be more precise, let us try to do the recurrence on the number of faces f. For this purpose, consider one of the faces of our graph, which looks as follows, with v_i

denoting the number of vertices on each side, with the endpoints excluded:



(8) Now let us collapse this chosen face to a single point, in the obvious way. In this process, the total number of vertices v, edges e, and faces f, evolves as follows:

$$v \to v - k + 1 - \sum v_i$$

$$e \to e - \sum (v_i + 1)$$

$$f \to f - 1$$

Thus, in this process, the Euler quantity $v - e + f$ evolves as follows:

$$\begin{aligned} v - e + f &\to v - k + 1 - \sum v_i - e + \sum (v_i + 1) + f - 1 \\ &= v - k + 1 - \sum v_i - e + \sum v_i + k + f - 1 \\ &= v - e + f \end{aligned}$$

So, done with the recurrence, and the Euler formula is proved. □

As a famous application, or rather version, of the Euler formula, let us record:

Proposition 13.10. *For a convex polyhedron we have the Euler formula*

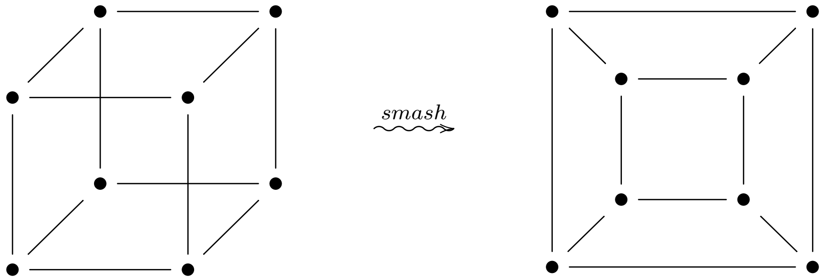
$$v - e + f = 2$$

with v, e, f being the number of vertices, edges and faces.

Proof. This is more or less the same thing as Theorem 13.9, save for getting rid of the internal trees of the planar graph there, the idea being as follows:

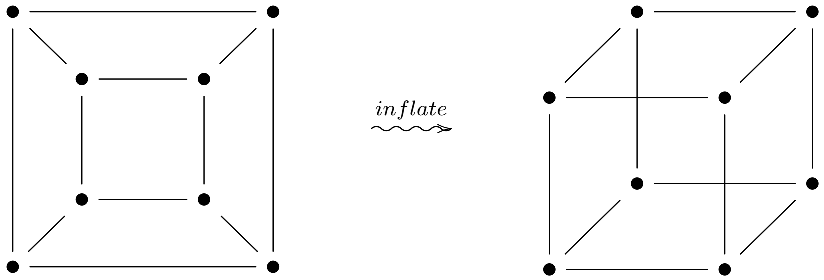
(1) In one sense, consider a convex polyhedron P. We can then enlarge one face, as much as needed, and then smash our polyhedron with a big hammer, as to get a planar

graph X. As an illustration, here is how this method works, for a cube:



But, in this process, each of the numbers v, e, f stays the same, so we get the Euler formula for P, as a consequence of the Euler formula for X, from Theorem 13.9.

(2) Conversely, consider a connected planar graph X. Then, save for getting rid of the internal trees, as explained in the proof of Theorem 13.9, we can assume that we are dealing with an agglomeration of N-gons, again as explained in the proof of Theorem 13.9. But now, we can inflate our graph as to obtain a convex polyhedron P, as follows:



Again, in this process, each of the numbers v, e, f will stay the same, and so we get the Euler formula for X, as a consequence of the Euler formula for P. □

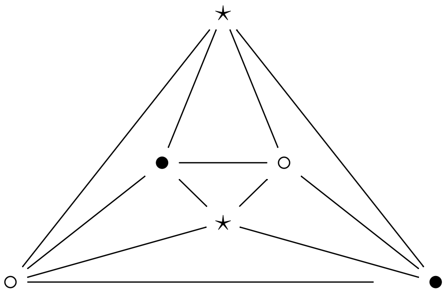
Summarizing, Euler formula understood, but as a matter of making sure that we didn't mess up anything with our mathematics, let us do some direct checks as well:

Proposition 13.11. *The Euler formula $v - e + f = 2$ holds indeed for the five possible regular polyhedra, as follows:*

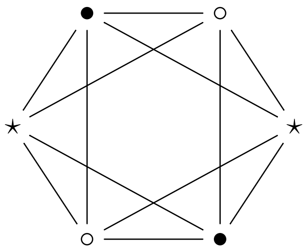
1. *Tetrahedron:* $4 - 6 + 4 = 2$.
2. *Cube:* $8 - 12 + 6 = 2$.
3. *Octahedron:* $6 - 12 + 8 = 2$.
4. *Dodecahedron:* $20 - 30 + 12 = 2$.
5. *Isocahedron:* $12 - 30 + 20 = 2$.

Proof. The figures in the statement are certainly the good ones for the tetrahedron and the cube. Regarding now the octahedron, again the figures are the good ones, by thinking in 3D, but as an interesting exercise for us, which is illustrating for the above, let us attempt to find a nice way of drawing the corresponding graph:

(1) To start with, the "smashing" method from the proof of Proposition 13.10 provides us with a graph which is certainly planar, but which, even worse than before for the cube, sort of misses the whole point with the 3D octahedron, its symmetries, and so on:



(2) Much nicer, instead, is the following picture, which still basically misses the 3D beauty of the octahedron, but at least reveals some of its symmetries:



In short, you get the point, quite subjective all this, and as a conclusion, drawing graphs in an appropriate way remains an art. As for the dodecahedron and isocahedron, exercise here for you, and if failing, take some drawing classes. Math is not everything. □

As a third main result now about the planar graphs, we have:

Theorem 13.12. *Any planar graph has the following properties:*

(1) *It is vertex* 4*-colorable.*
(2) *It is a* 4*-partite graph.*

Proof. Heavy theorem that we have here, and exercise for you to learn more about this, and why not, come with a simpler proof, of your own, for this theorem. □

And we will stop here with graphs. As you can see, just by looking at our main results, Theorems 13.8, 13.9 and 13.12, what can be said about them varies wildly in difficulty, and is quite unpredictable. Quite fascinating all this, hope you agree with me.

13c. Vector products

Ready for some physics? That takes place in 3D, and our knowledge accumulated so far can be very useful, in understanding how basic physics works. However, before getting started with this, we will need one more mathematical notion, as follows:

Definition 13.13. *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:

$$\begin{matrix} & \uparrow_{x\times y} \\ \leftarrow_x & \\ \swarrow_y & \end{matrix}$$

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 13.14. *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 13.15. *Assume that a 3D body rotates along an axis, with angular speed w. For a fixed point of the body, with position vector x, the usual 3D speed is*

$$v = \omega \times x$$

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

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

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

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

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

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

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

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

$$\begin{aligned}V &= \dot{x}_1u_1 + \dot{x}_2u_2 + \dot{x}_3u_3 + x_1\dot{u}_1 + x_2\dot{u}_2 + x_3\dot{u}_3\\ &= \dot{x} + (x_1 \cdot \omega \times u_1 + x_2 \cdot \omega \times u_2 + x_3 \cdot \omega \times u_3)\\ &= \dot{x} + w \times (x_1u_1 + x_2u_2 + x_3u_3)\\ &= \dot{x} + \omega \times x\end{aligned}$$

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

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

Theorem 13.16. *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 13.15, 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 13.17. *The acceleration of an object m subject to a force F is given by*

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

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

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

We refer to any standard undergraduate mechanics book, such as Feynman [31], Kibble [53] or Taylor [89] for more on the above, including various numerics on what happens here on Earth, the Foucault pendulum, history of all this, and many other things. Let us just mention here, as a basic illustration for all this, that a rock dropped from 100m deviates about 1cm from its intended target, due to the formula in Theorem 13.17.

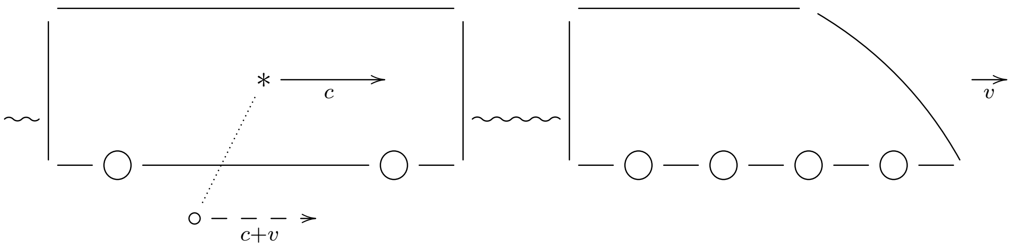
13d. Speed addition

Time now for some Einstein relativity theory, or perhaps foundational mathematics. Based on experiments by Fizeau, then Michelson-Morley and others, and on some physics by Maxwell and Lorentz too, Einstein came upon the following principles:

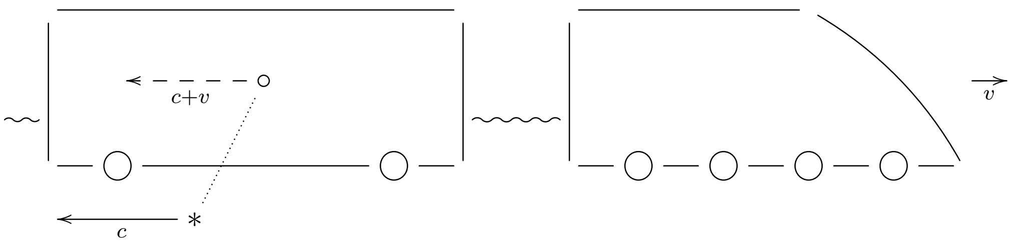
Fact 13.18 (Einstein principles). *The following happen:*

(1) *Light travels in vacuum at a finite speed, $c < \infty$.*
(2) *This speed c is the same for all inertial observers.*
(3) *In non-vacuum, the light speed is lower, $v < c$.*
(4) *Nothing can travel faster than light, $v \not> c$.*

The point now is that, obviously, something is wrong here. Indeed, assuming for instance that we have a train, running in vacuum at speed $v > 0$, and someone on board lights a flashlight $*$ towards the locomotive, an observer $\circ$ on the ground will see the light traveling at speed $c + v > c$, which is a contradiction:



Equivalently, with the same train running, in vacuum at speed $v > 0$, if the observer on the ground lights a flashlight $*$ towards the back of the train, then viewed from the train, that light will travel at speed $c + v > c$, which is a contradiction again:



Summarizing, Fact 13.18, while physically true, implies $c + v = c$, so contradicts classical mechanics, which needs a fix. In the classical case, to start with, we have:

Proposition 13.19. *The classical speeds add according to the Galileo formula*

$$v_{AC} = v_{AB} + v_{BC}$$

where v_{AB} denotes the relative speed of A with respect to B.

Proof. This is clear indeed from the definition of speed, and very intuitive. □

In order to find the fix, let us first discuss the 1D case. We will use two tricks. First, let us forget about absolute speeds, and talk about relative speeds only. In this case we are allowed to sum only quantities of type v_{AB}, v_{BC}, and we denote by $v_{AB} +_g v_{BC}$ the corresponding sum v_{AC}. With this convention, the Galileo formula becomes:

$$u +_g v = u + v$$

As a second trick now, observe that this Galileo formula holds in any system of units. In order now to deal with our problems, basically involving high speeds, it is convenient to change the system of units, as to have $c = 1$. With this convention our $c + v = c$ problem becomes $1 + v = 1$, and the solution to it is quite obvious, as follows:

Proposition 13.20. *If we define the Einstein sum $+_e$ of relative speeds by*

$$u +_e v = \frac{u+v}{1+uv}$$

in $c = 1$ units, then we have the formula $1 +_e v = 1$, valid for any v.

Proof. This is obvious indeed from our definition of $+_e$, because if we plug in $u = 1$ in the above formula, we obtain as result:

$$1 +_e v = \frac{1+v}{1+v} = 1$$

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

Summarizing, we have solved our problem. In order now to formulate a final result, we must do some reverse engineering, by waiving the above two tricks. First, by getting back to usual units, $v \to v/c$, our new addition formula becomes:

$$\frac{u}{c} +_e \frac{v}{c} = \frac{\frac{u}{c} + \frac{v}{c}}{1 + \frac{u}{c} \cdot \frac{v}{c}}$$

By multiplying by c, we can write this formula in a better way, as follows:

$$u +_e v = \frac{u+v}{1+uv/c^2}$$

In order now to finish, it remains to go back to absolute speeds, and we are led to:

Theorem 13.21. *If we sum the speeds according to the Einstein formula*

$$v_{AC} = \frac{v_{AB} + v_{BC}}{1 + v_{AB}v_{BC}/c^2}$$

then the Galileo formula still holds, approximately, for low speeds

$$v_{AC} \simeq v_{AB} + v_{BC}$$

and if we have $v_{AB} = c$ or $v_{BC} = c$, the resulting sum is $v_{AC} = c$.

Proof. This is indeed self-explanatory, coming from the above discussion. □

All the above is very nice, but remember, takes place in 1D. So, time now to get seriously to work, and see what all this becomes in 3D. We have here:

Question 13.22. *What is the correct analogue of the Einstein summation formula*

$$u +_e v = \frac{u+v}{1+uv}$$

in $c = 1$ units, in 2 and 3 dimensions?

In order to discuss this question, let us attempt to construct $u +_e v$ in arbitrary dimensions, just by using our common sense and intuition. When the vectors $u, v \in \mathbb{R}^N$ are proportional, we are basically in 1D, and so our addition formula must satisfy:

$$u \sim v \implies u +_e v = \frac{u+v}{1+ < u, v >}$$

However, the formula on the right will not work as such in general, for arbitrary speeds $u, v \in \mathbb{R}^N$, and this because we have, as natural requirements for our operation:

$$||u|| = 1 \implies u +_e v = u \quad , \quad ||v|| = 1 \implies u +_e v = v$$

Summarizing, our $u \sim v$ formula above is not bad, as a start, but we must add a correction term to it, for the above requirements to be satisfied, and of course with the correction term vanishing when $u \sim v$. So, we are led to a math puzzle:

Puzzle 13.23. *What vanishes when $u \sim v$, and then how to correctly define*

$$u +_e v = \frac{u+v+\gamma_{uv}}{1+ < u, v >}$$

as for the correction term γ_{uv} to vanish when $u \sim v$?

But the solution to the first question is well-known in 3D. Indeed, here we can use the vector product $u \times v$, that we met before, which notoriously satisfies:

$$u \sim v \implies u \times v = 0$$

Thus, our correction term γ_{uv} must be something containing $w = u \times v$, which vanishes when this vector w vanishes, and in addition arranged such that $||u|| = 1$ produces a simplification, with $u +_e v = u$ as end result, and with $||v|| = 1$ producing a simplification too, with $u +_e v = v$ as end result. Thus, our vector calculus puzzle becomes:

Puzzle 13.24. *How to correctly define the Einstein summation in 3 dimensions,*

$$u +_e v = \frac{u+v+\gamma_{uvw}}{1+ < u, v >}$$

with $w = u \times v$, in such a way as for the correction term γ_{uvw} to satisfy

$$w = 0 \implies \gamma_{uvw} = 0$$

and also such that $||u|| = 1 \implies u +_e v = u$, and $||v|| = 1 \implies u +_e v = v$?

In order to solve this latter puzzle, we must "transport" the vector w to the plane spanned by u, v. But this is simplest done by taking the vector product with any vector in this plane, and so as a reasonable candidate for our correction term, we have:

$$\gamma_{uvw} = (\alpha u + \beta v) \times w$$

Here $\alpha, \beta \in \mathbb{R}$ are some scalars to be determined, but let us take a break, and leave the computations for later. We did some good work, time to update our puzzle:

PUZZLE 13.25. *How to define the Einstein summation in* 3 *dimensions,*

$$u +_e v = \frac{u + v + \gamma_{uvw}}{1 + < u, v >}$$

with the correction term being of the following form, with $w = u \times v$, and $\alpha, \beta \in \mathbb{R}$,

$$\gamma_{uvw} = (\alpha u + \beta v) \times w$$

in such a way as to have $||u|| = 1 \implies u +_e v = u$, and $||v|| = 1 \implies u +_e v = v$?

In order to investigate what happens when $||u|| = 1$ or $||v|| = 1$, we must compute the vector products $u \times w$ and $v \times w$. So, pausing now our study for consulting the vector calculus database, and then coming back, here is the formula that we need:

$$u \times (u \times v) = < u, v > u - < u, u > v$$

As for the formula of $v \times w$, that I forgot to record, we can recover it from the one above of $u \times w$, by using the basic properties of the vector products, as follows:

$$\begin{aligned} v \times (u \times v) &= -v \times (v \times u) \\ &= -(< v, u > v - < v, v > u) \\ &= < v, v > u - < u, v > v \end{aligned}$$

With these formulae in hand, we can now compute the correction term, with the result here, that we will need several times in what comes next, being as follows:

PROPOSITION 13.26. *The correction term $\gamma_{uvw} = (\alpha u + \beta v) \times w$ is given by*

$$\gamma_{uvw} = (\alpha < u, v > + \beta < v, v >)u - (\alpha < u, u > + \beta < u, v >)v$$

for any values of the scalars $\alpha, \beta \in \mathbb{R}$.

PROOF. According to our vector product formulae above, we have:

$$\begin{aligned} \gamma_{uvw} &= (\alpha u + \beta v) \times w \\ &= \alpha(< u, v > u - < u, u > v) + \beta(< v, v > u - < u, v > v) \\ &= (\alpha < u, v > + \beta < v, v >)u - (\alpha < u, u > + \beta < u, v >)v \end{aligned}$$

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

Time now to get into the real thing, see what happens when $||u||=1$ and $||v||=1$, if we can get indeed $u+_e v=u$ and $u+_e v=v$. It is convenient here to do some reverse engineering. Regarding the first desired formula, namely $u+_e v=u$, we have:

$$\begin{aligned}u+_e v=u &\iff u+v+\gamma_{uvw}=(1+<u,v>)u\\ &\iff \gamma_{uvw}=<u,v>u-v\\ &\iff \alpha=1,\ \beta=0,\ ||u||=1\end{aligned}$$

Thus, with the parameter choice $\alpha=1,\beta=0$, we will have, as desired:

$$||u||=1\implies u+_e v=u$$

In what regards now the second desired formula, namely $u+_e v=v$, here the computation is almost identical, save for a sign switch, which after some thinking comes from our choice $w=u\times v$ instead of $w=v\times u$, clearly favoring u, as follows:

$$\begin{aligned}u+_e v=v &\iff u+v+\gamma_{uvw}=(1+<u,v>)v\\ &\iff \gamma_{uvw}=-u+<u,v>v\\ &\iff \alpha=0,\ \beta=-1,\ ||v||=1\end{aligned}$$

Thus, with the parameter choice $\alpha=0,\beta=-1$, we will have, as desired:

$$||v||=1\implies u+_e v=v$$

All this is mixed news, because we managed to solve both our problems, at $||u||=1$ and at $||v||=1$, but our solutions are different. So, time to breathe, decide that we did enough interesting work for the day, and formulate our conclusion as follows:

Proposition 13.27. *When defining the Einstein speed summation in* 3D *as*

$$u+_e v=\frac{u+v+u\times(u\times v)}{1+<u,v>}$$

in $c=1$ units, the following happen:

1. *When $u\sim v$, we recover the previous* 1D *formula.*
2. *When $||u||=1$, speed of light, we have $u+_e v=u$.*
3. *However, $||v||=1$ does not imply $u+_e v=v$.*
4. *Also, the formula $u+_e v=v+_e u$ fails.*

Proof. Here (1) and (2) follow from the above discussion, with the following choice for the correction term, by favoring the $||u||=1$ problem over the $||v||=1$ one:

$$\gamma_{uvw}=u\times w$$

As for the last two assertions, (3) and (4), these are also clear from the above discussion, coming from the obvious lack of symmetry of our summation formula. □

Looking now at Proposition 13.27 from an abstract, mathematical perspective, there are still many things missing from there, which can be summarized as follows:

QUESTION 13.28. *Can we fine-tune the Einstein speed summation in* 3D *into*

$$u +_e v = \frac{u + v + \lambda \cdot u \times (u \times v)}{1 + < u, v >}$$

with $\lambda \in \mathbb{R}$, *chosen such that* $||u|| = 1 \implies \lambda = 1$, *as to have:*

(1) $||u||, ||v|| < 1 \implies ||u +_e v|| < 1$.
(2) $||v|| = 1 \implies ||u +_e v|| = 1$.

Obviously, as simplest answer, λ must be some well-chosen function of $||u||$, or rather of $||u||^2$, because it is always better to use square norms, when possible. But then, with this idea in mind, after a few computations, we are led to the following solution:

THEOREM 13.29. *When defining the Einstein speed summation in* 3D *as*

$$u +_e v = \frac{1}{1 + < u, v >}\left(u + v + \frac{u \times (u \times v)}{1 + \sqrt{1 - ||u||^2}}\right)$$

in $c = 1$ *units, the following happen:*

(1) *When* $u \sim v$, *we recover the previous* 1D *formula.*
(2) *We have* $||u||, ||v|| < 1 \implies ||u +_e v|| < 1$.
(3) *When* $||u|| = 1$, *we have* $u +_e v = u$.
(4) *When* $||v|| = 1$, *we have* $||u +_e v|| = 1$.
(5) *However,* $||v|| = 1$ *does not imply* $u +_e v = v$.
(6) *Also, the formula* $u +_e v = v +_e u$ *fails.*

PROOF. This follows from the above discussion, as follows:

(1) This is something that we know from Proposition 13.27.

(2) In order to simplify notation, let us set $\delta = \sqrt{1 - ||u||^2}$, which is the inverse of the quantity $\gamma = 1/\sqrt{1 - ||u||^2}$. With this convention, we have:

$$\begin{aligned} u +_e v &= \frac{1}{1 + < u, v >}\left(u + v + \frac{< u, v > u - ||u||^2 v}{1 + \delta}\right) \\ &= \frac{(1 + \delta + < u, v >)u + (1 + \delta - ||u||^2)v}{(1 + < u, v >)(1 + \delta)} \end{aligned}$$

Taking now the squared norm and computing gives the following formula:

$$||u +_e v||^2 = \frac{(1 + \delta)^2 ||u + v||^2 + (||u||^2 - 2(1 + \delta))(||u||^2 ||v||^2 - < u, v >^2)}{(1 + < u, v >)^2 (1 + \delta)^2}$$

But this formula can be further processed by using $\delta = \sqrt{1 - ||u||^2}$, and by navigating through the various quantities which appear, we obtain, as a final product:

$$||u +_e v||^2 = \frac{||u + v||^2 - ||u||^2 ||v||^2 + < u, v >^2}{(1 + < u, v >)^2}$$

But this type of formula is exactly what we need, for what we want to do. Indeed, by assuming $||u||,||v||<1$, we have the following estimate:

$$\begin{aligned}||u+_e v||^2<1 &\iff ||u+v||^2-||u||^2||v||^2+<u,v>^2<(1+<u,v>)^2\\ &\iff ||u+v||^2-||u||^2||v||^2<1+2<u,v>\\ &\iff ||u||^2+||v||^2-||u||^2||v||^2<1\\ &\iff (1-||u||^2)(1-||v||^2)>0\end{aligned}$$

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

(3) This is something that we know from Proposition 13.27.

(4) This comes from the squared norm formula established in the proof of (2) above, because when assuming $||v||=1$, we obtain:

$$\begin{aligned}||u+_e v||^2 &= \frac{||u+v||^2-||u||^2+<u,v>^2}{(1+<u,v>)^2}\\ &= \frac{||u||^2+1+2<u,v>-||u||^2+<u,v>^2}{(1+<u,v>)^2}\\ &= \frac{1+2<u,v>+<u,v>^2}{(1+<u,v>)^2}\\ &= 1\end{aligned}$$

(5) This is clear, from the obvious lack of symmetry of our formula.

(6) This is again clear, from the obvious lack of symmetry of our formula. □

That was nice, all this mathematics, and hope you're still with me. And good news, the formula in Theorem 13.29 is the good one, confirmed by experimental physics.

13e. Exercises

Space geometry is a quite tricky business, and as exercises here, we have:

Exercise 13.30. *Learn more about the mathematics of tetrahedra.*

Exercise 13.31. *Learn also about solid angles, and their applications.*

Exercise 13.32. *Learn more about planar graphs, and their properties.*

Exercise 13.33. *Learn also about toral graphs, such as the Petersen graph.*

Exercise 13.34. *Practice with regular polyhedra, on problems of your choice.*

Exercise 13.35. *Get to know everything about the Coriolis force.*

Exercise 13.36. *Learn more about Fizeau, Einstein, and 1D relativity.*

Exercise 13.37. *Learn more about 3D relativity, and its various consequences.*

As bonus exercise, find and read an old-style, dusty 3D geometry book.

CHAPTER 14

Linear algebra

14a. Linear maps

We have already met vectors, linear maps and matrices in this book, and time now to have a more systematic look at this. As a starting point, we have the following fact:

Definition 14.1. *The points $x\in\mathbb R^N$ can be represented as vectors*

$$x=\begin{pmatrix}x_1\\ \vdots\\ x_N\end{pmatrix}$$

and are subject to the addition and multiplication by scalars operations

$$x+y=\begin{pmatrix}x_1+y_1\\ \vdots\\ x_N+y_N\end{pmatrix}\quad,\quad\lambda x=\begin{pmatrix}\lambda x_1\\ \vdots\\ \lambda x_N\end{pmatrix}$$

geometrically corresponding to forming a parallelogram, and dilating by λ.

To be more precise, in what regards the notation for vectors, for some reasons that will become clear in a moment, we prefer the above vertical notation to the horizontal one. As for the two operations on vectors, this is something that we discussed in some detail in chapter 7, in two dimensions, and in general N dimensions the situation is similar.

As explained in chapter 7, and then in chapter 8, again in two dimensions, most of the vector mathematics comes from the above two operations on vectors. In general N dimensions the situation is quite similar, and this suggests the following definition:

Definition 14.2. *A map $f:\mathbb R^N\to\mathbb R^M$ is called linear when it satisfies:*

$$f(x+y)=f(x)+f(y)\quad,\quad f(\lambda x)=\lambda f(x)$$

That is, f must behave well with respect to the basic operations on vectors.

As a first question that you might have, why calling linear such beasts? In answer, observe that the above linearity conditions can be merged into one, as follows:

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

But this latter condition tells us that our map f must map lines into lines, or rather points moving on lines to points moving on lines, as follows:

$$f:\ [x-y]\ \ \rightsquigarrow\ [f(x)-f(y)]$$

Thus, the terminology is justified. In what regards now the mathematics of the linear maps, again by following the material from chapters 7 and 8, we have:

THEOREM 14.3. *The linear maps $f:\mathbb{R}^N\to\mathbb{R}^M$ are in correspondence with the matrices $A\in M_{M\times N}(\mathbb{R})$, with the linear map associated to such a matrix being*

$$f(x)=Ax$$

and with the matrix associated to a linear map being given by the formula

$$A_{ij}=<f(e_j),e_i>$$

with $\{e_i\}$ being the standard bases, and $<x,y>=\sum_i x_iy_i$ being the scalar product.

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

(1) According to Definition 14.2, a linear map $f:\mathbb{R}^N\to\mathbb{R}^M$ must send a vector $x\in\mathbb{R}^N$ to a certain vector $f(x)\in\mathbb{R}^M$, all whose components are linear combinations of the components of x. Thus, we can write, for certain numbers $A_{ij}\in\mathbb{R}$:

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

Now observe that the parameters $A_{ij}\in\mathbb{R}$ can be regarded as being the entries of a rectangular matrix $A\in M_{M\times N}(\mathbb{R})$. Thus, we have a correspondence, as follows:

$$f\ \leftrightarrow\ A$$

(2) In order to understand now how this correspondence works, let us make the following convention, for the multiplication of the rectangular matrices:

$$(AB)_{ij}=\sum_k A_{ik}B_{kj}$$

To be more precise, we assume here $A\in M_{M\times N}(\mathbb{R})$ and $B\in M_{N\times K}(\mathbb{R})$, and we obtain in this way a certain matrix $AB\in M_{M\times K}(\mathbb{R})$. Now observe that in the case $K=1$, and by omitting the corresponding trivial index j, our multiplication formula reads:

$$(AB)_i=\sum_k A_{ik}B_k$$

But this is quite similar to what we have in (1), with the formula there taking the following form, which is the one in the statement, with our present conventions:

$$f(x)=Ax$$

(3) Regarding now the second assertion, with $f(x) = Ax$ as above, if we denote by $e_1, \ldots, e_N$ the standard basis of $\mathbb{R}^N$, then we have the following formula:

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

But this gives the formula $< f(e_j), e_i >= A_{ij}$ in the statement, as desired. □

The above result is something quite deep, and many comments can be made, in relation with it. To start with, we have the following warning, to be duly recorded:

Warning 14.4. *Always write your vectors $x \in \mathbb{R}^N$ vertically, as in Definition 14.1, otherwise the basic linear algebra formula $f(x) = Ax$ won't work.*

And I'm saying this in view of the fact that many math professors, myself included, quite often do the bad thing, and write the vectors horizontally. With the main reason for this coming from our Latex software, that we use for writing math, where an horizontal vector is easy to write, while a vertical vector needs some substantial lines of code.

As a second comment now, coming rather as a mathematical theorem, we have:

Theorem 14.5. *Regarding the linear maps, written as $f_A(x) = Ax$:*

1. *These compose according to $f_A f_B = f_{AB}$.*
2. *f_A is invertible when A is invertible, and $f_A^{-1} = f_{A^{-1}}$.*
3. *When A is invertible, $f_A(x) = y$ is solved by $x = f_{A^{-1}}(y)$.*

Proof. This is something self-explanatory, with (1) being clear from definitions, (2) coming from (1), and (3) coming from (2). As a comment, however, in order to understand the meaning of this, let us see what (3) tells us. The equation $f_A(x) = y$ reads:

$$\begin{cases} A_{11}x_1 + \ldots + A_{1N}x_N = y_1 \\ \qquad\qquad \vdots \\ A_{N1}x_1 + \ldots + A_{NN}x_N = y_N \end{cases}$$

We recognize here an arbitrary linear system, which is something that is certainly not easy to solve, with bare hands. But with our linear algebra technology, assuming that $A = (A_{ij})$ is invertible, say with inverse $B = (B_{ij})$, the solution is given by:

$$\begin{cases} x_1 = B_{11}y_1 + \ldots + B_{1N}y_N \\ \qquad\qquad \vdots \\ x_N = B_{N1}y_1 + \ldots + B_{NN}y_N \end{cases}$$

Which sounds quite amazing, hope you agree with me. In practice, however, inverting matrices is something non-trivial. We will be back to this later, with a solution. □

What is next? Lots of examples, and geometry, in order to understand how all the above works. Let us first discuss the linear maps $f:\mathbb{R}^2\to\mathbb{R}^2$. We have here:

Proposition 14.6. *The rotation of angle $t\in\mathbb{R}$, and the symmetry with respect to the Ox axis rotated by an angle $t/2\in\mathbb{R}$, are given by the matrices*

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

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

Proof. The rotation being linear, it must correspond to a certain matrix:

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

We can guess this matrix, via its action on the basic coordinate vectors $\binom{1}{0}$ and $\binom{0}{1}$. Indeed, a quick picture in the plane shows that we must have:

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

Guessing now the matrix is not complicated, because the first equality gives us the first column, and the second equality gives us the second column:

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

Thus, we can just put together these two vectors, and we obtain our matrix R_t. As for the symmetry, the proof here is similar, again by computing $S_t\binom{1}{0}$ and $S_t\binom{0}{1}$. □

Let us record as well a result regarding the projections, as follows:

Proposition 14.7. *The projection on the Ox axis rotated by an angle $t/2\in\mathbb{R}$ is*

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

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

Proof. A quick picture in the plane, using similarity of triangles, and the basic trigonometry formulae for the duplication of angles, show that we must have:

$$P_t\begin{pmatrix}1\\ 0\end{pmatrix}=\cos\frac{t}{2}\begin{pmatrix}\cos\frac{t}{2}\\ \sin\frac{t}{2}\end{pmatrix}=\frac{1}{2}\begin{pmatrix}1+\cos t\\ \sin t\end{pmatrix}$$

Similarly, another quick picture plus trigonometry show that we must have:

$$P_t\begin{pmatrix}0\\ 1\end{pmatrix}=\sin\frac{t}{2}\begin{pmatrix}\cos\frac{t}{2}\\ \sin\frac{t}{2}\end{pmatrix}=\frac{1}{2}\begin{pmatrix}\sin t\\ 1-\cos t\end{pmatrix}$$

Now by putting together these two vectors, and we obtain our matrix. □

As a continuation of this, let us try now to understand the rotations, symmetries and projections in $\mathbb{R}^N$. For this purpose, we will need the following standard fact:

THEOREM 14.8. *We have the following formula, valid for any $x, y \in \mathbb{R}^N$,*

$$< Ax, y >=< x, A^t y >$$

with A^t being the transpose matrix, given by the following formula:

$$(A^t)_{ij} = A_{ji}$$

Also, the transpose matrices are subject to the formula $(AB)^t = B^t A^t$.

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

(1) We recall from Theorem 14.3 that we have the following formula:

$$< Ae_j, e_i >= A_{ij}$$

By using this for both the matrix A and its transpose A^t, we obtain:

$$< Ae_j, e_i >=< e_j, A^t e_i >$$

Now by linearity, this gives the formula in the statement, for any $x, y \in \mathbb{R}^N$.

(2) Regarding now the second assertion, this can be checked as follows:

$$\begin{aligned} ((AB)^t)_{ij} &= (AB)_{ji} \\ &= \sum_k A_{jk} B_{ki} \\ &= \sum_k (B^t)_{ik} (A^t)_{kj} \\ &= (B^t A^t)_{ij} \end{aligned}$$

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

We can solve now our isometry and projection questions in $\mathbb{R}^N$, as follows:

THEOREM 14.9. *The following happen:*

1. *$f(x) = Ux$ with $U \in M_N(\mathbb{R})$ is an isometry precisely when $U^t = U^{-1}$.*
2. *$f(x) = Px$ with $P \in M_N(\mathbb{R})$ is a projection precisely when $P^2 = P^t = P$.*

PROOF. Let us first recall that the lengths, or norms, of the vectors $x \in \mathbb{R}^N$ can be recovered from the knowledge of the scalar products, as follows:

$$||x|| = \sqrt{< x, x >}$$

Conversely, and quite remarkably, we can compute the scalar products in terms of distances, by using the parallelogram identity, which is as follows:

$$\begin{aligned} ||x+y||^2 - ||x-y||^2 &= ||x||^2 + ||y||^2 + 2 < x, y > -||x||^2 - ||y||^2 + 2 < x, y > \\ &= 4 < x, y > \end{aligned}$$

Finally, we will make use of the formulae in Theorem 14.8, namely:

$$< Ax, y >=< x, A^t y > \quad , \quad (AB)^t = B^t A^t$$

(1) Given a matrix $U \in M_N(\mathbb{R})$, we have indeed the following equivalences:

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

(2) Given a matrix $P \in M_N(\mathbb{R})$, in order for $x \to Px$ to be an oblique projection, we must have $P^2 = P$. Now observe that this projection is orthogonal when:

$$\begin{aligned}
< Px - Py, Px - x >= 0 &\iff < x - y, P^t Px - P^t x >= 0 \\
&\iff P^t Px - P^t x = 0 \\
&\iff P^t P - P^t = 0 \\
&\iff P^t = P^t P
\end{aligned}$$

The point now is that by transposing this latter formula, we must have as well:

$$P = (P^t P)^t = P^t (P^t)^t = P^t P$$

Thus we must have $P = P^t$, and this gives the result. □

Here is now a key computation of projections, in arbitrary N dimensions:

THEOREM 14.10. *The rank* 1 *projections are given by the formula*

$$P_x = \frac{1}{||x||^2}(x_i x_j)_{ij}$$

with the constant $||x||$ *being as usual the length of the vector.*

PROOF. Consider a vector $y \in \mathbb{R}^N$. Its projection on $\mathbb{R}x$ must be a certain multiple of x, and we are led in this way to the following formula:

$$P_x y = \frac{< y, x >}{< x, x >} x = \frac{1}{||x||^2} < y, x > x$$

With this in hand, we can now compute the entries of P_x, as follows:

$$\begin{aligned}
(P_x)_{ij} &= < P_x e_j, e_i > \\
&= \frac{1}{||x||^2} < e_j, x >< x, e_i > \\
&= \frac{x_j x_i}{||x||^2}
\end{aligned}$$

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

As an application, we can recover a result that we already know, namely:

PROPOSITION 14.11. *In 2 dimensions, the rank 1 projections, which are the projections on the Ox axis rotated by an angle $t/2 \in [0,\pi)$, are given by the following formula:*

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

Together with the following two matrices, which are the rank 0 and 2 projections in $\mathbb{R}^2$,

$$0 = \begin{pmatrix}0 & 0\\ 0 & 0\end{pmatrix} \quad , \quad 1 = \begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}$$

these are all the projections in 2 dimensions.

PROOF. The first assertion follows from the general formula in Theorem 14.10, by plugging in the following vector, depending on a parameter $s \in [0,\pi)$:

$$x = \begin{pmatrix}\cos s\\ \sin s\end{pmatrix}$$

We obtain in this way the following matrix, which with $t = 2s$ is the one in the statement, via some trigonometry:

$$P_{2s} = \begin{pmatrix}\cos^2 s & \cos s\sin s\\ \cos s\sin s & \sin^2 s\end{pmatrix}$$

As for the second assertion, this is clear from the first one, because outside rank 1 we can only have rank 0 or rank 2, corresponding to the matrices in the statement. □

Here is another interesting application, this time in N dimensions:

PROPOSITION 14.12. *The projection on the all-1 vector $\xi \in \mathbb{R}^N$ is*

$$P_\xi = \frac{1}{N}\begin{pmatrix}1 & \ldots & 1\\ \vdots & & \vdots\\ 1 & \ldots & 1\end{pmatrix}$$

with the all-1 matrix on the right being called the flat matrix.

PROOF. The matrix in the statement acts in the following way:

$$P_\xi\begin{pmatrix}x_1\\ \vdots\\ x_N\end{pmatrix} = \frac{x_1+\ldots+x_N}{N}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$$

Thus P_ξ is indeed a projection onto $\mathbb{R}\xi$, and the fact that this projection is indeed the orthogonal one follows either by a direct orthogonality computation, or by using the general formula in Theorem 14.10, by plugging in the all-1 vector ξ. □

14b. The determinant

We have seen so far that most of the interesting maps $f:\mathbb{R}^N\to\mathbb{R}^N$ that we know, such as the rotations, symmetries and projections, are linear, and can be written in the following form, with $A\in M_N(\mathbb{R})$ being a square matrix:

$$f(v)=Av$$

Motivated by Theorem 14.5, let us study now invertibility questions, for matrices or linear maps. In the simplest case, in 2 dimensions, the result is as follows:

Theorem 14.13. *We have the following inversion formula, for the 2×2 matrices:*

$$\begin{pmatrix}a&b\\ c&d\end{pmatrix}^{-1}=\frac{1}{ad-bc}\begin{pmatrix}d&-b\\ -c&a\end{pmatrix}$$

When $ad-bc=0$, the matrix is not invertible.

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

(1) As a first observation, when $ad-bc=0$ we must have, for some $\lambda\in\mathbb{R}$:

$$b=\lambda a\quad,\quad d=\lambda c$$

Thus our matrix must be of the following special type:

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

But in this case the columns are proportional, so the linear map associated to the matrix is not invertible, and so the matrix itself is not invertible either.

(2) When $ad-bc\neq 0$, let us look for an inversion formula of the following type:

$$\begin{pmatrix}a&b\\ c&d\end{pmatrix}^{-1}=\frac{1}{ad-bc}\begin{pmatrix}*&*\\ *&*\end{pmatrix}$$

We must therefore solve the following system of equations:

$$\begin{pmatrix}a&b\\ c&d\end{pmatrix}\begin{pmatrix}*&*\\ *&*\end{pmatrix}=\begin{pmatrix}ad-bc&0\\ 0&ad-bc\end{pmatrix}$$

But the solution to these equations is obvious, is as follows:

$$\begin{pmatrix}a&b\\ c&d\end{pmatrix}\begin{pmatrix}d&-b\\ -c&a\end{pmatrix}=\begin{pmatrix}ad-bc&0\\ 0&ad-bc\end{pmatrix}$$

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

In order to deal now with the inversion problem in general, for the arbitrary matrices $A\in M_N(\mathbb{R})$, we will use the same method as the one above, at $N=2$. The difficult point is that of constructing the analogue of $ad-bc$, and this can be done as follows:

Definition 14.14. *The determinant of a square real matrix is the signed volume of the parallelepiped formed by its column vectors. That is, we have*

$$\det(v_1\ldots v_N)=\pm vol<v_1,\ldots,v_N>$$

with the sign being $+$ *when* $\{v_1,\ldots,v_N\}\subset\mathbb R^N$ *is positively oriented, in the sense that one can continuously pass from it to the standard basis of* $\mathbb R^N$*, and is* $-$ *otherwise.*

As a first observation, it follows from Theorem 14.5 that we have indeed:

$$\exists A^{-1}\iff\det A\neq0$$

As for the compatibility with what we did in Theorem 14.13, this comes from:

Theorem 14.15. *In* 2 *dimensions we have the following formula,*

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

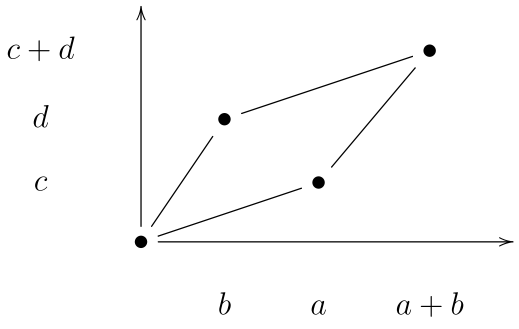
with $\det:M_2(\mathbb R)\to\mathbb R$ *being the function constructed above.*

Proof. We must do two computations, for the volume itself, taken unsigned, and then for the sign, and both these computations are elementary, as follows:

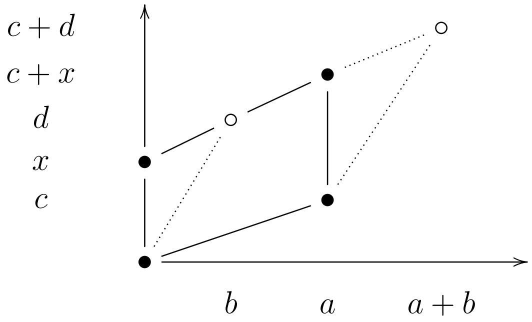
(1) First, we must show that the area of the parallelogram formed by $\binom{a}{c},\binom{b}{d}$ equals $|ad-bc|$. We can assume $a,b,c,d>0$ for simplifying, the proof in general being similar. Moreover, by switching if needed the vectors $\binom{a}{c},\binom{b}{d}$, we can assume that we have:

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

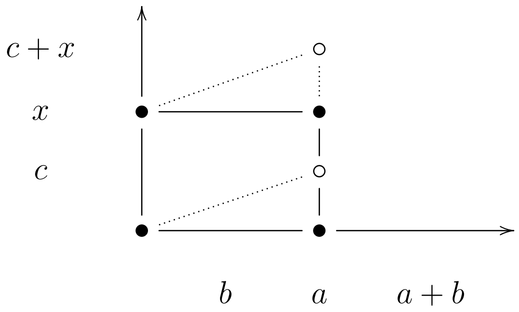
According to these conventions, the picture of our parallelogram is as follows:



Now let us slide the upper side downwards left, until we reach the Oy axis. Our parallelogram, which has not changed its area in this process, becomes:



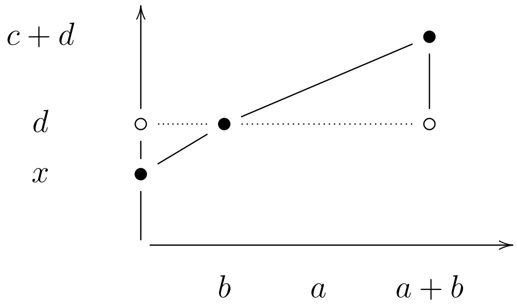
We can further modify this parallelogram, once again by not altering its area, by sliding the right side downwards, until we reach the Ox axis:



Let us compute now the area. Since our two sliding operations have not changed the area of the original parallelogram, this area is given by:

$$A = ax$$

In order to compute the quantity x, observe that in the context of the first move, we have two similar triangles, according to the following picture:



Thus, we are led to the following equation for the number x:

$$\frac{d-x}{b}=\frac{c}{a}$$

By solving this equation, we obtain the following value for x:

$$x=d-\frac{bc}{a}$$

Thus the area of our parallelogram, or rather of the final rectangle obtained from it, which has the same area as the original parallelogram, is given by, as desired:

$$A=ax=ad-bc$$

(2) It remains to compute the sign. According to our general conventions from Definition 14.14, the sign of the system of vectors $\binom{a}{c},\binom{b}{d}$ is as follows:

– The sign is $+$ when these vectors come in this order with respect to the counterclockwise rotation in the plane, around 0.

– The sign is $-$ otherwise, meaning when these vectors come in this order with respect to the clockwise rotation in the plane, around 0.

If we assume now $a,b,c,d>0$ for simplifying, we are left with comparing the angles having the numbers c/a and d/b as tangents, and we obtain in this way:

$$sgn\left[\begin{pmatrix}a\\c\end{pmatrix},\begin{pmatrix}b\\d\end{pmatrix}\right]=\begin{cases}+&\text{if }\frac{c}{a}<\frac{d}{b}\\-&\text{if }\frac{c}{a}>\frac{d}{b}\end{cases}$$

But this gives the formula in the statement, and the proof in general is similar. □

In the general case now, by similarly playing with the Thales theorem and other geometry methods, we are led to some rules for computing determinants, as follows:

Theorem 14.16. *The determinant has the following properties:*

(1) *When adding two columns, the determinants get added:*

$$\det(\ldots,u+v,\ldots)=\det(\ldots,u,\ldots)+\det(\ldots,v,\ldots)$$

(2) *When multiplying columns by scalars, the determinant gets multiplied:*

$$\det(\lambda_1v_1,\ldots,\lambda_Nv_N)=\lambda_1\ldots\lambda_N\det(v_1,\ldots,v_N)$$

(3) *When permuting two columns, the determinant changes the sign:*

$$\det(\ldots,u,\ldots,v,\ldots)=-\det(\ldots,v,\ldots,u,\ldots)$$

(4) *The determinant* $\det(e_1,\ldots,e_N)$ *of the standard basis of* $\mathbb{R}^N$ *is* 1.

Proof. As already mentioned, this follows a bit as in the proof of Theorem 14.15, by playing with the Thales theorem and other elementary geometry tools. The details can be found in any old-style linear algebra book, including mine [10]. □

As a basic application of the above result, we have:

THEOREM 14.17. *We have the following results:*

(1) *The determinant of a diagonal matrix is the product of diagonal entries.*
(2) *The same is true for the upper triangular matrices.*
(3) *The same is true for the lower triangular matrices.*

PROOF. Here (1) is clear from definitions, and (2) can be established as follows:

$$\begin{aligned}
\begin{vmatrix}\lambda_1&&&*\\&\lambda_2&&\\&&\ddots&\\0&&&\lambda_N\end{vmatrix}
&=\begin{vmatrix}\lambda_1&0&&*\\&\lambda_2&&\\&&\ddots&\\0&&&\lambda_N\end{vmatrix}\\
&\ \vdots\\
&\ \vdots\\
&=\begin{vmatrix}\lambda_1&&&0\\&\lambda_2&&\\&&\ddots&\\0&&&\lambda_N\end{vmatrix}\\
&=\lambda_1\ldots\lambda_N
\end{aligned}$$

As for the proof of (3), this is similar, by proceeding from right to left. □

As an important theoretical result now, which can ultimately lead to an algebraic reformulation of the whole determinant problematics, we have:

THEOREM 14.18. *The determinant of square matrices is the unique map*

$$\det : M_N(\mathbb{R}) \to \mathbb{R}$$

satisfying the conditions in Theorem 14.16.

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

(1) Our first claim is that any map $\det' : M_N(\mathbb{R}) \to \mathbb{R}$ satisfying the conditions in Theorem 14.16 must coincide with det on the upper triangular matrices. But this is clear from the proof of Theorem 14.17, which only uses the rules in Theorem 14.16.

(2) Our second claim is that we have $\det' = \det$, on all matrices. But this can be proved by putting the matrix in upper triangular form, by using operations on the columns, in the spirit of the manipulations from the proof of Theorem 14.17. □

Moving on, here is another important result, both in theory and practice:

Theorem 14.19. *The determinant is subject to the row expansion formula*

$$\begin{aligned}\begin{vmatrix}a_{11}&\ldots&a_{1N}\\ \vdots&&\vdots\\ a_{N1}&\ldots&a_{NN}\end{vmatrix}\ &=\ a_{11}\begin{vmatrix}a_{22}&\ldots&a_{2N}\\ \vdots&&\vdots\\ a_{N2}&\ldots&a_{NN}\end{vmatrix}-a_{12}\begin{vmatrix}a_{21}&a_{23}&\ldots&a_{2N}\\ \vdots&\vdots&&\vdots\\ a_{N1}&a_{N3}&\ldots&a_{NN}\end{vmatrix}\\ &+\ldots\ldots+(-1)^{N+1}a_{1N}\begin{vmatrix}a_{21}&\ldots&a_{2,N-1}\\ \vdots&&\vdots\\ a_{N1}&\ldots&a_{N,N-1}\end{vmatrix}\end{aligned}$$

and this method fully computes it, by recurrence.

Proof. This follows from the fact that the formula in the statement produces a certain function $\det : M_N(\mathbb{R}) \to \mathbb{R}$, which has the 4 properties in Theorem 14.16. □

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

Theorem 14.20. *The determinant is subject to the column expansion formula*

$$\begin{aligned}\begin{vmatrix}a_{11}&\ldots&a_{1N}\\ \vdots&&\vdots\\ a_{N1}&\ldots&a_{NN}\end{vmatrix}\ &=\ a_{11}\begin{vmatrix}a_{22}&\ldots&a_{2N}\\ \vdots&&\vdots\\ a_{N2}&\ldots&a_{NN}\end{vmatrix}-a_{21}\begin{vmatrix}a_{12}&\ldots&a_{1N}\\ a_{32}&\ldots&a_{3N}\\ \vdots&&\vdots\\ a_{N2}&\ldots&a_{NN}\end{vmatrix}\\ &+\ldots\ldots+(-1)^{N+1}a_{N1}\begin{vmatrix}a_{12}&\ldots&a_{1N}\\ \vdots&&\vdots\\ a_{N-1,2}&\ldots&a_{N-1,N}\end{vmatrix}\end{aligned}$$

and this method fully computes it, by recurrence.

Proof. This follows indeed by using the same argument as for the rows. □

We can now complement Theorem 14.16 with a similar result for the rows:

Theorem 14.21. *The determinant has the following properties:*

(1) *When adding two rows, the determinants get added:*

$$\det\begin{pmatrix}\vdots\\ u+v\\ \vdots\end{pmatrix}=\det\begin{pmatrix}\vdots\\ u\\ \vdots\end{pmatrix}+\det\begin{pmatrix}\vdots\\ v\\ \vdots\end{pmatrix}$$

(2) *When multiplying row by scalars, the determinant gets multiplied:*

$$\det\begin{pmatrix}\lambda_1v_1\\ \vdots\\ \lambda_Nv_N\end{pmatrix}=\lambda_1\ldots\lambda_N\det\begin{pmatrix}v_1\\ \vdots\\ v_N\end{pmatrix}$$

(3) *When permuting two rows, the determinant changes the sign.*

Proof. This follows indeed by using the using various formulae established above, and is best seen by using the column expansion formula from Theorem 14.20. □

In order to reach now to some further results, let us adopt the linear map point of view. In this setting, the definition of the determinant reformulates as follows:

Theorem 14.22. *Given a linear map, written as $f(v) = Av$, its "inflation coefficient", obtained as the signed volume of the image of the unit cube, is given by:*

$$I_f = \det A$$

More generally, I_f is the inflation ratio of any parallelepiped in $\mathbb{R}^N$, via the transformation f. In particular f is invertible precisely when $\det A \neq 0$.

Proof. The only non-trivial thing in all this is the fact that the inflation coefficient I_f, as defined above, is independent of the choice of the parallelepiped. But this is a generalization of the Thales theorem, which follows from the Thales theorem itself. □

As a first application of the above linear map viewpoint, we have:

Theorem 14.23. *We have the following formula, valid for any matrices A, B:*

$$\det(AB) = \det A \cdot \det B$$

In particular, we have $\det(AB) = \det(BA)$.

Proof. The decomposition formula in the statement follows from $f_{AB} = f_A f_B$. Indeed, when computing the determinant, by using the "inflation coefficient" viewpoint from Theorem 14.22, we obtain the same thing on both sides. As for the formula $\det(AB) = \det(BA)$, this is clear from the first formula, which is symmetric in A, B. □

Moving on, in order to further build on what we have, we will need:

Theorem 14.24. *The permutations have a signature function*

$$\varepsilon : S_N \to \{\pm 1\}$$

which can be defined in the following equivalent ways:

1. *As $(-1)^c$, where c is the number of inversions.*
2. *As $(-1)^t$, where t is the number of transpositions.*
3. *As $(-1)^o$, where o is the number of odd cycles.*
4. *As $(-1)^x$, where x is the number of crossings.*
5. *As the sign of the corresponding permuted basis of $\mathbb{R}^N$.*

Proof. We have explain what the numbers c, t, o, x appearing in (1-4) exactly are, then why they are well-defined modulo 2, then why they are equal to each other, and finally why the constructions (1-4) yield the same sign as (5). Let us begin with the first two steps, namely precise definition of the numbers c, t, o, x, modulo 2:

(1) The idea here is that given any two numbers $i < j$ among $1, \ldots, N$, the permutation can either keep them in the same order, $\sigma(i) < \sigma(j)$, or invert them:

$$\sigma(j) > \sigma(i)$$

Now by making $i < j$ vary over all pairs of numbers in $1, \ldots, N$, we can count the number of inversions, and call it c. This is an integer, $c \in \mathbb{N}$, which is well-defined.

(2) Here the idea, which is something quite intuitive, is that any permutation appears as a product of switches, also called transpositions:

$$i \leftrightarrow j$$

The decomposition as a product of transpositions is not unique, but the number t of the needed transpositions is unique, when considered modulo 2. This follows for instance from the equivalence of (2) with (1,3,4,5), explained below.

(3) Here the point is that any permutation decomposes, in a unique way, as a product of cycles, which are by definition permutations of the following type:

$$i_1 \to i_2 \to i_3 \to \ldots\ldots \to i_k \to i_1$$

Some of these cycles have even length, and some others have odd length. By counting those having odd length, we obtain a well-defined number $o \in \mathbb{N}$.

(4) Here the method is that of drawing the permutation, as we usually do, and by avoiding triple crossings, and then counting the number of crossings. This number x depends on the way we draw the permutations, but modulo 2, we always get the same number. Indeed, this follows from the fact that we can continuously pass from a drawing to each other, and that when doing so, the number of crossings can only jump by ±2.

Summarizing, we have 4 different definitions for the signature of the permutations, which all make sense, constructed according to (1-4) above. Regarding now the fact that we always obtain the same number, this can be established as follows:

(1)=(2) This is clear, because any transposition inverts once, modulo 2.

(1)=(3) This is clear as well, because the odd cycles invert once, modulo 2.

(1)=(4) This comes from the fact that the crossings correspond to inversions.

(2)=(3) This follows by decomposing the cycles into transpositions.

(2)=(4) This comes from the fact that the crossings correspond to transpositions.

(3)=(4) This follows by drawing a product of cycles, and counting the crossings.

Finally, in what regards the equivalence of all these constructions with (5), here simplest is to use (2). Indeed, we already know that the sign of a system of vectors switches when interchanging two vectors, and so the equivalence between (2,5) is clear. □

We can now formulate the ultimate result regarding determinants, as follows:

Theorem 14.25. *We have the following formula for the determinant,*

$$\det A=\sum_{\sigma\in S_N}\varepsilon(\sigma)A_{1\sigma(1)}\ldots A_{N\sigma(N)}$$

with the signature function being the one introduced above.

Proof. This follows by recurrence over $N\in\mathbb{N}$, as follows:

(1) When developing the determinant over the first column, we obtain a signed sum of N determinants of size $(N-1)\times(N-1)$. But each of these determinants can be computed by developing over the first column too, and so on, and we are led to the conclusion that we have a formula as in the statement, with $\varepsilon(\sigma)\in\{-1,1\}$ being certain coefficients.

(2) But these latter coefficients $\varepsilon(\sigma)\in\{-1,1\}$ can only be the signatures of the corresponding permutations $\sigma\in S_N$, with this being something that can be viewed again by recurrence, with either of the definitions (1-5) in Theorem 14.24 for the signature. □

The above result is something quite tricky. As a first illustration, in 2 dimensions we recover the usual formula of the determinant, the details being as follows:

$$\begin{vmatrix}a&b\\c&d\end{vmatrix}=\varepsilon(|\,|)\cdot ad+\varepsilon(\times)\cdot cb=ad-bc$$

In 3 dimensions now, we recover the well-known Sarrus formula:

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

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

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

Finally, as a theoretical application of the formula in Theorem 14.25, we have:

Theorem 14.26. *We have the formula*

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

valid for any square matrix A.

Proof. This follows from the formula in Theorem 14.25. Indeed, we have:

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

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

And with this, good news, we know everything, or almost, about det.

14c. Diagonalization

Let us discuss now the diagonalization question for linear maps and matrices. The basic diagonalization theory, formulated in terms of matrices, is as follows:

Proposition 14.27. *A vector $v\in\mathbb{R}^N$ is called eigenvector of $A\in M_N(\mathbb{R})$, with corresponding eigenvalue λ, when A multiplies by λ in the direction of v:*

$$Av=\lambda v$$

In the case where $\mathbb{R}^N$ has a basis $v_1,\ldots,v_N$ formed by eigenvectors of A, with corresponding eigenvalues $\lambda_1,\ldots,\lambda_N$, in this new basis A becomes diagonal, as follows:

$$A\sim\begin{pmatrix}\lambda_1&&\\&\ddots&\\&&\lambda_N\end{pmatrix}$$

Equivalently, if we denote by $D=diag(\lambda_1,\ldots,\lambda_N)$ the above diagonal matrix, and by $P=[v_1\ldots v_N]$ the square matrix formed by the eigenvectors of A, we have:

$$A=PDP^{-1}$$

In this case we say that the matrix A is diagonalizable.

Proof. This is something which is clear, the idea being as follows:

(1) The first assertion is clear, because the matrix which multiplies each basis element v_i by a number λ_i is precisely the diagonal matrix $D=diag(\lambda_1,\ldots,\lambda_N)$.

(2) The second assertion follows from the first one, by changing the basis. We can prove this by a direct computation as well, because we have $Pe_i = v_i$, and so:

$$PDP^{-1}v_i = PDe_i = P\lambda_i e_i = \lambda_i Pe_i = \lambda_i v_i$$

Thus, the matrices A and PDP^{-1} coincide, as stated. □

In order to study the diagonalization problem, the idea is that the eigenvectors can be grouped into linear spaces, called eigenspaces, as follows:

Theorem 14.28. *Let $A \in M_N(\mathbb{R})$, and for any eigenvalue $\lambda \in \mathbb{R}$ define the corresponding eigenspace as being the vector space formed by the corresponding eigenvectors:*

$$E_\lambda = \left\{ v \in \mathbb{R}^N \Big| Av = \lambda v \right\}$$

These eigenspaces E_λ are then in a direct sum position, in the sense that given vectors $v_1 \in E_{\lambda_1}, \ldots, v_k \in E_{\lambda_k}$ corresponding to different eigenvalues $\lambda_1, \ldots, \lambda_k$, we have:

$$\sum_i c_i v_i = 0 \implies c_i = 0$$

In particular, we have $\sum_\lambda \dim(E_\lambda) \leq N$, with the sum being over all the eigenvalues, and our matrix is diagonalizable precisely when we have equality.

Proof. We prove the first assertion by recurrence on $k \in \mathbb{N}$. Assume by contradiction that we have a formula as follows, with the scalars $c_1, \ldots, c_k$ being not all zero:

$$c_1 v_1 + \ldots + c_k v_k = 0$$

By dividing by one of these scalars, we can assume that our formula is:

$$v_k = c_1 v_1 + \ldots + c_{k-1} v_{k-1}$$

Now let us apply A to this vector. On the left we obtain:

$$Av_k = \lambda_k v_k = \lambda_k c_1 v_1 + \ldots + \lambda_k c_{k-1} v_{k-1}$$

On the right we obtain something different, as follows:

$$\begin{aligned} A(c_1 v_1 + \ldots + c_{k-1} v_{k-1}) &= c_1 A v_1 + \ldots + c_{k-1} A v_{k-1} \\ &= c_1 \lambda_1 v_1 + \ldots + c_{k-1} \lambda_{k-1} v_{k-1} \end{aligned}$$

We conclude from this that the following equality must hold:

$$\lambda_k c_1 v_1 + \ldots + \lambda_k c_{k-1} v_{k-1} = c_1 \lambda_1 v_1 + \ldots + c_{k-1} \lambda_{k-1} v_{k-1}$$

On the other hand, we know by recurrence that the vectors $v_1, \ldots, v_{k-1}$ must be linearly independent. Thus, the coefficients must be equal, at right and at left, and since at least one of the numbers c_i must be nonzero, from $\lambda_k c_i = c_i \lambda_i$ we obtain $\lambda_k = \lambda_i$, which is a contradiction. As for the second assertion, this follows from the first one. □

In order to reach now to more advanced results, we can use the following fact:

THEOREM 14.29. *Given $A \in M_N(\mathbb{R})$, consider its characteristic polynomial:*

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

The eigenvalues of A are then the roots of P. Also, we have the inequality

$$\dim(E_\lambda) \leq m_\lambda$$

where m_λ is the multiplicity of λ, as root of P.

PROOF. The first assertion follows from the following computation, using the fact that a linear map is bijective when the determinant of the associated matrix is nonzero:

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

Regarding now the second assertion, given an eigenvalue λ of our matrix A, consider the dimension $d_\lambda = \dim(E_\lambda)$ of the corresponding eigenspace. By changing the basis of $\mathbb{R}^N$, as for the eigenspace E_λ to be spanned by the first d_λ basis elements, our matrix becomes as follows, with B being a certain smaller matrix:

$$A \sim \begin{pmatrix}\lambda 1_{d_\lambda} & 0\\ 0 & B\end{pmatrix}$$

We conclude that the characteristic polynomial of A is of the following form:

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

Thus the multiplicity m_λ of our eigenvalue λ, as a root of P, satisfies $m_\lambda \geq d_\lambda$, and this leads to the conclusion in the statement. □

In view of Theorem 14.29, it is better to trade now $\mathbb{R}$ for $\mathbb{C}$, where all polynomials have roots. What we learned so far linear algebra holds over $\mathbb{C}$ too, with the determinant being best introduced via the formula in Theorem 14.25, and we have:

THEOREM 14.30. *Given a matrix $A \in M_N(\mathbb{C})$, consider its characteristic polynomial*

$$P(X) = \det(A - X1_N)$$

then factorize this polynomial, by computing the complex roots, with multiplicities,

$$P(X) = (-1)^N (X - \lambda_1)^{n_1} \ldots (X - \lambda_k)^{n_k}$$

and finally compute the corresponding eigenspaces, for each eigenvalue found:

$$E_i = \left\{v \in \mathbb{C}^N \Big| Av = \lambda_i v\right\}$$

The dimensions of these eigenspaces satisfy then the following inequalities,

$$\dim(E_i) \leq n_i$$

and A is diagonalizable precisely when we have equality for any i.

Proof. This follows by combining the above results. Indeed, by summing the inequalities $\dim(E_\lambda)\leq m_\lambda$ from Theorem 14.29, we obtain an inequality as follows:

$$\sum_\lambda\dim(E_\lambda)\leq\sum_\lambda m_\lambda\leq N$$

On the other hand, we know from Theorem 14.28 that our matrix is diagonalizable when we have global equality. Thus, we are led to the conclusion in the statement. □

As an illustration for all this, which is a must-know computation, we have:

Proposition 14.31. *The rotation of angle $t\in\mathbb R$ in the plane diagonalizes as:*

$$\begin{pmatrix}\cos t&-\sin t\\ \sin t&\cos t\end{pmatrix}=\frac{1}{2}\begin{pmatrix}1&1\\ i&-i\end{pmatrix}\begin{pmatrix}e^{-it}&0\\ 0&e^{it}\end{pmatrix}\begin{pmatrix}1&-i\\ 1&i\end{pmatrix}$$

Over the reals this is impossible, unless $t=0,\pi$, where the rotation is diagonal.

Proof. Observe first that, as indicated, unlike we are in the case $t=0,\pi$, where our rotation is $\pm 1_2$, our rotation is a "true" rotation, having no eigenvectors in the plane. Fortunately the complex numbers come to the rescue, via the following computation:

$$\begin{pmatrix}\cos t&-\sin t\\ \sin t&\cos t\end{pmatrix}\begin{pmatrix}1\\ i\end{pmatrix}=\begin{pmatrix}\cos t-i\sin t\\ i\cos t+\sin t\end{pmatrix}=e^{-it}\begin{pmatrix}1\\ i\end{pmatrix}$$

We have as well a second complex eigenvector, coming from:

$$\begin{pmatrix}\cos t&-\sin t\\ \sin t&\cos t\end{pmatrix}\begin{pmatrix}1\\ -i\end{pmatrix}=\begin{pmatrix}\cos t+i\sin t\\ -i\cos t+\sin t\end{pmatrix}=e^{it}\begin{pmatrix}1\\ -i\end{pmatrix}$$

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

At the level of basic examples of diagonalizable matrices, we first have the following result, which provides us with the "generic" examples:

Theorem 14.32. *For a matrix $A\in M_N(\mathbb C)$ the following conditions are equivalent,*

1. *The eigenvalues are different, $\lambda_i\neq\lambda_j$,*
2. *The characteristic polynomial P has simple roots,*
3. *The characteristic polynomial satisfies $(P,P')=1$,*
4. *The resultant of P,P' is nonzero, $R(P,P')\neq 0$,*
5. *The discriminant of P is nonzero, $\Delta(P)\neq 0$,*

and in this case, the matrix is diagonalizable.

Proof. The last assertion holds indeed, due to Theorem 14.30. As for the various equivalences, these are all standard, by using the theory of R,Δ from chapter 9. □

As already mentioned, one can prove that the matrices having distinct eigenvalues are "generic", and so the above result basically captures the whole situation. We have in fact the following collection of density results, which are quite advanced:

Theorem 14.33. *The following happen, inside $M_N(\mathbb{C})$:*

1. *The invertible matrices are dense.*
2. *The matrices having distinct eigenvalues are dense.*
3. *The diagonalizable matrices are dense.*

Proof. These are quite advanced results, which can be proved as follows:

(1) This is clear, intuitively speaking, because the invertible matrices are given by the condition $\det A \neq 0$. Thus, the set formed by these matrices appears as the complement of the hypersurface $\det A = 0$, and so must be dense inside $M_N(\mathbb{C})$, as claimed.

(2) Here we can use a similar argument, this time by saying that the set formed by the matrices having distinct eigenvalues appears as the complement of the hypersurface given by $\Delta(P_A) = 0$, and so must be dense inside $M_N(\mathbb{C})$, as claimed.

(3) This follows from (2), via the fact that the matrices having distinct eigenvalues are diagonalizable, that we know from Theorem 14.32. There are of course some other proofs as well, for instance by putting the matrix in Jordan form. □

14d. Some applications

As a first application of our linear algebra technology, which is actually related to what we just did in Theorem 14.33, let us go back to the resultant $R(P, Q)$, constructed in chapter 9. We know from there that the computation of $R(P, Q)$ is not an easy question. However, and remarkably, with linear algebra we can solve this question, as follows:

Theorem 14.34. *The resultant of two polynomials, written as*

$$P = p_k X^k + \ldots + p_1 X + p_0 \quad , \quad Q = q_l X^l + \ldots + q_1 X + q_0$$

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 material from chapter 9. □

We can formulate as well a result regarding the discriminants, as follows:

THEOREM 14.35. *Given a polynomial $P \in \mathbb{C}[X]$, written as*

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

its discriminant is given by the following formula,

$$\Delta(P) = \frac{(-1)^{\binom{N}{2}}}{a} \begin{vmatrix} a & & & Na & & \\ \vdots & \ddots & & \vdots & \ddots & \\ z & & a & y & & Na \\ & \ddots & \vdots & & \ddots & \vdots \\ & & z & & & y \end{vmatrix}$$

with the normalization being there for having $\Delta(P) = b^2 - 4ac$ at $N = 2$.

PROOF. This follows indeed from the following formula, from chapter 9, standing as a definition for the discriminant, by computing the resultant using Theorem 14.34:

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

As for the numerics at $N = 2$, these go as follows, using Sarrus:

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

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

At $N = 3$ now, we can recover the formula from chapter 9, with a simpler proof:

THEOREM 14.36. *The discriminant of a degree* 3 *polynomial,*

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

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

PROOF. We have indeed the following computation, using Theorem 14.34:

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

Thus, the discriminant of our polynomial is given by the following formula:

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

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

We can do as well the computation in degree 4, the result being as follows:

THEOREM 14.37. *The discriminant of $P = ax^4 + bx^3 + cx^2 + dx + e$ is given by*

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

and with this being nearly impossible to prove, with bare hands.

PROOF. We have indeed the following formula, coming from Theorem 14.35, which gives the result, after some routine computations:

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

As for the last assertion, this is obviously something subjective, which actually depends on whether you solved my chapter 9 exercise, about this, or not. □

14e. Exercises

This was a very nice and pleasant chapter, and as exercises on this, we have:

EXERCISE 14.38. *Is an arbitrary map, mapping lines to lines, linear?*

EXERCISE 14.39. *Learn more about scalar products in $\mathbb{R}^N$, and their use.*

EXERCISE 14.40. *Hide this book and recover the formulae of R_t, S_t, P_t, by yourself.*

EXERCISE 14.41. *Write in matrix form some basic transformations, in $3D$.*

EXERCISE 14.42. *Diagonalize, without computations, the matrices S_t, P_t.*

EXERCISE 14.43. *Diagonalize the all-one matrix over $\mathbb{R}$. Try also over $\mathbb{C}$.*

EXERCISE 14.44. *Learn about complex scalar products, $< x, y >= \sum_i x_i\bar{y}_i$.*

EXERCISE 14.45. *Learn also about adjoint matrices, and about unitaries.*

As a bonus exercise, of course, read more linear algebra, as much as you can.

CHAPTER 15

Advanced calculus

15a. First derivatives

We discuss here the study of the functions $f:\mathbb{R}^N\to\mathbb{R}^M$, as a continuation of what we know from Part III, regarding the functions $f:\mathbb{R}\to\mathbb{R}$. We must first talk about continuity and intermediate values, and for this purpose, it is most convenient to use the advanced approach to continuity, using open and closed sets. Let us start with:

Proposition 15.1. *We can talk about open and closed sets in $\mathbb{R}^N$, in the obvious way, exactly as we did it before in $\mathbb{R}$, and the following happen:*

(1) *Open balls are open, closed balls are closed.*
(2) *Union of open sets is open, intersection of closed sets is closed.*
(3) *Finite intersection of open sets is open, finite union of closed sets is closed.*
(4) *The open sets are exactly the complements of closed sets.*

Proof. This is something that we know well from chapter 10, in the case $N=1$, and the proof in general is nearly identical. We will leave this as an instructive exercise. □

Getting now to the functions, we have the following result about them:

Theorem 15.2. *Given a function $f:\mathbb{R}^N\to\mathbb{R}^M$, the following are equivalent:*

(1) *f is continuous.*
(2) *If O is open, then $f^{-1}(O)$ is open.*
(3) *If C is closed, then $f^{-1}(C)$ is closed.*

Proof. This is again something that we know well from chapter 10, in the case $N=1$, and the proof in general is similar, with the equivalence (1) $\iff$ (2) coming from definitions, and with (2) $\iff$ (3) coming from Proposition 15.1 (4). □

Regarding now the compact and connected sets, their basic theory is as follows:

Proposition 15.3. *We can talk about compact and connected sets in $\mathbb{R}^N$, in the obvious way, exactly as we did it before in $\mathbb{R}$, and the following happen:*

(1) *Compact is the same as being closed and bounded.*
(2) *Convex $\implies$ path connected $\implies$ connected.*

Proof. Again, this is a routine extension of things that we know from chapter 10:

(1) Let us call indeed $K \subset \mathbb{R}^N$ compact when any open cover $K \subset \cup_i O_i$ has a finite subcover. It is clear then, by using suitable covers, exactly as in the 1-dimensional case, that K must be closed, and bounded as well. As for the converse, this follows again as in the 1-dimensional case, with the main ingredient here, which again can be proved by using a suitable cover, being the fact that the unit cube is indeed compact.

(2) Let us call indeed $E \subset \mathbb{R}^N$ connected when it is not possible to split it into two parts, that is, when it is not possible to have $E \subset A \sqcup B$, with A, B open. It is then clear that if E is path connected, then it must be connected, because when assuming $E \subset A \sqcup B$ with A, B open, we cannot have any path from $a \in A$ to $b \in B$. □

Getting now to the functions, we have the following result about them:

Theorem 15.4. *Assuming that $f : \mathbb{R}^N \to \mathbb{R}^M$ is continuous:*

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

Proof. This is again very standard, as in 1 dimension, with (1) coming from the definition of compactness, and (2) coming from the definition of connectedness. □

Getting now to what we really wanted to say about continuous functions, intermediate value theorem, this is (2) above, so let us have this highlighted, as follows:

Theorem 15.5 (Intermediate values). *Assuming that a function*

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

with $X \subset \mathbb{R}^N$ is continuous, if the domain X is connected, so is its image $f(X)$.

Proof. We have already stated this in Theorem 15.4 (2), but let us see now how the detailed proof goes as well. Assume by contradiction that $f(X)$ is not connected, which in practice means that we can find two disjoint open sets A, B such that:

$$f(X) \subset A \sqcup B$$

By taking inverse images, we obtain from this a disjoint union as follows:

$$X \subset f^{-1}(A) \sqcup f^{-1}(B)$$

Now since inverse image of an open set is open, wich this being something which is clear from definitions, both the above sets $f^{-1}(A)$ and $f^{-1}(B)$ are open. Thus we have managed to split X into two parts, contradicting its connectivity, as desired. □

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

Theorem 15.6 (Heine, Cantor). *Any continuous function*

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

with $X \subset \mathbb{R}^N$ compact is automatically uniformly continuous.

Proof. This is again something that we know well from chapter 10, in the case $N=M=1$, and the proof in general is similar, by using a suitable open cover. □

Getting now to more advanced analysis, let us discuss the differentiability in several variables. At order 1, the situation is quite simple, as follows:

Theorem 15.7. *The derivative of a function $f:\mathbb{R}^N\to\mathbb{R}^M$, making the formula*

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

work, must be the matrix of partial derivatives at x, namely

$$f'(x)=\left(\frac{df_i}{dx_j}(x)\right)_{ij}\in M_{M\times N}(\mathbb{R})$$

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

Proof. As a first observation, the formula in the statement makes sense indeed, as an equality, or rather approximation, of vectors in $\mathbb{R}^M$, as follows:

$$f\begin{pmatrix}x_1+t_1\\ \vdots\\ x_N+t_N\end{pmatrix}\simeq f\begin{pmatrix}x_1\\ \vdots\\ x_N\end{pmatrix}+\begin{pmatrix}\frac{df_1}{dx_1}(x)&\ldots&\frac{df_1}{dx_N}(x)\\ \vdots&&\vdots\\ \frac{df_M}{dx_1}(x)&\ldots&\frac{df_M}{dx_N}(x)\end{pmatrix}\begin{pmatrix}t_1\\ \vdots\\ t_N\end{pmatrix}$$

In order to prove now this formula, we can proceed by recurrence, as follows:

(1) First of all, at $N=M=1$ what we have is a usual 1-variable function $f:\mathbb{R}\to\mathbb{R}$, and the formula in the statement is something that we know well, namely:

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

(2) Let us discuss now the case $N=2,M=1$. Here what we have is a function $f:\mathbb{R}^2\to\mathbb{R}$, and by using twice the basic approximation result from (1), we obtain:

$$\begin{aligned}f\begin{pmatrix}x_1+t_1\\ x_2+t_2\end{pmatrix}&\simeq f\begin{pmatrix}x_1+t_1\\ x_2\end{pmatrix}+\frac{df}{dx_2}(x)t_2\\ &\simeq f\begin{pmatrix}x_1\\ x_2\end{pmatrix}+\frac{df}{dx_1}(x)t_1+\frac{df}{dx_2}(x)t_2\\ &= f\begin{pmatrix}x_1\\ x_2\end{pmatrix}+\begin{pmatrix}\frac{df}{dx_1}(x)&\frac{df}{dx_2}(x)\end{pmatrix}\begin{pmatrix}t_1\\ t_2\end{pmatrix}\end{aligned}$$

(3) More generally, we can deal in this way with the case $N\in\mathbb{N},M=1$, by recurrence. But this gives the result in the general case $N,M\in\mathbb{N}$ too. Indeed, let us write:

$$f=\begin{pmatrix}f_1\\ \vdots\\ f_M\end{pmatrix}$$

We can apply our result to each of the components $f_i:\mathbb R^N\to\mathbb R$, and we get:

$$f_i\begin{pmatrix}x_1+t_1\\ \vdots\\ x_N+t_N\end{pmatrix}\simeq f_i\begin{pmatrix}x_1\\ \vdots\\ x_N\end{pmatrix}+\begin{pmatrix}\frac{df_i}{dx_1}(x)&\ldots&\frac{df_i}{dx_N}(x)\end{pmatrix}\begin{pmatrix}t_1\\ \vdots\\ t_N\end{pmatrix}$$

But this is precisely what we want, at the level of the global map $f:\mathbb R^N\to\mathbb R^M$. $\square$

As a technical complement to the above result, further clarifying things, we have:

Theorem 15.8. *For a function $f:X\to\mathbb R^M$, with $X\subset\mathbb R^N$, the following conditions are equivalent, and in this case we say that f is continuously differentiable:*

1. *f is differentiable, and the map $x\to f'(x)$ is continuous.*
2. *f has partial derivatives, which are continuous with respect to $x\in X$.*

If these conditions are satisfied, $f'(x)$ is the matrix fomed by the partial derivatives at x.

Proof. We already know, from Theorem 15.7, that the last assertion holds. Regarding now the proof of the equivalence, this goes as follows:

(1) $\implies$ (2) Assuming that f is differentiable, we know from Theorem 15.7 that $f'(x)$ is the matrix fomed by the partial derivatives at x. Thus, for any $x,y\in X$:

$$\frac{df_i}{dx_j}(x)-\frac{df_i}{dx_j}(y)=f'(x)_{ij}-f'(y)_{ij}$$

By applying now the absolute value, we obtain from this the following estimate:

$$\begin{aligned}\left|\frac{df_i}{dx_j}(x)-\frac{df_i}{dx_j}(y)\right|&=|f'(x)_{ij}-f'(y)_{ij}|\\&=|(f'(x)-f'(y))_{ij}|\\&\le||f'(x)-f'(y)||\end{aligned}$$

But this gives the result, because if the map $x\to f'(x)$ is assumed to be continuous, then the partial derivatives follow to be continuous with respect to $x\in X$.

(2) $\implies$ (1) This is something more technical. For simplicity, let us assume $M=1$, the proof in general being similar. Given $x\in X$ and $\varepsilon>0$, let us pick $r>0$ such that the ball $B=B_x(r)$ belongs to X, and such that the following happens, over B:

$$\left|\frac{df}{dx_j}(x)-\frac{df}{dx_j}(y)\right|<\frac{\varepsilon}{N}$$

Our claim is that, with this choice made, we have the following estimate, for any $t\in\mathbb R^N$ satisfying $||t||<r$, with A being the vector of partial derivatives at x:

$$|f(x+t)-f(x)-At|\le\varepsilon||t||$$

In order to prove this claim, the idea will be that of suitably applying the mean value theorem, over the N directions of $\mathbb{R}^N$. Indeed, consider the following vectors:

$$t^{(k)} = \begin{pmatrix} t_1 \\ \vdots \\ t_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

In terms of these vectors, we have the following formula:

$$f(x+t) - f(x) = \sum_{j=1}^N f(x + t^{(j)}) - f(x + t^{(j-1)})$$

Also, the mean value theorem gives a formula as follows, with $s_j \in [0,1]$:

$$f(x + t^{(j)}) - f(x + t^{(j-1)}) = \frac{df}{dx_j}(x + s_j t^{(j)} + (1 - s_j)t^{(j-1)}) \cdot t_j$$

But, according to our assumption on $r > 0$ from the beginning, the derivative on the right differs from $\frac{df}{dx_j}(x)$ by something which is smaller than ε/N:

$$\left| \frac{df}{dx_j}(x + s_j t^{(j)} + (1 - s_j)t^{(j-1)}) - \frac{df}{dx_j}(x) \right| < \frac{\varepsilon}{N}$$

Now by putting everything together, we obtain the following estimate:

$$\begin{aligned} |f(x+t) - f(x) - At| &= \left| \sum_{j=1}^N f(x + t^{(j)}) - f(x + t^{(j-1)}) - \frac{df}{dx_j}(x) \cdot t_j \right| \\ &\leq \sum_{j=1}^N \left| f(x + t^{(j)}) - f(x + t^{(j-1)}) - \frac{df}{dx_j}(x) \cdot t_j \right| \\ &= \sum_{j=1}^N \left| \frac{df}{dx_j}(x + s_j t^{(j)} + (1 - s_j)t^{(j-1)}) \cdot t_j - \frac{df}{dx_j}(x) \cdot t_j \right| \\ &\leq \sum_{j=1}^N \frac{\varepsilon}{N} \cdot |t_j| \\ &\leq \varepsilon ||t|| \end{aligned}$$

Thus we have proved our claim, and this gives the result. □

Moving on, with this done, our next task will be that of extending to several variables our basic results from one-variable calculus. As a standard result here, we have:

THEOREM 15.9. *We have the chain derivative formula*

$$(f\circ g)'(x)=f'(g(x))\cdot g'(x)$$

as an equality of matrices.

PROOF. This is something standard in one variable, and in several variables the proof is similar, by using the abstract notion of derivative coming from Theorem 15.7. To be more precise, consider a composition of functions, as follows:

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

According to Theorem 15.7, the derivatives of these functions are certain linear maps, corresponding to certain rectangular matrices, as follows:

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

Thus, our formula makes sense indeed. As for proof, this comes from:

$$\begin{aligned}(f\circ g)(x+t)&=f(g(x+t))\\&\simeq f(g(x)+g'(x)t)\\&\simeq f(g(x))+f'(g(x))g'(x)t\end{aligned}$$

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

As a standard application of the above chain rule differentiation result, generalizing some basic things that we know from one-variable calculus, we have:

THEOREM 15.10. *Assuming that $f:X\to\mathbb R^M$ is differentiable, with $X\subset\mathbb R^N$ being convex, we have the estimate*

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

for any $x,y\in X$, where the quantity on the right is given by:

$$M=\sup_{x\in X}||f'(x)||$$

Moreover, this estimate can be sharp, for instance for the linear functions.

PROOF. This is something quite tricky, which in several variables cannot be proved with bare hands. However, we can get it by using our chain derivative formula. Consider indeed the path $\gamma:[0,1]\to\mathbb R^M$ given by the following formula:

$$\gamma(t)=tx+(1-t)y$$

Now let us set $g(t)=f(\gamma(t))$. We have then, according to the chain rule formula:

$$\begin{aligned}g'(t)&=f'(\gamma(t))\gamma'(t)\\&=f'(\gamma(t))(x-y)\end{aligned}$$

But this gives the following estimate, with $M > 0$ being as in the statement:

$$\begin{aligned} |g'(t)| &\leq ||f'(\gamma(t))|| \cdot ||x-y|| \\ &\leq M||x-y|| \end{aligned}$$

Now by using one-variable results that we know, we obtain from this:

$$||g(1)-g(0)|| \leq ||M|| \cdot ||x-y||$$

Thus, we obtain the formula in the statement. Finally, the last assertion is clear. □

Finally, again in analogy with what we know well from chapter 10, we have:

Theorem 15.11. *The Taylor formula at order* 1 *for a function* $f : \mathbb{R}^N \to \mathbb{R}$ *is*

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

and in particular, in order for x to be a local extremum, we must have $f'(x) = 0$.

Proof. Here the first assertion is something that we know, and the second assertion follows from it. Indeed, let us look at the order 1 term, given by:

$$f'(x)t = \sum_{i=1}^N \frac{df}{dx_i} t_i$$

Now since this linear combination of the entries of $t \in \mathbb{R}^N$ can range among positives and negatives, unless all the coefficients are zero, which means $f'(x) = 0$, we are led to the conclusion that local extremum needs $f'(x) = 0$ to hold, as stated. □

15b. Second derivatives

Moving on, we can talk as well about higher derivatives, simply by performing the operation of taking derivatives recursively. As a first result here, we have:

Theorem 15.12. *The double derivatives of a function* $f : \mathbb{R}^2 \to \mathbb{R}$ *satisfy*

$$\frac{d^2f}{dxdy} = \frac{d^2f}{dydx}$$

called Clairaut formula.

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

(1) Before pulling out a formal proof, as an intuitive justification for our formula, let us consider a product of power functions, $f(z) = x^p y^q$. We have then:

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

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

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

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

Thus, we can see that our commutation formula for derivatives holds indeed, and this due to the fact that the functions in x and y commute. Of course, this does not prove our formula, in general. But exercise for you, to have this idea further working.

(2) Getting now to more standard techniques, given a point in the plane, $z = (a, b)$, consider the following functions, depending on $h, k \in \mathbb{R}$ small:

$$u(h,k) = f(a+h, b+k) - f(a+h, b)$$

$$v(h,k) = f(a+h, b+k) - f(a, b+k)$$

$$w(h,k) = f(a+h, b+k) - f(a+h, b) - f(a, b+k) + f(a, b)$$

By the mean value theorem, for $h, k \neq 0$ we can find $\alpha, \beta \in \mathbb{R}$ such that:

$$\begin{aligned} w(h,k) &= u(h,k) - u(0,k) \\ &= h \cdot \frac{d}{dx} u(\alpha h, k) \\ &= h\left(\frac{d}{dx} f(a+\alpha h, b+k) - \frac{d}{dx} f(a+\alpha h, b)\right) \\ &= hk \cdot \frac{d}{dy} \cdot \frac{d}{dx} f(a+\alpha h, b+\beta k) \end{aligned}$$

Similarly, again for $h, k \neq 0$, we can find $\gamma, \delta \in \mathbb{R}$ such that:

$$\begin{aligned} w(h,k) &= v(h,k) - v(h,0) \\ &= k \cdot \frac{d}{dy} v(h, \delta k) \\ &= k\left(\frac{d}{dy} f(a+h, b+\delta k) - \frac{d}{dy} f(a, b+\delta k)\right) \\ &= hk \cdot \frac{d}{dx} \cdot \frac{d}{dy} f(a+\gamma h, b+\delta k) \end{aligned}$$

Now by dividing everything by $hk \neq 0$, we conclude from this that the following equality holds, with the numbers $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ being found as above:

$$\frac{d}{dy} \cdot \frac{d}{dx} f(a+\alpha h, b+\beta k) = \frac{d}{dx} \cdot \frac{d}{dy} f(a+\gamma h, b+\delta k)$$

But with $h, k \to 0$ we get from this the Clairaut formula, at $z = (a, b)$, as desired. □

In arbitrary dimensions now, we have the following result:

Theorem 15.13. *Given $f:\mathbb{R}^N\to\mathbb{R}$, we can talk about its higher derivatives,*

$$\frac{d^kf}{dx_{i_1}\ldots dx_{i_k}}=\frac{d}{dx_{i_1}}\cdots\frac{d}{dx_{i_k}}(f)$$

provided that these derivatives exist indeed. Moreover, due to the Clairaut formula,

$$\frac{d^2f}{dx_idx_j}=\frac{d^2f}{dx_jdx_i}$$

the order in which these higher derivatives are computed is irrelevant.

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

(1) First of all, we can talk about the quantities in the statement, with the remark however that at each step of our recursion, the corresponding partial derivative can exist of not. We will say in what follows that our function is k times differentiable if the quantities in the statement exist at any $l\leq k$, and smooth, if this works with $k=\infty$.

(2) Regarding now the second assertion, this is something more tricky. Let us first recall from Theorem 15.12 that the second derivatives of a twice differentiable function of two variables $f:\mathbb{R}^2\to\mathbb{R}$ are subject to the Clairaut formula, namely:

$$\frac{d^2f}{dxdy}=\frac{d^2f}{dydx}$$

(3) But this result clearly extends to our function $f:\mathbb{R}^N\to\mathbb{R}$, simply by ignoring the unneeded variables, so we have the Clairaut formula in general, also called Schwarz formula, which is the one in the statement, namely:

$$\frac{d^2f}{dx_idx_j}=\frac{d^2f}{dx_jdx_i}$$

(4) Now observe that this tells us that the order in which the higher derivatives are computed is irrelevant. That is, we can permute the order of our partial derivative computations, and a standard way of doing this is by differentiating first with respect to x_1, as many times as needed, then with respect to x_2, and so on. Thus, the collection of partial derivatives can be written, in a more convenient form, as follows:

$$\frac{d^kf}{dx_1^{k_1}\ldots dx_N^{k_N}}=\frac{d^{k_1}}{dx_1^{k_1}}\cdots\frac{d^{k_N}}{dx_N^{k_N}}(f)$$

(5) To be more precise, here $k\in\mathbb{N}$ is as usual the global order of our derivatives, the exponents $k_1,\ldots,k_N\in\mathbb{N}$ are subject to the condition $k_1+\ldots+k_N=k$, and the operations on the right are the familiar one-variable higher derivative operations.

(6) This being said, for certain tricky questions it is more convenient not to order the indices, or rather to order them according to what order best fits our computation, so what we have in the statement is the good formula, and (4-5) are mere remarks. □

All this is very nice, and as an illustration, let us work out in detail the case $k = 2$. Here things are quite special, and we can formulate the following definition:

Definition 15.14. *Given a twice differentiable function $f : \mathbb{R}^N \to \mathbb{R}$, we set*

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

which is a symmetric matrix, called Hessian matrix of f at the point $x \in \mathbb{R}^N$.

To be more precise, we know that when $f : \mathbb{R}^N \to \mathbb{R}$ is twice differentiable, its order $k = 2$ partial derivatives are the numbers in the statement. Now since these numbers naturally form a $N \times N$ matrix, the temptation is high to call this matrix $f''(x)$, and so we will do. And finally, we know from Clairaut that this matrix is symmetric:

$$f''(x)_{ij} = f''(x)_{ji}$$

Observe that at $N = 1$ this is compatible with the usual definition of the second derivative f'', because in this case, the 1×1 matrix from Definition 15.14 is:

$$f''(x) = (f''(x)) \in M_{1\times 1}(\mathbb{R})$$

As a word of warning, however, never use Definition 15.14 for functions $f : \mathbb{R}^N \to \mathbb{R}^M$, where the second derivative can only be something more complicated. Also, never attempt either to do something similar at $k = 3$ or higher, for functions $f : \mathbb{R}^N \to \mathbb{R}$ with $N > 1$, because again, that beast has too many indices, for being a true, honest matrix.

Back now to our usual business, approximation, we have the following result:

Theorem 15.15. *Given a twice differentiable function $f : \mathbb{R}^N \to \mathbb{R}$, we have*

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

where $f''(x) \in M_N(\mathbb{R})$ stands as usual for the Hessian matrix.

Proof. This is something more tricky, the idea being as follows:

(1) As a first observation, at $N = 1$ the Hessian matrix as constructed in Definition 15.14 is the 1×1 matrix having as entry the second derivative $f''(x)$, and the formula in the statement is something that we know well from basic calculus, namely:

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

(2) In general now, this is in fact something which does not need a new proof, because it follows from the one-variable formula above, applied to the restriction of f to the following segment in $\mathbb{R}^N$, which can be regarded as being a one-variable interval:

$$I = [x, x + t]$$

To be more precise, let $y \in \mathbb{R}^N$, and consider the following function, with $r \in \mathbb{R}$:

$$g(r) = f(x + ry)$$

We know from (1) that the Taylor formula for g, at the point $r = 0$, reads:

$$g(r) \simeq g(0) + g'(0)r + \frac{g''(0)r^2}{2}$$

And our claim is that, with $t = ry$, this is precisely the formula in the statement.

(3) So, let us see if our claim is correct. By using the chain rule, we have the following formula, with on the right, as usual, a row vector multiplied by a column vector:

$$g'(r) = f'(x + ry) \cdot y$$

By using again the chain rule, we can compute the second derivative as well:

$$\begin{aligned} g''(r) &= (f'(x+ry) \cdot y)' \\ &= \left(\sum_i \frac{df}{dx_i}(x+ry) \cdot y_i\right)' \\ &= \sum_i \sum_j \frac{d^2 f}{dx_i dx_j}(x+ry) \cdot \frac{d(x+ry)_j}{dr} \cdot y_i \\ &= \sum_i \sum_j \frac{d^2 f}{dx_i dx_j}(x+ry) \cdot y_i y_j \\ &= < f''(x+ry)y, y > \end{aligned}$$

(4) Time now to conclude. We know that we have $g(r) = f(x + ry)$, and according to our various computations above, we have the following formulae:

$$g(0) = f(x) \quad , \quad g'(0) = f'(x) \quad , \quad g''(0) =< f''(x)y, y >$$

Buit with this data in hand, the usual Taylor formula for our one variable function g, at order 2, at the point $r = 0$, takes the following form, with $t = ry$:

$$\begin{aligned} f(x+ry) &\simeq f(x) + f'(x)ry + \frac{< f''(x)y, y > r^2}{2} \\ &= f(x) + f'(x)t + \frac{< f''(x)t, t >}{2} \end{aligned}$$

Thus, we have obtained the formula in the statement.

(5) Finally, for completness, let us record as well a more numeric formulation of what we found. According to our usual rules for matrix calculus, what we found is:

$$f(x+t) \simeq f(x) + \sum_{i=1}^N \frac{df}{dx_i} t_i + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \frac{d^2 f}{dx_i dx_j} t_i t_j$$

Observe that, since the Hessian matrix $f''(x)$ is symmetric, most of the terms on the right will appear in pairs, making it clear what the $1/2$ is there for, namely avoiding redundancies. However, this is only true for the off-diagonal terms, so instead of further messing up our numeric formula above, we will just leave it like this. □

We can go back now to local extrema, and we have, improving Theorem 15.11:

THEOREM 15.16. *In order for a twice differentiable function $f:\mathbb{R}^N\to\mathbb{R}$ to have a local minimum or maximum at $x\in\mathbb{R}^N$, the first derivative must vanish there,*

$$f'(x)=0$$

and the Hessian must be positive or negative, in the sense that the quantities

$$<f''(x)t,t>\in\mathbb{R}$$

must keep a constant sign, positive or negative, when $t\in\mathbb{R}^N$ varies.

PROOF. This comes from Theorem 15.15. Consider indeed the formula there, namely:

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

We know from Theorem 15.11 that, in order for our function to have a local minimum or maximum at $x\in\mathbb{R}^N$, the first derivative must vanish there, $f'(x)=0$. Moreover, with this assumption made, the approximation that we have around x becomes:

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

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

As a conclusion to our study so far, our analytic questions lead us into a linear algebra question, regarding the square matrices of type $f''(x)\in M_N(\mathbb{R})$, and more specifically the positivity and negativity properties of the following quantities, when $t\in\mathbb{R}^N$ varies:

$$<f''(x)t,t>\in\mathbb{R}$$

This is actually a quite subtle question, and many things can be said here, after some linear algebra work. We will be back to this in a moment.

At higher order now, things become more complicated, as follows:

THEOREM 15.17. *Given an order k differentiable function $f:\mathbb{R}^N\to\mathbb{R}$, we have*

$$f(x+t)\simeq f(x)+f'(x)t+\frac{<f''(x)t,t>}{2}+\ldots$$

and this helps in identifying the local extrema, a bit as in the one-variable case.

PROOF. The study here is very similar to that at $k = 2$, from the proof of Theorem 15.15, with everything coming from the usual Taylor formula, applied on:

$$I = [x, x + t]$$

Thus, it is pretty much clear that we are led to the conclusion in the statement. We will leave some study here as an instructive exercise. □

Getting back now to Theorem 15.16, and the comments afterwards, we will need:

THEOREM 15.18. *Any matrix $A \in M_N(\mathbb{R})$ which is symmetric, $A = A^t$, is diagonalizable, with the diagonalization being of the following type,*

$$A = UDU^t$$

with $U \in M_N(\mathbb{R})$ orthogonal, and $D \in M_N(\mathbb{R})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because if we take a matrix of the form $A = UDU^t$, with U orthogonal and D diagonal, we have:

$$A^t = (UDU^t)^t = UDU^t = A$$

In the other sense now, assume that A is symmetric, $A = A^t$. Our first claim is that the eigenvalues are real. Indeed, assuming $Av = \lambda v$, we have:

$$\begin{aligned}
\lambda < v, v > &= < \lambda v, v > \\
&= < Av, v > \\
&= < v, Av > \\
&= < v, \lambda v > \\
&= \bar{\lambda} < v, v >
\end{aligned}$$

Thus we obtain $\lambda \in \mathbb{R}$, as claimed. Our next claim now is that the eigenspaces corresponding to different eigenvalues are pairwise orthogonal. Assume indeed that:

$$Av = \lambda v \quad , \quad Aw = \mu w$$

We have then the following computation, using $\lambda, \mu \in \mathbb{R}$:

$$\begin{aligned}
\lambda < v, w > &= < \lambda v, w > \\
&= < Av, w > \\
&= < v, Aw > \\
&= < v, \mu w > \\
&= \mu < v, w >
\end{aligned}$$

Thus $\lambda \neq \mu$ implies $v \perp w$, as claimed. In order now to finish the proof, it remains to prove that the eigenspaces of A span the whole space $\mathbb{R}^N$. For this purpose, we will use a recurrence method. Let us pick an eigenvector of our matrix:

$$Av = \lambda v$$

Assuming now that we have a vector w orthogonal to it, $v \perp w$, we have:

$$\begin{aligned} < Aw, v > &= < w, Av > \\ &= < w, \lambda v > \\ &= \lambda < w, v > \\ &= 0 \end{aligned}$$

Thus, if v is an eigenvector, then the vector space $v^\perp$ is invariant under A. We can therefore proceed by recurrence, and we obtain the result. □

Now back to our analysis questions, armed with the above linear algebra theorem, we have the following result, complementing what was said in Theorem 15.16:

Theorem 15.19. *Given a symmetric matrix $A \in M_N(\mathbb{R})$, as for instance a Hessian matrix $A = f''(x)$, with eigenvalues $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$, the following happen,*

(1) $< At, t > \geq 0$ *for any* $t \in \mathbb{R}^N$ *precisely when* $\lambda_1, \ldots, \lambda_N \geq 0$.
(2) $< At, t >> 0$ *for any* $t \neq 0$ *precisely when* $\lambda_1, \ldots, \lambda_N > 0$.
(3) $< At, t > \leq 0$ *for any* $t \in \mathbb{R}^N$ *precisely when* $\lambda_1, \ldots, \lambda_N \leq 0$.
(4) $< At, t >< 0$ *for any* $t \neq 0$ *precisely when* $\lambda_1, \ldots, \lambda_N < 0$.

and with this helping identifying the minima and maxima of functions $f : \mathbb{R}^N \to \mathbb{R}$.

Proof. This is something self-explanatory, coming from Theorem 15.18, and with the last assertion being something that we already know, from Theorem 15.16. □

As a comment, the above result is of course not the end of the story with the extrema of functions $f : \mathbb{R}^N \to \mathbb{R}$, because depending on how the Hessian $A = f''(x)$ looks like, we might been in need of a study at higher order, as suggested in Theorem 15.17. We will leave some exploration here, examples and conclusions, as an instructive exercise.

As another comment on all this, Theorem 15.18, which is something quite powerful, can be useful for a wide variety of other purposes, as follows:

(1) To start with, this can be used in order to fully justify what we said in chapter 8 about quadrics, with the Sylvester theorem used there coming from this.

(2) As another application, the adjacency matrices of the graphs considered in chapter 13 are symmetric, so Theorem 15.18 applies to these matrices too.

(3) In relation with the relativity theory computations in chapter 13, there is some underlying linear algebra there too, in connection with the curvature of spacetime.

Summarizing, many things to be learned, and as a rule of thumb, the more linear algebra you know, the better your mathematics and physics will be. And among the many things that you can do here, you can read about more general spectral theorems, for self-adjoint matrices, and for normal matrices, generalizing Theorem 15.18.

15c. Spherical coordinates

With the derivatives of the functions $f:\mathbb{R}^N\to\mathbb{R}^M$ understood, time now to discuss the integrals. Obviously, the integral of a function $f:\mathbb{R}^N\to\mathbb{R}^M$ can only be the vector of $\mathbb{R}^M$ formed by the integrals of its components $f_i:\mathbb{R}^N\to\mathbb{R}$, so in order to construct the integral, we can assume $M=1$. Thus, we are led to the following question:

QUESTION 15.20. *How to integrate the functions $f:\mathbb{R}^N\to\mathbb{R}$,*

$$f\to\int_{\mathbb{R}^N}f(z)dz$$

in analogy with what we know about integrating functions $f:\mathbb{R}\to\mathbb{R}$?

In answer, and taking $N=2$ for simplifying, I bet that your answer would be that we can define the multivariable integral by iterating, as follows:

$$\int_{\mathbb{R}^2}f(z)dz=\int_{\mathbb{R}}\int_{\mathbb{R}}f(x,y)dxdy$$

Which looks fine, at a first glance, but there is in fact a bug, with this. Indeed, assuming so, we would have by symmetry the following formula too:

$$\int_{\mathbb{R}^2}f(z)dz=\int_{\mathbb{R}}\int_{\mathbb{R}}f(x,y)dydx$$

Thus, and forgetting now about what we wanted to do, we can see that our method is based on the Fubini formula, stating that we must have:

$$\int_{\mathbb{R}}\int_{\mathbb{R}}f(x,y)dxdy=\int_{\mathbb{R}}\int_{\mathbb{R}}f(x,y)dydx$$

But, and here comes the point, this Fubini formula does not always work, with the counterexamples being not very difficult to construct, as follows:

THEOREM 15.21. *The Fubini formula, namely*

$$\int_{\mathbb{R}}\int_{\mathbb{R}}f(x,y)dxdy=\int_{\mathbb{R}}\int_{\mathbb{R}}f(x,y)dydx$$

can fail, for certain suitably chosen functions.

PROOF. We have indeed the following computation, no question about this:

$$\begin{aligned}\int_0^1\int_0^1\frac{y^2-x^2}{(x^2+y^2)^2}\,dxdy&=\int_0^1\left[\frac{x}{x^2+y^2}\right]_0^1dy\\&=\int_0^1\frac{1}{1+y^2}\,dy\\&=\frac{\pi}{4}\end{aligned}$$

On the other hand, by using this, and symmetry, we have as well:

$$\begin{aligned}\int_0^1\int_0^1\frac{y^2-x^2}{(x^2+y^2)^2}\,dydx&=\int_0^1\int_0^1\frac{x^2-y^2}{(x^2+y^2)^2}\,dxdy\\&=-\int_0^1\int_0^1\frac{y^2-x^2}{(x^2+y^2)^2}\,dxdy\\&=-\frac{\pi}{4}\end{aligned}$$

Thus Fubini can fail for certain functions, as said in the statement. Damn. □

What do do? Well, there is a mathematical answer to this, which is however something quite complicated, whose essentials can be summarized as follows:

Theorem 15.22 (Measure theory). *We can rigorously integrate the functions*

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

and assuming that f is measurable and integrable, in the sense that we have

$$\int_{\mathbb{R}^N}|f(z)|dz<\infty$$

we have the following equalities, for any decomposition $N=N_1+N_2$:

$$\int_{\mathbb{R}^{N_1}}\int_{\mathbb{R}^{N_2}}f(x,y)dydx=\int_{\mathbb{R}^{N_2}}\int_{\mathbb{R}^{N_1}}f(x,y)dxdy=\int_{\mathbb{R}^N}f(z)dz$$

Moreover, the same holds when $f:\mathbb{R}^N\to\mathbb{R}$ is assumed positive, and measurable.

Proof. This is something quite long and complicated, due to Lebesgue, Riesz, Borel, Fubini, Tonelli and others, traditionally learned in measure theory class. Alternatively, have a look at the first dozen pages of Rudin [74], which explain all this. □

Summarizing, we can talk about multiple integrals. Getting now to the general theory and rules, for computing such integrals, the key result here is the change of variable formula. In order to discuss this, let us start with something that we know well, in 1D:

Proposition 15.23. *We have the change of variable formula*

$$\int_a^b f(x)dx=\int_c^d f(\varphi(t))\varphi'(t)dt$$

where $c=\varphi^{-1}(a)$ and $d=\varphi^{-1}(b)$.

Proof. This follows with $f=F'$, via the following differentiation rule:

$$(F\varphi)'(t)=F'(\varphi(t))\varphi'(t)$$

Indeed, by integrating between c and d, we obtain the result. □

In several variables now, we can only expect the above $\varphi'(t)$ factor to be replaced by something similar, a sort of "derivative of φ, arising as a real number". But this can only be the Jacobian $\det(\varphi'(t))$, and with this in mind, we are led to:

THEOREM 15.24. *Given a transformation $\varphi = (\varphi_1, \ldots, \varphi_N)$, we have*

$$\int_E f(x)dx = \int_{\varphi^{-1}(E)} f(\varphi(t))|J_\varphi(t)|dt$$

with the J_φ quantity, called Jacobian, being given by

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

and with this generalizing the formula from Proposition 15.23.

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

(1) Observe first that this generalizes indeed the change of variable formula in 1 dimension, from Proposition 15.23, the point here being that the absolute value on the derivative appears as to compensate for the lack of explicit bounds for the integral.

(2) As a second observation, we can assume if we want, by linearity, that we are dealing with the constant function $f = 1$. For this function, our formula reads:

$$vol(E) = \int_{\varphi^{-1}(E)} |J_\varphi(t)|dt$$

In terms of $D = \varphi^{-1}(E)$, this amounts in proving that we have:

$$vol(\varphi(D)) = \int_D |J_\varphi(t)|dt$$

Now since this latter formula is additive with respect to D, it is enough to prove it for small cubes D. And here, as a first remark, our formula is clear for the linear maps φ, by using the definition of the determinant of real matrices, as a signed volume.

(3) However, the extension of this to the case of non-linear maps φ is something which looks non-trivial, so we will not follow this path, in what follows. So, while the above $f = 1$ discussion is certainly something nice, our theorem is still in need of a proof.

(4) In order to prove the theorem, as stated, let us rather focus on the transformations used φ, instead of the functions to be integrated f. Our first claim is that the validity of the theorem is stable under taking compositions of such transformations φ.

(5) In order to prove this claim, consider a composition, as follows:

$$\varphi : E \to F \quad , \quad \psi : D \to E \quad , \quad \varphi \circ \psi : D \to F$$

Assuming that the theorem holds for φ, ψ, we have the following computation:

$$\begin{aligned}\int_F f(x)dx &= \int_E f(\varphi(s))|J_\varphi(s)|ds\\ &= \int_D f(\varphi\circ\psi(t))|J_\varphi(\psi(t))|\cdot|J_\psi(t)|dt\\ &= \int_D f(\varphi\circ\psi(t))|J_{\varphi\circ\psi}(t)|dt\end{aligned}$$

Thus, our theorem holds as well for $\varphi\circ\psi$, and we have proved our claim.

(6) Next, as a key ingredient, let us examine the case where we are in $N=2$ dimensions, and our transformation φ has one of the following special forms:

$$\varphi(x,y)=(\psi(x,y),y)\quad,\quad\varphi(x,y)=(x,\psi(x,y))$$

By symmetry, it is enough to deal with the first case. Here the Jacobian is $d\psi/dx$, and by replacing if needed $\psi\to-\psi$, we can assume that this Jacobian is positive, $d\psi/dx>0$. Now by assuming as before that $D=\varphi^{-1}(E)$ is a rectangle, $D=[a,b]\times[c,d]$, we can prove our formula by using the change of variables in 1 dimension, as follows:

$$\begin{aligned}\int_E f(s)ds &= \int_{\varphi(D)} f(x,y)dxdy\\ &= \int_c^d\int_{\psi(a,y)}^{\psi(b,y)} f(x,y)dxdy\\ &= \int_c^d\int_a^b f(\psi(x,y),y)\frac{d\psi}{dx}\,dxdy\\ &= \int_D f(\varphi(t))J_\varphi(t)dt\end{aligned}$$

(7) But with this, we can now prove the theorem, in $N=2$ dimensions. Indeed, given a transformation $\varphi=(\varphi_1,\varphi_2)$, consider the following two transformations:

$$\phi(x,y)=(\varphi_1(x,y),y)\quad,\quad\psi(x,y)=(x,\varphi_2\circ\phi^{-1}(x,y))$$

We have then $\varphi=\psi\circ\phi$, and by using (6) for ψ,ϕ, which are of the special form there, and then (3) for composing, we conclude that the theorem holds for φ, as desired.

(8) Thus, theorem proved in $N=2$ dimensions, and the extension of the above proof to arbitrary N dimensions is straightforward, that we will leave this as an exercise. □

And with this, good news, we have all the needed integration tools in our bag. To be more precise, still missing would be an analogue of the fundamental theorem of calculus, but in several variables this is something fairly complicated, related to physics.

Time now do some exciting computations, with the technology that we have. In what regards the applications of Theorem 15.24, these often come via:

PROPOSITION 15.25. *We have polar coordinates in* 2 *dimensions,*

$$\begin{cases}x &= r\cos t\\ y &= r\sin t\end{cases}$$

the corresponding Jacobian being $J = r$.

PROOF. This is elementary, the Jacobian being:

$$\begin{aligned}J &= \begin{vmatrix}\frac{d(r\cos t)}{dr} & \frac{d(r\cos t)}{dt}\\ \\ \frac{d(r\sin t)}{dr} & \frac{d(r\sin t)}{dt}\end{vmatrix}\\ &= \begin{vmatrix}\cos t & -r\sin t\\ \sin t & r\cos t\end{vmatrix}\\ &= r\cos^2 t + r\sin^2 t\\ &= r\end{aligned}$$

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

We can now compute the Gauss integral, which is the best calculus formula ever:

THEOREM 15.26. *We have the following formula,*

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

called Gauss integral formula.

PROOF. Let I be the above integral. By using polar coordinates, we obtain:

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

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

Moving now to 3 dimensions, we have here the following result:

Proposition 15.27. *We have spherical coordinates in 3 dimensions,*

$$\begin{cases}x&=&r\cos s\\ y&=&r\sin s\cos t\\ z&=&r\sin s\sin t\end{cases}$$

the corresponding Jacobian being $J(r,s,t)=r^2\sin s$.

Proof. The fact that we have indeed spherical coordinates is clear. Regarding now the Jacobian, this is given by the following formula:

$$\begin{aligned}&J(r,s,t)\\
=&\begin{vmatrix}\cos s&-r\sin s&0\\ \sin s\cos t&r\cos s\cos t&-r\sin s\sin t\\ \sin s\sin t&r\cos s\sin t&r\sin s\cos t\end{vmatrix}\\
=&r^2\sin s\sin t\begin{vmatrix}\cos s&-r\sin s\\ \sin s\sin t&r\cos s\sin t\end{vmatrix}+r\sin s\cos t\begin{vmatrix}\cos s&-r\sin s\\ \sin s\cos t&r\cos s\cos t\end{vmatrix}\\
=&r\sin s\sin^2 t\begin{vmatrix}\cos s&-r\sin s\\ \sin s&r\cos s\end{vmatrix}+r\sin s\cos^2 t\begin{vmatrix}\cos s&-r\sin s\\ \sin s&r\cos s\end{vmatrix}\\
=&r\sin s(\sin^2 t+\cos^2 t)\begin{vmatrix}\cos s&-r\sin s\\ \sin s&r\cos s\end{vmatrix}\\
=&r\sin s\times 1\times r\\
=&r^2\sin s\end{aligned}$$

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

Let us work out now the general spherical coordinate formula, in arbitrary N dimensions. The formula here, which generalizes those at $N=2,3$, is as follows:

Theorem 15.28. *We have spherical coordinates in N dimensions,*

$$\begin{cases}x_1&=&r\cos t_1\\ x_2&=&r\sin t_1\cos t_2\\ \vdots\\ x_{N-1}&=&r\sin t_1\sin t_2\ldots\sin t_{N-2}\cos t_{N-1}\\ x_N&=&r\sin t_1\sin t_2\ldots\sin t_{N-2}\sin t_{N-1}\end{cases}$$

the corresponding Jacobian being given by the following formula,

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

and with this generalizing the known formulae at $N=2,3$.

Proof. As before, the fact that we have spherical coordinates is clear. Regarding now the Jacobian, also as before, by developing over the last column, we have:

$$\begin{aligned} J_N &= r\sin t_1\ldots\sin t_{N-2}\sin t_{N-1}\times\sin t_{N-1}J_{N-1}\\ &+ r\sin t_1\ldots\sin t_{N-2}\cos t_{N-1}\times\cos t_{N-1}J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}(\sin^2 t_{N-1}+\cos^2 t_{N-1})J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}J_{N-1} \end{aligned}$$

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

As a comment here, the above convention for spherical coordinates is one among many, designed to best work in arbitrary N dimensions. Also, in what regards the precise range of the angles $t_1,\ldots,t_{N-1}$, we will leave this to you, as an instructive exercise.

As an application, let us compute the volumes of spheres. For this purpose, we must understand how the products of coordinates integrate over spheres. Let us start with the case $N=2$. Here the sphere is the unit circle $\mathbb{T}$, and with $z=e^{it}$ the coordinates are $\cos t,\sin t$. We can first integrate arbitrary powers of these coordinates, as follows:

Theorem 15.29 (Wallis). *We have the following formulae,*

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

where $\varepsilon(p)=1$ if p is even, and $\varepsilon(p)=0$ if p is odd, and where

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

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

Proof. Let us first compute the integral on the left in the statement:

$$I_p=\int_0^{\pi/2}\cos^p t\,dt$$

We do this by partial integration. We have the following formula:

$$\begin{aligned} (\cos^p t\sin t)' &= p\cos^{p-1}t(-\sin t)\sin t+\cos^p t\cos t\\ &= p\cos^{p+1}t-p\cos^{p-1}t+\cos^{p+1}t\\ &= (p+1)\cos^{p+1}t-p\cos^{p-1}t \end{aligned}$$

By integrating between 0 and $\pi/2$, we obtain the following formula:

$$(p+1)I_{p+1}=pI_{p-1}$$

Thus we can compute I_p by recurrence, and we obtain:

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

But $I_0=\frac{\pi}{2}$ and $I_1=1$, so we get the result. As for the second formula, this follows from the first one, with $t=\frac{\pi}{2}-s$. Thus, we have proved both formulae in the statement. □

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

Theorem 15.30. *The volume of the unit sphere in $\mathbb{R}^N$ is given by*

$$V=\left(\frac{\pi}{2}\right)^{[N/2]}\frac{2^N}{(N+1)!!}$$

with our usual convention $N!!=(N-1)(N-3)(N-5)\ldots$

Proof. Let us denote by B^+ the positive part of the unit sphere, or rather unit ball B, obtained by cutting this unit ball in 2^N parts. At the level of volumes, we have:

$$V=2^NV^+$$

We have the following computation, using spherical coordinates:

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

Thus, we obtain the formula in the statement. □

15d. Normal variables

We have kept the best for the end. By using the Gauss formula $\int_{\mathbb{R}} e^{-x^2} = \sqrt{\pi}$ from Theorem 15.26, we can now introduce the normal laws, as follows:

Definition 15.31. *The normal law of parameter* 1 *is the following measure:*

$$g_1 = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

More generally, the normal law of parameter $t > 0$ *is the following measure:*

$$g_t = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t} dx$$

These are also called Gaussian distributions, with "g" standing for Gauss.

Observe that the above laws have indeed mass 1, as they should. This follows indeed from the Gauss formula, which gives, with $x = \sqrt{2t}\, y$:

$$\begin{aligned}
\int_{\mathbb{R}} e^{-x^2/2t} dx &= \int_{\mathbb{R}} e^{-y^2} \sqrt{2t}\, dy \\
&= \sqrt{2t} \int_{\mathbb{R}} e^{-y^2} dy \\
&= \sqrt{2t} \times \sqrt{\pi} \\
&= \sqrt{2\pi t}
\end{aligned}$$

Generally speaking, the normal laws appear as bit everywhere, in real life. The reasons behind this phenomenon come from the Central Limit Theorem (CLT):

Theorem 15.32 (CLT). *Given random variables* $f_1, f_2, f_3, \ldots \in L^\infty(X)$ *which are i.i.d., centered, and with variance* $t > 0$, *we have, with* $n \to \infty$, *in moments,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n f_i \sim g_t$$

with g_t *being the Gaussian law of parameter* t.

Proof. The Fourier transform $F_f(x) = E(e^{ixf})$ is given by the following formula:

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

Thus, the Fourier transform of the variable in the statement is given by:

$$\begin{aligned}
F(x) &= \left[F_f\left(\frac{x}{\sqrt{n}}\right)\right]^n\\
&= \left[1-\frac{tx^2}{2n}+O(n^{-2})\right]^n\\
&\simeq \left[1-\frac{tx^2}{2n}\right]^n\\
&\simeq e^{-tx^2/2}
\end{aligned}$$

On the other hand, the Fourier transform of g_t is given by:

$$\begin{aligned}
F_{g_t}(x) &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} e^{-y^2/2t+ixy}dy\\
&= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} e^{-(y/\sqrt{2t}-\sqrt{t/2}ix)^2-tx^2/2}dy\\
&= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} e^{-z^2-tx^2/2}\sqrt{2t}dz\\
&= \frac{1}{\sqrt{\pi}}e^{-tx^2/2}\int_{\mathbb{R}} e^{-z^2}dz\\
&= \frac{1}{\sqrt{\pi}}e^{-tx^2/2}\cdot\sqrt{\pi}\\
&= e^{-tx^2/2}
\end{aligned}$$

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

15e. Exercises

Welcome to multivariable calculus, such a joy, and as exercises here, we have:

Exercise 15.33. *Clarify everything that we said, in relation with continuity.*

Exercise 15.34. *Learn some other formulations of the chain rule formula.*

Exercise 15.35. *Try writing a Taylor formula at order* 3*, in several variables.*

Exercise 15.36. *Learn some other proofs of the change of variable formula.*

Exercise 15.37. *Clarify the range and meaning of spherical coordinate angles.*

Exercise 15.38. *Learn about the Stirling formula, and its applications.*

Exercise 15.39. *Learn about further Wallis formulae, and hyperspherical laws.*

Exercise 15.40. *Learn more about the CLT, and its various applications.*

As bonus exercise, as previously suggested, learn more linear algebra.

CHAPTER 16

Physics, equations

16a. Gravity basics

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

FACT 16.1. *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 before. And we are looking at a particle whose position is described by a function $x = x(t)$. Then, it is quite clear that the speed of this particle should be described by the first derivative $v = x'(t)$, and that the acceleration of the particle should be described by the second derivative $a = v'(t) = x''(t)$.

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

THEOREM 16.2. *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:

$$\begin{tikzcd} \circ_m \arrow[d] \\ \bullet_M \end{tikzcd}$$

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 exterted by M, which is fixed in our formalism, on the object m which moves, is subject to the following equations:

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

To be more precise, in the first equation $G\simeq 6.674\times 10^{-11}$ is the gravitational constant, in usual SI units, and the sign is $-$ because F is attractive. The second equation is something standard and very intuitive, and the last two equations are those from Fact 16.1. Now observe that, with the above data for F, the equation $F=ma$ reads:

$$-G\cdot\frac{Mm}{x^2}=m\ddot{x}$$

Thus, by simplifying, we are led to the equation in the statement. □

In two dimensions now, we first have the following result:

THEOREM 16.3. *In the context of a free fall from distance $x_0=R>>0$, with initial velocity $v_0=0$, the equation of the trajectory is*

$$x\simeq R-\frac{gt^2}{2}$$

with the constant being $g=GM/R^2$, called gravity of M, at distance R from it.

PROOF. As before, the equation for gravity is as follows, with $K=GM$:

$$\ddot{x}=-\frac{K}{d^2}\cdot\frac{x}{||x||}=-\frac{Kx}{||x||^3}$$

In one dimension now, things get simpler, and the equation of motion reads:

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

Since we assumed $R>>0$, we must look for a solution of type $x\simeq R+ct^2$, with the lack of the t term coming from $v_0=0$. But with $x\simeq R+ct^2$, our equation reads:

$$2c\simeq-\frac{K}{R^2}$$

Now by multiplying by $t^2/2$, and adding R, we obtain as solution:

$$x\simeq R-\frac{Kt^2}{2R^2}$$

Thus, we have indeed $x\simeq R-gt^2/2$, with g being the following number:

$$g=\frac{K}{R^2}=\frac{GM}{R^2}$$

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

As an illustration for the above result, let us do a numeric terrestrial check, based on it. The gravitational constant, the mass of the Earth, and the average radius of the Earth are as follows, expressed as usual in meters and kilograms:

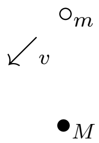
$$G = 6.674 \times 10^{-11} \quad , \quad M = 5.972 \times 10^{24} \quad , \quad R = 6.371 \times 10^6$$

We obtain the following value for the number g computed above:

$$g = \frac{6.674 \times 5.972}{6.371 \times 6.371} \times 10 = 9.819$$

Which is quite decent, when compared to the observed value, $g = 9.806$.

As a second toy example now for our 3D gravitation theory, which is more advanced, lying somewhere between 1D and 2D, let us add an arbitrary initial speed $v_0 = v$ to the above situation, which in addition is allowed to be a vector in $\mathbb{R}^2$, as follows:



We obtain in this way the following generalization of Theorem 16.3:

Theorem 16.4. *In the context of a free fall from distance $x_0 = R >> 0$, with initial plane velocity vector $v_0 = v$, the equation of the trajectory is*

$$x \simeq R + vt - \frac{gt^2}{2}$$

where $g = GM/R^2$ as usual, and with the quantities R, g in the above being regarded now as vectors, pointing upwards. The approximate trajectory is a parabola.

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

(1) Let us first discuss the simpler case where we are still in 1D, as in Theorem 16.3, but with an initial velocity $v_0 = v$ added. In order to find the equation of motion, we can just redo the computations from the proof of Theorem 16.3, with now looking for a general solution of type $x \simeq R + vt + ct^2$, and we get, as stated above:

$$x \simeq R + vt - \frac{gt^2}{2}$$

Alternatively, we can simply argue that, by linearity, what we have to do is to take the solution $x \simeq R - gt^2/2$ found in Theorem 16.3, and add an extra vt term to it.

(2) In the general 2D case now, where the initial velocity $v_0 = v$ is a vector in $\mathbb{R}^2$, the same arguments apply, either by redoing the computations from the proof of Theorem 16.3, or simply by arguing that by linearity we can just take the solution $x \simeq R - gt^2/2$ found there, and add an extra vt term to it. Thus, we have our solution.

(3) Let us study now the solution that we found. In standard (x,y) coordinates, with $v=(p,q)$, and with R,g being now back scalars, our solution looks as follows:

$$x=pt \quad , \quad y\simeq R+qt-\frac{gt^2}{2}$$

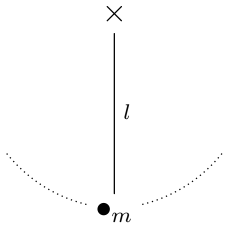
From the first equation we get $t=x/p$, and by substituting into the second:

$$y\simeq R+\frac{qx}{p}-\frac{gx^2}{2p^2}$$

We recognize here the approximate equation of a parabola, and we are done. □

Along the same lines, we can discuss as well confined motion, under a uniform gravitational field. Let us start our discussion with something very basic, namely:

Definition 16.5. *A simple pendulum is a device of type*



consisting of a bob of mass m, attached to a rigid rod of length l.

In order to study the physics of the pendulum, which can easily lead to a lot of complicated computations, when approached with bare hands, the most convenient is to use the notion of energy. For a particle moving under the influence of a force F, the position x, speed v and acceleration a are related by the following formulae:

$$v=\dot{x} \quad , \quad a=\dot{v}=\ddot{x} \quad , \quad F=ma$$

The kinetic energy of our particle is then given by the following formula:

$$T=\frac{mv^2}{2}$$

By differentiating with respect to time t, we obtain the following formula:

$$\dot{T}=mv\dot{v}=mva=Fv$$

Now by integrating, also with respect to t, this gives the following formula:

$$T=\int Fvdt=\int F\dot{x}dt=\int Fdx$$

But this suggests to define the potential energy V by the following formula, up to a constant, with the derivative being with respect to the space variable x:

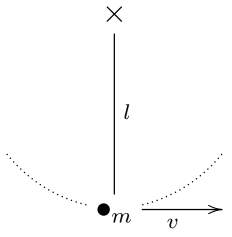
$$V'=-F$$

Indeed, we know from the above that we have $T' = F$, so if we define the total energy to be $E = T + V$, then this total energy is constant, as shown by:

$$E' = T' + V' = 0$$

Very nice all this, and by getting back now to the pendulum from Definition 16.5, we can have this understood with not many computations involved, as follows:

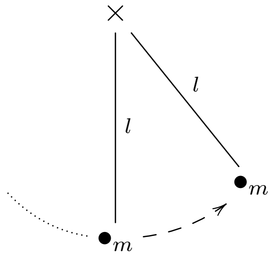
THEOREM 16.6. *For a pendulum starting with speed v from the equilibrium position,*



the motion will be confined if $v^2 < 4gl$, and circular if $v^2 > 4gl$.

PROOF. There are many ways of proving this result, along with working out several other useful related formulae, for which we will refer to the proof below, and with a quite elegant approach to this, using no computations or almost, being as follows:

(1) Let us first examine what happens when the bob has traveled an angular distance $\theta > 0$, with respect to the vertical. The picture here is as follows:



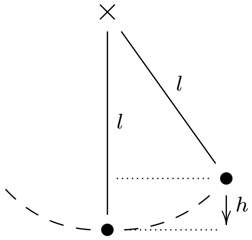
The distance traveled is then $x = l\theta$. As for the force acting, this is $F_{total} = mg$ oriented downwards, with the component alongside x being given by:

$$\begin{aligned} F &= -\|F_{total}\| \sin\theta \\ &= -mg\sin\theta \\ &= -mg\sin\left(\frac{x}{l}\right) \end{aligned}$$

(2) But with this, we can compute the potential energy. With the convention that this vanishes at the equilibrium position, $V(0)=0$, we obtain the following formula:

$$\begin{aligned}
V'=-F \quad &\implies \quad V'=mg\sin\left(\frac{x}{l}\right)\\
&\implies \quad V=mgl\left(1-\cos\left(\frac{x}{l}\right)\right)\\
&\implies \quad V=mgl(1-\cos\theta)
\end{aligned}$$

(3) Alternatively, in case this sounds too wizarding, we can compute the potential energy in the old fashion, by letting the bob fall, the picture being as follows:



The height of the fall is then $h=l-l\cos\theta$, and since for this fall the force is constant, $\mathcal{F}=-mg$, we obtain the following formula for the potential energy:

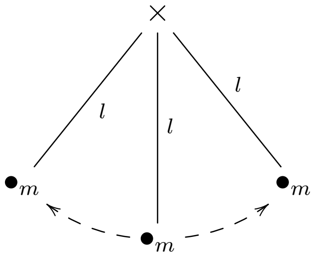
$$\begin{aligned}
V'=-\mathcal{F} \quad &\implies \quad V'=mg\\
&\implies \quad V=mgh\\
&\implies \quad V=mgl(1-\cos\theta)
\end{aligned}$$

Summarizing, one way or another we have our formula for the potential energy V.

(4) Now comes the discussion. The motion will be confined when the initial kinetic energy, namely $E=mv^2/2$, satisfies the following condition:

$$\begin{aligned}
E<\sup_\theta V=2mgl \quad &\iff \quad \frac{mv^2}{2}<2mgl\\
&\iff \quad v^2<4gl
\end{aligned}$$

In this case, the motion will be confined between two angles $-\theta,\theta$, as follows:



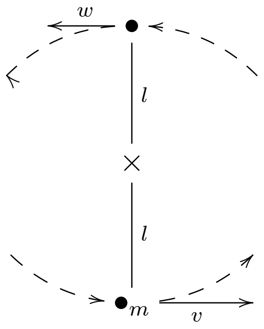
To be more precise here, the two extreme angles $-\theta,\theta\in(-\pi,\pi)$ can be explicitly computed, as being solutions of the following equation:

$$\begin{aligned}V=E\quad&\iff\quad mgl(1-\cos\theta)=\frac{mv^2}{2}\\&\iff\quad 1-\cos\theta=\frac{v^2}{2gl}\end{aligned}$$

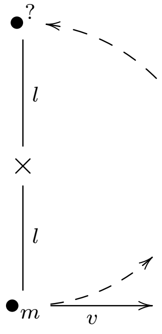
(5) Regarding now the case $v^2>4gl$, here the bob will certainly reach the upwards position, with the speed $w>0$ there being given by the following formula:

$$\begin{aligned}\frac{mw^2}{2}=E-2mgl\quad&\implies\quad\frac{mw^2}{2}=\frac{mv^2}{2}-2mgl\\&\implies\quad w^2=v^2-4gl\\&\implies\quad w=\sqrt{v^2-4gl}\end{aligned}$$

Thus, with the convention in the statement for v, that is, going to the right, the motion of the pendulum will be counterclockwise circular, and perpetual:



(6) Finally, in the case $v^2=4gl$, the bob will also reach the upwards position, but with speed $w=0$ there, and then, at least theoretically, will remain there:



(7) Actually, it is quite interesting in this latter situation, $v^2=4gl$, to further speculate on what can happen, when making our problem more realistic. For instance, we can add to our setting the assumption that when the bob is stuck on top, with speed 0, there is a

33% chance for it to keep going, to the left, a 33% chance for it to come back, to the right, and a 33% chance for it to remain stuck. In this case there are infinitely many possible trajectories, which are best investigated by using probability. Welcome to chaos.

(8) As a final comment, yes I know that the figures in (7) don't add up to 100%. This is because there is as well a remaining 1% possibility, where a relativistic black cat appears, with a continuous effect on the bob, via a paw slap, when on top, with speed $w' \in (0.3c, 0.7c)$, with c being the speed of light. In this case, the set of possible trajectories becomes uncountable, and is again best investigated by using probability. □

And good news, done with the pendulum. Never ever will we be scared by it, all the above was very nice, and the continuation of this chapter will be the same, nice too.

16b. Kepler and Newton

Getting now to the real thing, astronomy, the result here, which is the pride of mathematics, physics, and human knowledge in general, is the following theorem:

Theorem 16.7 (Kepler, Newton). *Planets and other celestial bodies move around the Sun on conics, that is, on curves of type*

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

with $P \in \mathbb{R}[x,y]$ being of degree 2. The same is true for any body moving around another body, provided that we are not in the situation of a free fall.

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

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

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

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

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

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

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

(2) Obviously, the problem happens in 2 dimensions, and you can even find, as an exercise, a formal proof of that, based on the above equation. Now here the most convenient is to use standard x, y coordinates, and denote our point as $z = (x, y)$. With this change made, and by setting $K = GM$, the equation of motion becomes:

$$\ddot{z} = -\frac{Kz}{||z||^3}$$

In other words, in terms of the coordinates x, y, the equations are:

$$\ddot{x} = -\frac{Kx}{(x^2+y^2)^{3/2}} \quad , \quad \ddot{y} = -\frac{Ky}{(x^2+y^2)^{3/2}}$$

(3) Let us begin with a simple particular case, that of the circular solutions. To be more precise, we are interested in solutions of the following type:

$$x = r\cos\alpha t \quad , \quad y = r\sin\alpha t$$

In this case we have $||z|| = r$, so our equation of motion becomes:

$$\ddot{z} = -\frac{Kz}{r^3}$$

On the other hand, differentiating x, y leads to the following formula:

$$\ddot{z} = (\ddot{x}, \ddot{y}) = -\alpha^2(x, y) = -\alpha^2 z$$

Thus, we have a circular solution when the parameters r, α satisfy:

$$r^3\alpha^2 = K$$

(4) In the general case now, the problem can be solved via some calculus. Let us write indeed our vector $z = (x, y)$ in polar coordinates, as follows:

$$x = r\cos\theta \quad , \quad y = r\sin\theta$$

We have then $||z|| = r$, and our equation of motion becomes, as in (3):

$$\ddot{z} = -\frac{Kz}{r^3}$$

Let us differentiate now x, y. By using the standard calculus rules, we have:

$$\dot{x} = \dot{r}\cos\theta - r\sin\theta\cdot\dot{\theta}$$

$$\dot{y} = \dot{r}\sin\theta + r\cos\theta\cdot\dot{\theta}$$

Differentiating one more time gives the following formulae:

$$\ddot{x} = \ddot{r}\cos\theta - 2\dot{r}\sin\theta\cdot\dot{\theta} - r\cos\theta\cdot\dot{\theta}^2 - r\sin\theta\cdot\ddot{\theta}$$

$$\ddot{y} = \ddot{r}\sin\theta + 2\dot{r}\cos\theta\cdot\dot{\theta} - r\sin\theta\cdot\dot{\theta}^2 + r\cos\theta\cdot\ddot{\theta}$$

Consider now the following two quantities, appearing as coefficients in the above:

$$a = \ddot{r} - r\dot{\theta}^2 \quad , \quad b = 2\dot{r}\dot{\theta} + r\ddot{\theta}$$

In terms of these quantities, our second derivative formulae read:

$$\ddot{x} = a\cos\theta - b\sin\theta$$

$$\ddot{y} = a\sin\theta + b\cos\theta$$

(5) We can now solve the equation of motion from (4). Indeed, with the formulae that we found for $\ddot{x}, \ddot{y}$, our equation of motion takes the following form:

$$a\cos\theta - b\sin\theta = -\frac{K}{r^2}\cos\theta$$

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

But these two formulae can be written in the following way:

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

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

By making now the product, and assuming that we are in a non-degenerate case, where the angle θ varies indeed, we obtain by positivity that we must have:

$$a + \frac{K}{r^2} = b = 0$$

(6) We are almost there. Let us first examine the second equation, $b = 0$. Remembering who b is, from (4), this equation can be solved as follows:

$$\begin{aligned} b = 0 \quad &\iff \quad 2\dot{r}\dot{\theta} + r\ddot{\theta} = 0 \\ &\iff \quad \frac{\ddot{\theta}}{\dot{\theta}} = -2\frac{\dot{r}}{r} \\ &\iff \quad (\log\dot{\theta})' = (-2\log r)' \\ &\iff \quad \log\dot{\theta} = -2\log r + c \\ &\iff \quad \dot{\theta} = \frac{\lambda}{r^2} \end{aligned}$$

As for the first equation the we found, namely $a + K/r^2 = 0$, remembering from (4) that a was by definition given by $a = \ddot{r} - r\dot{\theta}^2$, this equation now becomes:

$$\ddot{r} - \frac{\lambda^2}{r^3} + \frac{K}{r^2} = 0$$

(7) As a conclusion to all this, in polar coordinates, $x = r\cos\theta$, $y = r\sin\theta$, our equations of motion are as follows, with λ being a constant, not depending on t:

$$\ddot{r} = \frac{\lambda^2}{r^3} - \frac{K}{r^2} \quad , \quad \dot{\theta} = \frac{\lambda}{r^2}$$

Even better now, by writing $K=\lambda^2/c$, these equations read:

$$\ddot{r}=\frac{\lambda^2}{r^2}\left(\frac{1}{r}-\frac{1}{c}\right)\quad,\quad\dot{\theta}=\frac{\lambda}{r^2}$$

(8) As an illustration, let us quickly work out the case of a circular motion, where r is constant. Here $\ddot{r}=0$, so the first equation gives $c=r$. Also we have $\dot{\theta}=\alpha$, with:

$$\alpha=\frac{\lambda}{r^2}$$

Assuming $\theta=0$ at $t=0$, from $\dot{\theta}=\alpha$ we obtain $\theta=\alpha t$, and so, as in (3) above:

$$x=r\cos\alpha t\quad,\quad y=r\sin\alpha t$$

Observe also that the condition found in (3) is indeed satisfied:

$$r^3\alpha^2=\frac{\lambda^2}{r}=\frac{\lambda^2}{c}=K$$

(9) Back to the general case now, our claim is that we have the following formula, for the distance $r=r(t)$ as function of the angle $\theta=\theta(t)$, for some $\varepsilon,\delta\in\mathbb{R}$:

$$r=\frac{c}{1+\varepsilon\cos\theta+\delta\sin\theta}$$

Let us first check that this formula works indeed. With r being as above, and by using our second equation found before, $\dot{\theta}=\lambda/r^2$, we have the following computation:

$$\begin{aligned}\dot{r}&=\frac{c(\varepsilon\sin\theta-\delta\cos\theta)\dot{\theta}}{(1+\varepsilon\cos\theta+\delta\sin\theta)^2}\\&=\frac{\lambda c(\varepsilon\sin\theta-\delta\cos\theta)}{r^2(1+\varepsilon\cos\theta+\delta\sin\theta)^2}\\&=\frac{\lambda(\varepsilon\sin\theta-\delta\cos\theta)}{c}\end{aligned}$$

Thus, the second derivative of the above function r is given, as desired, by:

$$\begin{aligned}\ddot{r}&=\frac{\lambda(\varepsilon\cos\theta+\delta\sin\theta)\dot{\theta}}{c}\\&=\frac{\lambda^2(\varepsilon\cos\theta+\delta\sin\theta)}{r^2c}\\&=\frac{\lambda^2}{r^2}\left(\frac{1}{r}-\frac{1}{c}\right)\end{aligned}$$

(10) The above check was something quite informal, and now we must prove that our formula is indeed the correct one. For this purpose, we use a trick. Let us write:

$$r(t)=\frac{1}{f(\theta(t))}$$

Abbreviated, and by always reminding that f takes $\theta = \theta(t)$ as variable, this reads:

$$r = \frac{1}{f}$$

With the convention that dots mean as usual derivatives with respect to t, and that the primes will denote derivatives with respect to $\theta = \theta(t)$, we have:

$$\dot{r} = -\frac{f'\dot{\theta}}{f^2} = -\frac{f'}{f^2} \cdot \frac{\lambda}{r^2} = -\lambda f'$$

By differentiating one more time with respect to t, we obtain:

$$\ddot{r} = -\lambda f''\dot{\theta} = -\lambda f'' \cdot \frac{\lambda}{r^2} = -\frac{\lambda^2}{r^2} f''$$

On the other hand, our equation for $\ddot{r}$ found in (7) reads:

$$\ddot{r} = \frac{\lambda^2}{r^2}\left(\frac{1}{r} - \frac{1}{c}\right) = \frac{\lambda^2}{r^2}\left(f - \frac{1}{c}\right)$$

Thus, in terms of $f = 1/r$ as above, our equation for $\ddot{r}$ simply reads:

$$f'' + f = \frac{1}{c}$$

But this latter equation is elementary to solve. Indeed, both functions $\cos t, \sin t$ satisfy g” $+ g = 0$, so any linear combination of them satisfies as well this equation. But the solutions of $f'' + f = 1/c$ being those of $g'' + g = 0$ shifted by $1/c$, we obtain:

$$f = \frac{1 + \varepsilon \cos\theta + \delta \sin\theta}{c}$$

Now by inverting, we obtain the formula announced in (9), namely:

$$r = \frac{c}{1 + \varepsilon \cos\theta + \delta \sin\theta}$$

(11) But this leads to the conclusion that the trajectory is a conic. Indeed, in terms of the parameter θ, the formulae of the coordinates are:

$$x = \frac{c\cos\theta}{1 + \varepsilon \cos\theta + \delta \sin\theta}$$

$$y = \frac{c\sin\theta}{1 + \varepsilon \cos\theta + \delta \sin\theta}$$

But these are precisely the equations of conics in polar coordinates.

(12) To be more precise, in order to find the precise equation of the conic, observe that the two functions x, y that we found above satisfy the following formula:

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

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

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

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

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

But what we have here is an equation of a conic, as claimed. □

The above was theory, and for further applications, here is a sort of "best of" the formulae found in the proof of Theorem 16.7, which are all very useful in practice:

Theorem 16.8 (Kepler, Newton). *In the context of a 2-body problem, with M fixed at 0, and m starting its movement from Ox, the equation of motion of m, namely*

$$\ddot{z}=-\frac{Kz}{||z||^3}$$

with $K=GM$, and $z=(x,y)$, becomes in polar coordinates, $x=r\cos\theta$, $y=r\sin\theta$,

$$\ddot{r}=\frac{\lambda^2}{r^2}\left(\frac{1}{r}-\frac{1}{c}\right)\quad,\quad\dot{\theta}=\frac{\lambda}{r^2}$$

for some $\lambda, c\in\mathbb{R}$, related by $\lambda^2=Kc$. The value of r in terms of θ is given by

$$r=\frac{c}{1+\varepsilon\cos\theta+\delta\sin\theta}$$

for some $\varepsilon,\delta\in\mathbb{R}$. At the level of the affine coordinates x, y, this means

$$x=\frac{c\cos\theta}{1+\varepsilon\cos\theta+\delta\sin\theta}\quad,\quad y=\frac{c\sin\theta}{1+\varepsilon\cos\theta+\delta\sin\theta}$$

with $\theta=\theta(t)$ being subject to $\dot{\theta}=\lambda^2/r$, as above. Finally, we have

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

which is a degree 2 equation, and so the resulting trajectory is a conic.

Proof. As already mentioned, this is a sort of "best of" the formulae found in the proof of Theorem 16.7. And in the hope of course that we have not forgotten anything. Finally, let us mention that the simplest illustration for this is the circular motion, and for details on this, not included in the above, we refer to the proof of Theorem 16.7. □

As a next question, we would like to understand how the various parameters appearing above, namely $\lambda, c, \varepsilon, \delta$, which via some basic math can only tell us more about the shape of the orbit, appear from the initial data. The formulae here are as follows:

Theorem 16.9. *In the context of Theorem 16.8, and in polar coordinates, $x = r\cos\theta$, $y = r\sin\theta$, the initial data is as follows, with $R = r_0$:*

$$r_0 = \frac{c}{1+\varepsilon} \quad , \quad \theta_0 = 0$$

$$\dot{r}_0 = -\frac{\delta\sqrt{K}}{\sqrt{c}} \quad , \quad \dot{\theta}_0 = \frac{\sqrt{Kc}}{R^2}$$

$$\ddot{r}_0 = \frac{\varepsilon K}{R^2} \quad , \quad \ddot{\theta}_0 = \frac{4\delta K}{R^2}$$

The corresponding formulae for the affine coordinates x, y can be deduced from this. Also, the various motion parameters c, ε, δ and $\lambda = \sqrt{Kc}$ can be recovered from this data.

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

(1) As mentioned in Theorem 16.8, the object m begins its movement on Ox. Thus we have $\theta_0 = 0$, and from this we get the formula of r_0 in the statement.

(2) Regarding the initial speed now, the formula of $\dot{\theta}_0$ follows from:

$$\dot{\theta} = \frac{\lambda}{r^2} = \frac{\sqrt{Kc}}{r^2}$$

Also, in what concerns the radial speed, the formula of $\dot{r}_0$ follows from:

$$\begin{aligned}
\dot{r} &= \frac{c(\varepsilon\sin\theta - \delta\cos\theta)\dot{\theta}}{(1+\varepsilon\cos\theta+\delta\sin\theta)^2} \\
&= \frac{c(\varepsilon\sin\theta - \delta\cos\theta)}{c^2/r^2} \cdot \frac{\sqrt{Kc}}{r^2} \\
&= \frac{\sqrt{K}(\varepsilon\sin\theta - \delta\cos\theta)}{\sqrt{c}}
\end{aligned}$$

(3) Regarding now the initial acceleration, by using $\dot{\theta} = \sqrt{Kc}/r^2$ we find:

$$\ddot{\theta} = -2\sqrt{Kc} \cdot \frac{2r\dot{r}}{r^3} = -\frac{4\sqrt{Kc} \cdot \dot{r}}{r^2}$$

In particular at $t=0$ we obtain the formula in the statement, namely:

$$\ddot{\theta}_0=-\frac{4\sqrt{Kc}\cdot\dot{r}_0}{R^2}=\frac{4\sqrt{Kc}}{R^2}\cdot\frac{\delta\sqrt{K}}{\sqrt{c}}=\frac{4\delta K}{R^2}$$

(4) Also regarding acceleration, with $\lambda=\sqrt{Kc}$ our main motion formula reads:

$$\ddot{r}=\frac{Kc}{r^2}\left(\frac{1}{r}-\frac{1}{c}\right)$$

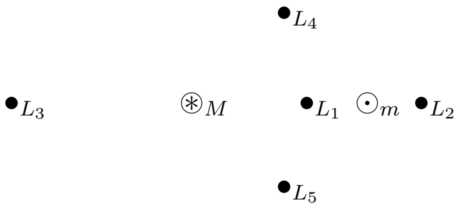
In particular at $t=0$ we obtain the formula in the statement, namely:

$$\ddot{r}_0=\frac{Kc}{R^2}\left(\frac{1}{R}-\frac{1}{c}\right)=\frac{Kc}{R^2}\cdot\frac{\varepsilon}{c}=\frac{\varepsilon K}{R^2}$$

(5) Finally, the last assertion is clear, and since the formulae look better anyway in polar coordinates than in affine coordinates, we will not get into details here. □

With the above formulae in hand, which are a precious complement to Theorem 16.8, we can do some reverse engineering at the level of parameters, and work out how various inital speeds and accelerations lead to various types of conics. There are many things that can be said here, and we refer here to any standard mechanics book.

Finally, a word about the 3-body problem. An interesting question here is how to position a specialized scientific satellite, deep in space, and away from the dust and radiation of the usual orbits around the Earth, as to stay there, under the joint influence of the gravity of the Sun M and of the Earth m. And there are 5 possible solutions here, called Lagrange points L1-L5, whose positions with respect to M,m are as follows:



Moreover, and here comes another interesting point, L4, L5 are stable, in the sense that a satellite installed there will really stay there, regardless of the various tiny little things that might happen, like an asteroid passing by, while L1, L2, L3 are unstable, in the sense that a satellite installed there will need constant tiny adjustments, in order to really stay there. So, which one would you choose for installing your satellite?

You would probably say L4, L5, but this is precisely the wrong answer, because due to their stability, these points attract a lot of asteroids and space garbage, and our satellite will certainly not perform well there, in that crowd. So, with L4, L5 ruled out, and with L3 ruled out too, being too far, the correct choices are L1, L2. But here, due to instability, you still need to learn a lot more mechanics, for knowing how to do this, in practice.

16c. Wave equation

As more physics, which can lead us into many interesting things, including light, we can talk about waves. Let us start with a discussion in 1 dimension, as follows:

Theorem 16.10. *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. □

More generally, we can talk about waves in N dimensions, as follows:

Theorem 16.11. *The wave equation in $\mathbb{R}^N$ is as follows,*

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

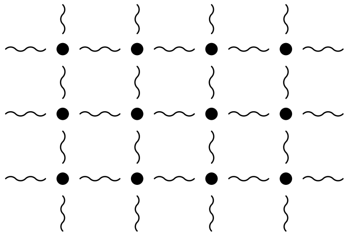
with $v>0$ being the propagation speed of the wave, and with Δ given by

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

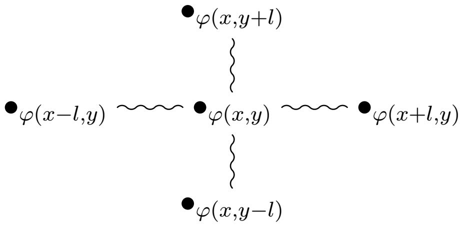
being the Laplace operator, playing the role of a numeric second derivative.

Proof. We can use here a lattice model as before, as follows:

(1) In 2 dimensions, to start with, the same argument as before carries on. Indeed, we can use a lattice model as follows, with all the edges standing for small springs:



As before in one dimension, we send an impulse, and we zoom on one ball. The situation here is as follows, with l being the spring length:



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

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

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

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

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

$$\begin{aligned}m\cdot\ddot{\varphi}(x,y) &= k(\varphi(x+l,y)-2\varphi(x,y)+\varphi(x-l,y))\\ &+ k(\varphi(x,y+l)-2\varphi(x,y)+\varphi(x,y-l))\end{aligned}$$

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

$$m = \frac{M}{B^2} \quad , \quad k = K \quad , \quad l = \frac{L}{B}$$

With these changes, our equation of motion found in (3) reads:

$$\begin{aligned}\ddot{\varphi}(x,y) &= \frac{KB^2}{M}(\varphi(x+l,y)-2\varphi(x,y)+\varphi(x-l,y))\\ &+ \frac{KB^2}{M}(\varphi(x,y+l)-2\varphi(x,y)+\varphi(x,y-l))\end{aligned}$$

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

$$\begin{aligned}\ddot{\varphi}(x,y) &= \frac{KL^2}{M}\times\frac{\varphi(x+l,y)-2\varphi(x,y)+\varphi(x-l,y)}{l^2}\\ &+ \frac{KL^2}{M}\times\frac{\varphi(x,y+l)-2\varphi(x,y)+\varphi(x,y-l)}{l^2}\end{aligned}$$

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

$$\ddot{\varphi}(x,y) = \frac{KL^2}{M}\left(\frac{d^2\varphi}{dx^2}+\frac{d^2\varphi}{dy^2}\right)(x,y)$$

As a conclusion to this, we are led to the following wave equation in two dimensions, with $v=\sqrt{K/M}\cdot L$ being the propagation speed of our wave:

$$\ddot{\varphi}(x,y)=v^2\left(\frac{d^2\varphi}{dx^2}+\frac{d^2\varphi}{dy^2}\right)(x,y)$$

But we recognize at right the Laplace operator, and we are done. As before in 1D, there is of course some discussion to be made here, arguing that our spring model in (1) is indeed the correct one. But do not worry, experiments confirm our findings.

(3) In 3 dimensions now, which is the case of the main interest, corresponding to our real-life world, the same argument carries over, and the wave equation is as follows:

$$\ddot{\varphi}(x,y,z)=v^2\left(\frac{d^2\varphi}{dx^2}+\frac{d^2\varphi}{dy^2}+\frac{d^2\varphi}{dz^2}\right)(x,y,z)$$

(4) Finally, the same argument, namely a lattice model, carries on in arbitrary N dimensions, and the wave equation here is as follows:

$$\ddot{\varphi}(x_1,\ldots,x_N)=v^2\sum_{i=1}^N\frac{d^2\varphi}{dx_i^2}(x_1,\ldots,x_N)$$

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

Regarding now the solution of the wave equation, in 1D we have:

Theorem 16.12. *The solution of the 1D wave equation with initial value conditions $\varphi(x,0)=f(x)$ and $\dot{\varphi}(x,0)=g(x)$ is given by the d'Alembert formula, namely:*

$$\varphi(x,t)=\frac{f(x-vt)+f(x+vt)}{2}+\frac{1}{2v}\int_{x-vt}^{x+vt}g(s)ds$$

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 must show that our solution is unique, but instead of going here into abstract arguments, we will 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 in the statement.

(3) In regards now with our discretization questions, by using a 1D lattice model with balls and springs as before, what happens to all the above is more or less that the above d'Alembert integral gets computed via Riemann sums, in our model, as stated. □

In $N\geq 2$ dimensions things get more complicated, among others requiring the use of spherical coordinates, and we refer here to any reasonably advanced mechanics book.

16d. Heat equation

Time now for heat, which is intimately related to the waves, via light, which is a wave. The general equation here is quite similar to the one for the waves, as follows:

Theorem 16.13. *Heat diffusion in $\mathbb{R}^N$ is described by the heat equation*

$$\dot{\varphi}=\alpha\Delta\varphi$$

where $\alpha>0$ is the thermal diffusivity of the medium, and Δ is the Laplace operator.

Proof. The study here is quite similar to the study of waves, as follows:

(1) To start with, as an intuitive explanation for the equation, since the second derivative φ'' in one dimension, or the quantity $\Delta\varphi$ in general, computes the average value of a function φ around a point, minus the value of φ at that point, 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.

(2) The heat equation as formulated above is of course something approximative, and several improvements can be made to it, first by incorporating a term accounting for heat radiation, and then doing several fine-tunings, depending on the material involved. But more on this later, for the moment let us focus on the heat equation above.

(3) In relation with our modeling questions, we can recover this equation a bit as we did before for the wave equation, by using a basic lattice model. Indeed, let us first assume, for simplifying, that we are in the one-dimensional case, $N = 1$. Here our model looks as follows, with distance $l > 0$ between neighbors:

$$\text{——}\ \circ_{x-l}\overset{l}{\text{——}}\circ_x\overset{l}{\text{——}}\circ_{x+l}\ \text{——}$$

In order to model heat diffusion, we have to implement the intuitive mechanism explained above, namely "the rate of change of the temperature of the material at any given point must be proportional, with proportionality factor $\alpha > 0$, to the average difference of temperature between that given point and the surrounding material".

(4) In practice, this leads to a condition as follows, expressing the change of the temperature φ, over a small period of time $\delta > 0$:

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

To be more precise, we have made several assumptions here, as follows:

– General heat diffusion assumption: the change of temperature at any given point x is proportional to the average over neighbors, $y \sim x$, of the differences $\varphi(y, t) - \varphi(x, t)$ between the temperatures at x, and at these neighbors y.

– Infinitesimal time and length conditions: in our model, the change of temperature at a given point x is proportional to small period of time involved, $\delta > 0$, and is inverse proportional to the square of the distance between neighbors, l^2.

(5) Regarding these latter assumptions, the one regarding the proportionality with the time elapsed $\delta > 0$ is something quite natural, physically speaking, and mathematically speaking too, because we can rewrite our equation as follows, making it clear that we have here an equation regarding the rate of change of temperature at x:

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

As for the second assumption that we made above, namely inverse proportionality with l^2, this can be justified on physical grounds too, but again, perhaps the best is to do the math, which will show right away where this proportionality comes from.

(6) So, let us do the math. In the context of our 1D model the neighbors of x are the points $x \pm l$, and so the equation that we wrote above takes the following form:

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

Now observe that we can write this equation as follows:

$$\frac{\varphi(x,t+\delta)-\varphi(x,t)}{\delta}=\alpha\cdot\frac{\varphi(x+l,t)-2\varphi(x,t)+\varphi(x-l,t)}{l^2}$$

(7) As it was the case with the wave equation before, we recognize on the right the usual approximation of the second derivative, coming from calculus. Thus, when taking the continuous limit of our model, $l \to 0$, we obtain the following equation:

$$\frac{\varphi(x,t+\delta)-\varphi(x,t)}{\delta}=\alpha\cdot\varphi''(x,t)$$

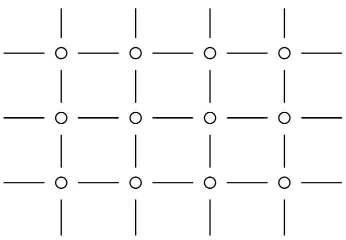
Now with $t \to 0$, we are led in this way to the heat equation, namely:

$$\dot{\varphi}(x,t)=\alpha\cdot\varphi''(x,t)$$

Summarizing, we are done with the 1D case, with our proof being quite similar to the one for the wave equation, from before.

(8) In practice now, there are of course still a few details to be discussed, in relation with all this, for instance at the end, in relation with the precise order of the limiting operations $l \to 0$ and $\delta \to 0$ to be performed, but these remain minor aspects, because our equation makes it clear, right from the beginning, that time and space are separated, and so that there is no serious issue with all this. And so, fully done with 1D.

(9) With this done, let us discuss now 2 dimensions. Here, as before for the waves, we can use a lattice model as follows, with all lengths being $l > 0$, for simplifying:



(10) We have to implement now the physical heat diffusion mechanism, namely "the rate of change of the temperature of the material at any given point must be proportional, with proportionality factor $\alpha > 0$, to the average difference of temperature between that given point and the surrounding material". In practice, this leads to a condition as follows,

expressing the change of the temperature φ, over a small period of time $\delta > 0$:

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

In fact, we can rewrite our equation as follows, making it clear that we have here an equation regarding the rate of change of temperature at x:

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

(11) So, let us do the math. In the context of our 2D model the neighbors of x are the points $(x \pm l, y \pm l)$, so the equation above takes the following form:

$$\begin{aligned}
&\frac{\varphi(x,y,t+\delta) - \varphi(x,y,t)}{\delta}\\
= \ &\frac{\alpha}{l^2}\Big[(\varphi(x+l,y,t) - \varphi(x,y,t)) + (\varphi(x-l,y,t) - \varphi(x,y,t))\Big]\\
+ \ &\frac{\alpha}{l^2}\Big[(\varphi(x,y+l,t) - \varphi(x,y,t)) + (\varphi(x,y-l,t) - \varphi(x,y,t))\Big]
\end{aligned}$$

Now observe that we can write this equation as follows:

$$\begin{aligned}
\frac{\varphi(x,y,t+\delta) - \varphi(x,y,t)}{\delta} &= \alpha \cdot \frac{\varphi(x+l,y,t) - 2\varphi(x,y,t) + \varphi(x-l,y,t)}{l^2}\\
&+ \alpha \cdot \frac{\varphi(x,y+l,t) - 2\varphi(x,y,t) + \varphi(x,y-l,t)}{l^2}
\end{aligned}$$

(12) As it was the case when modeling the wave equation before, we recognize on the right the usual approximation of the second derivative, coming from calculus. Thus, when taking the continuous limit of our model, $l \to 0$, we obtain the following equation:

$$\frac{\varphi(x,y,t+\delta) - \varphi(x,y,t)}{\delta} = \alpha\left(\frac{d^2\varphi}{dx^2} + \frac{d^2\varphi}{dy^2}\right)(x,y,t)$$

Now with $t \to 0$, we are led in this way to the heat equation, namely:

$$\dot{\varphi}(x,y,t) = \alpha \cdot \Delta\varphi(x,y,t)$$

Finally, in arbitrary N dimensions the same argument carries over, namely a straightforward lattice model, and gives the heat equation, as formulated in the statement. □

Regarding now the mathematics of the heat equation, many things can be said. As a first result here, often used by mathematicians, as to assume $\alpha = 1$, we have:

Proposition 16.14. *Up to a time rescaling, we can assume $\alpha = 1$, as to deal with*

$$\dot{\varphi} = \Delta\varphi$$

called normalized heat equation.

Proof. This is clear physically speaking, because according to our model, changing the parameter $\alpha > 0$ will result in accelerating or slowing the heat diffusion, in time $t > 0$. Mathematically, this follows via a change of variables, for the time variable t. □

Regarding now the resolution of the heat equation, we have here:

Theorem 16.15. *The heat equation, normalized as $\dot{\varphi} = \Delta\varphi$, and with initial condition $\varphi(x, 0) = f(x)$, has as solution the function*

$$\varphi(x, t) = (K_t * f)(x)$$

where the function $K_t : \mathbb{R}^N \to \mathbb{R}$, called heat kernel, is given by

$$K_t(x) = (4\pi t)^{-N/2} e^{-||x||^2/4t}$$

with $||x||$ being the usual norm of vectors $x \in \mathbb{R}^N$.

Proof. To start with, we can define the convolution of functions as follows:

$$(f * g)(x) = \int_{\mathbb{R}^N} f(x - y)g(y)dy$$

In relation now with our heat equation question, we have to check that the following function satisfies $\dot{\varphi} = \Delta\varphi$, with initial condition $\varphi(x, 0) = f(x)$:

$$\varphi(x, t) = (4\pi t)^{-N/2} \int_{\mathbb{R}^N} e^{-||x-y||^2/4t} f(y)dy$$

But both checks are elementary, coming from definitions. □

Many other things can be said, as a continuation of this, and quite often in relation with the Central Limit Theorem from probability, that we learned in chapter 15. In fact, Theorem 16.15 is something quite foundational, for modern mathematics.

16e. Exercises

Congratulations for having read this book, and no exercises for this final chapter. However, if interested in more, have a look at the various books referenced below.

Bibliography

[1] V.I. Arnold, Ordinary differential equations, Springer (1973).

[2] V.I. Arnold, Mathematical methods of classical mechanics, Springer (1974).

[3] V.I. Arnold, Catastrophe theory, Springer (1984).

[4] V.I. Arnold, Lectures on partial differential equations, Springer (1997).

[5] V.I. Arnold and B.A. Khesin, Topological methods in hydrodynamics, Springer (1998).

[6] M.F. Atiyah, K-theory, CRC Press (1964).

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

[8] M.F. Atiyah and I.G. MacDonald, Introduction to commutative algebra, Addison-Wesley (1969).

[9] T. Banica, Calculus and applications (2024).

[10] T. Banica, Linear algebra and group theory (2024).

[11] T. Banica, Introduction to modern physics (2025).

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

[13] S.J. Blundell and K.M. Blundell, Concepts in thermal physics, Oxford Univ. Press (2006).

[14] B. Bollobás, Modern graph theory, Springer (1998).

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

[16] A.R. Choudhuri, Astrophysics for physicists, Cambridge Univ. Press (2012).

[17] D.D. Clayton, Principles of stellar evolution and nucleosynthesis, Univ. of Chicago Press (1968).

[18] A. Connes, Noncommutative geometry, Academic Press (1994).

[19] J.B. Conway, A course in functional analysis, Springer (1985).

[20] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).

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

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

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

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

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

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

[27] A. Einstein, Relativity: the special and the general theory, Dover (1916).

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

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

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

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

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

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

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

[35] P. Flajolet and R. Sedgewick, Analytic combinatorics, Cambridge Univ. Press (2009).

[36] A.P. French, Special relativity, Taylor and Francis (1968).

[37] W. Fulton, Algebraic topology, Springer (1995).

[38] W. Fulton and J. Harris, Representation theory, Springer (1991).

[39] C. Godsil and G. Royle, Algebraic graph theory, Springer (2001).

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

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

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

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

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

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

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

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

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

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

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

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

[52] V.F.R. Jones, Subfactors and knots, AMS (1991).

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

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

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

[56] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).

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

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

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

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

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

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

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

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

[65] J.M. Lee, Introduction to topological manifolds, Springer (2011).

[66] J.M. Lee, Introduction to smooth manifolds, Springer (2012).

[67] J.M. Lee, Introduction to Riemannian manifolds, Springer (2018).

[68] M.L. Mehta, Random matrices, Elsevier (2004).

[69] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge Univ. Press (2000).

[70] R.K. Pathria and and P.D. Beale, Statistical mechanics, Elsevier (1972).

[71] P. Petersen, Linear algebra, Springer (2012).

[72] P. Petersen, Riemannian geometry, Springer (1998).

[73] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[74] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[75] W. Rudin, Functional analysis, McGraw-Hill (1973).

[76] W. Rudin, Fourier analysis on groups, Dover (1974).

[77] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[78] B. Ryden and B.M. Peterson, Foundations of astrophysics, Cambridge Univ. Press (2010).

[79] B. Ryden and R.W. Pogge, Interstellar and intergalactic medium, Cambridge Univ. Press (2021).

[80] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).

[81] J.P. Serre, A course in arithmetic, Springer (1973).

[82] J.P. Serre, Linear representations of finite groups, Springer (1977).

[83] I.R. Shafarevich, Basic algebraic geometry, Springer (1974).

[84] R. Shankar, Fundamentals of physics I: mechanics, relativity, and thermodynamics, Yale Univ. Press (2014).

[85] R. Shankar, Fundamentals of physics II: electromagnetism, optics, and quantum mechanics, Yale Univ. Press (2016).

[86] R. Shankar, Principles of quantum mechanics, Springer (1980).

[87] R. Shankar, Quantum field theory and condensed matter: an introduction, Cambridge Univ. Press (2017).

[88] A.M. Steane, Thermodynamics, Oxford Univ. Press (2016).

[89] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[90] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[91] J. von Neumann and O. Morgenstern, Theory of games and economic behavior, Princeton Univ. Press (1944).

[92] J. Watrous, The theory of quantum information, Cambridge Univ. Press (2018).

[93] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[94] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[95] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).

[96] S. Weinberg, Cosmology, Oxford Univ. Press (2008).

[97] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).

[98] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).

[99] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).

[100] B. Zwiebach, A first course in string theory, Cambridge Univ. Press (2004).

Index