Principles of mathematics

Teo Banica

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

2010 *Mathematics Subject Classification.* 01A67

Key words and phrases. Algebra, Geometry

Abstract. This is an introduction to mathematics, with emphasis on algebra and geometry. We first discuss numbers, 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

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

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

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

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

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

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

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

Part III goes into more advanced mathematics, namely functions and analysis, again with motivations coming from craftsmanship, and geometry.

Part IV is an introduction to vector calculus, space geometry and basic mechanics, which is certainly quite advanced material, that you need to know too.

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

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

Cergy, April 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

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

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:

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

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,*

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

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:

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

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,*

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

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:

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

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:

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

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

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

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:

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

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:

$$\begin{array}{ccc} + & 1 & 2 \\ 1 & 2 & 10 \\ 2 & 10 & 11 \end{array}$$

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

$$\begin{array}{ccc} \times & 1 & 2 \\ 1 & 1 & 2 \\ 2 & 2 & 11 \end{array}$$

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*

$$\begin{array}{cccc} + & 1 & 2 & 3 \\ 1 & 2 & 3 & 10 \\ 2 & 3 & 10 & 11 \\ 3 & 10 & 11 & 12 \end{array}$$

the multiplication table is

$$\begin{array}{cccc} \times & 1 & 2 & 3 \\ 1 & 1 & 2 & 3 \\ 2 & 2 & 10 & 12 \\ 3 & 3 & 12 & 21 \end{array}$$

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*

$$n = a_1 a_2 \ldots a_k$$

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

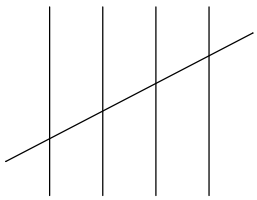
$+$	1	2	3	4
1	2	3	4	10
2	3	4	10	11
3	4	10	11	12
4	10	11	12	13

the multiplication table is

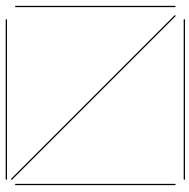
$\times$	1	2	3	4
1	1	2	3	4
2	2	4	11	13
3	3	11	14	22
4	4	13	22	31

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

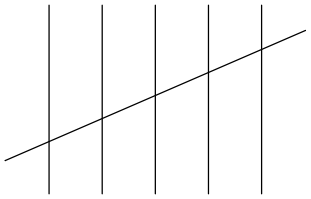
the multiplication table is

$\times$	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

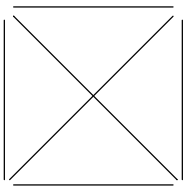
and in practice, this beats both basis two, and basis three.

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

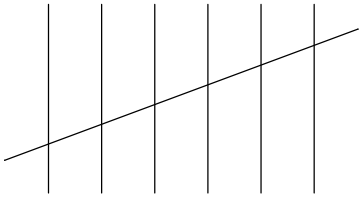
$\times$	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	11	13	15
3	3	6	12	15	21	24
4	4	11	15	22	26	33
5	5	13	21	26	34	42
6	6	15	24	33	42	51

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

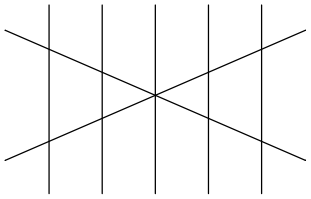
To be more precise here, in what regards the last assertion, remember that damn 6×7 and 7×8 computations from school, that we all had big troubles with. Well, in basis seven these two computations take a very simple form, as follows:

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

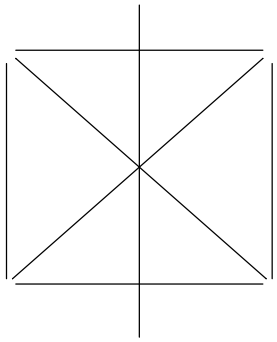
In what regards the graphics, however, not very good news here, because with the heptagon being hard to draw, we are basically left with ugly pictures, as follows:



Alternatively, we have pictures as follows, which look ugly as well:



Yet another unpleasant convention, which can be used as well, is as follows:



And so on. In fact, feel free to have some thinking at this, how to best count, pictorially speaking, in numeration basis seven, with this being a quite interesting question, normally taking you into plane geometry, and many other interesting things.

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, as everyone uses nowadays:

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

$$n = a_1 a_2 \ldots a_k$$

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*

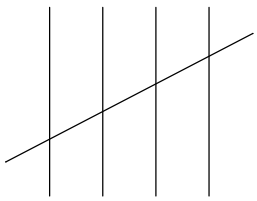
$+$	1	2	3	4	5	6	7	8	9
1	2	3	4	5	6	7	8	9	10
2	3	4	5	6	7	8	9	10	11
3	4	5	6	7	8	9	10	11	12
4	5	6	7	8	9	10	11	12	13
5	6	7	8	9	10	11	12	13	14
6	7	8	9	10	11	12	13	14	15
7	8	9	10	11	12	13	14	15	16
8	9	10	11	12	13	14	15	16	17
9	10	11	12	13	14	15	16	17	18

the multiplication table is

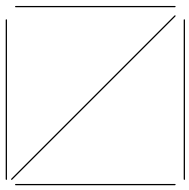
$\times$	1	2	3	4	5	6	7	8	9
1	1	2	3	4	5	6	7	8	9
2	2	4	6	8	10	12	14	16	18
3	3	6	9	12	15	18	21	24	27
4	4	8	12	16	20	24	28	32	36
5	5	10	15	20	25	30	35	40	45
6	6	12	18	24	30	36	42	48	54
7	7	14	21	28	35	42	49	56	63
8	8	16	24	32	40	48	56	64	72
9	9	18	27	36	45	54	63	72	81

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

Which sounds very nice, save for the memorization of the above tables, and good luck with this, and with the remark that for graphics, we can still use, up to some extent, our basis 5 signs, with the well-known convention for this being as follows:



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



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 $\times$:

$$a_1 = a_1$$

$$a_1a_2 = 10 \times a_1 + a_2$$

$$a_1a_2a_3 = 100 \times a_1 + 10 \times a_2 + a_3$$

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

$$\ldots$$

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.

And with this, good news, done with the hard mathematics, with the rest of the present chapter being more or less trivialities, coming from the above.

1c. Basic arithmetic

We say that b divides a, and write $b|a$, when there is a number c such that $a = bc$. In this case we also use the following notation, for designating this quotient number c:

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

These beasts, called "fractions", are subject to a number of simple formulae, which are all useful, in the real life. For addition, the formula is as follows:

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

For substraction, the formula here is quite similar, is as follows:

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

For multiplication, here the formula is something very simple, as follows:

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

As for division, here the formula is again something simple, as follows:

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

And more on this, divisibility of numbers, and on fractions too, in the above sense, and in some generalized sense too, when $a \not| b$, later in this book.

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

THEOREM 1.18. *We can talk about:*

(1) *Greatest common divisors* (a, b).
(2) *Least common multiples* $[a, b]$.

PROOF. This is indeed quite clear from definitions. Many things can be said here, the general idea being that given two numbers a, b, we can write them as follows, with d being a certain number, and with a', b' being numbers having no common divisors:

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

But with this writing in hand, the formulae that we are looking for are:

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

Observe now that we have the following formula:

$$\begin{aligned} ab &= d^2a'b' \\ &= d \times da'b' \\ &= (a, b)[a, b] \end{aligned}$$

Many other things can be said, as a continuation of this. □

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

THEOREM 1.19. *We can talk about:*

(1) *Greatest common divisors* (a, b, c).
(2) *Least common multiples* $[a, b, c]$.

PROOF. This is again quite clear from definitions. Many things can be said here, the general idea being that given three numbers a, b, c, we can write them as follows, with d being a certain number, and with a', b', c' being numbers having no common divisor:

$$a = da'$$

$$b = db'$$

$$c = dc'$$

But with this writing in hand, the formulae that we are looking for are:

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

In order to compute now $[a', b', c']$, we can apply to the pairs (a', b'), (b', c'), (a', c'), we can apply to them the theory that we learned in Theorem 1.18, and its proof. We are led in this way to decomposition results as follows, for the numbers a', b', c':

$$a' = pqx$$

$$b' = pry$$

$$c' = qrz$$

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

$$a = dpqx$$

$$b = dpry$$

$$c = dqrz$$

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

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

$$[a, b, c] = dpqrxyz$$

Which is something more complicated that what was happening before, for just two numbers, for instance because when multiplying, we have the following formula:

$$\begin{aligned} abc &= dpqx \cdot dpry \cdot dqrz \\ &= d^3(pqr)^2xyz \\ &= d \cdot dpqrxyz \cdot dpqr \\ &= (a, b, c) \cdot [a, b, c] \cdot dpqr \end{aligned}$$

Many other things can be said, as a continuation of this. □

More generally now, we have the following result:

THEOREM 1.20. *We can talk about:*

(1) *Greatest common divisors* $(a_1, \ldots, a_k)$.
(2) *Least common multiples* $[a_1, \ldots, a_k]$.

PROOF. This generalizes the above, the idea being as follows:

(1) Again, the fact that we can talk indeed about greatest common divisors, and about least common multiples, is quite clear from definitions.

(2) However, when it comes to suitably decomposing our numbers $a_1, \ldots, a_k$, by using their various common divisors, as we did in Theorem 1.18 at $k = 2$, and in Theorem 1.19 at $k = 3$, things become considerably more complicated at $k = 4$, and higher.

(3) And here, we can only recommend some thinking and computations at $k = 4$, which are something very instructive. □

Summarizing, we have up and working a useful theory of greatest common divisors, and least common multiples, but obviously this is just the tip of the iceberg, and many interesting questions remain open. We will be back to them, later in this book.

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

DEFINITION 1.21. *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:

$$6|a \iff 2|a \text{ and } 3|a$$
$$10|a \iff 2|a \text{ and } 5|a$$
$$12|a \iff 3|a \text{ and } 4|a$$
$$14|a \iff 2|a \text{ and } 7|a$$
$$15|a \iff 3|a \text{ and } 5|a$$
$$18|a \iff 2|a \text{ and } 9|a$$
$$20|a \iff 4|a \text{ and } 5|a$$
$$21|a \iff 3|a \text{ and } 7|a$$
$$22|a \iff 2|a \text{ and } 11|a$$
$$24|a \iff 3|a \text{ and } 8|a$$

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 divisior, we just have to solve this question for certain special values of n, excluding $n = 6, 10, 12, 14, 15, 18, 20, 21, 22, 24, \ldots$

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

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

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

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

THEOREM 1.22. *Given a positive integer $a = a_1 \ldots a_r$, we have:*

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

PROOF. This is something well-known, and easy to deduce from definitions, the idea being that the $q = 2^k, 5$ assertions follow from $10 = 2 \times 5$, the $q = 3, 9$ assertions follow from $10 = 9 + 1$, and the $q = 11$ assertion follows from $10 = 11 - 1$. □

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

THEOREM 1.23. *Assuming that we have convinced mankind to change the numeration basis from* 10 *to* 14*, given a positive integer* $a = a_1 \ldots a_r$*, we have:*

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

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

As a conclusion to this, good news, 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, and launch an online petition if enthusiastic about it.

Be said in passing, our Theorem 1.23 above is a bit ill-formulated, mixing things written in basis 10 and basis 14, and we will leave fixing all this, with a fully correct mathematical statement, as another instructive exercise for you.

1d. Prime numbers

Time now to get into prime numbers, which will be a main theme of discussion, in this book. And as a first question, from me to you, how many primes do you know?

With this being something very serious. The more primes you know, the better for your computations and science, believe me, and we will soon understand why.

In any case, to start with, knowing the prime numbers under 100 is something mandatory, at the beginner level, and here they are, in all their beauty:

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47$$

$$53, 59, 61, 67, 71, 73, 79, 83, 89, 97$$

Actually those between 101 and 200 are mandatory too, here they are:

$$101, 103, 107, 109, 113, 127, 131, 137, 139, 149$$

$$151, 157, 163, 167, 173, 179, 181, 191, 193, 197, 199$$

But then, can we ignore those between 201 and 300. These are as follows:

$$211, 223, 227, 229, 233, 239, 241$$

$$251, 257, 263, 269, 271, 277, 281, 283, 293$$

Not to forget the primes between 301 and 400, which are as follows:

$$307, 311, 313, 317, 331, 337, 347, 349$$

$$353, 359, 367, 373, 379, 383, 389, 397$$

And we have kept the best for the end, primes between 401 and 500, which are:

$$401, 409, 419, 421, 431, 433, 439, 443, 449$$

$$457, 461, 463, 467, 479, 487, 491, 499$$

So, these are our beasts, in arithmetic, and try to get familiar with them, and learn some of their powers too, because these prime powers are very useful too.

We have already met prime numbers in the above, when talking divisibility, and even used some of their basic properties, that you were certainly very familiar with, but time now to review all this, on a more systematic basis, with proofs and everything.

First, as definition for the prime numbers, we have:

DEFINITION 1.24. *The prime numbers are the integers $p > 1$ satisfying*

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

with each of these properties uniquely determining them.

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

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

THEOREM 1.25. *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 that you certainly know, related to the equivalent conditions (1,2,3) in Definition 1.24, and exercise for you, to remember how all this works. Exercise as well, work out this for all integers $n \leq 100$, with no calculators allowed. □

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

THEOREM 1.26. *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 1.27. *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:

$$\begin{array}{cccccccccccccccccccc}
\underline{2} & 3 & \not{4} & 5 & \not{6} & 7 & \not{8} & 9 & \not{10} & 11 & \not{12} & 13 & \not{14} & 15 & \not{16} & 17 & \not{18} & 19 & \not{20} \\
 & \underline{3} & & 5 & & 7 & & \not{9} & & 11 & & 13 & & \not{15} & & 17 & & 19 & \\
 & & & \underline{5} & & 7 & & & & 11 & & 13 & & & & 17 & & 19 & \\
 & & & & & \underline{7} & & & & 11 & & 13 & & & & 17 & & 19 & \\
 & & & & & & & & & \underline{11} & & 13 & & & & 17 & & 19 & \\
 & & & & & & & & & & & \underline{13} & & & & 17 & & 19 & \\
 & & & & & & & & & & & \vdots & & & & & & &
\end{array}$$

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 1.28. *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

Exercises:

Exercise 1.29.

Exercise 1.30.

Exercise 1.31.

Exercise 1.32.

Exercise 1.33.

Exercise 1.34.

Exercise 1.35.

Exercise 1.36.

Bonus exercise.

CHAPTER 2

Fractions

2a. Fractions

Time now for some more complicated mathematics, going beyond what we know about the positive integers. We will be talking here about the mathematics of fractions.

We recall from chapter 1 that given an integer dividing another integer, $b|a$, we can talk about the corresponding quotient c, given by $a = bc$, which is denoted as follows:

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

The above beasts, called "fractions", are subject to a number of simple formulae, 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 $b|a$ does not hold, in the obvious way. And, with this convention, the above formulae still hold. Good to know.

Let us formulate now the following definition, based on the above:

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".

Getting now to algebra, the integers add and multiply of course according to the rules that you know well. As for the rational numbers, these add according to the usual rule for quotients, which is as follows, and never ever forget it:

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

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

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

We can invert the nonzero rational numbers, according to the following formula:

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

Finally, we can divide the rationals, the formula being as follows:

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

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

As a basic result about the rational numbers, in relation with what we like to do the most, since the beginning of this book, namely counting, we have:

THEOREM 2.2. *The set of rational numbers $\mathbb{Q}$ is countable.*

PROOF. This can be proved by using a standard diagonal trick. Consider indeed the following table, containing all quotients of type a/b, with $a,b\in\mathbb{N}$:

$$\begin{array}{cccccccc}
\frac{1}{1} & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \cdots\\
\frac{2}{1} & \frac{2}{2} & \frac{2}{3} & \frac{2}{4} & \frac{2}{5} & \frac{2}{6} & \cdots\\
\frac{3}{1} & \frac{3}{2} & \frac{3}{3} & \frac{3}{4} & \frac{3}{5} & \frac{3}{6} & \cdots\\
\frac{4}{1} & \frac{4}{2} & \frac{4}{3} & \frac{4}{4} & \frac{4}{5} & \frac{4}{6} & \cdots\\
\frac{5}{1} & \frac{5}{2} & \frac{5}{3} & \frac{5}{4} & \frac{5}{5} & \frac{5}{6} & \cdots\\
\frac{6}{1} & \frac{6}{2} & \frac{6}{3} & \frac{6}{4} & \frac{6}{5} & \frac{6}{6} & \cdots\\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

We can then snake our way inside this table, in the obvious way, starting from top left, and we count in this way $\mathbb{Q}_+$, with some redundancies. Thus, theorem proved. □

2b. Binomials, factorials

Time now to get into some interesting mathematics, by using our knowledge of numbers and fractions. We will be interested in counting sets. We first have here:

THEOREM 2.3. *A finite set E has*

$$|P(E)| = 2^{|E|}$$

possible subsets.

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

(1) In the case $|E| = 0$, meaning $E = \emptyset$, we have $2^0 = 1$ subsets, namely:

$$\emptyset$$

(2) In the case $|E| = 1$, meaning $E = \{e\}$, we have $2^1 = 2$ subsets, as follows:

$$\emptyset \quad , \quad \{e\}$$

(3) In the case $|E| = 2$, meaning $E = \{a, b\}$, we have $2^2 = 4$ subsets, as follows:

$$\emptyset \quad , \quad \{a\} \quad , \quad \{b\} \quad , \quad \{a, b\}$$

(4) Next, at $|E| = 3$, where $E = \{a, b, c\}$, we have $2^3 = 8$ subsets, as follows:

$$\emptyset \ , \ \{a\} \ , \ \{b\} \ , \ \{c\} \ , \ \{a, b\} \ , \ \{b, c\} \ , \ \{a, c\} \ , \ \{a, b, c\}$$

(5) In the general case now, the simplest is to say that the choice of a subset $G \subset E$ requires a binary choice for each of the elements $e \in E$, either in, or out. Now since these binary choices will multiply, we are led to the formula in the statement. □

In relation with the above, at a speculatory level, we are led to the following question, asking whether there is something between $\mathbb{N}$, and the functions $\mathbb{N} \to \{0, 1\}$:

$$2^\infty >^? \infty$$

And more on this, which is actually something quite scary, in relation with lots of mathematical logic, and with real numbers and other numbers too, later in this book.

Getting back now to the real life, and to concrete mathematics, as our first true theorem, solving a problem which often appears in real life, we have:

THEOREM 2.4. *The number of possibilities of choosing k objects among n objects is*

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

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

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

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

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

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

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

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

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

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

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

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

ADVICE 2.5. *Always doublecheck what you're doing, regularly, and definitely at the end, either with an alternative proof, or with some numerics.*

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

Back to work now, as an important adding to Theorem 2.4, we have:

CONVENTION 2.6. *By definition,* $0! = 1$.

This convention comes, and no surprise here, from Advice 2.5. Indeed, we obviously have $\binom{n}{n} = 1$, but if we want to recover this formula via Theorem 2.4 we are a bit in trouble, and so we must declare that $0! = 1$, as for the following computation to work:

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

Going ahead now with more mathematics and less philosophy, with Theorem 2.4 complemented by Convention 2.6 being in final form (trust me), we have:

THEOREM 2.7. *We have the binomial formula*

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

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

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

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

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

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

In order to finish, it remains to compute the coefficients C_k. But, according to our product formula, C_k is the number of choices for the k needed a variables among the n available a variables. Thus, according to Theorem 2.4, we have:

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

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

Theorem 2.7 is something quite interesting, so let us doublecheck it with some numerics. At small values of n we obtain the following formulae, which are all correct:

$$(a+b)^0 = 1$$
$$(a+b)^1 = a+b$$
$$(a+b)^2 = a^2+2ab+b^2$$
$$(a+b)^3 = a^3+3a^2b+3ab^2+b^3$$
$$(a+b)^4 = a^4+4a^3b+6a^2b^2+4ab^3+b^4$$
$$(a+b)^5 = a^5+5a^4b+10a^3b^2+10a^2b^3+5a^4b+b^5$$
$$\vdots$$

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

THEOREM 2.8. *The Pascal triangle, formed by the binomial coefficients* $\binom{n}{k}$,

$$1$$
$$1\ ,\ 1$$
$$1\ ,\ 2\ ,\ 1$$
$$1\ ,\ 3\ ,\ 3\ ,\ 1$$
$$1\ ,\ 4\ ,\ 6\ ,\ 4\ ,\ 1$$
$$1\ ,\ 5\ ,\ 10\ ,\ 10\ ,\ 5\ ,\ 1$$
$$\vdots$$

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

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

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

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

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

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

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

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

By using the binomial formula, this formula becomes:

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

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

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

There are many more things that can be said about binomial coefficients, with all sorts of interesting formulae, and we will be back to this, later in this book, on a regular basis, and with the idea being always the same, namely that in order to find such formulae you have a choice between algebra and combinatorics, a bit as in the above, and that when it comes to formal proofs, the brute-force computation method is something useful too.

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

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

DEFINITION 2.9. *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 2.10. *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.

More modestly now, but along the same lines, let us attempt to work out some basic arithmetic properties of the general binomial coefficients. We have here the following result, which is something very useful, in practice, for various purposes:

THEOREM 2.11. *Given a prime* $p \geq 2$*, the exponent of* p *inside* $n!$ *is*

$$a_n = \left[\frac{n}{p}\right] + \left[\frac{n}{p^2}\right] + \left[\frac{n}{p^3}\right] + \ldots$$

and so the exponent of p *inside* $\binom{n}{k}$ *is given by the formula*

$$b_{n,k} = a_n - a_k - a_{n-k}$$

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

PROOF. This is something quite self-explanatory, with the first assertion being elementary, coming from definitions, and with the second assertion following from it. □

There are many interesting illustrations for the above result. We will be back to such things later in this book, when doing more advanced arithmetics.

2c. Further counts

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.

So, as a first philosophical question, what are the simplest graphs X, that we can try to do some loop computations for? And here, we have 3 possible answers, as follows:

FACT 2.12. *The following are graphs X, with a distinguished vertex $0 \in X$:*

(1) *The circle graph, having N vertices, with 0 being one of the vertices.*
(2) *The segment graph, having N vertices, with 0 being the vertex at left.*
(3) *The segment graph, having $2N + 1$ vertices, with 0 being in the middle.*

So, let us start with these. For the circle, the computations are quite non-trivial, and you can try doing some, in order to understand what I am talking about. The problem comes from the fact that loops of length $k = 0, 2, 4, 6, \ldots$ are quite easy to count, but then, once we pass $k = N$, the loops can turn around the circle or not, and they can even turn several times, and so on, and all this makes the count too complicated. In addition, again due to loop turning, when N is odd, we have as well loops of odd length.

As for the two segment graphs, here the computations look again complicated, and even more complicated than for the circle, because, again, once we pass $k = N$ many things can happen, and this makes the count too complicated. And here, again you can try doing some computations, in order to understand what I am talking about.

So, shall we give up, in waiting for more advanced techniques, say coming from diagonalization? This would be a wise decision, but before that, let us pull an analysis trick, and formulate the following result, which is of course something informal, and modest:

THEOREM 2.13. *For the circle graph, having N vertices, the number of length k loops based at one of the vertices is approximately*

$$L_k \simeq \frac{2^k}{N}$$

in the $k \to \infty$ limit, when N is odd, and is approximately

$$L_k \simeq \begin{cases} \frac{2^{k+1}}{N} & (k \text{ even}) \\ 0 & (k \text{ odd}) \end{cases}$$

also with $k \to \infty$, when N is even. However, in what regards the two segment graphs, we can expect here things to be more complicated.

PROOF. This is something not exactly trivial, and with the way the statement is written, which is clearly informal, witnessing for that. The idea is as follows:

(1) Consider the circle graph X, with vertices denoted $0, 1, \ldots, N-1$. Since each vertex has valence 2, any length k path based at 0 will consist of a binary choice at the beginning, then another binary choice afterwards, and so on up to a k-th binary choice at the end. Thus, there is a total of 2^k such paths, based at 0, and having length k.

(2) But now, based on the obvious "uniformity" of the circle, we can argue that, in the $k \to \infty$ limit, the endpoint of such a path will become random among the vertices $0, 1, \ldots, N-1$. Thus, if we want this endpoint to be 0, as to have a loop, we have $1/N$ chances for this to happen, so the total number of loops is $L_k \simeq 2^k/N$, as stated.

(3) With the remark, however, that the above argument works fine only when N is odd. Indeed, when N is even, the endpoint of a length k path will be random among $0, 2, \ldots, 2N-2$ when k is even, and random among $1, 3, \ldots, 2N-1$ when k is odd. Thus for getting a loop we must assume that k is even, and in this case the number of such loops is the total number of length k paths, namely 2^k, approximately divided by $N/2$, the number of points in $\{0, 2, \ldots, 2N-2\}$, which gives $L_k = 2^k/(N/2)$, as stated.

(4) Moving ahead now to the segment graphs, it is pretty much clear that for both, we lack the "uniformity" needed in (2), and this due to the 2 endpoints of the segment. In fact, thinking well, these graphs are no longer 2-valent, again due to the 2 endpoints, each having valence 1, and so even (1) must be fixed. And so, we will stop here. □

All this is obviously not very good news, and so again, as question, shall we give up, in waiting for more advanced techniques, say coming from diagonalization? Well, instead of giving up, let us look face-to-face at the difficulties that we met. We are led this way, after analyzing the situation, to the following thought:

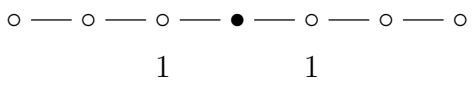
THOUGHT 2.14. *The difficulties that we met, with the circle and the two segments, come from the fact that our loops are not "free to move",*

(1) *for the circle, because these can circle around the circle,*
(2) *for the segments, obviously because of the endpoints,*

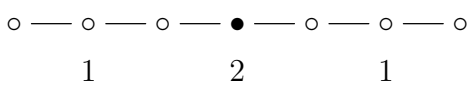
and so our difficulties will dissapear, and we will be able to do our exact loop count, once we find a graph X where the loops are truly "free to move".

Thinking some more, all this definitely buries the first interval graph, where the vertex 0 is one of the endpoints. However, we can still try to recycle the circle, by unwrapping it, or extend our second interval graph up to ∞. But in both cases what we get is the graph $\mathbb{Z}$ formed by the integers. So, let us try to count the length k paths on $\mathbb{Z}$, based at 0. At $k = 1$ we have 2 such paths, ending at -1 and 1, and the count results can be

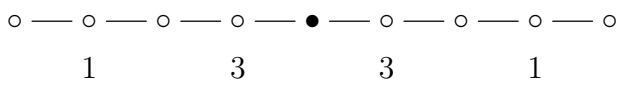
pictured as follows, with everything being self-explanatory:



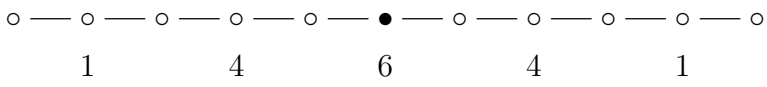
At $k = 2$ now, we have 4 paths, one of which ends at -2, two of which end at 0, and one of which ends at 2. The results can be pictured as follows:



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



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



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, eventually, we found the simplest graph ever, finite or not, namely $\mathbb{Z}$, and we have the following beautiful result about it:

THEOREM 2.15. *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 third example, let us try 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 2.16. *The Catalan numbers C_k, counting the loops on $\mathbb{N}$ based at 0,*

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

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:

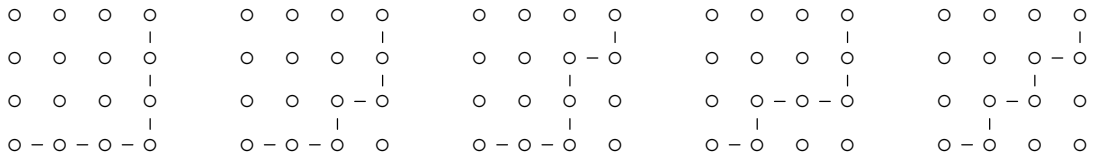
THEOREM 2.17. *The Catalan numbers C_k count:*

(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 2.16 with the loops on $\mathbb{N}$, you always end up with the same sequence of numbers, namely those found in Proposition 2.16:

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

(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 2.18. *The Catalan numbers are given by the formula*

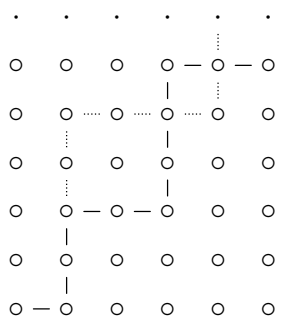
$$C_k = \frac{1}{k+1}\binom{2k}{k}$$

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. □

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. We first have:

THEOREM 2.19. *The probabilities at poker are as follows:*

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

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

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

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

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

Thus, the probability of having a pair is:

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

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

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

Thus, the probability of having two pairs is:

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

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

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

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

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

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

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

Thus, the probability of having full house is:

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

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

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

Thus, the probability of having a straight is:

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

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

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

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

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

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

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

Thus, the probability of having a flush is:

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

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

$$N = 4 \cdot 4$$

Thus, the probability of having a straight flush is:

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

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

Along the same lines, here is now a theorem about flipping coins:

THEOREM 2.20. *When flipping a coin k times what you can win are quantities of type $\$k-2s$, with $s=0,1,\ldots,k$, with the probability for this to happen being:*

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

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

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

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

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

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

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

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

But shall we really trust this. Imagine for instance that you play your game for \$1000 instead of \$1 as basic gain, your life is obviously at stake, so all this is worth a second doublecheck, before being used in practice. So, as second doublecheck, let us verify that, on average, what you win is exactly \$0, which is something very intuitive, the game itself

obviously not favoring you, nor your partner. But this can be checked as follows:

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

Summarizing, done with all our checks, and so theorem proved. □

Now with Theorem 2.20 in hand, it is quite clear that the basic $1/2-1/2$ probabilities there can be repaced with something of type $p-(1-p)$, with $p\in[0,1]$ being arbitrary. We are led in this way to the following notions, which are quite general:

DEFINITION 2.21. *Given $p\in[0,1]$, the Bernoulli law of parameter p is given by:*

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

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

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

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

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

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

PROOF. Obviously, this is something a bit informal, but let us prove this as stated, and we will come back later to it, with precise definitions, theorems and everything. In what regards the first assertion, nothing to be said there, this is what life teaches us. As

for the second assertion, the formula $\mu_{bin} = \mu_{ber}^{*k}$ there certainly looks like mathematics, so job for us to figure out what this exactly means. And, this can be done as follows:

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

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

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

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

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

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

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

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

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

Getting to formal mathematical work now, let us start with:

DEFINITION 2.23. *A random variable on a probability space X is a function*

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

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.

Let us complement now this definition with something finer, regarding the "quality" of the expectation $E(f) \in \mathbb{R}$ appearing there. And the first thought here, which is the correct one, goes to the following number, called variance of our variable:

$$\begin{aligned} V(f) &= E\left((f - E(f))^2\right) \\ &= E(f^2) - E(f)^2 \end{aligned}$$

However, let us not stop here. For a total control of your business, be that of financial, mathematical, physical or chemical type, you will certainly want to know more about your variable $f : X \to \mathbb{R}$. Which leads us into general moments, constructed as follows:

DEFINITION 2.24. *The moments of a variable $f : X \to \mathbb{R}$ are the numbers*

$$M_k = E(f^k)$$

which satisfy $M_0 = 1$, then $M_1 = E(f)$, and then $V(f) = M_2 - M_1^2$.

And, good news, with this we have all the needed tools in our bag for doing some good business. To put things in a very compacted way, M_0 is about foundations, M_1 is about running some business, M_2 is about running that business well, and M_3 and higher are advanced level, about ruining all the competing businesses.

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

THEOREM 2.25. *Given a random variable $f : X \to \mathbb{R}$, if we define its law as being*

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

regarded as probability measure on $\mathbb{R}$, then the moments are given by the formula

$$E(f^k) = \int_{\mathbb{R}} y^k d\mu(y)$$

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 $f : X \to \mathbb{R}$, we can certainly talk about its law μ, as being the formal linear combination of Dirac masses in the statement.

(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:

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

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

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

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

Next, we have the following straightforward definition, inspired by games:

DEFINITION 2.26. *We say that two variables $f, g : X \to \mathbb{R}$ are independent when*

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

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

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 2.27. *Assuming that $f, g : X \to \mathbb{R}$ are independent, we have:*

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

More generally, we have the following formula, for the mixed moments,

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

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:

$$\begin{aligned} E(f^k g^l) &= \sum_{xy} x^k y^l P(f = x, g = y) \\ &= \sum_{xy} x^k y^l P(f = x)P(g = y) \\ &= \sum_{x} x^k P(f = x) \sum_{y} y^l P(g = y) \\ &= E(f^k)E(g^l) \end{aligned}$$

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 2.22, let us start with the following abstract definition:

DEFINITION 2.28. *Given a space X with a sum operation $+$, we can define the convolution of any two discrete probability measures on it,*

$$\mu = \sum_i a_i \delta_{x_i} \quad , \quad \nu = \sum_j b_j \delta_{y_j}$$

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

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

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

As a first observation, our operation is well-defined, with $\mu * \nu$ being indeed a discrete probability measure, because the weights are positive, $a_i b_j \geq 0$, and their sum is:

$$\sum_{ij} a_i b_j = \sum_i a_i \sum_j b_j = 1 \times 1 = 1$$

Also, the above definition agrees with what we did before with coins, and Bernoulli and binomial laws. We have in fact the following general result:

THEOREM 2.29. *Assuming that $f, g : X \to \mathbb{R}$ 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:

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

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

Before going further, let us attempt as well to find a proof of Theorem 2.29, based on the moment characterization of independence, from Theorem 2.27. For this purpose, we will need the following standard fact, which is of certain theoretical interest:

THEOREM 2.30. *The sequence of moments*

$$M_k = \int_{\mathbb{R}} x^k d\mu(x)$$

uniquely determines the law.

PROOF. Indeed, assume that the law of our variable is as follows:

$$\mu = \sum_i \lambda_i \delta_{x_i}$$

The sequence of moments is then given by the following formula:

$$M_k = \sum_i \lambda_i x_i^k$$

But it is then standard calculus to recover the numbers $\lambda_i, x_i \in \mathbb{R}$, and so the measure μ, out of the sequence of numbers M_k. Indeed, assuming that the numbers x_i are $0 < x_1 < \ldots < x_n$ for simplifying, in the $k \to \infty$ limit we have the following formula:

$$M_k \sim \lambda_n x_n^k$$

Thus, we got the parameters $\lambda_n, x_n \in \mathbb{R}$ of our measure μ, and then by substracting them and doing an obvious recurrence, we get the other parameters $\lambda_i, x_i \in \mathbb{R}$ as well. □

Getting back now to our philosophical question above, namely recovering Theorem 2.25 via moment technology, we can now do this, the result being as follows:

THEOREM 2.31. *Assuming that $f, g : X \to \mathbb{R}$ are independent, the measures*

$$\mu_{f+g} \quad , \quad \mu_f * \mu_g$$

have the same moments, and so, they coincide.

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

$$\begin{aligned} M_k(f+g) &= E((f+g)^k) \\ &= \sum_r \binom{k}{r} E(f^r g^{k-r}) \\ &= \sum_r \binom{k}{r} M_r(f) M_{k-r}(g) \end{aligned}$$

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

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

Thus, job done, and theorem proved, or rather Theorem 2.29 reproved. □

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

2e. Exercises

Exercises:

Exercise 2.32.

Exercise 2.33.

Exercise 2.34.

Exercise 2.35.

Exercise 2.36.

Exercise 2.37.

Exercise 2.38.

Exercise 2.39.

Bonus exercise.

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:

$$12.456\,783\,872$$

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

This computation can surely be done, but, annoyingly, it must be done from right to left, instead of left to right, as we would prefer. I mean, personally I would be most

interested in knowing first what happens at left, if the integer part is 39 or 40, but go do all the computation, starting from the right, in order to figure out that. In short, my feeling is that this addition algorithm, while certainly good, is a bit deceiving.

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

$$12.456\,783\,872$$

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

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

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

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

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

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

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

$$\ldots$$

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

FACT 3.2. *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:

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

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

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

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

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

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

THEOREM 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:

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

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

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

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

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

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

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

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

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

$$x=\pm a_1\ldots a_n.b_1b_2b_3\ldots\ldots$$

with $a_i,b_i\in\{0,1,\ldots,9\}$, with the convention $\ldots b999\ldots=\ldots(b+1)000\ldots$

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

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

(2) Moreover, we have a second claim as well, namely that any expression of type $x=\pm a_1\ldots a_n.b_1b_2b_3\ldots\ldots$ corresponds to a real number $x\in\mathbb{R}$, and that with the convention $\ldots b999\ldots=\ldots(b+1)000\ldots$, the correspondence is bijective.

(3) In order to prove now these two assertions, our first claim is that we can restrict the attention to the case $x\in[0,1)$, and with this meaning of course $0\leq x<1$, with respect to the order relation for the reals discussed in the above.

(4) Getting started now, let $x\in\mathbb{R}$, coming from a cut $\mathbb{Q}=A_x\sqcup B_x$. Since the set $A_x\cap\mathbb{Z}$ consists of integers, and is bounded from above by any element $q\in B_x$ of your choice, this set has a maximal element, that we can denote $[x]$:

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

It follows from definitions that $[x]$ has the usual properties of the integer part, namely:

$$[x]\leq x<[x]+1$$

Thus we have $x = [x] + y$ with $[x] \in \mathbb{Z}$ and $y \in [0, 1)$, and getting back now to what we want to prove, namely (1,2) above, it is clear that it is enough to prove these assertions for the remainder $y \in [0, 1)$. Thus, we have proved (3), and we can assume $x \in [0, 1)$.

(5) So, assume $x \in [0, 1)$. We are first looking for a best approximation from below of type $0.b_1$, with $b_1 \in \{0, \ldots, 9\}$, and it is clear that such an approximation exists, simply by comparing x with the numbers $0.0, 0.1, \ldots, 0.9$. Thus, we have our first digit b_1, and then we can construct the second digit b_2 as well, by comparing x with the numbers $0.b_10, 0.b_11, \ldots, 0.b_19$. And so on, which finishes the proof of our claim (1).

(6) In order to prove now the remaining claim (2), let us restrict again the attention, as explained in (4), to the case $x \in [0, 1)$. First, it is clear that any expression of type $x = 0.b_1b_2b_3 \ldots$ defines a real number $x \in [0, 1]$, simply by declaring that the corresponding cut $\mathbb{Q} = A_x \sqcup B_x$ comes from the following set, and its complement:

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

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

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

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

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

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

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

The above theorem was of course quite difficult, but this is how things are.

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 $(\cup_i A_i)^c$, we certainly have to start with $|A|$.

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

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

(4) 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 $\sigma(N-2)$.
⋮
– 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:

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

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

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

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

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,*

$$p_1 = \frac{1}{e}\sum_{k\geq 0}\frac{\delta_k}{k!}$$

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

$$p_t = e^{-t}\sum_{k\geq 0}\frac{t^k}{k!}\,\delta_k$$

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:

$$e^t = \sum_{k\geq 0}\frac{t^k}{k!}$$

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 result, regarding their mean and variance:

PROPOSITION 3.32. *The mean and variance of* p_t *are given by:*

$$E = t \quad , \quad V = t$$

In particular for the Poisson law p_1 *we have* $E = 1, V = 1$.

PROOF. We have two computations to be performed, as follows:

(1) Regarding the mean, this can be computed as follows:

$$\begin{aligned}
E &= e^{-t}\sum_{k\geq 0}\frac{t^k}{k!}\cdot k\\
&= e^{-t}\sum_{k\geq 1}\frac{t^k}{(k-1)!}\\
&= e^{-t}\sum_{l\geq 0}\frac{t^{l+1}}{l!}\\
&= te^{-t}\sum_{l\geq 0}\frac{t^l}{l!}\\
&= t
\end{aligned}$$

(2) For the variance, we first compute the second moment, as follows:

$$\begin{aligned}
M_2 &= e^{-t}\sum_{k\geq 0}\frac{t^k}{k!}\cdot k^2\\
&= e^{-t}\sum_{k\geq 1}\frac{t^k k}{(k-1)!}\\
&= e^{-t}\sum_{l\geq 0}\frac{t^{l+1}(l+1)}{l!}\\
&= te^{-t}\sum_{l\geq 0}\frac{t^l l}{l!}+te^{-t}\sum_{l\geq 0}\frac{t^l}{l!}\\
&= te^{-t}\sum_{l\geq 1}\frac{t^l}{(l-1)!}+t\\
&= t^2e^{-t}\sum_{m\geq 0}\frac{t^m}{m!}+t\\
&= t^2+t
\end{aligned}$$

Thus the variance is $V=M_2-E^2=(t^2+t)-t^2=t$, as claimed. □

At the theoretical level now, we first have the following result:

THEOREM 3.33. *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 3.34. *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. □

We can now establish the Poisson Limit Theorem, as follows:

THEOREM 3.35 (PLT). *We have the following convergence, in moments,*

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

for any $t > 0$.

PROOF. Let us denote by ν_n the measure under the convolution sign. We have the following computation, for the Fourier transform of the limit:

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

Thus, we obtain indeed the Fourier transform of p_t, as desired. □

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

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

THEOREM 3.36. *The main character of S_N, which counts the fixed points, given by*

$$\chi = \sum_i \sigma_{ii}$$

via the standard embedding $S_N \subset O_N$, follows the Poisson law p_1, in the $N \to \infty$ limit. More generally, the truncated characters of S_N, given by

$$\chi_t = \sum_{i=1}^{[tN]} \sigma_{ii}$$

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

PROOF. Let us construct the main character of S_N, as in the statement. The permutation matrices being given by $\sigma_{ij} = \delta_{i\sigma(j)}$, we have the following formula:

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

In order to establish now the asymptotic result in the statement, regarding these characters, we must prove the following formula, for any $r \in \mathbb{N}$, in the $N \to \infty$ limit:

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

We already know that this formula holds at $r = 0$. In the general case now, we have to count the permutations $\sigma \in S_N$ having exactly r points. Now since having such a

permutation amounts in choosing r points among $1,\ldots,N$, and then permuting the $N-r$ points left, without fixed points allowed, we have:

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

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

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

Now by using the computation at $r=0$, that we already have, from (1), it follows that with $N\to\infty$ we have the following estimate:

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

Thus, we obtain as limiting measure the Poisson law of parameter 1, as stated. Finally, the last assertion follows in a similar way, with all this being quite standard. □

3e. Exercises

Exercises:

Exercise 3.37.

Exercise 3.38.

Exercise 3.39.

Exercise 3.40.

Exercise 3.41.

Exercise 3.42.

Exercise 3.43.

Exercise 3.44.

Bonus exercise.

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:*

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

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:

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

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

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

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

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

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

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

Very nice all this, so have have two equivalent definitions for the real numbers. 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. Number fields

Switching topics now, but still remaining quite philosophical, about numbers, in more advanced mathematical terms, the basic operations on the rationals, namely sum, product and inversion, tell us that $\mathbb{Q}$ is a field, in the following abstract sense:

DEFINITION 4.5. *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 $a(b + c) = ab + ac$.*

Apparently, what we did before, with our philosophical discussion regarding creation, $\diamondsuit \to \mathbb{N} \to \mathbb{Z} \to \mathbb{Q}$, was to construct the simplest possible field, $\mathbb{Q}$. However, this is not

exactly true, because, by a strange twist of fate, the numbers $0, 1$, whose presence in a field is mandatory, $0, 1 \in F$, can form themselves a field, with addition as follows:

$$1 + 1 = 0$$

To be more precise, according to our field axioms, we certainly must have:

$$0 + 0 = 0 \times 0 = 0 \times 1 = 1 \times 0 = 0$$

$$0 + 1 = 1 + 0 = 1 \times 1 = 1$$

Thus, everything regarding the addition and multiplication of $0, 1$ is uniquely determined, except for the value of $1 + 1$. And here, you would say that we should normally set $1 + 1 = 2$, with $2 \neq 0$ being a new field element, but the point is that $1 + 1 = 0$ is something natural too, this being the addition modulo 2:

$$1 + 1 = 0(2)$$

And, what we get in this way is a field, denoted as follows:

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

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

THEOREM 4.6. *The following happen:*

(1) $\mathbb{Q}$ *is the simplest field having the property* $1 + \ldots + 1 \neq 0$*, in the sense that any field* F *having this property must contain it,* $\mathbb{Q} \subset F$*.*

(2) *The property* $1 + \ldots + 1 \neq 0$ *can hold or not, and if not, the smallest number of terms needed for having* $1 + \ldots + 1 = 0$ *is a certain prime number* p*.*

(3) $\mathbb{F}_p = \{0, 1, \ldots, p - 1\}$*, with* p *prime, is the simplest field having the property* $1 + \ldots + 1 = 0$*, with* p *terms, in the sense that this implies* $\mathbb{F}_p \subset F$*.*

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

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

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

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

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

Getting back now to our philosophical discussion regarding numbers, what we have in Theorem 4.6 is not exactly good news, suggesting that, on purely mathematical grounds, there is a certain rivalry between $\mathbb{Q}$ and $\mathbb{F}_p$, as being the simplest field.

So, which of these two fields shall we study here, say as having been created first? Not an easy question, and as an answer to this, we have:

ANSWER 4.7. *Ignoring what pure mathematics might say, and trusting instead physics and chemistry, we will choose to trust in $\mathbb{Q}$, as being the simplest field.*

Moving ahead with some more arithmetic, inside $\mathbb{Q}$ and perhaps other fields too, let us start with the following key theorem of Fermat, for the usual integers:

THEOREM 4.8. *We have the following congruence, for any prime p,*

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

called Fermat's little theorem.

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

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

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

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

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

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 4.9. *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 4.10. *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:

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

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:

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

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

$$a^q = a$$

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

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

THEOREM 4.11. *Given a field F, any finite subgroup of its multiplicative group*

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

must be cyclic.

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

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

(2) In order to prove this claim, let $d = (m, n)$, write $d = am + bn$ with $a, b \in \mathbb{Z}$, and set $k = g^a h^b$. 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 4.12. *Given a finite field F, with $q = |F|$ we have*

$$a^q = a$$

for any $a \in F$.

PROOF. According to Theorem 4.11 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 4.13. *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 4.14. *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 4.9 must be an isomorphism. Thus, done with this.

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

(3) To be more precise, this polynomial $X^q - X$ must have some sort of abstract, minimal "splitting field", and this is how $\mathbb{F}_q$ comes, both existence and uniqueness. We will be back to this, which is something non-trivial, later in this book, with details. □

4c. Legendre symbol

Getting back now to what we did before in chapter 2, in relation with the equation $a = b^2(c)$, we have the following definition, putting everything on a solid basis:

DEFINITION 4.15. *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 4.16. *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 4.15, we obtain the formula in the statement. □

As a first consequence of the Euler formula, we have the following result:

PROPOSITION 4.17. *We have the following formula, valid for any $a, b \in \mathbb{Z}$:*

$$\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right)$$

That is, the Legendre symbol is multiplicative in its upper variable.

PROOF. This is clear indeed from the Euler formula, because $a^{\frac{p-1}{2}}(p)$ is obviously multiplicative in $a \in \mathbb{Z}$. Alternatively, this can be proved as well directly, with no need for the Fermat formula used in the proof of Euler, just by thinking at what is quadratic residue and what is not in $\mathbb{F}_p$, along the lines of (2) in the proof of Theorem 4.16. □

The above result looks quite conceptual, and as consequences, we have:

PROPOSITION 4.18. *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 4.17, 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 4.17. □

So, computing the Legendre symbol. There are many things to be known here, and all must be known, for efficient application, to the real life. We have opted to present them all, of course with full proofs, when these proofs are easy, and leave the more complicated proofs for later. As a first and main result, which is something heavy, we have:

THEOREM 4.19. *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:

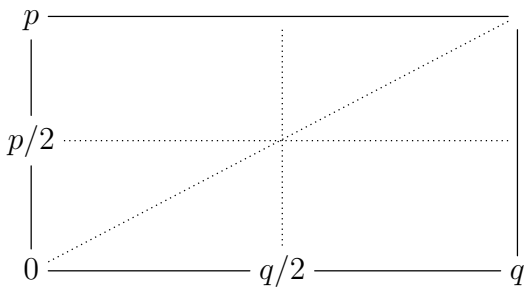
$$\left(\frac{a}{q}\right) = (-1)^n$$

(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:

$$\left(\frac{3}{173}\right) = (-1)^{\frac{3-1}{2}\cdot\frac{173-1}{2}} \left(\frac{173}{3}\right) = \left(\frac{173}{3}\right) = \left(\frac{2}{3}\right) = -1$$

In fact, when combining Theorem 4.19 with Proposition 4.17, it is quite clear that, no matter how big p is, if a has only small prime factors, we are saved.

Besides Proposition 4.17, the quadratic reciprocity formula comes accompanied by two other statements, which are very useful in practice. First, at $a=-1$, we have:

PROPOSITION 4.20. *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 4.21. *We have the following formula,*

$$\left(\frac{2}{p}\right)=\begin{cases}1 & \text{if } p=1,7(8)\\ -1 & \text{if } p=3,5(8)\end{cases}$$

solving in practice the equation $b^2=2(p)$.

PROOF. This is actually a bit complicated. The Euler formula at $a=2$ gives:

$$\left(\frac{2}{p}\right)=2^{\frac{p-1}{2}}(p)$$

However, with more work, we have the following formula, which gives the result:

$$\left(\frac{2}{p}\right)=(-1)^{\frac{p^2-1}{8}}$$

We will be back to this later in this book, with a full proof for it. □

As a continuation of this, speaking Legendre symbol for small values of the upper variable, we can try to compute these for $a=\pm\,3,4,5,6,7,8,\ldots$ But by multiplicativity plus Proposition 4.20 plus Theorem 4.21 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 4.19. We first have:

PROPOSITION 4.22. *We have the following formula,*

$$\left(\frac{3}{p}\right)=\begin{cases}1 & \text{if } p=1,11(12)\\ -1 & \text{if } p=5,7(8)\end{cases}$$

valid for any prime $p\geq 5$.

PROOF. By quadratic reciprocity, we have the following formula:

$$\left(\frac{3}{p}\right)=(-1)^{\frac{3-1}{2}\cdot\frac{p-1}{2}}\left(\frac{p}{3}\right)=(-1)^{\frac{p-1}{2}}\left(\frac{p}{3}\right)$$

Now since the sign depends on p modulo 4, and the symbol on the right depends on p modulo 3, we conclude that our symbol depends on p modulo 12, and the computation gives the formula in the statement. Finally, we have the following formula too:

$$\left(\frac{3}{p}\right)=(-1)^{\left[\frac{p+1}{6}\right]}$$

Indeed, the quantity on the right is something which depends on p modulo 12, and is in fact the simplest functional implementation of the formula in the statement. □

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

PROPOSITION 4.23. *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 4.24. *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 4.25. *The theory of Jacobi symbols can be further extended into a theory of Kronecker symbols, according to the formula*

$$\left(\frac{a}{\pm p_1^{s_1}\ldots p_k^{s_k}}\right)=\left(\frac{a}{\pm 1}\right)\left(\frac{a}{p_1}\right)^{s_1}\ldots\left(\frac{a}{p_k}\right)^{s_k}$$

with the denominator being an arbitrary integer, via suitable values for

$$\left(\frac{a}{2}\right)\quad,\quad\left(\frac{a}{-1}\right)\quad,\quad\left(\frac{a}{0}\right)$$

and this theory has as results those imported from the Jacobi theory.

PROOF. Unlike the extension from Legendre to Jacobi, which was something straightforward, here we have some work to be done, in order to figure out the correct values of the 3 symbols in the statement. The answer for the first symbol is as follows:

$$\left(\frac{a}{2}\right)=\begin{cases}1 & \text{if } a=\pm 1(8)\\ 0 & \text{if } a=0(2)\\ -1 & \text{if } a=\pm 3(8)\end{cases}$$

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:

$$\left(\frac{a}{0}\right)=\begin{cases}1 & \text{if } a=\pm 1\\ 0 & \text{if } a\neq\pm 1\end{cases}$$

And we will leave this as an instructive exercise, to figure out what the puzzle exactly is, and why these are the correct answers. And for an even better exercise, cover with a cloth the present proof, and try to figure out everything by yourself. □

4d. Primes, revised

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

THEOREM 4.26. *We have the following formula, implying $|P| = \infty$:*

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

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

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

valid for any integer $N \geq 2$.

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

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

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

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

In what concerns the sum on the left, this is well-known to be ∞. In what concerns now the product on the right, this can be estimated, and we are led to the result. □

Let us discuss now some wild arithmetic tricks, for dealing with equations over the rationals, and with the rational numbers themselves, based on the notion of p-adic number. The idea will be very simple, namely that of completing $\mathbb{Q}$ with respect to a different norm, which privileges the prime number p that we have chosen in advance.

Before that, some motivational talk. The dream in arithmetics, usually concerned with solving equations $f = 0$ over the rationals, is something very simple, namely:

DREAM 4.27. *I checked that my equation $f = 0$ has solutions modulo p, for any prime p, so my equation must have solutions over $\mathbb{Q}$.*

As a first observation, the dream holds when f is constant, $f = c$. Indeed, ignoring a bit the differences between integers and rationals, $c = 0(p)$ for any prime p means $c = 0$, so our equation is $c = 0$, having any rational number $x \in \mathbb{Q}$ as solution.

Along the same lines, there are some other examples of very simple equations $f = 0$ for which the dream holds. However, such equations are usually so simple, that we can solve them right away, and so our dream for them is not useful. In general, for more complicated equations, our dream remains wrong, and must be fine-tuned.

As a second piece of motivation, let us talk some analysis too. Everything in analytic number theory comes from the Euler formula, namely:

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

But this is again something of "local-global" type, with on the left the global quantity, that is, a usual number, which actually happens to be ∞, in our case, and on the right the "local" versions of this number, with respect to the various primes p.

Summarizing, our dream is something important, both from the algebraic and analytic perspective, and is definitely worth a second look, with the aim of fixing it. We are led in this way to the following update to it, which is a bit more modest:

HOPE 4.28. *I checked that my equation $f=0$ has solutions with respect to any prime p, in a suitable sense, so my equation must have solutions over $\mathbb{Q}$.*

So, this will be our plan for what follows, doing some mathematics, as for this hope come true. We will see that this can indeed be done, with our vague wording above "with respect to any prime p, in a suitable sense" being replaced by something very precise and mathematical, namely "over the p-adics, for any prime p", and with the statement itself being a deep principle in number theory, called Hasse local-global principle.

Getting to work now, let us further reformulate our dreams and hopes, as follows:

QUESTION 4.29. *What are the p-adic numbers, defined with respect to a chosen prime number p, making the local-global principle work?*

In answer, let us temporarily forget about equations, and the local-global principle, and simply pick a prime number p, and look at the world from the perspective of p. So, imagining that we are p, both me and you, what we see is something as follows:

(1) First, we see all sorts of integers $a\in\mathbb{Z}$. Some appear friendly, namely those of the form $a\in p\mathbb{Z}$, while the others, of the form $a\notin p\mathbb{Z}$, appear bizarre and distant.

(2) Moreover, between friends $a\in p\mathbb{Z}$, those of the form $a\in p^2\mathbb{Z}$ appear particularly close. And among them, $a\in p^3\mathbb{Z}$ are truly very close friends. And so on.

(3) Then, we see all sorts of rationals, $r=a/b$, and again, some are close, some are distant, depending on the exact p^k factor, with $k\in\mathbb{Z}$, appearing inside r.

(4) In particular, the rationals of the form $r=1/p^k$ with $k>>0$ appear really frightening. Fortunately they are very far away from us, we can barely see them.

(5) And finally, we can see some irrationals $x\notin\mathbb{Q}$ too, but these being uncountable, it is quite hard to figure out how they look like, and are distributed in space.

Very good, so getting back to Earth now, let us write down a definition, based on what we saw in our Prime Number Experience. By focusing on the integers, and more generally the rationals, and leaving the irrationals for later, we have:

DEFINITION 4.30. *Given p prime, we define the p-adic norm of $r \in \mathbb{Q}$ as being:*

$$|r| = p^{-k} \quad , \quad r = p^k \frac{a}{b} \quad , \quad a, b \neq 0(p)$$

Also, we call the integer $k \in \mathbb{Z}$ the p-adic valuation of r, and denote it $k = v(r)$.

As a comment here, $|r| = p^{-k}$ is the natural choice, because according to our Prime Number Experience, the bigger $k \in \mathbb{Z}$ is, the smaller $|r| > 0$ must be, and so we are looking for a formula of type $|r| = \beta^{-k}$ with $\beta > 1$, as for this to happen. Of course, there is still a question left, in regards with the value of $\beta > 1$. But, again coming from our Prime Number Experience, if I am for instance $p = 11$, why shall I use $\beta = 17$.

Of course you might argue here that there might be some mighty universal number, such as $e = 2.7182\ldots$ or $\pi = 3.1415\ldots$ or $1/\alpha = 137.0359\ldots$ doing the job for all prime numbers p. But this cannot work, as we will see next, with some simple math.

Going ahead now with math, the question is, is our Definition 4.30 correct? That is, is $|r|$ indeed a norm? And here, is depends a bit on your background, with mathematicians being a bit dissapointed, to the point of even choosing to stop calling $|r|$ a norm, but physicists and others being fully happy with it, the result being as follows:

THEOREM 4.31. *The p-adic norm $|r| = p^{-k}$ is not exactly a norm, but satisfies the following conditions, which are even better:*

(1) *First axiom: $|x| \geq 0$, with $|x| = 0$ when $x = 0$.*
(2) *Modified second axiom: $|xy| = |x| \cdot |y|$.*
(3) *Strong triangle inequality: $|x + y| \leq \max(|x|, |y|)$.*

PROOF. All this follows indeed from some simple arithmetics modulo p:

(1) That axiom clearly holds, with the remark that we forgot to say in Definition 4.30 that $v(0) = \infty$, by definition, because any p^k, no matter how big $k \in \mathbb{N}$ is, divides 0.

(2) As a first observation, the usual second norm axiom, namely $|\lambda x| = ||\lambda|| \cdot |x|$, with $||.||$ standing here for the usual absolute value of the numbers, definitely fails, and this because all the p-adic norms $|r|$ are by definition integer powers of p, and an arbitrary $\lambda \in \mathbb{Q}$ will mess up this. However, we have instead $|xy| = |x| \cdot |y|$, coming from:

$$v(xy) = v(x)v(y)$$

And is this good news or not. After some thinking, this modified second axiom is just as good as the failed usual second axiom, because who cares about arbitrary numbers $\lambda \in \mathbb{Q}$, not viewed from the perspective of p, I mean. More on this in a moment.

(3) Finally, let us look at sums $x+y$. Over the integers $p^k|x,y$ implies $p^k|x+y$, and with a bit of fractions arithmetic, that we will leave here as an easy exercise, the same holds for rationals, in the sense that we have, in terms of the p-adic valuation:

$$v(x+y)\geq\min(v(x),v(y))$$

Thus the p-adic norm itself, $|r|=p^{-v(r)}$, satisfies the following inequality:

$$|x+y|\leq\max(|x|,|y|)$$

Now, what does this inequality mean, geometrically? Good question, and as a first remark, since this is obviously something stronger than the usual triangle inequality satisfied by the norms, $|x+y|\leq|x|+|y|$, we will call it strong triangle inequality. □

Before going ahead, let us further examine the strong triangle inequality found in the above. This is something new to us, and as a further result on it, we have:

PROPOSITION 4.32. *The strong triangle inequality implies*

$$|x|\neq|y|\implies|x+y|=\max(|x|,|y|)$$

and with this being valid for any modified norm, in the sense of Theorem 4.31.

PROOF. This is again something elementary, the idea being as follows:

(1) In what regards the p-adic norm, going back to (3) in the proof of Theorem 4.31, we can add there the observation that, trivially over the integers, and then over the rationals too, with a bit of fraction work, the p-adic valuation satisfies:

$$v(x)\neq v(y)\implies v(x+y)=\min(v(x),v(y))$$

Thus the p-adic norm itself satisfies the condition in the statement.

(2) More generally now, and with this being something quite interesting, our claim is that this phenomenon is valid for any generalized norm in the sense of Theorem 4.31. Indeed, assume that $|x|\geq 0$, with $|x|=0$ when $x=0$, as usual, and that:

$$|xy|=|x|\cdot|y|\quad,\quad|x+y|\leq\max(|x|,|y|)$$

In order to prove our result, assume $|x|>|y|$. We then have, trivially:

$$|x+y|\leq\max(|x|,|y|)=|x|$$

(3) In the other sense now, we have to work a bit. We have the following computation, with at the end the observation that the max cannot be $|y|$, because if that would be the case, the inequality that we would obtain would be $|x|\leq|y|$, contradicting $|x|>|y|$:

$$\begin{aligned}|x|&=|(x+y)-y|\\&\leq\max(|x+y|,|y|)\\&=|x+y|\end{aligned}$$

Thus, we have equality in the estimate in (2), as desired. □

Very nice all this, and getting back now to what we have in Theorem 4.31, namely the modified norm axioms there, we can formulate, as a simple consequence:

PROPOSITION 4.33. *The p-adic norm $|r| = p^{-k}$ is not exactly a norm, but*

$$d(x, y) = |x - y|$$

is a distance. Thus, the rationals $\mathbb{Q}$ become in this way a metric space.

PROOF. With the conditions satisfied by the p-norm $|r|$ in hand, it follows, trivially, that $d(x, y) = |x - y|$ is indeed a distance, making $\mathbb{Q}$ a metric space. □

Now let us turn to irrationals. The quite blurry picture that we saw during our Prime Number Experience, and with the blame at that time being on the uncountability of these beasts, in the lack of something better, can be now explained. Indeed, what we saw were not the "usual" irrationals $x \in \mathbb{R} - \mathbb{Q}$, but rather some irrationals $x \in \mathbb{Q}_p - \mathbb{Q}$ viewed from the perspective of p, constructed according to the following result:

THEOREM 4.34. *By completing $\mathbb{Q}$ with respect to the p-adic distance*

$$d(x, y) = |x - y|$$

we obtain a certain field $\mathbb{Q}_p$, called field of p-adic numbers.

PROOF. This is something very standard, with the passage $\mathbb{Q} \to \mathbb{Q}_p$ being very similar to the passage $\mathbb{Q} \to \mathbb{R}$, that we are very familiar with. In fact, some things get even simpler for p-adics, due to the strong triangle inequality satisfied by the norm. □

What is next? Many things, especially in relation with understanding what the p-adic irrationals $x \in \mathbb{Q}_p - \mathbb{Q}$ really are, concretely speaking. But before that, inspired by the theory of usual numbers, $\mathbb{Z} \subset \mathbb{Q}$, we can introduce the p-adic integers, as follows:

THEOREM 4.35. *We can introduce the p-adic integers $\mathbf{Z}_p \subset \mathbb{Q}_p$ as being*

$$\mathbf{Z}_p = \left\{ x \in \mathbb{Q}_p \,\middle|\, |x| \leq 1 \right\}$$

not to be confused with $\mathbb{Z}_p$, and this is a ring, appearing as completion of $\mathbb{Z} \subset \mathbf{Z}_p$.

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

(1) We can certainly introduce a set $\mathbf{Z}_p \subset \mathbb{Q}_p$ by the condition in the statement, and the ring axioms are all clear from the modified norm conditions, from Theorem 4.31, the verifications of the fact that $\mathbf{Z}_p$ is stable under sums and products being as follows:

$$|x|, |y| \leq 1 \implies |x + y| \leq \max(|x|, |y|) \leq 1$$

$$|x|, |y| \leq 1 \implies |xy| = |x| \cdot |y| \leq 1$$

(2) Next, since the valuation of a usual integer $x \in \mathbb{Z}$ satisfies $v(x) \geq 0$, the norm satisfies $|x| \leq 1$, and so we have an inclusion $\mathbb{Z} \subset \mathbf{Z}_p$, as in the statement.

(3) With a bit more work, we can see that $\mathbf{Z}_p$ is closed with respect to the p-adic norm, and also, that is appears as the completion of its subring $\mathbb{Z} \subset \mathbf{Z}_p$.

(4) Finally, and getting now into hot stories and other funny facts, the ring of p-adic integers $\mathbf{Z}_p$ is obviously not to be confused with the cyclic group $\mathbb{Z}_p$. There are actually two schools of thought here, with the other school denoting the p-adic integers by $\mathbb{Z}_p$, and using for the cyclic group all sorts of bizarre notations, such as C_p.

(5) In what regards our philosophy, that is very simple. If you need some sort of integers with respect to p, for your mathematics, this is a no-brainer, go with the remainders modulo p, or even better, with the p-th roots of unity, and that will solve your mathematical question, in 99% of the cases. And in the remaining 1% cases, what you need are probably the p-adic integers. So, assuming at least a little bit of modesty and common sense, the simplest notation, $\mathbb{Z}_p$, should be attributed to the cyclic group. □

With this understood, let us get now to the irrationals, and non-integers, and the p-adic numbers in general, viewed as a whole. Obviously, in order to understand them, we must understand well the Cauchy sequences and convergence in $\mathbb{Q}_p$. But here, many surprises are waiting for us, as for instance the following notorious formula:

PROPOSITION 4.36. *We have the following formula,*

$$\sum_{k=0}^{\infty} p^k = \frac{1}{1-p}$$

with respect to the p-adic norm.

PROOF. By using $p^n \to 0$, with respect to the p-adic norm, we have:

$$\begin{aligned}
\sum_{k=0}^{n-1} p^k &= \frac{1-p^n}{1-p} \\
&= \frac{1}{1-p} - \frac{p^n}{1-p} \\
&\simeq \frac{1}{1-p} - \frac{0}{1-p} \\
&= \frac{1}{1-p}
\end{aligned}$$

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

Quite cool the above formula, we are learning new things here, aren't we, and even more spectacular is its $p = 2$ particular case, which reads:

$$\sum_{k=0}^{\infty} 2^k = -1$$

But we will not get scared by this. Moving ahead now with our general program, of understanding the Cauchy sequences and convergence in $\mathbb{Q}_p$, we have:

THEOREM 4.37. *Convergence in $\mathbb{Q}_p$, and corresponding picture of $\mathbb{Q}_p$.*

PROOF. This follows, as usual, from some elementary arithmetic modulo p, with the conclusion being that the arbitrary p-adic numbers $x \in \mathbb{Q}_p$ have, after all, a quite intuitive interpretation, when it comes to their decimal, or rather p-adic, expansion. □

Finally, again in the analogy with what we know about numbers, we have:

THEOREM 4.38. *The field of p-adic numbers $\mathbb{Q}_p$ can be further enlarged,*

$$\mathbb{Q}_p \subset \bar{\mathbb{Q}}_p$$

into an algebrically closed field $\bar{\mathbb{Q}}_p$, having many interesting properties.

PROOF. This follows indeed by using the general $F \to \bar{F}$ technology from Galois theory, and with this being quite similar to the construction $\mathbb{R} \to \mathbb{C}$. □

Getting back now to our original motivations, namely equations for the integers and rationals, and the local-global principle for them, that we are dreaming of, we have:

THEOREM 4.39. *Hasse local-global principle, and Hasse-Minkowski theorem.*

PROOF. Many things can be said here, but the proofs use a lot of non-trivial algebra. We will present here the main ideas, behind these proofs, with some details missing. □

So long for completions of $\mathbb{Q}$. We will be back to this, on several occasions.

4e. Exercises

Exercises:

EXERCISE 4.40.

EXERCISE 4.41.

EXERCISE 4.42.

EXERCISE 4.43.

EXERCISE 4.44.

EXERCISE 4.45.

EXERCISE 4.46.

EXERCISE 4.47.

Bonus exercise.

Part II

Geometry

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

CHAPTER 5

Triangles

5a. Parallel lines

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

CAT 5.5. *What about a formula of type*

$$PQ = \lambda P + (1 - \lambda)Q$$

proving your axiom.

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

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

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

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

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

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

$$K \cap L = \emptyset$$

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

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

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

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

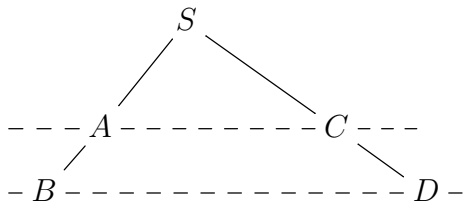
Very good, and now with Axiom 5.1 and Axiom 5.8 in hand, we are potentially ready for doing some geometry. However, this is not exactly true, and we will need as well:

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

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

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

THEOREM 5.10 (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 $AC||BD$.

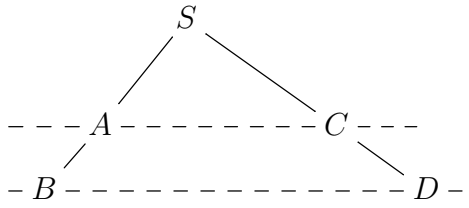
PROOF. We have indeed the following computation, based on the usual area formula for the triangles, that is, half of side times height, used multiple times:

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

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

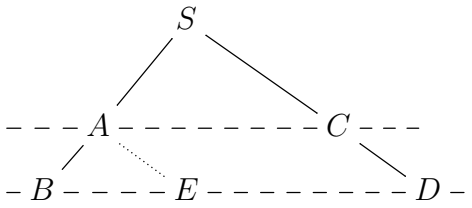


the following equality, involving the same number, holds as well:

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

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:



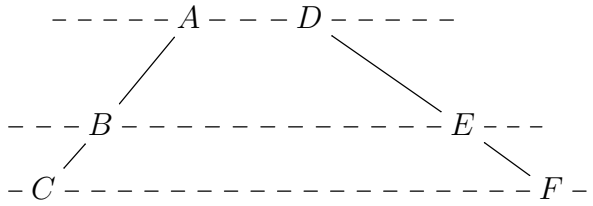
By using Theorem 5.10, we have then the following computation, as desired:

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

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 5.12 (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:

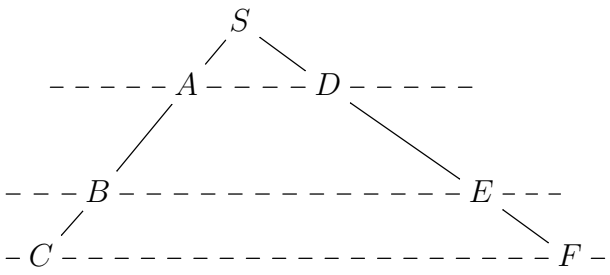
$$\frac{AB}{BC}=\frac{DE}{EF}$$

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 5.10 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. □

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

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

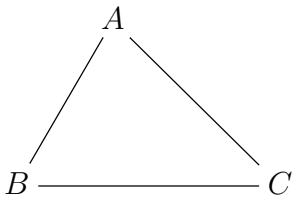
5b. Angles, triangles

Welcome to 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 5.13. *Given a triangle ABC, the following happen:*

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

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



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

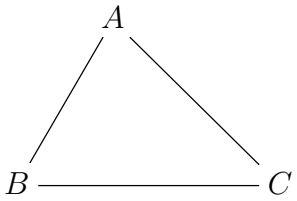
(1) Come with a small circle, inside ABC, and then inflate it, as to touch all 3 edges. The center of the circle will be then at equal distance from all 3 edges, so it will lie on all 3 angle bisectors. Thus, we have constructed the incenter, as required.

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

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

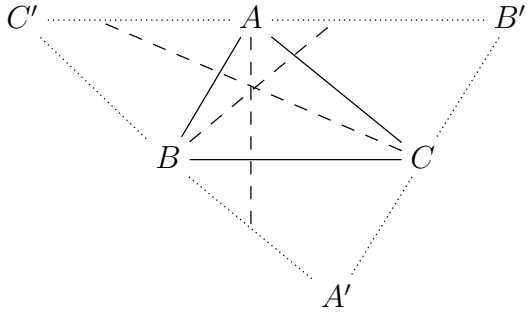
Thus P lies on the median emanating from A, and a similar argument shows that P lies as well on the medians emanating from B, C. Thus, we have our barycenter.

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



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

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

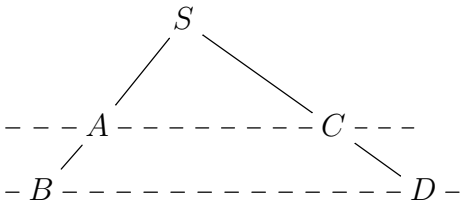


Thus, we are led to the conclusions in the statement. □

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 5.14. *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 5.15. *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:

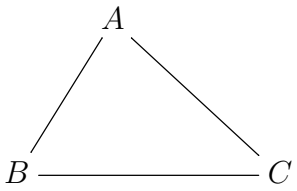
DEFINITION 5.16. *We can talk about the numeric value of angles, as follows:*

(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.

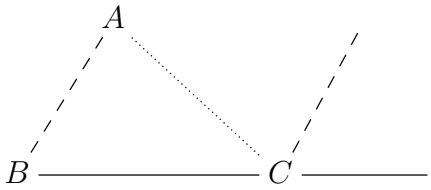
Getting back to work now, theorems and proofs, in relation with the above, here is a key result, which will be our main tool for the study of the angles:

THEOREM 5.17. *In an arbitrary triangle*



the sum of all three angles is 180°.

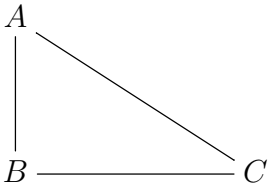
PROOF. This does not seem obvious to prove, with bare hands, but as usual, in such situations, some tricky parallels can come to the rescue. Let us prolong indeed the segment BC a bit, on the C side, and then draw a parallel at C, to the line AB, as follows:



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

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

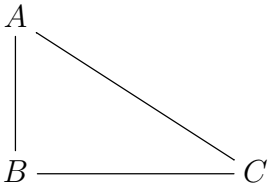
DEFINITION 5.18. *We call right triangle a triangle of type*



having one of the angles equal to 90°.

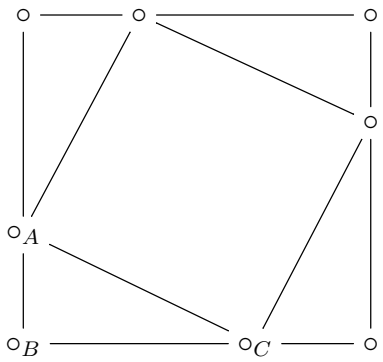
Many things can be said about right triangles, in particular with:

THEOREM 5.19 (Pythagoras). *In a right triangle ABC,*



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

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



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

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

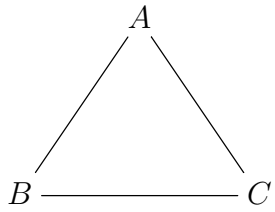
On the other hand, the outer square is made of the smaller square, having side AC, and of four identical right triangles, having sizes AB, BC. Thus:

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

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

As a second important angle, we have the 60° angle, which usually appears via:

THEOREM 5.20. *In an equilateral triangle, having all sides equal,*

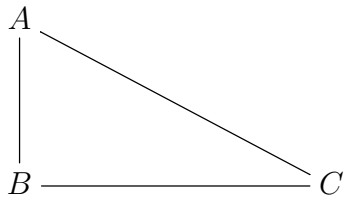


all angles equal 60°*.*

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

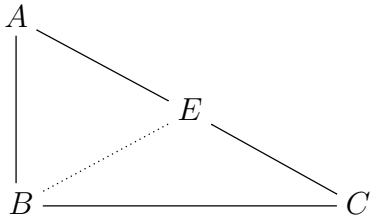
Another interesting angle is the 30° one. About it, we have:

THEOREM 5.21. *In a right triangle having small angles* $30°, 60°$*,*



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

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



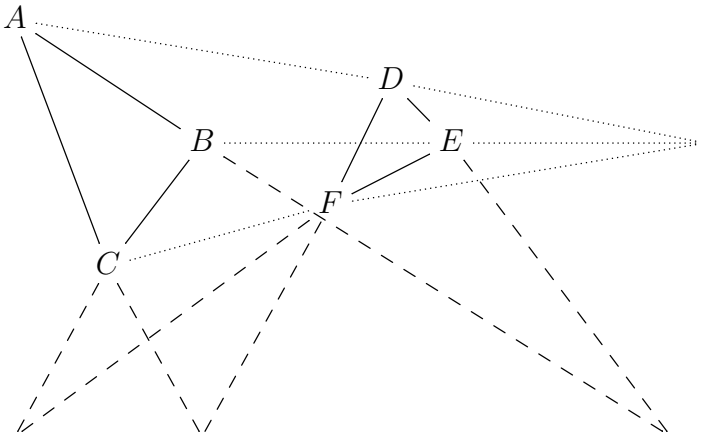
Thus, we are led to the conclusion in the statement. □

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

5c. Advanced results

Moving forward, as a basic statement now, which is something more subtle, due to Desargues, we have the following fact, that we will prove in what follows:

FACT 5.22 (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 5.23. *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, and we will choose here to use something quite conceptual. So, temporarily forgetting about Desargues, we have the following result, which is something having its own interest:

THEOREM 5.24. *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. Here the fact that we have a duality is something quite self-explanatory, and the statement at the end is something which holds too, the idea being as follows:

(1) We can certainly construct the correspondence $P \to p$ in the statement, which maps points $P \neq O$ to lines p not containing O, and which is clearly injective.

(2) Conversely, given a line p not containing O, we can project O on this line, to a point Q, and then construct $P \in OQ$ by the formula in the statement, $OP \cdot OQ = r^2$.

(3) We conclude from this that we have indeed a bijection $P \to p$ as in the statement, which maps points $P \neq O$ to lines p not containing O.

(4) Before getting further, let us make a few simple observations. As a first remark, when P belongs to the circle, p is the tangent to the circle, drawn at that point P.

(5) Along the same lines, some further basic observations include the fact that when P is inside the circle, p is outside of it, meaning not intersecting it, and vice versa.

(6) Getting now to the last assertion, this is something which holds indeed. We will be back to this later, with details, once we will know more about circles. □

The point now is that the Desargues configuration is self-dual, so we obtain:

THEOREM 5.25. *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 5.24:

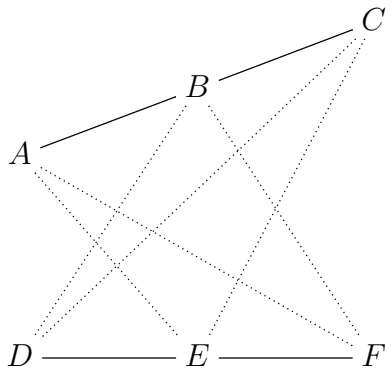
– The original triangles ABC, DEF are in central perspective precisely when the dual triangles abc, def are in axial perspective.

– 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 5.23 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 5.26 (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 5.27. *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 5.24, 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) To start with, the notion of cross ratio, as constructed in the statement, is something very natural. Observe first that we can write the cross ratio as follows:

$$(A, B, C, D) = \frac{AC}{BC} \cdot \frac{BD}{AD}$$

On the other hand, we can write as well the cross ratio as follows:

$$(A, B, C, D) = \frac{AC}{AD} \cdot \frac{BD}{BC}$$

But are these quantities really the same? Hell yes, the theory of fractions says, but go see that geometrically, and have it all the time in mind, when working with the cross ratio, that ain't no easy task, which takes a lot of practice. Welcome to geometry.

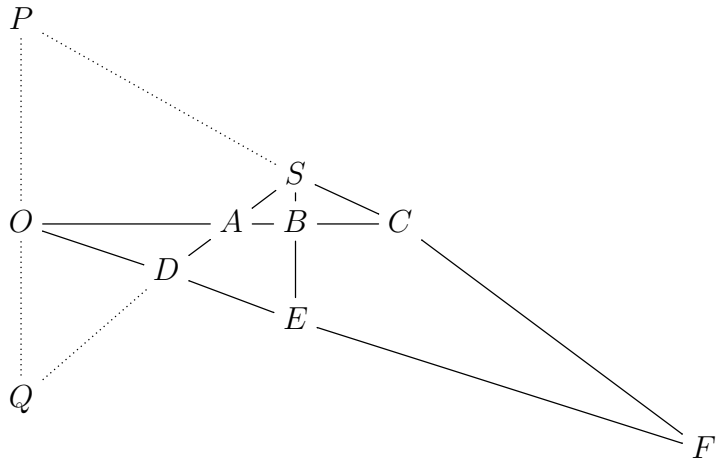
(2) Next, many other things can be said, as for instance being the fact that A, B, C, D are somehow "nicely positioned" on their line when their cross ratio is -1:

$$(A, B, C, D) = -1$$

Again, try getting familiar with this, by working out some examples, doing some computations and so on. All this is first-class geometry, that you should know.

(3) Getting now to what our statement says, in relation with points in central perspective, 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:



(4) 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:

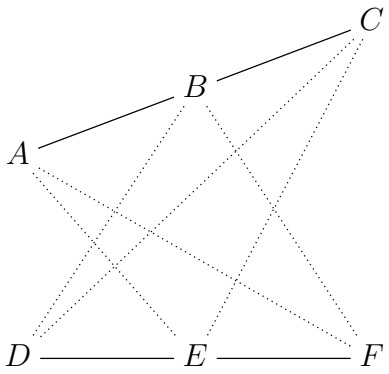
$$(O,B,C,A)=(O,E,F,D)$$

(5) 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, and we will leave working out the details here as an instructive exercise.

(6) As for the second assertion, this follows from the first one, in a standard way, and we will leave working out the details here as an instructive exercise too. □

Good news, we can now prove the Pappus theorem, as follows:

THEOREM 5.28 (Pappus). *Given a hexagon $AFBDCE$ with both the odd and the even vertices being collinear*



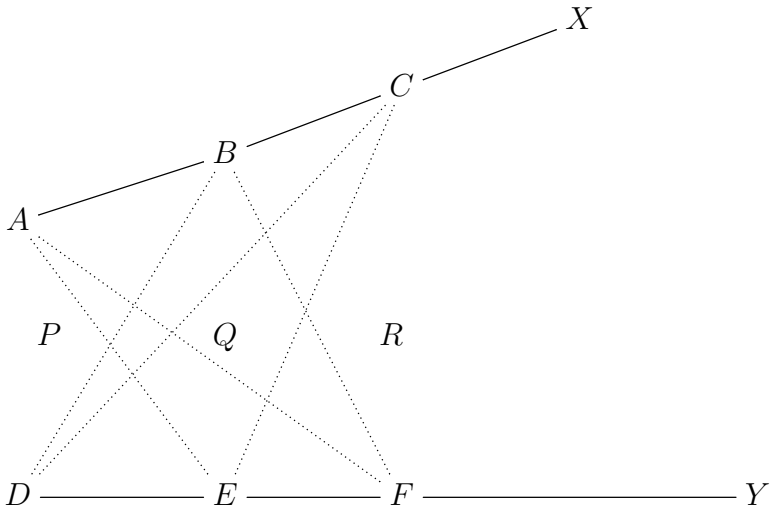
the pairs of opposite sides cross into three collinear points.

PROOF. Observe first the fancier formulation of the statement, with respect to what we had before in Fact 5.26, but this was of course more for fun, of perhaps for some deeper reasons too, these mysterious hexagons sort of rule in plane geometry, and more on this later in this book too, on several occasions. In practice now, what we have to prove remains as in Fact 5.26, and the idea is that can be proved by refining the picture, by adding some extra points, and using the cross ratio technology from Theorem 5.27:

(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 5.27, 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 5.27, this time in reverse form, we conclude that the images of X, Y via the above projections must coincide:

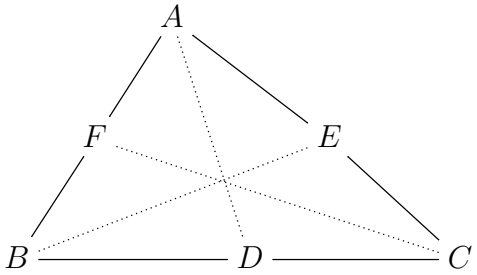
$$(DX \cap AY) \in PQ$$

But, according to our conventions above, $DX \cap AY = R$, so we obtain, as desired:

$$R \in PQ$$

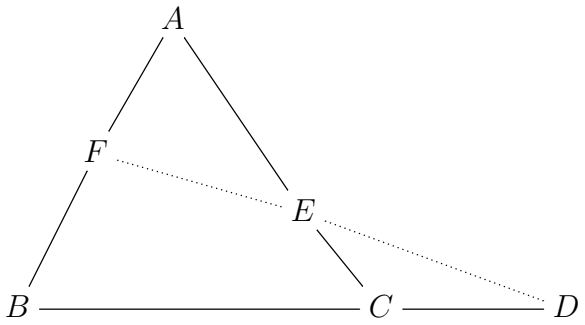
(5) Thus, result proved. As a further comment, observe that there is a relation with Desargues too. Finally, note that the Pappus configuration is self-dual. □

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 5.29 (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:

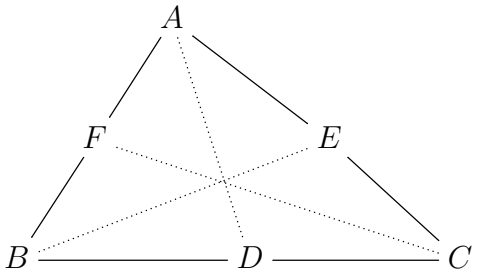
$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=-1$$

Moreover, the converse holds, with this formula guaranteeing that F, E, D are colinear.

PROOF. This is indeed something very standard, by drawing some altitudes. As for the converse, this follows from the main result, in the obvious way. □

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 5.30 (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, which is obviously related to the previous theorem of Menelaus, and which is best seen by computing some areas. As for the converse, this follows from the main result, in the obvious way. □

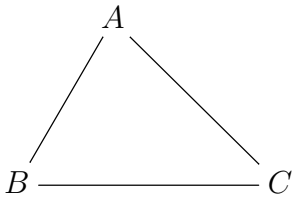
As a basic application of the Ceva theorem, we have now a new point of view on the barycenter. Indeed, the fact that the medians of a triangle cross can be seen as coming from the Ceva theorem, via the following trivial computation:

$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=1\times 1\times 1=1$$

Which is very nice, but needless to say, there is still a lot of work to be done, on the barycenter, in order to understand what cats and physicists know about it, in relation with what was said in the beginning of this chapter. More on this later in this book.

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

THEOREM 5.31. *Besides the* 4 *main centers of a triangle, discussed in the above, many more remarkable points can be associated to a triangle ABC,*



and most of these lie on a line, called Euler line of ABC.

PROOF. This is something more technical, which can be proved as well, via some work, the idea with this being as follows:

(1) To start with, it is possible to prove, via some tricks and computations, that the barycenter, the circumcenter and the orthocenter of a triangle are colinear. With this being a key result, among others providing a definition for the Euler line.

(2) Needless to say, in order for that Euler line to exist, as defined above, the triangle ABC must be assumed to be not equilateral. As for the basic example, for this, for an isosceles triangle, not equilateral, the Euler line is of course the symmetry axis.

(3) At a more advanced level now, as indicated in the statement, it is possible to construct other interesting centers of a triangle, which usually lie on the Euler line. We will be back to this in the next theorem, when discussing the nine-point circle.

(4) Finally, again at the level of more advanced results, we have the question of understanding how these various points lie on the Euler line, meaning understanding the ratios between the distances between them. Again, many things can be said here. □

So long for triangles and their centers. This was a very fashionable business long ago, but in more modern times the goals of mathematicians have slightly deviated towards arithmetic, with the must-do thing, instead of constructing a new triangle center, being that of joining the list of generalizators of the Legendre symbol. As for the truly modern times, here the goal is that of having your own version of quantum field theory.

5d. Projective geometry

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

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

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

DEFINITION 5.32. *A projective space is a space consisting of points and lines, subject to the following conditions:*

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

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

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

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

Thus, what we have is a projective space in the sense of Definition 5.32. More generally now, we have the following construction, in arbitrary dimensions:

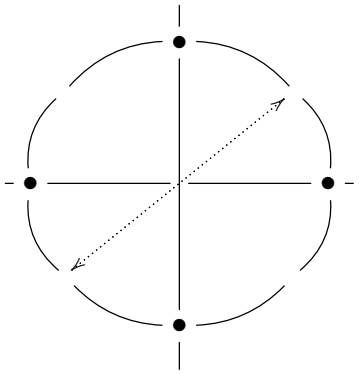
THEOREM 5.33. *We can define the projective space $P^{N-1}_{\mathbb{R}}$ as being the space of lines in $\mathbb{R}^N$ passing through the origin, and in small dimensions:*

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

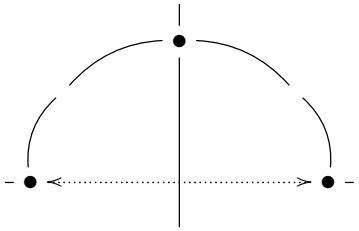
PROOF. We have several assertions here, with all this being of course a bit informal, and self-explanatory, the idea and some further details being as follows:

(1) To start with, the fact that the space $P^{N-1}_{\mathbb{R}}$ constructed in the statement is indeed a projective space in the sense of Definition 5.32 follows from definitions, exactly as in the discussion preceding the statement, regarding the case $N = 3$.

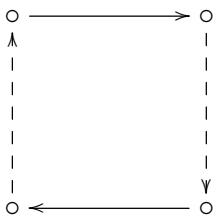
(2) At $N = 2$ now, a line in $\mathbb{R}^2$ passing through the origin corresponds to 2 opposite points on the unit circle $\mathbb{T} \subset \mathbb{R}^2$, according to the following scheme:



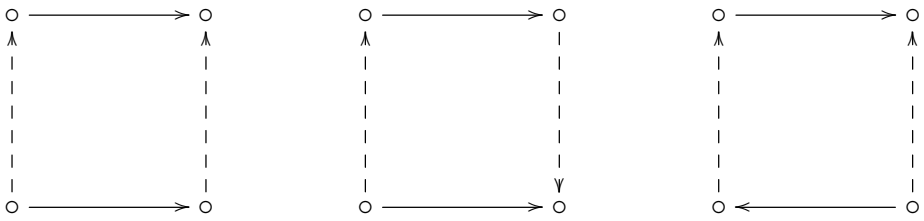
Thus, $P^1_{\mathbb{R}}$ corresponds to the upper semicircle of $\mathbb{T}$, with the endpoints identified, and so we obtain a circle, $P^1_{\mathbb{R}} = \mathbb{T}$, according to the following scheme:



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



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



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

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

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

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

5e. Exercises

Exercises:

Exercise 5.34.

Exercise 5.35.

Exercise 5.36.

Exercise 5.37.

Exercise 5.38.

Exercise 5.39.

Exercise 5.40.

Exercise 5.41.

Bonus exercise.

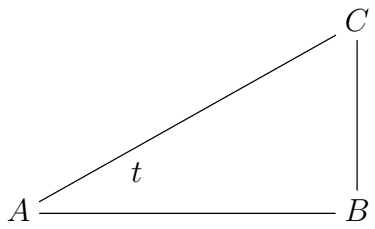
CHAPTER 6

Trigonometry

6a. Sine, cosine

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

DEFINITION 6.1. *Given a right triangle ABC,*



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

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

We call the sine and cosine basic trigonometric functions.

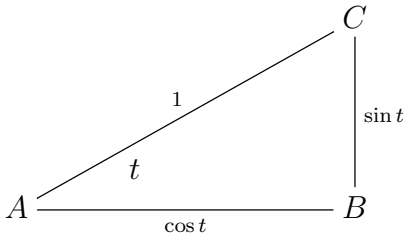
As a first observation, the sine and cosine do not depend on the choice of the given right triangle ABC having an angle t at A, and this due to the Thales theorem. In view of this, whenever possible, we will choose the right triangle ABC as to have:

$$AC = 1$$

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

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

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



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

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

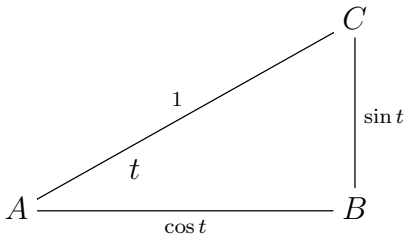
Observe that the numbers in the above two lists are the same, but written backwards in the second list. In fact, we have the following result, regarding this:

THEOREM 6.2. *The sines and cosines are subject to the formulae*

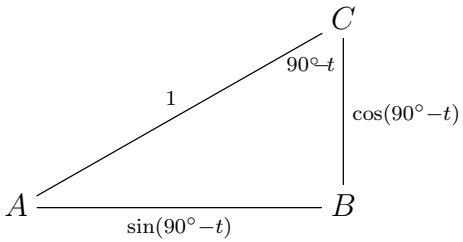
$$\sin(90^\circ - t) = \cos t \quad , \quad \cos(90^\circ - t) = \sin t$$

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:



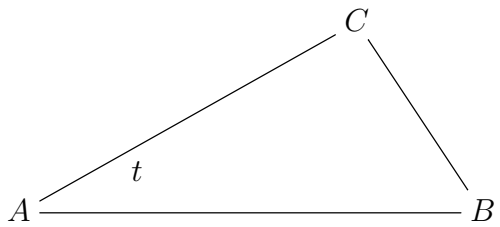
On the other hand, by focusing now at the angle at C, and perhaps twisting a bit our minds too, we have as well the following picture, for the same triangle:



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,*

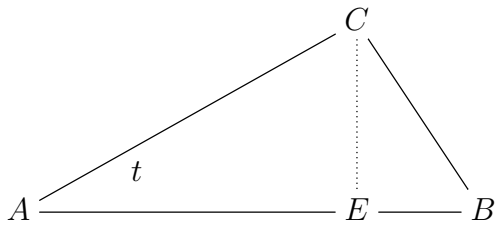


is given by the following formula, making appear the sine:

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

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

Thanks cat, interesting all this, 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, hope you agree with me. There is no way of getting into more trigonometry, before having an answer to this.

In answer now, thinking a bit, 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 of the triangle 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

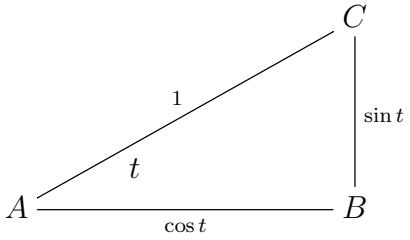
$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

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 $t > 90°$, when the triangle is drawn at the left of A, the lower side AB changes orientation:

$$AB \to BA$$

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

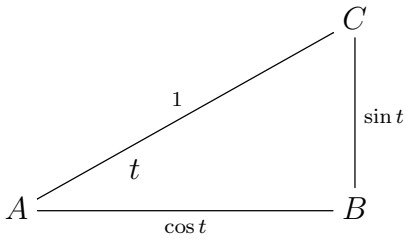
$$AB \to -AB$$

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

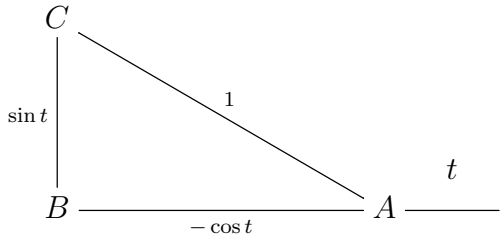
$$\cos t = -\cos(180° - t)$$

Very good all this, so let us update now Theorem 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 $t \in [0°, 180°]$, according to the following picture,*



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



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

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

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

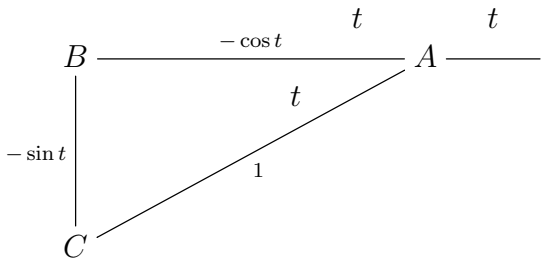
$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

holds for any triangle, without any assumption on it.

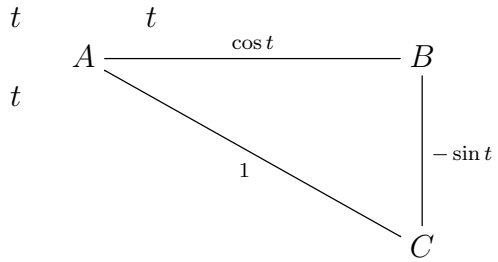
PROOF. This follows indeed by putting together all the above. □

Which sounds quite good, and normally end of the story, but let us be crazy now, and try to talk as well about the sine or cosine of angles $t < 0°$, or $t > 180°$.

Indeed, we know the recipe, namely suitably drawing our right triangle, with attention to positive and negatives. Thus, for $t \in [180°, 270°]$, our picture should be as follows:



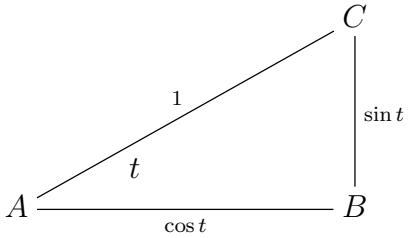
As for the next case, $t \in [270°, 360°]$, here our picture should be as follows:



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.

So, good work that we did for the day, and time now to further improve Theorem 6.7, into something truly final, trust me here, 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 $t < 0°$, 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. This follows indeed by putting together all the above, and with the basic formulae in question being as follows, and in the hope that I forgot none:

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

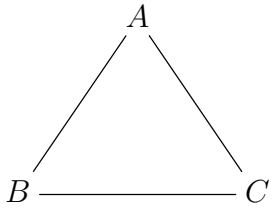
$$\sin(180^\circ - t) = \sin t \quad , \quad \cos(180^\circ - t) = -\cos t$$
$$\sin(180^\circ + t) = -\sin t \quad , \quad \cos(180^\circ + t) = -\cos t$$

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

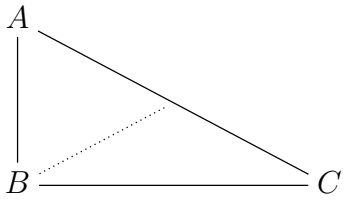
$$\sin(360^\circ - t) = -\sin t \quad , \quad \cos(360^\circ - t) = \cos t$$
$$\sin(360^\circ + t) = \sin t \quad , \quad \cos(360^\circ + t) = \cos t$$

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

In order to study now sin and cos, let us first update the numerics that we already have, for very simple angles in $[0^\circ, 90^\circ]$, to more angles, in $[0^\circ, 360^\circ]$. Let us first recall that, in order to reach to some basic formulae, our basic tool is the equilateral triangle:



Equivalently, we can use a right triangle having small angles $30^\circ, 60^\circ$:



As explained before, the study of these triangles leads to the following formulae:

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

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

By using now the formalism from Theorem 6.8, we are led in this way to:

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 of the basic angles are as follows,

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

$$\cos 120^\circ = -\frac{1}{2} \quad , \quad \cos 135^\circ = -\frac{1}{\sqrt{2}} \quad , \quad \cos 150^\circ = -\frac{\sqrt{3}}{2} \quad , \quad \cos 180^\circ = -1$$

$$\cos 210^\circ = -\frac{\sqrt{3}}{2} \quad , \quad \cos 225^\circ = -\frac{1}{\sqrt{2}} \quad , \quad \cos 240^\circ = -\frac{1}{2} \quad , \quad \cos 270^\circ = 0$$

$$\cos 300^\circ = \frac{1}{2} \quad , \quad \cos 315^\circ = \frac{1}{\sqrt{2}} \quad , \quad \cos 330^\circ = \frac{\sqrt{3}}{2} \quad , \quad \cos 360^\circ = 1$$

with this coming from the basic geometry of right triangles.

PROOF. This is indeed self-explanatory, with input coming from the above. □

6b. Trigonometry

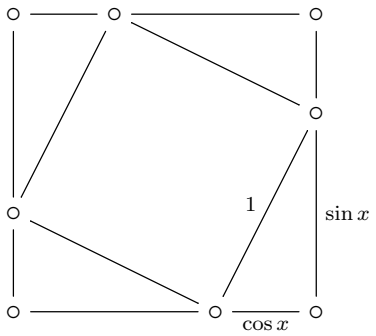
The problem is now, how to get beyond the above formulae? Not an easy question, but do not worry, we will be back to this, in due time. For the moment, as a complement to the above, let us record the following key formula, coming from Pythagoras:

THEOREM 6.10. *The sines and cosines are subject to the formula*

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

coming from Pythagoras' theorem.

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



When computing the area of the outer square, we obtain:

$$(\sin x+\cos x)^2=1+4\times\frac{\sin x\cos x}{2}$$

Now when expanding we obtain $\sin^2 x+\cos^2 x=1$, as claimed. □

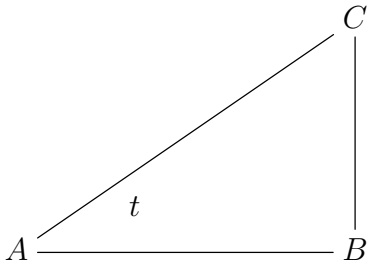
It is possible to say many more things about angles and $\sin x$, $\cos x$, and also talk about some supplementary quantities, such as the tangent:

DEFINITION 6.11. *We can talk about the tangent of angles $t\in\mathbb{R}$, as being given by*

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

with $\sin x,\cos x$ 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:

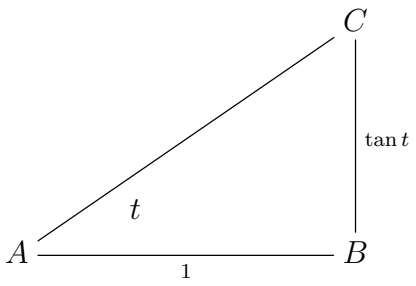
$$\tan t=\frac{\sin t}{\cos t}=\frac{BC}{AC}\Big/\frac{AB}{AC}=\frac{BC}{AB}$$

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 6.12. *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:

$$\tan(-t) = -\tan t$$

$$\tan(90° - t) = \frac{1}{\tan t} \quad , \quad \cos(90° + t) = -\frac{1}{\tan t}$$

$$\tan(180° - t) = -\tan t \quad , \quad \tan(180° + t) = \tan t$$

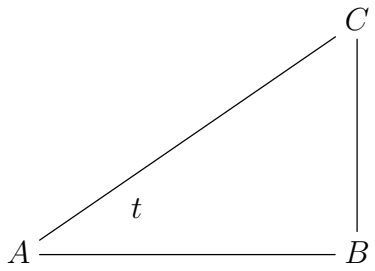
Let us record as well the formulae for the basic angles. These are as follows:

$$\tan 0° = 0 \quad , \quad \tan 30° = \frac{1}{\sqrt{3}} \quad , \quad \sin 45° = 1 \quad , \quad \sin 60° = \sqrt{3}$$

$$\tan 120° = -\sqrt{3} \quad , \quad \tan 135° = -1 \quad , \quad \tan 150° = -\frac{1}{\sqrt{3}} \quad , \quad \tan 180° = 0$$

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 6.13. *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 $\sin, \cos, \tan$.

Very nice all this, so shall we study now all these new functions too, in the spirit of what we did in the aboves for $\sin, \cos, \tan$, after all we can potentially have some mathematical fun, with our enlarged collection of 6 trigonometric functions, which looks complete, symmetric and beautiful. Not clear, so time to ask the cat. And cat says:

CAT 6.14. *These* $\sec, \csc, \cot$ *functions sound more like pure mathematics. What about trying instead* $\arcsin, \arccos, \arctan$*? Or* $\sinh, \cosh, \tanh$*? Or* $\operatorname{arcsinh}, \operatorname{arccosh}, \operatorname{arctanh}$*?*

Which sounds quite interesting, not that I fully understand what cat says, but one thing is sure, namely that we won't potentially get to any new mathematics by applying $x \to x^{-1}$, so I kind of agree with the first advice, forget about $\sec, \csc, \cot$.

As for the rest, yes I have this feeling too that many more interesting trigonometric functions are waiting for us, and more on this later in this book, once we will have some appropriate tools, beyond basic triangle geometry, in order to discuss them.

Getting back now to the basics, sine and cosine, how these can be computed, and what can be done with them, we have the following key result:

THEOREM 6.15. *The sines and cosines of sums are given by*

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

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

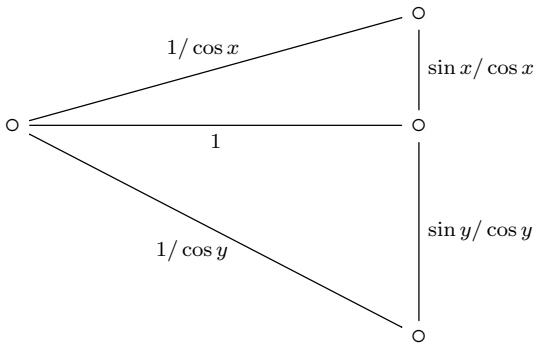
and these formulae give a formula for the tangent too, namely

$$\tan(x+y) = \frac{\tan x + \tan y}{1 - \tan x \tan y}$$

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 6.16. *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 6.15, 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 illustrations for the above formulae, we can now compute the sine, cosine and tangent of various interesting new angles, appearing as sums and differences, such as:

$$15^\circ=45^\circ-30^\circ$$

$$75^\circ=45^\circ+30^\circ$$

In fact, thinking well, this is pretty much it, modulo periodicity formulae. So, all in all, we can now compute the sine, cosine and tangent of all the multiples of 15°. Nice.

Time now for more advanced trigonometry. Indeed, by taking $x=y$ in Theorem 6.15 we obtain some interesting formulae for the duplication of angles, as follows:

THEOREM 6.17. *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,

$$\begin{aligned}\cos(2t) &= \cos^2t-\sin^2t\\ &= 2\cos^2t-1\\ &= 1-2\sin^2t\end{aligned}$$

with all these three formulae being useful, in practice.

PROOF. By taking $x=y=t$ in the formulae from Theorem 6.15, we obtain:

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

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

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 6.18. *The tangents of the doubles of angles are given by*

$$\tan(2t)=\frac{2\tan t}{1-\tan^2t}$$

provided as usual that the denominator is nonzero.

PROOF. This follows indeed by taking $x = y = t$ in the formula for tangents from Theorem 6.15. 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 6.17. □

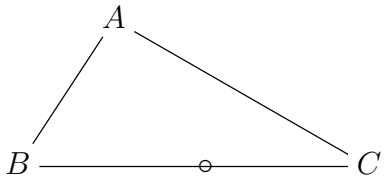
The point now is that, with this, we can substantially improve our data from Theorem 6.9, by computing the cosines of the halves of the angles there, using the above formula for $\cos(2t)$, and then computing the sines of these angles too, by using Pythagoras.

Along the same lines, we can talk as well about $\sin(3t)$ and $\cos(3t)$, again with interesting numeric applications. We will be back to such questions later, with better tools.

6c. The number pi

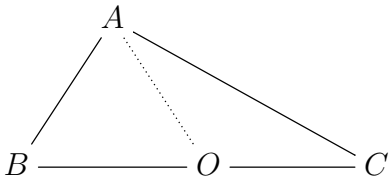
Let us get now into a more advanced study of the angles, by using circles, which are quite advanced technology. We have here the following key result, to start with:

THEOREM 6.19. *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 full 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 is as follows, with r, s being respectively the angles at B, C:

$$2r + 2s = 180°$$

Thus, we obtain $r + s = 90°$, as claimed. □

Many other things can be said, as a continuation of this, the idea being that Theorem 6.19 provides us with a number of circle methods for studying the angles.

At a more advanced level, we have many interesting plane geometry results featuring circles, due to Monge, Apollonius and others, somehow in analogy with what we know about points and lines. Again, many interesting things can be said here.

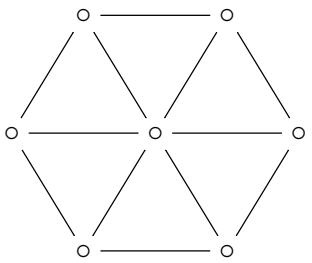
Let us get now into an even more advanced study of the angles. For this purpose, the best is to talk first about circles, more in detail, and about the number π.

But, do we really know what the number π is. So here, to start with, we have the following result, which can be regarded as being something quite axiomatic:

THEOREM 6.20. *The following two definitions of π are equivalent:*

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

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



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

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

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

In what regards now the precise value of π, the above picture at $N = 6$ shows that we have $\pi > 3$, but not by much. More can be said by using some basic trigonometry, for instance by replacing the hexagon used in the above with higher polygons:

THEOREM 6.21. *We can work out approximations of*

$$\pi = 3.14\ldots$$

by using various polygons, and basic trigonometry.

PROOF. This is indeed quite standard, as explained above. In practice, we are led in this way into estimating square roots of integers, and even square roots of reals containing square roots, which is not a simple question, either. We will be back to this, more in detail, at the end of the present chapter, with full computations for the small polygons. □

Getting now to really reliable results and data, which are actually known since long, obtained via lots of efforts, the precise figure for π is as follows:

$$\pi = 3.14159\ldots$$

We will come back to such approximation questions for π later, once we will have appropriate tools for dealing with them, coming from more advanced analysis.

It is also possible to prove that π is irrational, $\pi \notin \mathbb{Q}$, and even transcendental, but this is not trivial either. Again, we will be back to such questions later.

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

DEFINITION 6.22. *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:

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

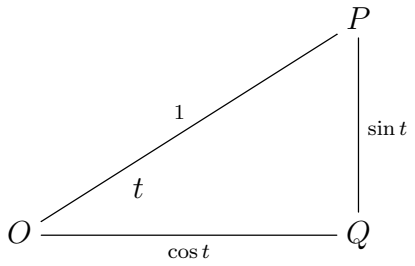
$$210^\circ = \frac{7\pi}{6} \quad , \quad 240^\circ = \frac{4\pi}{3} \quad , \quad 300^\circ = \frac{5\pi}{3} \quad , \quad 330^\circ = \frac{11\pi}{6}$$

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 in chapter 5:

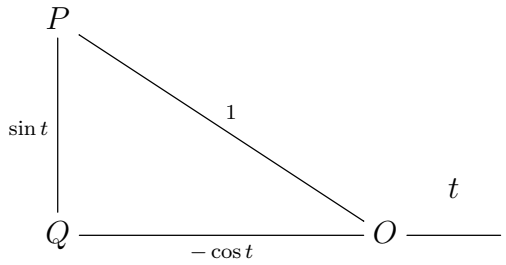
THEOREM 6.23. *The sine and cosine of an angle are obtained by putting the angle on the unit circle, as above, then projecting on the vertical and the horizontal, and then measuring the oriented segments that we get, on the vertical and horizontal.*

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

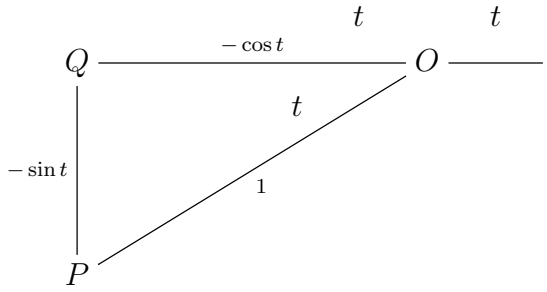
(1) In the simplest case, namely $t \in [0, \pi/2]$, the sine and cosine are indeed computed according to the following picture, which is the one in the statement:



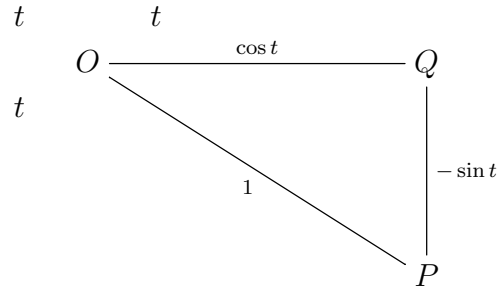
(2) In the case of obtuse angles, $t \in [\pi/2, \pi]$, the picture becomes as follows:



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



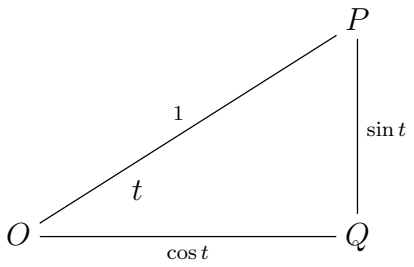
(4) As for the last case, namely $t \in [3\pi/2, 2\pi]$, here our picture is as follows:



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

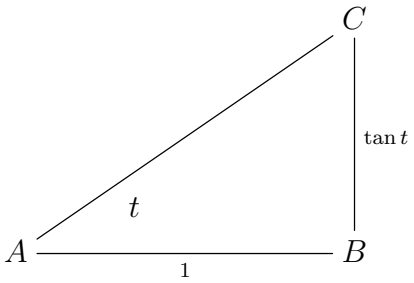
As an interesting fact, we can complement Theorem 6.23 with a statement regarding the tangent, the trigonometric function that we often forgot so far, as follows:

THEOREM 6.24. *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. □

Now that we know well about sine, cosine and tangent, time perhaps to introduce the remaining trigonometric functions too. These are as follows:

DEFINITION 6.25. *Reciprocals of* $\sin, \cos, \tan$.

We will be actually not using much these latter functions, which rather bring confusion into the math formulae, two many definitions being, generally speaking, a bad thing.

On the opposite now, here are some truly interesting functions:

DEFINITION 6.26. *Inverses of* $\sin, \cos, \tan$*, and of their reciprocals too.*

Here a bijectivity discussion is of course needed. There are actually many things can that be said about these inverses. We will be back to this.

6d. Basic estimates

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:

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

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

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 6.27. *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 6.28. *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 $\sin^2 t + \cos^2 t = 1$ 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. □

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

THEOREM 6.29. *The following happen, for small angles:*

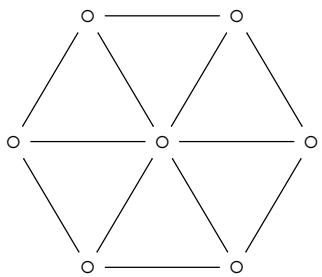
(1) $\sin t \simeq t - t^3/6$.
(2) $\cos t \simeq 1 - t^2/2 + t^4/24$.
(3) $\tan t \simeq t + t^3/3$.

PROOF. This is something which is substantially harder to prove, and with the comment that, as before with the estimates in Theorem 6.28, there are some relations between the above estimates, at various orders, due to Pythagoras, and to the formula for the tangent, in terms of the sine and cosine. Here is for instance the verification for the fact that the above formulae for the sine and cosine are compatible indeed with Pythagoras:

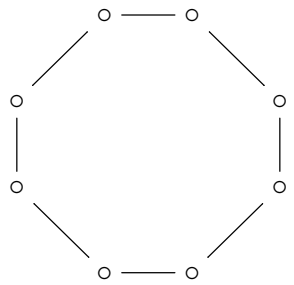
$$\begin{aligned}
\sin^2 t + \cos^2 t &\simeq \left(t - \frac{t^3}{6}\right)^2 + \left(1 - \frac{t^2}{2} + \frac{t^4}{24}\right)^2\\
&= \left(t^2 - \frac{t^4}{3} + \frac{t^6}{36}\right) + \left(1 + \frac{t^4}{4} + \frac{t^8}{576} - t^2 + \frac{t^4}{12} - \frac{t^6}{24}\right)\\
&\simeq \left(t^2 - \frac{t^4}{3} + \frac{t^6}{36}\right) + \left(1 + \frac{t^4}{4} - t^2 + \frac{t^4}{12} - \frac{t^6}{24}\right)\\
&= \left(t^2 - \frac{t^4}{3} + \frac{t^6}{36}\right) + \left(1 - t^2 + \frac{t^4}{3} - \frac{t^6}{24}\right)\\
&\simeq \left(t^2 - \frac{t^4}{3}\right) + \left(1 - t^2 + \frac{t^4}{3}\right)\\
&= 1
\end{aligned}$$

Quite wild all this, hope you agree with me. We will be back to such questions later in this book, once we will have better tools for dealing with them. □

Finally, still talking analysis, we have a lot of interesting estimates, of varying levels of difficuty, regarding π itself. As already mentioned in the beginning of this chapter, we can see right away that we have $\pi > 3$, and not by much, by using a hexagon:



Leaving the heptagon aside, next we have the octagon, which is as follows:



And here, with some trigonometry help from Conor and Khabib, and with GSP helping with the square roots, we can have some approximations for π going. To be more precise, we can use here the following formulae, established before:

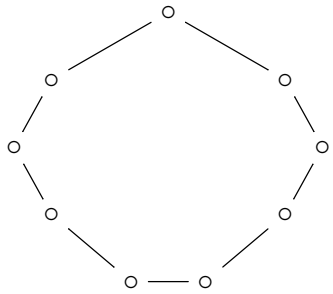
$$\sin(2t) = 2\sin t\cos t$$

$$\begin{aligned}\cos(2t) &= \cos^2 t - \sin^2 t\\ &= 2\cos^2 t - 1\\ &= 1 - 2\sin^2 t\end{aligned}$$

$$\tan(2t) = \frac{2\tan t}{1-\tan^2 t}$$

Thus, we can compute the octagon edge, and then approximate π.

Next, we can have some computations for the nonagon, which looks as follows:



And so on, with increasingly more complex computations, which are all interesting, hiding all sorts of mysterious mathematics, and with all this being quite addictive.

We will be back to such things later in this book, with some better methods for approximating π, when systematically doing analysis.

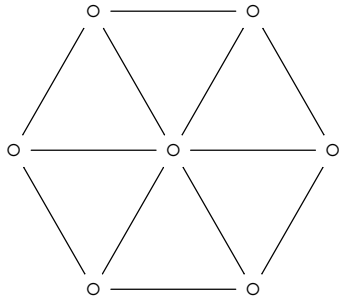
Finally, getting back to π, in analogy with our previous theory of e, we have the following well-known, and beautiful, probabilistic interpretation of π, due to Buffon:

THEOREM 6.30. *The probability for a needle of length* 1*, when trown on a grid of parallel* 1*-spaced lines, to intersect one line, is* $P = 2/\pi$*.*

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

As a last topic for this chapter, let us further discuss polar geometry, as a continuation of the material from chapter 5. We first have the following theorem:

THEOREM 6.31 (Pascal). *Given a hexagon lying on a circle*

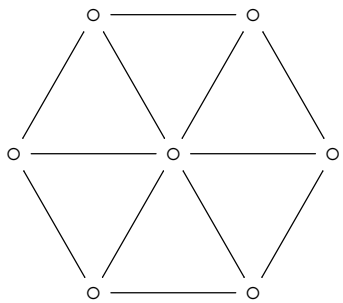


the pairs of opposite sides intersect in points which are collinear.

PROOF. This can be proved indeed, with some tricks. Observe the similarity with Pappus. We will see in fact, later in this book, when talking about conics, that the Pascal theorem generalizes to the case of conics, and with this fully generalizing Pappus. □

And here is now, at a truly advanced level, a quite scary theorem:

THEOREM 6.32 (Brianchon). *Given a hexagon circumscribed around on a circle*



the main diagonals intersect.

PROOF. This is nearly impossible to prove, with bare hands, but follows by duality from Pascal. As before with Pascal, we will see later that this extends to conics. □

Quite interesting all these results about hexagons, and the relation between them. We will be back to such things, with better tools for dealing with them, later in this book.

6e. Exercises

Exercises:

EXERCISE 6.33.

EXERCISE 6.34.

EXERCISE 6.35.

EXERCISE 6.36.

EXERCISE 6.37.

EXERCISE 6.38.

EXERCISE 6.39.

EXERCISE 6.40.

Bonus exercise.

CHAPTER 7

Coordinates

7a. Real plane

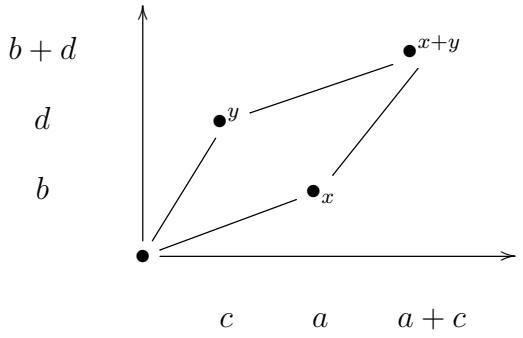
Still doing geometry, but at a more advanced level, we can talk about the real plane, in the obvious way, with each point $x \in \mathbb{R}^2$ being written as a vector, as follows:

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

Of particular interest is the summing operation for such vectors, which, according to the usual calculus rules for the vectors, is 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, the idea behind this is something very simple, namely that the vectors add by forming a parallelogram, according to the following picture:



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 formula, allowing us to compute the length of any vector:

$$x = \begin{pmatrix} a \\ b \end{pmatrix} \implies ||x|| = \sqrt{a^2 + b^2}$$

So far, so good, and time now to see how our coordinate technology works, if that is worth something, or not. We will review here all the triangle and basic geometry material from before, with new proofs for everything, using coordinates, no less than that.

So, God bless, and let us get started. 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 this book with, the following theorem, coming along with a trivial proof:

THEOREM 7.1. *Any two distinct points $P \neq Q$ determine a line, denoted PQ.*

PROOF. This is clear indeed, with coordinates, because we have:

$$PQ = \lambda P + (1 - \lambda)Q$$

So, very good news, axiom becoming theorem, what more can we wish for. □

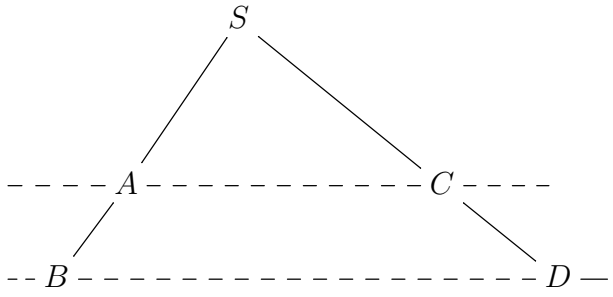
Same situation for the second axiom, which becomes a theorem too:

THEOREM 7.2. *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. This is again clear with coordinates. □

Getting now to the next thing that we did in chapter 5, namely the Thales theorem, and as further good news, that drastically simplifies with coordinates, as follows:

THEOREM 7.3 (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:

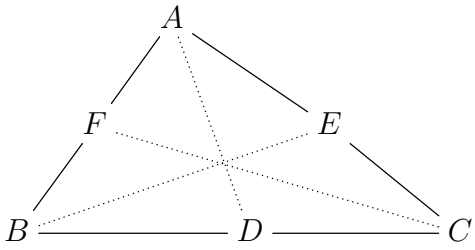
$$\frac{SA}{SB} = \frac{SC}{SD}$$

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. Again, this is clear with coordinates, and in fact the other formulations of the Thales theorem, also from chapter 5, are clear as well too, again with coordinates. □

Getting now to the barycenter theorem, this drastically simplifies, as follows:

THEOREM 7.4 (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. □

We can prove now as well some things claimed in chapter 5, as follows:

THEOREM 7.5. *The gravity center of a triangle ABC is as follows:*

(1) *In the 0-dimensional case, that is, when putting equal weigths at the vertices A, B, C, and computing the center, this is the barycenter.*
(2) *In the 1-dimensional case, that is, with the sides AB, BC, AC have weigths proportional with their length, this is, in general, different from the barycenter.*
(3) *In the 2-dimensional case, that is, with the triangle ABC itself, as an area, having a weight, uniformly distributed, this is again the barycenter.*

PROOF. Again, this is clear with coordinates. □

Getting now to the other centers of a triangle, we have here:

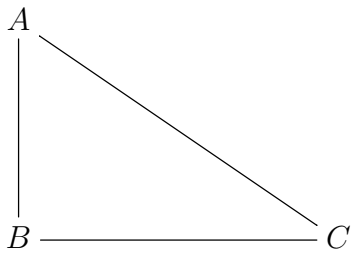
THEOREM 7.6. *Given a triangle ABC, the following happen:*

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

PROOF. Again, such things can be proved with coordinates, and patience. We will actually leave some of the calculations here as an instructive exercise for you, reader. □

Coming next, we have the theorem of Pythagoras:

THEOREM 7.7 (Pythagoras). *In a right triangle ABC,*

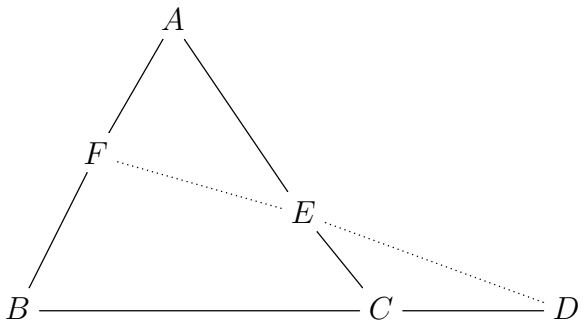


we have $AB^2 + BC^2 = AC^2$.

PROOF. Again, this is clear with coordinates. □

Next, we have the following key result, due to Menelaus:

THEOREM 7.8 (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:

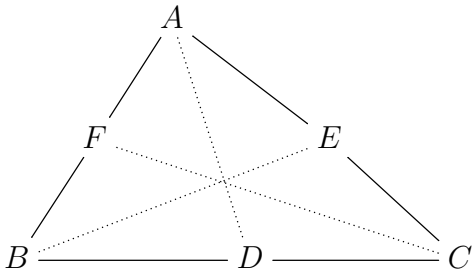
$$\frac{AF}{FB} \cdot \frac{BD}{DC} \cdot \frac{CE}{EA} = -1$$

Moreover, the converse holds, with this formula guaranteeing that F, E, D are colinear.

PROOF. Again, this is clear with coordinates. □

Next, we have the following remarkable result, due to Ceva:

THEOREM 7.9 (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:

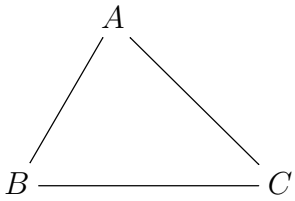
$$\frac{AF}{FB} \cdot \frac{BD}{DC} \cdot \frac{CE}{EA} = 1$$

Moreover, the converse holds, with this formula guaranteeing that AD, BE, CF cross.

PROOF. Again, this is clear with coordinates. □

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

THEOREM 7.10. *Besides the 4 main centers of a triangle, discussed in the above, many more remarkable points can be associated to a triangle ABC,*

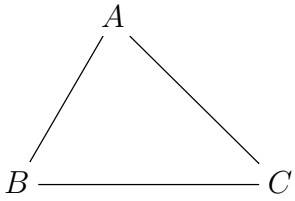


and most of these lie on a line, called Euler line of ABC.

PROOF. Proving this with coordinates is a good exercise for you, reader. □

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

THEOREM 7.11. *Associated to a triangle ABC,*

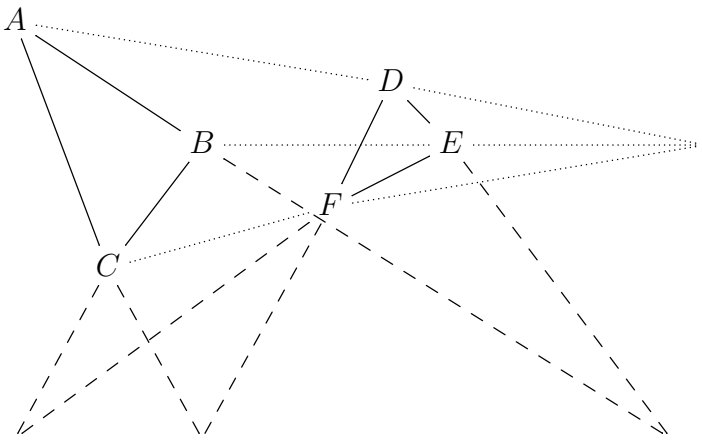


we have as well a nine-point circle, whose center lies on the Euler line.

PROOF. Again, proving this with coordinates is a good exercise for you, reader. □

Moving ahead, we have the Desargues theorem:

THEOREM 7.12 (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.

PROOF. Again, this is clear with coordinates. □

We have as well the following result, that we met before in relation with our previous proof of Desargues, and which is something having its own interest:

THEOREM 7.13. *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,*

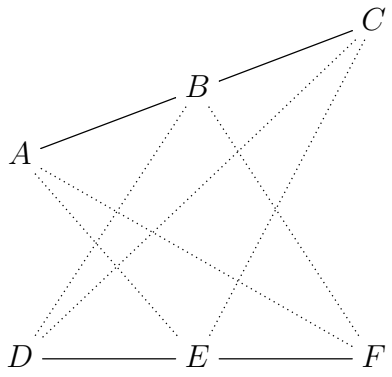
(1) *Given a point P, construct Q on the line OP, as to have $OP \cdot OQ = r^2$,*
(2) *Draw the perpendicular at Q on the line OQ. This is the dual line p,*

and this duality $P \leftrightarrow p$ transforms collinear points into concurrent lines.

PROOF. Again, this is clear with coordinates. □

Next, we have the Pappus theorem:

THEOREM 7.14 (Pappus). *Given a configuration as follows,*



the three middle points are collinear.

PROOF. Again, this is clear with coordinates. □

We have as well the following result, that we met before in relation with our previous proof of Pappus, and which is something having its own interest:

THEOREM 7.15. *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. Again, this is clear with coordinates. □

In relation with the projective geometry considerations from the end of chapter 5, coordinates can halp in that setting too, and we have the following result:

THEOREM 7.16. *Projective coordinates.*

PROOF. This is something that can be done too, and many interesting things can be said here. We will be back to this on several occasions, in what follows. □

As a conclusion to all this, coordinates seem to perfom quite well, and you might probably have this question right now, why not having started the present book with coordinates. In answer, modesty and patience, this is how math is best learned. We will actually see in what follows that our present $\mathbb{R}^2$ coordinates can be beaten themselves by some better coordinates, namely the $\mathbb{C}$ ones. So, long story still to go, and ho hurry.

Our idea in what follows will be that of improving part of the vector technology, by using polar coordinates. The idea here is very simple, as follows:

THEOREM 7.17. *The points of the plane $x \in \mathbb{R}^2$, written as vectors*

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

can be written in polar coordinates, as follows,

$$x = \begin{pmatrix} r\cos t \\ r\sin t \end{pmatrix}$$

with the connecting formulae being as follows,

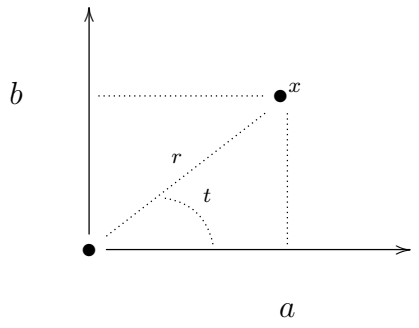
$$a = r\cos t \quad , \quad b = r\sin t$$

and in the other sense being as follows,

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

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

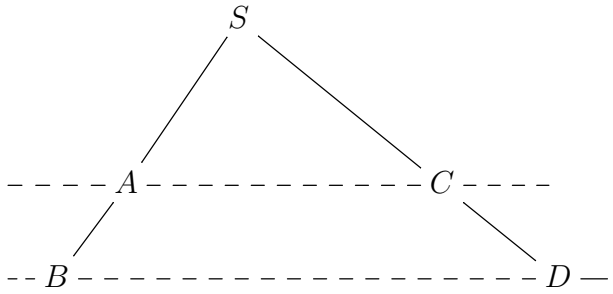
PROOF. This is something self-explanatory and intuitive, with $r = \sqrt{a^2 + b^2}$ being as usual the length of the vector, and with t being the angle made by the vector with the Ox axis. That is, with the picture for what is going on in the above being as follows:



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

Many interesting things can be done with polar coordinates. As a matter of getting familiar with this technology, let us reprove some of our previous results. We first have:

THEOREM 7.18 (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:

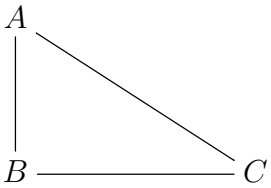
$$\frac{SA}{SB} = \frac{SC}{SD}$$

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. This is clear indeed with polar coordinates. □

In relation with triangle geometry, we first have:

THEOREM 7.19 (Pythagoras). *In a right triangle ABC,*



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

PROOF. This is clear indeed with polar coordinates. □

At a more advanced level now, we have:

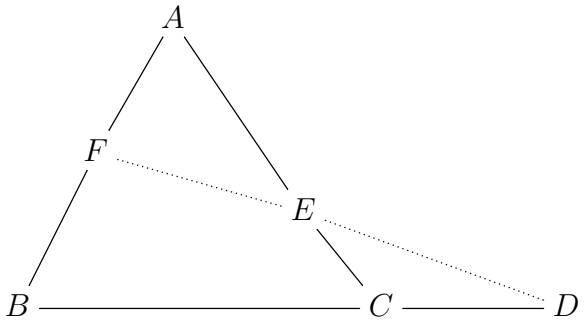
THEOREM 7.20. *Given a triangle ABC, the following happen:*

(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. Again, this is clear with polar coordinates. □

Next, let us have a look at the Menelaus theorem:

THEOREM 7.21 (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:

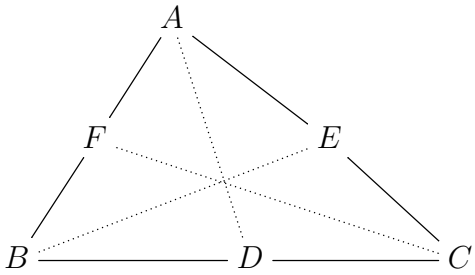
$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=-1$$

Moreover, the converse holds, with this formula guaranteeing that F, E, D are colinear.

PROOF. Again, this is clear with polar coordinates. □

Next, we have the following remarkable result, due to Ceva:

THEOREM 7.22 (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:

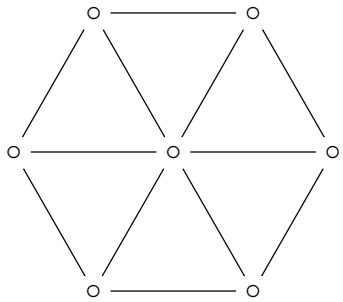
$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=1$$

Moreover, the converse holds, with this formula guaranteeing that AD, BE, CF cross.

PROOF. Again, this is clear with polar coordinates. □

At the level of more advanced results, we first have:

THEOREM 7.23 (Pascal). *Given a hexagon lying on a circle*

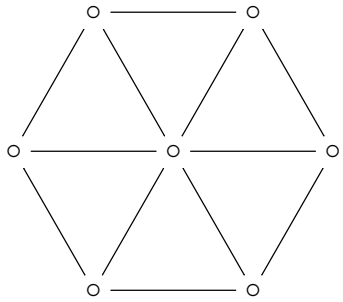


the pairs of opposite sides intersect in points which are collinear.

PROOF. This can be proved indeed, with some tricks. Observe the similarity with Pappus. We will see in fact, later in this book, when talking about conics, that the Pascal theorem generalizes to the case of conics, and with this fully generalizing Pappus. □

And here is now, at a truly advanced level, a quite scary theorem:

THEOREM 7.24 (Brianchon). *Given a hexagon circumscribed around on a circle*



the main diagonals intersect.

PROOF. This is nearly impossible to prove, with bare hands, but follows by duality from Pascal. As before with Pascal, we will see later that this extends to conics. □

7b. Complex plane

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

DEFINITION 7.25. *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 7.26. *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 7.27. *The complex numbers, written as usual*

$$x = a + ib$$

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

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

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

PROOF. Consider indeed two arbitrary complex numbers:

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

Their sum is then by definition the following complex number:

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

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

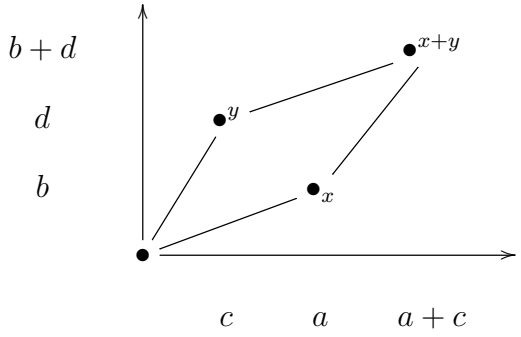
$$x = \begin{pmatrix} a \\ b \end{pmatrix} \quad , \quad y = \begin{pmatrix} c \\ d \end{pmatrix}$$

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

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

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

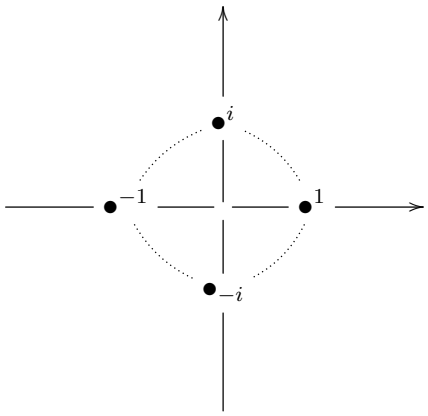
Here we have assumed that you are a bit familiar with vector calculus. If not, no problem, the idea is simply that vectors add by forming a parallelogram, as follows:



Observe that in our geometric picture from Proposition 7.27, the real numbers correspond to the numbers on the Ox axis. As for the purely imaginary numbers, these lie on the Oy axis, with the number i itself being given by the following formula:

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

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



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

ADVICE 7.28. *When drawing complex numbers, always begin with the coordinate axes Ox, Oy, and with a copy of the unit circle.*

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

DEFINITION 7.29. *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,

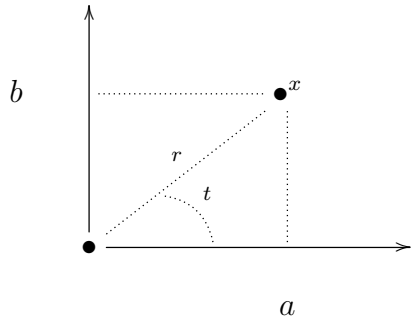
$$a = r \cos t \quad , \quad b = r \sin t$$

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 7.27, 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 7.29 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 7.30. *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 7.31. *The powers of a complex number, written in polar form,*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

As a basic application of Theorem 7.31, we have the following result:

PROPOSITION 7.32. *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 7.31, at $k = 1/2$. As for its particular cases with $x \in \mathbb{R}$, these are clear from it. □

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

With the above results in hand, and notably with the square root formula from Proposition 7.32, we can now go back to the degree 2 equations, and we have:

THEOREM 7.33. *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 7.32 that any complex number has a square root, we are led to the conclusion in the statement. □

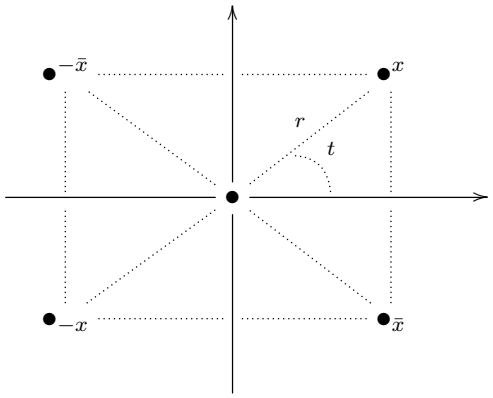
As a last general topic regarding the complex numbers, let us discuss conjugation. This is something quite tricky, complex number specific, as follows:

DEFINITION 7.34. *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 7.35. *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 7.34.

(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 7.26, 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:

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

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

7c. Analysis tricks

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

THEOREM 7.36. *We have the following formula,*

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

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

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

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

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

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

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

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

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

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

THEOREM 7.37. *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 7.36, by setting $t = \pi$. Indeed, we obtain:

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

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

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

THEOREM 7.38. *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 7.26, by using the formula $e^{it} = \cos t + i\sin t$ from Theorem 7.37, 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 7.39. *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 7.30, reformulated by using the notations from Theorem 7.38. Observe that this follows as well directly, from the fact that we have $e^{a+b} = e^a e^b$, that we know from analysis. □

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

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

THEOREM 7.40. *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 7.31, 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 7.38:

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

(2) Once again by using Theorem 7.38, 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 7.38, this operation $x\to x^a$ is indeed the correct one.

(4) This comes from the fact, that we know from Theorem 7.35, 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 7.33 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 7.41. *Any polynomial $P\in\mathbb{C}[X]$ decomposes as*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

which are called roots of unity of order N.

PROOF. This follows from the general multiplication formula for complex numbers from Theorem 7.39. Indeed, with $x = re^{it}$ our equation reads:

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

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

As an illustration here, the roots of unity of small order are as follows:

$\underline{N = 1}$. Here the unique root of unity is 1.

$\underline{N = 2}$. Here we have two roots of unity, namely 1 and -1.

$\underline{N = 3}$. Here we have 1, then $w = e^{2\pi i/3}$, and then $w^2 = \bar{w} = e^{4\pi i/3}$.

$\underline{N = 4}$. Here the roots of unity, read as usual counterclockwise, are $1, i, -1, -i$.

$\underline{N = 5}$. Here, with $w = e^{2\pi i/5}$, the roots of unity are $1, w, w^2, w^3, w^4$.

$\underline{N = 6}$. Here a useful alternative writing is $\{\pm 1, \pm w, \pm w^2\}$, with $w = e^{2\pi i/3}$.

The roots of unity are very useful variables, and have many interesting properties. As a first application, we can now solve the ambiguity questions related to the extraction of N-th roots, from Theorem 7.31 and Theorem 7.30, the statement being as follows:

THEOREM 7.43. *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. □

7e. Exercises

Exercises:

EXERCISE 7.44.

EXERCISE 7.45.

EXERCISE 7.46.

EXERCISE 7.47.

EXERCISE 7.48.

EXERCISE 7.49.

EXERCISE 7.50.

EXERCISE 7.51.

Bonus exercise.

CHAPTER 8

Plane curves

8a. Ellipses

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

THEOREM 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:*

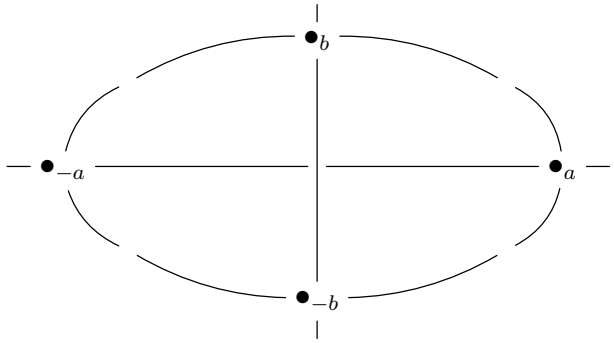
$$\bigcirc \quad \to \quad ?$$

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

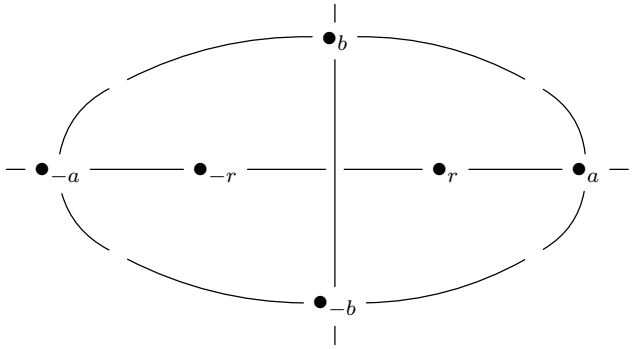
PROOF. This might look a bit confusing, and you might say, what exactly is to be proved here. Good point, and in answer, what is to be proved is that the above constructions (1-4) give rise to the same class of curves. And this can be done as follows:

(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 ellipsis:



(2) Let us prove now that such an ellipsis has two focal points, as stated in (2). We must look for a number $r > 0$, and a number $l > 0$, such that our ellipsis appears as $d(z, p) + d(z, q) = l$, with $p = (0, -r)$ and $q = (0, r)$, according to the following picture:



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

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

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

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

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

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

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

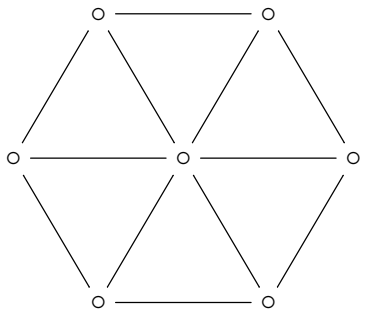
$$\begin{aligned}
(4x^2 - l^2)(4r^2 - l^2) = 4l^2y^2 \quad & \iff \quad \frac{4x^2 - l^2}{l^2} = \frac{4y^2}{4r^2 - l^2} \\
& \iff \quad \frac{4x^2 - l^2}{l^2} = \frac{y^2}{r^2 - l^2/4} \\
& \iff \quad \left(\frac{x}{2l}\right)^2 - 1 = \left(\frac{y}{\sqrt{r^2 - l^2/4}}\right)^2 \\
& \iff \quad \left(\frac{x}{2l}\right)^2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

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

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

Many other things can be said about the ellipses, as a continuation of the above. As a first result, coming as a continuation of what we knew before for circles, we have:

THEOREM 8.2 (Pascal). *Given a hexagon lying on an ellipsis*

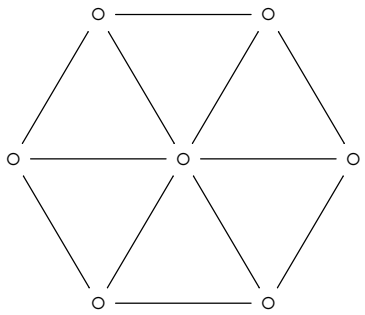


the pairs of opposite sides intersect in points which are collinear.

PROOF. This can be proved indeed, with some tricks. Observe the similarity with Pappus. We will see in fact later, when talking about conics, that the Pascal theorem generalizes to the case of conics, and with this fully generalizing Pappus. □

And here is now, at a truly advanced level, a quite scary theorem:

THEOREM 8.3 (Brianchon). *Given a hexagon circumscribed around on an ellipsis*



the main diagonals intersect.

PROOF. This is nearly impossible to prove, with bare hands, but follows by duality from Pascal. As before with Pascal, we will see later that this extends to conics. □

There are several other results of this type. We will be back to this.

8b. The conics

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$dx + ey + f = 0$$

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

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

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

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

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

$$ax + by + cz = d$$

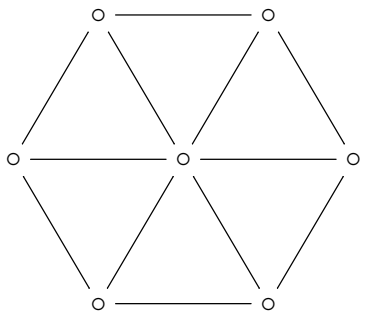
(7) However, before getting into computations, observe that what we want to find is a certain degree 2 equation in the above plane, for the intersection. Thus, it is convenient to change the coordinates, as for our plane to be given by the following equation:

$$z = 0$$

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

Many other things can be said about the ellipses, as a continuation of the above. As a first result, coming as a continuation of what we knew before for ellipses, we have:

THEOREM 8.5 (Pascal). *Given a hexagon lying on a conic*

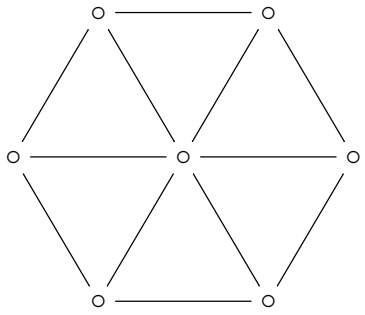


the pairs of opposite sides intersect in points which are collinear.

PROOF. This can be proved indeed, with some tricks. Observe the similarity with Pappus. In fact, the Pascal theorem for conics fully generalizes Pappus. □

And here is now, at a truly advanced level, a quite scary theorem:

THEOREM 8.6 (Brianchon). *Given a hexagon circumscribed around on a conic*



the main diagonals intersect.

PROOF. This is nearly impossible to prove, with bare hands, but follows by duality from Pascal. As before with Pascal, many other things can be said here. □

There are several other results of this type. We will be back to this.

Ready for some physics? We have the following result:

THEOREM 8.7. *Planets and other celestial bodies move around the Sun on conics,*

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

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

PROOF. This is something very standard, 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:

$$\begin{aligned} b = 0 &\iff 2\dot{r}\dot{\theta} + r\ddot{\theta} = 0 \\ &\iff \frac{\ddot{\theta}}{\dot{\theta}} = -2\frac{\dot{r}}{r} \\ &\iff (\log\dot{\theta})' = (-2\log r)' \\ &\iff \log\dot{\theta} = -2\log r + c \\ &\iff \dot{\theta} = \frac{\lambda}{r^2} \end{aligned}$$

As for the first equation the we found, namely $a + K/r^2 = 0$, 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. □

Still with me, I hope, after all these computations. Good work that we did.

The above was theory, and for further applications, here is a sort of "best of" the formulae found in the proof of Theorem 8.7, which are all very useful in practice:

THEOREM 8.8 (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 8.7. 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 8.7. □

As a first 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 8.9. *In the context of Theorem 8.8, 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 8.8, 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 8.8, 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.

8c. 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 8.10. *An algebraic curve in $\mathbb{R}^2$ is the vanishing set*

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

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 8.11. *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 8.12. *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 8.13. *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 8.14. *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 8.15. *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 8.16. *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 8.17. *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 8.18. *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:

$$\begin{aligned}
r^3 = a^3 \cos 3\theta &\iff r^3 \cos 3\theta = \left(\frac{r^2}{a}\right)^3 \\
&\iff z^3 + \bar{z}^3 = 2\left(\frac{z\bar{z}}{a}\right)^3 \\
&\iff (x+iy)^3 + (x-iy)^3 = 2\left(\frac{x^2+y^2}{a}\right)^3 \\
&\iff x^3 - 3xy^2 = \left(\frac{x^2+y^2}{a}\right)^3 \\
&\iff (x^2+y^2)^3 = a^3(x^3 - 3xy^2)
\end{aligned}$$

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 8.19. *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:

$$\begin{aligned}
r = a\cos^3\left(\frac{\theta}{3}\right) &\iff ar\cos^3\left(\frac{\theta}{3}\right) = r^2 \\
&\iff a\left(\frac{u+\bar{u}}{2}\right)^3 = r^2 \\
&\iff a(u^3 + \bar{u}^3 + 3u\bar{u}(u+\bar{u})) = 8r^2 \\
&\iff 3au\bar{u} \cdot \frac{u+\bar{u}}{2} = 4r^2 - ax \\
&\iff 27a^3r^6 \cdot \frac{r^2}{a} = (4r^2 - ax)^3 \\
&\iff 27a^2(x^2+y^2)^2 = (4x^2 + 4y^2 - ax)^3
\end{aligned}$$

Thus, we have indeed a sextic, as claimed. □

8d. 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 8.20. *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:

$$\begin{aligned} r^n = a^n \cos n\theta &\iff r^n \cos n\theta = \left(\frac{r^2}{a}\right)^n \\ &\iff z^n + \bar{z}^n = 2\left(\frac{z\bar{z}}{a}\right)^n \\ &\iff (x+iy)^n + (x-iy)^n = 2\left(\frac{x^2+y^2}{a}\right)^n \end{aligned}$$

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 8.21. *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 $(z^{1/m}+\bar{z}^{1/m})^m=2^m a$.*
(4) *With $n=1/m$, $m\in\mathbb{N}$, the equation is $(z^{1/m}+\bar{z}^{1/m})^m=2^m z\bar{z}/a$.*

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:

$$\begin{aligned}z^{-1/m}+\bar{z}^{-1/m}=2\left(\frac{z\bar{z}}{a}\right)^{-1/m}&\iff z^{1/m}+\bar{z}^{1/m}=2a^{1/m}\\&\iff (z^{1/m}+\bar{z}^{1/m})^m=2^m a\end{aligned}$$

(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 8.22. *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 8.23. *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 8.24. *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.

8e. Exercises

Exercises:

Exercise 8.25.

Exercise 8.26.

Exercise 8.27.

Exercise 8.28.

Exercise 8.29.

Exercise 8.30.

Exercise 8.31.

Exercise 8.32.

Bonus exercise.

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

We have seen that many number theory questions lead us into computing roots of polynomials $P \in \mathbb{Q}[X]$. We will investigate here such questions, with a detailed study of the arbitrary polynomials $P \in \mathbb{C}[X]$, and their roots, often by using analytic methods.

Let us start with something that we know well, but is always good to remember:

THEOREM 9.1. *The solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{C}$ are*

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

with the square root of complex numbers being defined as $\sqrt{re^{it}} = \sqrt{r}e^{it/2}$.

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

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

Here we have used the fact, mentioned in the statement, and that we know well from chapter 7, that any complex number $z = re^{it}$ has indeed a square root, given by:

$$\sqrt{z} = \sqrt{r}e^{it/2}$$

In fact, as we know from chapter 7 too, in the case $z \neq 0$ we have as well a second square root, namely $-\sqrt{z}$. Thus, we are led to the conclusion in the statement. □

Very nice all this, and you would probably say that the story is over here, with degree 2 equations. However, not really, and this for a number of reasons:

(1) First, it takes some skill in order to quickly draw parabolas, $y = ax^2 + bx + c$, and I can only recommend here doing a lot of exercises, of this type.

(2) Still talking parabolas, these appear in physics, in relation with gravity, and many interesting things can be said here. More on this in Part IV below.

(3) Getting now to Theorem 9.1 as stated, when a, b, c are all of the form $p + q\sqrt{r}$ with $p, q, r \in \mathbb{Z}$, some arithmetics comes into play, for computing $\sqrt{b^2 - 4ac}$.

(4) And there is no telling of what happens when a, b, c are more complicated numbers, say involving iterated square roots, with all this getting us again into arithmetic.

In addition to this, Theorem 9.1 as stated is quite often not the ideal way of viewing things. Here are indeed some useful tricks, in order to deal with degree 2 questions:

TRICKS 9.2. *The following happen:*

(1) *The roots of $x^2 - ax + b$ can be computed by using the following formulae:*

$$r + s = a \quad , \quad rs = b$$

(2) *Also, the eigenvalues of $A \in M_2(\mathbb{C})$ can be computed by using the formulae*

$$r + s = Tr(A) \quad , \quad rs = \det A$$

with $Tr(A)$ being the sum of the diagonal entries, called trace of the matrix.

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

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

In order to find the eigenvalues r, s, you are certainly very used to compute the characteristic polynomial, then apply Theorem 9.1. But my point is that this characteristic polynomial is of the form $x^2 - ax + b$, with $a = Tr(a)$ and $b = \det A$, so we can normally apply the trick in (1), provided of course that r, s are simple numbers, say integers.

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

WARNING 9.3. *The above tricks work in pure mathematics, where the numbers r, s that we can meet are usually integers, or rationals. In applied mathematics, however, the numbers that we meet are integers or rationals with probability $P = 0$, so no tricks.*

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

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

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

THEOREM 9.4. *Given a monic polynomial $P \in \mathbb{C}[X]$, factorized as*

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

the following happen:

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

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

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

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

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

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

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

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

$$S_r = \sum_i a_i^r$$

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

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

(3) The point now is that, alternatively, the symmetric functions in our variables $a_1, \ldots, a_k$ appear as well as linear combinations of the functions f_r that we found in (1),

and that when allowing polynomials instead of linear combinations, we need in fact only the first k functions, namely $f_1, \ldots, f_k$. That is, we have as well:

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

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

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

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

THEOREM 9.5. *Given two polynomials $P, Q \in \mathbb{C}[X]$, written as*

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

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

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

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

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

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

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

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

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

(4) Getting started now, let us expand the formula of $R(P, Q)$, by making all the multiplications there, abstractly, in our head. Everything being symmetric in $a_1, \ldots, a_k$, we obtain in this way certain symmetric functions in these variables, which will be therefore

certain polynomials in the coefficients of P. Moreover, due to our normalization factor c^l, these polynomials in the coefficients of P will have integer coefficients.

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

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

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

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

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

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

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

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

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

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

Thus, we have Theorem 9.5 fully verified and illustrated in the case of a polynomial of degree 2, and a polynomial of degree 1. More such computations in a moment.

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

We have here the following result, which is something more advanced:

THEOREM 9.6. *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 $< k$:

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

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

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

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

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

(4) Finally, our claim is that we have $\det \Phi = R(P,Q)$. But this follows from the uniqueness of the resultant, up to a scalar, and with this uniqueness property being elementary to establish, along the lines of the proofs of Theorems 9.4 and 9.5. □

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

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

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

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

We will compute many other resultants, in what follows.

9b. The discriminant

We can go back now to our original question, from the beginning of this chapter, that of finding the analogue of the discriminant $\Delta = b^2 - 4ac$ for higher degree polynomials. We have here the following result, providing a complete answer to this question:

THEOREM 9.7. *Given a polynomial $P \in \mathbb{C}[X]$, written as*

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

its discriminant, defined as being the following quantity,

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

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

PROOF. The fact that the discriminant $\Delta(P)$ is a polynomial in the coefficients of P, with integer coefficients, comes from Theorem 9.5, coupled with the fact that the division by the leading coefficient a is indeed possible, under $\mathbb{Z}$, as being shown by the following formula, which is of course a bit informal, coming from Theorem 9.6:

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

Also, the fact that we have $\Delta(P) = 0$ precisely when P has a double root is clear from Theorem 9.5. Finally, let us mention that the sign $(-1)^{\binom{N}{2}}$ is there for various reasons, including the compatibility with some well-known formulae, at small values of $N \in \mathbb{N}$, such as $\Delta(P) = b^2 - 4ac$ in degree 2, that we will discuss in a moment. □

As already mentioned, by using Theorem 9.6, we have an explicit formula for the discriminant, as the determinant of a certain matrix. There is a lot of theory here, and in order to get into this, let us first see what happens in degree 2. Here we have:

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

Thus, the resultant is given by the following formula:

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

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

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

Alternatively, we can use the formula in Theorem 9.6, and we obtain:

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

We will be back later to such formulae, in degree 3, and in degree 4 as well, with the comment however, coming in advance, that these formulae are not very beautiful.

At the theoretical level now, we have the following result, which is not trivial:

THEOREM 9.8. *The discriminant of a polynomial P is given by the formula*

$$\Delta(P) = a^{2N-2}\prod_{i<j}(r_i - r_j)^2$$

where a is the leading coefficient, and $r_1, \ldots, r_N$ are the roots.

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

(1) The first thought goes to the formula in Theorem 9.5, so let us see what that formula teaches us, in the case $Q = P'$. Let us write P, P' as follows:

$$P = a(x - r_1)\ldots(x - r_N)$$

$$P' = Na(x - p_1)\ldots(x - p_{N-1})$$

According to Theorem 9.5, the resultant of P, P' is then given by:

$$R(P, P') = a^{N-1}(Na)^N\prod_{ij}(r_i - p_j)$$

And bad news, this is not exactly what we wished for, namely the formula in the statement. That is, we are on the good way, but certainly have to work some more.

(2) Obviously, we must get rid of the roots $p_1, \ldots, p_{N-1}$ of the polynomial P'. In order to do this, let us rewrite the formula that we found in (1) in the following way:

$$\begin{aligned} R(P,P') &= N^N a^{2N-1} \prod_i \left(\prod_j (r_i - p_j) \right) \\ &= N^N a^{2N-1} \prod_i \frac{P'(r_i)}{Na} \\ &= a^{N-1} \prod_i P'(r_i) \end{aligned}$$

(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 9.7, and then using the formula found in (3), we obtain:

$$\begin{aligned} \Delta(P) &= (-1)^{\binom{N}{2}} a^{N-2} \prod_i P'(r_i) \\ &= (-1)^{\binom{N}{2}} a^{2N-2} \prod_{i \neq j} (r_i - r_j) \end{aligned}$$

(5) This is already a nice formula, which is very useful in practice, and that we can safely keep as a conclusion, to our computations. However, we can do slightly better, by

grouping opposite terms. Indeed, this gives the following formula:

$$\begin{aligned}\Delta(P) &= (-1)^{\binom{N}{2}}a^{2N-2}\prod_{i\neq j}(r_i-r_j)\\ &= (-1)^{\binom{N}{2}}a^{2N-2}\prod_{i<j}(r_i-r_j)\cdot\prod_{i>j}(r_i-r_j)\\ &= (-1)^{\binom{N}{2}}a^{2N-2}\prod_{i<j}(r_i-r_j)\cdot(-1)^{\binom{N}{2}}\prod_{i<j}(r_i-r_j)\\ &= a^{2N-2}\prod_{i<j}(r_i-r_j)^2\end{aligned}$$

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

As applications now, the formula in Theorem 9.8 is quite useful for the real polynomials $P\in\mathbb{R}[X]$ in small degree, because it allows to say when the roots are real, or complex, or at least have some partial information about this. For instance, we have:

PROPOSITION 9.9. *Consider a polynomial with real coefficients, $P\in\mathbb{R}[X]$, assumed for simplicity to have nonzero discriminant, $\Delta\neq 0$.*

(1) *In degree 2, the roots are real when $\Delta>0$, and complex when $\Delta<0$.*
(2) *In degree 3, all roots are real precisely when $\Delta>0$.*

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

(1) The first assertion is something that you certainly know, coming from Theorem 9.1, but let us see how this comes via the formula in Theorem 9.8, namely:

$$\Delta(P)=a^{2N-2}\prod_{i<j}(r_i-r_j)^2$$

In degree $N=2$, this formula looks as follows, with r_1,r_2 being the roots:

$$\Delta(P)=a^2(r_1-r_2)^2$$

Thus $\Delta>0$ amounts in saying that we have $(r_1-r_2)^2>0$. Now since r_1,r_2 are conjugate, and with this being something trivial, meaning no need here for the computations in Theorem 9.1, we conclude that $\Delta>0$ means that r_1,r_2 are real, as stated.

(2) In degree $N=3$ now, we know from analysis that P has at least one real root, and the problem is whether the remaining 2 roots are real, or complex conjugate. For this purpose, we can use the formula in Theorem 9.8, which in degree 3 reads:

$$\Delta(P)=a^4(r_1-r_2)^2(r_1-r_3)^2(r_2-r_3)^2$$

We can see that in the case $r_1,r_2,r_3\in\mathbb{R}$, we have $\Delta(P)>0$. Conversely now, assume that $r_1=r$ is the real root, coming from analysis, and that the other roots are $r_2=z$

and $r_3 = \bar{z}$, with z being a complex number, which is not real. We have then:

$$\begin{aligned}\Delta(P) &= a^4(r-z)^2(r-\bar{z})^2(z-\bar{z})^2\\ &= a^4|r-z|^4(2iIm(z))^2\\ &= -4a^4|r-z|^4Im(z)^2\\ &< 0\end{aligned}$$

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

In relation with the above, for our result to be truly useful, we must of course compute the discriminant in degree 3. We will do this, along with applications, right next.

9c. Cardano formula

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

THEOREM 9.10. *The discriminant of a degree* 3 *polynomial,*

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

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

PROOF. We have two methods available, based on Theorem 9.5 and Theorem 9.6, and both being instructive, we will try them both. The computations are as follows:

(1) Let us first go the pedestrian way, based on the definition of the resultant, from Theorem 9.5. Consider two polynomials, of degree 3 and degree 2, written as follows:

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

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

The resultant of these two polynomials is then given by:

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

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

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

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

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

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

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

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

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

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

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

Thus, we have the following formula for the resultant:

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

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

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

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

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

But this gives the formula in the statement, namely:

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

(2) Let us see as well how the computation does, by using Theorem 9.6, which is our most advanced tool, so far. Consider a polynomial of degree 3, and its derivative:

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

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

By using now Theorem 9.6 and computing the determinant, we obtain:

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

Now according to Theorem 9.7, the discriminant of our polynomial is given by:

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

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

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

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

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

THEOREM 9.11. *For a normalized degree 3 equation, namely*

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

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

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

is a real solution of our equation.

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

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

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

Regarding the other roots, we know from Proposition 9.9 that these are both real when $\Delta<0$, and complex conjugate when $\Delta<0$. Thus, in the context of Theorem 9.11, the other two roots are complex conjugate, the formula for them being as follows:

PROPOSITION 9.12. *For a normalized degree 3 equation, namely*

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

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

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

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

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

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

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

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

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

THEOREM 9.13. *For a normalized degree* 3 *equation, namely*

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

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

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

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

PROOF. This follows indeed from Theorem 9.11 and Proposition 9.12. Alternatively, we can redo the computation in their proof, which was nearly identical anyway, in the present setting, with x being given by the above formula, by using $w^3=1$. □

As a comment here, the formula in Theorem 9.13 holds of course in the case $\Delta>0$ too, and also when the coefficients are complex numbers, $p,q\in\mathbb{C}$, and this due to the fact that the proof rests on the nearly trivial computation from the proof of Theorem 9.11, or of Proposition 9.12. Thus, degree 3 equations fully understood, at least in theory.

However, in practice, these latter extensions are quite often not very useful, because when it comes to extract all the above square and cubic roots, for complex numbers, you can well end up with the initial question, the one that you started with.

9d. Higher degree

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

THEOREM 9.14. *The discriminant of* $P=ax^4+bx^3+cx^2+dx+e$ *is given by the following formula:*

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

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

PROOF. The formula of Δ follows from the definition of the discriminant, from Theorem 9.7, with the resultant computed via Theorem 9.6, as follows:

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

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

In practice, as in degree 3, we can do first some manipulations on our polynomials, as to have them in simpler form, and we have the following version of Theorem 9.14:

PROPOSITION 9.15. *The discriminant of $P=x^4+cx^2+dx+e$, normalized degree 4 polynomial, is given by the following formula:*

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

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

PROOF. This is a consequence of Theorem 9.14, with $a=1,b=0$, but we can deduce this as well directly. Indeed, the formula of Δ follows, quite easily, from:

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

As for the last assertion, this is something that we know, from Theorem 9.14. □

We still have some work to do. Indeed, looking back at what we did in degree 3, the passage there from Theorem 9.10 to Theorem 9.11 was made of two operations, namely "depressing" the equation, that is, getting rid of the next-to-highest term, and then rescaling the coefficients, as for the formula of Δ to become as simple as possible.

In our present setting now, degree 4, with the depressing done as above, in Proposition 9.15, it remains to rescale the coefficients, as for the formula of Δ to become as simple as possible. And here, a bit of formula hunting, in relation with $2,3$ powers, leads to:

THEOREM 9.16. *The discriminant of a normalized degree* 4 *polynomial, written as*

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

is given by the following formula:

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

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

PROOF. This follows from Proposition 9.15, with $c = 6p, d = 4q, e = 3r$, but we can deduce this as well directly. Indeed, the formula of Δ follows, quite easily, from:

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

As for the last assertion, this is something that we know from Theorem 9.14. □

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

THEOREM 9.17. *The roots of a normalized degree* 4 *equation, written as*

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

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

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

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

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

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

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

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

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

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

The idea will be that of adding a suitable common term, to both sides, as to make square on both sides, as to eventually end with a sort of double quadratic equation. For this purpose, our claim is that what we need is a number y satisfying:

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

Indeed, assuming that we have this number y, our equation becomes:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We still have to compute the number y appearing in the above via Cardano, and the result here, adding to what we already have in Theorem 9.17, is as follows:

THEOREM 9.18 (continuation). *The value of y in the previous theorem is*

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

where the number t is given by the formula

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

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

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

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

With this in hand, we have the following computation:

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

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

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

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

In degree 5 and more, things become fairly complicated. We will need:

THEOREM 9.19. *Given a field extension $E\subset F$, we can talk about its Galois group G, as the group of automorphisms of F fixing E. The intermediate fields*

$$E\subset K\subset F$$

are then in correspondence with the subgroups $H\subset G$, with such a field K corresponding to the subgroup H consisting of automorphisms $g\in G$ fixing K.

PROOF. This is something self-explanatory, and follows indeed from some algebra, under suitable assumptions, in order for that algebra to properly apply. □

Getting now towards polynomials and their roots, we have here:

THEOREM 9.20. *Given a field F and a polynomial $P\in F[X]$, we can talk about the abstract splitting field of P, where this polynomial decomposes as:*

$$P(X)=c\prod_i(X-a_i)$$

In particular, any field F has a certain algebraic closure $\bar{F}$, where all the polynomials $P\in F[X]$, and in fact all polynomials $P\in\bar{F}[X]$ too, have roots.

PROOF. This is again something self-explanatory, which follows from Theorem 9.19 and from some extra algebra, under suitable assumptions, in order for that extra algebra to properly apply. Regarding the construction at the end, as main example here we have $\bar{\mathbb{R}}=\mathbb{C}$. However, as an interesting fact, $\bar{\mathbb{Q}}\subset\mathbb{C}$ is a proper subfield. □

As an illustration for this, we can now elucidate the structure of finite fields:

THEOREM 9.21. *For any prime power $q=p^k$ there is a unique field $\mathbb{F}_q$ having q elements, which appears as the splitting field of the polynomial $P=X^q-X$.*

PROOF. We know from basic arithmetics that given a finite field, $|F| = q$ with $k \in \mathbb{N}$, the corresponding Fermat polynomial $P = X^q - X$ factorizes as follows:

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

But this shows, via the general theory from Theorem 9.20, that our field F must be the splitting field of P, and so is unique. As for the existence, this follows again from Theorem 9.20, telling us that the splitting field always exists. □

Getting back now to degree 5 equations, we have the following result:

THEOREM 9.22. *There is no general formula for the roots of polynomials of degree $N = 5$ and higher, with the reason for this, coming from Galois theory, being that the group S_5 is not solvable. The simplest numeric example is $P = X^5 - X - 1$.*

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

(1) Given a field F, assume that the roots of $P \in F[X]$ can be computed by using iterated roots, a bit as for the degree 2 equation, or the degree 3 and 4 equations. Then, algebrically speaking, this gives rise to a tower of fields as folows, with $F_0 = F$, and each F_{i+1} being obtained from F_i by adding a root, $F_{i+1} = F_i(x_i)$, with $x_i^{n_i} \in F_i$:

$$F_0 \subset F_1 \subset \ldots \subset F_k$$

(2) In order for Galois theory to apply to this situation, we must make all the extensions normal, which amounts in replacing each $F_{i+1} = F_i(x_i)$ by its extension $K_i(x_i)$, with K_i extending F_i by adding a n_i-th root of unity. Thus, with this replacement, we can assume that the tower in (1) in normal, meaning that all Galois groups are cyclic.

(3) Now by Galois theory, at the level of the corresponding Galois groups we obtain a tower of groups as follows as follows, which is a resolution of the last group G_k, the Galois group of P, in the sense of group theory, in the sense that all quotients are cyclic:

$$G_1 \subset G_2 \subset \ldots \subset G_k$$

As a conclusion, Galois theory tells us that if the roots of a polynomial $P \in F[X]$ can be computed by using iterated roots, then its Galois group $G = G_k$ must be solvable.

(4) In the generic case, the conclusion is that Galois theory tells us that, in order for all polynomials of degree 5 to be solvable, via square roots, the group S_5, which appears there as Galois group, must be solvable, in the sense of group theory. But this is wrong, because the alternating subgroup $A_5 \subset S_5$ is simple, and therefore not solvable.

(5) Finally, regarding the polynomial $P = X^5 - X - 1$, some elementary computations here, based on arithmetic over $\mathbb{F}_2, \mathbb{F}_3$, and involving various cycles of length $2, 3, 5$, show that its Galois group is S_5. Thus, we have our counterexample.

(6) Finally, let us mention that all this shows as well that a random polynomial of degree 5 or higher is not solvable by square roots, and with this being an elementary consequence of the main result from (4), via some standard analysis arguments. □

There is a lot of further interesting theory that can be developed here, following Galois and others. For more on all this, we recommend any number theory book.

9e. Exercises

Exercises:

Exercise 9.23.

Exercise 9.24.

Exercise 9.25.

Exercise 9.26.

Exercise 9.27.

Exercise 9.28.

Exercise 9.29.

Exercise 9.30.

Bonus exercise.

CHAPTER 10

Functions

10a. Functions, continuity

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

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, and as a basic example here, we have the inverse function, which cannot be defined at $x = 0$:

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

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

DEFINITION 10.2 (update). *More generally, we call real function any correspondence*

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

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

In practice, however, this 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 is not really possible, in a general way, convenient for all functions, because depending on the exact function $f : \mathbb{R} \to \mathbb{R}$ that we have in mind, restricting the domain to this or that $X \subset \mathbb{R}$, as to have f injective, remains something subjective. We will be back to this, with some explicit examples, when knowing more about functions.

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

QUESTION 10.4. *How to suitably represent our functions $f : \mathbb{R} \to \mathbb{R}$?*

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

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. □

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 10.7. *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*.*

We focus now our study on the functions $f : \mathbb{R} \to \mathbb{R}$ which are suitably regular. And, in what regards these regularity properties, the most basic of them is continuity:

DEFINITION 10.8. *A function $f : \mathbb{R} \to \mathbb{R}$, or more generally $f : X \to \mathbb{R}$, with $X \subset \mathbb{R}$ being a subset, is called continuous when, for any $x_n, x \in X$:*

$$x_n \to x \implies f(x_n) \to f(x)$$

Also, we say that $f : X \to \mathbb{R}$ is continuous at a given point $x \in X$ when the above condition is satisfied, for that point x.

Observe that a function $f : X \to \mathbb{R}$ is continuous precisely when it is continuous at any point $x \in X$. We will see examples in a moment. Still speaking theory, there are many equivalent formulations of the notion of continuity, with a well-known one, coming by reminding in the above definition what convergence of a sequence means, twice, for both the convergences $x_n \to x$ and $f(x_n) \to f(x)$, being as follows:

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

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

THEOREM 10.9. *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:

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

Alternatively, let us prove this as well by using that scary ε, δ condition given after Definition 10.8. So, let us pick $\varepsilon > 0$. We want in the end to have something of type $|f(g(x)) - f(g(y))| < \varepsilon$, so we must first use that ε, δ condition for the function f. So, let us start in this way. Since f is continuous at $g(x)$, we can find $\delta > 0$ such that:

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

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

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

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

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

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

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

THEOREM 10.10. *The following functions are continuous:*

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

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

(1) Since $f(x) = x$ is continuous, by using Theorem 10.9 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 10.9.

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

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

Indeed, with this formula in hand, we can establish the continuity of $\sin x$, as follows, with the limits at 0 which are used being both clear on pictures:

$$\begin{aligned}
\lim_{y\to 0} \sin(x+y) &= \lim_{y\to 0} (\sin x \cos y + \cos x \sin y) \\
&= \sin x \lim_{y\to 0} \cos y + \cos x \lim_{y\to 0} \sin y \\
&= \sin x \cdot 1 + \cos x \cdot 0 \\
&= \sin x
\end{aligned}$$

(4) Moving ahead now with $\cos x$, here the continuity follows from the continuity of $\sin x$, by using the following formula, which is obvious from definitions:

$$\cos x = \sin\left(\frac{\pi}{2} - x\right)$$

(5) Alternatively, and let us do this because we will need later the formula, by using the formula for $\sin(x+y)$ we can deduce a formula for $\cos(x+y)$, as follows:

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

But with this, we can use the same method as in (4), and we get, as desired:

$$\begin{aligned}
\lim_{y\to 0} \cos(x+y) &= \lim_{y\to 0} (\cos x \cos y - \sin x \sin y) \\
&= \cos x \lim_{y\to 0} \cos y - \sin x \lim_{y\to 0} \sin y \\
&= \cos x \cdot 1 - \sin x \cdot 0 \\
&= \cos x
\end{aligned}$$

(6) Finally, the fact that $\tan x$, $\cot x$ are continuous is clear from the fact that $\sin x$, $\cos x$ are continuous, by using the result regarding quotients from Theorem 10.9. □

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

PROPOSITION 10.11. *If a function $f : X \to \mathbb{R}$ has the Lipschitz property*

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

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

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

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

DEFINITION 10.12. *A function $f : X \to \mathbb{R}$ is called uniformly continuous when:*

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

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

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

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

Indeed, it is clear by looking at the graph of f that, the further our point $x \in \mathbb{R}$ is from 0, the smaller our $\delta > 0$ must be, compared to $\varepsilon > 0$, in our ε, δ definition of continuity. Thus, given an $\varepsilon > 0$, we have no $\delta > 0$ doing the $|x - y| < \delta \implies |f(x) - f(y)| < \varepsilon$ job at any $x \in \mathbb{R}$, and so our function is indeed not uniformly continuous.

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

THEOREM 10.13. *Any continuous function defined on a closed, bounded interval*

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

is automatically uniformly continuous.

PROOF. This is something quite subtle, the idea with this being as follows:

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

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

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

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

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

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

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

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

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

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

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

There are many functions which are not continuous everywhere, such as $f(x) = 1/x$ at $x = 0$. And the question is, what to do with them? That is, can we have some mathematical theory going on for them as well, inspired by what we did in the above?

In answer, here is something that we can do, in general:

DEFINITION 10.14. *Given a function $f : X \to \mathbb{R}$ and $x \in X$, we set*

$$f(x_-) = \lim_{y \nearrow x} f(y) \quad , \quad f(x_+) = \lim_{y \searrow x} f(y)$$

provided that these two limits exist indeed, and we call the quantity

$$J_f(x) = f(x_+) - f(x_-)$$

which does not depend on $f(x)$, the jump of f at the given point $x \in X$.

As a first observation, assuming that a function $f : X \to \mathbb{R}$ is continuous at $x \in X$, its jump there is zero, so that we have the following implications:

$$f \text{ continuous at } x \implies J_f(x) = 0$$

$$f \text{ continuous} \implies J_f(x) = 0, \ \forall x \in X$$

Observe also that the converses of these implications do not necessarily hold, and this because the jump $J_f(x)$, as constructed above, does not depend on $f(x)$, so we can easily construct counterexamples, just by modifying the value $f(x)$. More on this later.

Here are now some basic computations of jumps:

PROPOSITION 10.15. *For the basic step function, given by*

$$f(x)=\begin{cases}0 & \text{if } x<0\\ 1 & \text{if } x>0\end{cases}$$

we have $J_f(0)=1$, as we should. Also, for the inverse function

$$g(x)=\frac{1}{x}$$

we have $J_g(0)=\infty$, again as we should.

PROOF. Both the above formulae are indeed clear from definitions. □

Going ahead now with more theory, we can complement the basic notions introduced in Definition 10.14 with some more notions, which are equally useful. First we have:

DEFINITION 10.16. *Given a function $f:X\to\mathbb{R}$ and $x\in X$, we say that:*

(1) *f is left continuous at x, if $f(x_-)=f(x)$.*
(2) *f is right continuous at x, if $f(x_+)=f(x)$.*

As a first observation, a function $f:X\to\mathbb{R}$ is continuous at x precisely when it is left and right continuous there, so that we have the following equivalences:

$$f \text{ continuous at } x \iff f(x_-)=f(x)=f(x_+)$$

$$f \text{ continuous} \iff f(x_-)=f(x)=f(x_+),\ \forall x\in X$$

Which sounds quite interesting, in relation with the issues with the jump, discussed after Definition 10.14. So, let us update as well the definition of the jump, as follows:

THEOREM 10.17. *Given a function $f:X\to\mathbb{R}$ and $x\in X$, we call the quantities*

$$J_f(x_-)=f(x)-f(x_-)\quad,\quad J_f(x_+)=f(x_+)-f(x)$$

the left and right jumps at x, so that the total jump there is given by:

$$J_f(x)=J_f(x_+)+J_f(x_-)$$

The function f is then continuous at x when both its jumps there vanish,

$$f \text{ continuous at } x \iff J(x_-)=J(x_+)=0$$

and globally continuous, when we have $J(x_-)=J(x_+)=0$, for any $x\in X$.

PROOF. This is something quite self-explanatory, with a lot of talking, basically definitions, and with the jump formula being something obvious, as follows:

$$\begin{aligned}J_f(x)&=f(x_+)-f(x_-)\\ &=(f(x_+)-f(x))+(f(x)-f(x_-))\\ &=J_f(x_+)+J_f(x_-)\end{aligned}$$

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

As an illustration now, for the same functions as in Proposition 10.15, we have:

PROPOSITION 10.18. *For the basic step function, given by*

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

we have $J_f(0_-) = 1$ and $J_f(0_+) = 0$, as we should. Also, for the inverse function

$$g(x) = \frac{1}{x}$$

with the convention $g(0) = \alpha \in \mathbb{R}$, we have $J_g(0-) = J_g(0_+) = \infty$, also as we should.

PROOF. The above formulae are indeed all clear from definitions. □

As a last topic of discussion, let us go back to our original notion of jump, from Definition 10.14. We can say now something interesting here, as follows:

THEOREM 10.19. *Assuming that a function $f : X \to \mathbb{R}$ does not jump at $x \in X$,*

$$J_f(x) = 0$$

we can modify our function by forgetting the old value $f(x)$, and setting

$$f(x) = f(x_-) = f(x_+)$$

and we obtain in this way a function which is continuous at x.

PROOF. This is something which is clear from Theorem 10.17, because our modified function f has both its left and right jumps vanishing at x:

$$J_f(x_-) = J_f(x_+) = 0$$

Thus Theorem 10.17 applies, and tells us that our modified f is continuous at x. □

The above result can be applied to the various points where f is discontinous, provided that these points are "isolated" from each other. In the case where f needs as "fix" on a more substantial set of points, things are more complicated. We will be back to this.

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 10.20. *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 10.21. *A set $O \subset \mathbb{R}$ is open precisely when its complement $C \subset \mathbb{R}$ is closed, and vice versa.*

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

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

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

As 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 10.22. *A function is continuous precisely when $f^{-1}(O)$ is open, for any O open. Equivalently, $f^{-1}(C)$ must be closed, for any C closed.*

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

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

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

As a test for the above criterion, let us reprove the fact, that we know from Theorem 10.9, that if f, g are continuous, so is $f \circ g$. But this is clear, coming from:

$$(f \circ g)^{-1}(O) = g^{-1}(f^{-1}(O))$$

In short, not bad, because at least in relation with this specific problem, our proof using open sets is as simple as the simplest proof, namely the one using $f(x_n) \to f(x)$, and is simpler than the other proof that we know, namely the one with ε, δ.

In order to reach to true applications of Theorem 10.22, we will need to know more about the open and closed sets. Let us begin with a useful result, as follows:

PROPOSITION 10.23. *The following happen:*

(1) *Union of open sets is open.*
(2) *Intersection of closed sets is closed.*
(3) *Finite intersection of open sets is open.*
(4) *Finite union of closed sets is closed.*

PROOF. Here (1) is clear from definitions, (3) is clear from definitions too, and (2,4) follow from (1,3) by taking complements $E \to E^c$, using the following formulae:

$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c \quad , \quad \left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c$$

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

As an important comment, (3,4) above do not hold when removing the finiteness assumption. Indeed, in what regards (3), the simplest counterexample here is:

$$\bigcap_{n\in\mathbb{N}} \left(-\frac{1}{n}\,,\,\frac{1}{n}\right) = \{0\}$$

As for (4), here the simplest counterexample is as follows:

$$\bigcup_{n\in\mathbb{N}} \left[0\,,\,1-\frac{1}{n}\right] = [0,1)$$

All this is quite interesting, and leads us to the question about what the open and closed sets really are. And fortunately, this question can be answered, as follows:

THEOREM 10.24. *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 10.23 a union of such intervals is open.

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

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

(4) Finally, the second assertion comes from Proposition 10.21. □

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

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

As basic examples, the closed bounded intervals $[a, b]$ are compact, as we know from the proof of Theorem 10.13, and so are the finite unions of such intervals. As for connected sets, the basic examples here are the various types of intervals, namely (a, b), $(a, b]$, $[a, b)$, $[a, b]$, and it looks impossible to come up with more examples. In fact, we have:

THEOREM 10.26. *The compact and connected sets are as follows:*

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

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

(1) The fact that compact implies both closed and bounded is clear from our definition of compactness, because assuming non-closedness or non-boundedness leads to an open cover having no finite subcover. As for the converse, we know from the proof of Theorem 10.13 that any closed bounded interval $[a, b]$ is compact, and it follows that any $K \subset \mathbb{R}$ closed and bounded is a closed subset of a compact set, which follows to be compact.

(2) This is something which is obvious, and this regardless of what "cannot be broken into parts" in Definition 10.25 exactly means, mathematically speaking, with several possible definitions being possible here, all being equivalent. Indeed, $E \subset \mathbb{R}$ having this property is equivalent to $a, b \in E \implies [a, b] \subset E$, and this gives the result. □

We will be back to all this later in this book, when looking at open, closed, compact and connected sets in $\mathbb{R}^N$, or more general spaces, where things are more complicated than in $\mathbb{R}$. Now with this discussed, let us go back to continuous functions. We have:

THEOREM 10.27. *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 10.25. To be more precise:

(1) This comes from the fact that if a function f is continuous, then the inverse function f^{-1} returns an open cover into an open cover.

(2) This is something clear as well, because if $f(E)$ can be split into two parts, then by applying f^{-1} we can split as well E into two parts. □

Let us record as well the following useful generalization of Theorem 10.13:

THEOREM 10.28. *Any continuous function defined on a compact set*

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

is automatically uniformly continuous.

PROOF. We can prove this as Theorem 10.13, by using the compactness of X. □

You might perhaps ask at this point, were Theorems 10.27 and 10.28 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 10.29. *The following happen for a continuous function $f : [a, b] \to \mathbb{R}$:*

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

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

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

THEOREM 10.30. *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 10.29. Regarding now the continuity of f^{-1}, we want to prove that we have:

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

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

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

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

And with this, we have now all the needed generalities in our bag.

10c. Elementary functions

As a basic application of Theorem 10.30, we have:

PROPOSITION 10.31. *The various usual inverse functions, such as the inverse trigonometric functions* arcsin*,* arccos*,* arctan*,* arccot*, are all continuous.*

PROOF. This follows indeed from Theorem 10.30, with a course the full discussion needing some explanations on bijectivity and domains. But you surely know all that, and in what concerns us, our claim is simply that these beasts are all continuous, proved. □

As another basic application of this, we have:

PROPOSITION 10.32. *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 10.29, as follows:

(1) This is clear from Theorem 10.29 (3), applied on $[-\infty,\infty]$.

(2) This follows from (1), by using the polynomial $P(z) = z^n - x$.

(3) This follows as well by applying Theorem 10.29 (3) to the polynomial $P(z) = z^n - x$, but this time on $[0,\infty)$. □

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

THEOREM 10.33. *The function x^a is defined and continuous on $(0,\infty)$, for any $a \in \mathbb{R}$. Moreover, when trying to extend it to $\mathbb{R}$, we have 4 cases, as follows,*

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

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

PROOF. The idea is that we know how to extract roots by using Proposition 10.32, and all the rest follows by continuity. To be more precise:

(1) Assume $a = p/q$, with $p, q \in \mathbb{N}$, $p \neq 0$ and q odd. Given a number $x \in \mathbb{R}$, we can construct the power x^a in the following way, by using Proposition 10.32:

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

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

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

(3) Assume first $a \in \mathbb{Q}_{even}$, $a > 0$. This means $a = p/q$ with $p, q \in \mathbb{N}$, $p \neq 0$ and q even, and as before in (1), we can set $x^a = \sqrt[q]{x^p}$ for $x \geq 0$, by using Proposition 10.32. 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 10.34. *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 10.33, by exchanging the variables, $x \leftrightarrow a$. To be more precise, the situation is as follows:

(1) We know from Theorem 10.33 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 10.33 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. □

Our goal now will be to extend the material from chapter 4 regarding the numeric sequences and series, to the case of the sequences and series of functions. To start with, we can talk about the convergence of sequences of functions, $f_n \to f$, as follows:

DEFINITION 10.35. *We say that f_n converges pointwise to f, and write $f_n \to f$, if*

$$f_n(x) \to f(x)$$

for any x. Equivalently, $\forall x, \forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, |f_n(x) - f(x)| < \varepsilon$.

The question is now, assuming that f_n are continuous, does it follow that f is continuous? I am pretty much sure that you think that the answer is "yes", based on:

$$\begin{aligned}
\lim_{y \to x} f(y) &= \lim_{y \to x} \lim_{n \to \infty} f_n(y) \\
&= \lim_{n \to \infty} \lim_{y \to x} f_n(y) \\
&= \lim_{n \to \infty} f_n(x) \\
&= f(x)
\end{aligned}$$

However, this proof is wrong, because we know well from chapter 1 that we cannot intervert limits, with this being a common beginner mistake. In fact, the result itself is wrong in general, because if we consider the functions $f_n : [0,1] \to \mathbb{R}$ given by $f_n(x) = x^n$, which are obviously continuous, their limit is discontinuous, given by:

$$\lim_{n\to\infty} x^n = \begin{cases} 0 & , \quad x \in [0,1) \\ 1 & , \quad x = 1 \end{cases}$$

Of course, you might say here that allowing $x = 1$ in all this might be a bit unnatural, for whatever reasons, but there is an answer to this too. We can do worse, as follows:

PROPOSITION 10.36. *The basic step function, namely the sign function*

$$sgn(x) = \begin{cases} -1 & , \quad x < 0 \\ 0 & , \quad x = 0 \\ 1 & , \quad x > 0 \end{cases}$$

can be approximated by suitable modifications of $\arctan(x)$*. Even worse, there are examples of* $f_n \to f$ *with each* f_n *continuous, and with* f *totally discontinuous.*

PROOF. To start with, $\arctan(x)$ looks a bit like $sgn(x)$, so to say, but one problem comes from the fact that its image is $[-\pi/2, \pi/2]$, instead of the desired $[-1, 1]$. Thus, we must first rescale $\arctan(x)$ by $\pi/2$. Now with this done, we can further stretch the variable x, as to get our function closer and closer to $sgn(x)$, as desired. This proves the first assertion, and the second assertion, which is a bit more technical, and that we will not really need in what follows, is left as an exercise for you, reader. □

Sumarizing, we are a bit in trouble, because we would like to have in our bag of theorems something saying that $f_n \to f$ with f_n continuous implies f continuous. Fortunately, this can be done, with a suitable refinement of the notion of convergence, as follows:

DEFINITION 10.37. *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 10.35 and Definition 10.37. As a first observation, we have:

PROPOSITION 10.38. *Uniform convergence implies pointwise convergence,*

$$f_n \to_u f \implies f_n \to f$$

but the converse is not true, in general.

PROOF. Here the first assertion is clear from definitions, just by thinking at what is going on, with no computations needed. As for the second assertion, the simplest counterexamples here are the functions $f_n : [0,1] \to \mathbb{R}$ given by $f_n(x) = x^n$, that we met before in Proposition 10.36. Indeed, uniform convergence on $[0,1)$ would mean:

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, x^n < \varepsilon, \forall x \in [0,1)$$

But this is wrong, because no matter how big N is, we have $\lim_{x\to 1} x^N = 1$, and so we can find $x \in [0,1)$ such that $x^N > \varepsilon$. Thus, we have our counterexample. □

Moving ahead now, let us state our main theorem on uniform convergence, as follows:

THEOREM 10.39. *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 10.37 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 10.37 is something quite interesting, worth some more study. As a first result, we have:

PROPOSITION 10.40. *The following happen, regarding uniform limits:*

(1) $f_n \to_u f$, $g_n \to_u g$ *imply* $f_n + g_n \to_u f + g$.
(2) $f_n \to_u f$, $g_n \to_u g$ *imply* $f_n g_n \to_u fg$.
(3) $f_n \to_u f$, $f \neq 0$ *imply* $1/f_n \to_u 1/f$.
(4) $f_n \to_u f$, g *continuous imply* $f_n \circ g \to_u f \circ g$.
(5) $f_n \to_u f$, g *continuous imply* $g \circ f_n \to_u g \circ f$.

PROOF. All this is routine, exactly as for the results for numeric sequences from chapter 4, that we know well, with no difficulties or tricks involved. □

Finally, there is some abstract mathematics to be done as well. Indeed, observe that the notion of uniform convergence, as formulated in Definition 10.37, means that:

$$\sup_x \left|f_n(x) - f(x)\right| \longrightarrow_{n\to\infty} 0$$

This suggests measuring the distance between functions via a supremum as above, and in relation with this, we have the following result:

THEOREM 10.41. *The uniform convergence, $f_n \to_u f$, means that we have $f_n \to f$ with respect to the following distance,*

$$d(f,g) = \sup_x \left|f(x) - g(x)\right|$$

which is indeed a distance function.

PROOF. Here the fact that d is indeed a distance, in the sense that it satisfies all the intuitive properties of a distance, including the triangle inequality, follows from definitions, and the fact that the uniform convergence can be interpreted as above is clear as well. □

10d. Binomial formula

With the above theory in hand, let us get now to interesting things, namely computations. Among others, because this is what a mathematician's job is, doing all sorts of computations. We will be mainly interested in the functions x^a and a^x, which remain something quite mysterious. Regarding x^a, we first have the following result:

THEOREM 10.42. *We have the generalized binomial formula*

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

with the generalized binomial coefficients being given by

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

valid for any exponent $a \in \mathbb{Z}$, and any $|x| < 1$.

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

(1) For exponents $a \in \mathbb{N}$, this is something that we know well from chapter 1, and which is valid for any $x \in \mathbb{R}$, coming from the usual binomial formula, namely:

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

(2) For the exponent $a=-1$ this is something that we know from Part I too, coming from the following formula, valid for any $|x|<1$:

$$\frac{1}{1+x}=1-x+x^2-x^3+\ldots$$

Indeed, this is exactly our generalized binomial formula at $a=-1$, because:

$$\binom{-1}{k}=\frac{(-1)(-2)\ldots(-k)}{k!}=(-1)^k$$

(3) Let us discuss now the general case $a\in-\mathbb{N}$. With $a=-n$, and $n\in\mathbb{N}$, the generalized binomial coefficients are given by the following formula:

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

Thus, our generalized binomial formula at $a=-n$, and $n\in\mathbb{N}$, reads:

$$\frac{1}{(1+t)^n}=\sum_{k=0}^{\infty}(-1)^k\binom{n+k-1}{n-1}t^k$$

(4) In order to prove this formula, it is convenient to write it with $-t$ instead of t, in order to get rid of signs. The formula to be proved becomes:

$$\frac{1}{(1-t)^n}=\sum_{k=0}^{\infty}\binom{n+k-1}{n-1}t^k$$

We prove this by recurrence on n. At $n=1$ this formula definitely holds, as explained in (2) above. So, assume that the formula holds at $n\in\mathbb{N}$. We have then:

$$\begin{aligned}\frac{1}{(1-t)^{n+1}}&=\frac{1}{1-t}\cdot\frac{1}{(1-t)^n}\\&=\sum_{k=0}^{\infty}t^k\sum_{l=0}^{\infty}\binom{n+l-1}{n-1}t^l\\&=\sum_{s=0}^{\infty}t^s\sum_{l=0}^{s}\binom{n+l-1}{n-1}\end{aligned}$$

On the other hand, the formula that we want to prove is:

$$\frac{1}{(1-t)^{n+1}}=\sum_{s=0}^{\infty}\binom{n+s}{n}t^k$$

Thus, in order to finish, we must prove the following formula:

$$\sum_{l=0}^{s}\binom{n+l-1}{n-1}=\binom{n+s}{n}$$

(5) In order to prove this latter formula, we proceed by recurrence on $s\in\mathbb{N}$. At $s=0$ the formula is trivial, $1=1$. So, assume that the formula holds at $s\in\mathbb{N}$. In order to prove the formula at $s+1$, we are in need of the following formula:

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

But this is the Pascal formula, that we know from chapter 2, and we are done. □

Let us discuss now some further generalizations of what we have. Quite interestingly, we have as well the following result, which is something very useful, in practice:

THEOREM 10.43. *The generalized binomial formula, namely*

$$(1+x)^a=\sum_{k=0}^{\infty}\binom{a}{k}x^k$$

holds as well at $a=\pm 1/2$. In practice, at $a=1/2$ we obtain the formula

$$\sqrt{1+t}=1-2\sum_{k=1}^{\infty}C_{k-1}\left(\frac{-t}{4}\right)^k$$

with $C_k=\frac{1}{k+1}\binom{2k}{k}$ being the Catalan numbers, and at $a=-1/2$ we obtain

$$\frac{1}{\sqrt{1+t}}=\sum_{k=0}^{\infty}D_k\left(\frac{-t}{4}\right)^k$$

with $D_k=\binom{2k}{k}$ being the central binomial coefficients.

PROOF. This can be done in several steps, as follows:

(1) At $a = 1/2$, the generalized binomial coefficients are as follows:

$$\begin{aligned}\binom{1/2}{k} &= \frac{1/2(-1/2)\ldots(3/2-k)}{k!}\\ &= (-1)^{k-1}\frac{1\cdot 3\cdot 5\ldots(2k-3)}{2^k k!}\\ &= (-1)^{k-1}\frac{(2k-2)!}{2^{k-1}(k-1)!2^k k!}\\ &= -2\left(\frac{-1}{4}\right)^k C_{k-1}\end{aligned}$$

(2) At $a = -1/2$, the generalized binomial coefficients are as follows:

$$\begin{aligned}\binom{-1/2}{k} &= \frac{-1/2(-3/2)\ldots(1/2-k)}{k!}\\ &= (-1)^k\frac{1\cdot 3\cdot 5\ldots(2k-1)}{2^k k!}\\ &= (-1)^k\frac{(2k)!}{2^k k!2^k k!}\\ &= \left(\frac{-1}{4}\right)^k D_k\end{aligned}$$

(3) Summarizing, we have proved so far that the binomial formula at $a = \pm 1/2$ is equivalent to the explicit formulae in the statement, involving the Catalan numbers C_k, and the central binomial coefficients D_k. It remains now to prove that these two explicit formulae hold indeed. For this purpose, let us write these formulae as follows:

$$\sqrt{1-4t} = 1 - 2\sum_{k=1}^{\infty} C_{k-1}t^k \quad , \quad \frac{1}{\sqrt{1-4t}} = \sum_{k=0}^{\infty} D_k t^k$$

In order to check these latter formulae, we must prove the following identities:

$$\left(1 - 2\sum_{k=1}^{\infty} C_{k-1}t^k\right)^2 = 1 - 4t \quad , \quad \left(\sum_{k=0}^{\infty} D_k t^k\right)^2 = \frac{1}{1-4t}$$

(4) As a first observation, the formula on the left is equivalent to:

$$\sum_{k+l=n} C_k C_l = C_{n+1}$$

By using the series for $1/(1-4t)$, the formula on the right is equivalent to:

$$\sum_{k+l=n} D_k D_l = 4^n$$

Finally, observe that if our formulae hold indeed, by multiplying we must have:

$$\sum_{k+l=n} C_k D_l = \frac{D_{n+1}}{2}$$

(5) Summarizing, we have to understand 3 formulae, which look quite similar. Let us first attempt to prove $\sum_{k+l=n} D_k D_l = 4^n$, by recurrence. We have:

$$D_{k+1} = \binom{2k+2}{k+1} = \frac{4k+2}{k+1}\binom{2k}{k} = \left(4 - \frac{2}{k+1}\right) D_k$$

Thus, assuming that we have $\sum_{k+l=n} D_k D_l = 4^n$, we obtain:

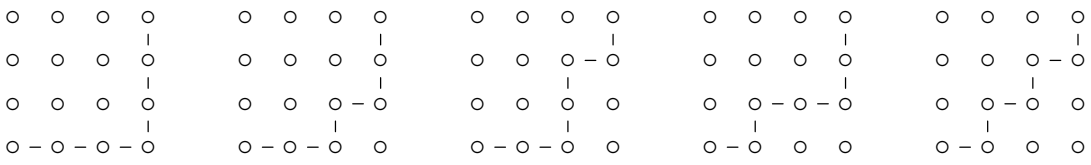
$$\begin{aligned}
\sum_{k+l=n+1} D_k D_l &= D_0 D_{n+1} + \sum_{k+l=n} \left(4 - \frac{2}{k+1}\right) D_k D_l \\
&= D_{n+1} + 4 \sum_{k+l=n} D_k D_l - 2 \sum_{k+l=n} \frac{D_k D_l}{k+1} \\
&= D_{n+1} + 4^{n+1} - 2 \sum_{k+l=n} C_k D_l
\end{aligned}$$

Thus, this leads to a sort of half-failure, the conclusion being that for proving by recurrence the second formula in (4), we need the third formula in (4).

(6) All this suggests a systematic look at the three formulae in (4). According to our various observations above, these three formulae are equivalent, and so it is enough to prove one of them. We will chose here to prove the first one, namely:

$$\sum_{k+l=n} C_k C_l = C_{n+1}$$

(7) For this purpose, we will trick. Let us count the Dyck paths in the plane, which are by definition the paths from $(0,0)$ to (n,n), marching North-East over the integer lattice $\mathbb{Z}^2 \subset \mathbb{R}^2$, by staying inside the square $[0,n] \times [0,n]$, and staying as well under the diagonal of this square. As an example, here are the 5 possible Dyck paths at $n = 3$:



In fact, the number C'_n of these paths is as follows, coinciding with C_n:

$$1, 1, 2, 5, 14, 42, 132, 429, \ldots$$

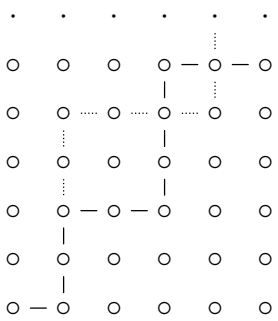
(8) We will prove that the numbers C'_n satisfy the recurrence for the numbers C_n that we want to prove, from (6), and on the other hand we will prove that we have $C'_n = C_n$.

Getting to work, in what regards our first task, this is easy, because when looking at where our path last intersects the diagonal of the square, we obtain, as desired:

$$C'_n = \sum_{k+l=n-1} C'_k C'_l$$

(9) In what regards now our second task, proving that we have $C'_n = C_n$, this is more tricky. If we ignore the assumption that our path must stay under the diagonal of the square, we have $\binom{2n}{n}$ such paths. And among these, we have the "good" ones, those that we want to count, and then the "bad" ones, those that we want to ignore.

(10) So, let us count the bad paths, those crossing the diagonal of the square, and reaching the higher diagonal next to it, the one joining $(0,1)$ and $(n,n+1)$. In order to count these, the trick is to "flip" their bad part over that higher diagonal, as follows:



(11) Now observe that, as it is obvious on the above picture, due to the flipping, the flipped bad path will no longer end in (n,n), but rather in $(n-1,n+1)$. Moreover, more is true, in the sense that, by thinking a bit, we see that the flipped bad paths are precisely those ending in $(n-1,n+1)$. Thus, good news, we are done with the count.

(12) To finish now, by putting everything together, we have:

$$\begin{aligned} C'_n &= \binom{2n}{n} - \binom{2n}{n-1} \\ &= \binom{2n}{n} - \frac{n}{n+1}\binom{2n}{n} \\ &= \frac{1}{n+1}\binom{2n}{n} \end{aligned}$$

Thus we have indeed $C'_n = C_n$, and this finishes the proof. □

The generalized binomial formula holds in fact for any exponent $a \in \mathbb{Z}/2$, after some combinatorial pain, and even for any $a \in \mathbb{R}$, but this is non-trivial. More on this later.

10e. Exercises

Exercises:

Exercise 10.44.

Exercise 10.45.

Exercise 10.46.

Exercise 10.47.

Exercise 10.48.

Exercise 10.49.

Exercise 10.50.

Exercise 10.51.

Bonus exercise.

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 $\varepsilon(t) \to 0$ for $t \to 0$:

$$f(x+t) = f(x) + \varepsilon(t)$$

Thus, what we are looking for is a better approximation, of the following type, with the function $\nu(t)$ being some sort of simple approximation of the error term $\varepsilon(t)$:

$$f(x+t) \simeq f(x) + \nu(t)$$

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*

$$f'(x) = \lim_{t \to 0} \frac{f(x+t) - f(x)}{t}$$

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:

$$\lim_{t \to 0} [f(x+t) - f(x)] = 0$$

Thus, f must be continuous at x. However, the converse is not true, a basic counterexample being $f(x) = |x|$ at $x = 0$. Let us summarize these findings as follows:

PROPOSITION 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:

$$\begin{aligned}(\sin x)' &= \lim_{t\to 0}\frac{\sin(x+t)-\sin x}{t}\\ &= \lim_{t\to 0}\frac{\sin x\cos t+\cos x\sin t-\sin x}{t}\\ &= \lim_{t\to 0}\sin x\cdot\frac{\cos t-1}{t}+\cos x\cdot\frac{\sin t}{t}\\ &= \cos x\end{aligned}$$

Here we have used the fact, which is clear on pictures, by drawing the trigonometric circle, that we have $\sin t \simeq t$ for $t \simeq 0$, plus the fact, which follows from this and from Pythagoras, $\sin^2 + \cos^2 = 1$, that we have as well $\cos t \simeq 1 - t^2/2$, for $t \simeq 0$.

(2) The computation for cos is similar, as follows:

$$\begin{aligned}(\cos x)' &= \lim_{t\to 0}\frac{\cos(x+t)-\cos x}{t}\\ &= \lim_{t\to 0}\frac{\cos x\cos t-\sin x\sin t-\cos x}{t}\\ &= \lim_{t\to 0}\cos x\cdot\frac{\cos t-1}{t}-\sin x\cdot\frac{\sin t}{t}\\ &= -\sin x\end{aligned}$$

(3) For the exponential, the derivative can be computed as follows:

$$\begin{aligned}(e^x)' &= \left(\sum_{k=0}^{\infty}\frac{x^k}{k!}\right)'\\ &= \sum_{k=0}^{\infty}\frac{kx^{k-1}}{k!}\\ &= e^x\end{aligned}$$

(4) As for the logarithm, the computation here is as follows, using $\log(1+y)\simeq y$ for $y\simeq 0$, which follows from $e^y\simeq 1+y$ that we found in (3), by taking the logarithm:

$$\begin{aligned}(\log x)' &= \lim_{t\to 0}\frac{\log(x+t)-\log x}{t}\\ &= \lim_{t\to 0}\frac{\log(1+t/x)}{t}\\ &= \frac{1}{x}\end{aligned}$$

Thus, we are led to the formulae in the statement. □

Speaking exponentials, we can now formulate a nice result about them:

THEOREM 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, and so our function must look as follows:

$$f(x)=1+\sum_{k=1}^{\infty}c_kx^k$$

According to our differentiation rules, the derivative of this series is given by:

$$f(x)=\sum_{k=1}^{\infty}kc_kx^{k-1}$$

Thus, the equation $f'=f$ is equivalent to the following equalities:

$$c_1=1\quad,\quad 2c_2=c_1\quad,\quad 3c_3=c_2\quad,\quad 4c_4=c_3\quad,\quad\ldots$$

But this system of equations can be solved by recurrence, as follows:

$$c_1 = 1 \quad , \quad c_2 = \frac{1}{2} \quad , \quad c_3 = \frac{1}{2 \times 3} \quad , \quad c_4 = \frac{1}{2 \times 3 \times 4} \quad , \quad \ldots$$

Thus we have $c_k = 1/k!$, leading to the conclusion in the statement. □

Observe that the above result leads to a more conceptual explanation for the number e itself. To be more precise, $e \in \mathbb{R}$ is the unique number satisfying:

$$(e^x)' = e^x$$

Which is 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:

$$\begin{aligned}
(f + g)'(x) &= \lim_{t \to 0} \frac{(f + g)(x + t) - (f + g)(x)}{t} \\
&= \lim_{t \to 0} \left(\frac{f(x + t) - f(x)}{t} + \frac{g(x + t) - g(x)}{t} \right) \\
&= \lim_{t \to 0} \frac{f(x + t) - f(x)}{t} + \lim_{t \to 0} \frac{g(x + t) - g(x)}{t} \\
&= f'(x) + g'(x)
\end{aligned}$$

(2) This follows from definitions too, the computation, by using the more convenient formula $f(x + t) \simeq f(x) + f'(x)t$ as a definition for the derivative, being as follows:

$$\begin{aligned}
(fg)(x + t) &= f(x + t)g(x + t) \\
&\simeq (f(x) + f'(x)t)(g(x) + g'(x)t) \\
&\simeq f(x)g(x) + (f'(x)g(x) + f(x)g'(x))t
\end{aligned}$$

Indeed, we obtain from this that the derivative is the coefficient of t, namely:

$$(fg)'(x) = f'(x)g(x) + f(x)g'(x)$$

(3) Regarding compositions, the computation here is as follows, again by using the more convenient formula $f(x+t) \simeq f(x) + f'(x)t$ as a definition for the derivative:

$$\begin{aligned}(f \circ g)(x+t) &= f(g(x+t)) \\ &\simeq f(g(x) + g'(x)t) \\ &\simeq f(g(x)) + f'(g(x))g'(x)t\end{aligned}$$

Indeed, we obtain from this that the derivative is the coefficient of t, namely:

$$(f \circ g)'(x) = f'(g(x))g'(x)$$

Thus, we are led to the conclusions in the statement. □

We can of course combine the above formulae, and we obtain for instance:

PROPOSITION 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. □

All the above might seem to start to be a bit too complex, with too many things to be memorized and so on, and as a piece of advice here, we have:

ADVICE 11.9. *Memorize and cherish the formula for fractions*

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$$

along with the usual addition formula, that you know well

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$$

and generally speaking, never mess with fractions.

With this coming from a lifelong calculus teacher and scientist, mathematics can be difficult, and many things can be pardoned, but not messing with fractions. And with this going beyond mathematics too, say if you want to make a living by selling apples or tomatoes at the market, fine, but you'll need to know well fractions, trust me.

Back to work now, with the above formulae in hand, we can do all sorts of computations for other basic functions that we know, including $\tan x$, or $\arctan x$:

PROPOSITION 11.10. *We have the following formulae,*

$$(\tan x)' = \frac{1}{\cos^2 x} \quad , \quad (\arctan x)' = \frac{1}{1 + x^2}$$

and the derivatives of the remaining trigonometric functions can be computed as well.

PROOF. For tan, we have the following computation:

$$\begin{aligned}(\tan x)' &= \left(\frac{\sin x}{\cos x}\right)' \\ &= \frac{\sin' x \cos x - \sin x \cos' x}{\cos^2 x} \\ &= \frac{\cos^2 x + \sin^2 x}{\cos^2 x} \\ &= \frac{1}{\cos^2 x}\end{aligned}$$

As for arctan, we can use here the following computation:

$$\begin{aligned}(\tan \circ \arctan)'(x) &= \tan'(\arctan x) \arctan'(x) \\ &= \frac{1}{\cos^2(\arctan x)} \arctan'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x' = 1$, we obtain from this:

$$\arctan'(x) = \cos^2(\arctan x)$$

On the other hand, with $t = \arctan x$ we know that we have $\tan t = x$, and so:

$$\cos^2(\arctan x) = \cos^2 t = \frac{1}{1 + \tan^2 t} = \frac{1}{1 + x^2}$$

Thus, we are led to the formula in the statement, namely:

$$(\arctan x)' = \frac{1}{1 + x^2}$$

As for the last assertion, we will leave this as an exercise. □

At the theoretical level now, further building on Theorem 11.3, we have:

THEOREM 11.11. *The local minima and maxima of a differentiable function $f : \mathbb{R} \to \mathbb{R}$ appear at the points $x \in \mathbb{R}$ where:*

$$f'(x) = 0$$

However, the converse of this fact is not true in general.

PROOF. The first assertion follows from the formula 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 11.11, we have:

THEOREM 11.12. *Assuming that $f : [a, b] \to \mathbb{R}$ is differentiable, we have*

$$\frac{f(b) - f(a)}{b - a} = f'(c)$$

for some $c \in (a, b)$, called mean value property of f.

PROOF. In the case $f(a) = f(b)$, the result, called Rolle theorem, states that we have $f'(c) = 0$ for some $c \in (a, b)$, and follows from Theorem 11.11. Now in what regards our statement, due to Lagrange, this follows from Rolle, applied to the following function:

$$g(x) = f(x) - \frac{f(b) - f(a)}{b - a} \cdot x$$

Indeed, we have $g(a) = g(b)$, due to our choice of the constant on the right, so we get $g'(c) = 0$ for some $c \in (a, b)$, which translates into the formula in the statement. □

In practice, Theorem 11.11 can be used in order to find the maximum and minimum of any differentiable function, and this method is best recalled as follows:

ALGORITHM 11.13. *In order to find the minimum and maximum of $f : [a, b] \to \mathbb{R}$:*

(1) *Compute the derivative f'.*
(2) *Solve the equation $f'(x) = 0$.*
(3) *Add a, b to your set of solutions.*
(4) *Compute $f(x)$, for all your solutions.*
(5) *Compute the min/max of all these $f(x)$ values.*
(6) *Then this is the min/max of your function.*

To be more precise, we are using here Theorem 11.11, 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 11.14. *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 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 11.15. *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 11.15 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 11.16. *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 11.15, in order to see what exactly is going on. Also, we have tan and the inverse trigonometric functions too. In short, plenty of good exercises here, for you, and the more you solve, the better your calculus will be.

(2) Regarding now the counterexample, recall first that the simplest example of a function which is continuous, but not differentiable, was $f(x) = |x|$, the idea behind this being to use a "piecewise linear function whose branches do not fit well". In connection now with our question, piecewise linear will not do, but we can use a similar idea, namely "piecewise quadratic function whose branches do not fit well". So, let us set:

$$f(x) = \begin{cases} ax^2 & (x \leq 0) \\ bx^2 & (x \geq 0) \end{cases}$$

This function is then differentiable, with its derivative being:

$$f'(x) = \begin{cases} 2ax & (x \leq 0) \\ 2bx & (x \geq 0) \end{cases}$$

Now for getting our counterexample, we can set $a = -1, b = 1$, so that f is:

$$f(x) = \begin{cases} -x^2 & (x \leq 0) \\ x^2 & (x \geq 0) \end{cases}$$

Indeed, the derivative is $f'(x) = 2|x|$, which is not differentiable, as desired. □

Getting now to theory, we first have the following key result:

THEOREM 11.17. *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 11.15 (3), we have the following statement, which is a bit heuristic, but we will call it however Proposition:

PROPOSITION 11.18. *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 11.17, as such, and with $t \to -t$ too:

$$f(x+t) \simeq f(x) + f'(x)t + \frac{f''(x)}{2}\,t^2$$

$$f(x-t) \simeq f(x) - f'(x)t + \frac{f''(x)}{2}\,t^2$$

By making the average, we obtain the following formula:

$$\frac{f(x+t) + f(x-t)}{2} \simeq f(x) + \frac{f''(x)}{2}\,t^2$$

But this is what our statement says, save for some uncertainties regarding the averaging method, and for the precise value of $I(t^2/2)$. We will leave this for later. □

Back to rigorous mathematics, we can improve as well Theorem 11.11, as follows:

THEOREM 11.19. *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 11.17, 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 11.11, 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 11.13 as well. Normally that algorithm stays strong, because Theorem 11.19 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 11.19 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 11.15 (4), we have the following result:

THEOREM 11.20. *Given a convex function $f : \mathbb{R} \to \mathbb{R}$, we have the following Jensen inequality, for any $x_1, \ldots, x_N \in \mathbb{R}$, and any $\lambda_1, \ldots, \lambda_N > 0$ summing up to 1,*

$$f(\lambda_1 x_1 + \ldots + \lambda_N x_N) \leq \lambda_1 f(x_1) + \ldots + \lambda_N x_N$$

with equality when $x_1 = \ldots = x_N$. In particular, by taking the weights λ_i to be all equal, we obtain the following Jensen inequality, valid for any $x_1, \ldots, x_N \in \mathbb{R}$,

$$f\left(\frac{x_1 + \ldots + x_N}{N}\right) \leq \frac{f(x_1) + \ldots + f(x_N)}{N}$$

and once again with equality when $x_1 = \ldots = x_N$. A similar statement holds for the concave functions, with all the inequalities being reversed.

PROOF. This is indeed something quite routine, the idea being as follows:

(1) First, we can talk about convex functions in a usual, intuitive way, with this meaning by definition that the following inequality must be satisfied:

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x) + f(y)}{2}$$

(2) But this means, via a simple argument, by approximating numbers $t \in [0, 1]$ by sums of powers 2^{-k}, that for any $t \in [0, 1]$ we must have:

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

Alternatively, via yet another simple argument, this time by doing some geometry with triangles, this means that we must have:

$$f\left(\frac{x_1 + \ldots + x_N}{N}\right) \leq \frac{f(x_1) + \ldots + f(x_N)}{N}$$

But then, again alternatively, by combining the above two simple arguments, the following must happen, for any $\lambda_1, \ldots, \lambda_N > 0$ summing up to 1:

$$f(\lambda_1 x_1 + \ldots + \lambda_N x_N) \leq \lambda_1 f(x_1) + \ldots + \lambda_N x_N$$

(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 11.21. *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 11.20, because we have:

$$(x^p)'' = p(p-1)x^{p-2}$$

Thus x^p is convex for $p > 1$ and concave for $p < 1$, which gives the results. □

Observe that at $p = 2$ we obtain as particular case of the above inequality the Cauchy-Schwarz inequality, or rather something equivalent to it, namely:

$$\left(\frac{x_1+\ldots+x_N}{N}\right)^2 \leq \frac{x_1^2+\ldots+x_N^2}{N}$$

We will be back to this later on in this book, when talking scalars products and Hilbert spaces, with some more conceptual proofs for such inequalities.

Finally, as yet another important application of the Jensen inequality, we have:

THEOREM 11.22. *We have the Young inequality,*

$$ab \leq \frac{a^p}{p}+\frac{b^q}{q}$$

valid for any $a,b \geq 0$, and any exponents $p,q > 1$ satisfying $\frac{1}{p}+\frac{1}{q} = 1$.

PROOF. We use the logarithm function, which is concave on $(0,\infty)$, due to:

$$(\log x)'' = \left(-\frac{1}{x}\right)' = -\frac{1}{x^2}$$

Thus we can apply the Jensen inequality, and we obtain in this way:

$$\begin{aligned}\log\left(\frac{a^p}{p}+\frac{b^q}{q}\right) &\geq \frac{\log(a^p)}{p}+\frac{\log(b^q)}{q}\\ &= \log(a)+\log(b)\\ &= \log(ab)\end{aligned}$$

Now by exponentiating, we obtain the Young inequality. □

In general, the Young inequality is something non-trivial, and the idea with it is that "when stuck with a problem, and with $ab \leq \frac{a^2+b^2}{2}$ not working, try Young". We will be back to this general principle, later in this book, with some illustrations.

11d. The Taylor formula

Back now to the general theory of the derivatives, and their theoretical applications, we can further develop our basic approximation method, at order 3, at order 4, and so on. Let us start with something nice and intuitive, as follows:

FACT 11.23. *The third derivatives are related to the jerk.*

Here the terminology comes from real life and classical mechanics, where the jerk is by definition the derivative of the acceleration, and so is the second derivative of the speed, and so is the third derivative of the position, according to the following formulae:

$$j = \dot{a} = \ddot{v} = \dddot{x}$$

As before with second derivatives, many other things can be said. Let us also record the formulae of the third derivatives of the basic functions, which are as follows:

THEOREM 11.24. *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 11.25. *The fourth derivatives of the basic functions are as follows:*

(1) $(x^p)'''' = p(p-1)(p-2)(p-3)x^{p-4}$.
(2) $\sin'''' = \sin$.
(3) $\cos'''' = \cos$.
(4) $\exp'''' = \exp$.
(5) $\log''''(x) = -6/x^4$.

PROOF. The various formulae in the statement all follow from the various formulae for the third derivatives established before, as follows:

$$(x^p)'''' = (p(p-1)(p-2)x^{p-3})' = p(p-1)(p-2)(p-3)x^{p-4}$$

$$(\sin x)'''' = (-\cos x)' = \sin x$$

$$(\cos x)'''' = (\sin x)' = \cos x$$

$$(e^x)'''' = (e^x)' = e^x$$

$$(\log x)'''' = (2/x^3)' = -6/x^4$$

Thus, we are led to the formulae in the statement. □

Observe the magic brought by the fourth derivative at the level of basic trigonometric functions. This is something quite subtle, and we will be back to it, later in this book.

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 11.26. *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 11.17. 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 11.27. *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 11.28. *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 a something self-explanatory, the idea being as follows:

(1) In order to compute the local maxima and minima, we know that the method is by using the following formula, which comes from the definition of the derivative:

$$f(x+t) \simeq f(x) + f'(x)t$$

Indeed, this formula shows that when $f'(x) \neq 0$, the point x cannot be a local minimum or maximum, due to the fact that $t \to -t$ will invert the growth.

(2) In relation with the problems left, the second derivative comes to the rescue. Indeed, we can use the following more advanced formula, coming via l'Hôpital's rule:

$$f(x+t) \simeq f(x) + f'(x)t + \frac{f''(x)}{2}\, t^2$$

To be more precise, assume that we have $f'(x) = 0$, as required by the study in (1). Then this second order formula simply reads:

$$f(x+t) \simeq f(x) + \frac{f''(x)}{2}\, t^2$$

But this is something very useful, telling us that when $f''(x) < 0$, what we have is a local maximum, and when $f''(x) > 0$, what we have is a local minimum. As for the remaining case, that when $f''(x) = 0$, things here remain open.

(3) All this is very useful in practice, and with what we have in (1), complemented if needed with what we have in (2), we can in principle compute the local minima and maxima, without much troubles. However, if really needed, more tools are available. Indeed, we can use if we want the order 3 Taylor formula, which is as follows:

$$f(x+t) \simeq f(x) + f'(x)t + \frac{f''(x)}{2}\, t^2 + \frac{f'''(x)}{6}\, t^3$$

To be more precise, assume that we are in the case $f'(x) = f''(x) = 0$, which is where our joint algorithm coming from (1) and (2) fails. In this case, our formula becomes:

$$f(x+t) \simeq f(x) + \frac{f'''(x)}{6}\, t^3$$

But this solves the problem in the case $f'''(x) \neq 0$, because here we cannot have a local minimum or maximum, due to $t \to -t$ which switches growth. As for the remaining case, $f'''(x) = 0$, things here remain open, and we have to go at higher order.

(4) Summarizing, we have a recurrence method for solving our problem. In order to formulate now an abstract result about this, we can use the Taylor formula at order n:

$$f(x+t)\simeq\sum_{k=0}^{n}\frac{f^{(k)}(x)}{k!}\,t^k$$

Indeed, assume that we started to compute the derivatives $f'(x),f''(x),f''(x),\ldots$ of our function at the point x, with the goal of finding the first such derivative which does not vanish, and we found this derivative, as being the order n one:

$$f'(x)=f''(x)=\ldots=f^{(n-1)}(x)=0\quad,\quad f^{(n)}(x)\neq 0$$

Then, the Taylor formula at x at order n takes the following form:

$$f(x+t)\simeq f(x)+\frac{f^{(n)}(x)}{n!}\,t^n$$

But this is exactly what we need, in order to fully solve our local extremum problem. Indeed, when n is even, if $f^{(n)}(x)<0$ what we have is a local maximum, and if $f^{(n)}(x)>0$, what we have is a local minimum. As for the case where n is odd, here we cannot have a local minimum or maximum, due to $t\to -t$ which switches growth. □

As a concrete application now of the Taylor formula, we have:

THEOREM 11.29. *We have the following formulae,*

$$\sin x=\sum_{l=0}^{\infty}(-1)^l\frac{x^{2l+1}}{(2l+1)!}\quad,\quad\cos x=\sum_{l=0}^{\infty}(-1)^l\frac{x^{2l}}{(2l)!}$$

as well as the following formulae,

$$e^x=\sum_{k=0}^{\infty}\frac{x^k}{k!}\quad,\quad\log(1+x)=\sum_{k=0}^{\infty}(-1)^{k+1}\frac{x^k}{k}$$

as Taylor series, and in general as well, with $|x|<1$ needed for log.

PROOF. There are several statements here, the proofs being as follows:

(1) Regarding sin and cos, we can use here the following formulae:

$$(\sin x)'=\cos x\quad,\quad(\cos x)'=-\sin x$$

Thus, we can differentiate sin and cos as many times as we want to, so we can compute the corresponding Taylor series, and we obtain the formulae in the statement.

(2) Regarding exp and log, here the needed formulae, which lead to the formulae in the statement for the corresponding Taylor series, are as follows:

$$(e^x)'=e^x\quad,\quad(\log x)'=x^{-1}\quad,\quad(x^p)'=px^{p-1}$$

(3) Finally, the fact that the formulae in the statement extend beyond the small t setting, coming from Taylor series, is something standard too. □

We can improve as well the binomial formula, in the following way:

THEOREM 11.30. *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) Case $p = -1$. Here we can use the following formula, valid for $|t| < 1$:

$$\frac{1}{1+t} = 1 - t + t^2 - t^3 + \ldots$$

But this is exactly our generalized binomial formula at $p = -1$, because:

$$\binom{-1}{k} = \frac{(-1)(-2)\ldots(-k)}{k!} = (-1)^k$$

(3) Case $p \in -\mathbb{N}$. This is a continuation of our study at $p = -1$, which will finish the study at $p \in \mathbb{Z}$. With $p = -m$, the generalized binomial coefficients are:

$$\begin{aligned}
\binom{-m}{k} &= \frac{(-m)(-m-1)\ldots(-m-k+1)}{k!} \\
&= (-1)^k \frac{m(m+1)\ldots(m+k-1)}{k!} \\
&= (-1)^k \frac{(m+k-1)!}{(m-1)!k!} \\
&= (-1)^k \binom{m+k-1}{m-1}
\end{aligned}$$

Thus, our generalized binomial formula at $p = -m$ reads:

$$\frac{1}{(1+t)^m} = \sum_{k=0}^{\infty} (-1)^k \binom{m+k-1}{m-1} t^k$$

But this is something which holds indeed, as we know from chapter 10.

(4) General case, $p \in \mathbb{R}$. As we can see, things escalate quickly, so we will skip the next step, $p \in \mathbb{Q}$, and discuss directly the case $p \in \mathbb{R}$. Consider the following function:

$$f(x) = x^p$$

The derivatives at $x = 1$ are then given by the following formula:

$$f^{(k)}(1) = p(p-1)\ldots(p-k+1)$$

Thus, the Taylor approximation at $x = 1$ is as follows:

$$f(1+t) = \sum_{k=0}^{\infty} \frac{p(p-1)\ldots(p-k+1)}{k!}\, t^k$$

But this is exactly our generalized binomial formula, so we are done with the case where t is small. With a bit more care, we obtain that this holds for any $|t| < 1$, and we will leave this as an instructive exercise, and come back to it, later in this book. □

As a main application now of our generalized binomial formula, which is something very useful in practice, we can extract square roots, as follows:

THEOREM 11.31. *We have the following formula,*

$$\sqrt{1+t} = 1 - 2\sum_{k=1}^{\infty} C_{k-1} \left(\frac{-t}{4}\right)^k$$

with $C_k = \frac{1}{k+1}\binom{2k}{k}$ *being the Catalan numbers. Also, we have*

$$\frac{1}{\sqrt{1+t}} = \sum_{k=0}^{\infty} D_k \left(\frac{-t}{4}\right)^k$$

with $D_k = \binom{2k}{k}$ *being the central binomial coefficients.*

PROOF. This is something that we already know from chapter 10, but time now to review all this. At $p = 1/2$, the generalized binomial coefficients are:

$$\begin{aligned}\binom{1/2}{k} &= \frac{1/2(-1/2)\dots(3/2-k)}{k!}\\ &= (-1)^{k-1}\frac{(2k-2)!}{2^{k-1}(k-1)!2^k k!}\\ &= -2\left(\frac{-1}{4}\right)^k C_{k-1}\end{aligned}$$

Also, at $p = -1/2$, the generalized binomial coefficients are:

$$\begin{aligned}\binom{-1/2}{k} &= \frac{-1/2(-3/2)\dots(1/2-k)}{k!}\\ &= (-1)^k\frac{(2k)!}{2^k k!2^k k!}\\ &= \left(\frac{-1}{4}\right)^k D_k\end{aligned}$$

Thus, Theorem 11.30 at $p = \pm 1/2$ gives the formulae in the statement. □

11e. Exercises

Exercises:

EXERCISE 11.32.

EXERCISE 11.33.

EXERCISE 11.34.

EXERCISE 11.35.

EXERCISE 11.36.

EXERCISE 11.37.

EXERCISE 11.38.

EXERCISE 11.39.

Bonus exercise.

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 book, 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:

$$a_{i+1} - a_i < \frac{\delta}{2}\ ,\ a'_{i+1} - a'_i < \delta_2 \implies \left|\Sigma_I(f) - \Sigma_{I'}(f)\right| < \varepsilon$$

(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:

$$\begin{aligned}\left|\Sigma_I(f) - \Sigma_{I'}(f)\right| &= \left|\sum_{ij} l_{ij} f(a_i) - \sum_{ij} l_{ij} f(a'_j)\right| \\ &= \left|\sum_{ij} l_{ij}(f(a_i) - f(a'_j))\right|\end{aligned}$$

(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:

$$\begin{aligned} \left|\Sigma_I(f) - \Sigma_{I'}(f)\right| &= \left|\sum_{ij} l_{ij}(f(a_i) - f(a'_j))\right| \\ &\leq \varepsilon \sum_{ij} l_{ij} \\ &= \varepsilon(b - a) \end{aligned}$$

Thus our two Riemann sums are close enough, provided that they are both chosen to be fine enough, and this finishes the proof of the first assertion.

(5) Regarding now the second assertion, this is something more technical, that we will not really need in what follows. We will leave the proof here, which uses similar ideas to those in the proof of (1) above, namely subdivisions and estimates, as an exercise. □

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:

$$1^0 + 2^0 + \ldots + N^0 = N$$

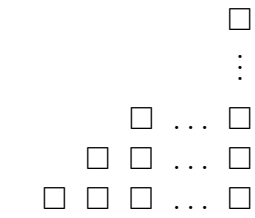
(4) At $p = 1$ now, we are confronted with a well-known question, namely the computation of $1 + 2 + \ldots + N$. But this is simplest done by arguing that the average of the numbers $1, 2, \ldots, N$ being the number in the middle, we have:

$$\frac{1 + 2 + \ldots + N}{N} = \frac{N + 1}{2}$$

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:



Now if we take two copies of this, and put them one on the top of the other, with a twist, in the obvious way, we obtain a rectangle having size $N \times (N + 1)$. Thus:

$$2(1 + 2 + \ldots + N) = N(N + 1)$$

But this gives the same formula as before, solving our question, namely:

$$1 + 2 + \ldots + N = \frac{N(N + 1)}{2} \simeq \frac{N^2}{2}$$

(6) Armed with this new method, let us attack now the case $p = 2$. Here we obviously need to do some 3D geometry, namely taking the picture P formed by a succession of solid squares, having sizes 1×1, 2×2, 3×3, and so on up to $N \times N$. Some quick thinking suggests that stacking 3 copies of P, with some obvious twists, will lead us to a

parallelepiped. But this is not exactly true, and some further thinking shows that what we have to do is to add 3 more copies of P, leading to the following formula:

$$1^2+2^2+\ldots+N^2=\frac{N(N+1)(2N+1)}{6}$$

Or at least, that's how the legend goes. In practice, the above formula holds indeed, and you can check it for instance by recurrence, and this solves our problem:

$$1^2+2^2+\ldots+N^2\simeq\frac{2N^3}{6}=\frac{N^3}{3}$$

(7) At $p=3$ now, the legend goes that by deeply thinking in 4D we are led to the following formula, a bit as in the cases $p=1,2$, explained above:

$$1^3+2^3+\ldots+N^3=\left(\frac{N(N+1)}{2}\right)^2$$

Alternatively, assuming that the gods of combinatorics are with us, we can see right away the following formula, which coupled with (4) gives the result:

$$1^3+2^3+\ldots+N^3=(1+2+\ldots+N)^2$$

In any case, in practice, the above formula holds indeed, and you can check it for instance by recurrence, and this solves our problem:

$$1^3+2^3+\ldots+N^3\simeq\frac{N^4}{4}$$

(8) Thus, good news, we proved our theorem. Of course, I can hear you screaming, that what about $p=4$ and higher. But the thing is that, by a strange twist of fate, there is no exact formula for $1^p+2^p+\ldots+N^p$, at $p=4$ and higher. Thus, game over. □

What happened above, with us unable to integrate x^p at $p=4$ and higher, not to mention the exponents $p\in\mathbb{R}-\mathbb{N}$ that we have not even dared to talk about, is quite annoying. As a conclusion to all this, however, let us formulate:

Conjecture 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,

$$\int_a^b x^p dx=\frac{b^{p+1}-a^{p+1}}{p+1}$$

valid for any exponent $p\in\mathbb{N}$, and perhaps for some other $p\in\mathbb{R}$.

We will see later that this conjecture is indeed true, and with the exact details regarding the exponents $p\in\mathbb{R}-\mathbb{N}$ too. Now, instead of struggling with this, let us look at some other functions, which are not polynomial. And here, as good news, we have:

THEOREM 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. Advanced results

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 is quite puzzling, usually my cat is quite helpful. Guess he must be either a reincarnation of Newton or Leibnitz, these gentlemen used to talk like that, or that I should take care at some point of my garden, remove catnip and other weeds.

This being said, wait. There is suggestion to connect integrals and derivatives, and this is in fact what we have, coming from Conjecture 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 10, 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 10, 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, 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 10, and whose proof is something elementary:

$$(F\varphi)'(t) = F'(\varphi(t))\varphi'(t)$$

Indeed, by integrating between c and d, we obtain the result. □

As a main application now of our theory, in relation with advanced calculus, and more specifically with the Taylor formula from chapter 11, we have:

THEOREM 12.26. *Given a function* $f : \mathbb{R} \to \mathbb{R}$, *we have the formula*

$$f(x+t) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k + \int_x^{x+t} \frac{f^{(n+1)}(s)}{n!}(x+t-s)^n\, ds$$

called Taylor formula with integral formula for the remainder.

PROOF. This is something which looks a bit complicated, so we will first do some verifications, and then we will go for the proof in general:

(1) At $n = 0$ the formula in the statement is as follows, and certainly holds, due to the fundamental theorem of calculus, which gives $\int_x^{x+t} f'(s)ds = f(x+t) - f(x)$:

$$f(x+t) = f(x) + \int_x^{x+t} f'(s)ds$$

(2) At $n = 1$, the formula in the statement becomes more complicated, as follows:

$$f(x+t) = f(x) + f'(x)t + \int_x^{x+t} f''(s)(x+t-s)ds$$

As a first observation, this formula holds indeed for the linear functions, where we have $f(x+t) = f(x) + f'(x)t$, and $f'' = 0$. So, let us try $f(x) = x^2$. Here we have:

$$f(x+t) - f(x) - f'(x)t = (x+t)^2 - x^2 - 2xt = t^2$$

On the other hand, the integral remainder is given by the same formula, namely:

$$\begin{aligned}
\int_x^{x+t} f''(s)(x+t-s)ds &= 2\int_x^{x+t} (x+t-s)ds \\
&= 2t(x+t) - 2\int_x^{x+t} sds \\
&= 2t(x+t) - ((x+t)^2 - x^2) \\
&= 2tx + 2t^2 - 2tx - t^2 \\
&= t^2
\end{aligned}$$

(3) Still at $n = 1$, let us try now to prove the formula in the statement, in general. Since what we have to prove is an equality, this cannot be that hard, and the first thought goes towards differentiating. But this method works indeed, and we obtain the result.

(4) In general, the proof is similar, by differentiating, the computations being similar to those at $n = 1$, and we will leave this as an instructive exercise. □

So long for basic integration theory. As a first concrete application now, we can compute all sorts of areas and volumes. Normally such computations are the business of multivariable calculus, and we will be back to this later, but with the technology that we have so far, we can do a number of things. As a first such computation, we have:

PROPOSITION 12.27. *The area of an ellipsis, given by the equation*

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$$

with $a, b > 0$ being half the size of a box containing the ellipsis, is $A = \pi ab$.

PROOF. The idea is that of cutting the ellipsis into vertical slices. First observe that, according to our equation $(x/a)^2 + (y/b)^2 = 1$, the x coordinate can range as follows:

$$x \in [-a, a]$$

For any such x, the other coordinate y, satisfying $(x/a)^2 + (y/b)^2 = 1$, is given by:

$$y = \pm b\sqrt{1 - \frac{x^2}{a^2}}$$

Thus the length of the vertical ellipsis slice at x is given by the following formula:

$$l(x) = 2b\sqrt{1 - \frac{x^2}{a^2}}$$

We conclude from this discussion that the area of the ellipsis is given by:

$$\begin{aligned} A &= 2b\int_{-a}^{a}\sqrt{1 - \frac{x^2}{a^2}}\,dx \\ &= \frac{4b}{a}\int_{0}^{a}\sqrt{a^2 - x^2}\,dx \\ &= 4ab\int_{0}^{1}\sqrt{1 - y^2}\,dy \\ &= 4ab \cdot \frac{\pi}{4} \\ &= \pi ab \end{aligned}$$

Finally, as a verification, for $a = b = 1$ we get $A = \pi$, as we should. □

Moving now to 3D, as an obvious challenge here, we can try to compute the volume of the sphere. This can be done a bit as for the ellipsis, the answer being as follows:

THEOREM 12.28. *The volume of the unit sphere is given by:*

$$V = \frac{4\pi}{3}$$

More generally, the volume of the sphere of radius R is $V = 4\pi R^3/3$.

PROOF. We proceed a bit as for the ellipsis. The equation of the sphere is:

$$x^2 + y^2 + z^2 = 1$$

Thus, the range of the first coordinate x is as follows:

$$x \in [-1, 1]$$

Now when this first coordinate x is fixed, the other coordinates y, z vary on a circle, given by the equation $y^2 + z^2 = 1 - x^2$, and so having radius as follows:

$$r(x) = \sqrt{1 - x^2}$$

Thus, the vertical slice of our sphere at x has area as follows:

$$a(x) = \pi r(x)^2 = \pi(1 - x^2)$$

We conclude from this discussion that the volume of the sphere is given by:

$$\begin{aligned} V &= \pi \int_{-1}^{1} 1 - x^2 \, dx \\ &= \pi \int_{-1}^{1} \left(x - \frac{x^3}{3} \right)' dx \\ &= \pi \left[\left(1 - \frac{1}{3} \right) - \left(-1 + \frac{1}{3} \right) \right] \\ &= \pi \left(\frac{2}{3} + \frac{2}{3} \right) \\ &= \frac{4\pi}{3} \end{aligned}$$

Finally, the last assertion is clear too, by multiplying everything by R, which amounts in multiplying the final result of our volume computation by R^3. □

As another application of our integration methods, we can now solve the 1D wave equation. In order to explain this, we will need a standard calculus result, as follows:

PROPOSITION 12.29. *The derivative of a function of type*

$$\varphi(x) = \int_{g(x)}^{h(x)} f(s) ds$$

is given by the formula $\varphi'(x) = f(h(x))h'(x) - f(g(x))g'(x)$.

PROOF. Consider a primitive of the function that we integrate, $F' = f$. We have:

$$\begin{aligned} \varphi(x) &= \int_{g(x)}^{h(x)} f(s) ds \\ &= \int_{g(x)}^{h(x)} F'(s) ds \\ &= F(h(x)) - F(g(x)) \end{aligned}$$

By using now the chain rule for derivatives, we obtain from this:

$$\begin{aligned} \varphi'(x) &= F'(h(x))h'(x) - F'(g(x))g'(x) \\ &= f(h(x))h'(x) - f(g(x))g'(x) \end{aligned}$$

Thus, we are led to the formula in the statement. □

Now back to the 1D waves, the result here, due to d'Alembert, is as follows:

THEOREM 12.30. *The solution of the 1D wave equation $\ddot{\varphi} = v^2\varphi''$ with initial value conditions $\varphi(x,0) = f(x)$ and $\dot{\varphi}(x,0) = g(x)$ is given by the d'Alembert formula:*

$$\varphi(x,t) = \frac{f(x-vt) + f(x+vt)}{2} + \frac{1}{2v}\int_{x-vt}^{x+vt} g(s)ds$$

Moreover, in the context of our previous lattice model discretizations, what happens is more or less that the above d'Alembert integral gets computed via Riemann sums.

PROOF. There are several things going on here, the idea being as follows:

(1) Let us first check that the d'Alembert solution is indeed a solution of the wave equation $\ddot{\varphi} = v^2\varphi''$. The first time derivative is computed as follows:

$$\dot{\varphi}(x,t) = \frac{-vf'(x-vt) + vf'(x+vt)}{2} + \frac{1}{2v}(vg(x+vt) + vg(x-vt))$$

The second time derivative is computed as follows:

$$\ddot{\varphi}(x,t) = \frac{v^2f''(x-vt) + v^2f(x+vt)}{2} + \frac{vg'(x+vt) - vg'(x-vt)}{2}$$

Regarding now space derivatives, the first one is computed as follows:

$$\varphi'(x,t) = \frac{f'(x-vt) + f'(x+vt)}{2} + \frac{1}{2v}(g'(x+vt) - g'(x-vt))$$

As for the second space derivative, this is computed as follows:

$$\varphi''(x,t) = \frac{f''(x-vt) + f''(x+vt)}{2} + \frac{g''(x+vt) - g''(x-vt)}{2v}$$

Thus we have indeed $\ddot{\varphi} = v^2\varphi''$. As for the initial conditions, $\varphi(x,0) = f(x)$ is clear from our definition of φ, and $\dot{\varphi}(x,0) = g(x)$ is clear from our above formula of $\dot{\varphi}$.

(2) Conversely now, we can simply solve our equation, which among others will doublecheck the computations in (1). Let us make the following change of variables:

$$\xi = x - vt \quad , \quad \eta = x + vt$$

With this change of variables, which is quite tricky, mixing space and time variables, our wave equation $\ddot{\varphi} = v^2\varphi''$ reformulates in a very simple way, as follows:

$$\frac{d^2\varphi}{d\xi d\eta} = 0$$

But this latter equation tells us that our new ξ, η variables get separated, and we conclude from this that the solution must be of the following special form:

$$\varphi(x,t) = F(\xi) + G(\eta) = F(x-vt) + G(x+vt)$$

Now by taking into account the intial conditions $\varphi(x,0) = f(x)$ and $\dot{\varphi}(x,0) = g(x)$, and then integrating, we are led to the d'Alembert formula. Finally, in what regards the last assertion, we will leave the study here as an instructive exercise. □

12d. Some probability

As another application of the integration theory developed above, let us develop now some theoretical probability theory. You probably know, from real life, what probability is. But in practice, when trying to axiomatize this, in mathematical terms, things can be quite tricky. So, here comes our point, the definition saving us is as follows:

DEFINITION 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 you 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 from coins, dice, cards and so on. For instance when rolling a die, the corresponding density is as follows:

$$\varphi = \frac{1}{6}\left(\delta_1 + \delta_2 + \delta_3 + \delta_4 + \delta_5 + \delta_6\right)$$

In what regards now the random variables f, described as above by densities φ, the first questions regard their mean and variance, constructed as follows:

DEFINITION 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 $P(f \in [a,b]) = \int_a^b d\mu(x)$. In this case

$$M_k = \int_{\mathbb{R}} x^k d\mu(x)$$

are called moments of f, and $M = M_1$ and $V = M_2 - M_1^2$ are called mean, and variance.

In practice now, let us look for some illustrations for this. The simplest random variables are those following discrete laws, $\psi = 0$, and as a basic example here, when flipping a coin and being rewarded \$0 for heads, and \$1 for tails, the corresponding law is $\mu = \frac{1}{2}(\delta_0 + \delta_1)$. More generally, playing the same game with a biased coin, which lands on heads with probability $p \in (0,1)$, leads to the following law, called Bernoulli law:

$$\mu = p\delta_0 + (1-p)\delta_1$$

Many more things can be said here, notably with a study of what happens when you play the game n times in a row, leading to some sort of powers of the Bernoulli laws, called binomial laws. Skipping some discussion here, and getting straight to the point, the most important laws in discrete probability are the Poisson laws, constructed as follows:

DEFINITION 12.35. *The Poisson law of parameter 1 is the following measure,*

$$p_1 = \frac{1}{e}\sum_{k\in\mathbb{N}} \frac{\delta_k}{k!}$$

and more generally, the Poisson law of parameter $t > 0$ is the following measure,

$$p_t = e^{-t}\sum_{k\in\mathbb{N}} \frac{t^k}{k!}\,\delta_k$$

with the letter "p" standing for Poisson.

Observe that our laws have indeed mass 1, as they should, and this due to:

$$e^t = \sum_{k\in\mathbb{N}} \frac{t^k}{k!}$$

In general, the idea with the Poisson laws is that these appear a bit everywhere, in the real life, the reasons for this coming from the Poisson Limit Theorem (PLT). However, this theorem uses advanced calculus, and we will leave it for later. In the meantime, however, we can have some fun with moments, the result here being as follows:

THEOREM 12.36. *The moments of p_1 are the Bell numbers,*

$$M_k(p_1) = |P(k)|$$

where $P(k)$ is the set of partitions of $\{1,\ldots,k\}$. More generally, we have

$$M_k(p_t) = \sum_{\pi\in P(k)} t^{|\pi|}$$

for any $t > 0$, where $|.|$ is the number of blocks.

PROOF. The moments of p_1 satisfy the following recurrence formula:

$$\begin{aligned}
M_{k+1} &= \frac{1}{e}\sum_r \frac{(r+1)^{k+1}}{(r+1)!}\\
&= \frac{1}{e}\sum_r \frac{r^k}{r!}\left(1+\frac{1}{r}\right)^k\\
&= \frac{1}{e}\sum_r \frac{r^k}{r!}\sum_s \binom{k}{s} r^{-s}\\
&= \sum_s \binom{k}{s}\cdot\frac{1}{e}\sum_r \frac{r^{k-s}}{r!}\\
&= \sum_s \binom{k}{s} M_{k-s}
\end{aligned}$$

With this done, let us try now to find a recurrence for the Bell numbers, $B_k = |P(k)|$. Since a partition of $\{1,\ldots,k+1\}$ appears by choosing s neighbors for 1, among the k numbers available, and then partitioning the $k-s$ elements left, we have:

$$B_{k+1} = \sum_s \binom{k}{s} B_{k-s}$$

Since the initial values coincide, $M_1 = B_1 = 1$ and $M_2 = B_2 = 2$, we obtain by recurrence $M_k = B_k$, as claimed. Regarding now the law p_t with $t > 0$, we have here a

similar recurrence formula for the moments, as follows:

$$\begin{aligned}
M_{k+1} &= e^{-t}\sum_r \frac{t^{r+1}(r+1)^{k+1}}{(r+1)!}\\
&= e^{-t}\sum_r \frac{t^{r+1}r^k}{r!}\left(1+\frac{1}{r}\right)^k\\
&= e^{-t}\sum_r \frac{t^{r+1}r^k}{r!}\sum_s \binom{k}{s}r^{-s}\\
&= \sum_s \binom{k}{s}\cdot e^{-t}\sum_r \frac{t^{r+1}r^{k-s}}{r!}\\
&= t\sum_s \binom{k}{s}M_{k-s}
\end{aligned}$$

Regarding the initial values, the first moment of p_t is given by:

$$M_1 = e^{-t}\sum_r \frac{t^r r}{r!} = e^{-t}\sum_r \frac{t^r}{(r-1)!} = t$$

Now by using the above recurrence we obtain from this:

$$M_2 = t\sum_s \binom{1}{s}M_{k-s} = t(1+t) = t+t^2$$

On the other hand, some standard combinatorics, a bit as before at $t=1$, shows that the numbers in the statement $S_k = \sum_{\pi\in P(k)} t^{|\pi|}$ satisfy the same recurrence relation, and with the same initial values. Thus we have $M_k = S_k$, as claimed. □

12e. Exercises

Exercises:

Exercise 12.37.

Exercise 12.38.

Exercise 12.39.

Exercise 12.40.

Exercise 12.41.

Exercise 12.42.

Exercise 12.43.

Exercise 12.44.

Bonus exercise.

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

Space geometry, in that usual 3 dimensions that we live in. Many interesting things can be said here, in analogy with what we know from chapter 5 about triangles.

At a more advanced level, we can do some algebraic geometry in $\mathbb{R}^3$, in continuation to what we did before in $\mathbb{R}^2$. 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 13.1. *An algebraic curve in $\mathbb{R}^3$ is a curve as follows,*

$$C = \left\{(x, y, z) \in \mathbb{R}^3 \middle| P(x, y, z) = 0,\ Q(x, y, z) = 0\right\}$$

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. We will be back to this, later 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 13.2. *An algebraic surface in $\mathbb{R}^3$ is a surface as follows,*

$$S = \left\{(x, y, z) \in \mathbb{R}^3 \middle| P(x, y, z) = 0\right\}$$

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 $C \subset \mathbb{R}^2$ in chapter 8, we have the following result:

THEOREM 13.3. *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 explicitely listed.

PROOF. We will be quite brief here, because we intend to rediscuss all this in a moment, with full 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 13.4. *An algebraic hypersurface in $\mathbb{R}^N$ is a space of the form*

$$S = \left\{(x_1, \ldots, x_N) \in \mathbb{R}^N \middle| P(x_1, \ldots, x_N) = 0, \forall i\right\}$$

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 13.5. *An algebraic manifold in $\mathbb{R}^N$ is a space of the form*

$$X = \left\{(x_1, \ldots, x_N) \in \mathbb{R}^N \middle| P_i(x_1, \ldots, x_N) = 0, \forall i\right\}$$

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:

$$X = S_1 \cap \ldots \cap S_r$$

There is actually a bit of a discussion needed here, regarding the parameter $r \in \mathbb{N}$, shall we allow this parameter to be $r = \infty$ too, or not. We will discuss this later, with some algebra helping, the idea being that allowing $r = \infty$ forces in fact $r < \infty$.

As an announcement now, good news, what we have in Definition 13.5 is the good and final notion of algebraic manifold, very general, and with the branch of mathematics studying such manifolds being called algebraic geometry. In what follows we will discuss a bit what can be done with this, as a continuation of our previous work on the plane curves, at the elementary level. All this will lead us into the conclusion that we must first develop commutative algebra, and come back to algebraic geometry afterwards.

Let us first look more in detail at the hypersurfaces. We have here:

THEOREM 13.6. *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 $x_i x_j$ 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. □

Regarding now the examples of hypersurfaces $S \subset \mathbb{R}^N$, or of more general algebraic manifolds $X \subset \mathbb{R}^N$, there are countless of them, and it is impossible to have some discussion started here, without being subjective. The unit sphere $S_{\mathbb{R}}^{N-1} \subset \mathbb{R}^N$ gets of course the crown from everyone, as being the most important manifold after $\mathbb{R}^N$ itself. But then, passed this sphere, things ramify, depending on what exact applications of algebraic geometry you have in mind. In what concerns me, here is my next favorite example:

THEOREM 13.7. *The invertible matrices $A \in M_N(\mathbb{R})$ lie outside the hypersurface*

$$\det A = 0$$

and are therefore dense, in the space of all matrices $M_N(\mathbb{R})$.

PROOF. This is something self-explanatory, but with this result being some key in linear algebra, all this is worth a detailed discussion, as follows:

(1) We certainly know from basic linear algebra that a matrix $A \in M_N(\mathbb{R})$ is invertible precisely when it has nonzero determinant, $\det A \neq 0$. Thus, the invertible matrices $A \in M_N(\mathbb{R})$ are located precisely in the complement of the following space:

$$S = \left\{ A \in M_N(\mathbb{R}) \middle| \det A = 0 \right\}$$

(2) We also know from basic linear algebra, or perhaps not so basic linear algebra, that the determinant $\det A$ is a certain polynomial in the entries of A, of degree N:

$$\det \in \mathbb{R}[X_{11}, \ldots, X_{NN}]$$

(3) We conclude from this that the above set S is a degree N algebraic hypersurface in our sense, in the Euclidean space $M_N(\mathbb{R}) \simeq \mathbb{R}^n$, with $n = N^2$.

(4) Now since the complements of non-trivial hypersurfaces $S \subset \mathbb{R}^n$ are obviously dense, and if needing a formal proof here, for our above hypersurface S this is clear,

simply by suitably perturbing the matrix, and in general do not worry, we will be back to this, with full details, we are led to the conclusions in the statement. □

As an illustration for the power of our density result, we have:

THEOREM 13.8. *Given two matrices $A, B \in M_N(\mathbb{R})$, their products*

$$AB, BA \in M_N(\mathbb{R})$$

have the same characteristic polynomial, $P_{AB} = P_{BA}$.

PROOF. This is something quite hard to prove with bare hands, but we can trick by using Theorem 13.7. Indeed, it follows from definitions that the characteristic polynomial of a matrix is invariant under conjugation, in the sense that we have:

$$P_C = P_{ACA^{-1}}$$

Now observe that, when assuming that A is invertible, we have:

$$AB = A(BA)A^{-1}$$

Thus, we obtain the following formula, in the case where A is invertible:

$$P_{AB} = P_{BA}$$

Now by using the density result from Theorem 13.7, we conclude that this formula holds in fact for any matrix A, by continuity, as desired. □

Summarizing, we have some algebraic geometry theory going on, with applications, at least to questions in linear algebra, and presumably in calculus too. Getting back now to the basics, it is in fact possible to do even more generally, as follows:

DEFINITION 13.9. *An algebraic manifold over a field F is a space of the form*

$$X = \left\{(x_1, \ldots, x_N) \in F^N \middle| P_i(x_1, \ldots, x_N) = 0, \forall i\right\}$$

with $P_i \in F[x_1, \ldots, x_N]$ being a family of polynomials.

This might seem a bit abstract, but as a first observation, recall that $F = \mathbb{C}$ is a field too, on par with $F = \mathbb{R}$, and even better than it, in certain contexts. For instance quantum mechanics naturally lives over $F = \mathbb{C}$, instead of our usual $F = \mathbb{R}$. Also, in relation with questions in linear algebra, a matrix $A \in M_N(\mathbb{R})$ is much better viewed as matrix $A \in M_N(\mathbb{C})$, because here it has all N eigenvalues, when counted with multiplicities.

In fact, based on this linear algebra observation, and as our first result in complex algebraic geometry, we can improve Theorem 13.8, as follows:

THEOREM 13.10. *Given two matrices $A, B \in M_N(\mathbb{C})$, their products*

$$AB, BA \in M_N(\mathbb{C})$$

have the same eigenvalues, with the same multiplicities.

PROOF. To start with, Theorem 13.7 holds over $\mathbb{C}$ too, with the invertible matrices $A \in M_N(\mathbb{C})$ being dense, as being complementary to the following hypersurface:

$$\det A = 0$$

But with this in hand, the trick from the proof of Theorem 13.8 applies, and gives:

$$P_{AB} = P_{BA}$$

But this gives the result, because in the complex matrix setting the characteristic polynomial P encodes the eigenvalues, with multiplicities. □

This was for a first result in complex algebraic geometry, perhaps a bit advanced. At the level of more elementary things, the first thought goes to the plane algebraic curves, in a complex sense. But, surprise here, these are the spaces as follows:

$$C = \left\{(x, y) \in \mathbb{C}^2 \middle| P(x, y) = 0\right\}$$

Now when looking at this formula, we realize that our curve $C \subset \mathbb{C}^2$ is in fact something quite complicated, corresponding to a 2-dimensional surface $X \subset \mathbb{R}^4$. But, no worries, we will come back to this regularly. In fact, in what follows, we will be jointly developing our theory over both $F = \mathbb{R}$ and $F = \mathbb{C}$, with such questions in mind.

Getting back now to Definition 13.9 as stated, what about other fields F? Good question, and in answer, I would have a quick exercise for you, as follows:

EXERCISE 13.11. *Prove that for $n \geq 3$ the following curve,*

$$x^n + y^n = 1$$

has no non-trivial points, $x, y \neq 0$, over $F = \mathbb{Q}$.

Such ideas are very old, going back to the ancient Greeks, and there are many things that can be said about algebraic geometry in its "arithmetic" version, over arbitrary fields F as above. In fact, this is a point where algebraic geometry really shines, with many known advanced results in number theory having been obtained in this way.

Also, importantly, such arithmetic questions have been a strong motivation for the development of modern algebraic geometry, starting from the 1950s, by Grothendieck and others, basically over arbitrary fields F, but with important consequences regarding $F = \mathbb{R}$ and $F = \mathbb{C}$ too. But more on Grothendieck and modern geometry later, once we will know a bit more about algebraic manifolds, and their basic algebraic theory.

13b. Regular polyhedra

Switching topics now, let us first discuss, still in relation with space geometry questions, the graphs. As a fundamental result about them, we have:

THEOREM 13.12. *For a connected planar graph we have the Euler formula*

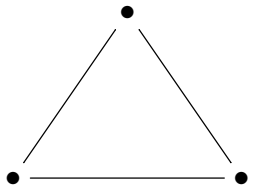
$$v - e + f = 2$$

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.

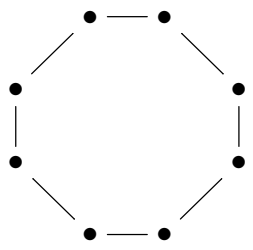
(2) As a first illustration for how this formula works, consider a triangle:



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:

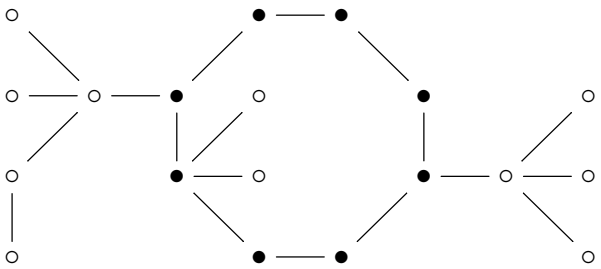
$$N - N + 2 = 2$$

(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:

$$N - (N - 1) + 1 = 2$$

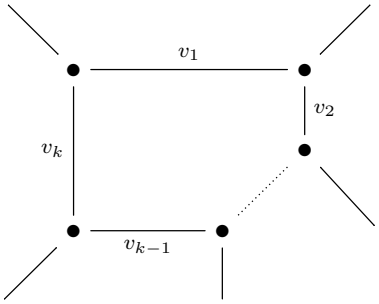
(5) At $f = 2$ now, our graph must be an N-gon as above, but with some trees allowed to grow from the vertices, with an illustrating example here being as follows:



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:

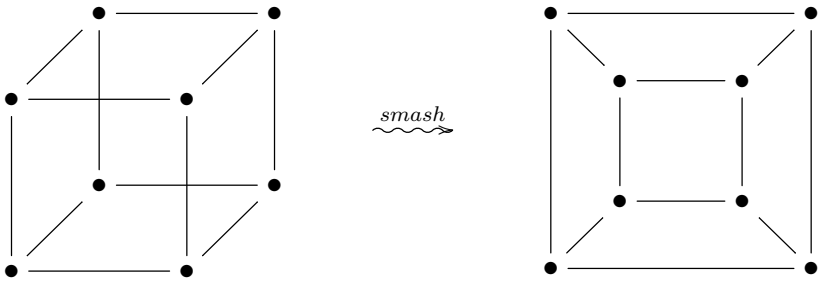
PROPOSITION 13.13. *For a convex polyhedron we have the Euler formula*

$$v-e+f=2$$

with v,e,f being the number of vertices, edges and faces.

PROOF. This is more or less the same thing as Theorem 13.12, save for getting rid of the internal trees of the planar graph there, the idea being as follows:

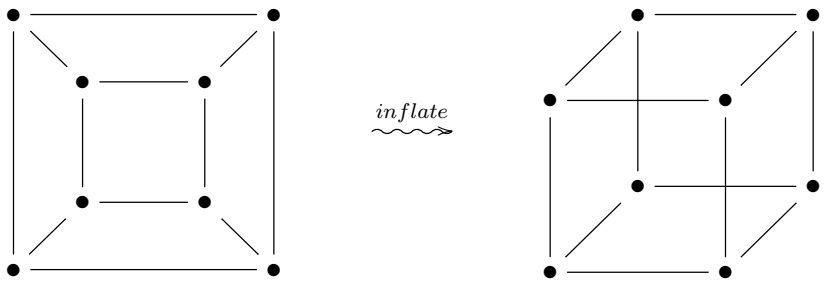
(1) In one sense, consider a convex polyhedron P. We can then enlarge one face, as much as needed, and then smash our polyhedron with a big hammer, as to get a planar graph X. As an illustration, here is how this method works, for a cube:



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 13.12.

(2) Conversely, consider a connected planar graph X. Then, save for getting rid of the internal trees, as explained in the proof of Theorem 13.12, we can assume that we are dealing with an agglomeration of N-gons, again as explained in the proof of Theorem

13.12. 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. □

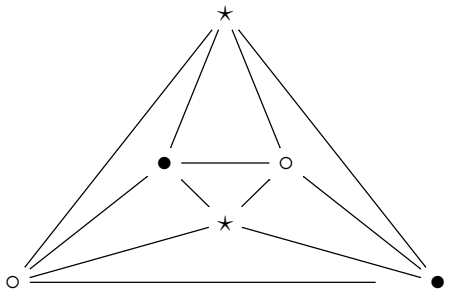
Summarizing, Euler formula understood, but as a matter of making sure that we didn't mess up anything with our mathematics, let us do some direct checks as well:

PROPOSITION 13.14. *The Euler formula $v-e+f=2$ holds indeed for the five possible regular polyhedra, as follows:*

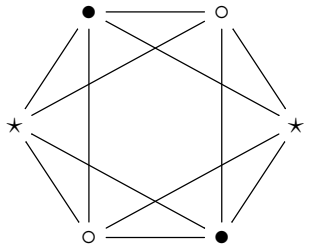
(1) *Tetrahedron:* $4-6+4=2$.
(2) *Cube:* $8-12+6=2$.
(3) *Octahedron:* $6-12+8=2$.
(4) *Dodecahedron:* $20-30+12=2$.
(5) *Isocahedron:* $12-30+20=2$.

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 13.13 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. □

The Euler formula $v - e + f = 2$, in both its above formulations, the graph one from Theorem 13.12, and the polyhedron one from Proposition 13.13, is something very interesting, at the origin of modern pure mathematics, and having countless other versions and generalizations. We will be back to it on several occasions, in what follows.

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 the basic physics works.

We will be talking here about all sorts of mechanics, all taking place in 3D.

However, before getting started with some physics and applications, we will need one more mathematical notion, which is something 3D specific, as follows:

DEFINITION 13.15. *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 13.16. *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 13.17. *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:

$$\begin{aligned} V &= \dot{x}_1 u_1 + \dot{x}_2 u_2 + \dot{x}_3 u_3 + x_1 \dot{u}_1 + x_2 \dot{u}_2 + x_3 \dot{u}_3 \\ &= \dot{x} + (x_1 \cdot \omega \times u_1 + x_2 \cdot \omega \times u_2 + x_3 \cdot \omega \times u_3) \\ &= \dot{x} + w \times (x_1 u_1 + x_2 u_2 + x_3 u_3) \\ &= \dot{x} + \omega \times x \end{aligned}$$

Thus, we are led to the conclusions in the statement. □

In what regards now the acceleration, the result, which is famous, is as follows:

THEOREM 13.18. *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 13.17, 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 13.19. *The acceleration of an object m subject to a force F is given by*

$$ma = F - mg - 2m\omega \times v - m\omega \times (\omega \times x)$$

with g pointing upwards, and with the last terms being the Coriolis and centrifugal forces.

PROOF. This follows indeed from the above discussion, by assuming that the acceleration A there comes from the combined effect of a force F, and of the usual g. □

We refer to any standard undergraduate mechanics book, such as Feynman [33], Kibble [57] or Taylor [91] for more on the above, including various numerics on what happens here on Earth, the Foucault pendulum, history of all this, and many other things. Let us just mention here, as a basic illustration for all this, that a rock dropped from 100m deviates about 1cm from its intended target, due to the formula in Theorem 13.19.

Switching topics, let us talk now about electrodynamics, and the Maxwell equations:

THEOREM 13.20. *Electrodynamics is governed by the formulae*

$$< \nabla, E >= \frac{\rho}{\varepsilon_0} \quad , \quad < \nabla, B >= 0$$

$$\nabla \times E = -\dot{B} \quad , \quad \nabla \times B = \mu_0 J + \mu_0 \varepsilon_0 \dot{E}$$

called Maxwell equations.

PROOF. This is something fundamental, appearing as a tricky mixture of physics facts and mathematical results, the idea being as follows:

(1) To start with, electrodynamics is the science of moving electrical charges. And this is something quite complicated, because unlike in classical mechanics, where the Newton law is good for both the static and the dynamic setting, the Coulomb law, which is actually very similar to the Newton law, does the job when the charges are static, but no longer describes well the situation when the charges are moving.

(2) The problem comes from the fact that moving charges produce magnetism, and with this being visible when putting together two electric wires, which will attract or repel,

depending on orientation. Thus, in contrast with classical mechanics, where static or dynamic problems are described by a unique field, the gravitational one, in electrodynamics we have two fields, namely the electric field E, and the magnetic field B.

(3) Fortunately, there is a full set of equations relating the electric field E and the magnetic field B, those above. Regarding the math, the dots denote derivatives with respect to time, and ∇ is the gradient operator, or space derivative, given by:

$$\nabla = \begin{pmatrix} \frac{d}{dx} \\ \frac{d}{dy} \\ \frac{d}{dz} \end{pmatrix}$$

(4) Regarding the physics, the first formula is the Gauss law, ρ being the charge, and ε_0 being a constant, and with this Gauss law more or less replacing the Coulomb law from electrostatics. The second formula is something basic, and anonymous. The third formula is the Faraday law. As for the fourth formula, this is the Ampère law, as modified by Maxwell, with J being the volume current density, and μ_0 being a constant. □

Quite surprisingly, the constants μ_0, ε_0 appearing above are related as follows:

FACT 13.21. *The constants μ_0, ε_0 are related by the Biot-Savart formula*

$$\mu_0\varepsilon_0 = \frac{1}{c^2}$$

with $c = 299,792,458$ being as usual the speed of light in vacuum.

The point indeed is that, in relation with the Maxwell equations, we have:

THEOREM 13.22. *In regions of space where there is no charge or current present the Maxwell equations for electrodynamics read*

$$< \nabla, E >=< \nabla, B >= 0$$

$$\nabla \times E = -\dot{B} \quad , \quad \nabla \times B = \dot{E}/c^2$$

and both the electric field E and magnetic field B are subject to the wave equation

$$\ddot{\varphi} = c^2\Delta\varphi$$

where $\Delta = \sum_i d^2/dx_i^2$ is the Laplace operator, and c is the speed of light.

PROOF. Under the circumstances in the statement, namely no charge or current present, the Maxwell equations from Theorem 13.20 simply read:

$$< \nabla, E >=< \nabla, B >= 0$$

$$\nabla \times E = -\dot{B} \quad , \quad \nabla \times B = \dot{E}/c^2$$

Now by applying the curl operator to the third equation, we obtain:

$$\begin{aligned}\nabla\times(\nabla\times E) &= -\nabla\times\dot{B}\\ &= -(\nabla\times B)'\\ &= -\ddot{E}/c^2\end{aligned}$$

Also, by applying the curl operator to the fourth equation, we obtain:

$$\begin{aligned}\nabla\times(\nabla\times B) &= \nabla\times\dot{E}/c^2\\ &= (\nabla\times E)'/c^2\\ &= -\ddot{B}/c^2\end{aligned}$$

But the double curl operator is subject to the following formula:

$$\nabla\times(\nabla\times\varphi)=\nabla<\nabla,\varphi>-\Delta\varphi$$

Now by using the first two equations, we are led to the conclusion in the statement. □

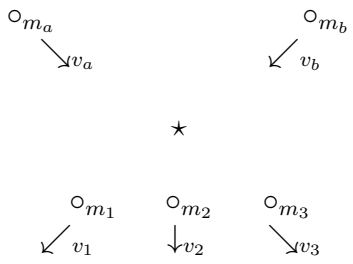
So, what is light? Light is the wave predicted by Theorem 13.22, traveling in vacuum at the maximum possible speed, c, and with an important extra property being that it depends on a real positive parameter, that can be called, upon taste, frequency, wavelength, or color. And in what regards the creation of light, the mechanism here is as follows:

FACT 13.23. *An accelerating or decelerating charge produces electromagnetic radiation, called light, whose frequency and wavelength can be explicitly computed.*

This phenomenon can be observed is a variety of situations, such as the usual light bulbs, where electrons get decelerated by the filament, acting as a resistor, or in usual fire, which is a chemical reaction, with the electrons moving around, as they do in any chemical reaction, or in more complicated machinery like nuclear plants, particle accelerators, and so on, leading there to all sorts of eerie glows, of various colors.

13d. Solid angles

Let us talk now about interactions between particles. But here, we have some experience from classical mechanics, with the typical picture of what can happen being:



This was for basic interactions in classical mechanics. In our present setting, particle physics, things are a bit more complicated than this, due to a variety of reasons, and experimental physics suggests looking at two main types of interactions, as follows:

FACT 13.24. *In particle physics, we have two main types of interactions, namely:*

(1) *Decay. This is when a particle decomposes, as a result of whatever internal mechanism, into a sum of other particles,* $*_0 \to *_1 + \ldots + *_n$.
(2) *Scattering. This is when two particles meet, by colliding, or almost, and combine and decompose into a sum of other particles,* $*_a + *_b \to *_1 + \ldots + *_n$.

Obviously, all this departs a bit from our classical mechanics knowledge, as explained above, and several comments are in order here, as follows:

(1) In what regards decay, something that we talked a lot about, when doing thermodynamics, and then quantum mechanics, is an electron of an atom changing its energy level, and emitting a photon. But this can be regarded as being decay.

(2) As for scattering, the simplest example here appears again from an electron of an atom, changing its energy level, but this time by absorbing a photon. Of course, there are many other possible examples, such as the electron-positron annihilation.

Getting to work for good now, decay and its mathematics. Ignoring the physics, this is basically a matter of probability and statistics, and the basics here are as follows:

THEOREM 13.25. *In the context of decay, the quantity to look at is the decay rate* λ*, which is the probability per unit time that the particle will disintegrate. With this:*

(1) *The number of particles remaining at time* $t > 0$ *is* $N_t = e^{-\lambda t} N_0$.
(2) *The mean lifetime of a particle is* $\tau = 1/\lambda$.
(3) *The half-life of the substance is* $t_{1/2} = (\log 2)/\lambda$.

PROOF. As said above, this is basic probability, as follows:

(1) In mathematical terms, our definition of the decay rate reads:

$$\frac{dN}{dt} = -\lambda N$$

By integrating, we are led to the formula in the statement, namely:

$$N_t = e^{-\lambda t} N_0$$

(2) Let us first convert what we have into a probability law. We have:

$$\int_0^\infty N_t dt = \int_0^\infty N_0 e^{-\lambda t} dt = \frac{N_0}{\lambda}$$

Thus, the density of the probability decay function is given by:

$$f(t) = \frac{\lambda}{N_0} \cdot N_0 e^{-\lambda t} = \lambda e^{-\lambda t}$$

We can now compute the mean lifetime, by integrating by parts, as follows:

$$\begin{aligned}
\tau &= < t > \\
&= \int_0^\infty t f(t) dt \\
&= \int_0^\infty \lambda t e^{-\lambda t} dt \\
&= \int_0^\infty t(-e^{-\lambda t})' dt \\
&= \int_0^\infty e^{-\lambda t} dt \\
&= \frac{1}{\lambda}
\end{aligned}$$

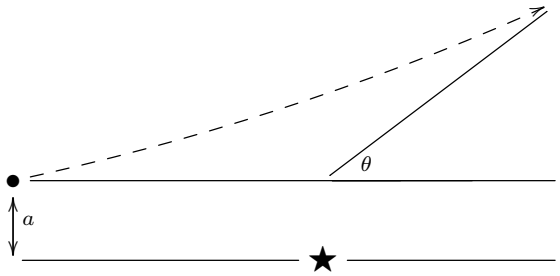
(3) Finally, regarding the half-life, this is by definition the time $t_{1/2}$ required for the decaying quantity to fall to one-half of its initial value. Mathematically, this means:

$$N_t = 2^{-\frac{t}{t_{1/2}}} N_0$$

Now by comparing with $N_t = e^{-\lambda t} N_0$, this gives $t_{1/2} = (\log 2)/\lambda$, as stated. □

Getting now to scattering, this is something far more familiar, because we can fully use here our experience from classical mechanics. Let us start with:

DEFINITION 13.26. *The generic picture of scattering is as follows,*



with $a \geq 0$ being the impact parameter, and $\theta \in [0, \pi]$ being the scattering angle.

In other words, we assume here that the particle misses its target by $a \geq 0$, with the limiting case $a = 0$ corresponding of course to exactly hitting the target, and we are interested in computing the scattering angle $\theta \in [0, \pi]$ as a function $\theta = \theta(a)$.

Many things can be said here, and more on this in a moment, but as an answer to a question that you might certainly have, we are interested in $a > 0$ because this is what happens in particle physics, there is no need for exactly hitting the target for having a collision-type interaction. By the case, the limiting case $a = 0$ is rather unwanted in the context of our scaterring question, because by symmetry this would normally force the scattering angle to be $\theta = 0$ or $\theta = \pi$, which does not look very interesting.

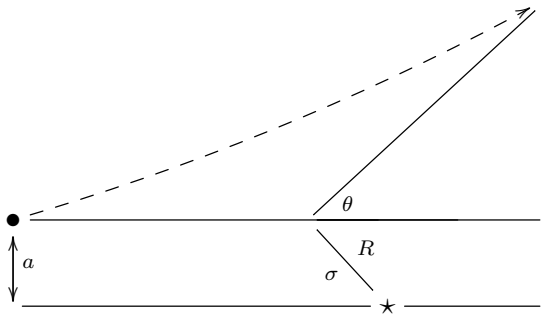
But probably too much talking, let us do a computation. We have here:

PROPOSITION 13.27. *In the context of classical particle colliding elastically with a hard sphere of radius $R > 0$, we have the formula*

$$a = R \cos \frac{\theta}{2}$$

and so the scattering angle is given by $\theta = 2 \arccos(a/R)$.

PROOF. In the context from the statement, which is all classical mechanics, and more specifically is a basic elastic collision, between a point particle and a hard sphere, if the impact factor is $a > R$, nothing happens. In the case $a \leq R$ we do have an impact, and a bounce of our particle on the hard sphere, the picture of the event being as follows:



Here the sphere is missing, due to budget cuts, with only its center $\star$ being pictured, but you get the point. Now with σ being the angle in the statement, we have the following two formulae, with the first one being clear on the above picture, and with the second one coming from the fact that, at the rebound, the various angles must sum up to π:

$$a = R \sin \sigma \quad , \quad 2\sigma + \theta = \pi$$

We deduce that the impact factor is given by the following formula:

$$a = R \sin \left(\frac{\pi}{2} - \frac{\theta}{2} \right) = R \cos \frac{\theta}{2}$$

Thus, we are led to the conclusions in the statement. □

With this understood, let us try to make something more 3D, and statistical, out of this. We can indeed further build on Definition 13.26, as follows:

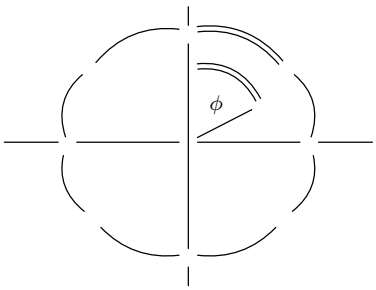
DEFINITION 13.28. *In the general context of scattering, we can:*

(1) *Extend our length/angle correspondence $a \to \theta$ into an infinitesimal area/solid angle correspondence $d\sigma \to d\Omega$.*
(2) *Talk about the inverse derivative $D(\theta)$ of this correspondence, called differential cross section, according to the formula $d\sigma = D(\theta)d\Omega$.*
(3) *And finally, define the total cross section of the scattering event as being the quantity $\sigma = \int d\sigma = \int D(\theta)d\Omega$.*

And good news, the notion of total cross section σ, as constructed above, is the one that we will need, in what follows, with this being to scattering something a bit similar to what the decay rate λ was to decay, that is, the main quantity to look at.

In order to understand how the cross section works, we have:

PROPOSITION 13.29. *Assuming that the incoming beam comes as follows,*



subtending a certain angle ϕ, the differential cross section is given by

$$D(\theta) = \left| \frac{a}{\sin\theta} \cdot \frac{da}{d\theta} \right|$$

and the total cross section is given by $\sigma = \int D(\theta)d\Omega$.

PROOF. Assume indeed that we have a uniform beam as the one pictured in the statement, enclosed by the double lines appearing there, and with the need for a beam instead of a single particle coming from what we do in Definition 13.28, which is rather of continuous nature. Our claim is that we have the following formulae:

$$d\sigma = |a \cdot da \cdot d\phi| \quad , \quad d\Omega = |\sin\theta \cdot d\theta \cdot d\phi|$$

Indeed, the first formula, at departure, is clear from the picture above, and the second formula is clear from a similar picture at the arrival. Now with these formulae in hand,

by dividing them, we obtain the following formula for the differential cross section:

$$\begin{aligned} D(\theta) &= \frac{d\sigma}{d\Omega} \\ &= \left|\frac{a\cdot da\cdot d\phi}{\sin\theta\cdot d\theta\cdot d\phi}\right| \\ &= \left|\frac{a}{\sin\theta}\cdot\frac{da}{d\theta}\right| \end{aligned}$$

As for the total cross section, this is given as usual by $\sigma=\int D(\theta)d\Omega$. □

As an illustration for this, in the case of a hard sphere scattering, we have:

THEOREM 13.30. *In the case of a hard sphere scattering, the cross section is*

$$\sigma=\pi R^2$$

with $R>0$ being the radius of the sphere.

PROOF. We know from Proposition 13.27 that, with the notations there, we have:

$$a=R\cos\frac{\theta}{2}$$

At the level of the corresponding differentials, this gives the following formula:

$$\frac{da}{d\theta}=-\frac{R}{2}\sin\frac{\theta}{2}$$

We can now compute the differential cross section, as above, and we obtain:

$$\begin{aligned} D(\theta) &= \left|\frac{a}{\sin\theta}\cdot\frac{da}{d\theta}\right| \\ &= \frac{R\cos(\theta/2)}{\sin\theta}\cdot\frac{R\sin(\theta/2)}{2} \\ &= \frac{R^2(\sin\theta)/2}{2\sin\theta} \\ &= \frac{R^2}{4} \end{aligned}$$

Now by integrating, we obtain from this, via some calculus, the following formula:

$$\sigma=\int\frac{R^2}{4}\,d\Omega=\pi R^2$$

Thus, we are led to the conclusion in the statement. □

13e. Exercises

Exercises:

Exercise 13.31.

Exercise 13.32.

Exercise 13.33.

Exercise 13.34.

Exercise 13.35.

Exercise 13.36.

Exercise 13.37.

Exercise 13.38.

Bonus exercise.

CHAPTER 14

Vector calculus

14a. Matrices, rotations

Welcome to advanced vector calculus, and hang on, tough algebra to come. At a more advanced level, indeed, we can try to understand the various transformations of $\mathbb{R}^N$. Let us start with the plane. The transformations that we are interested in are as follows:

DEFINITION 14.1. *A map $f : \mathbb{R}^2 \to \mathbb{R}^2$ is called affine when it maps lines to lines,*

$$f(tx + (1-t)y) = tf(x) + (1-t)f(y)$$

for any $x, y \in \mathbb{R}^2$ and any $t \in \mathbb{R}$. If in addition $f(0) = 0$, we call f linear.

As a first observation, our "maps lines to lines" interpretation of the equation in the statement assumes that the points are degenerate lines, and this in order for our interpretation to work when $x = y$, or when $f(x) = f(y)$. Also, what we call line is not exactly a set, but rather a dynamic object, think trajectory of a point on that line. We will be back to this later, once we will know more about such maps.

Here are some basic examples of symmetries, all being linear in the above sense:

PROPOSITION 14.2. *The symmetries with respect to Ox and Oy are:*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x \\ -y \end{pmatrix} \quad , \quad \begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} -x \\ y \end{pmatrix}$$

The symmetries with respect to the $x = y$ and $x = -y$ diagonals are:

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} y \\ x \end{pmatrix} \quad , \quad \begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} -y \\ -x \end{pmatrix}$$

All these maps are linear, in the above sense.

PROOF. The fact that all these maps are linear is clear, because they map lines to lines, in our sense, and they also map 0 to 0. As for the explicit formulae in the statement, these are clear as well, by drawing pictures for each of the maps involved. □

Here are now some basic examples of rotations, once again all being linear:

PROPOSITION 14.3. *The rotations of angle* 0° *and of angle* 90° *are:*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x \\ y \end{pmatrix} \quad , \quad \begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} -y \\ x \end{pmatrix}$$

The rotations of angle 180° *and of angle* 270° *are:*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} -x \\ -y \end{pmatrix} \quad , \quad \begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} y \\ -x \end{pmatrix}$$

All these maps are linear, in the above sense.

PROOF. As before, these rotations are all linear, for obvious reasons. As for the formulae in the statement, these are clear as well, by drawing pictures. □

Here are some basic examples of projections, once again all being linear:

PROPOSITION 14.4. *The projections on* Ox *and* Oy *are:*

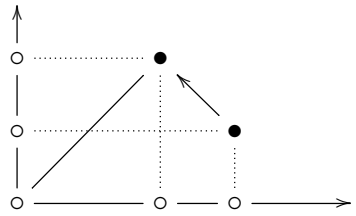
$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x \\ 0 \end{pmatrix} \quad , \quad \begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} 0 \\ y \end{pmatrix}$$

The projections on the $x = y$ *and* $x = -y$ *diagonals are:*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \frac{1}{2}\begin{pmatrix} x+y \\ x+y \end{pmatrix} \quad , \quad \begin{pmatrix} x \\ y \end{pmatrix} \to \frac{1}{2}\begin{pmatrix} x-y \\ y-x \end{pmatrix}$$

All these maps are linear, in the above sense.

PROOF. Again, these projections are all linear, and the formulae are clear as well, by drawing pictures, with only the last 2 formulae needing some explanations. In what regards the projection on the $x = y$ diagonal, the picture here is as follows:



But this gives the result, since the 45° triangle shows that this projection leaves invariant $x + y$, so we can only end up with the average $(x + y)/2$, as double coordinate. As for the projection on the $x = -y$ diagonal, the proof here is similar. □

Finally, we have the translations, which are as follows:

PROPOSITION 14.5. *The translations are exactly the maps of the form*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x+p \\ y+q \end{pmatrix}$$

with $p, q \in \mathbb{R}$*, and these maps are all affine, in the above sense.*

PROOF. A translation $f : \mathbb{R}^2 \to \mathbb{R}^2$ is clearly affine, because it maps lines to lines. Also, such a translation is uniquely determined by the following vector:

$$f\begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}p\\q\end{pmatrix}$$

To be more precise, f must be the map which takes a vector $\binom{x}{y}$, and adds this vector $\binom{p}{q}$ to it. But this gives the formula in the statement. □

Summarizing, we have many interesting examples of linear and affine maps. Let us develop now some general theory, for such maps. As a first result, we have:

THEOREM 14.6. *For a map $f : \mathbb{R}^2 \to \mathbb{R}^2$, the following are equivalent:*

(1) *f is linear in our sense, mapping lines to lines, and 0 to 0.*
(2) *f maps sums to sums, $f(x+y) = f(x) + f(y)$, and satisfies $f(\lambda x) = \lambda f(x)$.*

PROOF. This is something which comes from definitions, as follows:

(1) $\implies$ (2) We know that f satisfies the following equation, and $f(0) = 0$:

$$f(tx + (1-t)y) = tf(x) + (1-t)f(y)$$

By setting $y = 0$, and by using our assumption $f(0) = 0$, we obtain, as desired:

$$f(tx) = tf(x)$$

As for the first condition, regarding sums, this can be established as follows:

$$\begin{aligned} f(x+y) &= f\left(2 \cdot \frac{x+y}{2}\right) \\ &= 2f\left(\frac{x+y}{2}\right) \\ &= 2 \cdot \frac{f(x) + f(y)}{2} \\ &= f(x) + f(y) \end{aligned}$$

(2) $\implies$ (1) Conversely now, assuming that f satisfies $f(x+y) = f(x) + f(y)$ and $f(\lambda x) = \lambda f(x)$, then f must map lines to lines, as shown by:

$$\begin{aligned} f(tx + (1-t)y) &= f(tx) + f((1-t)y) \\ &= tf(x) + (1-t)f(y) \end{aligned}$$

Also, we have $f(0) = f(2 \cdot 0) = 2f(0)$, which gives $f(0) = 0$, as desired. □

The above result is very useful, and in practice, we will often use the condition (2) there, somewhat as a new definition for the linear maps.

Let us record this finding as an upgrade of our formalism, as follows:

DEFINITION 14.7 (upgrade). *A map $f : \mathbb{R}^2 \to \mathbb{R}^2$ is called:*

(1) *Linear, when it satisfies $f(x + y) = f(x) + f(y)$ and $f(\lambda x) = \lambda f(x)$.*
(2) *Affine, when it is of the form $f = g + x$, with g linear, and $x \in \mathbb{R}^2$.*

Before getting into the mathematics of linear maps, let us comment a bit more on the "maps lines to lines" feature of such maps. As mentioned after Definition 14.1, this feature requires thinking at lines as being "dynamic" objects, the point being that, when thinking at lines as being sets, this interpretation fails, as shown by the following map:

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x^3 \\ 0 \end{pmatrix}$$

However, in relation with all this we have the following useful result:

THEOREM 14.8. *For a continuous injective $f : \mathbb{R}^2 \to \mathbb{R}^2$, the following are equivalent:*

(1) *f is affine in our sense, mapping lines to lines.*
(2) *f maps set-theoretical lines to set-theoretical lines.*

PROOF. By composing f with a translation, we can assume that we have $f(0) = 0$. With this assumption made, the proof goes as follows:

(1) $\implies$ (2) This is clear from definitions.

(2) $\implies$ (1) Let us first prove that we have $f(x + y) = f(x) + f(y)$. We do this first in the case where our vectors are not proportional, $x \not\sim y$. In this case we have a proper parallelogram $(0, x, y, x+y)$, and since f was assumed to be injective, it must map parallel lines to parallel lines, and so must map our parallelogram into a parallelogram $(0, f(x), f(y), f(x + y))$. But this latter parallelogram shows that we have:

$$f(x + y) = f(x) + f(y)$$

In the remaining case where our vectors are proportional, $x \sim y$, we can pick a sequence $x_n \to x$ satisfying $x_n \not\sim y$ for any n, and we obtain, as desired:

$$\begin{aligned} x_n \to x, x_n \not\sim y, \forall n &\implies f(x_n + y) = f(x_n) + f(y), \forall n \\ &\implies f(x + y) = f(x) + f(y) \end{aligned}$$

Regarding now $f(\lambda x) = \lambda f(x)$, since f maps lines to lines, it must map the line $0 - x$ to the line $0 - f(x)$, so we have a formula as follows, for any λ, x:

$$f(\lambda x) = \varphi_x(\lambda) f(x)$$

But since f maps parallel lines to parallel lines, by Thales the function $\varphi_x : \mathbb{R} \to \mathbb{R}$ does not depend on x. Thus, we have a formula as follows, for any λ, x:

$$f(\lambda x) = \varphi(\lambda) f(x)$$

We know that we have $\varphi(0) = 0$ and $\varphi(1) = 1$, and we must prove that we have $\varphi(\lambda) = \lambda$ for any λ. For this purpose, we use a trick. On one hand, we have:

$$f((\lambda + \mu)x) = \varphi(\lambda + \mu)f(x)$$

On the other hand, since f maps sums to sums, we have as well:

$$\begin{aligned} f((\lambda + \mu)x) &= f(\lambda x) + f(\mu x) \\ &= \varphi(\lambda)f(x) + \varphi(\mu)f(x) \\ &= (\varphi(\lambda) + \varphi(\mu))f(x) \end{aligned}$$

Thus our rescaling function $\varphi : \mathbb{R} \to \mathbb{R}$ satisfies the following conditions:

$$\varphi(0) = 0 \quad , \quad \varphi(1) = 1 \quad , \quad \varphi(\lambda + \mu) = \varphi(\lambda) + \varphi(\mu)$$

But with these conditions in hand, it is clear that we have $\varphi(\lambda) = \lambda$, first for all the inverses of integers, $\lambda = 1/n$ with $n \in \mathbb{N}$, then for all rationals, $\lambda \in \mathbb{Q}$, and finally by continuity for all reals, $\lambda \in \mathbb{R}$. Thus, we have proved the following formula:

$$f(\lambda x) = \lambda f(x)$$

But this finishes the proof of (2) $\implies$ (1), and we are done. □

All this is very nice, and there are some further things that can be said, but getting to business, Definition 14.7 is what we need. Indeed, we have the following powerful result, stating that the linear/affine maps $f : \mathbb{R}^2 \to \mathbb{R}^2$ are fully described by 4/6 parameters:

THEOREM 14.9. *The linear maps $f : \mathbb{R}^2 \to \mathbb{R}^2$ are precisely the maps of type*

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix}$$

and the affine maps $f : \mathbb{R}^2 \to \mathbb{R}^2$ are precisely the maps of type

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix} + \begin{pmatrix} p \\ q \end{pmatrix}$$

with the conventions from Definition 14.7 for such maps.

PROOF. Assuming that f is linear in the sense of Definition 14.7, we have:

$$\begin{aligned} f\begin{pmatrix} x \\ y \end{pmatrix} &= f\left(\begin{pmatrix} x \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ y \end{pmatrix}\right) \\ &= f\begin{pmatrix} x \\ 0 \end{pmatrix} + f\begin{pmatrix} 0 \\ y \end{pmatrix} \\ &= f\left(x\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) + f\left(y\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) \\ &= xf\begin{pmatrix} 1 \\ 0 \end{pmatrix} + yf\begin{pmatrix} 0 \\ 1 \end{pmatrix} \end{aligned}$$

Thus, we obtain the formula in the statement, with $a, b, c, d \in \mathbb{R}$ being given by:

$$f\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}a\\c\end{pmatrix} \quad , \quad f\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}b\\d\end{pmatrix}$$

In the affine case now, we have as extra piece of data a vector, as follows:

$$f\begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}p\\q\end{pmatrix}$$

Indeed, if $f : \mathbb{R}^2 \to \mathbb{R}^2$ is affine, then the following map is linear:

$$f - \begin{pmatrix}p\\q\end{pmatrix} : \mathbb{R}^2 \to \mathbb{R}^2$$

Thus, by using the formula in (1) we obtain the result. □

Moving ahead now, Theorem 14.9 is all that we need for doing some non-trivial mathematics, and so in practice, that will be our "definition" for the linear and affine maps. In order to simplify now all that, which might be a bit complicated to memorize, the idea will be to put our parameters a, b, c, d into a matrix, in the following way:

DEFINITION 14.10. *A matrix $A \in M_2(\mathbb{R})$ is an array as follows:*

$$A = \begin{pmatrix}a & b\\c & d\end{pmatrix}$$

These matrices act on the vectors in the following way,

$$\begin{pmatrix}a & b\\c & d\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}ax+by\\cx+dy\end{pmatrix}$$

the rule being "multiply the rows of the matrix by the vector".

The above multiplication formula might seem a bit complicated, at a first glance, but it is not. Here is an example for it, quickly worked out:

$$\begin{pmatrix}1 & 2\\5 & 6\end{pmatrix}\begin{pmatrix}3\\1\end{pmatrix} = \begin{pmatrix}1\cdot 3+2\cdot 1\\5\cdot 3+6\cdot 1\end{pmatrix} = \begin{pmatrix}5\\21\end{pmatrix}$$

As already mentioned, all this comes from our findings from Theorem 14.9. Indeed, with the above multiplication convention for matrices and vectors, we can turn Theorem 14.9 into something much simpler, and better-looking, as follows:

THEOREM 14.11. *The linear maps $f : \mathbb{R}^2 \to \mathbb{R}^2$ are precisely the maps of type*

$$f(v) = Av$$

and the affine maps $f : \mathbb{R}^2 \to \mathbb{R}^2$ are precisely the maps of type

$$f(v) = Av + w$$

with A being a 2×2 matrix, and with $v, w \in \mathbb{R}^2$ being vectors, written vertically.

PROOF. With the above conventions, the formulae in Theorem 14.9 read:

$$f\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}a&b\\c&d\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

$$f\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}a&b\\c&d\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}+\begin{pmatrix}p\\q\end{pmatrix}$$

But these are exactly the formulae in the statement, with:

$$A=\begin{pmatrix}a&b\\c&d\end{pmatrix}\quad,\quad v=\begin{pmatrix}x\\y\end{pmatrix}\quad,\quad w=\begin{pmatrix}p\\q\end{pmatrix}$$

Thus, we have proved our theorem. □

Before going further, let us discuss some examples. First, we have:

PROPOSITION 14.12. *The symmetries with respect to Ox and Oy are given by:*

$$\begin{pmatrix}1&0\\0&-1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}\quad,\quad\begin{pmatrix}-1&0\\0&1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

The symmetries with respect to the $x=y$ and $x=-y$ diagonals are given by:

$$\begin{pmatrix}0&1\\1&0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}\quad,\quad\begin{pmatrix}0&-1\\-1&0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

PROOF. According to Proposition 14.2, the above transformations map $\binom{x}{y}$ to:

$$\begin{pmatrix}x\\-y\end{pmatrix}\quad,\quad\begin{pmatrix}-x\\y\end{pmatrix}\quad,\quad\begin{pmatrix}y\\x\end{pmatrix}\quad,\quad\begin{pmatrix}-y\\-x\end{pmatrix}$$

But this gives the formulae in the statement, by guessing in each case the matrix which does the job, in the obvious way. □

Regarding now the basic rotations, we have here:

PROPOSITION 14.13. *The rotations of angle 0° and of angle 90° are given by:*

$$\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}\quad,\quad\begin{pmatrix}0&-1\\1&0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

The rotations of angle 180° and of angle 270° are given by:

$$\begin{pmatrix}-1&0\\0&-1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}\quad,\quad\begin{pmatrix}0&1\\-1&0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

PROOF. As before, but by using Proposition 14.3, the vector $\binom{x}{y}$ maps to:

$$\begin{pmatrix}x\\y\end{pmatrix}\quad,\quad\begin{pmatrix}-y\\x\end{pmatrix}\quad,\quad\begin{pmatrix}-x\\-y\end{pmatrix}\quad,\quad\begin{pmatrix}y\\-x\end{pmatrix}$$

But this gives the formulae in the statement, as before by guessing the matrix. □

Finally, regarding the basic projections, we have here:

PROPOSITION 14.14. *The projections on Ox and Oy are given by:*

$$\begin{pmatrix}1&0\\0&0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}\quad,\quad\begin{pmatrix}0&0\\0&1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

The projections on the $x=y$ and $x=-y$ diagonals are given by:

$$\frac{1}{2}\begin{pmatrix}1&1\\1&1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}\quad,\quad\frac{1}{2}\begin{pmatrix}1&-1\\-1&1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

PROOF. As before, but according now to Proposition 14.4, the vector $\binom{x}{y}$ maps to:

$$\begin{pmatrix}x\\0\end{pmatrix}\quad,\quad\begin{pmatrix}0\\y\end{pmatrix}\quad,\quad\frac{1}{2}\begin{pmatrix}x+y\\x+y\end{pmatrix}\quad,\quad\frac{1}{2}\begin{pmatrix}x-y\\y-x\end{pmatrix}$$

But this gives the formulae in the statement, as before by guessing the matrix. □

In addition to the above transformations, there are many other examples. We have for instance the null transformation, which is given by:

$$\begin{pmatrix}0&0\\0&0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}$$

Here is now a more bizarre map, which can still be understood, however, as being the map which "switches the coordinates, then kills the second one":

$$\begin{pmatrix}0&1\\0&0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}y\\0\end{pmatrix}$$

Even more bizarrely now, here is a certain linear map, whose interpretation is more complicated, and is left to you, reader:

$$\begin{pmatrix}1&1\\0&0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}x+y\\0\end{pmatrix}$$

And here is another linear map, which once again, being something geometric, in 2 dimensions, can definitely be understood, at least in theory:

$$\begin{pmatrix}1&1\\0&1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}x+y\\y\end{pmatrix}$$

Let us discuss now the computation of the arbitrary symmetries, rotations and projections. We begin with the rotations, whose formula is a must-know:

THEOREM 14.15. *The rotation of angle $t\in\mathbb{R}$ is given by the matrix*

$$R_t=\begin{pmatrix}\cos t&-\sin t\\\sin t&\cos t\end{pmatrix}$$

depending on $t\in\mathbb{R}$ taken modulo 2π.

PROOF. The rotation being linear, it must correspond to a certain matrix:

$$R_t = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

We can guess this matrix, via its action on the basic coordinate vectors $\binom{1}{0}$ and $\binom{0}{1}$. A quick picture shows that we must have:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$$

Also, by paying attention to positives and negatives, we must have:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -\sin t \\ \cos t \end{pmatrix}$$

Guessing now the matrix is not complicated, because the first equation gives us the first column, and the second equation gives us the second column:

$$\begin{pmatrix} a \\ c \end{pmatrix} = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix} \quad , \quad \begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} -\sin t \\ \cos t \end{pmatrix}$$

Thus, we can just put together these two vectors, and we obtain our matrix. □

Regarding now the symmetries, the formula here is as follows:

THEOREM 14.16. *The symmetry with respect to the Ox axis rotated by an angle $t/2 \in \mathbb{R}$ is given by the matrix*

$$S_t = \begin{pmatrix} \cos t & \sin t \\ \sin t & -\cos t \end{pmatrix}$$

depending on $t \in \mathbb{R}$ taken modulo 2π.

PROOF. As before, we can guess the matrix via its action on the basic coordinate vectors $\binom{1}{0}$ and $\binom{0}{1}$. A quick picture shows that we must have:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$$

Also, by paying attention to positives and negatives, we must have:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \sin t \\ -\cos t \end{pmatrix}$$

Guessing now the matrix is not complicated, because we must have:

$$\begin{pmatrix} a \\ c \end{pmatrix} = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix} \quad , \quad \begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} \sin t \\ -\cos t \end{pmatrix}$$

Thus, we can just put together these two vectors, and we obtain our matrix. □

Finally, regarding the projections, the formula here is as follows:

THEOREM 14.17. *The projection on the Ox axis rotated by an angle $t/2 \in \mathbb{R}$ is given by the matrix*

$$P_t = \frac{1}{2}\begin{pmatrix} 1+\cos t & \sin t \\ \sin t & 1-\cos t \end{pmatrix}$$

depending on $t \in \mathbb{R}$ taken modulo 2π.

PROOF. We will need here some trigonometry, and more precisely the formulae for the duplication of the angles. Regarding the sine, the formula here is:

$$\sin(2t) = 2\sin t\cos t$$

Regarding the cosine, we have here 3 equivalent formulae, as follows:

$$\begin{aligned} \cos(2t) &= \cos^2 t - \sin^2 t \\ &= 2\cos^2 t - 1 \\ &= 1 - 2\sin^2 t \end{aligned}$$

Getting back now to our problem, some quick pictures, using similarity of triangles, and then the above trigonometry formulae, show that we must have:

$$P_t\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \cos\frac{t}{2}\begin{pmatrix} \cos\frac{t}{2} \\ \sin\frac{t}{2} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1+\cos t \\ \sin t \end{pmatrix}$$

$$P_t\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \sin\frac{t}{2}\begin{pmatrix} \cos\frac{t}{2} \\ \sin\frac{t}{2} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \sin t \\ 1-\cos t \end{pmatrix}$$

Now by putting together these two vectors, and we obtain our matrix. □

In order to formulate now our second theorem, dealing with compositions of maps, let us make the following multiplication convention, between matrices and matrices:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} p & q \\ r & s \end{pmatrix} = \begin{pmatrix} ap+br & aq+bs \\ cp+dr & cq+ds \end{pmatrix}$$

This might look a bit complicated, but as before, in what was concerning multiplying matrices and vectors, the idea is very simple, namely "multiply the rows of the first matrix by the columns of the second matrix". With this convention, we have:

THEOREM 14.18. *If we denote by $f_A : \mathbb{R}^2 \to \mathbb{R}^2$ the linear map associated to a matrix A, given by the formula*

$$f_A(v) = Av$$

then we have the following multiplication formula for such maps:

$$f_A f_B = f_{AB}$$

That is, the composition of linear maps corresponds to the multiplication of matrices.

PROOF. We want to prove that we have the following formula, valid for any two matrices $A, B \in M_2(\mathbb{R})$, and any vector $v \in \mathbb{R}^2$:

$$A(Bv) = (AB)v$$

For this purpose, let us write our matrices and vector as follows:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad , \quad B = \begin{pmatrix} p & q \\ r & s \end{pmatrix} \quad , \quad v = \begin{pmatrix} x \\ y \end{pmatrix}$$

The formula that we want to prove becomes:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \left[\begin{pmatrix} p & q \\ r & s \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \right] = \left[\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} p & q \\ r & s \end{pmatrix} \right] \begin{pmatrix} x \\ y \end{pmatrix}$$

But this is the same as saying that:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} px + qy \\ rx + sy \end{pmatrix} = \begin{pmatrix} ap + br & aq + bs \\ cp + dr & cq + ds \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

And this latter formula does hold indeed, because on both sides we get:

$$\begin{pmatrix} apx + aqy + brx + bsy \\ cpx + cqy + drx + dsy \end{pmatrix}$$

Thus, we have proved the result. □

As a verification for the above result, let us compose two rotations. The computation here is as follows, yieding a rotation, as it should, and of the correct angle:

$$\begin{aligned} R_s R_t &= \begin{pmatrix} \cos s & -\sin s \\ \sin s & \cos s \end{pmatrix} \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \\ &= \begin{pmatrix} \cos s \cos t - \sin s \sin t & -\cos s \sin t - \sin t \cos s \\ \sin s \cos t + \cos s \sin t & -\sin s \sin t + \cos s \cos t \end{pmatrix} \\ &= \begin{pmatrix} \cos(s+t) & -\sin(s+t) \\ \sin(s+t) & \cos(s+t) \end{pmatrix} \\ &= R_{s+t} \end{aligned}$$

We are ready now to pass to 3 dimensions. The idea is to select what we learned in 2 dimensions, nice results only, and generalize to 3 dimensions. We obtain:

THEOREM 14.19. *Consider a map $f : \mathbb{R}^3 \to \mathbb{R}^3$.*

(1) *f is linear when it is of the form $f(v) = Av$, with $A \in M_3(\mathbb{R})$.*
(2) *f is affine when $f(v) = Av + w$, with $A \in M_3(\mathbb{R})$ and $w \in \mathbb{R}^3$.*
(3) *We have the composition formula $f_A f_B = f_{AB}$, similar to the 2D one.*

PROOF. Here (1,2) can be proved exactly as in the 2D case, with the multiplication convention being as usual, "multiply the rows of the matrix by the vector":

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} ax+by+cz \\ dx+ey+fz \\ gx+hy+iz \end{pmatrix}$$

As for (3), once again the 2D idea applies, with the same product rule, "multiply the rows of the first matrix by the columns of the second matrix":

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} p & q & r \\ s & t & u \\ v & w & x \end{pmatrix} = \begin{pmatrix} ap+bs+cv & aq+bt+cw & ar+bu+cx \\ dp+es+fv & dq+et+fw & dr+eu+fx \\ gp+hs+iv & gq+ht+iw & gr+hu+ix \end{pmatrix}$$

Thus, we have proved our theorem. Of course, we are going a bit fast here, and some verifications are missing, but we will discuss all this in detail, in N dimensions. □

We are now ready to discuss 4 and more dimensions. Let us start with:

DEFINITION 14.20. *We can multiply the $M \times N$ matrices with $N \times K$ matrices,*

$$\begin{pmatrix} a_{11} & \ldots & a_{1N} \\ \vdots & & \vdots \\ a_{M1} & \ldots & a_{MN} \end{pmatrix} \begin{pmatrix} b_{11} & \ldots & b_{1K} \\ \vdots & & \vdots \\ b_{N1} & \ldots & b_{NK} \end{pmatrix}$$

the product being the $M \times K$ matrix given by the following formula,

$$\begin{pmatrix} a_{11}b_{11}+\ldots+a_{1N}b_{N1} & \ldots\ldots & a_{11}b_{1K}+\ldots+a_{1N}b_{NK} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ a_{M1}b_{11}+\ldots+a_{MN}b_{N1} & \ldots\ldots & a_{M1}b_{1K}+\ldots+a_{MN}b_{NK} \end{pmatrix}$$

obtained via the usual rule "multiply rows by columns".

In case the above formula looks hard to memorize, here is an alternative formulation of it, which is simpler and more powerful, by using the standard algebraic notation for the matrices, $A = (A_{ij})$, that we will heavily use, in what follows:

PROPOSITION 14.21. *The matrix multiplication is given by formula*

$$(AB)_{ij} = \sum_k A_{ik}B_{kj}$$

with A_{ij} standing for the entry of A at row i and column j.

PROOF. This is indeed just a shorthand for the formula in Definition 14.20, by following the rule there, namely "multiply the rows of A by the columns of B". □

As an illustration for the power of the convention in Proposition 14.21, we have:

PROPOSITION 14.22. *We have the following formula, valid for any matrices A, B, C,*

$$(AB)C = A(BC)$$

provided that the sizes of our matrices A, B, C fit.

PROOF. We have the following computation, using indices as above:

$$((AB)C)_{ij} = \sum_k (AB)_{ik} C_{kj} = \sum_{kl} A_{il} B_{lk} C_{kj}$$

On the other hand, we have as well the following computation:

$$(A(BC))_{ij} = \sum_l A_{il} (BC)_{lj} = \sum_{kl} A_{il} B_{lk} C_{kj}$$

Thus we have $(AB)C = A(BC)$, and we have proved our result. □

We can now talk about linear maps between spaces of arbitrary dimension, generalizing what we have been doing so far. The main result here is as follows:

THEOREM 14.23. *Consider a map $f : \mathbb{R}^N \to \mathbb{R}^M$.*

(1) *f is linear when it is of the form $f(v) = Av$, with $A \in M_{M \times N}(\mathbb{R})$.*
(2) *f is affine when $f(v) = Av + w$, with $A \in M_{M \times N}(\mathbb{R})$ and $w \in \mathbb{R}^M$.*
(3) *We have the composition formula $f_A f_B = f_{AB}$, whenever the sizes fit.*

PROOF. We already know that this happens at $M = N = 2$, and at $M = N = 3$ as well. In general, the proof is similar, by doing some elementary computations. □

14b. 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:

THEOREM 14.24. *A vector $v \in \mathbb{C}^N$ is called eigenvector of $A \in M_N(\mathbb{C})$, with corresponding eigenvalue λ, when A multiplies by λ in the direction of v:*

$$Av = \lambda v$$

In the case where $\mathbb{C}^N$ has a basis $v_1, \ldots, v_N$ formed by eigenvectors of A, with corresponding eigenvalues $\lambda_1, \ldots, \lambda_N$, in this new basis A becomes diagonal, as follows:

$$A \sim \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix}$$

Equivalently, if we denote by $D = diag(\lambda_1, \ldots, \lambda_N)$ the above diagonal matrix, and by $P = [v_1 \ldots v_N]$ the square matrix formed by the eigenvectors of A, we have:

$$A = PDP^{-1}$$

In this case we say that the matrix A is diagonalizable.

PROOF. This is something quite elementary, 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 14.25. *Let $A \in M_N(\mathbb{C})$, and for any eigenvalue $\lambda \in \mathbb{C}$ define the corresponding eigenspace as being the vector space formed by the corresponding eigenvectors:*

$$E_\lambda = \left\{ v \in \mathbb{C}^N \middle| Av = \lambda v \right\}$$

These eigenspaces E_λ are then in a direct sum position, in the sense that given vectors $v_1 \in E_{\lambda_1}, \ldots, v_k \in E_{\lambda_k}$ corresponding to different eigenvalues $\lambda_1, \ldots, \lambda_k$, we have:

$$\sum_i c_i v_i = 0 \implies c_i = 0$$

In particular, we have $\sum_\lambda \dim(E_\lambda) \leq N$, with the sum being over all the eigenvalues, and our matrix is diagonalizable precisely when we have equality.

PROOF. We prove the first assertion by recurrence on $k \in \mathbb{N}$. Assume by contradiction that we have a formula as follows, with the scalars $c_1, \ldots, c_k$ being not all zero:

$$c_1 v_1 + \ldots + c_k v_k = 0$$

By dividing by one of these scalars, we can assume that our formula is:

$$v_k = c_1 v_1 + \ldots + c_{k-1} v_{k-1}$$

Now let us apply A to this vector. On the left we obtain:

$$Av_k = \lambda_k v_k = \lambda_k c_1 v_1 + \ldots + \lambda_k c_{k-1} v_{k-1}$$

On the right we obtain something different, as follows:

$$\begin{aligned} A(c_1 v_1 + \ldots + c_{k-1} v_{k-1}) &= c_1 Av_1 + \ldots + c_{k-1} Av_{k-1} \\ &= c_1 \lambda_1 v_1 + \ldots + c_{k-1} \lambda_{k-1} v_{k-1} \end{aligned}$$

We conclude from this that the following equality must hold:

$$\lambda_k c_1 v_1 + \ldots + \lambda_k c_{k-1} v_{k-1} = c_1 \lambda_1 v_1 + \ldots + c_{k-1} \lambda_{k-1} v_{k-1}$$

On the other hand, we know by recurrence that the vectors $v_1, \ldots, v_{k-1}$ must be linearly independent. Thus, the coefficients must be equal, at right and at left:

$$\lambda_k c_1 = c_1 \lambda_1$$

$$\vdots$$

$$\lambda_k c_{k-1} = c_{k-1} \lambda_{k-1}$$

Now since at least one of the number c_i must be nonzero, from $\lambda_k c_i = c_i \lambda_i$ we obtain $\lambda_k = \lambda_i$, which is a contradiction. Thus our proof by recurrence of the first assertion is complete. As for the second assertion, this follows from the first one. □

In order to reach now to more advanced results, we can use the following key fact:

THEOREM 14.26. *Given a matrix $A \in M_N(\mathbb{C})$, consider its characteristic polynomial:*

$$P(x) = \det(A - x1_N)$$

The eigenvalues of A are then the roots of P. Also, we have the inequality

$$\dim(E_\lambda) \leq m_\lambda$$

where m_λ is the multiplicity of λ, as root of P.

PROOF. The first assertion follows from the following computation, using the fact that a linear map is bijective when the determinant of the associated matrix is nonzero:

$$\begin{aligned} \exists v, Av = \lambda v &\iff \exists v, (A - \lambda 1_N)v = 0 \\ &\iff \det(A - \lambda 1_N) = 0 \end{aligned}$$

Regarding now the second assertion, given an eigenvalue λ of our matrix A, consider the dimension $d_\lambda = \dim(E_\lambda)$ of the corresponding eigenspace. By changing the basis of $\mathbb{C}^N$, as for the eigenspace E_λ to be spanned by the first d_λ basis elements, our matrix becomes as follows, with B being a certain smaller matrix:

$$A \sim \begin{pmatrix} \lambda 1_{d_\lambda} & 0 \\ 0 & B \end{pmatrix}$$

We conclude that the characteristic polynomial of A is of the following form:

$$P_A = P_{\lambda 1_{d_\lambda}} P_B = (\lambda - x)^{d_\lambda} P_B$$

Thus the multiplicity m_λ of our eigenvalue λ, viewed as a root of P, is subject to the estimate $m_\lambda \geq d_\lambda$, and this leads to the conclusion in the statement. □

Now recall that we are over $\mathbb{C}$, where any polynomial equation, of arbitrary degree $N \in \mathbb{N}$, has exactly N complex solutions, counted with multiplicities. But with this, getting back now to linear algebra, we obtain the following result:

THEOREM 14.27. *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:

$$E_i = \left\{ v \in \mathbb{C}^N \middle| Av = \lambda_i v \right\}$$

The dimensions of these eigenspaces satisfy then the following inequalities,

$$\dim(E_i) \leq n_i$$

and A is diagonalizable precisely when we have equality for any i.

PROOF. This follows by combining the above results. By summing the inequalities $\dim(E_\lambda) \leq m_\lambda$ from Theorem 14.26, 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 14.25 that our matrix is diagonalizable when we have global equality. Thus, we are led to the conclusion in the statement. □

This was for the main result of linear algebra. There are countless applications of this, and generally speaking, advanced linear algebra consists in building on Theorem 14.27. Let us record as well a useful algorithmic version of the above result:

THEOREM 14.28. *The square matrices $A \in M_N(\mathbb{C})$ can be diagonalized as follows:*

(1) *Compute the characteristic polynomial.*
(2) *Factorize the characteristic polynomial.*
(3) *Compute the eigenvectors, for each eigenvalue found.*
(4) *If there are no N eigenvectors, A is not diagonalizable.*
(5) *Otherwise, A is diagonalizable, $A = PDP^{-1}$.*

PROOF. This is an informal reformulation of Theorem 14.27, with (4) referring to the total number of linearly independent eigenvectors found in (3), and with $A = PDP^{-1}$ in (5) being the usual diagonalization formula, with P, D being as before. □

As a remark here, in step (3) it is always better to start with the eigenvalues having big multiplicity. Indeed, a multiplicity 1 eigenvalue, for instance, can never lead to the end of the computation, via (4), simply because the eigenvectors always exist.

14c. Spectral theorems

Let us go back now to the diagonalization question. Here is a key result:

THEOREM 14.29. *Any matrix $A \in M_N(\mathbb{C})$ which is self-adjoint, $A = A^*$, is diagonalizable, with the diagonalization being of the following type,*

$$A = UDU^*$$

with $U \in U_N$, and with $D \in M_N(\mathbb{R})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because if we take a matrix of the form $A = UDU^*$, with U unitary and D diagonal and real, then we have:

$$\begin{aligned} A^* &= (UDU^*)^* \\ &= UD^*U^* \\ &= UDU^* \\ &= A \end{aligned}$$

In the other sense now, assume that A is self-adjoint, $A = A^*$. Our first claim is that the eigenvalues are real. Indeed, assuming $Av = \lambda v$, we have:

$$\begin{aligned} \lambda < v, v > &= < \lambda v, v > \\ &= < Av, v > \\ &= < v, Av > \\ &= < v, \lambda v > \\ &= \bar{\lambda} < v, v > \end{aligned}$$

Thus we obtain $\lambda \in \mathbb{R}$, as claimed. Our next claim now is that the eigenspaces corresponding to different eigenvalues are pairwise orthogonal. Assume indeed that:

$$Av = \lambda v \quad , \quad Aw = \mu w$$

We have then the following computation, using $\lambda, \mu \in \mathbb{R}$:

$$\begin{aligned} \lambda < v, w > &= < \lambda v, w > \\ &= < Av, w > \\ &= < v, Aw > \\ &= < v, \mu w > \\ &= \mu < v, w > \end{aligned}$$

Thus $\lambda \neq \mu$ implies $v \perp w$, as claimed. In order now to finish the proof, it remains to prove that the eigenspaces of A span the whole space $\mathbb{C}^N$. For this purpose, we will use a recurrence method. Let us pick an eigenvector of our matrix:

$$Av = \lambda v$$

Assuming now that we have a vector w orthogonal to it, $v \perp w$, we have:

$$\begin{aligned} < Aw, v > &= < w, Av > \\ &= < w, \lambda v > \\ &= \lambda < w, v > \\ &= 0 \end{aligned}$$

Thus, if v is an eigenvector, then the vector space $v^\perp$ is invariant under A. Moreover, since a matrix A is self-adjoint precisely when $< Av, v > \in \mathbb{R}$ for any vector $v \in \mathbb{C}^N$, as one can see by expanding the scalar product, the restriction of A to the subspace $v^\perp$ is self-adjoint. Thus, we can proceed by recurrence, and we obtain the result. □

Let us discuss now the case of the unitary matrices. We have here:

THEOREM 14.30. *Any matrix $U \in M_N(\mathbb{C})$ which is unitary, $U^* = U^{-1}$, is diagonalizable, with the eigenvalues on $\mathbb{T}$. More precisely we have*

$$U = VDV^*$$

with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because given a matrix of type $U = VDV^*$, with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ being diagonal, we have:

$$\begin{aligned} U^* &= (VDV^*)^* \\ &= VD^*V^* \\ &= VD^{-1}V^{-1} \\ &= (V^*)^{-1}D^{-1}V^{-1} \\ &= (VDV^*)^{-1} \\ &= U^{-1} \end{aligned}$$

Let us prove now the first assertion, stating that the eigenvalues of a unitary matrix $U \in U_N$ belong to $\mathbb{T}$. Indeed, assuming $Uv = \lambda v$, we have:

$$\begin{aligned} < v, v > &= < U^*Uv, v > \\ &= < Uv, Uv > \\ &= < \lambda v, \lambda v > \\ &= |\lambda|^2 < v, v > \end{aligned}$$

Thus we obtain $\lambda \in \mathbb{T}$, as claimed. Our next claim now is that the eigenspaces corresponding to different eigenvalues are pairwise orthogonal. Assume indeed that:

$$Uv = \lambda v \quad , \quad Uw = \mu w$$

We have then the following computation, using $U^* = U^{-1}$ and $\lambda, \mu \in \mathbb{T}$:

$$\begin{aligned}
\lambda < v, w > &= < \lambda v, w > \\
&= < Uv, w > \\
&= < v, U^* w > \\
&= < v, U^{-1} w > \\
&= < v, \mu^{-1} w > \\
&= \mu < v, w >
\end{aligned}$$

Thus $\lambda \neq \mu$ implies $v \perp w$, as claimed. In order now to finish the proof, it remains to prove that the eigenspaces of U span the whole space $\mathbb{C}^N$. For this purpose, we will use a recurrence method. Let us pick an eigenvector of our matrix:

$$Uv = \lambda v$$

Assuming that we have a vector w orthogonal to it, $v \perp w$, we have:

$$\begin{aligned}
< Uw, v > &= < w, U^* v > \\
&= < w, U^{-1} v > \\
&= < w, \lambda^{-1} v > \\
&= \lambda < w, v > \\
&= 0
\end{aligned}$$

Thus, if v is an eigenvector, then the vector space $v^\perp$ is invariant under U. Now since U is an isometry, so is its restriction to this space $v^\perp$. Thus this restriction is a unitary, and so we can proceed by recurrence, and we obtain the result. □

The self-adjoint matrices and the unitary matrices are particular cases of the general notion of a "normal matrix", and we have here:

THEOREM 14.31. *Any matrix $A \in M_N(\mathbb{C})$ which is normal, $AA^* = A^*A$, is diagonalizable, with the diagonalization being of the following type,*

$$A = UDU^*$$

with $U \in U_N$, and with $D \in M_N(\mathbb{C})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because if we take a matrix of the form $A = UDU^*$, with U unitary and D diagonal, then we have:

$$\begin{aligned}
AA^* &= UDU^* \cdot UD^*U^* \\
&= UDD^*U^* \\
&= UD^*DU^* \\
&= UD^*U^* \cdot UDU^* \\
&= A^*A
\end{aligned}$$

In the other sense now, this is something more technical. Our first claim is that a matrix A is normal precisely when the following happens, for any vector v:

$$||Av|| = ||A^*v||$$

Indeed, the above equality can be written as follows:

$$< AA^*v, v >=< A^*Av, v >$$

But this is equivalent to $AA^* = A^*A$, by expanding the scalar products. Our next claim is that A, A^* have the same eigenvectors, with conjugate eigenvalues:

$$Av = \lambda v \implies A^*v = \bar{\lambda}v$$

Indeed, this follows from the following computation, and from the trivial fact that if A is normal, then so is any matrix of type $A - \lambda 1_N$:

$$\begin{aligned} ||(A^* - \bar{\lambda}1_N)v|| &= ||(A - \lambda 1_N)^*v|| \\ &= ||(A - \lambda 1_N)v|| \\ &= 0 \end{aligned}$$

Let us prove now, by using this, that the eigenspaces of A are pairwise orthogonal. Assume that we have two eigenvectors, corresponding to different eigenvalues, $\lambda \neq \mu$:

$$Av = \lambda v \quad , \quad Aw = \mu w$$

We have the following computation, which shows that $\lambda \neq \mu$ implies $v \perp w$:

$$\begin{aligned} \lambda < v, w > &= < \lambda v, w > \\ &= < Av, w > \\ &= < v, A^*w > \\ &= < v, \bar{\mu}w > \\ &= \mu < v, w > \end{aligned}$$

In order to finish, it remains to prove that the eigenspaces of A span the whole $\mathbb{C}^N$. This is something that we have already seen for the self-adjoint matrices, and for unitaries, and we will use here these results, in order to deal with the general normal case. As a first observation, given an arbitrary matrix A, the matrix AA^* is self-adjoint:

$$(AA^*)^* = AA^*$$

Thus, we can diagonalize this matrix AA^*, as follows, with the passage matrix being a unitary, $V \in U_N$, and with the diagonal form being real, $E \in M_N(\mathbb{R})$:

$$AA^* = VEV^*$$

Now observe that, for matrices of type $A = UDU^*$, which are those that we supposed to deal with, we have the following formulae:

$$V = U \quad , \quad E = D\bar{D}$$

In particular, the matrices A and AA^* have the same eigenspaces. So, this will be our idea, proving that the eigenspaces of AA^* are eigenspaces of A. In order to do so, let us pick two eigenvectors v, w of the matrix AA^*, corresponding to different eigenvalues, $\lambda \neq \mu$. The eigenvalue equations are then as follows:

$$AA^*v = \lambda v \quad , \quad AA^*w = \mu w$$

We have the following computation, using the normality condition $AA^* = A^*A$, and the fact that the eigenvalues of AA^*, and in particular μ, are real:

$$\begin{aligned}
\lambda < Av, w > &= < \lambda Av, w > \\
&= < A\lambda v, w > \\
&= < AAA^*v, w > \\
&= < AA^*Av, w > \\
&= < Av, AA^*w > \\
&= < Av, \mu w > \\
&= \mu < Av, w >
\end{aligned}$$

We conclude that we have $< Av, w >= 0$. But this reformulates as follows:

$$\lambda \neq \mu \implies A(E_\lambda) \perp E_\mu$$

Now since the eigenspaces of AA^* are pairwise orthogonal, and span the whole $\mathbb{C}^N$, we deduce from this that these eigenspaces are invariant under A:

$$A(E_\lambda) \subset E_\lambda$$

But with this result in hand, we can finish. Indeed, we can decompose the problem, and the matrix A itself, following these eigenspaces of AA^*, which in practice amounts in saying that we can assume that we only have 1 eigenspace. Now by rescaling, this is the same as assuming that we have $AA^* = 1$. But with this, we are now into the unitary case, that we know how to solve, as explained in Theorem 14.30, and so done. □

14d. Some arithmetic

Getting back to the basics, we can develop geometry by using lines indexed by other types of numbers. And, regarding these numbers, these usually come from:

DEFINITION 14.32. *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 $a(b + c) = ab + ac$.*

As a basic example here, passed the reals that we know well, we have the field of rational numbers $\mathbb{Q}$, with its usual addition and multiplication, namely:

$$\frac{a}{b}+\frac{c}{d}=\frac{ad+bc}{ad}\quad,\quad\frac{a}{b}\cdot\frac{c}{d}=\frac{ac}{bd}$$

In fact, the simplest possible field seems to be $\mathbb{Q}$. However, this is not exactly true, because, by a strange twist of fate, the numbers $0,1$, whose presence in a field is mandatory, $0,1\in F$, can form themselves a field, with addition as follows:

$$1+1=0$$

Let us summarize this finding, along with a bit more, obtained by suitably replacing our 2, used for addition, with an arbitrary prime number p, as follows:

THEOREM 14.33. *The following happen:*

(1) $\mathbb{Q}$ *is the simplest field having the property* $1+\ldots+1\neq 0$*, in the sense that any field* F *having this property must contain it,* $\mathbb{Q}\subset F$*.*
(2) *The property* $1+\ldots+1\neq 0$ *can hold or not, and if not, the smallest number of terms needed for having* $1+\ldots+1=0$ *is a certain prime number* p*.*
(3) $\mathbb{F}_p=\{0,1,\ldots,p-1\}$*, with* p *prime, is the simplest field having the property* $1+\ldots+1=0$*, with* p *terms, in the sense that this implies* $\mathbb{F}_p\subset F$*.*

PROOF. All this is basic number theory, the idea being as follows:

(1) This is clear, because $1+\ldots+1\neq 0$ tells us that we have an embedding $\mathbb{N}\subset F$, and then by taking inverses with respect to $+$ and $\times$ we obtain $\mathbb{Q}\subset F$.

(2) Again, this is clear, because assuming $1+\ldots+1=0$, with $p=ab$ terms, chosen minimal, we would have a formula as follows, which is a contradiction:

$$(\underbrace{1+\ldots+1}_{a\ terms})(\underbrace{1+\ldots+1}_{b\ terms})=0$$

(3) This follows a bit as in (1), with the copy $\mathbb{F}_p\subset F$ consisting by definition of the various sums of type $1+\ldots+1$, which must cycle modulo p, as shown by (2). □

Now, let us do some basic geometry, say over $\mathbb{F}_p$. However, things are a bit bizarre here, and we have for instance the following result, to start with:

PROPOSITION 14.34. *The circle of radius zero* $x^2+y^2=0$ *over* $\mathbb{F}_p$ *is as follows:*

(1) *At* $p=2$*, this has* 2 *points.*
(2) *At* $p=1(4)$*, this has* $2p-1$ *points.*
(3) *At* $p=3(4)$*, this has* 1 *point.*

PROOF. Our circle $x^2+y^2=0$ is formed by the point $(0,0)$, and then of the solutions of $x^2+y^2=0$, with $x,y\neq 0$. But this latter equation is equivalent to $(x/y)^2+1=0$, and so to $(x/y)^2=-1$, so the number of points of our circle is:

$$N=1+(p-1)\#\{r|r^2=-1\}$$

But at $p = 2$ this gives $N = 1 + 1 \times 1 = 2$, then at $p = 1(4)$ this gives $N = 1 + (p-1) \times 2 = 2p - 1$, and finally at $p = 3(4)$ this gives $N = 1 + (p-1) \times 0 = 1$. □

When looking at more general conics, still over $\mathbb{F}_p$, things do not necessarily improve, and we have some other bizarre results, along the same lines, such as:

THEOREM 14.35. *Any curve over $\mathbb{F}_2$ is a conic. However, this is not the case for $\mathbb{F}_p$ with $p \geq 3$.*

PROOF. This is again something elementary, as follows:

(1) Let us find the conics over $\mathbb{F}_2$. These are given by equations as follows:

$$ax^2 + by^2 + cxy + dx + ey + f = 0$$

Since $x^2 = x$ holds in $\mathbb{F}_2$, the first 2 terms dissapear, and we are left with:

$$cxy + dx + ey + f = 0$$

– The first case, $c = 0$, corresponds to the lines over $\mathbb{F}_2$. But there are 8 such lines, all distinct, given by $r = 0$, $x = r$, $y = r$, $x + y = r$, with $r = 0, 1$.

– The second case, $c \neq 0$, corresponds to the non-degenerate conics over $\mathbb{F}_2$. But there are 8 such conics, all distinct, and distinct as well from the 8 lines found above, given by $xy = r$, $x(y+1) = r$, $(x+1)y = r$, $(x+1)(y+1) = r$, with $r = 0, 1$.

Summarizing, we have $8 + 8 = 16$ conics over $\mathbb{Z}_2$. But since the plane $\mathbb{F}_2 \times \mathbb{F}_2$ has $2 \times 2 = 4$ points, there are $2^4 = 16$ possible curves. Thus, all the curves are conics.

(2) Regarding now $\mathbb{F}_p$ with $p \geq 3$, here the plane $\mathbb{F}_p \times \mathbb{F}_p$ has p^2 points, so there are 2^{p^2} curves. Among these curves, the conics are given by equations as follows:

$$ax^2 + by^2 + cxy + dx + ey + f = 0$$

Thus, we have at most p^6 conics, and since we have $2^{p^2} > p^6$ for any $p \geq 4$, we are done with the case $p \geq 5$. In the remaining case now, $p = 3$, the $3^6 = 729$ possible conics split into the $2^5 = 243$ ones with $a = 0$, and the $2 \times 243 = 486$ ones with $a \neq 0$. But these latter conics appear twice, as we can see by dividing everything by a, and so there are only $1 \times 243 = 243$ of them. Thus, we have at most $243 + 243 = 486$ conics, and this is smaller than the number of curves of $\mathbb{F}_3 \times \mathbb{F}_3$, which is $2^9 = 512$, as desired. □

On the positive side, however, we have the following interesting result:

THEOREM 14.36. *Given a field F, we can talk about the projective space P_F^{N-1}, as being the space of lines in F^N passing through the origin. At $N = 3$ we have*

$$|P_F^2| = q^2 + q + 1$$

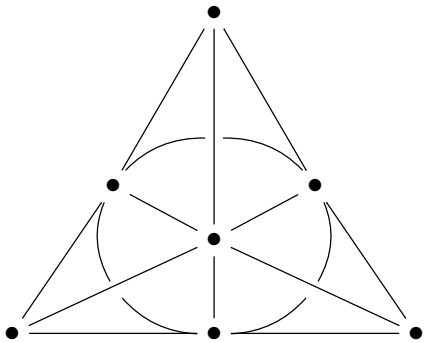
where $q = |F|$, in the case where our field F is finite.

PROOF. This is indeed clear from definitions, with the point count coming from:

$$\begin{aligned} |P_F^2| &= \frac{|F^3 - \{0\}|}{|F - \{0\}|} \\ &= \frac{q^3 - 1}{q - 1} \\ &= q^2 + q + 1 \end{aligned}$$

Thus, we are led to the conclusions in the statement. □

As an example, let us see what happens for the simplest finite field that we know, namely $F = \mathbb{Z}_2$. Here our projective plane, having $4 + 2 + 1 = 7$ points, and 7 lines, is a famous combinatorial object, called Fano plane, that we know from before:



Here the circle in the middle is by definition a line, and with this convention, the basic projective geometry axioms from chapter 1 are satisfied, in the sense that any two points determine a line, and any two lines determine a point. And isn't this beautiful.

14e. Exercises

Exercises:

EXERCISE 14.37.

EXERCISE 14.38.

EXERCISE 14.39.

EXERCISE 14.40.

EXERCISE 14.41.

EXERCISE 14.42.

EXERCISE 14.43.

EXERCISE 14.44.

Bonus exercise.

CHAPTER 15

Functions, revised

15a. Partial derivatives

Moving now to several variables, $N \geq 2$, as a first job, given a function $\varphi : \mathbb{R}^N \to \mathbb{R}$, we would like to find a quantity $\varphi'(x)$ making the following formula work:

$$\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h$$

But here, as in 1 variable, there are not so many choices, and the solution is that of defining $\varphi'(x)$ as being the row vector formed by the partial derivatives at x:

$$\varphi'(x) = \begin{pmatrix} \frac{d\varphi}{dx_1} & \ldots & \frac{d\varphi}{dx_N} \end{pmatrix}$$

To be more precise, with this value for $\varphi'(x)$, our approximation formula $\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h$ makes sense indeed, as an equality of real numbers, with $\varphi'(x)h \in \mathbb{R}$ being obtained as the matrix multiplication of the row vector $\varphi'(x)$, and the column vector h. As for the fact that our formula holds indeed, this follows by putting together the approximation properties of each of the partial derivatives $d\varphi/dx_i$, which give:

$$\varphi(x+h) \simeq \varphi(x) + \sum_{i=1}^{N} \frac{d\varphi}{dx_i} \cdot h_i = \varphi(x) + \varphi'(x)h$$

Before moving forward, you might say, why bothering with horizontal vectors, when it is so simple and convenient to have all vectors vertical, by definition. Good point, and in answer, we can indeed talk about the gradient of φ, constructed as follows:

$$\nabla\varphi = \begin{pmatrix} \frac{d\varphi}{dx_1} \\ \vdots \\ \frac{d\varphi}{dx_N} \end{pmatrix}$$

With this convention, $\nabla\varphi$ geometrically describes the slope of φ at the point x, in the obvious way. However, the approximation formula must be rewritten as follows:

$$\varphi(x+h) \simeq \varphi(x) + < \nabla\varphi(x), h >$$

In what follows we will use both φ' and $\nabla\varphi$, depending on the context. Moving now to second derivatives, the main result here is as follows:

THEOREM 15.1. *The second derivative of a function $\varphi : \mathbb{R}^N \to \mathbb{R}$, making the formula*

$$\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h + \frac{< \varphi''(x)h, h >}{2}$$

work, is its Hessian matrix $\varphi''(x) \in M_N(\mathbb{R})$, given by the following formula:

$$\varphi''(x) = \left(\frac{d^2\varphi}{dx_i dx_j} \right)_{ij}$$

Moreover, this Hessian matrix is symmetric, $\varphi''(x)_{ij} = \varphi'(x)_{ji}$.

PROOF. There are several things going on here, the idea being as follows:

(1) As a first observation, at $N = 1$ the Hessian matrix constructed above is simply the 1×1 matrix having as entry the second derivative $\varphi''(x)$, and the formula in the statement is something that we know well from chapter 10, namely:

$$\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h + \frac{\varphi''(x)h^2}{2}$$

(2) At $N = 2$ now, we obviously need to differentiate φ twice, and the point is that we come in this way upon the following formula, called Clairaut formula:

$$\frac{d^2\varphi}{dxdy} = \frac{d^2\varphi}{dydx}$$

But, is this formula correct or not? As an intuitive justification for it, let us consider a product of power functions, $\varphi(z) = x^p y^q$. We have then our formula, due to:

$$\frac{d^2\varphi}{dxdy} = \frac{d}{dx}\left(\frac{dx^p y^q}{dy} \right) = \frac{d}{dx}\left(qx^p y^{q-1} \right) = pqx^{p-1}y^{q-1}$$

$$\frac{d^2\varphi}{dydx} = \frac{d}{dy}\left(\frac{dx^p y^q}{dx} \right) = \frac{d}{dy}\left(px^{p-1} y^q \right) = pqx^{p-1}y^{q-1}$$

Next, let us consider a linear combination of power functions, $\varphi(z) = \sum_{pq} c_{pq} x^p y^q$, which can be finite or not. We have then, by using the above computation:

$$\frac{d^2\varphi}{dxdy} = \frac{d^2\varphi}{dydx} = \sum_{pq} c_{pq} pq x^{p-1} y^{q-1}$$

Thus, we can see that our commutation formula for derivatives holds indeed, due to the fact that the functions in x, y commute. Of course, all this does not fully prove our formula, in general. But exercise for you, to have this idea fully working, or to look up the standard proof of the Clairaut formula, using the mean value theorem.

(3) Moving now to $N = 3$ and higher, we can use here the Clairaut formula with respect to any pair of coordinates, which gives the Schwarz formula, namely:

$$\frac{d^2\varphi}{dx_i dx_j} = \frac{d^2\varphi}{dx_j dx_i}$$

Thus, the second derivative, or Hessian matrix, is symmetric, as claimed.

(4) Getting now to the main topic, namely approximation formula in the statement, in arbitrary N dimensions, this is in fact something which does not need a new proof, because it follows from the one-variable formula in (1), applied to the restriction of φ to the following segment in $\mathbb{R}^N$, which can be regarded as being a one-variable interval:

$$I = [x, x + h]$$

To be more precise, let $y \in \mathbb{R}^N$, and consider the following function, with $r \in \mathbb{R}$:

$$f(r) = \varphi(x + ry)$$

We know from (1) that the Taylor formula for f, at the point $r = 0$, reads:

$$f(r) \simeq f(0) + f'(0)r + \frac{f''(0)r^2}{2}$$

And our claim is that, with $h = ry$, this is precisely the formula in the statement.

(5) So, let us see if our claim is correct. By using the chain rule, we have the following formula, with on the right, as usual, a row vector multiplied by a column vector:

$$f'(r) = \varphi'(x + ry) \cdot y$$

By using again the chain rule, we can compute the second derivative as well:

$$\begin{aligned}
f''(r) &= (\varphi'(x + ry) \cdot y)' \\
&= \left(\sum_i \frac{d\varphi}{dx_i}(x + ry) \cdot y_i \right)' \\
&= \sum_i \sum_j \frac{d^2\varphi}{dx_i dx_j}(x + ry) \cdot \frac{d(x + ry)_j}{dr} \cdot y_i \\
&= \sum_i \sum_j \frac{d^2\varphi}{dx_i dx_j}(x + ry) \cdot y_i y_j \\
&= < \varphi''(x + ry)y, y >
\end{aligned}$$

(6) Time now to conclude. We know that we have $f(r) = \varphi(x + ry)$, and according to our various computations above, we have the following formulae:

$$f(0) = \varphi(x) \quad , \quad f'(0) = \varphi'(x) \quad , \quad f''(0) =< \varphi''(x)y, y >$$

Buit with this data in hand, the usual Taylor formula for our one variable function f, at order 2, at the point $r = 0$, takes the following form, with $h = ry$:

$$\begin{aligned}\varphi(x + ry) &\simeq \varphi(x) + \varphi'(x)ry + \frac{<\varphi''(x)y, y> r^2}{2}\\ &= \varphi(x) + \varphi'(x)t + \frac{<\varphi''(x)h, h>}{2}\end{aligned}$$

Thus, we have obtained the formula in the statement. □

As before in the one variable case, many more things can be said, as a continuation of the above. For instance the local minima and maxima of $\varphi : \mathbb{R}^N \to \mathbb{R}$ appear at the points $x \in \mathbb{R}^N$ where the derivative vanishes, $\varphi'(x) = 0$, and where the second derivative $\varphi''(x) \in M_N(\mathbb{R})$ is positive, respectively negative. But, you surely know all this.

As a key observation now, generalizing what we know in 1 variable, we have:

PROPOSITION 15.2. *Intuitively, the following quantity, called Laplacian of φ,*

$$\Delta\varphi = \sum_{i=1}^{N} \frac{d^2\varphi}{dx_i^2}$$

measures how much different is $\varphi(x)$, compared to the average of $\varphi(y)$, with $y \simeq x$.

PROOF. As before with 1 variable, this is something a bit heuristic, but good to know. Let us write the formula in Theorem 15.1, as such, and with $h \to -h$ too:

$$\varphi(x + h) \simeq \varphi(x) + \varphi'(x)h + \frac{<\varphi''(x)h, h>}{2}$$

$$\varphi(x - h) \simeq \varphi(x) - \varphi'(x)h + \frac{<\varphi''(x)h, h>}{2}$$

By making the average, we obtain the following formula:

$$\frac{\varphi(x + h) + \varphi(x - h)}{2} \simeq \varphi(x) + \frac{<\varphi''(x)h, h>}{2}$$

Thus, thinking a bit, we are led to the conclusion in the statement, modulo some discussion about integrating all this, that we will not really need, in what follows. □

With this understood, the problem is now, what can we say about the mathematics of Δ? As a first observation, which is a bit speculative, the Laplace operator appears by

applying twice the gradient operator, in a somewhat formal sense, as follows:

$$\begin{aligned}
\Delta\varphi &= \sum_{i=1}^{N}\frac{d^2\varphi}{dx_i^2}\\
&= \sum_{i=1}^{N}\frac{d}{dx_i}\cdot\frac{d\varphi}{dx_i}\\
&= \left\langle\begin{pmatrix}\frac{d}{dx_1}\\ \vdots\\ \frac{d}{dx_N}\end{pmatrix},\begin{pmatrix}\frac{d\varphi}{dx_1}\\ \vdots\\ \frac{d\varphi}{dx_N}\end{pmatrix}\right\rangle\\
&= <\nabla,\nabla\varphi>
\end{aligned}$$

Thus, it is possible to write a formula of type $\Delta=\nabla^2$, with the convention that the square of the gradient ∇ is taken in a scalar product sense, as above. However, this can be a bit confusing, and in what follows, we will not use this notation.

Instead of further thinking at this, and at double derivatives in general, let us formulate a more straightforward question, inspired by linear algebra, as follows:

QUESTION 15.3. *The Laplace operator being linear,*

$$\Delta(a\varphi+b\psi)=a\Delta\varphi+b\Delta\psi$$

what can we say about it, inspired by usual linear algebra?

In answer now, the space of functions $\varphi:\mathbb{R}^N\to\mathbb{R}$, on which Δ acts, being infinite dimensional, the usual tools from linear algebra do not apply as such, and we must be extremely careful. For instance, we cannot really expect to diagonalize Δ, via some sort of explicit procedure, as we usually do in linear algebra, for the usual matrices.

Thinking some more, there is actually a real bug too with our problem, because at $N=1$ this problem becomes "what can we say about the second derivatives $\varphi'':\mathbb{R}\to\mathbb{R}$ of the functions $\varphi:\mathbb{R}\to\mathbb{R}$, inspired by linear algebra", with answer "not much".

And by thinking even more, still at $N=1$, there is a second bug too, because if $\varphi:\mathbb{R}\to\mathbb{R}$ is twice differentiable, nothing will guarantee that its second derivative $\varphi'':\mathbb{R}\to\mathbb{R}$ is twice differentiable too. Thus, we have some issues with the domain and range of Δ, regarded as linear operator, and these problems will persist at higher N.

So, shall we trash Question 15.3? Not so quick, because, very remarkably, some magic comes at $N=2$ and higher in relation with complex analysis, according to:

PRINCIPLE 15.4. *The functions $\varphi : \mathbb{R}^N \to \mathbb{R}$ which are 0-eigenvectors of Δ,*

$$\Delta\varphi = 0$$

called harmonic functions, have the following properties:

(1) *At $N = 1$, nothing spectacular, these are just the linear functions.*
(2) *At $N = 2$, these are, locally, the real parts of holomorphic functions.*
(3) *At $N \geq 3$, these still share many properties with the holomorphic functions.*

In order to understand this, or at least get introduced to it, let us first look at the case $N = 2$. Here, any function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ can be regarded as function $\varphi : \mathbb{C} \to \mathbb{R}$, depending on $z = x + iy$. But, in view of this, it is natural to enlarge the attention to the functions $\varphi : \mathbb{C} \to \mathbb{C}$, and ask which of these functions are harmonic, $\Delta\varphi = 0$. And here, we have the following remarkable result, making the link with complex analysis:

THEOREM 15.5. *Any holomorphic function $\varphi : \mathbb{C} \to \mathbb{C}$, when regarded as function*

$$\varphi : \mathbb{R}^2 \to \mathbb{C}$$

is harmonic. Moreover, the conjugates $\bar{\varphi}$ of holomorphic functions are harmonic too.

PROOF. The first assertion comes from the following computation, with $z = x + iy$:

$$\begin{aligned}
\Delta z^n &= \frac{d^2 z^n}{dx^2} + \frac{d^2 z^n}{dy^2} \\
&= \frac{d(nz^{n-1})}{dx} + \frac{d(inz^{n-1})}{dy} \\
&= n(n-1)z^{n-2} - n(n-1)z^{n-2} \\
&= 0
\end{aligned}$$

As for the second assertion, this follows from $\Delta\bar{\varphi} = \overline{\Delta\varphi}$, which is clear from definitions, and which shows that if φ is harmonic, then so is its conjugate $\bar{\varphi}$. □

Many more things can be said, along these lines.

15b. Multiple integrals

We can talk about multiple integrals, in the obvious way. Getting now to the general theory and rules, for computing such integrals, the key result here is the change of variable formula. In order to discuss this, let us start with something that we know well, in 1D:

PROPOSITION 15.6. *We have the change of variable formula*

$$\int_a^b f(x)dx = \int_c^d f(\varphi(t))\varphi'(t)dt$$

where $c = \varphi^{-1}(a)$ and $d = \varphi^{-1}(b)$.

PROOF. This follows with $f = F'$, via the following differentiation rule:

$$(F\varphi)'(t) = F'(\varphi(t))\varphi'(t)$$

Indeed, by integrating between c and d, we obtain the result. □

In several variables now, we can only expect the above $\varphi'(t)$ factor to be replaced by something similar, a sort of "derivative of φ, arising as a real number". But this can only be the Jacobian $\det(\varphi'(t))$, and with this in mind, we are led to:

THEOREM 15.7. *Given a transformation $\varphi = (\varphi_1, \ldots, \varphi_N)$, we have*

$$\int_E f(x)dx = \int_{\varphi^{-1}(E)} f(\varphi(t))|J_\varphi(t)|dt$$

with the J_φ quantity, called Jacobian, being given by

$$J_\varphi(t) = \det\left[\left(\frac{d\varphi_i}{dx_j}(x)\right)_{ij}\right]$$

and with this generalizing the formula from Proposition 15.6.

PROOF. This is something quite tricky, the idea being as follows:

(1) Observe first that this generalizes indeed the change of variable formula in 1 dimension, from Proposition 15.6, the point here being that the absolute value on the derivative appears as to compensate for the lack of explicit bounds for the integral.

(2) In general now, we can first argue that, the formula in the statement being linear in f, we can assume $f = 1$. Thus we want to prove $vol(E) = \int_{\varphi^{-1}(E)} |J_\varphi(t)|dt$, and with $D = \varphi^{-1}(E)$, this amounts in proving $vol(\varphi(D)) = \int_D |J_\varphi(t)|dt$.

(3) Now since this latter formula is additive with respect to D, it is enough to prove that $vol(\varphi(D)) = \int_D J_\varphi(t)dt$, for small cubes D, and assuming $J_\varphi > 0$. But this follows by using the usual definition of the determinant, as a volume.

(4) The details and computations however are quite non-trivial, and can be found for instance in Rudin [79]. So, please read that. With this, reading the complete proof of the present theorem from Rudin, being part of the standard math experience. □

Many other things can be said, as a continuation of the above.

15c. Spherical coordinates

Time now do some exciting computations, with the technology that we have. In what regards the applications of Theorem 15.7, these often come via:

PROPOSITION 15.8. *We have polar coordinates in 2 dimensions,*

$$\begin{cases} x &= r\cos t \\ y &= r\sin t \end{cases}$$

the corresponding Jacobian being $J = r$.

PROOF. This is elementary, the Jacobian being:

$$\begin{aligned} J &= \begin{vmatrix} \frac{d(r\cos t)}{dr} & \frac{d(r\cos t)}{dt} \\ \frac{d(r\sin t)}{dr} & \frac{d(r\sin t)}{dt} \end{vmatrix} \\ &= \begin{vmatrix} \cos t & -r\sin t \\ \sin t & r\cos t \end{vmatrix} \\ &= r\cos^2 t + r\sin^2 t \\ &= r \end{aligned}$$

Thus, we have indeed the formula in the statement. □

We can now compute the Gauss integral, which is the best calculus formula ever:

THEOREM 15.9. *We have the following formula,*

$$\int_{\mathbb{R}} e^{-x^2} dx = \sqrt{\pi}$$

called Gauss integral formula.

PROOF. Let I be the above integral. By using polar coordinates, we obtain:

$$\begin{aligned} I^2 &= \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-x^2-y^2} dxdy \\ &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2} rdrdt \\ &= 2\pi\int_0^{\infty} \left(-\frac{e^{-r^2}}{2}\right)' dr \\ &= 2\pi\left[0 - \left(-\frac{1}{2}\right)\right] \\ &= \pi \end{aligned}$$

Thus, we are led to the formula in the statement. □

Moving now to 3 dimensions, we have here the following result:

PROPOSITION 15.10. *We have spherical coordinates in* 3 *dimensions,*

$$\begin{cases} x &= r\cos s \\ y &= r\sin s\cos t \\ z &= r\sin s\sin t \end{cases}$$

the corresponding Jacobian being $J(r,s,t) = r^2\sin s$.

PROOF. The fact that we have indeed spherical coordinates is clear. Regarding now the Jacobian, this is given by the following formula:

$$\begin{aligned} & J(r,s,t) \\ = & \begin{vmatrix} \cos s & -r\sin s & 0 \\ \sin s\cos t & r\cos s\cos t & -r\sin s\sin t \\ \sin s\sin t & r\cos s\sin t & r\sin s\cos t \end{vmatrix} \\ = & r^2\sin s\sin t\begin{vmatrix} \cos s & -r\sin s \\ \sin s\sin t & r\cos s\sin t \end{vmatrix} + r\sin s\cos t\begin{vmatrix} \cos s & -r\sin s \\ \sin s\cos t & r\cos s\cos t \end{vmatrix} \\ = & r\sin s\sin^2 t\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} + r\sin s\cos^2 t\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} \\ = & r\sin s(\sin^2 t+\cos^2 t)\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} \\ = & r\sin s\times 1\times r \\ = & r^2\sin s \end{aligned}$$

Thus, we have indeed the formula in the statement. □

Let us work out now the general spherical coordinate formula, in arbitrary N dimensions. The formula here, which generalizes those at $N = 2, 3$, is as follows:

THEOREM 15.11. *We have spherical coordinates in* N *dimensions,*

$$\begin{cases} x_1 &= r\cos t_1 \\ x_2 &= r\sin t_1\cos t_2 \\ \vdots & \\ x_{N-1} &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\cos t_{N-1} \\ x_N &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\sin t_{N-1} \end{cases}$$

the corresponding Jacobian being given by the following formula,

$$J(r,t) = r^{N-1}\sin^{N-2}t_1\sin^{N-3}t_2\ \ldots\ \sin^2 t_{N-3}\sin t_{N-2}$$

and with this generalizing the known formulae at $N = 2, 3$.

PROOF. As before, the fact that we have spherical coordinates is clear. Regarding now the Jacobian, also as before, by developing over the last column, we have:

$$\begin{aligned} J_N &= r\sin t_1\ldots\sin t_{N-2}\sin t_{N-1}\times\sin t_{N-1}J_{N-1}\\ &+ r\sin t_1\ldots\sin t_{N-2}\cos t_{N-1}\times\cos t_{N-1}J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}(\sin^2 t_{N-1}+\cos^2 t_{N-1})J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}J_{N-1}\end{aligned}$$

Thus, we obtain the formula in the statement, by recurrence. □

As a comment here, the above convention for spherical coordinates is one among many, designed to best work in arbitrary N dimensions. Also, in what regards the precise range of the angles $t_1,\ldots,t_{N-1}$, we will leave this to you, as an instructive exercise.

As an application, let us compute the volumes of spheres. For this purpose, we must understand how the products of coordinates integrate over spheres. Let us start with the case $N=2$. Here the sphere is the unit circle $\mathbb{T}$, and with $z=e^{it}$ the coordinates are $\cos t,\sin t$. We can first integrate arbitrary powers of these coordinates, as follows:

PROPOSITION 15.12. *We have the following formulae,*

$$\int_0^{\pi/2}\cos^p t\,dt=\int_0^{\pi/2}\sin^p t\,dt=\left(\frac{\pi}{2}\right)^{\varepsilon(p)}\frac{p!!}{(p+1)!!}$$

where $\varepsilon(p)=1$ if p is even, and $\varepsilon(p)=0$ if p is odd, and where

$$m!!=(m-1)(m-3)(m-5)\ldots$$

with the product ending at 2 if m is odd, and ending at 1 if m is even.

PROOF. Let us first compute the integral on the left in the statement:

$$I_p=\int_0^{\pi/2}\cos^p t\,dt$$

We do this by partial integration. We have the following formula:

$$\begin{aligned}(\cos^p t\sin t)' &= p\cos^{p-1}t(-\sin t)\sin t+\cos^p t\cos t\\ &= p\cos^{p+1}t-p\cos^{p-1}t+\cos^{p+1}t\\ &= (p+1)\cos^{p+1}t-p\cos^{p-1}t\end{aligned}$$

By integrating between 0 and $\pi/2$, we obtain the following formula:

$$(p+1)I_{p+1}=pI_{p-1}$$

Thus we can compute I_p by recurrence, and we obtain:

$$\begin{aligned} I_p &= \frac{p-1}{p} I_{p-2} \\ &= \frac{p-1}{p} \cdot \frac{p-3}{p-2} I_{p-4} \\ &= \frac{p-1}{p} \cdot \frac{p-3}{p-2} \cdot \frac{p-5}{p-4} I_{p-6} \\ &\vdots \\ &= \frac{p!!}{(p+1)!!} I_{1-\varepsilon(p)} \end{aligned}$$

But $I_0 = \frac{\pi}{2}$ and $I_1 = 1$, so we get the result. As for the second formula, this follows from the first one, with $t = \frac{\pi}{2} - s$. Thus, we have proved both formulae in the statement. □

We can now compute the volume of the sphere, as follows:

THEOREM 15.13. *The volume of the unit sphere in $\mathbb{R}^N$ is given by*

$$V = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!}$$

with our usual convention $N!! = (N-1)(N-3)(N-5)\ldots$

PROOF. Let us denote by B^+ the positive part of the unit sphere, or rather unit ball B, obtained by cutting this unit ball in 2^N parts. At the level of volumes, we have:

$$V = 2^N V^+$$

We have the following computation, using spherical coordinates:

$$\begin{aligned} V^+ &= \int_{B^+} 1 \\ &= \int_0^1 \int_0^{\pi/2} \ldots \int_0^{\pi/2} r^{N-1} \sin^{N-2} t_1 \ldots \sin t_{N-2}\, dr dt_1 \ldots dt_{N-1} \\ &= \int_0^1 r^{N-1}\, dr \int_0^{\pi/2} \sin^{N-2} t_1\, dt_1 \ldots \int_0^{\pi/2} \sin t_{N-2} dt_{N-2} \int_0^{\pi/2} 1 dt_{N-1} \\ &= \frac{1}{N} \times \left(\frac{\pi}{2}\right)^{[N/2]} \times \frac{(N-2)!!}{(N-1)!!} \cdot \frac{(N-3)!!}{(N-2)!!} \ldots \frac{2!!}{3!!} \cdot \frac{1!!}{2!!} \cdot 1 \\ &= \frac{1}{N} \times \left(\frac{\pi}{2}\right)^{[N/2]} \times \frac{1}{(N-1)!!} \\ &= \left(\frac{\pi}{2}\right)^{[N/2]} \frac{1}{(N+1)!!} \end{aligned}$$

Here we have used the following formula, for computing the exponent of $\pi/2$:

$$\begin{aligned}\varepsilon(0)+\varepsilon(1)+\varepsilon(2)+\ldots+\varepsilon(N-2) &= 1+0+1+\ldots+\varepsilon(N-2)\\ &= \left[\frac{N-2}{2}\right]+1\\ &= \left[\frac{N}{2}\right]\end{aligned}$$

Thus, we obtain the formula in the statement. □

As main particular cases of the above formula, we have:

THEOREM 15.14. *The volumes of the low-dimensional spheres are as follows:*

(1) *At $N=1$, the length of the unit interval is $V=2$.*
(2) *At $N=2$, the area of the unit disk is $V=\pi$.*
(3) *At $N=3$, the volume of the unit sphere is $V=\frac{4\pi}{3}$*
(4) *At $N=4$, the volume of the corresponding unit sphere is $V=\frac{\pi^2}{2}$.*

PROOF. Some of these results are well-known, but we can obtain all of them as particular cases of the general formula in Theorem 15.13, as follows:

(1) At $N=1$ we obtain $V=1\cdot\frac{2}{1}=2$.

(2) At $N=2$ we obtain $V=\frac{\pi}{2}\cdot\frac{4}{2}=\pi$.

(3) At $N=3$ we obtain $V=\frac{\pi}{2}\cdot\frac{8}{3}=\frac{4\pi}{3}$.

(4) At $N=4$ we obtain $V=\frac{\pi^2}{4}\cdot\frac{16}{8}=\frac{\pi^2}{2}$. □

The formula in Theorem 15.13 is certainly nice, but in practice, we would like to have estimates for that sphere volumes too. For this purpose, we will need:

THEOREM 15.15. *We have the Stirling formula*

$$N!\simeq\left(\frac{N}{e}\right)^N\sqrt{2\pi N}$$

valid in the $N\to\infty$ limit.

PROOF. This is something quite tricky, the idea being as follows:

(1) Let us first see what we can get with Riemann sums. We have:

$$\begin{aligned}\log(N!) &= \sum_{k=1}^{N}\log k\\ &\approx \int_1^N\log x\,dx\\ &= N\log N-N+1\end{aligned}$$

By exponentiating, this gives the following estimate, which is not bad:

$$N! \approx \left(\frac{N}{e}\right)^N \cdot e$$

(2) We can improve our estimate by replacing the rectangles from the Riemann sum approach to the integrals by trapezoids. In practice, this gives the following estimate:

$$\begin{aligned}
\log(N!) &= \sum_{k=1}^{N} \log k \\
&\approx \int_1^N \log x\, dx + \frac{\log 1 + \log N}{2} \\
&= N \log N - N + 1 + \frac{\log N}{2}
\end{aligned}$$

By exponentiating, this gives the following estimate, which gets us closer:

$$N! \approx \left(\frac{N}{e}\right)^N \cdot e \cdot \sqrt{N}$$

(3) In order to conclude, we must take some kind of mathematical magnifier, and carefully estimate the error made in (2). Fortunately, this mathematical magnifier exists, called Euler-Maclaurin formula, and after some computations, this leads to:

$$N! \simeq \left(\frac{N}{e}\right)^N \sqrt{2\pi N}$$

(4) However, all this remains a bit complicated, so we would like to present now an alternative approach to (3), which also misses some details, but better does the job, explaining where the $\sqrt{2\pi}$ factor comes from. First, by partial integration we have:

$$N! = \int_0^\infty x^N e^{-x} dx$$

Since the integrand is sharply peaked at $x = N$, as you can see by computing the derivative of $\log(x^N e^{-x})$, this suggests writing $x = N + y$, and we obtain:

$$\begin{aligned}
\log(x^N e^{-x}) &= N \log x - x \\
&= N \log(N + y) - (N + y) \\
&= N \log N + N \log\left(1 + \frac{y}{N}\right) - (N + y) \\
&\simeq N \log N + N\left(\frac{y}{N} - \frac{y^2}{2N^2}\right) - (N + y) \\
&= N \log N - N - \frac{y^2}{2N}
\end{aligned}$$

By exponentiating, we obtain from this the following estimate:

$$x^N e^{-x} \simeq \left(\frac{N}{e}\right)^N e^{-y^2/2N}$$

Now by integrating, and using the Gauss formula, we obtain from this:

$$\begin{aligned} N! &= \int_0^\infty x^N e^{-x} dx \\ &\simeq \int_{-N}^N \left(\frac{N}{e}\right)^N e^{-y^2/2N}\, dy \\ &\simeq \left(\frac{N}{e}\right)^N \int_{\mathbb{R}} e^{-y^2/2N}\, dy \\ &= \left(\frac{N}{e}\right)^N \sqrt{2N} \int_{\mathbb{R}} e^{-z^2}\, dz \\ &= \left(\frac{N}{e}\right)^N \sqrt{2\pi N} \end{aligned}$$

Thus, we have proved the Stirling formula, as formulated in the statement. □

With the above formula in hand, we have many useful applications, such as:

PROPOSITION 15.16. *We have the following estimate for binomial coefficients,*

$$\binom{N}{K} \simeq \left(\frac{1}{t^t(1-t)^{1-t}}\right)^N \frac{1}{\sqrt{2\pi t(1-t)N}}$$

in the $K \simeq tN \to \infty$ limit, with $t \in (0,1]$. In particular we have

$$\binom{2N}{N} \simeq \frac{4^N}{\sqrt{\pi N}}$$

in the $N \to \infty$ limit, for the central binomial coefficients.

PROOF. All this is very standard, by using the Stirling formula etablished above, for the various factorials which appear, the idea being as follows:

(1) This follows from the definition of the binomial coefficients, namely:

$$\begin{aligned}\binom{N}{K} &= \frac{N!}{K!(N-K)!}\\ &\simeq \left(\frac{N}{e}\right)^N\sqrt{2\pi N}\left(\frac{e}{K}\right)^K\frac{1}{\sqrt{2\pi K}}\left(\frac{e}{N-K}\right)^{N-K}\frac{1}{\sqrt{2\pi(N-K)}}\\ &= \frac{N^N}{K^K(N-K)^{N-K}}\sqrt{\frac{N}{2\pi K(N-K)}}\\ &\simeq \frac{N^N}{(tN)^{tN}((1-t)N)^{(1-t)N}}\sqrt{\frac{N}{2\pi tN(1-t)N}}\\ &= \left(\frac{1}{t^t(1-t)^{1-t}}\right)^N\frac{1}{\sqrt{2\pi t(1-t)N}}\end{aligned}$$

Thus, we are led to the conclusion in the statement.

(2) This estimate follows from a similar computation, as follows:

$$\begin{aligned}\binom{2N}{N} &= \frac{(2N)!}{N!N!}\\ &\simeq \left(\frac{2N}{e}\right)^{2N}\sqrt{4\pi N}\left(\frac{e}{N}\right)^{2N}\frac{1}{2\pi N}\\ &= \frac{4^N}{\sqrt{\pi N}}\end{aligned}$$

Alternatively, we can take $t=1/2$ in (1), then rescale. Indeed, we have:

$$\begin{aligned}\binom{N}{[N/2]} &\simeq \left(\frac{1}{(\frac{1}{2})^{1/2}(\frac{1}{2})^{1/2}}\right)^N\frac{1}{\sqrt{2\pi\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot N}}\\ &= 2^N\sqrt{\frac{2}{\pi N}}\end{aligned}$$

Thus with the change $N\to 2N$ we obtain the formula in the statement. □

We can now estimate the volumes of the spheres, as follows:

THEOREM 15.17. *The volume of the unit sphere in $\mathbb{R}^N$ is given by*

$$V\simeq\left(\frac{2\pi e}{N}\right)^{N/2}\frac{1}{\sqrt{\pi N}}$$

in the $N\to\infty$ limit.

PROOF. We use the formula for V found in Theorem 15.13, namely:

$$V=\left(\frac{\pi}{2}\right)^{[N/2]}\frac{2^N}{(N+1)!!}$$

In the case where N is even, the estimate goes as follows:

$$\begin{aligned}V&=\left(\frac{\pi}{2}\right)^{N/2}\frac{2^N}{(N+1)!!}\\&\simeq\left(\frac{\pi}{2}\right)^{N/2}2^N\left(\frac{e}{N}\right)^{N/2}\frac{1}{\sqrt{\pi N}}\\&=\left(\frac{2\pi e}{N}\right)^{N/2}\frac{1}{\sqrt{\pi N}}\end{aligned}$$

In the case where N is odd, the estimate goes as follows:

$$\begin{aligned}V&=\left(\frac{\pi}{2}\right)^{(N-1)/2}\frac{2^N}{(N+1)!!}\\&\simeq\left(\frac{\pi}{2}\right)^{(N-1)/2}2^N\left(\frac{e}{N}\right)^{N/2}\frac{1}{\sqrt{2N}}\\&=\sqrt{\frac{2}{\pi}}\left(\frac{2\pi e}{N}\right)^{N/2}\frac{1}{\sqrt{2N}}\\&=\left(\frac{2\pi e}{N}\right)^{N/2}\frac{1}{\sqrt{\pi N}}\end{aligned}$$

Thus, we are led to the uniform formula in the statement. □

Getting back now to our main result so far, Theorem 15.13, we can compute in the same way the area of the sphere, the result being as follows:

THEOREM 15.18. *The area of the unit sphere in $\mathbb{R}^N$ is given by*

$$A=\left(\frac{\pi}{2}\right)^{[N/2]}\frac{2^N}{(N-1)!!}$$

with the our usual convention for double factorials, namely:

$$N!!=(N-1)(N-3)(N-5)\ldots$$

In particular, at $N=2,3,4$ we obtain respectively $A=2\pi,4\pi,2\pi^2$.

PROOF. Regarding the first assertion, there is no need to compute again, because the formula in the statement can be deduced from Theorem 15.13, as follows:

(1) We can either use the "pizza" argument from chapter 6, which shows that the area and volume of the sphere in $\mathbb{R}^N$ are related by the following formula:

$$A=N\cdot V$$

Together with the formula in Theorem 15.13 for V, this gives the result.

(2) Or, we can start the computation in the same way as we started the proof of Theorem 15.13, the beginning of this computation being as follows:

$$vol(S^+) = \int_0^{\pi/2} \ldots \int_0^{\pi/2} \sin^{N-2} t_1 \ldots \sin t_{N-2}\, dt_1 \ldots dt_{N-1}$$

Now by comparing with the beginning of the proof of Theorem 15.13, the only thing that changes is the following quantity, which now dissapears:

$$\int_0^1 r^{N-1}\, dr = \frac{1}{N}$$

Thus, we have $vol(S^+) = N \cdot vol(B^+)$, and so we obtain the following formula:

$$vol(S) = N \cdot vol(B)$$

But this means $A = N \cdot V$, and together with the formula in Theorem 15.13 for V, this gives the result. As for the last assertion, this can be either worked out directly, or deduced from the results for volumes that we have so far, by multiplying by N. □

15d. Normal variables

We have kept the best for the end. As a starting point, we have:

DEFINITION 15.19. *Let X be a probability space, that is, a space with a probability measure, and with the corresponding integration denoted E, and called expectation.*

(1) *The random variables are the real functions $f \in L^\infty(X)$.*
(2) *The moments of such a variable are the numbers $M_k(f) = E(f^k)$.*
(3) *The law of such a variable is the measure given by $M_k(f) = \int_{\mathbb{R}} x^k d\mu_f(x)$.*

Here the fact that a measure μ_f as above exists indeed is not exactly trivial. But we can do this by looking at formulae of the following type:

$$E(\varphi(f)) = \int_{\mathbb{R}} \varphi(x) d\mu_f(x)$$

Indeed, having this for monomials $\varphi(x) = x^n$, as above, is the same as having it for polynomials $\varphi \in \mathbb{R}[X]$, which in turn is the same as having it for the characteristic functions $\varphi = \chi_I$ of measurable sets $I \subset \mathbb{R}$. Thus, in the end, what we need is:

$$P(f \in I) = \mu_f(I)$$

But this formula can serve as a definition for μ_f, and we are done.

Regarding now independence, we can formulate here the following definition:

DEFINITION 15.20. *Two variables $f, g \in L^\infty(X)$ are called independent when*

$$E(f^k g^l) = E(f^k)\,E(g^l)$$

happens, for any $k, l \in \mathbb{N}$.

Again, this definition hides some non-trivial things, the idea being a bit as before, namely that of looking at formulae of the following type:

$$E[\varphi(f)\psi(g)] = E[\varphi(f)]\,E[\psi(g)]$$

To be more precise, passing as before from monomials to polynomials, then to characteristic functions, we are led to the usual definition of independence, namely:

$$P(f \in I, g \in J) = P(f \in I)\,P(g \in J)$$

As a first result now, which is something very standard, we have:

THEOREM 15.21. *Assuming that $f, g \in L^\infty(X)$ are independent, we have*

$$\mu_{f+g} = \mu_f * \mu_g$$

where $*$ is the convolution of real probability measures.

PROOF. We have the following computation, using the independence of f, g:

$$\int_{\mathbb{R}} x^k d\mu_{f+g}(x) = E((f+g)^k) = \sum_r \binom{k}{r} M_r(f) M_{k-r}(g)$$

On the other hand, we have as well the following computation:

$$\begin{aligned}\int_{\mathbb{R}} x^k d(\mu_f * \mu_g)(x) &= \int_{\mathbb{R}\times\mathbb{R}} (x+y)^k d\mu_f(x) d\mu_g(y)\\ &= \sum_r \binom{k}{r} M_r(f) M_{k-r}(g)\end{aligned}$$

Thus μ_{f+g} and $\mu_f * \mu_g$ have the same moments, so they coincide, as claimed. □

As a second result on independence, which is more advanced, we have:

THEOREM 15.22. *Assuming that $f, g \in L^\infty(X)$ are independent, we have*

$$F_{f+g} = F_f F_g$$

where $F_f(x) = E(e^{ixf})$ is the Fourier transform.

PROOF. This is something which is very standard too, coming from:

$$\begin{aligned} F_{f+g}(x) &= \int_{\mathbb{R}} e^{ixz} d(\mu_f * \mu_g)(z) \\ &= \int_{\mathbb{R}\times\mathbb{R}} e^{ix(z+t)} d\mu_f(z) d\mu_g(t) \\ &= \int_{\mathbb{R}} e^{ixz} d\mu_f(z) \int_{\mathbb{R}} e^{ixt} d\mu_g(t) \\ &= F_f(x) F_g(x) \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Let us introduce now the normal laws. This can be done as follows:

DEFINITION 15.23. *The normal law of parameter* 1 *is the following measure:*

$$g_1 = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

More generally, the normal law of parameter $t > 0$ *is the following measure:*

$$g_t = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t} dx$$

These are also called Gaussian distributions, with "g" standing for Gauss.

Observe that the above laws have indeed mass 1, as they should. This follows indeed from the Gauss formula, which gives, with $x = \sqrt{2t}\, y$:

$$\begin{aligned} \int_{\mathbb{R}} e^{-x^2/2t} dx &= \int_{\mathbb{R}} e^{-y^2} \sqrt{2t}\, dy \\ &= \sqrt{2t} \int_{\mathbb{R}} e^{-y^2} dy \\ &= \sqrt{2t} \times \sqrt{\pi} \\ &= \sqrt{2\pi t} \end{aligned}$$

Generally speaking, the normal laws appear as bit everywhere, in real life. The reasons behind this phenomenon come from the Central Limit Theorem (CLT), that we will explain in a moment, after developing some general theory. As a first result, we have:

PROPOSITION 15.24. *We have the variance formula*

$$V(g_t) = t$$

valid for any $t > 0$.

PROOF. The first moment is 0, because our normal law g_t is centered. As for the second moment, this can be computed as follows:

$$\begin{aligned} M_2 &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} x^2 e^{-x^2/2t}dx \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} (tx)\left(-e^{-x^2/2t}\right)' dx \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} te^{-x^2/2t}dx \\ &= t \end{aligned}$$

We conclude from this that the variance is $V = M_2 = t$. □

Here is another result, which is the key one for the study of the normal laws:

THEOREM 15.25. *We have the following formula, valid for any $t > 0$:*

$$F_{g_t}(x) = e^{-tx^2/2}$$

In particular, the normal laws satisfy $g_s * g_t = g_{s+t}$, for any $s, t > 0$.

PROOF. The Fourier transform formula can be established as follows:

$$\begin{aligned} F_{g_t}(x) &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} e^{-y^2/2t+ixy}dy \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} e^{-(y/\sqrt{2t}-\sqrt{t/2}ix)^2-tx^2/2}dy \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} e^{-z^2-tx^2/2}\sqrt{2t}dz \\ &= \frac{1}{\sqrt{\pi}}e^{-tx^2/2}\int_{\mathbb{R}} e^{-z^2}dz \\ &= \frac{1}{\sqrt{\pi}}e^{-tx^2/2}\cdot\sqrt{\pi} \\ &= e^{-tx^2/2} \end{aligned}$$

As for the last assertion, this follows from the fact that $\log F_{g_t}$ is linear in t. □

We are now ready to state and prove the CLT, as follows:

THEOREM 15.26 (CLT). *Given random variables $f_1, f_2, f_3, \ldots \in L^\infty(X)$ which are i.i.d., centered, and with variance $t > 0$, we have, with $n \to \infty$, in moments,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f_i \sim g_t$$

where g_t is the Gaussian law of parameter t, having as density $\frac{1}{\sqrt{2\pi t}}e^{-y^2/2t}dy$.

PROOF. In terms of moments, we have the following formula:

$$\begin{aligned} F_f(x) &= E\left(\sum_{k=0}^{\infty}\frac{(ixf)^k}{k!}\right)\\ &= \sum_{k=0}^{\infty}\frac{(ix)^kE(f^k)}{k!}\\ &= \sum_{k=0}^{\infty}\frac{i^kM_k(f)}{k!}\,x^k \end{aligned}$$

Thus, the Fourier transform of the variable in the statement is:

$$\begin{aligned} F(x) &= \left[F_f\left(\frac{x}{\sqrt{n}}\right)\right]^n\\ &= \left[1-\frac{tx^2}{2n}+O(n^{-2})\right]^n\\ &\simeq \left[1-\frac{tx^2}{2n}\right]^n\\ &\simeq e^{-tx^2/2} \end{aligned}$$

But this latter function being the Fourier transform of g_t, we obtain the result. □

Finally, regarding the moments, we have here the following result:

THEOREM 15.27. *The even moments of the normal law are the numbers*

$$M_k(g_t)=t^{k/2}\times k!!$$

where $k!!=(k-1)(k-3)(k-5)\ldots$, *and the odd moments vanish.*

PROOF. We have the following computation, valid for any integer $k\in\mathbb{N}$:

$$\begin{aligned} M_k &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}}y^ke^{-y^2/2t}dy\\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}}(ty^{k-1})\left(-e^{-y^2/2t}\right)'dy\\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}}t(k-1)y^{k-2}e^{-y^2/2t}dy\\ &= t(k-1)\times\frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}}y^{k-2}e^{-y^2/2t}dy\\ &= t(k-1)M_{k-2} \end{aligned}$$

Now recall from the proof of Proposition 15.24 that we have $M_0=1$, $M_1=0$. Thus by recurrence, we are led to the formula in the statement. □

As a last topic for this chapter, let us discuss now the complex versions of the normal variables. To start with, we have the following definition:

DEFINITION 15.28. *The complex Gaussian law of parameter $t > 0$ is*

$$G_t = law\left(\frac{1}{\sqrt{2}}(a + ib)\right)$$

where a, b are independent, each following the law g_t.

As in the real case, these measures form convolution semigroups:

THEOREM 15.29. *The complex Gaussian laws have the property*

$$G_s * G_t = G_{s+t}$$

for any $s, t > 0$, and so they form a convolution semigroup.

PROOF. This follows indeed from the real result, namely $g_s * g_t = g_{s+t}$, established in Theorem 15.25, simply by taking real and imaginary parts. □

We have as well the following complex analogue of the CLT:

THEOREM 15.30 (CCLT). *Given complex variables $f_1, f_2, f_3, \ldots \in L^\infty(X)$ which are i.i.d., centered, and with common variance $t > 0$, we have*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f_i \sim G_t$$

with $n \to \infty$, in moments.

PROOF. This follows indeed from the real CLT, established in Theorem 15.26, simply by taking the real and imaginary parts of all variables involved. □

Regarding now the moments, the situation is more complicated than in the real case, because in order to have good results, we have to deal with both the complex variables, and their conjugates. Let us formulate the following definition:

DEFINITION 15.31. *The moments a complex variable $f \in L^\infty(X)$ are the numbers*

$$M_k = E(f^k)$$

depending on colored integers $k = \circ \bullet \bullet \circ \ldots$, with the conventions

$$f^\emptyset = 1 \quad , \quad f^\circ = f \quad , \quad f^\bullet = \bar{f}$$

and multiplicativity, in order to define the colored powers f^k.

Observe that, since $f, \bar{f}$ commute, we can permute terms, and restrict the attention to exponents of type $k = \ldots \circ \circ \circ \bullet \bullet \bullet \bullet \ldots$, if we want to. However, our result about the complex Gaussian laws, and other complex laws, later on, will actually look better without doing is, so we will use Definition 15.31 as stated. We have:

THEOREM 15.32. *The moments of the complex normal law are given by*

$$M_k(G_t) = \begin{cases} t^p p! & (k \text{ uniform, of length } 2p) \\ 0 & (k \text{ not uniform}) \end{cases}$$

where $k = \circ\bullet\bullet\circ\ldots$ is called uniform when it contains the same number of $\circ$ and $\bullet$.

PROOF. We must compute the moments, with respect to colored integer exponents $k = \circ\bullet\bullet\circ\ldots$, of the variable from Definition 15.28, namely:

$$f = \frac{1}{\sqrt{2}}(a + ib)$$

We can assume that we are in the case $t = 1$, and the proof here goes as follows:

(1) As a first observation, in the case where our exponent $k = \circ\bullet\bullet\circ\ldots$ is not uniform, a standard rotation argument shows that the corresponding moment of f vanishes. To be more precise, the variable $f' = wf$ is complex Gaussian too, for any complex number $w \in \mathbb{T}$, and from $M_k(f) = M_k(f')$ we obtain $M_k(f) = 0$, in this case.

(2) In the uniform case now, where the exponent $k = \circ\bullet\bullet\circ\ldots$ consists of p copies of $\circ$ and p copies of $\bullet$, the corresponding moment can be computed as follows:

$$\begin{aligned} M_k &= \int (f\bar{f})^p \\ &= \frac{1}{2^p}\int (a^2 + b^2)^p \\ &= \frac{1}{2^p}\sum_r \binom{p}{r}\int a^{2r}\int b^{2p-2r} \\ &= \frac{1}{2^p}\sum_r \binom{p}{r}(2r)!!(2p-2r)!! \\ &= \frac{1}{2^p}\sum_r \frac{p!}{r!(p-r)!}\cdot\frac{(2r)!}{2^r r!}\cdot\frac{(2p-2r)!}{2^{p-r}(p-r)!} \\ &= \frac{p!}{4^p}\sum_r \binom{2r}{r}\binom{2p-2r}{p-r} \end{aligned}$$

(3) In order to finish now the computation, let us recall that we have the following formula, coming from the generalized binomial formula, or from the Taylor formula:

$$\frac{1}{\sqrt{1+t}} = \sum_{q=0}^{\infty}\binom{2q}{q}\left(\frac{-t}{4}\right)^q$$

By taking the square of this series, we obtain the following formula:

$$\begin{aligned}\frac{1}{1+t} &= \sum_{qr}\binom{2q}{q}\binom{2r}{r}\left(\frac{-t}{4}\right)^{q+r}\\ &= \sum_{p}\left(\frac{-t}{4}\right)^{p}\sum_{r}\binom{2r}{r}\binom{2p-2r}{p-r}\end{aligned}$$

Now by looking at the coefficient of t^p on both sides, we conclude that the sum on the right equals 4^p. Thus, we can finish the moment computation in (2), as follows:

$$M_k=\frac{p!}{4^p}\times 4^p=p!$$

We are therefore led to the conclusion in the statement. □

15e. Exercises

Exercises:

Exercise 15.33.

Exercise 15.34.

Exercise 15.35.

Exercise 15.36.

Exercise 15.37.

Exercise 15.38.

Exercise 15.39.

Exercise 15.40.

Bonus exercise.

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 in the above. And we are looking at a particle whose position is described by a function $x = x(t)$. Then, it is quite clear that the speed of this particle should be described by the first derivative $v = x'(t)$, and that the acceleration of the particle should be described by the second derivative $a = v'(t) = x''(t)$.

Summarizing, with Newton's theory of derivatives, as we learned it in the previous chapter, we can certainly do some mathematics for the motion of bodies. But, for these bodies to move, we need them to be acted upon by some forces, right? The simplest such force is gravity, and in our present, modest 1 dimensional setting, we have:

THEOREM 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:

In order to reach to calculus as we know it, we must peform a rotation, as to have all this happening on the Ox axis. By doing this, and assuming that M is fixed at 0, our picture becomes as follows, with the attached numbers being now the coordinates:

Now comes the physics. The gravitational force exterted by M, which is fixed in our formalism, on the object m which moves, is subject to the following equations:

$$F = -G \cdot \frac{Mm}{x^2} \quad , \quad F = ma \quad , \quad a = \dot{v} \quad , \quad v = \dot{x}$$

To be more precise, in the first equation $G \simeq 6.674 \times 10^{-11}$ 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. We can use here the field equation for the gravity, namely:

$$f = \frac{K}{d^2}$$

This equation, with $d = ||x||$, describes the magnitude f of the acceleration a of our moving object m. Now since a points towards 0, which is opposite to x, we have:

$$a = -\frac{K}{d^2} \cdot \frac{x}{||x||} = -\frac{Kx}{||x||^3}$$

Moreover, since the acceleration a is by definition the second derivative of the position vector x, the equation of motion of our object m is as follows:

$$\ddot{x} = -\frac{Kx}{||x||^3}$$

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 basic 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:

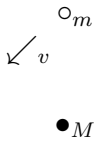
$$G = 6.674 \times 10^{-11} \quad , \quad M = 5.972 \times 10^{24} \quad , \quad R = 6.371 \times 10^6$$

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:

$$x = pt$$

$$y \simeq R + qt - \frac{gt^2}{2}$$

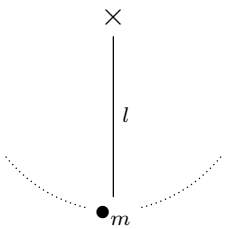
From the first equation we get $t = x/p$, and by substituting into the second:

$$y \simeq R + \frac{qx}{p} - \frac{gx^2}{2p^2}$$

We recognize here the approximate equation of a parabola, and we are done. □

Along the same lines, let us 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:

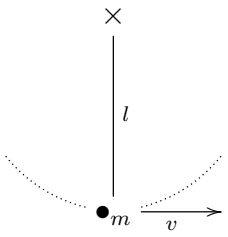
$$V' = -F$$

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:

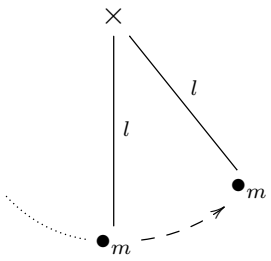
THEOREM 16.6. *For a pendulum starting with speed v from the equilibrium position,*



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:



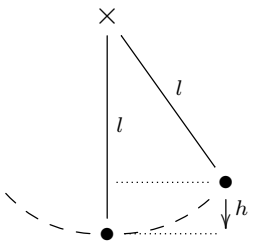
The distance traveled is then $x = l\theta$. As for the force acting, this is $F_{total} = mg$ oriented downwards, with the component alongside x being given by:

$$\begin{aligned} F &= -||F_{total}|| \sin\theta \\ &= -mg\sin\theta \\ &= -mg\sin\left(\frac{x}{l}\right) \end{aligned}$$

(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:

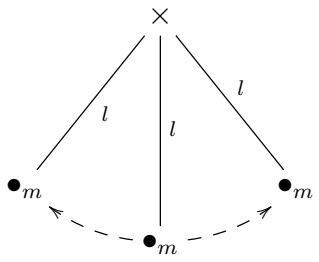
$$\begin{aligned} V' = -\mathcal{F} \quad &\implies \quad V' = mg \\ &\implies \quad V = mgh \\ &\implies \quad V = mgl(1 - \cos\theta) \end{aligned}$$

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:

$$\begin{aligned} E < \sup_\theta V = 2mgl &\iff \frac{mv^2}{2} < 2mgl \\ &\iff v^2 < 4gl \end{aligned}$$

In this case, the motion will be confined between two angles $-\theta, \theta$, as follows:



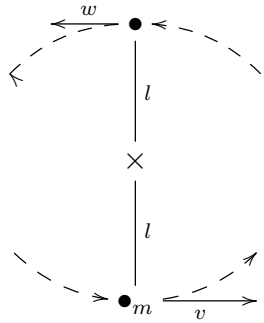
To be more precise here, the two extreme angles $-\theta, \theta \in (-\pi, \pi)$ can be explicitly computed, as being solutions of the following equation:

$$\begin{aligned} V = E &\iff mgl(1 - \cos\theta) = \frac{mv^2}{2} \\ &\iff 1 - \cos\theta = \frac{v^2}{2gl} \end{aligned}$$

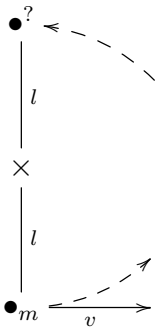
(5) Regarding now the case $v^2 > 4gl$, here the bob will certainly reach the upwards position, with the speed $w > 0$ there being given by the following formula:

$$\begin{aligned} \frac{mw^2}{2} = E - 2mgl &\implies \frac{mw^2}{2} = \frac{mv^2}{2} - 2mgl \\ &\implies w^2 = v^2 - 4gl \\ &\implies w = \sqrt{v^2 - 4gl} \end{aligned}$$

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*

$$C = \left\{(x, y) \in \mathbb{R}^2 \middle| P(x, y) = 0\right\}$$

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:

$$\begin{aligned} b = 0 &\iff 2\dot{r}\dot{\theta} + r\ddot{\theta} = 0 \\ &\iff \frac{\ddot{\theta}}{\dot{\theta}} = -2\frac{\dot{r}}{r} \\ &\iff (\log\dot{\theta})' = (-2\log r)' \\ &\iff \log\dot{\theta} = -2\log r + c \\ &\iff \dot{\theta} = \frac{\lambda}{r^2} \end{aligned}$$

As for the first equation the we found, namely $a + K/r^2 = 0$, 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. □

Still with me, I hope, after all these computations. Good work that we did.

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 first 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:

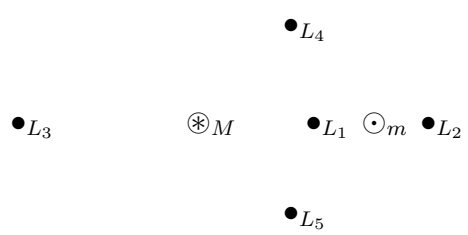
$$\ddot{r}_0 = \frac{Kc}{R^2}\left(\frac{1}{R} - \frac{1}{c}\right) = \frac{Kc}{R^2} \cdot \frac{\varepsilon}{c} = \frac{\varepsilon K}{R^2}$$

(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

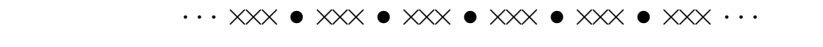
As more physics, we can talk as well about waves in 1 dimension, as follows:

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:



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:

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,*

$$\ddot{\varphi} = v^2 \Delta\varphi$$

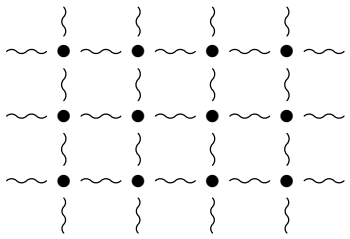
with $v > 0$ being the propagation speed of the wave, and with Δ given by

$$\Delta\varphi = \sum_{i=1}^{N} \frac{d^2\varphi}{dx_i^2}$$

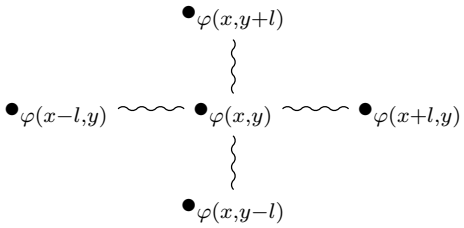
being the Laplace operator, playing the role of a numeric second derivative.

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. □

Let us record as well the following alternative approach to the waves:

THEOREM 16.12. *The wave equation can be understood as well directly, as a wave propagating through a linear elastic medium, via stress.*

PROOF. This is indeed something very standard, with the $N = 1$ picture involving a pulse propagating through a bar, and with at $N \geq 2$ something of a similar type:

(1) In the 1D case, assume that we have a bar of length L, made of linear elastic material. The stiffness of the bar is then the following quantity, with A being the cross-sectional area, and with E being the Young modulus of the material:

$$K = \frac{EA}{L}$$

Now when sending a pulse, this propagates as follows, M being the total mass:

$$\ddot{\varphi} = \frac{EAL}{M} \cdot \varphi''(x)$$

Bur since $V = AL$ is the volume, with $\rho = M/V$ being the density, we have:

$$\ddot{\varphi} = \frac{E}{\rho} \cdot \varphi''(x)$$

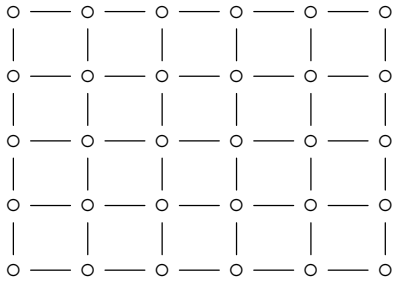
Thus, as a conclusion, the wave propagates with speed $v = \sqrt{E/\rho}$.

(2) In two or more dimensions, the study, and final result, are similar. □

As a continuation of this, the next question which appears is that of understanding how exactly the various mechanical waves propagate through solids, liquids and gases, and what corrections to the wave equation are needed, in each case.

16d. Heat equation

Time for heat. The simplest heat diffusion question, studied and understood since long, concerns a container containing two gases, having initial different temperatures $T_1 < T_2$, separated by a membrane. Heat transfer goes on, in this setting, and obviously, we can model this by focusing on the membrane, with a basic grid model for it:



There is some sort of "game" played by the two gases, over this grid, and we can model this, and then recover the known results about heat diffusion, in this setting.

At a more advanced level, we can remove the membrane. Again, there is some sort of "game" here, played by the two gases, which can be 2D or 3D, depending on modeling. Also, in this setting, we can actually keep the membrane, but allow it to inflate.

Let us go now into heavier, fully powerful models and equations for the heat diffusion mechanism, involving this time more advanced mathematics and physics. 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 $\delta > 0$:

$$\varphi(x, t+\delta) = \varphi(x,t) + \frac{\alpha\delta}{l^2}\sum_{x\sim y}[\varphi(y,t) - \varphi(x,t)]$$

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:

$$\frac{\varphi(x,t+\delta) - \varphi(x,t)}{\delta} = \frac{\alpha}{l^2} \sum_{x \sim y} [\varphi(y,t) - \varphi(x,t)]$$

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 in chapter 1, 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:

$$\frac{\varphi(x,t+\delta) - \varphi(x,t)}{\delta} = \alpha \cdot \varphi''(x,t)$$

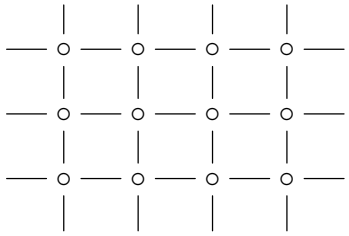
Now with $t \to 0$, we are led in this way to the heat equation, namely:

$$\dot{\varphi}(x,t) = \alpha \cdot \varphi''(x,t)$$

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 modellng 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. □

Many other things can be said, as a continuation of the above.

16e. Exercises

Congratulations for having read this book, and no exercises for this final chapter.

Bibliography

[1] A.A. Abrikosov, Fundamentals of the theory of metals, Dover (1988).

[2] V.I. Arnold, Ordinary differential equations, Springer (1973).

[3] V.I. Arnold, Lectures on partial differential equations, Springer (1997).

[4] V.I. Arnold, Catastrophe theory, Springer (1984).

[5] N.W. Ashcroft and N.D. Mermin, Solid state physics, Saunders College Publ. (1976).

[6] T. Banica, Calculus and applications (2024).

[7] T. Banica, Linear algebra and group theory (2024).

[8] T. Banica, Introduction to modern physics (2025).

[9] G.K. Batchelor, An introduction to fluid dynamics, Cambridge Univ. Press (1967).

[10] M.J. Benton, Vertebrate paleontology, Wiley (1990).

[11] M.J. Benton and D.A.T. Harper, Introduction to paleobiology and the fossil record, Wiley (2009).

[12] S.J. Blundell and K.M. Blundell, Concepts in thermal physics, Oxford Univ. Press (2006).

[13] B. Bollobás, Modern graph theory, Springer (1998).

[14] S.M. Carroll, Spacetime and geometry, Cambridge Univ. Press (2004).

[15] P.M. Chaikin and T.C. Lubensky, Principles of condensed matter physics, Cambridge Univ. Press (1995).

[16] A.R. Choudhuri, Astrophysics for physicists, Cambridge Univ. Press (2012).

[17] J. Clayden, S. Warren and N. Greeves, Organic chemistry, Oxford Univ. Press (2012).

[18] D.D. Clayton, Principles of stellar evolution and nucleosynthesis, Univ. of Chicago Press (1968).

[19] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).

[20] A. Cottrell, An introduction to metallurgy, CRC Press (1997).

[21] C. Darwin, On the origin of species (1859).

[22] P.A. Davidson, Introduction to magnetohydrodynamics, Cambridge Univ. Press (2001).

[23] P.A.M. Dirac, Principles of quantum mechanics, Oxford Univ. Press (1930).

[24] S. Dodelson, Modern cosmology, Academic Press (2003).

[25] S.T. Dougherty, Combinatorics and finite geometry, Springer (2020).

[26] M. Dresher, The mathematics of games of strategy, Dover (1981).

[27] R. Durrett, Probability: theory and examples, Cambridge Univ. Press (1990).

[28] F. Dyson, Origins of life, Cambridge Univ. Press (1984).

[29] A. Einstein, Relativity: the special and the general theory, Dover (1916).

[30] L.C. Evans, Partial differential equations, AMS (1998).

[31] W. Feller, An introduction to probability theory and its applications, Wiley (1950).

[32] E. Fermi, Thermodynamics, Dover (1937).

[33] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics I: mainly mechanics, radiation and heat, Caltech (1963).

[34] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics II: mainly electromagnetism and matter, Caltech (1964).

[35] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics III: quantum mechanics, Caltech (1966).

[36] R.P. Feynman and A.R. Hibbs, Quantum mechanics and path integrals, Dover (1965).

[37] P. Flajolet and R. Sedgewick, Analytic combinatorics, Cambridge Univ. Press (2009).

[38] A.P. French, Special relativity, Taylor and Francis (1968).

[39] J.H. Gillespie, Population genetics, Johns Hopkins Univ. Press (1998).

[40] C. Godsil and G. Royle, Algebraic graph theory, Springer (2001).

[41] H. Goldstein, C. Safko and J. Poole, Classical mechanics, Addison-Wesley (1980).

[42] D.L. Goodstein, States of matter, Dover (1975).

[43] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).

[44] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).

[45] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).

[46] D.J. Griffiths, Revolutions in twentieth-century physics, Cambridge Univ. Press (2012).

[47] V.P. Gupta, Principles and applications of quantum chemistry, Elsevier (2016).

[48] W.A. Harrison, Solid state theory, Dover (1970).

[49] W.A. Harrison, Electronic structure and the properties of solids, Dover (1980).

[50] R.A. Horn and C.R. Johnson, Matrix analysis, Cambridge Univ. Press (1985).

[51] C.E. Housecroft and A.G. Sharpe, Inorganic chemistry, Pearson (2018).

[52] K. Huang, Introduction to statistical physics, CRC Press (2001).

[53] K. Huang, Fundamental forces of nature, World Scientific (2007).

[54] S. Huskey, The skeleton revealed, Johns Hopkins Univ. Press (2017).

[55] L. Hyman, Comparative vertebrate anatomy, Univ. of Chicago Press (1942).

[56] L.P. Kadanoff, Statistical physics: statics, dynamics and renormalization, World Scientific (2000).

[57] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).

[58] C. Kittel, Introduction to solid state physics, Wiley (1953).

[59] D.E. Knuth, The art of computer programming, Addison-Wesley (1968).

[60] M. Kumar, Quantum: Einstein, Bohr, and the great debate about the nature of reality, Norton (2009).

[61] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).

[62] L.D. Landau and E.M. Lifshitz, Mechanics, Pergamon Press (1960).

[63] L.D. Landau and E.M. Lifshitz, The classical theory of fields, Addison-Wesley (1951).

[64] L.D. Landau and E.M. Lifshitz, Quantum mechanics: non-relativistic theory, Pergamon Press (1959).

[65] S. Lang, Algebra, Addison-Wesley (1993).

[66] P. Lax, Linear algebra and its applications, Wiley (2007).

[67] P. Lax, Functional analysis, Wiley (2002).

[68] P. Lax and M.S. Terrell, Calculus with applications, Springer (2013).

[69] P. Lax and M.S. Terrell, Multivariable calculus with applications, Springer (2018).

[70] S. Ling and C. Xing, Coding theory: a first course, Cambridge Univ. Press (2004).

[71] J.P. Lowe and K. Peterson, Quantum chemistry, Elsevier (2005).

[72] S.J. Marshall, The story of the computer: a technical and business history, Create Space Publ. (2022).

[73] M.L. Mehta, Random matrices, Elsevier (2004).

[74] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge Univ. Press (2000).

[75] R.K. Pathria and and P.D. Beale, Statistical mechanics, Elsevier (1972).

[76] T.D. Pollard, W.C. Earnshaw, J. Lippincott-Schwartz and G. Johnson, Cell biology, Elsevier (2022).

[77] J. Preskill, Quantum information and computation, Caltech (1998).

[78] R. Rojas and U. Hashagen, The first computers: history and architectures, MIT Press (2000).

[79] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[80] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[81] W. Rudin, Functional analysis, McGraw-Hill (1973).

[82] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[83] B. Ryden and B.M. Peterson, Foundations of astrophysics, Cambridge Univ. Press (2010).

[84] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).

[85] R. Shankar, Fundamentals of physics I: mechanics, relativity, and thermodynamics, Yale Univ. Press (2014).

[86] R. Shankar, Fundamentals of physics II: electromagnetism, optics, and quantum mechanics, Yale Univ. Press (2016).

[87] N.J.A. Sloane and S. Plouffe, Encyclopedia of integer sequences, Academic Press (1995).

[88] A.M. Steane, Thermodynamics, Oxford Univ. Press (2016).

[89] S. Sternberg, Dynamical systems, Dover (2010).

[90] D.R. Stinson, Combinatorial designs: constructions and analysis, Springer (2006).

[91] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[92] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[93] J. von Neumann and O. Morgenstern, Theory of games and economic behavior, Princeton Univ. Press (1944).

[94] J. Watrous, The theory of quantum information, Cambridge Univ. Press (2018).

[95] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[96] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[97] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).

[98] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).

[99] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).

[100] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).

Index