Angles and trigonometry

Teo Banica

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

2010 *Mathematics Subject Classification.* 01A20

Key words and phrases. Angles, Trigonometry

Abstract. This is an introduction to plane geometry, angles and trigonometry, starting from zero or almost, meaning basic knowledge of numbers and fractions, and with focus on the standard applications to science and engineering questions. We provide as well an introduction to space geometry, and to advanced trigonometry too.

Preface

Measuring angles is an art, mastered by artists, as well as craftsmen, scientists and engineers, requiring you to know quite a deal of advanced mathematics, that you can hopefully learn from this book. But, before anything, why measuring angles?

Leaving arts aside, where drawing obviously requires some good knowledge of angles and perspective, unless of course you are interested in doing some low-skill work, and sell that as modern art, angles appear naturally in any question related to building, or understanding all sorts of objects, devices and phenomena, typically at big scales.

Let us take for instance, talking big scales, the question of understanding the movements of the Sun, Moon, other planets, and stars, around our Earth. With this being not that philosophical as a question as it might seem, because when sailing at sea, or even walking on unknown land, the Sun, Moon and so on can be very useful in showing you the way. Well, in relation with this, with measuring distances being barred by the big scale of our objects, you are left with observing angles, and then hopefully produce from these angles, via some tricky math computations, the direction that you need.

So, this was for the main principle of angles and trigonometry, big things can only be observed, and used, via angles. As for the applications of this principle, no need of course to go to the astronomical scales evoked above, these abound in various big scale questions from real life, and engineering. Measuring land, or even smaller things, like trees, or building various things, such as bridges, roads, big houses and so on, all this will lead you into angles and trigonometry, exactly as our ship capitain above.

As a concrete illustration, you certainly know about that amazing pyramids built by the ancient Egyptians. Well, that pyramids were built by using an advanced knowledge of trigonometry, available at that time, and which dissapeared in the present modern ages. Or at least this is how one hypothesis about the pyramids goes, and looking around, at the trigonometry knowledge of my mathematics and engineering students, I am pretty much convinced that this is indeed the true explanation for the pyramids question.

Getting now to the present book, this will be an introduction to all this, plane geometry, angles and trigonometry, starting from zero or almost, meaning basic knowledge of numbers and fractions, and with focus on the standard applications to basic science and engineering questions, along the lines evoked above. We will provide as well a brief introduction to angles and geometry in space, and to advanced trigonometry too.

More in detail now, the book is organized in 4 parts, with Part I dealing with plane geometry and angles, starting from zero, then Part II dealing with coordinates and basic trigonometry, Part III dealing with advanced trigonometry, using tools from analysis, and finally Part IV dealing with geometry and angles in three dimensions.

Many thanks to my math professors, and now that I am a professor myself, to my students. Thanks as well to my cats, for their teachings regarding the angle of attack, which is a more advanced notion, to be discussed too in this book, at the end.

Cergy, January 2025
Teo Banica

Contents

Part I

Geometry, angles

Sometimes I feel so happy
Sometimes I feel so sad
Sometimes I feel so happy
But mostly you just make me mad

CHAPTER 1

Parallel lines

1a. 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 1.1. *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 1.1, 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 1.1 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 1.2. *In fact, spacetime itself is bent.*

Okay, thanks cat, so looks like we have multiple problems with the "rule proof" of Axiom 1.1, so that definitely does not qualify as a proof. And so Axiom 1.1 will be indeed an axiom, that is, a true and useful mathematical statement, coming without proof.

Getting now to more discussion, around Axiom 1.1, 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 1.3. *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 1.1, 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 1.1 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 1.1 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 1.1 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 1.4. *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 1.1, 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 1.1, 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 1.1. Please have mercy.

CAT 1.5. *What about $PQ = \lambda P + (1 - \lambda)Q$ proving your axiom.*

Okay, thanks cat, but I was already having this in mind, for chapter 5 below. So, Axiom 1.1 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 5 below, how cats and physicists can prove Axiom 1.1, 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 1.1:

OBSERVATION 1.6. *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 1.1, 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 1.7. *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 1.6 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 1.8. *Any two crossing lines $K \not| L$ determine a point, $P = K \cap L$.*

Very good, and now with Axiom 1.1 and Axiom 1.8 in hand, we are potentially ready for doing some geometry. However, this is not exactly true, and we will need as well:

AXIOM 1.9. *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.

1b. Thales theorem

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

THEOREM 1.10 (Thales). *Proportions are kept, along parallel lines.*

PROOF. This is indeed something very standard. □

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

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

1c. Projective space

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 1.11. *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 1.11. More generally now, we have the following construction, in arbitrary dimensions:

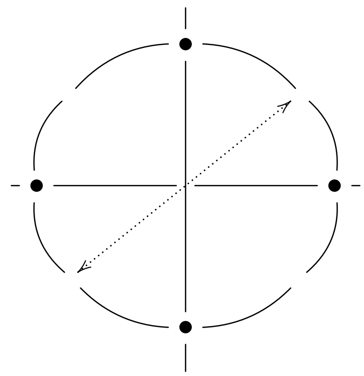
Theorem 1.12. *We can define the projective space $P^{N-1}_\mathbb{R}$ as being the space of lines in $\mathbb{R}^N$ passing through the origin, and in small dimensions:*

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

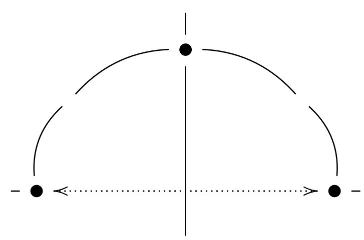
Proof. We have several assertions here, with all this being of course a bit informal, and self-explanatory, the idea and some further details being as follows:

(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 1.11 follows from definitions, exactly as in the discussion preceding the statement, regarding the case $N = 3$.

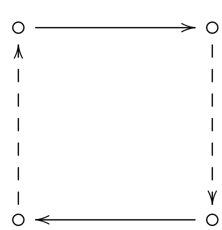
(2) At $N=2$ now, a line in $\mathbb{R}^2$ passing through the origin corresponds to 2 opposite points on the unit circle $\mathbb{T}\subset\mathbb{R}^2$, according to the following scheme:



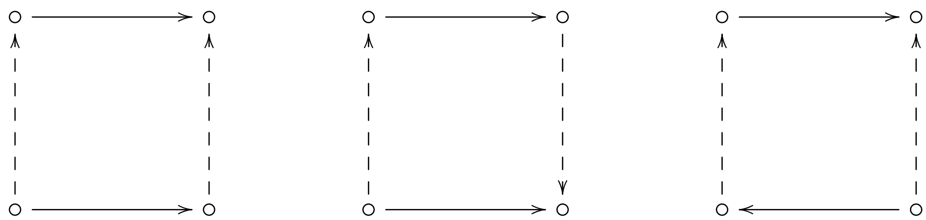
Thus, $P^1_\mathbb{R}$ corresponds to the upper semicircle of $\mathbb{T}$, with the endpoints identified, and so we obtain a circle, $P^1_\mathbb{R}=\mathbb{T}$, according to the following scheme:



(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. □

1d. Finite fields

All this is very nice, but we will pause our study here, because we still have many other things to say. Here is an interesting notion, that we can use for geometry:

Definition 1.13. *A field is a set F with a sum operation $+$ and a product operation $\times$, subject to the following conditions:*

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

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

$0, 1 \in F$, can form themselves a field, with addition as follows:

$$1 + 1 = 0$$

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

THEOREM 1.14. *The following happen:*

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

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

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

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

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

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

Getting now to geometry over finite fields, we have here:

THEOREM 1.15. *Given a field* F, *we can talk about the projective space* P^{N-1}_F, *as being the space of lines in* F^N *passing through the origin. At* $N = 3$ *we have*

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

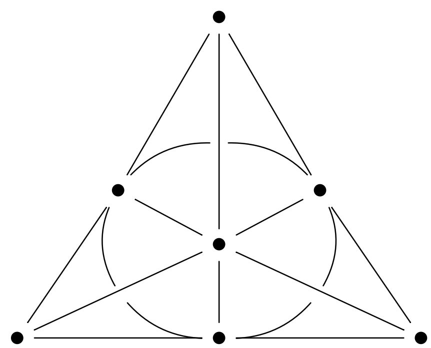
where $q = |F|$, *in the case where our field* F *is finite.*

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

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

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

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



Here the circle in the middle is by definition a line, and with this convention, the basic axioms in Definition 1.11 are satisfied, in the sense that any two points determine a line, and any two lines determine a point. And isn't this beautiful.

1e. Exercises

Exercises:

Exercise 1.16.

Exercise 1.17.

Exercise 1.18.

Exercise 1.19.

Exercise 1.20.

Exercise 1.21.

Exercise 1.22.

Exercise 1.23.

Bonus exercise.

CHAPTER 2

Triangles

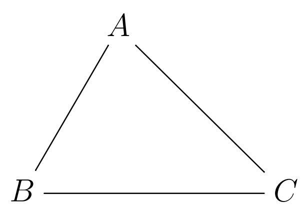
2a. Triangles

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

THEOREM 2.1. *Given a triangle ABC, the following happen:*

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

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



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

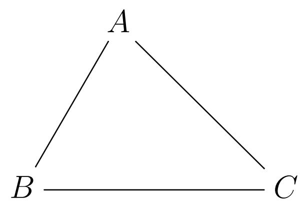
(1) Come with a small circle, inside ABC, and then inflate it, as to touch all 3 edges. The center of the circle will be then at equal distance from all 3 edges, so it will lie on all 3 angle bisectors. Thus, we have constructed the incenter, as required.

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

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

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

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



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

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

2b. More triangles

Along the same lines, but at a more advanced level now, we have:

Fact 2.2. *Besides the above 4 centers, many more remarkable points can be associated to a triangle ABC, and most of these lie on a line, called Euler line of ABC.*

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

2c. Angles, basics

Getting now to what we wanted to talk about in this book, angles and trigonometry, we can certain talk about angles, in the obvious way, by using triangles:

Fact 2.3. *We can talk about the angle between two crossing lines, and have some basic theory for the angles going, by using triangles.*

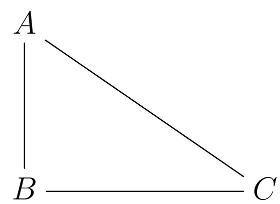
You might wonder of course what the values of these angles should be, say as real numbers. This is something quite tricky, that will take us some time to understand.

Getting started now with our study of angles, as a continuation of Fact 2.3, let us first talk about the simplest angle of them all, which is the right angle, denoted $90°$.

2d. Pythagoras theorem

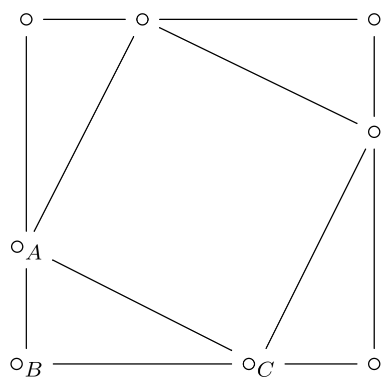
Many interesting things can be said about this right angle 90°, in particular with:

THEOREM 2.4 (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. □

The Pythagoras theorem has many applications. We will be back to this.

2e. Exercises

Exercises:

Exercise 2.5.

Exercise 2.6.

Exercise 2.7.

Exercise 2.8.

Exercise 2.9.

Exercise 2.10.

Exercise 2.11.

Exercise 2.12.

Bonus exercise.

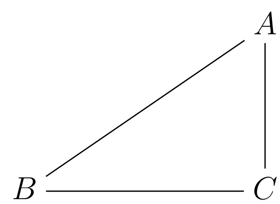
CHAPTER 3

Sine, cosine

3a. Sine, cosine

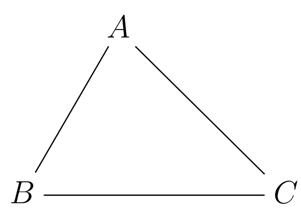
Now that we know about angles, and about Pythagoras' theorem too, it is tempting at this point to start talking about trigonometry. Let us begin with:

Definition 3.1. *We can talk about sines and cosines, by using a right triangle*



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

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



We will be back to this, on regular occassions, in what follows.

3b. Pythagoras, again

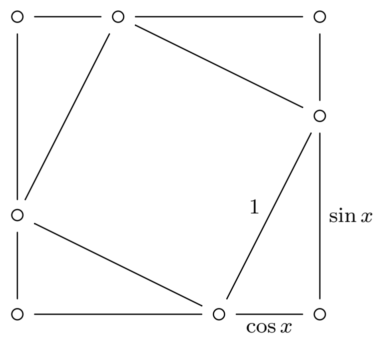
Getting now to more advanced theory, we first have:

Theorem 3.2. *The sines and cosines are subject to the formula*

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

coming from Pythagoras' theorem.

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



When computing the area of the outer square, we obtain:

$$(\sin x + \cos x)^2 = 1 + 4 \times \frac{\sin x \cos x}{2}$$

Now when expanding we obtain $\sin^2 x + \cos^2 x = 1$, as claimed. □

It is possible to say many more things about angles and $\sin x$, $\cos x$, and also talk about some supplementary quantities, such as the tangent:

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

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

3c. Sums of angles

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

THEOREM 3.3. *The sines and cosines of sums are given by*

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

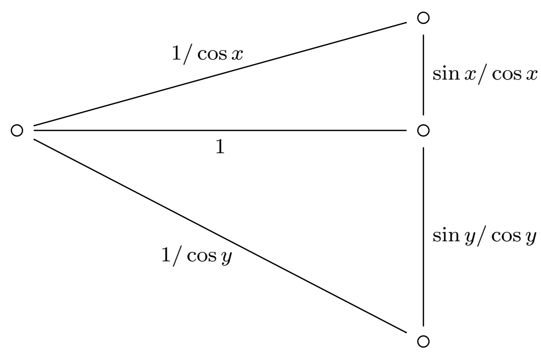
$$\cos(x + y) = \cos x \cos y - \sin x \sin y$$

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

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

(1) Let us first establish the formula for the sines. In order to do so, consider the following picture, consisting of a length 1 line segment, with angles x, y drawn on each

side, and with everything being completed, and lengths computed, as indicated:



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

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

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

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

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

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

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

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

3d. Duplication and more

Observe in particular that with $x = y$ we obtain some interesting formulae for the duplication of angles. We will be back to such questions later, with better tools.

3e. Exercises

Exercises:

Exercise 3.4.

Exercise 3.5.

Exercise 3.6.

Exercise 3.7.

Exercise 3.8.

Exercise 3.9.

Exercise 3.10.

Exercise 3.11.

Bonus exercise.

CHAPTER 4

Circle, angles

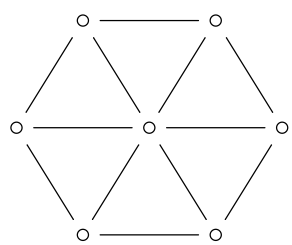
4a. Circles, pi

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

THEOREM 4.1. *The following two definitions of π are equivalent:*

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

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



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

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

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

In what regards now the precise value of π, the above picture at $N = 6$ shows that we have $\pi > 3$, but not by much. The precise figure is $\pi = 3.14159\ldots$, but we will come back to this later, once we will have appropriate tools for dealing with such questions. It is also possible to prove that π is irrational, $\pi \notin \mathbb{Q}$, but this is not trivial either.

4b. Numeric angles

Numeric angles.

4c. Basic estimates

Basic estimates.

4d. More about pi

More about pi.

4e. Exercises

Exercises:

Exercise 4.2.

Exercise 4.3.

Exercise 4.4.

Exercise 4.5.

Exercise 4.6.

Exercise 4.7.

Exercise 4.8.

Exercise 4.9.

Bonus exercise.

Part II

Basic trigonometry

In the clearing stands a boxer
And a fighter by his trade
And he carries the reminders
Of every glove that laid him down

CHAPTER 5

Affine coordinates

5a. Vector calculus

Vector calculus.

5b. Matrices, rotations

The transformations of the plane $\mathbb{R}^2$ that we are interested in are as follows:

Definition 5.1. *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 5.2. *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 5.3. *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 5.4. *The projections on* Ox *and* Oy *are:*

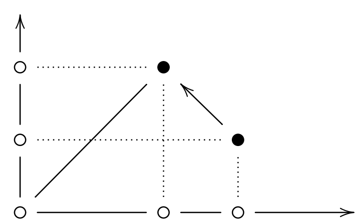
$$\begin{pmatrix}x\\ y\end{pmatrix}\to\begin{pmatrix}x\\ 0\end{pmatrix}\quad,\quad\begin{pmatrix}x\\ y\end{pmatrix}\to\begin{pmatrix}0\\ y\end{pmatrix}$$

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 5.5. *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 5.6. *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) $\Longrightarrow$ (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) $\Longrightarrow$ (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 5.7 (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 5.1, 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 5.8. *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:

$$\begin{aligned}x_n\to x,x_n\not\sim y,\forall n\quad&\implies\quad f(x_n+y)=f(x_n)+f(y),\forall n\\ &\implies\quad f(x+y)=f(x)+f(y)\end{aligned}$$

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 5.7 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 5.9. *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 5.7 for such maps.

Proof. Assuming that f is linear in the sense of Definition 5.7, 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 5.9 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 5.10. *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:

$$\begin{pmatrix}1&2\\5&6\end{pmatrix}\begin{pmatrix}3\\1\end{pmatrix}=\begin{pmatrix}1\cdot3+2\cdot1\\5\cdot3+6\cdot1\end{pmatrix}=\begin{pmatrix}5\\21\end{pmatrix}$$

As already mentioned, all this comes from our findings from Theorem 5.9. Indeed, with the above multiplication convention for matrices and vectors, we can turn Theorem 5.9 into something much simpler, and better-looking, as follows:

THEOREM 5.11. *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 5.9 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 5.12. *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 5.2, 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 5.13. *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 5.3, 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 5.14. *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 5.4, 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 5.15. *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 5.16. *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 5.17. *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 5.18. *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 will be back to this, with many applications, in what follows.

5c. Ellipses, conics

Time to discuss some applications. 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 about them were as follows:

THEOREM 5.19. *The ellipses, taken centered at the origin* 0*, and squarely oriented with respect to* Oxy*, can be defined in* 4 *possible ways, as follows:*

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

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

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

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

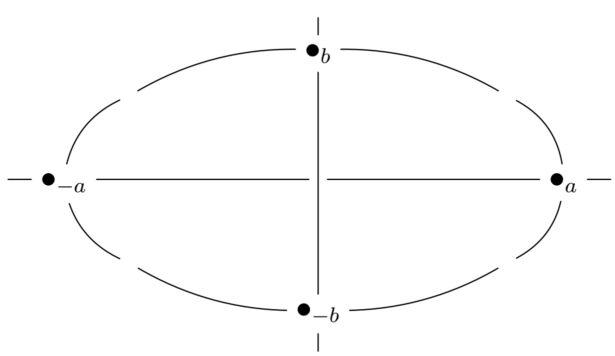
(3) *As the curves appearing when drawing a circle, from various perspectives:*

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

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

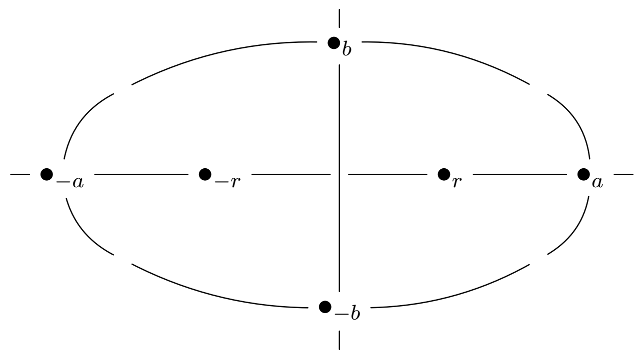
PROOF. This might look a bit confusing, and you might say, what exactly is to be proved here. Good point, and in answer, what is to be proved is that the above constructions (1-4) give rise to the same class of curves. And this can be done as follows:

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



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

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



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

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

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

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

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

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

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

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

(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 to our theorem, we have two other assertions there at the end, (3,4). But, thinking a bit, these assertions are equivalent, and (4) can be established by doing some 3D computations, that we will leave here as an instructive exercise, for you. □

All this is very nice, but before getting into physics, let us settle as well the question of wandering asteroids. 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 5.20. *The conics, which are the algebraic curves of degree* 2 *in the plane,*

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

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

PROOF. This follows by further building on Theorem 5.19, 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. □

5d. Kepler and Newton

Ready for some physics? We have the following result:

THEOREM 5.21. *Planets and other celestial bodies move around the Sun on conics,*

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

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

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

5e. Exercises

Exercises:

EXERCISE 5.22.

EXERCISE 5.23.

EXERCISE 5.24.

EXERCISE 5.25.

EXERCISE 5.26.

EXERCISE 5.27.

EXERCISE 5.28.

EXERCISE 5.29.

Bonus exercise.

CHAPTER 6

Basic trigonometry

6a. Triangles, revised

Triangles, revised.

6b. Polar coordinates

Polar coordinates.

6c. Circles and angles

Circles and angles.

6d. Basic trigonometry

Basic trigonometry.

6e. Exercises

Exercises:

Exercise 6.1.

Exercise 6.2.

Exercise 6.3.

Exercise 6.4.

Exercise 6.5.

Exercise 6.6.

Exercise 6.7.

Exercise 6.8.

Bonus exercise.

CHAPTER 7

Complex numbers

7a. Complex numbers

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

Definition 7.1. *The complex numbers are variables of the form*

$$x = a + ib$$

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

$$i^2 = -1$$

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

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

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

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

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

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

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

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

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

$$x = \pm i$$

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

THEOREM 7.2. *The complex solutions of* $ax^2 + bx + c = 0$ *with* $a, b, c \in \mathbb{R}$ *are*

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

with the square root of negative real numbers being defined as

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

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

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

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

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

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

PROPOSITION 7.3. *The complex numbers, written as usual*

$$x = a + ib$$

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

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

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

PROOF. Consider indeed two arbitrary complex numbers:

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

Their sum is then by definition the following complex number:

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

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

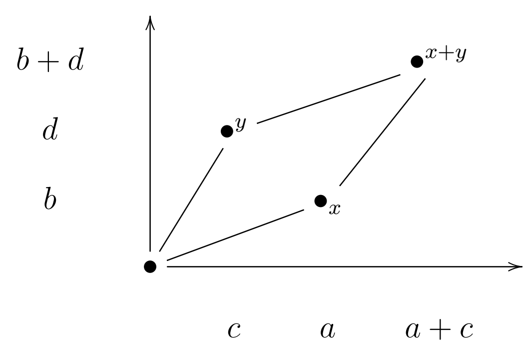
$$x=\begin{pmatrix}a\\ b\end{pmatrix}\quad,\quad y=\begin{pmatrix}c\\ d\end{pmatrix}$$

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

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

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

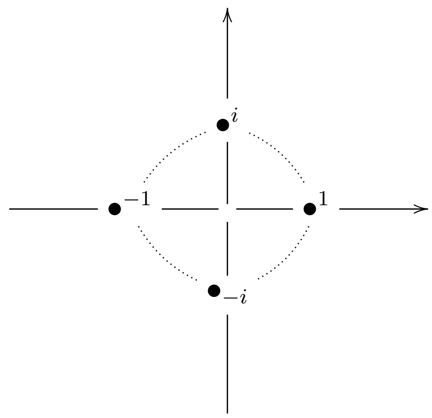
Here we have assumed that you are a bit familiar with vector calculus. If not, no problem, the idea is simply that vectors add by forming a parallelogram, as follows:



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

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

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



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

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

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

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

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

with the connecting formulae being as follows,

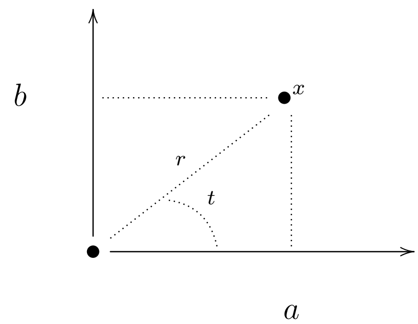
$$a = r\cos t \quad , \quad b = r\sin t$$

and in the other sense being as follows,

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

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

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



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

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

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

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

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

multiply according to the following formula:

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

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

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

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

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

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

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

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

7b. Powers, conjugates

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

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

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

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

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

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

PROOF. We have the following computation, with k terms everywhere:

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

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

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

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

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

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

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

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

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

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

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

As a basic application of Theorem 7.7, we have the following result:

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

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

has two square roots, given by the following formula:

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

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

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

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

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

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

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

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

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

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

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

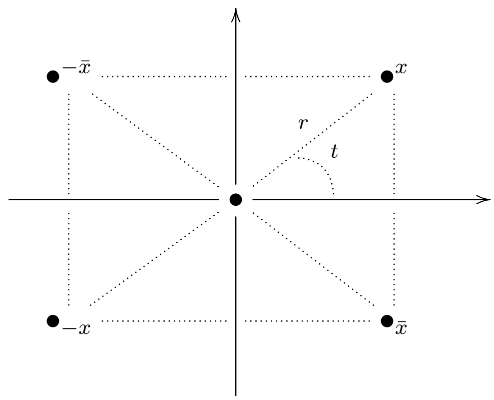
As a last general topic regarding the complex numbers, let us discuss conjugation. This is something quite tricky, complex number specific, as follows:

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

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(6) This comes from the formula of the solutions, that we know from Theorem 7.2, but we can deduce this as well directly, without computations. Indeed, by using our assumption that the coefficients are real, $a, b, c \in \mathbb{R}$, we have:

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

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

7c. Polynomials, roots

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

THEOREM 7.12. *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:

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

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:

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

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

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

which are called roots of unity of order N.

PROOF. This follows from the general multiplication formula for the complex numbers in polar form. Indeed, with the notation $x = re^{it}$, our equation reads:

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

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

As an illustration here, the roots of unity of small order, along with some of their basic properties, which are very useful for computations, are as follows:

$\underline{N=1}$. Here the unique root of unity is 1.

$\underline{N=2}$. Here we have two roots of unity, namely 1 and -1.

$\underline{N=3}$. Here we have 1, then $w=e^{2\pi i/3}$, and then $w^2=\bar{w}=e^{4\pi i/3}$.

$\underline{N=4}$. Here the roots of unity, read as usual counterclockwise, are $1,i,-1,-i$.

$\underline{N=5}$. Here, with $w=e^{2\pi i/5}$, the roots of unity are $1,w,w^2,w^3,w^4$.

$\underline{N=6}$. Here a useful alternative writing is $\{\pm 1,\pm w,\pm w^2\}$, with $w=e^{2\pi i/3}$.

$\underline{N=7}$. Here, with $w=e^{2\pi i/7}$, the roots of unity are $1,w,w^2,w^3,w^4,w^5,w^6$.

$\underline{N=8}$. Here the roots of unity, read as usual counterclockwise, are the numbers $1,w,i,iw,-1,-w,-i,-iw$, with $w=e^{\pi i/4}$, which is also given by $w=(1+i)/\sqrt{2}$.

The roots of unity are very useful variables, and have many interesting properties. As a first application, we can now solve the ambiguity questions related to the extraction of N-th roots, that we met in the above, the statement here being as follows:

THEOREM 7.14. *Any nonzero complex number, written as*

$$x=re^{it}$$

has exactly N roots of order N, which appear as

$$y=r^{1/N}e^{it/N}$$

multiplied by the N roots of unity of order N.

PROOF. We must solve the equation $z^N=x$, over the complex numbers. Since the number y in the statement clearly satisfies $y^N=x$, our equation is equivalent to:

$$z^N=y^N$$

Now observe that we can write this equation as follows:

$$\left(\frac{z}{y}\right)^N=1$$

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

The roots of unity appear in connection with many other interesting questions, and there are many useful formulae relating them, which are good to know. Here is a basic such formula, very beautiful, to be used many times in what follows:

Theorem 7.15. *The roots of unity, $\{w^k\}$ with $w = e^{2\pi i/N}$, have the property*

$$\sum_{k=0}^{N-1}(w^k)^s = N\delta_{N|s}$$

for any exponent $s \in \mathbb{N}$, where on the right we have a Kronecker symbol.

Proof. The numbers in the statement, when written more conveniently as $(w^s)^k$ with $k = 0, \ldots, N-1$, form a certain regular polygon in the plane P_s. Thus, if we denote by C_s the barycenter of this polygon, we have the following formula:

$$\frac{1}{N}\sum_{k=0}^{N-1} w^{ks} = C_s$$

Now observe that in the case $N \not| s$ our polygon P_s is non-degenerate, circling around the unit circle, and having center $C_s = 0$. As for the case $N|s$, here the polygon is degenerate, lying at 1, and having center $C_s = 1$. Thus, we have the following formula:

$$C_s = \delta_{N|s}$$

Thus, we obtain the formula in the statement. □

As an interesting philosophical fact, regarding the roots of unity, and the complex numbers in general, we can now solve the following equation, in a "uniform" way:

$$x_1 + \ldots + x_N = 0$$

With this being not a joke. Frankly, can you find some nice-looking family of real numbers $x_1, \ldots, x_N$ satisfying $x_1 + \ldots + x_N = 0$? Certainly not. But with complex numbers we have now our answer, the sum of the N-th roots of unity being zero.

7e. Exercises

Exercises:

Exercise 7.16.

Exercise 7.17.

Exercise 7.18.

Exercise 7.19.

Exercise 7.20.

Exercise 7.21.

Exercise 7.22.

Exercise 7.23.

Bonus exercise.

CHAPTER 8

Advanced trigonometry

8a. Complex geometry

Complex geometry.

8b. Rotation tricks

Rotation tricks.

8c. Advanced trigonometry

Advanced trigonometry.

8d. Plane curves

Recall from before that conics are at the core of everything, mathematics, physics, life. But, what is next? A natural answer to this question comes from:

Definition 8.1. *An algebraic curve in $\mathbb{R}^2$ is the vanishing set*

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

of a polynomial $P\in\mathbb{R}[X,Y]$ of arbitrary degree.

We already know well the algebraic curves in degree 2, which are the conics, and a first problem is, what results from what we learned about conics have a chance to be relevant to the arbitrary algebraic curves. And normally none, because the ellipses, parabolas and hyperbolas are obviously very particular curves, having very particular properties.

Let us record however a useful statement here, as follows:

Proposition 8.2. *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 8.3. *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 8.4. *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 8.5. *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 8.6. *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 8.7. *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 8.8. *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 8.9. *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 8.10. *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. □

We will be back to plane algebraic curves in chapter 15 below, with some generalizations, and more theory, when talking physics and field lines.

8e. Exercises

Exercises:

EXERCISE 8.11.

EXERCISE 8.12.

EXERCISE 8.13.

EXERCISE 8.14.

EXERCISE 8.15.

EXERCISE 8.16.

EXERCISE 8.17.

EXERCISE 8.18.

Bonus exercise.

Part III

Heavy calculus

There is a house in New Orleans
They call the Rising Sun
And it's been the ruin of many a poor boy
Dear God, I know I was one

CHAPTER 9

Functions, derivatives

9a. Functions, derivatives

The idea of calculus is very simple. We are interested in functions $f:\mathbb{R}\to\mathbb{R}$, and we already know that when f is continuous at a point x, we can write an approximation formula as follows, for the values of our function f around that point x:

$$f(x+t)\simeq f(x)$$

The problem is now, how to improve this? And a bit of thinking at all this suggests to look at the slope of f at the point x. Which leads us into the following notion:

Definition 9.1. *A function $f:\mathbb{R}\to\mathbb{R}$ is called differentiable at x when*

$$f'(x)=\lim_{t\to0}\frac{f(x+t)-f(x)}{t}$$

called derivative of f at that point x, exists.

As a first remark, in order for f to be differentiable at x, that is to say, in order for the above limit to converge, the numerator must go to 0, as the denominator t does:

$$\lim_{t\to0}[f(x+t)-f(x)]=0$$

Thus, f must be continuous at x. However, the converse is not true, a basic counterexample being $f(x)=|x|$ at $x=0$. Let us summarize these findings as follows:

Proposition 9.2. *If f is differentiable at x, then f must be continuous at x. However, the converse is not true, a basic counterexample being $f(x)=|x|$, at $x=0$.*

Proof. The first assertion is something that we already know, from the above. As for the second assertion, regarding $f(x)=|x|$, this is something quite clear on the picture of f, but let us prove this mathematically, based on Definition 9.1. We have:

$$\lim_{t\searrow0}\frac{|0+t|-|0|}{t}=\lim_{t\searrow0}\frac{t-0}{t}=1$$

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

$$\lim_{t\nearrow0}\frac{|0+t|-|0|}{t}=\lim_{t\nearrow0}\frac{-t-0}{t}=-1$$

Thus, the limit in Definition 9.1 does not converge, so we have our counterexample. □

Generally speaking, the last assertion in Proposition 9.2 should not bother us much, because most of the basic continuous functions are differentiable, and we will see examples in a moment. Before that, however, let us recall why we are here, namely improving the basic estimate $f(x+t)\simeq f(x)$. We can now do this, using the derivative, as follows:

THEOREM 9.3. *Assuming that f is differentiable at x, we have:*

$$f(x+t)\simeq f(x)+f'(x)t$$

In other words, f is, approximately, locally affine at x.

PROOF. Assume indeed that f is differentiable at x, and let us set, as before:

$$f'(x)=\lim_{t\to0}\frac{f(x+t)-f(x)}{t}$$

By multiplying by t, we obtain that we have, once again in the $t\to0$ limit:

$$f(x+t)-f(x)\simeq f'(x)t$$

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

All this is very nice, and before developing more theory, let us work out some examples. As a first illustration, the derivatives of the power functions are as follows:

THEOREM 9.4. *We have the differentiation formula*

$$(x^p)'=px^{p-1}$$

valid for any exponent $p\in\mathbb R$.

PROOF. We can do this in three steps, as follows:

(1) In the case $p\in\mathbb N$ we can use the binomial formula, which gives, as desired:

$$\begin{aligned}(x+t)^p&=\sum_{k=0}^n\binom{p}{k}x^{p-k}t^k\\&=x^p+px^{p-1}t+\ldots+t^p\\&\simeq x^p+px^{p-1}t\end{aligned}$$

(2) Let us discuss now the general case $p\in\mathbb Q$. We write $p=m/n$, with $m\in\mathbb Z$ and $n\in\mathbb N$. In order to do the computation, we use the following formula:

$$a^n-b^n=(a-b)(a^{n-1}+a^{n-2}b+\ldots+b^{n-1})$$

We set in this formula $a = (x+t)^{m/n}$ and $b = x^{m/n}$. We obtain, as desired:

$$\begin{aligned}(x+t)^{m/n} - x^{m/n} &= \frac{(x+t)^m - x^m}{(x+t)^{m(n-1)/n} + \ldots + x^{m(n-1)/n}} \\ &\simeq \frac{(x+t)^m - x^m}{nx^{m(n-1)/n}} \\ &\simeq \frac{mx^{m-1}t}{nx^{m(n-1)/n}} \\ &= \frac{m}{n} \cdot x^{m-1-m+m/n} \cdot t \\ &= \frac{m}{n} \cdot x^{m/n-1} \cdot t\end{aligned}$$

(3) In the general case now, where $p \in \mathbb{R}$ is real, we can use a similar argument. Indeed, given any integer $n \in \mathbb{N}$, we have the following computation:

$$\begin{aligned}(x+t)^p - x^p &= \frac{(x+t)^{pn} - x^{pn}}{(x+t)^{p(n-1)} + \ldots + x^{p(n-1)}} \\ &\simeq \frac{(x+t)^{pn} - x^{pn}}{nx^{p(n-1)}}\end{aligned}$$

Now observe that we have the following estimate, with [.] being the integer part:

$$(x+t)^{[pn]} \leq (x+t)^{pn} \leq (x+t)^{[pn]+1}$$

By using the binomial formula on both sides, for the integer exponents $[pn]$ and $[pn]+1$ there, we deduce that with $n >> 0$ we have the following estimate:

$$(x+t)^{pn} \simeq x^{pn} + pnx^{pn-1}t$$

Thus, we can finish our computation started above as follows:

$$(x+t)^p - x^p \simeq \frac{pnx^{pn-1}t}{nx^{pn-p}} = px^{p-1}t$$

But this gives $(x^p)' = px^{p-1}$, which finishes the proof. □

Here are some further computations, for other basic functions that we know:

THEOREM 9.5. *We have the following results:*

(1) $(\sin x)' = \cos x$.
(2) $(\cos x)' = -\sin x$.
(3) $(e^x)' = e^x$.
(4) $(\log x)' = x^{-1}$.

PROOF. This is quite tricky, as always when computing derivatives, as follows:

(1) Regarding sin, the computation here goes as follows:

$$\begin{aligned}
(\sin x)' &= \lim_{t\to 0}\frac{\sin(x+t)-\sin x}{t}\\
&= \lim_{t\to 0}\frac{\sin x\cos t+\cos x\sin t-\sin x}{t}\\
&= \lim_{t\to 0}\sin x\cdot\frac{\cos t-1}{t}+\cos x\cdot\frac{\sin t}{t}\\
&= \cos x
\end{aligned}$$

Here we have used the fact, which is clear on pictures, by drawing the trigonometric circle, that we have $\sin t\simeq t$ for $t\simeq 0$, plus the fact, which follows from this and from Pythagoras, $\sin^2+\cos^2=1$, that we have as well $\cos t\simeq 1-t^2/2$, for $t\simeq 0$.

(2) The computation for cos is similar, as follows:

$$\begin{aligned}
(\cos x)' &= \lim_{t\to 0}\frac{\cos(x+t)-\cos x}{t}\\
&= \lim_{t\to 0}\frac{\cos x\cos t-\sin x\sin t-\cos x}{t}\\
&= \lim_{t\to 0}\cos x\cdot\frac{\cos t-1}{t}-\sin x\cdot\frac{\sin t}{t}\\
&= -\sin x
\end{aligned}$$

(3) For the exponential, the derivative can be computed as follows:

$$\begin{aligned}
(e^x)' &= \left(\sum_{k=0}^\infty\frac{x^k}{k!}\right)'\\
&= \sum_{k=0}^\infty\frac{kx^{k-1}}{k!}\\
&= e^x
\end{aligned}$$

(4) As for the logarithm, the computation here is as follows, using $\log(1+y)\simeq y$ for $y\simeq 0$, which follows from $e^y\simeq 1+y$ that we found in (3), by taking the logarithm:

$$\begin{aligned}
(\log x)' &= \lim_{t\to 0}\frac{\log(x+t)-\log x}{t}\\
&= \lim_{t\to 0}\frac{\log(1+t/x)}{t}\\
&= \frac{1}{x}
\end{aligned}$$

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

Speaking exponentials, we can now formulate a nice result about them:

THEOREM 9.6. *The exponential function, namely*

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

is the unique power series satisfying $f' = f$ and $f(0) = 1$.

PROOF. Consider indeed a power series satisfying $f' = f$ and $f(0) = 1$. Due to $f(0) = 1$, the first term must be 1, and so our function must look as follows:

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

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

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

Thus, the equation $f' = f$ is equivalent to the following equalities:

$$c_1 = 1 \quad , \quad 2c_2 = c_1 \quad , \quad 3c_3 = c_2 \quad , \quad 4c_4 = c_3 \quad , \quad \ldots$$

But this system of equations can be solved by recurrence, as follows:

$$c_1 = 1 \quad , \quad c_2 = \frac{1}{2} \quad , \quad c_3 = \frac{1}{2 \times 3} \quad , \quad c_4 = \frac{1}{2 \times 3 \times 4} \quad , \quad \ldots$$

Thus we have $c_k = 1/k!$, leading to the conclusion in the statement. □

Observe that the above result leads to a more conceptual explanation for the number e itself. To be more precise, $e \in \mathbb{R}$ is the unique number satisfying:

$$(e^x)' = e^x$$

Let us work out now some general results. We have here the following statement:

THEOREM 9.7. *We have the following formulae:*

(1) $(f+g)' = f' + g'$.
(2) $(fg)' = f'g + fg'$.
(3) $(f \circ g)' = (f' \circ g) \cdot g'$.

PROOF. All these formulae are elementary, the idea being as follows:

(1) This follows indeed from definitions, the computation being as follows:

$$\begin{aligned}(f+g)'(x) &= \lim_{t\to 0}\frac{(f+g)(x+t)-(f+g)(x)}{t}\\ &= \lim_{t\to 0}\left(\frac{f(x+t)-f(x)}{t}+\frac{g(x+t)-g(x)}{t}\right)\\ &= \lim_{t\to 0}\frac{f(x+t)-f(x)}{t}+\lim_{t\to 0}\frac{g(x+t)-g(x)}{t}\\ &= f'(x)+g'(x)\end{aligned}$$

(2) This follows from definitions too, the computation, by using the more convenient formula $f(x+t)\simeq f(x)+f'(x)t$ as a definition for the derivative, being as follows:

$$\begin{aligned}(fg)(x+t) &= f(x+t)g(x+t)\\ &\simeq (f(x)+f'(x)t)(g(x)+g'(x)t)\\ &\simeq f(x)g(x)+(f'(x)g(x)+f(x)g'(x))t\end{aligned}$$

Indeed, we obtain from this that the derivative is the coefficient of t, namely:

$$(fg)'(x)=f'(x)g(x)+f(x)g'(x)$$

(3) Regarding compositions, the computation here is as follows, again by using the more convenient formula $f(x+t)\simeq f(x)+f'(x)t$ as a definition for the derivative:

$$\begin{aligned}(f\circ g)(x+t) &= f(g(x+t))\\ &\simeq f(g(x)+g'(x)t)\\ &\simeq f(g(x))+f'(g(x))g'(x)t\end{aligned}$$

Indeed, we obtain from this that the derivative is the coefficient of t, namely:

$$(f\circ g)'(x)=f'(g(x))g'(x)$$

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

We can of course combine the above formulae, and we obtain for instance:

THEOREM 9.8. *The derivatives of fractions are given by:*

$$\left(\frac{f}{g}\right)'=\frac{f'g-fg'}{g^2}$$

In particular, we have the following formula, for the derivative of inverses:

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

In fact, we have $(f^p)'=pf^{p-1}$, *for any exponent* $p\in\mathbb{R}$.

PROOF. This statement is written a bit upside down, and for the proof it is better to proceed backwards. To be more precise, by using $(x^p)' = px^{p-1}$ and Theorem 9.7 (3), we obtain the third formula. Then, with $p = -1$, we obtain from this the second formula. And finally, by using this second formula and Theorem 9.7 (2), we obtain:

$$\begin{aligned}\left(\frac{f}{g}\right)' &= \left(f \cdot \frac{1}{g}\right)' \\ &= f' \cdot \frac{1}{g} + f\left(\frac{1}{g}\right)' \\ &= \frac{f'}{g} - \frac{fg'}{g^2} \\ &= \frac{f'g - fg'}{g^2}\end{aligned}$$

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

With the above formulae in hand, we can do all sorts of computations for other basic functions that we know, including $\tan x$, or $\arctan x$:

THEOREM 9.9. *We have the following formulae,*

$$(\tan x)' = \frac{1}{\cos^2 x} \quad , \quad (\arctan x)' = \frac{1}{1 + x^2}$$

and the derivatives of the remaining trigonometric functions can be computed as well.

PROOF. For tan, we have the following computation:

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

As for arctan, we can use here the following computation:

$$\begin{aligned}(\tan \circ \arctan)'(x) &= \tan'(\arctan x) \arctan'(x) \\ &= \frac{1}{\cos^2(\arctan x)} \arctan'(x)\end{aligned}$$

Indeed, since the term on the left is simply $x' = 1$, we obtain from this:

$$\arctan'(x) = \cos^2(\arctan x)$$

On the other hand, with $t = \arctan x$ we know that we have $\tan t = x$, and so:

$$\cos^2(\arctan x) = \cos^2 t = \frac{1}{1+\tan^2 t} = \frac{1}{1+x^2}$$

Thus, we are led to the formula in the statement, namely:

$$(\arctan x)' = \frac{1}{1+x^2}$$

As for the last assertion, we will leave this as an exercise. □

At the theoretical level now, further building on Theorem 9.3, we have:

THEOREM 9.10. *The local minima and maxima of a differentiable function $f : \mathbb{R} \to \mathbb{R}$ appear at the points $x \in \mathbb{R}$ where:*

$$f'(x) = 0$$

However, the converse of this fact is not true in general.

PROOF. The first assertion follows from the formula in Theorem 9.3, 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 9.10, we have:

THEOREM 9.11. *Assuming that $f:[a,b]\to\mathbb{R}$ is differentiable, we have*

$$\frac{f(b)-f(a)}{b-a}=f'(c)$$

for some $c\in(a,b)$, called mean value property of f.

PROOF. In the case $f(a)=f(b)$, the result, called Rolle theorem, states that we have $f'(c)=0$ for some $c\in(a,b)$, and follows from Theorem 9.10. Now in what regards our statement, due to Lagrange, this follows from Rolle, applied to the following function:

$$g(x)=f(x)-\frac{f(b)-f(a)}{b-a}\cdot x$$

Indeed, we have $g(a)=g(b)$, due to our choice of the constant on the right, so we get $g'(c)=0$ for some $c\in(a,b)$, which translates into the formula in the statement. □

In practice, Theorem 9.10 can be used in order to find the maximum and minimum of any differentiable function, and this method is best recalled as follows:

ALGORITHM 9.12. *In order to find the minimum and maximum of $f:[a,b]\to\mathbb{R}$:*
(1) *Compute the derivative f'.*
(2) *Solve the equation $f'(x)=0$.*
(3) *Add a,b to your set of solutions.*
(4) *Compute $f(x)$, for all your solutions.*
(5) *Compute the min/max of all these $f(x)$ values.*
(6) *Then this is the min/max of your function.*

Needless to say, all this is very interesting, and powerful. The general problem in any type of applied mathematics is that of finding the minimum or maximum of some function, and we have now an algorithm for dealing with such questions. Very nice.

9b. 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 9.13. *We say that $f:\mathbb{R}\to\mathbb{R}$ is twice differentiable if it is differentiable, and its derivative $f':\mathbb{R}\to\mathbb{R}$ is differentiable too. The derivative of f' is denoted*

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

and is called second derivative of f.

You might probably wonder why coming with this definition, which looks a bit abstract and complicated, instead of further developing the theory of the first derivative, which looks like something very reasonable and useful. Good point, and answer to this coming in a moment. But before that, let us get a bit familiar with f''. We first have:

Interpretation 9.14. *The second derivative $f''(x)\in\mathbb R$ is the number which:*

(1) *Expresses the growth rate of the slope $f'(z)$ at the point x.*
(2) *Gives us the acceleration of the function f at the point x.*
(3) *Computes how much different is $f(x)$, compared to $f(z)$ with $z\simeq x$.*
(4) *Tells us how much convex or concave is f, around the point x.*

So, this is the truth about the second derivative, making it clear that what we have here is a very interesting notion. In practice now, (1) follows from the usual interpretation of the derivative, as both a growth rate, and a slope. Regarding (2), this is some sort of reformulation of (1), using the intuitive meaning of the word "acceleration", with the relevant physics equations, due to Newton, being as follows:

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

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

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

Proposition 9.15. *The second derivatives of the basic functions are as follows:*

(1) $(x^p)''=p(p-1)x^{p-2}$.
(2) $\sin''=-\sin$.
(3) $\cos''=-\cos$.
(4) $\exp'=\exp$.
(5) $\log'(x)=-1/x^2$.

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

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

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

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

Of course, this is not the end of the story, because these formulae remain quite opaque, and must be examined in view of Interpretation 9.14, in order to see what exactly is going on. Also, we have tan and the inverse trigonometric functions too. In short, plenty of good exercises here, for you, and the more you solve, the better your calculus will be.

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

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

This function is then differentiable, with its derivative being:

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

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

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

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

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

THEOREM 9.16. *Any twice differentiable function $f : \mathbb{R} \to \mathbb{R}$ is locally quadratic,*

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

with $f''(x)$ being as usual the derivative of the function $f' : \mathbb{R} \to \mathbb{R}$ at the point x.

PROOF. Assume indeed that f is twice differentiable at x, and let us try to construct an approximation of f around x by a quadratic function, as follows:

$$f(x+t) \simeq a + bt + ct^2$$

We must have $a = f(x)$, and we also know from Theorem 9.3 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 9.3. 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 9.3, and there are many applications of it. As a first such application, justifying Interpretation 9.14 (3), we have the following statement, which is a bit heuristic, but we will call it however Proposition:

PROPOSITION 9.17. *Intuitively speaking, the second derivative $f''(x) \in \mathbb{R}$ computes how much different is $f(x)$, compared to the average of $f(z)$, with $z \simeq x$.*

PROOF. As already mentioned, this is something a bit heuristic, but which is good to know. Let us write the formula in Theorem 9.17, as such, and with $t \to -t$ too:

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

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

By making the average, we obtain the following formula:

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

But this is what our statement says, save for some uncertainties regarding the averaging method, and for the precise value of $I(t^2/2)$. We will leave this for later. □

Back to rigorous mathematics, we can improve as well Theorem 9.10, as follows:

Theorem 9.18. *The local minima and local maxima of a twice differentiable function $f:\mathbb{R}\to\mathbb{R}$ appear at the points $x\in\mathbb{R}$ where*

$$f'(x)=0$$

with the local minima corresponding to the case $f'(x)\geq 0$, and with the local maxima corresponding to the case $f''(x)\leq 0$.

Proof. The first assertion is something that we already know. As for the second assertion, we can use the formula in Theorem 9.16, which in the case $f'(x)=0$ reads:

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

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

As before with Theorem 9.10, the above result is not the end of the story with the mathematics of the local minima and maxima, because things are undetermined when:

$$f'(x)=f''(x)=0$$

For instance the functions $\pm x^n$ with $n\in\mathbb{N}$ all satisfy this condition at $x=0$, which is a minimum for the functions of type x^{2m}, a maximum for the functions of type $-x^{2m}$, and not a local minimum or local maximum for the functions of type $\pm x^{2m+1}$.

There are some comments to be made in relation with Algorithm 9.12 as well. Normally that algorithm stays strong, because Theorem 9.18 can only help in relation with the final steps, and is it worth it to compute the second derivative f'', just for getting rid of roughly $1/2$ of the $f(x)$ values to be compared. However, in certain cases, this method proves to be useful, so Theorem 9.18 is good to know, when applying that algorithm.

9c. Convex functions

As a main concrete application now of the second derivative, which is something very useful in practice, and related to Interpretation 9.14 (4), we have the following result:

Theorem 9.19. *Given a convex function $f:\mathbb{R}\to\mathbb{R}$, we have the following Jensen inequality, for any $x_1,\ldots,x_N\in\mathbb{R}$, and any $\lambda_1,\ldots,\lambda_N>0$ summing up to 1,*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$f''(x)\geq0$$

(4) Thus, we are led to the conclusions in the statement, regarding the convex functions. As for the concave functions, the proof here is similar. Alternatively, we can say that f is concave precisely when $-f$ is convex, and get the results from what we have. □

As a basic application of the Jensen inequality, which is very classical, we have:

THEOREM 9.20. *For any $p\in(1,\infty)$ we have the following inequality,*

$$\left|\frac{x_1+\ldots+x_N}{N}\right|^p\leq\frac{|x_1|^p+\ldots+|x_N|^p}{N}$$

and for any $p\in(0,1)$ we have the following inequality,

$$\left|\frac{x_1+\ldots+x_N}{N}\right|^p\geq\frac{|x_1|^p+\ldots+|x_N|^p}{N}$$

with in both cases equality precisely when $|x_1|=\ldots=|x_N|$.

PROOF. This follows indeed from Theorem 9.19, because we have:

$$(x^p)''=p(p-1)x^{p-2}$$

Thus x^p is convex for $p>1$ and concave for $p<1$, which gives the results. □

Observe that at $p=2$ we obtain as particular case of the above inequality the Cauchy-Schwarz inequality, or rather something equivalent to it, namely:

$$\left(\frac{x_1+\ldots+x_N}{N}\right)^2\leq\frac{x_1^2+\ldots+x_N^2}{N}$$

We will be back to this later on in this book, when talking scalars products and Hilbert spaces, with some more conceptual proofs for such inequalities.

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

THEOREM 9.21. *We have the Young inequality,*

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

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

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

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

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

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

Now by exponentiating, we obtain the Young inequality. □

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

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

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

9d. Taylor formula

Back now to the general theory of the derivatives, and their theoretical applications, we can further develop our basic approximation method, at order 3, at order 4, and so on, the ultimate result on the subject, called Taylor formula, being as follows:

Theorem 9.22. *Any function $f:\mathbb{R}\to\mathbb{R}$ can be locally approximated as*

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

where $f^{(k)}(x)$ are the higher derivatives of f at the point x.

Proof. Consider the function to be approximated, namely:

$$\varphi(t)=f(x+t)$$

Let us try to best approximate this function at a given order $n\in\mathbb{N}$. We are therefore looking for a certain polynomial in t, of the following type:

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

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

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

In order to prove now that this approximation holds indeed, we can use L'Hôpital's rule, applied several times, as in the proof of Theorem 14.16. To be more precise, if we denote by $\varphi(t)\simeq P(t)$ the approximation to be proved, we have:

$$\begin{aligned}\frac{\varphi(t)-P(t)}{t^n}&\simeq\frac{\varphi'(t)-P'(t)}{nt^{n-1}}\\&\simeq\frac{\varphi''(t)-P''(t)}{n(n-1)t^{n-2}}\\&\ \vdots\\&\simeq\frac{\varphi^{(n)}(t)-P^{(n)}(t)}{n!}\\&=\frac{f^{(n)}(x)-f^{(n)}(x)}{n!}\\&=0\end{aligned}$$

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

Here is a related interesting statement, inspired from the above proof:

Proposition 9.23. *For a polynomial of degree n, the Taylor approximation*

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

is an equality. The converse of this statement holds too.

PROOF. By linearity, it is enough to check the equality in question for the monomials $f(x) = x^p$, with $p \leq n$. But here, the formula to be proved is as follows:

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

We recognize the binomial formula, so our result holds indeed. As for the converse, this is clear, because the Taylor approximation is a polynomial of degree n. □

There are many other things that can be said about the Taylor formula, at the theoretical level, notably with a study of the remainder, when truncating this formula at a given order $n \in \mathbb{N}$. We will be back to this, later in this book.

As an application of the Taylor formula, we can now improve the binomial formula, which was actually our main tool so far, in the following way:

THEOREM 9.24. *We have the following generalized binomial formula, with $p \in \mathbb{R}$,*

$$(x+t)^p = \sum_{k=0}^{\infty} \binom{p}{k} x^{p-k} t^k$$

with the generalized binomial coefficients being given by the formula

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

valid for any $|t| < |x|$. With $p \in \mathbb{N}$, we recover the usual binomial formula.

PROOF. It is customary to divide everything by x, which is the same as assuming $x = 1$. The formula to be proved is then as follows, under the assumption $|t| < 1$:

$$(1+t)^p = \sum_{k=0}^{\infty} \binom{p}{k} t^k$$

Let us discuss now the validity of this formula, depending on $p \in \mathbb{R}$:

(1) Case $p \in \mathbb{N}$. According to our definition of the generalized binomial coefficients, we have $\binom{p}{k} = 0$ for $k > p$, so the series is stationary, and the formula to be proved is:

$$(1+t)^p = \sum_{k=0}^{p} \binom{p}{k} t^k$$

But this is the usual binomial formula, which holds for any $t \in \mathbb{R}$.

(2) Case $p = -1$. Here we can use the following formula, valid for $|t| < 1$:

$$\frac{1}{1+t} = 1 - t + t^2 - t^3 + \ldots$$

But this is exactly our generalized binomial formula at $p=-1$, because:

$$\binom{-1}{k}=\frac{(-1)(-2)\ldots(-k)}{k!}=(-1)^k$$

(3) Case $p\in-\mathbb{N}$. This is a continuation of our study at $p=-1$, which will finish the study at $p\in\mathbb{Z}$. With $p=-m$, the generalized binomial coefficients are:

$$\begin{aligned}\binom{-m}{k} &= \frac{(-m)(-m-1)\ldots(-m-k+1)}{k!}\\ &= (-1)^k\frac{m(m+1)\ldots(m+k-1)}{k!}\\ &= (-1)^k\frac{(m+k-1)!}{(m-1)!k!}\\ &= (-1)^k\binom{m+k-1}{m-1}\end{aligned}$$

Thus, our generalized binomial formula at $p=-m$ reads:

$$\frac{1}{(1+t)^m}=\sum_{k=0}^\infty(-1)^k\binom{m+k-1}{m-1}t^k$$

But this is something which holds indeed, and not difficult to prove.

(4) General case, $p\in\mathbb{R}$. As we can see, things escalate quickly, so we will skip the next step, $p\in\mathbb{Q}$, and discuss directly the case $p\in\mathbb{R}$. Consider the following function:

$$f(x)=x^p$$

The derivatives at $x=1$ are then given by the following formula:

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

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

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

But this is exactly our generalized binomial formula, so we are done with the case where t is small. With a bit more care, we obtain that this holds for any $|t|<1$, and we will leave this as an instructive exercise, and come back to it, later in this book. □

We can see from the above the power of the Taylor formula, saving us from quite complicated combinatorics. Remember indeed the mess when trying to directly establish particular cases of the generalized binomial formula. Gone all that.

As a main application now of our generalized binomial formula, which is something very useful in practice, we can extract square roots, as follows:

Proposition 9.25. *We have the following formula,*

$$\sqrt{1+t}=1-2\sum_{k=1}^\infty C_{k-1}\left(\frac{-t}{4}\right)^k$$

with $C_k=\frac{1}{k+1}\binom{2k}{k}$ *being the Catalan numbers. Also, we have*

$$\frac{1}{\sqrt{1+t}}=\sum_{k=0}^\infty D_k\left(\frac{-t}{4}\right)^k$$

with $D_k=\binom{2k}{k}$ *being the central binomial coefficients.*

Proof. Indeed, at $p=1/2$, the generalized binomial coefficients are:

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

Also, at $p=-1/2$, the generalized binomial coefficients are:

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

Thus, Theorem 9.24 at $p=\pm1/2$ gives the formulae in the statement. □

As another basic application of the Taylor series, we have:

Theorem 9.26. *We have the following formulae,*

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

as well as the following formulae,

$$e^x=\sum_{k=0}^\infty\frac{x^k}{k!}\quad,\quad\log(1+x)=\sum_{k=0}^\infty(-1)^{k+1}\frac{x^k}{k}$$

as Taylor series, and in general as well, with $|x|<1$ *needed for* log.

PROOF. There are several statements here, the proofs being as follows:

(1) Regarding sin and cos, we can use here the following formulae:

$$(\sin x)' = \cos x \quad , \quad (\cos x)' = -\sin x$$

Thus, we can differentiate sin and cos as many times as we want to, so we can compute the corresponding Taylor series, and we obtain the formulae in the statement.

(2) Regarding exp and log, here the needed formulae, which lead to the formulae in the statement for the corresponding Taylor series, are as follows:

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

(3) Finally, the fact that the formulae in the statement extend beyond the small t setting, coming from Taylor series, is something standard too. We will leave this as an instructive exercise, and come back to it later, in chapter 10 below. □

9e. Exercises

Exercises:

EXERCISE 9.27.

EXERCISE 9.28.

EXERCISE 9.29.

EXERCISE 9.30.

EXERCISE 9.31.

EXERCISE 9.32.

EXERCISE 9.33.

EXERCISE 9.34.

Bonus exercise.

CHAPTER 10

Trigonometric functions

10a. Complex exponential

We discuss now the theory of complex functions $f : \mathbb{C} \to \mathbb{C}$, in analogy with the theory of the real functions $f : \mathbb{R} \to \mathbb{R}$. We will see that many results that we know from the real setting extend to the complex setting. Let us start with something basic:

Definition 10.1. *A complex function $f : \mathbb{C} \to \mathbb{C}$, or more generally $f : X \to \mathbb{C}$, with $X \subset \mathbb{C}$ being a subset, is called continuous when, for any $x_n, x \in X$:*

$$x_n \to x \implies f(x_n) \to f(x)$$

where the convergence of the sequences of complex numbers, $x_n \to x$, means by definition that for n big enough, the quantity $|x_n - x|$ becomes arbitrarily small.

Observe that in real coordinates, $x = (a, b)$, the distances appearing in the definition of the convergence $x_n \to x$ are given by the following formula:

$$|x_n - x| = \sqrt{(a_n - a)^2 + (b_n - b)^2}$$

Thus $x_n \to x$ in the complex sense means that $(a_n, b_n) \to (a, b)$ in the usual, intuitive sense, with respect to the usual distance in the plane $\mathbb{R}^2$, and as a consequence, a function $f : \mathbb{C} \to \mathbb{C}$ is continuous precisely when it is continuous, in an intuitive sense, when regarded as function $f : \mathbb{R}^2 \to \mathbb{R}^2$. But more on this, later in this chapter.

At the level of examples, we have the following result:

Theorem 10.2. *We can exponentiate the complex numbers, according to the formula*

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

and the function $x \to e^x$ is continuous, and satisfies $e^{x+y} = e^x e^y$.

Proof. We must first prove that the series converges. But this follows from:

$$\begin{aligned} |e^x| &= \left|\sum_{k=0}^\infty \frac{x^k}{k!}\right| \\ &\leq \sum_{k=0}^\infty \left|\frac{x^k}{k!}\right| \\ &= \sum_{k=0}^\infty \frac{|x|^k}{k!} \\ &= e^{|x|} < \infty \end{aligned}$$

Regarding the formula $e^{x+y} = e^x e^y$, this follows too as in the real case, as follows:

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

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

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

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

$$\lim_{t\to 0} e^{x+t} = \lim_{t\to 0} e^x e^t = e^x \lim_{t\to 0} e^t = e^x \cdot 1 = e^x$$

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

We will be back to more functions later. As an important fact, however, let us point out that, contrary to what the above might suggest, everything does not always extend trivally from the real to the complex case. For instance, we have:

PROPOSITION 10.3. *We have the following formula, valid for any $|x|<1$,*

$$\frac{1}{1-x}=1+x+x^2+\ldots$$

but, unlike in the real case, the geometric meaning of this formula is quite unclear.

PROOF. Here the formula in the statement holds indeed, by multiplying and cancelling terms, and with the convergence being justified by the following estimate:

$$\left|\sum_{n=0}^\infty x^n\right|\leq\sum_{n=0}^\infty|x|^n=\frac{1}{1-|x|}$$

As for the last assertion, this is something quite informal. To be more precise, for $x=1/2$ our formula is clear, by cutting the interval $[0,2]$ into half, and so on:

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

More generally, for $x\in(-1,1)$ the meaning of the formula in the statement is something quite clear and intuitive, geometrically speaking, by using a similar argument. However, when x is complex, and not real, we are led into a kind of mysterious spiral there, and the only case where the formula is "obvious", geometrically speaking, is that when $x=rw$, with $r\in[0,1)$, and with w being a root of unity. To be more precise here, by anticipating a bit, assume that we have a number $w\in\mathbb C$ satisfying $w^N=1$, for some $N\in\mathbb N$. We have then the following formula, for our infinite sum:

$$\begin{aligned}1+rw+r^2w^2+\ldots&=&&(1+rw+\ldots+r^{N-1}w^{N-1})\\&+&&(r^N+r^{N+1}w\ldots+r^{2N-1}w^{N-1})\\&+&&(r^{2N}+r^{2N+1}w\ldots+r^{3N-1}w^{N-1})\\&+&&\ldots\end{aligned}$$

Thus, by grouping the terms with the same argument, our infinite sum is:

$$\begin{aligned}1+rw+r^2w^2+\ldots&=&&(1+r^N+r^{2N}+\ldots)\\&+&&(r+r^{N+1}+r^{2N+1}+\ldots)w\\&+&&\ldots\\&+&&(r^{N-1}+r^{2N-1}+r^{3N-1}+\ldots)w^{N-1}\end{aligned}$$

But the sums of each ray can be computed with the real formula for geometric series, that we know and understand well, and with an extra bit of algebra, we get:

$$\begin{aligned}1+rw+r^2w^2+\ldots &= \frac{1}{1-r^N}+\frac{rw}{1-r^N}+\ldots+\frac{r^{N-1}w^{N-1}}{1-r^N}\\ &= \frac{1}{1-r^N}\left(1+rw+\ldots+r^{N-1}w^{N-1}\right)\\ &= \frac{1}{1-r^N}\cdot\frac{1-r^N}{1-rw}\\ &= \frac{1}{1-rw}\end{aligned}$$

Summarizing, as claimed above, the geometric series formula can be understood, in a purely geometric way, for variables of type $x=rw$, with $r\in[0,1)$, and with w being a root of unity. In general, however, this formula tells us that the numbers on a certain infinite spiral sum up to a certain number, which remains something quite mysterious. □

10b. Polar writing

Getting back now to less mysterious mathematics, which in fact will turn to be quite mysterious as well, as is often the case with things involving complex numbers, as an application of all this, let us discuss the final and most convenient writing of the complex numbers, which is a variation on the polar writing, as follows:

$$x=re^{it}$$

The point with this formula comes from the following deep result:

THEOREM 10.4. *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 from Theorem 10.2 that we have the following formula, for the exponential of an arbitrary complex number $x\in\mathbb{C}$:

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

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

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

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

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

THEOREM 10.5. *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 10.4, 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 10.6. *The complex numbers $x = a + ib$ can be written in polar coordinates,*

$$x = re^{it}$$

with the connecting formulae being

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

and in the other sense being

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

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

PROOF. This is a reformulation of our previous polar writing notions, by using the formula $e^{it}=\cos t+i\sin t$ from Theorem 10.4, 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 10.7. *In polar coordinates, the complex numbers multiply as*

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

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

PROOF. This is something that we already know, from chapter 7, reformulated by using the notations from Theorem 10.6. Observe that this follows as well directly, from the fact that we have $e^{a+b}=e^ae^b$, 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 10.8. *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:* $(re^{it})^a=r^ae^{ita}$.
(4) *Conjugation:* $\overline{re^{it}}=re^{-it}$.

PROOF. This is something that we already know, from chapter 7, but we can now discuss all this, from a more conceptual viewpoint, the idea being as follows:

(1) We have indeed the following computation, using Theorem 10.7:

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

(2) Once again by using Theorem 10.7, 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 10.7, this operation $x\to x^a$ is indeed the correct one.

(4) This comes from the fact, that we know from chapter 7, that the conjugation operation $x\to\bar{x}$ keeps the modulus, and switches the sign of the argument. □

10c. Trigonometric functions

Getting now to more complicated functions, such as sin, cos, exp, log, again many things extend well from real to complex, the basic theory here being as follows:

THEOREM 10.9. *The functions* $\sin, \cos, \exp, \log$ *have complex extensions, given by*

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

$$e^x = \sum_{k=0}^\infty \frac{x^k}{k!} \quad , \quad \log(1+x) = \sum_{k=1}^\infty (-1)^{k+1} \frac{x^k}{k}$$

with $|x| < 1$ *needed for* $\log$, *which are continuous over their domain, and satisfy the formulae* $e^{x+y} = e^x e^y$ *and* $e^{ix} = \cos x + i \sin x$.

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

(1) We already know about e^x from before, the idea being that the convergence of the series, and then the continuity of e^x, come from the following estimate:

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

(2) Regarding $\sin x$, the same method works, with the following estimate:

$$|\sin x| \leq \sum_{l=0}^\infty \frac{|x|^{2l+1}}{(2l+1)!} \leq \sum_{k=0}^\infty \frac{|x|^k}{k!} = e^{|x|}$$

(3) The same goes for $\cos x$, the estimate here being as follows:

$$|\cos x| \leq \sum_{l=0}^\infty \frac{|x|^{2l}}{(2l)!} \leq \sum_{k=0}^\infty \frac{|x|^k}{k!} = e^{|x|}$$

(4) Regarding now the formulae satisfied by $\sin, \cos, \exp$, we already know from chapter 5 that the exponential has the following property, exactly as in the real case:

$$e^{x+y} = e^x e^y$$

We also have the following formula, connecting $\sin, \cos, \exp$, again as before:

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

(5) In order to discuss now the complex logarithm function log, let us first study some more the complex exponential function exp. By using $e^{x+y} = e^x e^y$ we obtain $e^x \neq 0$ for any $x \in \mathbb{C}$, so the complex exponential function is as follows:

$$\exp : \mathbb{C} \to \mathbb{C} - \{0\}$$

Now since we have $e^{x+iy} = e^x e^{iy}$ for $x, y \in \mathbb{R}$, with e^x being surjective onto $(0, \infty)$, and with e^{iy} being surjective onto the unit circle $\mathbb{T}$, we deduce that $\exp : \mathbb{C} \to \mathbb{C} - \{0\}$ is surjective. Also, again by using $e^{x+iy} = e^x e^{iy}$, we deduce that we have:

$$e^x = e^y \iff x - y \in 2\pi i \mathbb{Z}$$

(6) With these ingredients in hand, we can now talk about log. Indeed, let us fix a horizontal strip in the complex plane, having width 2π:

$$S = \left\{ x + iy \middle| x \in \mathbb{R}, y \in [a, a + 2\pi) \right\}$$

We know from the above that the restriction map $\exp : S \to \mathbb{C} - \{0\}$ is bijective, so we can define log as to be the inverse of this map:

$$\log = \exp^{-1} : \mathbb{C} - \{0\} \to S$$

(7) In practice now, the best is to choose for instance $a = 0$, or $a = -\pi$, as to have the whole real line included in our strip, $\mathbb{R} \subset S$. In this case on $\mathbb{R}_+$ we recover the usual logarithm, while on $\mathbb{R}_-$ we obtain complex values, as for instance $\log(-1) = \pi i$ in the case $a = 0$, or $\log(-1) = -\pi i$ in the case $a = -\pi$, coming from $e^{\pi i} = -1$.

(8) Finally, assuming $|x| < 1$, we can consider the following series, which converges:

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

We have then $e^{f(x)} = 1 + x$, and so $f(x) = \log(1 + x)$, when $1 + x \in S$. □

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

METHOD 10.10. *Knowing* $e^x=\sum_k x^k/k!$ *and* $e^{ix}=\cos x+i\sin x$ *gives you*

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

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

right away, in case you forgot these formulae, as well as

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

again, right away, in case you forgot these formulae.

To be more precise, assume that we forgot everything trigonometry, which is something that can happen to everyone, in the real life, but still know the formulae $e^x=\sum_k x^k/k!$ and $e^{ix}=\cos x+i\sin x$. Then, we can recover the formulae for sums, as follows:

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

And isn't this smart. Also, and even more impressively, we can recover the Taylor formulae for sin, cos, which are certainly difficult to memorize, as follows:

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

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

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

To be more precise, $\log(1+y)\simeq y$ is clear, and with a bit more work, that we will leave here as an instructive exercise, you can recover $\log(1+y)=y-y^2/2$ too. Of course, the higher terms can be recovered too, with enough work involved, at each step.

10d. Hyperbolic functions

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

THEOREM 10.11. *The following functions, called hyperbolic sine and cosine,*

$$\sinh x = \frac{e^x - e^{-x}}{2} \quad , \quad \cosh x = \frac{e^x + e^{-x}}{2}$$

are subject to the following formulae:

(1) $e^x = \cosh x + \sinh x$.
(2) $\sinh(ix) = i\sin x$, $\cosh(ix) = \cos x$, *for* $x \in \mathbb{R}$.
(3) $\sinh(x+y) = \sinh x \cosh y + \cosh x \sinh y$.
(4) $\cosh(x+y) = \cosh x \cosh y + \sinh x \sinh y$.
(5) $\sinh x = \sum_l \frac{x^{2l+1}}{(2l+1)!}$, $\cosh x = \sum_l \frac{x^{2l}}{(2l)!}$.

PROOF. The formula (1) follows from definitions. As for (2), this follows from:

$$\sinh(ix) = \frac{e^{ix} - e^{-ix}}{2} = \frac{\cos x + i\sin x}{2} - \frac{\cos x - i\sin x}{2} = i\sin x$$

$$\cosh(ix) = \frac{e^{ix} + e^{-ix}}{2} = \frac{\cos x + i\sin x}{2} + \frac{\cos x - i\sin x}{2} = \cos x$$

Regarding now (3,4), observe first that the formula $e^{x+y} = e^x + e^y$ reads:

$$\cosh(x+y) + \sinh(x+y) = (\cosh x + \sinh x)(\cosh y + \sinh y)$$

Thus, we have some good explanation for (3,4), and in practice, these formulae can be checked by direct computation, as follows:

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

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

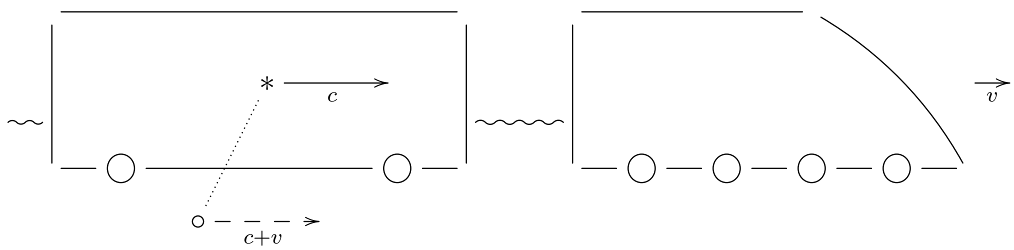
Finally, (5) is clear from the definition of sinh, cosh, and from $e^x = \sum_k \frac{x^k}{k!}$. □

Ready for some physics? Based on experiments by Fizeau, then Michelson-Morley and others, and some physics by Maxwell and Lorentz too, Einstein came upon:

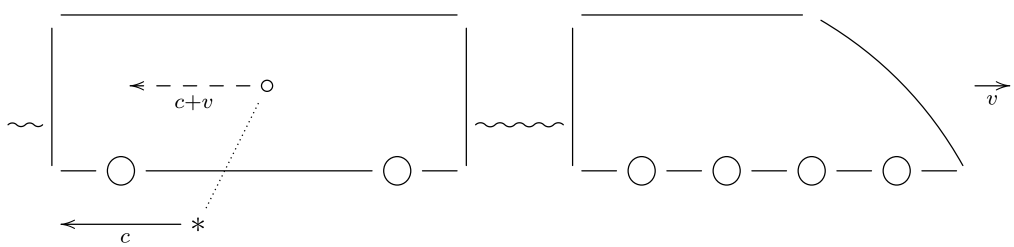
FACT 10.12 (Einstein principles). *The following happen:*

(1) *Light travels in vacuum at a finite speed,* $c < \infty$.
(2) *This speed* c *is the same for all inertial observers.*
(3) *In non-vacuum, the light speed is lower,* $v < c$.
(4) *Nothing can travel faster than light,* $v \not> c$.

The point now is that, obviously, something is wrong here. Indeed, assuming for instance that we have a train, running in vacuum at speed $v > 0$, and someone on board lights a flashlight $*$ towards the locomotive, then an observer $\circ$ on the ground will see the light travelling at speed $c+v>c$, which is a contradiction:



Equivalently, with the same train running, in vacuum at speed $v>0$, if the observer on the ground lights a flashlight $*$ towards the back of the train, then viewed from the train, that light will travel at speed $c+v>c$, which is a contradiction again:



Summarizing, Fact 10.12 implies $c+v=c$, so contradicts classical mechanics, which therefore needs a fix. By dividing all speeds by c, as to have $c=1$, and by restricting the attention to the 1D case, to start with, we are led to the following puzzle:

Puzzle 10.13. *How to define speed addition on the space of* 1D *speeds, which is*

$$I=[-1,1]$$

with our $c=1$ convention, as to have $1+c=1$, as required by physics?

In view of our geometric knowledge so far, a natural idea here would be that of wrapping $[-1,1]$ into a circle, and then stereographically projecting on $\mathbb{R}$. Indeed, we can then "import" to $[-,1,1]$ the usual addition on $\mathbb{R}$, via the inverse of this map.

So, let us see where all this leads us. First, the formula of our map is as follows:

Proposition 10.14. *The map wrapping $[-1,1]$ into the unit circle, and then stereographically projecting on $\mathbb{R}$ is given by the formula*

$$\varphi(u)=\tan\left(\frac{\pi u}{2}\right)$$

with the convention that our wrapping is the most straightforward one, making correspond $\pm 1\to i$, with negatives on the left, and positives on the right.

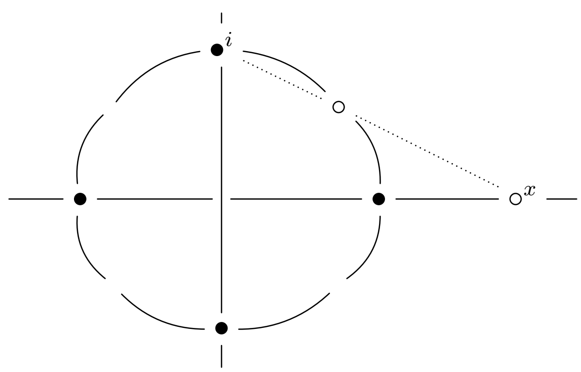
Proof. Regarding the wrapping, as indicated, this is given by:

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

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

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

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



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

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

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

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

Thus, we are led to the conclusion in the statement. $\square$

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

So, what to do? Obviously, trash Proposition 10.14, and start all over again. Getting back now to Puzzle 10.13, this has in fact a simpler solution, based this time on algebra, and which in addition is the good, physically correct solution, as follows:

THEOREM 10.15. *If we sum the speeds according to the Einstein formula*

$$u +_e v = \frac{u + v}{1 + uv}$$

then the Galileo formula still holds, approximately, for low speeds

$$u +_e v \simeq u + v$$

and if we have $u = 1$ or $v = 1$, the resulting sum is $u +_e v = 1$.

PROOF. All this is self-explanatory, and clear from definitions, and with the Einstein formula of $u +_e v$ itself being just an obvious solution to Puzzle 10.13, provided that, importantly, we know 0 geometry, and rely on very basic algebra only. □

So, very nice, problem solved, at least in 1D. But, shall we give up with geometry, and the stereographic projection? Certainly not, let us try to recycle that material. In order to do this, let us recall that the usual trigonometric functions are given by:

$$\sin x = \frac{e^{ix} - e^{-ix}}{2i} \quad , \quad \cos x = \frac{e^{ix} + e^{-ix}}{2} \quad , \quad \tan x = \frac{e^{ix} - e^{-ix}}{i(e^{ix} + e^{-ix})}$$

The point now is that, and you might know this from calculus, the above functions have some natural "hyperbolic" or "imaginary" analogues, constructed as follows:

$$\sinh x = \frac{e^{x} - e^{-x}}{2} \quad , \quad \cosh x = \frac{e^{x} + e^{-x}}{2} \quad , \quad \tanh x = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

But the function on the right, tanh, starts reminding the formula of Einstein addition, from Theorem 10.15. So, we have our idea, and we are led to the following result:

THEOREM 10.16. *The Einstein speed summation in 1D is given by*

$$\tanh x +_e \tanh y = \tanh(x + y)$$

with $\tanh : [-\infty, \infty] \to [-1, 1]$ *being the hyperbolic tangent function.*

PROOF. This follows by putting together our various formulae above, but it is perhaps better, for clarity, to prove this directly. Our claim is that we have:

$$\tanh(x + y) = \frac{\tanh x + \tanh y}{1 + \tanh x \tanh y}$$

But this can be checked via direct computation, from the definitions, as follows:

$$\begin{aligned}
&\frac{\tanh x+\tanh y}{1+\tanh x\tanh y}\\
&= \left(\frac{e^x-e^{-x}}{e^x+e^{-x}}+\frac{e^y-e^{-y}}{e^y+e^{-y}}\right)\Big/\left(1+\frac{e^x-e^{-x}}{e^x+e^{-x}}\cdot\frac{e^y-e^{-y}}{e^y+e^{-y}}\right)\\
&= \frac{(e^x-e^{-x})(e^y+e^{-y})+(e^x+e^{-x})(e^y-e^{-y})}{(e^x+e^{-x})(e^y+e^{-y})+(e^x-e^{-x})(e^y+e^{-y})}\\
&= \frac{2(e^{x+y}-e^{-x-y})}{2(e^{x+y}+e^{-x-y})}\\
&= \tanh(x+y)
\end{aligned}$$

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

Very nice all this, hope you agree. As a conclusion, passing from the Riemann stereographic projection sum to the Einstein summation basically amounts in replacing:

$$\tan\to\tanh$$

Let us formulate as well this finding more philosophically, as follows:

Conclusion 10.17. *The Einstein speed summation in 1D is the imaginary analogue of the summation on* $[-1,1]$ *obtained via Riemann's stereographic projection.*

Which looks quite deep, and we will stop here. More on this later in this book, when discussing curved spacetime, in full generality, and with more advanced tools.

10e. Exercises

Exercises:

Exercise 10.18.

Exercise 10.19.

Exercise 10.20.

Exercise 10.21.

Exercise 10.22.

Exercise 10.23.

Exercise 10.24.

Exercise 10.25.

Bonus exercise.

CHAPTER 11

Sums, estimates

11a. More about e

Time for some tough calculus. We first have the following result, about e:

THEOREM 11.1. *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 e, with the old style one, which is quite cool, and that you can still find in fine calculus books, being:

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

The definition in the statement is the modern one. Indeed, imagine that you are looking for a function exp, satisfying $\exp'=\exp$, and $\exp(0)=1$. With $\exp(x)=\sum c_kx^k$, you must have $c_0=1$, then $c_1=1$, $c_2=1/2$, $c_3=1/6$ and so on, meaning:

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

But now, it is an easy exercice to show that $\exp(x+y)=\exp(x)\exp(y)$, which gives $\exp(x)=e^x$, for a certain number $e>0$. Which number e can only be $e=\exp(1)$.

(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. □

As a continuation, we have the following result, which is substantially harder:

THEOREM 11.2. *The number e is transcendental.*

PROOF. Assume by contradiction that e is algebraic, with this meaning that it is a root of a polynomial with integer coefficients, $c_i \in \mathbb{Z}$, as follows:

$$c_0 + c_1 e + \ldots + c_n e^n = 0$$

(1) Following Hermite, consider the following polynomials, and we will see later why:

$$f_k(x) = x^k\left[(x-1)\ldots(x-n)\right]^{k+1}$$

Consider also the following quantities, that we will study more in detail later:

$$A_k = \int_0^\infty f_k(x)e^{-x}dx$$

By multiplying our equation for e by this quantity A_k, we obtain:

$$c_0\int_0^\infty f_k(x)e^{-x}dx+c_1\int_0^\infty f_k(x)e^{1-x}dx+\ldots c_n\int_0^\infty f_k(x)e^{n-x}dx=0$$

(2) Here comes the trick. Consider the following two quantities:

$$P=c_0\int_0^\infty f_k(x)e^{-x}dx+c_1\int_1^\infty f_k(x)e^{1-x}dx+\ldots+c_n\int_n^\infty f_k(x)e^{n-x}dx$$

$$Q=c_1\int_0^1 f_k(x)e^{-x}dx+\ldots+c_n\int_0^n f_k(x)e^{n-x}dx$$

In terms of these quantities, the formula that we found in (1) reads:

$$P+Q=0$$

(3) Now let us look at P. Our claim is that this is an integer, $P\in\mathbb{Z}$, and that there are arbitrarily large numbers $k>>0$ for which the following holds:

$$\frac{P}{k!}\in\mathbb{Z}-\{0\}$$

Indeed, according to our formula above defining P, we have:

$$\begin{aligned}P&=\sum_{r=0}^n c_r\int_r^\infty f_k(x)e^{r-x}dx\\&=\sum_{r=0}^n c_r\int_0^\infty f_k(x+r)e^{-x}dx\\&=\int_0^\infty\left(\sum_{r=0}^n c_rf_k(x+r)\right)e^{-x}dx\end{aligned}$$

On the other hand, integrating such functions is easy, according to:

$$\begin{aligned}\int_0^\infty x^se^{-x}dx&=\int_0^\infty\left(\frac{x^{s+1}}{s+1}\right)'e^{-x}dx\\&=\int_0^\infty\frac{x^{s+1}}{s+1}e^{-x}dx\\&=\frac{1}{s+1}\int_0^\infty x^{s+1}e^{-x}dx\end{aligned}$$

Thus, we are led by recurrence on $s\in\mathbb{N}$ to the following formula:

$$\int_0^\infty x^se^{-x}dx=s!$$

For a linear combination now, we are led to the following formula:

$$g(x)=\sum_s a_sx^s\implies\int_0^\infty g(x)e^{-x}dx=\sum_s a_ss!$$

But this shows that we have indeed $P\in\mathbb{Z}$, and also, via a bit of study based on the exact formula of f_k, from the beginning of (1), that we have in fact:

$$\frac{P}{k!}\in\mathbb{Z}$$

Finally, we still have to prove that we have $P\neq 0$, for arbitrarily large numbers $k>>0$. But the point here is that for $k+1>n,|c_0|$, chosen prime, a detailed study of our integral shows that we have $(k+1)\not| P$, and so $P\neq 0$ indeed, as desired.

(4) With this done, let us look now at Q. Our claim is that for $k>>0$ we have:

$$\left|\frac{Q}{k!}\right|<1$$

Indeed, by using the exact formula of f_k, from the beginning of (1), we have:

$$\begin{aligned}
f_k(x)e^{-x}&=x^k\left[(x-1)\ldots(x-n)\right]^{k+1}e^{-x}\\
&=\left[x(x-1)\ldots(x-n)\right]^k\times(x-1)\ldots(x-n)e^{-x}
\end{aligned}$$

We conclude that for $x\in[0,n]$ we have an estimate as follows, with $G,H>0$ being certain constants, appearing as maxima of the two components appearing above:

$$|f_k(x)e^{-x}|<G^kH$$

Now by integrating, we obtain from this the following estimate for Q itself:

$$\begin{aligned}
|Q|&=\left|c_1\int_0^1 f_k(x)e^{-x}dx+\ldots+c_ne^n\int_0^n f_k(x)e^{-x}dx\right|\\
&\leq |c_1|\int_0^1|f_k(x)e^{-x}|dx+\ldots+|c_n|e^n\int_0^n|f_k(x)e^{-x}|dx\\
&\leq |c_1|\cdot G^kH+\ldots+|c_n|e^n\cdot nG^kH\\
&=(|c_1|e+\ldots+|c_n|e^n)\,\frac{n(n+1)}{2}\,G^kH
\end{aligned}$$

But in this estimate the only term depending on k is the power G^k, and since since $k!$ grows much faster than this power G^k, this proves our claim:

$$\left|\frac{Q}{k!}\right|\approx\frac{G^k}{k!}\to 0$$

(5) And with this, done, because what we found in (3,4) contradicts the formula $P+Q=0$ from (2). Thus e is indeed transcendental, as claimed. $\square$

11b. More about pi

Let us prove now, a bit as for e before, that π is irrational, and even transcendental. Let us start with:

THEOREM 11.3. *The number π is irrational.*

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

As a continuation, we have the following result, which is substantially harder:

THEOREM 11.4. *The number π is transcendental.*

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

11c. Sums, estimates

Sums, estimates.

11d. Special functions

Special functions.

11e. Exercises

Exercises:

EXERCISE 11.5.

EXERCISE 11.6.

EXERCISE 11.7.

EXERCISE 11.8.

EXERCISE 11.9.

EXERCISE 11.10.

EXERCISE 11.11.

EXERCISE 11.12.

Bonus exercise.

CHAPTER 12

Into arithmetic

12a. Squares, residues

Let us go back to what we did before with congruences. Our aim here will be that of further building on some of the theorems there. To be more precise, we will be interested in solving the following ubiquitous equation, over the integers:

$$a=b^2(c)$$

Many things can be said here, of various levels of difficulty. Inspired by all this, we have the following definition, putting everything on a solid basis:

Definition 12.1. *The Legendre symbol is defined as follows,*

$$\left(\frac{a}{p}\right)=\begin{cases}1&\text{if }\exists\, b\neq 0, a=b^2(p)\\ 0&\text{if }a=0(p)\\ -1&\text{if }\nexists\, b, a=b^2(p)\end{cases}$$

with $p\geq 3$ prime.

Now leaving aside all sorts of nice and amateurish things that can be said about $a=b^2(c)$, and going straight to the point, what we want to do is to compute this symbol. I mean, if we manage to have this symbol computed, that would be a big win.

As a first result on the subject, due to Euler, we have:

Theorem 12.2. *The Legendre symbol is given by the formula*

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

called Euler formula for the Legendre symbol.

Proof. This is something not that complicated, the idea being as follows:

(1) We know from Fermat that we have $a^p=a(p)$, and leaving aside the case $a=0(p)$, which is trivial, and therefore solved, this tells us that $a^{p-1}=1(p)$. But since our prime p was assumed to be odd, $p\geq 3$, we can write this formula as follows:

$$\left(a^{\frac{p-1}{2}}-1\right)\left(a^{\frac{p-1}{2}}+1\right)=0(p)$$

(2) Now let us think a bit at the elements of $\mathbb{F}_p - \{0\}$, which can be a quadratic residue, and which cannot. Since the squares b^2 with $b \neq 0$ are invariant under $b \to -b$, and give different b^2 values modulo p, up to this symmetry, we conclude that there are exactly $(p-1)/2$ quadratic residues, and with the remaining $(p-1)/2$ elements of $\mathbb{F}_p - \{0\}$ being non-quadratic residues. So, as a conclusion, $\mathbb{F}_p - \{0\}$ splits as follows:

$$\mathbb{F}_p - \{0\} = \left\{\frac{p-1}{2} \ squares\right\} \bigsqcup \left\{\frac{p-1}{2} \ non\!-\!squares\right\}$$

(3) Now by comparing what we have in (1) and in (2), the splits there must correspond to each other, so we are led to the following formula, valid for any $a \in \mathbb{F}_p - \{0\}$:

$$a^{\frac{p-1}{2}} = \begin{cases} 1 & \text{if } \exists\, b, a = b^2 \\ -1 & \text{if } \nexists\, b, a = b^2 \end{cases}$$

By comparing now with Definition 12.1, we obtain the formula in the statement. □

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

PROPOSITION 12.3. *We have the following formula, valid for any $a, b \in \mathbb{Z}$:*

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

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

PROOF. This is clear indeed from the Euler formula, because $a^{\frac{p-1}{2}}(p)$ is obviously multiplicative in $a \in \mathbb{Z}$. Alternatively, this can be proved as well directly, 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 12.2. □

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

PROPOSITION 12.4. *We have the following formula, telling us that modulo any prime number p, a product of non-squares is a square:*

$$\left(\frac{a}{p}\right) = -1 \ , \ \left(\frac{b}{p}\right) = -1 \implies \left(\frac{ab}{p}\right) = 1$$

Also, the Legendre symbol, regarded as a function

$$\chi : \mathbb{F}_p - \{0\} \to \{-1, 1\} \quad , \quad \chi(a) = \left(\frac{a}{p}\right)$$

is a character, in the sense that it is multiplicative.

PROOF. The first asssertion is a consequence of Proposition 12.3, more or less equivalent to it, and with the remark that this formally holds at $p = 2$ too, as $\emptyset \implies \emptyset$. As for the second assertion, this is just a fancy reformulation of Proposition 12.3. □

It is possible to say some further conceptual things, some sounding very fancy, in relation with Proposition 12.3 and Proposition 12.4. But remember that, according to the plan made in the beginning of this chapter, we are here for the kill, namely computing the Legendre symbol, no matter what, and with no prisoners taken.

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 12.5. *We have the quadratic reciprocity formula*

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

valid for any primes $p,q\geq 3$.

PROOF. This is something quite tricky, one proof being as follows:

(1) First we have a combinatorial formula for the Legendre symbol, called Gauss lemma. Given a prime number $q\geq 3$, and $a\neq 0(q)$, consider the following sequence:

$$a\ ,\ 2a\ ,\ 3a\ ,\ \ldots\ ,\ \frac{q-1}{2}\,a$$

The Gauss lemma tells us that if we look at these numbers modulo q, and denote by n the number of residues modulo q which are greater than $q/2$, then:

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

(2) In order to prove this lemma, the idea is to look at the following product:

$$Z=a\times 2a\times 3a\times\ldots\times\frac{q-1}{2}\,a$$

Indeed, on one hand we have the following formula, with Euler used at the end:

$$Z=a^{\frac{q-1}{2}}\left(\frac{q-1}{2}\right)!=\left(\frac{a}{q}\right)\left(\frac{q-1}{2}\right)!$$

(3) On the other hand, we can compute Z in more complicated way, but leading to a simpler answer. Indeed, let us define the following function:

$$|x|=\begin{cases}x & \text{if } 0<x<q/2\\ q-x & \text{if } q/2<x<q\end{cases}$$

With this convention, our product Z is given by the following formula, with n being as in (1), namely the number of residues modulo q which are greater than $q/2$:

$$Z=(-1)^n\times|a|\times|2a|\times|3a|\times\ldots\times\left|\frac{q-1}{2}\,a\right|$$

(4) But, the numbers $|ra|$ appearing in the above formula are all distinct, so up to a permutation, these must be exactly the numbers $1,2,\ldots,\frac{q-1}{2}$. That is, we have:

$$\left\{|a|\,,\,|2a|\,,\,|3a|\,,\,\ldots\,,\,\left|\frac{q-1}{2}\,a\right|\right\}=\left\{1,2,3,\ldots,\frac{q-1}{2}\right\}$$

Now by multiplying all these numbers, we obtain, via the formula in (3):

$$Z=(-1)^n\left(\frac{q-1}{2}\right)!$$

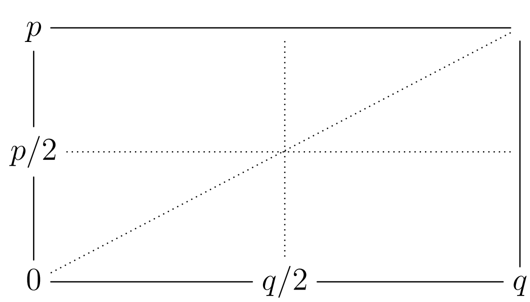
(5) But this is what we need, because when comparing with what we have in (2), we obtain the following formula, which is exactly the one claimed by the Gauss lemma:

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

(6) Next, we have a variation of this formula, due to Eisenstein. His formula for the Legendre symbol, this time involving a prime number numerator $p\geq 3$ in the symbol, is as follows, with the quantities on the right being integer parts, and with the proof being very similar to the proof of the Gauss lemma, that we will leave here as an exercise:

$$\left(\frac{p}{q}\right)=(-1)^n\quad,\quad n=\sum_{k=0}^{(q-1)/2}\left[\frac{2kp}{q}\right]$$

(7) The key point now is that, in this latter formula of Eisenstein, the number n itself counts the points of the lattice $\mathbb{Z}^2$ lying in the triangle $(0,0),(q,0),(q,p)$. So, based on this observation, let us draw a picture, as follows:



(8) We must count the points of $\mathbb{Z}^2$ lying in the triangle $(0,0),(q,0),(q,p)$, modulo 2. This triangle has 3 components, when split by the dotted lines above. Since the points at right, in the small rectangle, and in the small triangle above it, will cancel modulo 2,

we are left with the points at left, in the small triangle there, and the conclusion is that, if we denote by m the number of integer points there, we have the following formula:

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

(9) Now by flipping the diagram, we have as well the following formula, with r being the number of integer points in the small triangle above the small triangle in (8):

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

(10) But, since our two small triangles add up to a small rectangle, we have:

$$m + r = \frac{p-1}{2} \cdot \frac{q-1}{2}$$

Thus, by multiplying the formulae in (8) and (9), we are led to the result. □

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 12.5 with Proposition 12.3, it is quite clear that, no matter how big p is, if a has only small prime factors, we are saved.

Besides Proposition 12.3, the quadratic reciprocity formula comes accompanied by two other statements, which are very useful in practice. First, at $a = -1$, we have:

PROPOSITION 12.6. *We have the following formula,*

$$\left(\frac{-1}{p}\right) = \begin{cases} 1 & \text{if } p = 1(4) \\ -1 & \text{if } p = 3(4) \end{cases}$$

solving in practice the equation $b^2 = -1(p)$.

PROOF. This follows from the Euler formula, which at $a = -1$ reads:

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

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

As a second useful result, this time at $a = 2$, we have:

THEOREM 12.7. *We have the following formula,*

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

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

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

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

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

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

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

As a continuation of this, speaking Legendre symbol for small values of the upper variable, we can try to compute these for $a=\pm\,3,4,5,6,7,8,\ldots$ But by multiplicativity plus Proposition 12.6 plus Theorem 12.7 we are left with the case where $a=q$ is an odd prime, and we can solve the problem with quadratic reciprocity, so done.

Let us record however a few statements here, which can be useful in practice, and with this being mostly for illustration purposes, for Theorem 12.5. We first have:

PROPOSITION 12.8. *We have the following formula,*

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

valid for any prime $p\geq 5$.

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

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

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

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

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

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

PROPOSITION 12.9. *We have the following formula,*

$$\left(\frac{5}{p}\right)=\begin{cases}1 & \text{if } p=1,4(5)\\ -1 & \text{if } p=2,3(5)\end{cases}$$

valid for any odd prime $p\neq 5$.

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

$$\left(\frac{5}{p}\right)=(-1)^{\frac{5-1}{2}\cdot\frac{p-1}{2}}\left(\frac{p}{5}\right)=\left(\frac{p}{5}\right)$$

Thus, we have the result. Alternatively, we have the following formula:

$$\left(\frac{5}{p}\right)=(-1)^{\left[\frac{2p+2}{5}\right]}$$

Indeed, this is the simplest implementation of the formula in the statement. □

Moving ahead now, we have the following interesting generalization of the Legendre symbol, to the case of denominators not necessarily prime, due to Jacobi:

THEOREM 12.10. *The theory of Legendre symbols can be extended by multiplicativity into a theory of Jacobi symbols, according to the formula*

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

with the denominator being not necessarily prime, but just an arbitrary odd number, and this theory has as results those imported from the Legendre theory.

PROOF. This is something self-explanatory, and we will leave listing the basic properties of the Jacobi symbols, based on the theory of Legendre symbols, as an exercise. □

The story is not over with Jacobi, because the denominator there is still odd, and positive. So, we have a problem to be solved, the solution to it being as follows:

THEOREM 12.11. *The theory of Jacobi symbols can be further extended into a theory of Kronecker symbols, according to the formula*

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

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

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

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

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

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

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. □

12b. Gauss sums

Time for the roots of unity to strike again, this time with some non-trivial applications to the Legendre symbols. Going back to what we learned about these symbols, there were several mysterious things there, that we will attempt to elucidate now.

Let us start with the $a=2$ case. The result here is as follows:

THEOREM 12.12. *We have the following formula,*

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

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

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

(1) As a first observation, the Euler formula at $a=2$ is as follows, obviously well below the quality of the very precise formula in the statement:

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

As a second observation, the quadratic reciprocity formula, assuming that known, cannot help either, because in that formula $p,q\geq 3$ are odd primes.

(2) Thus, we must prove the result. As already mentioned before, the proof will come via the following formula, which is equivalent to the formula in the statement:

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

Finally, let us mention too that, despite 2 being an even prime, the problematics here is a bit similar to the one of the quadratic reciprocity formula, and the proof below will contain many good ideas, that we will use later in the proof of quadratic reciprocity.

(3) Getting started now, let us set $w=e^{\pi i/4}$, so that $w^2=i$, do not ask me why, and then $t=w+w^{-1}$. We have of course $t=\sqrt{2}$, but it is better to forget this, and do formal arithmetics instead, with integers as scalars, based on the following computation:

$$\begin{aligned}t^2&=2+w^2+w^{-2}\\&=2+i-i\\&=2\end{aligned}$$

Now by using the Euler formula for the Legendre symbol, we have:

$$\begin{aligned}\left(\frac{2}{p}\right)&=2^{\frac{p-1}{2}}\ (p)\\&=(t^2)^{\frac{p-1}{2}}\ (p)\\&=t^{p-1}\ (p)\end{aligned}$$

(4) By multiplying now by t we obtain from this, in a formal sense, and I will leave it you to clarify all the details here, namely what this formal sense exactly means:

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

(5) On the other hand, by using the binomial formula, and the standard fact that all non-trivial binomial coefficients are multiples of p, we obtain, again formally:

$$\begin{aligned}t^p&=(w+w^{-1})^p\\&=\sum_{k=0}^p\binom{k}{p}w^kw^{k-p}\\&=w^p+w^{-p}\ (p)\end{aligned}$$

(6) Now let us look at w^p+w^{-p}, as usual complex number. Since $w=e^{\pi i/4}$, this quantity will depend only on p modulo 8, and more precisely, we have:

$$w^p+w^{-p}=\begin{cases}w+w^{-1}&\text{if } p=\pm1(8)\\-w-w^{-1}&\text{if } p=\pm3(8)\end{cases}$$

Thus $w^p + w^{-p} = \pm t$, with the sign depending on p modulo 8, and more specifically:

$$w^p + w^{-p} = (-1)^{\frac{p^2-1}{8}} t$$

(7) Time now to put everything together. By combining (4,5,6) we obtain:

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

By dividing by t, this gives the following formula:

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

But the mod p symbol can now be dropped, because our equality is between two ± 1 quantities, and we obtain the formula in the statement. □

With the same idea, we can prove as well the quadratic reciprocity theorem:

THEOREM 12.13. *We have the quadratic reciprocity formula*

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

valid for any primes $p, q \geq 3$.

PROOF. This is something already advertised in the above, and we refer to the discussion there for the mighty power of this formula, and its enigmatic nature. However, thinking a bit, our $t = w + w^{-1}$ trick above can be adapted, as follows:

(1) To start with, we need an analogue of that $t = w + w^{-1}$ variable. For this purpose, let us set $w = e^{2\pi i/q}$, now that we have a prime $q \geq 3$ involved, and then:

$$t = \sum_{k=0}^{q-1} w^{k^2}$$

Observe that at $q = 2$, excluded by the statement, we have $w = -1$, and so $t = 1 + (-1) = 0$, instead of the $t = w + w^{-1}$ with $w = e^{\pi i/4}$ used before. However, believe me, this is due to some bizarre reasons, and the above t is the good variable, at $q \geq 3$.

(2) The above variable t is called Gauss sum, can be defined for any $q \in \mathbb{N}$, not necessarily prime, and can be explicitely computed, the formula being as follows:

$$t = \begin{cases} \sqrt{q} & \text{if } q = 1(4) \\ 0 & \text{if } q = 2(4) \\ \sqrt{q}\, i & \text{if } q = 3(4) \\ \sqrt{q}(1+i) & \text{if } q = 0(4) \end{cases}$$

In particular, assuming that q is odd, as is our $q \geq 3$ prime, we have:

$$t^2 = \begin{cases} q & \text{if } q = 1(4) \\ -q & \text{if } q = 3(4) \end{cases}$$

(3) In what follows we will only need this latter formula, for $q \geq 3$ prime, so let us prove this now, and with the comment that the proof of the first formula in (2) is something quite complicated, and better avoid that. We have, by definition of our variable t:

$$\begin{aligned} |t|^2 &= \sum_{kl} w^{k^2-l^2} \\ &= \sum_{kl} w^{(k+l)(k-l)} \\ &= \sum_{lr} w^{r(2l+r)} \\ &= \sum_r w^{r^2} \sum_l (w^{2r})^l \\ &= q \end{aligned}$$

(4) On the other hand, it is easy to see that t^2 is real, so $t^2 = \pm q$. With a bit more work it is possible to compute the sign too, $t^2 = (-1)^{\frac{q-1}{2}} q$, but we will not need this here, because the sign will come for free at the end of the proof, via a symmetry argument. So, as a conclusion, we have a formula as follows, for a certain $e_q \in \{0, 1\}$:

$$t^2 = (-1)^{e_q} q$$

(5) With this done, let us turn to the proof of our theorem, by using the variable t a bit as before, in the proof of Theorem 12.12. By using the Euler formula, we have:

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

By multiplying now by t we obtain from this, in a formal sense:

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

(6) In order to compute now t^p by other means, observe first that, if we denote by $\mathbb{Z}_q - \{0\} = S \sqcup N$ the partition into squares and non-squares, we have:

$$\begin{aligned} t &= \sum_{k=0}^{q-1} w^{k^2} \\ &= 1 + 2\sum_{s\in S} w^s \\ &= \sum_{s\in S} w^s - \sum_{s\in N} w^s \\ &= \sum_{r=0}^{k-1} \left(\frac{r}{q}\right) w^r \end{aligned}$$

(7) By using now the multinomial formula, with the observation that all the non-trivial multinomial coefficients are multiples of p, we obtain, in a formal sense:

$$\begin{aligned} t^p &= \left(\sum_r \left(\frac{r}{q}\right) w^r\right)^p \\ &= \sum_r \left(\frac{r}{q}\right) w^{rp}\ (p) \\ &= \sum_s \left(\frac{p^{-1}s}{q}\right) w^s\ (p) \\ &= \left(\frac{p^{-1}}{q}\right) \sum_s \left(\frac{s}{q}\right) w^s\ (p) \\ &= \left(\frac{p}{q}\right) t\ (p) \end{aligned}$$

(8) Time now to put everything together. By combining (5,7) we obtain:

$$\left(\frac{t^2}{p}\right) t = \left(\frac{p}{q}\right) t\ (p)$$

We can divide by t, and then drop the modulo p symbol, because our new equality, without t, is between two ± 1 quantities, and we obtain:

$$\left(\frac{t^2}{p}\right) = \left(\frac{p}{q}\right)$$

Now by taking into account the formula found in (4), this reads:

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

By using the Euler formula for the symbol on the left, we obtain from this:

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

Now by symmetry we must have $e_q = \frac{q-1}{2}$, and this finishes the proof. □

As a conclusion, the quadratic reciprocity theorem can be established via Gauss sums t, and this is certainly excellent news. However, we have mentioned in step (2) of our proof above a very nice, powerful formula for the Gauss sum t itself, and this even in the general case, where $q \in \mathbb{N}$ is not necessarily prime. We refer here to the literature.

12c. Prime numbers

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

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

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

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

valid for any integer $N \geq 2$.

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

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

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

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

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

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

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

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

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

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

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

12d. Zeta function

We have already met the Riemann zeta function on several occasions, in the above, at values $s > 1$ of the parameter, with the conclusion every time that this function is intimately related to the primes. In this chapter we discuss a systematic approach to this phenomenon, by using complex analysis. As a first observation, we can talk without much pain about zeta at complex values of s as well, in the following way:

THEOREM 12.15. *We can talk about the Riemann zeta function*

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

at any $s\in\mathbb{C}$ *with* $Re(z)>1$.

PROOF. We have the following computation, assuming $s=r+it$ with $r>1$:

$$\begin{aligned}|\zeta(s)| &= \left|\sum_{n=1}^\infty\frac{1}{n^s}\right|\\ &\leq \sum_{n=1}^\infty\frac{1}{|n^s|}\\ &\leq \sum_{n=1}^\infty\frac{1}{n^r}\\ &< 1+\int_1^\infty\frac{1}{x^r}\,dx\\ &= 1+\left[\frac{x^{1-r}}{1-r}\right]_1^\infty\\ &= 1+\frac{1}{r-1}\end{aligned}$$

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

As a first result, we can write zeta as an Euler product, as follows:

PROPOSITION 12.16. *We have the following formula,*

$$\zeta(s)=\prod_p\left(1-\frac{1}{p^s}\right)^{-1}$$

valid for any exponent $s\in\mathbb{C}$ *with* $Re(s)>1$.

Proof. We have the following computation, with everything converging:

$$\begin{aligned}\zeta(s) &= \sum_{n=1}^\infty \frac{1}{n^s}\\ &= \prod_p \left(1+\frac{1}{p^s}+\frac{1}{p^{2s}}+\frac{1}{p^{3s}}+\ldots\right)\\ &= \prod_p \left(1-\frac{1}{p^s}\right)^{-1}\end{aligned}$$

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

We have as well the following formula, which is elementary too:

Proposition 12.17. *We have the following formula,*

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

with μ being the Möbius function, given by the formula

$$\mu(n) = \begin{cases} (-1)^k & \text{if } n = p_1 \ldots p_k \\ 0 & \text{if } n \text{ is not square} - \text{free} \end{cases}$$

valid for any exponent $s \in \mathbb{C}$ with $Re(s) > 1$.

Proof. We have the following computation, with everything converging:

$$\begin{aligned}\frac{1}{\zeta(s)} &= \prod_p \left(1-\frac{1}{p^s}\right)\\ &= \sum_{k=0}^\infty (-1)^k \prod_{p_1\ldots p_k} \frac{1}{p_1^s \ldots p_k^s}\\ &= \sum_{n=1}^\infty \frac{\mu(n)}{n^s}\end{aligned}$$

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

Along the same lines, as another elementary result, we have:

Proposition 12.18. *The square of the zeta function is given by*

$$\zeta^2(s) = \sum_{n=1}^\infty \frac{\tau(n)}{n^s}$$

with $\tau(n)$ being the number of divisors of n, for any $s \in \mathbb{C}$ with $Re(s) > 1$.

Proof. We have the following computation, with everything converging:

$$\zeta(s)^2=\sum_{k=1}^\infty\sum_{l=1}^\infty\frac{1}{(kl)^s}=\sum_{n=1}^\infty\frac{\tau(n)}{n^s}$$

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

In order to present now a more advanced result, we will need:

Proposition 12.19. *We can talk about the gamma function*

$$\Gamma(s)=\int_0^\infty x^{s-1}e^{-x}\,dx$$

extending the usual factorial of integers, $\Gamma(s)=(s-1)!$.

Proof. The integral converges indeed, and by partial integration we have:

$$\begin{aligned}\Gamma(s+1) &= \int_0^\infty x^se^{-x}\,dx\\ &= \int_0^\infty sx^{s-1}e^{-x}\,dx\\ &= s\Gamma(s)\end{aligned}$$

Regarding now the case $s\in\mathbb N$, for the initial value $s=1$ we have:

$$\Gamma(1)=\int_0^\infty e^{-x}dx=1$$

Thus, for $s\in\mathbb N$ we have indeed $\Gamma(s)=(s-1)!$, as claimed. □

We can now formulate a key result about zeta, as follows:

Theorem 12.20. *We have the following formula,*

$$\zeta(s)=\frac{1}{\Gamma(s)}\int_0^\infty\frac{x^{s-1}}{e^x-1}\,dx$$

valid for any $s\in\mathbb C$ *with* $Re(s)>1$.

Proof. We have indeed the following computation:

$$\begin{aligned}
\int_0^\infty\frac{x^{s-1}}{e^x-1}\,dx &= \int_0^\infty\frac{x^{s-1}}{e^x}\cdot\frac{1}{1-e^{-x}}\,dx\\
&= \int_0^\infty x^{s-1}(e^{-x}+e^{-2x}+e^{-3x}+\ldots)\\
&= \sum_{n=1}^\infty\int_0^\infty x^{s-1}e^{-nx}\,dx\\
&= \sum_{n=1}^\infty\int_0^\infty\left(\frac{y}{n}\right)^{s-1}e^{-y}\,\frac{dy}{n}\\
&= \sum_{n=1}^\infty\frac{1}{n^s}\int_0^\infty y^{s-1}e^{-y}\,dy\\
&= \zeta(s)\Gamma(s)
\end{aligned}$$

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

At a more advanced level, we can try to compute particular values of ζ. Things are quite tricky here, and we have the following result, briefly discussed before:

Theorem 12.21. *We have the following formula, for the even integers $s=2k$,*

$$\zeta(2k)=(-1)^{k+1}\frac{(2\pi)^{2k}B_{2k}}{2\cdot(2k)!}$$

with B_n being the Bernoulli numbers, which in practice gives the formulae

$$\zeta(2)=\frac{\pi^2}{6}\quad,\quad\zeta(4)=\frac{\pi^4}{90}\quad,\quad\zeta(6)=\frac{\pi^6}{945}\quad,\quad\zeta(8)=\frac{\pi^8}{9450}\quad,\quad\ldots$$

generalizing the formula $\zeta(2)=\pi^2/6$ of Euler, solving the Basel problem.

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

(1) To start with, at $s=2$ the Euler computation, from before, was as follows:

$$\begin{aligned}
\frac{\sin x}{x} &= 1-\frac{x^2}{3!}+\frac{x^4}{5!}-\frac{x^6}{7!}+\ldots\\
&= \left(1-\frac{x}{\pi}\right)\left(1+\frac{x}{\pi}\right)\left(1-\frac{x}{2\pi}\right)\left(1+\frac{x}{2\pi}\right)\ldots\\
&= \left(1-\frac{x^2}{\pi^2}\right)\left(1-\frac{x^2}{4\pi^2}\right)\left(1-\frac{x^2}{9\pi^2}\right)\ldots\\
&= 1-\frac{1}{\pi^2}\sum_{n=1}^\infty\frac{1}{n^2}\,x^2+\ldots
\end{aligned}$$

It is possible to use the same idea for dealing with $\zeta(2k)$ with $k \in \mathbb{N}$, but this is quite complicated, and in addition the above method of Euler needs some justification, that we have not really provided before, so in short, not a path to be followed.

(2) Instead, we have the following luminous computation, based on Theorem 12.20:

$$\begin{aligned}
\zeta(2k) &= \frac{1}{\Gamma(2k)}\int_0^\infty \frac{x^{2k-1}}{e^x-1}\,dx \\
&= \frac{1}{(2k-1)!}\int_0^\infty \frac{x^{2k-1}}{e^x-1}\,dx \\
&= \frac{1}{(2k-1)!}\int_0^\infty \frac{(2\pi t)^{2k-1}}{e^{2\pi t}-1}\,2\pi dt \\
&= \frac{(2\pi)^{2k}}{(2k-1)!}\int_0^\infty \frac{t^{2k-1}}{e^{2\pi t}-1}\,dt
\end{aligned}$$

(3) But, we recognize on the right the integral giving rise to the even Bernoulli numbers, with one of the many definitions of these numbers being as follows:

$$B_{2k} = 4k(-1)^{k+1}\int_0^\infty \frac{t^{2k-1}}{e^{2\pi t}-1}\,dt$$

Thus, we can finish our computation of the values $\zeta(2k)$ as follows:

$$\begin{aligned}
\zeta(2k) &= \frac{(2\pi)^{2k}}{(2k-1)!}\cdot(-1)^{k+1}\frac{B_{2k}}{4k} \\
&= (-1)^{k+1}\frac{(2\pi)^{2k}B_{2k}}{2\cdot(2k)!}
\end{aligned}$$

(4) Regarding now the Bernoulli numbers, there is a long story here. At the beginning, we have the following formula of Bernoulli, standing as a definition for them:

$$\sum_{k=0}^{n-1} k^m = \frac{1}{m+1}\sum_{k=0}^{m} B_k n^{m+1-k}$$

This leads to the following recurrence relation, which computes them:

$$B_m = -\frac{1}{m+1}\sum_{k=0}^{m-1}\binom{m+1}{k}B_k$$

In practice, we can see that the odd Bernoulli numbers all vanish, except for the first one, $B_1 = -1/2$, and that the even Bernoulli numbers are as follows:

$$\frac{1}{6}\ ,\quad -\frac{1}{30}\ ,\quad \frac{1}{42}\ ,\quad -\frac{1}{30}\ ,\quad \frac{5}{66}\ ,\quad -\frac{691}{2730}\ ,\quad \frac{7}{6}\ ,\quad \ldots$$

(5) For analytic purposes, the Bernoulli numbers are best viewed as follows, with this coming from the fact that the coefficients satisfy the above recurrence relation:

$$\begin{aligned}\frac{x}{e^x-1} &= \sum_{n=0}^\infty B_n\frac{x^n}{n!}\\ &= 1-\frac{1}{2}x+\frac{1}{6}\cdot\frac{x^2}{2!}-\frac{1}{30}\cdot\frac{x^4}{4!}+\frac{1}{42}\cdot\frac{x^6}{6!}-\frac{1}{30}\cdot\frac{x^8}{8!}+\ldots\end{aligned}$$

Observe that all this is related as well to the hyperbolic functions, via:

$$\frac{x}{2}\left(\coth\frac{x}{2}-1\right)=\frac{x}{e^x-1}=\sum_{n=0}^\infty B_n\frac{x^n}{n!}$$

The point now is that, in relation with our zeta business, the above analytic formulae give, after some calculus, the formula that we used in (3), namely:

$$B_{2k}=4k(-1)^{k+1}\int_0^\infty\frac{t^{2k-1}}{e^{2\pi t}-1}\,dt$$

(6) Finally, no discussion about the Bernoulli numbers would be complete without mentioning the Euler-Maclaurin formula, involving them, which is as follows:

$$\begin{aligned}\sum_{k=0}^{n-1}f(x) &\simeq \int_0^n f(x)dx-\frac{1}{2}(f(n)-f(0))\\ &+\frac{1}{6}\cdot\frac{f'(n)-f'(0)}{2!}-\frac{1}{30}\cdot\frac{f^{(3)}(n)-f^{(3)}(0)}{4!}\\ &+\frac{1}{42}\cdot\frac{f^{(5)}(n)-f^{(5)}(0)}{6!}-\frac{1}{30}\cdot\frac{f^{(7)}(n)-f^{(7)}(0)}{8!}+\ldots\end{aligned}$$

(7) And there is more coming from the complex extension of the zeta function, by analytic continuation, that we will discuss later. An an announcement here, the values of zeta at the negative integers $0,-1,-2,-3,\ldots$ will not be ∞, but rather given by:

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

Alternatively, we have the following formula for the Bernoulli numbers:

$$B_n=(-1)^{n-1}n\zeta(1-n)$$

(8) In any case, we are led to the various conclusions in the statement, both theoretical and numeric. And exercise for you of course to learn more about the Bernoulli numbers, and beware of the freakish notations used by mathematicians there. □

As a more digest form of Theorem 12.21, let us record as well:

Theorem 12.22. *The generating function of the numbers $\zeta(2k)$ with $k\in\mathbb N$ is*

$$\sum_{k=0}^\infty\zeta(2k)x^{2k}=-\frac{\pi x}{2}\cot(\pi x)$$

and with this generalizing the formula involving Bernoulli numbers.

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

(1) A version of the recurrence formula for Bernoulli numbers is as follows:

$$B_{2n}=-\frac{1}{n+1/2}\sum_{k=1}^{n-1}\binom{2n}{2k}B_{2k}B_{2n-2k}$$

Now observe that this formula can be written in the following way:

$$\frac{B_{2n}}{(2n)!}=-\frac{1}{n+1/2}\sum_{k=1}^{n-1}\frac{B_{2k}}{(2k)!}\cdot\frac{B_{2n-2k}}{(2n-2k)!}$$

In view of Theorem 12.21, we obtain the following formula, valid at any $n>1$:

$$\zeta(2n)=\frac{1}{n+1/2}\sum_{k=1}^{n-1}\zeta(2k)\zeta(2n-2k)$$

(2) But this allows the computation of the series in the statement, by squaring that series. Indeed, consider the following modified version of that series:

$$f(x)=2\sum_{k=0}^\infty\zeta(2k)\left(\frac{x}{\pi}\right)^{2k}$$

By squaring, and using the recurrence formula for the numbers $\zeta(2n)$ found in (1), with some care at the values $n=0,1$, not covered by that formula, we obtain:

$$f^2+f+x^2=xf'$$

(3) But this is precisely the functional equation satisfied by $g(x)=-x\cot x$. Indeed, by using the well-known formula $\cot'=-\cot^2-1$, we have:

$$\begin{aligned}xg'&=x(-\cot x-x\cot' x)\\&=x(-\cot x+x\cot^2x+x)\\&=g+g^2+x^2\end{aligned}$$

(4) We conclude that we have $f=g$, which reads:

$$2\sum_{k=0}^\infty\zeta(2k)\left(\frac{x}{\pi}\right)^{2k}=-x\cot x$$

Now by replacing $x\to\pi x$, we obtain the formula in the statement. □

Regarding now the values $\zeta(2k+1)$ with $k \in \mathbb{N}$, the story here is more complicated, with the first such number being the Apéry constant, given by:

$$\zeta(3) = \sum_{n=1}^{\infty} \frac{1}{n^3}$$

There has been a lot of work on this number, by Apéry and others, and on the higher $\xi(2k+1)$ values as well. Let us record here the following result, a bit of physics flavor:

THEOREM 12.23. *We have the following formula,*

$$\zeta(s) = \int_0^1 \ldots \int_0^1 \frac{dx_1 \ldots dx_s}{1 - x_1 \ldots x_s}$$

valid for any $s \in \mathbb{N}$, $s \geq 2$.

PROOF. This follows as usual from some calculus, the idea being as follows:

(1) At $s = 2$ we have indeed the following computation, using Theorem 12.20:

$$\begin{aligned}
\int_0^1 \int_0^1 \frac{1}{1-xy}\, dxdy &= \int_0^1 \left[-\frac{\log(1-xy)}{y} \right]_0^1 dy \\
&= -\int_0^1 \frac{\log(1-y)}{y}\, dy \\
&= -\int_0^{\infty} \frac{\log(e^{-t})}{1-e^{-t}}\, e^{-t} dt \\
&= \int_0^{\infty} \frac{t}{e^t - 1}\, dt \\
&= \zeta(2)\Gamma(2) \\
&= \zeta(2)
\end{aligned}$$

In general the proof is similar, and we will leave this as an instructive exercise.

(2) Before leaving, however, let us see as well, out of mathematical curiosity, what happens at the exponent $s = 1$. Here the integral in the statement is:

$$\begin{aligned}
\int_0^1 \frac{1}{1-x}\, dx &= \left[-\log(1-x) \right]_0^1 \\
&= -\log(1-1) + \log(1-0) \\
&= \infty + 0 \\
&= \zeta(1)
\end{aligned}$$

Not a big deal, you would say, but as an interesting remark, since $\log(1-x) \simeq -x$, we are led to the conclusion that ζ, when suitably extended by analytic continuation, should have a simple pole at $s = 1$, with residue 1. We will be back to this, in a moment. □

Many other things can be said about ζ and its special values. In what concerns us, we will rather head towards the analytic left half-plane $Re(s) \leq 1$, using complex analysis. The idea will be that of "forcing" zeta to converge in the strip $0 < Re(s) < 1$, by adding signs, and then recovering zeta, or rather its analytic continuation, in this same strip, by removing the signs. This leads to the following remarkable result:

THEOREM 12.24. *We have the following formula,*

$$\zeta(s) = \frac{1}{1-2^{1-s}} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s}$$

which can stand as definition for ζ, in the strip $0 < Re(s) < 1$.

PROOF. This is something elementary, known since Dirichlet and Euler, but of key importance, and with many consequences, the idea being as follows:

(1) We follow the trick mentioned above. To start with, we can define the Dirichlet function η as being the signed version of ζ, in the following way:

$$\eta(s) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s}$$

Observe that this function converges indeed in the strip $0 < Re(s) < 1$.

(2) We must now connect ζ and η, at $Re(s) > 1$, and this can be done as follows:

$$\begin{aligned}\zeta(s) + \eta(s) &= \sum_{n=1}^{\infty} \frac{1}{n^s} + \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s} \\ &= 2 \sum_{k=1}^{\infty} \frac{1}{(2k)^s} \\ &= 2^{1-s} \sum_{k=1}^{\infty} \frac{1}{k^s} \\ &= 2^{1-s} \zeta(s)\end{aligned}$$

(3) But this gives the following formula, valid at any exponent $s \in \mathbb{C}$ satisfying $Re(s) > 1$, and which is the formula in the statement:

$$\zeta(s) = \frac{1}{1-2^{1-s}} \, \eta(s)$$

(4) In order now to conclude, we can invoke the theory of analytic continuation. Skipping some theoretical details here, and we refer for instance to Rudin [80] for all this, what we have in the statement is a formula for ζ in the whole right half-plane, $Re(s) > 0$, which is analytic, and more specifically meromorphic, with a single pole, at $s = 1$, and which coincides with the usual formula of ζ on the usual domain of definition, $Re(s) > 1$.

But, in this situation, the theory of analytic continuation tells us that we can redefine ζ all over the right half-plane, $Re(s) > 0$, by the formula in the statement, and with this extension being unique, as per the general properties of the meromorphic functions. □

With a bit more care, we have in fact the following result:

THEOREM 12.25. *We can talk about the Riemann zeta, as a meromorphic function $\zeta : \mathbb{C} \to \mathbb{C}$, with a single pole, at $s = 1$ with residue 1. At $Re(s) > 1$ we have*

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

and more generally at $Re(s) > 0$ we have the following formula:

$$\zeta(s) = \frac{1}{1 - 2^{1-s}} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s}$$

Also, the values of zeta at any s and $1 - s$ are related by the Riemann formula

$$\zeta(s) = 2^s \pi^{s-1} \sin\left(\frac{\pi s}{2}\right) \Gamma(1-s)\zeta(1-s)$$

with Γ being as usual the gamma function.

PROOF. This is something quite heavy, due to Riemann himself. □

The zeta function has trivial zeroes at $-2, -4, -6, \ldots$, and the nontrivial zeroes must lie in the closed critical strip $0 \leq Re(s) \leq 1$. The Riemann hypothesis states that the nontrivial zeroes must satisfy $Re(s) = 1/2$. With this being important, because many questions in arithmetic reformulate in terms of sums at the zeroes of zeta.

12e. Exercises

Exercises:

EXERCISE 12.26.

EXERCISE 12.27.

EXERCISE 12.28.

EXERCISE 12.29.

EXERCISE 12.30.

EXERCISE 12.31.

EXERCISE 12.32.

EXERCISE 12.33.

Bonus exercise.

Part IV

Three dimensions

If you're going to San Francisco
Be sure to wear some flowers in your hair
If you're going to San Francisco
You're gonna meet some gentle people there

CHAPTER 13

Space geometry

13a. Space geometry

Space geometry. Many things can be said here.

13b. Vector products

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

DEFINITION 13.1. *The vector product of two vectors in* $\mathbb{R}^3$ *is given by*

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

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

$$\begin{matrix} & \uparrow_{x \times y} \\ \leftarrow_x & \\ \swarrow_y & \end{matrix}$$

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

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} x_2 y_3 - x_3 y_2 \\ x_3 y_1 - x_1 y_3 \\ x_1 y_2 - x_2 y_1 \end{pmatrix}$$

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

$$J = x \times p$$

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

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

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

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

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

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

13c. Rotating bodies

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

$$v = \omega \times x$$

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

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

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

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

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

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

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

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

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

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

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

THEOREM 13.4. *Assuming as before that a 3D body rotates along an axis, the acceleration of a moving point on the body, computed by O as before, is given by*

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

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

PROOF. This comes by using twice the formulae in Theorem 13.3, as follows:

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

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

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

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

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

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

THEOREM 13.5. *The acceleration of an object m subject to a force F is given by*

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

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

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

As a basic illustration for all this, a rock dropped from 100m deviates about 1cm from its intended target, due to the formula in Theorem 13.5.

13d. Curved spacetime

Still talking basic 3D geometry, let us discuss now curved spacetime, as a continuation of our 1D study from chapter 10. We would like to solve the following question:

QUESTION 13.6. *What is the Einstein speed summation formula in 3D? And, what does this tell us about our usual spacetime $\mathbb{R}^4$, how does this exactly get curved?*

And we will stop with the philosophy here, we have a very good and concrete question now, and time to get to work. Let us attempt to construct $u +_e v$ in arbitrary dimensions, just by using our common sense and intuition. When the vectors $u, v \in \mathbb{R}^N$ are proportional, we are basically in 1D, and so our addition formula must satisfy:

$$u \sim v \implies u +_e v = \frac{u+v}{1+ < u, v >}$$

However, the formula on the right will not work as such in general, for arbitrary speeds $u, v \in \mathbb{R}^N$, and this because we have, as main requirement for our operation, in analogy with the $1 +_e v = 1$ formula from 1D, the following condition:

$$||u|| = 1 \implies u +_e v = u$$

Equivalently, in analogy with $u +_e 1 = 1$ from 1D, we would like to have:

$$||v|| = 1 \implies u +_e v = v$$

Summarizing, our $u \sim v$ formula above is not bad, as a start, but we must add a correction term to it, for the above requirements to be satisfied, and of course with the correction term vanishing when $u \sim v$. So, we are led to a math puzzle:

PUZZLE 13.7. *What vanishes when $u \sim v$, and then how to correctly define*

$$u +_e v = \frac{u + v + \gamma_{uv}}{1+ < u, v >}$$

as for the correction term γ_{uv} to vanish when $u \sim v$?

But the solution to the first question is well-known in 3D. Indeed, here we can use the vector product $u \times v$, that we met before, which notoriously satisfies:

$$u \sim v \implies u \times v = 0$$

Thus, our correction term γ_{uv} must be something containing $w = u \times v$, which vanishes when this vector w vanishes, and in addition arranged such that $||u|| = 1$ produces a simplification, with $u +_e v = u$ as end result, and with $||v|| = 1$ producing a simplification too, with $u +_e v = v$ as end result. Thus, our vector calculus puzzle becomes:

PUZZLE 13.8. *How to correctly define the Einstein summation in 3 dimensions,*

$$u +_e v = \frac{u + v + \gamma_{uvw}}{1 + < u, v >}$$

with $w = u \times v$, in such a way as for the correction term γ_{uvw} to satisfy

$$w = 0 \implies \gamma_{uvw} = 0$$

and also such that $||u|| = 1 \implies u +_e v = u$, and $||v|| \implies u +_e v = v$?

As an obvious task, we must "transport" the vector w to the plane spanned by u, v. But this is simplest done by taking the vector product with any vector in this plane. So, time to update our mathematical puzzle, in the following way:

PUZZLE 13.9. *How to define the Einstein summation in 3 dimensions,*

$$u +_e v = \frac{u + v + \gamma_{uvw}}{1 + < u, v >}$$

with the correction term being of the following form, with $w = u \times v$, and $\alpha, \beta \in \mathbb{R}$,

$$\gamma_{uvw} = (\alpha u + \beta v) \times w$$

in such a way as to have $||u|| = 1 \implies u +_e v = u$, and $||v|| \implies u +_e v = v$?

In order to investigate what happens when $||u|| = 1$ or $||v|| = 1$, we must compute the vector products $u \times w$ and $v \times w$. So, pausing now our study for consulting the vector calculus database, and then coming back, here is the formula that we need:

$$u \times (u \times v) = < u, v > u - < u, u > v$$

With this formula in hand, we can now compute the correction term, with the result here, that we will need several times in what comes next, being as follows:

PROPOSITION 13.10. *The correction term $\gamma_{uvw} = (\alpha u + \beta v) \times w$ is given by*

$$\gamma_{uvw} = (\alpha < u, v > + \beta < v, v >)u - (\alpha < u, u > + \beta < u, v >)v$$

for any values of the scalars $\alpha, \beta \in \mathbb{R}$.

Proof. According to our vector product formula above, we have:

$$\begin{aligned}\gamma_{uvw} &= (\alpha u+\beta v)\times w\\ &= \alpha(<u,v>u-<u,u>v)+\beta(<v,v>u-<u,v>v)\\ &= (\alpha<u,v>+\beta<v,v>)u-(\alpha<u,u>+\beta<u,v>)v\end{aligned}$$

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

Time now to get into the real thing, see what happens when $||u||=1$ and $||v||=1$, if we can get indeed $u+_e v=u$ and $u+_e v=v$. It is convenient here to do some reverse engineering. Regarding the first desired formula, namely $u+_e v=u$, we have:

$$\begin{aligned}u+_e v=u &\iff u+v+\gamma_{uvw}=(1+<u,v>)u\\ &\iff \gamma_{uvw}=<u,v>u-v\\ &\iff \alpha=1,\ \beta=0,\ ||u||=1\end{aligned}$$

Thus, with the parameter choice $\alpha=1,\beta=0$, we will have, as desired:

$$||u||=1\implies u+_e v=u$$

In what regards now the second desired formula, namely $u+_e v=v$, here the computation is almost identical, save for a sign switch, which after some thinking comes from our choice $w=u\times v$ instead of $w=v\times u$, clearly favoring u, as follows:

$$\begin{aligned}u+_e v=v &\iff u+v+\gamma_{uvw}=(1+<u,v>)v\\ &\iff \gamma_{uvw}=-u+<u,v>v\\ &\iff \alpha=0,\ \beta=-1,\ ||v||=1\end{aligned}$$

Thus, with the parameter choice $\alpha=0,\beta=-1$, we will have, as desired:

$$||v||=1\implies u+_e v=v$$

All this is mixed news, because we managed to solve both our problems, at $||u||=1$ and at $||v||=1$, but our solutions are different. So, time to breathe, decide that we did enough interesting work for the day, and formulate our conclusion as follows:

Proposition 13.11. *When defining the Einstein speed summation in* 3D *as*

$$u+_e v=\frac{u+v+u\times(u\times v)}{1+<u,v>}$$

in $c=1$ units, the following happen:

1. *When $u\sim v$, we recover the previous* 1D *formula.*
2. *When $||u||=1$, speed of light, we have $u+_e v=u$.*
3. *However, $||v||=1$ does not imply $u+_e v=v$.*
4. *Also, the formula $u+_e v=v+_e u$ fails.*

PROOF. Here (1) and (2) follow from the above discussion, with the following choice for the correction term, by favoring the $||u||=1$ problem over the $||v||=1$ one:

$$\gamma_{uvw} = u \times w$$

In fact, with this choice made, the computation is very simple, as follows:

$$\begin{aligned} ||u||=1 \quad &\implies \quad \gamma_{uvw} =< u,v > u - v \\ &\implies \quad u+v+\gamma_{uvw} = u+ < u,v > u \\ &\implies \quad \frac{u+v+\gamma_{uvw}}{1+ < u,v >} = u \end{aligned}$$

As for (3) and (4), these are also clear from the above discussion, coming from the obvious lack of symmetry of our summation formula. □

Looking now at Proposition 13.11 from an abstract, mathematical perspective, there are still many things missing from there, which can be summarized as follows:

QUESTION 13.12. *Can we fine-tune the Einstein speed summation in* 3D *into*

$$u +_e v = \frac{u+v+\lambda \cdot u \times (u \times v)}{1+ < u,v >}$$

with $\lambda \in \mathbb{R}$, chosen such that $||u||=1 \implies \lambda = 1$, as to have:

(1) $||u||, ||v|| < 1 \implies ||u +_e v|| < 1$.
(2) $||v|| = 1 \implies ||u +_e v|| = 1$.

All this is quite tricky, and deserves some explanations. First, if we add a scalar $\lambda \in \mathbb{R}$ into our formula, as above, we will still have, exactly as before:

$$u \sim v \implies u +_e v = \frac{1+uv}{1+ < u,v >}$$

On the other hand, we already know from our previous computations, those preceding Proposition 13.11, that if we ask for $\lambda \in \mathbb{R}$ to be a plain constant, not depending on u, v, then $\lambda = 1$ is the only good choice, making the following formula happen:

$$||u|| = 1 \implies u +_e v = u$$

But, and here comes our point, $\lambda = 1$ is not an ideal choice either, because it would be nice to have the properties (1,2) in the statement, and these properties have no reason to be valid for $\lambda = 1$, as you can check for instance by yourself by doing some computations. Thus, the solution to our problem most likely involves a scalar $\lambda \in \mathbb{R}$ depending on u, v, and satisfying the following condition, as to still have $||u|| = 1 \implies u +_e v = u$:

$$||u|| = 1 \implies \lambda = 1$$

Obviously, as simplest answer, λ must be some well-chosen function of $||u||$, or rather of $||u||^2$, because it is always better to use square norms, when possible. But then, with this idea in mind, after a few computations we are led to the following solution:

$$\lambda = \frac{1}{1+\sqrt{1-||u||^2}}$$

Summarizing, final correction done, and we are led to the following theorem:

THEOREM 13.13. *When defining the Einstein speed summation in* 3D *as*

$$u +_e v = \frac{1}{1+ < u, v >}\left(u + v + \frac{u \times (u \times v)}{1+\sqrt{1-||u||^2}}\right)$$

in $c = 1$ *units, the following happen:*

(1) *When* $u \sim v$, *we recover the previous* 1D *formula.*
(2) *We have* $||u||, ||v|| < 1 \implies ||u +_e v|| < 1$.
(3) *When* $||u|| = 1$, *we have* $u +_e v = u$.
(4) *When* $||v|| = 1$, *we have* $||u +_e v|| = 1$.
(5) *However,* $||v|| = 1$ *does not imply* $u +_e v = v$.
(6) *Also, the formula* $u +_e v = v +_e u$ *fails.*

PROOF. This follows from the above discussion, and a few more computations. □

13e. Exercises

Exercises:

EXERCISE 13.14.

EXERCISE 13.15.

EXERCISE 13.16.

EXERCISE 13.17.

EXERCISE 13.18.

EXERCISE 13.19.

EXERCISE 13.20.

EXERCISE 13.21.

Bonus exercise.

CHAPTER 14

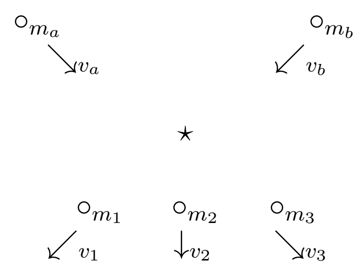
Solid angles

14a. Solid angles

Solid angles. Many things can be said here.

14b. Particle physics

We need to talk 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 14.1. *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.

14c. Decay, scattering

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 14.2. *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:

$$\int_0^\infty N_t dt=\int_0^\infty N_0e^{-\lambda t}dt=\frac{N_0}{\lambda}$$

Thus, the density of the probability decay function is given by:

$$f(t)=\frac{\lambda}{N_0}\cdot N_0e^{-\lambda t}=\lambda e^{-\lambda t}$$

We can now compute the mean lifetime, by integrating by parts, as follows:

$$\begin{aligned}\tau &= <t>\\ &= \int_0^\infty tf(t)dt\\ &= \int_0^\infty \lambda te^{-\lambda t}dt\\ &= \int_0^\infty t(-e^{-\lambda t})'dt\\ &= \int_0^\infty e^{-\lambda t}dt\\ &= \frac{1}{\lambda}\end{aligned}$$

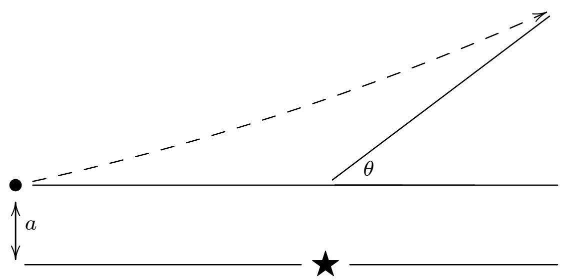
(3) Finally, regarding the half-life, this is by definition the time $t_{1/2}$ required for the decaying quantity to fall to one-half of its initial value. Mathematically, this means:

$$N_t = 2^{-\frac{t}{t_{1/2}}} N_0$$

Now by comparing with $N_t = e^{-\lambda t} N_0$, this gives $t_{1/2} = (\log 2)/\lambda$, as stated. $\square$

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 14.3. *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.

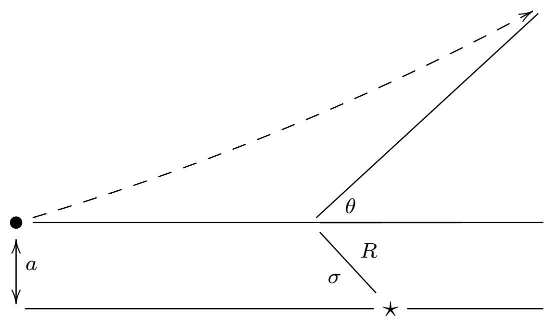
But probably too much talking, let us do a computation. We have here:

Proposition 14.4. *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 14.3, as follows:

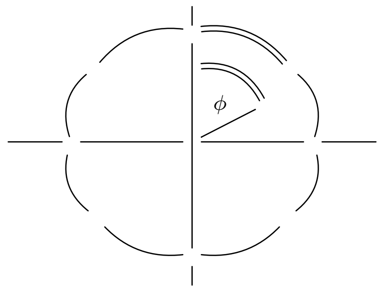
DEFINITION 14.5. *In the general context of scattering, we can:*

(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 14.6. *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 14.5, 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 14.7. *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 14.4 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. □

14d. Golden Rule

Golden Rule.

14e. Exercises

Exercises:

EXERCISE 14.8.

EXERCISE 14.9.

EXERCISE 14.10.

EXERCISE 14.11.

EXERCISE 14.12.

EXERCISE 14.13.

EXERCISE 14.14.

EXERCISE 14.15.

Bonus exercise.

CHAPTER 15

Field lines

15a. Electrostatics

Time for electricity. Let us start with something very basic, namely:

FACT 15.1. *Each piece of matter has a charge $q \in \mathbb{R}$, which is normally neutral, $q = 0$, but that we can make positive or negative, by using various methods. We say that responsible for the charge is the amount of electrons present, as follows:*

(1) *When the matter lacks electrons, the charge is positive, $q > 0$.*
(2) *When there are more electrons than needed, the charge is negative, $q < 0$.*

As our first result, due to Coulomb, and that will come as a physics fact instead of a mathematics theorem, because, well, I must admit that what we have in Fact 15.1 is indeed more than borderline, as axiomatics for a theory, we have:

FACT 15.2 (Coulomb law). *Any pair of charges $q_1, q_2 \in \mathbb{R}$ is subject to a force as follows, which is attractive if $q_1 q_2 < 0$ and repulsive if $q_1 q_2 > 0$,*

$$||F|| = K \cdot \frac{|q_1 q_2|}{d^2}$$

where $d > 0$ is the distance between the charges, and $K > 0$ is a certain constant.

Observe the amazing similarity with the Newton law for gravity. However, as we will discover soon, passed a few simple facts, things will be far more complicated here.

As in the gravity case, the force F appearing above is understood to be parallel to the vector $x_2 - x_1 \in \mathbb{R}^3$ joining as $x_1 \to x_2$ the locations $x_1, x_2 \in \mathbb{R}^3$ of our charges, and by taking into account the attraction/repulsion rules above, we have:

PROPOSITION 15.3. *The Coulomb force of q_1 at x_1 acting on q_2 at x_2 is*

$$F = K \cdot \frac{q_1 q_2 (x_2 - x_1)}{||x_2 - x_1||^3}$$

with $K > 0$ being the Coulomb constant, as above.

PROOF. We have indeed the following computation:

$$\begin{aligned} F &= sgn(q_1q_2)\cdot||F||\cdot\frac{x_2-x_1}{||x_2-x_1||} \\ &= sgn(q_1q_2)\cdot K\cdot\frac{|q_1q_2|}{||x_2-x_1||^2}\cdot\frac{x_2-x_1}{||x_2-x_1||} \\ &= K\cdot\frac{q_1q_2(x_2-x_1)}{||x_2-x_1||^3} \end{aligned}$$

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

In analogy with the usual study of gravity, let us start with:

DEFINITION 15.4. *Given charges $q_1,\ldots,q_k\in\mathbb{R}$ located at positions $x_1,\ldots,x_k\in\mathbb{R}^3$, we define their electric field to be the vector function*

$$E(x)=K\sum_i\frac{q_i(x-x_i)}{||x-x_i||^3}$$

so that their force applied to a charge $Q\in\mathbb{R}$ positioned at $x\in\mathbb{R}^3$ is given by $F=QE$.

More generally, we will be interested in electric fields of various non-discrete configurations of charges, such as charged curves, surfaces and solid bodies. Indeed, things like wires or metal sheets or solid bodies coming in all sorts of shapes, tailored for their purpose, play a key role, so this extension is essential. So, let us go ahead with:

DEFINITION 15.5. *The electric field of a charge configuration $L\subset\mathbb{R}^3$, with charge density function $\rho:L\to\mathbb{R}$, is the vector function*

$$E(x)=K\int_L\frac{\rho(z)(x-z)}{||x-z||^3}\,dz$$

so that the force of L applied to a charge Q positioned at x is given by $F=QE$.

And good news, with this viewpoint, we are now into advanced physics.

15b. Field lines

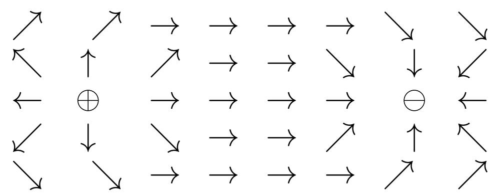
With the above definitions in hand, it is most convenient now to forget about the charges, and focus on the study of the corresponding electric fields E.

These fields are by definition vector functions $E:\mathbb{R}^3\to\mathbb{R}^3$, with the convention that they take $\pm\infty$ values at the places where the charges are located, and intuitively, are best represented by their field lines, which are constructed as follows:

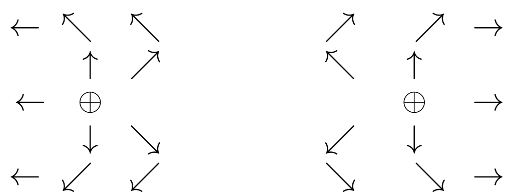
DEFINITION 15.6. *The field lines of an electric field $E:\mathbb{R}^3\to\mathbb{R}^3$ are the oriented curves $\gamma\subset\mathbb{R}^3$ pointing at every point $x\in\mathbb{R}^3$ at the direction of the field, $E(x)\in\mathbb{R}^3$.*

As a basic example here, for one charge the field lines are the half-lines emanating from its position, oriented according to the sign of the charge:

For two charges now, if these are of opposite signs, $+$ and $-$, you get a picture that you are very familiar with, namely that of the field lines of a bar magnet:



If the charges are $+,+$ or $-,-$, you get something of similar type, but repulsive this time, with the field lines emanating from the charges being no longer shared:



These pictures, and notably the last one, with $+,+$ charges, are quite interesting, because the repulsion situation does not appear in the context of gravity. Thus, we can only expect our geometry here to be far more complicated than that of gravity.

The field lines obviously do not encapsulate the whole information about the field, with the direction of each vector $E(x) \in \mathbb{R}^3$ being there, but with the magnitude $||E(x)|| \geq 0$ of this vector missing. However, say when drawing, when picking up uniformly radially spaced field lines around each charge, and with the number of these lines proportional to the magnitude of the charge, and then completing the picture, the density of the field lines around each point $x \in \mathbb{R}$ will give you then the magnitude $||E(x)|| \geq 0$ of the field there, up to a scalar. Let us summarize these observations as follows:

PROPOSITION 15.7. *Given an electric field $E : \mathbb{R}^3 \to \mathbb{R}^3$, the knowledge of its field lines is the same as the knowledge of the composition*

$$nE : \mathbb{R}^3 \to \mathbb{R}^3 \to S$$

where $S \subset \mathbb{R}^3$ is the unit sphere, and $n : \mathbb{R}^3 \to S$ is the rescaling map, namely:

$$n(x) = \frac{x}{||x||}$$

However, in practice, when the field lines are accurately drawn, the density of the field lines gives you the magnitude of the field, up to a scalar.

PROOF. The first assertion is clear from definitions, with our usual convention that the electric field and its problematics take place outside the locations of the charges. As for the last assertion, this basically follows from the above discussion. □

15c. Gauss law

Let us introduce now a key definition, as follows:

DEFINITION 15.8. *The flux of an electric field $E:\mathbb R^3\to\mathbb R^3$ through a surface $S\subset\mathbb R^3$, assumed to be oriented, is the quantity*

$$\Phi_E(S)=\int_S<E(x),n(x)>dx$$

with $n(x)$ being unit vectors orthogonal to S, following the orientation of S. Intuitively, the flux measures the signed number of field lines crossing S.

Here by orientation of S we mean precisely the choice of unit vectors $n(x)$ as above, orthogonal to S, which must vary continuously with x. For instance a sphere has two possible orientations, one with all these vectors $n(x)$ pointing inside, and one with all these vectors $n(x)$ pointing outside. More generally, any surface has locally two possible orientations, so if it is connected, it has two possible orientations. In what follows the convention is that the closed surfaces are oriented with each $n(x)$ pointing outside.

As a first illustration, let us do a basic computation, as follows:

PROPOSITION 15.9. *For a point charge $q\in\mathbb R$ at the center of a sphere S,*

$$\Phi_E(S)=\frac{q}{\varepsilon_0}$$

where the constant is $\varepsilon_0=1/(4\pi K)$, independently of the radius of S.

PROOF. Assuming that S has radius r, we have the following computation:

$$\begin{aligned}\Phi_E(S)&=\int_S<E(x),n(x)>dx\\&=\int_S\left\langle\frac{Kqx}{r^3},\frac{x}{r}\right\rangle dx\\&=\int_S\frac{Kq}{r^2}\,dx\\&=\frac{Kq}{r^2}\times4\pi r^2\\&=4\pi Kq\end{aligned}$$

Thus with $\varepsilon_0=1/(4\pi K)$ as above, we obtain the result. □

More generally now, we have the following result:

THEOREM 15.10. *The flux of a field E through a sphere S is given by*

$$\Phi_E(S)=\frac{Q_{enc}}{\varepsilon_0}$$

where Q_{enc} is the total charge enclosed by S, and $\varepsilon_0=1/(4\pi K)$.

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

(1) Before jumping into computations, let us do some manipulations. First, by discretizing the problem, we can assume that we are dealing with a system of point charges. Moreover, by additivity, we can assume that we are dealing with a single charge. And if we denote by $q\in\mathbb{R}$ this charge, located at $v\in\mathbb{R}^3$, we want to prove that we have the following formula, where $B\subset\mathbb{R}^3$ denotes the ball enclosed by S:

$$\Phi_E(S)=\frac{q}{\varepsilon_0}\,\delta_{v\in B}$$

(2) By linearity we can assume that we are dealing with the unit sphere S. Moreover, by rotating we can assume that our charge q lies on the Ox axis, that is, that we have $v=(r,0,0)$ with $r\geq 0$, $r\neq 1$. The formula that we want to prove becomes:

$$\Phi_E(S)=\frac{q}{\varepsilon_0}\,\delta_{r<1}$$

(3) Let us start now the computation. With $u=(x,y,z)$, we have:

$$\begin{aligned}
\Phi_E(S) &= \int_S <E(u),u>du\\
&= \int_S\left\langle\frac{Kq(u-v)}{||u-v||^3},u\right\rangle du\\
&= Kq\int_S\frac{<u-v,u>}{||u-v||^3}\,du\\
&= Kq\int_S\frac{1-<v,u>}{||u-v||^3}\,du\\
&= Kq\int_S\frac{1-rx}{(1-2xr+r^2)^{3/2}}\,du
\end{aligned}$$

(4) In order to compute the above integral, we will use spherical coordinates for the unit sphere S, which are as follows, with $s\in[0,\pi]$ and $t\in[0,2\pi]$:

$$\begin{cases}
x=\cos s\\
y=\sin s\cos t\\
z=\sin s\sin t
\end{cases}$$

We recall that the corresponding Jacobian, computed before, is given by:

$$J=\sin s$$

(5) With the above change of coordinates, our integral from (3) becomes:

$$\begin{aligned}\Phi_E(S) &= Kq\int_S \frac{1-rx}{(1-2xr+r^2)^{3/2}}\,du\\ &= Kq\int_0^{2\pi}\int_0^{\pi}\frac{1-r\cos s}{(1-2r\cos s+r^2)^{3/2}}\cdot\sin s\,ds\,dt\\ &= 2\pi Kq\int_0^{\pi}\frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}}\,ds\\ &= \frac{q}{2\varepsilon_0}\int_0^{\pi}\frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}}\,ds\end{aligned}$$

(6) The point now is that the integral on the right can be computed with the change of variables $x=\cos s$. Indeed, we have $dx=-\sin s\,ds$, and we obtain:

$$\begin{aligned}\int_0^{\pi}\frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}}\,ds &= \int_{-1}^{1}\frac{1-rx}{(1-2rx+r^2)^{3/2}}\,dx\\ &= \left[\frac{x-r}{\sqrt{1-2rx+r^2}}\right]_{-1}^{1}\\ &= \frac{1-r}{\sqrt{1-2r+r^2}}-\frac{-1-r}{\sqrt{1+2r+r^2}}\\ &= \frac{1-r}{|1-r|}+1\\ &= 2\delta_{r<1}\end{aligned}$$

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

More generally now, we have the following key result, due to Gauss:

THEOREM 15.11 (Gauss law). *The flux of a field E through a surface S is given by*

$$\Phi_E(S)=\frac{Q_{enc}}{\varepsilon_0}$$

where Q_{enc} is the total charge enclosed by S, and $\varepsilon_0=1/(4\pi K)$.

PROOF. This basically follows from Theorem 15.10, or even from Proposition 15.9, by adding to the results there a number of new ingredients, as follows:

(1) Our first claim is that given a closed surface S, with no charges inside, the flux through it of any choice of external charges vanishes:

$$\Phi_E(S)=0$$

This claim is indeed supported by the intuitive interpretation of the flux, as corresponding to the signed number of field lines crossing S. Indeed, any field line entering as $+$ must exit somewhere as $-$, and vice versa, so when summing we get 0.

(2) In practice now, in order to prove this rigorously, there are several ways. A standard argument, which is quite elementary, is the one used by Feynman in [34], based on the fact that, due to $F \sim 1/d^2$, local deformations of S will leave invariant the flux, and so in the end we are left with a rotationally invariant surface, where the result is clear.

(3) The point now is that, with this and Proposition 15.9 in hand, we can finish by using a standard math trick. Let us assume indeed, by discretizing, that our system of charges is discrete, consisting of enclosed charges $q_1, \ldots, q_k \in \mathbb{R}$, and an exterior total charge Q_{ext}. We can surround each of $q_1, \ldots, q_k$ by small disjoint spheres $U_1, \ldots, U_k$, chosen such that their interiors do not touch S, and we have:

$$\begin{aligned}
\Phi_E(S) &= \Phi_E(S - \cup U_i) + \Phi_E(\cup U_i)\\
&= 0 + \Phi_E(\cup U_i)\\
&= \sum_i \Phi_E(U_i)\\
&= \sum_i \frac{q_i}{\varepsilon_0}\\
&= \frac{Q_{enc}}{\varepsilon_0}
\end{aligned}$$

(4) To be more precise, in the above the union $\cup U_i$ is a usual disjoint union, and the flux is of course additive over components. As for the difference $S - \cup U_i$, this is by definition the disjoint union of S with the disjoint union $\cup(-U_i)$, with each $-U_i$ standing for U_i with orientation reversed, and since this difference has no enclosed charges, the flux through it vanishes by (2). Finally, the end makes use of Proposition 15.9. □

We have the following point of view on the Gauss formula, more conceptual:

THEOREM 15.12 (Gauss). *Given an electric potential E, its divergence is given by*

$$< \nabla, E >= \frac{\rho}{\varepsilon_0}$$

where ρ denotes as usual the charge distribution. Also, we have

$$\nabla \times E = 0$$

meaning that the curl of E vanishes.

PROOF. The first formula, called Gauss law in differential form, is clear. Regarding now the curl, by discretizing and linearity we can assume that we are dealing with a single charge q, positioned at 0. We have, by using spherical coordinates r,s,t:

$$\begin{aligned}\int_a^b < E(x), dx > &= \int_a^b \left\langle \frac{Kqx}{||x||^3}, dx \right\rangle \\ &= \int_a^b \left\langle \frac{Kq}{r^2} \cdot \frac{x}{||x||}, dx \right\rangle \\ &= \int_a^b \frac{Kq}{r^2}\, dr \\ &= \left[-\frac{Kq}{r} \right]_a^b \\ &= Kq \left(\frac{1}{r_a} - \frac{1}{r_b} \right)\end{aligned}$$

In particular the integral of E over any closed loop vanishes, and by using now the Stokes theorem, we conclude that the curl of E vanishes, as stated. □

15d. Algebraic curves

In order to further advance, let us go back to the various plane curves discussed in chapter 8. Quite remarkably, most of that curves are sinusoidal spirals, in the following sense, and with actually the term "sinusoidal spiral" being a bit unfortunate:

THEOREM 15.13. *The sinusoidal spirals, which are as follows,*

$$r^n = a^n \cos n\theta$$

with $a \neq 0$ and $n \in \mathbb{Q} - \{0\}$, include the following curves:

(1) *$n = -1$ line.*
(2) *$n = 1$ circle, $n = -1/2$ parabola, $n = -2$ hyperbola.*
(3) *$n = -3$ Humbert cubic, $n = -1/3$ Tschirnhausen curve.*
(4) *$n = 1/2$ cardioid, $n = 2$ Bernoulli lemniscate.*
(5) *$n = 3$ Kiepert trefoil, $n = 1/3$ Cayley sextic.*

PROOF. We first have to prove that the sinusoidal spirals are indeed algebraic curves. But this is best done by using the complex coordinate $z = re^{i\theta}$, as follows:

$$\begin{aligned}r^n = a^n \cos n\theta \quad &\Longleftrightarrow \quad r^n \cos n\theta = \left(\frac{r^2}{a} \right)^n \\ &\Longleftrightarrow \quad z^n + \bar{z}^n = 2 \left(\frac{z\bar{z}}{a} \right)^n \\ &\Longleftrightarrow \quad (x+iy)^n + (x-iy)^n = 2 \left(\frac{x^2+y^2}{a} \right)^n\end{aligned}$$

As a first observation now, in the case $n\in\mathbb N$ we can simply use the binomial formula, and we get an algebraic equation of degree $2n$, as follows:

$$\sum_{k=0}^{[n/2]}(-1)^k\binom{n}{2k}x^{n-2k}y^{2k}=\left(\frac{x^2+y^2}{a}\right)^n$$

In general, things are a bit more complicated, as shown for instance by our computation for the Cayley sextic. However, the same idea as there applies, and we are led in this way to the equation of an algebraic curve, as claimed. Regarding now the examples:

(1) At $n=-1$ the equation is as follows, producing a line:

$$r\cos\theta=a\iff x=a$$

(2) At $n=1$ the equation is as follows, producing a circle:

$$r=a\cos\theta\iff r^2=ax\iff x^2+y^2=ax$$

(3) At $n=-1/2$ the equation is as follows, producing a parabola:

$$a=r\cos^2(\theta/2)\iff r+x=2a\iff y^2=4a(a-x)$$

(4) At $n=-2$ the equation is as follows, producing a hyperbola:

$$a^2=r\cos^2 2\theta\iff a^2=2x^2-r^2\iff(x+y)(x-y)=a^2$$

(5) At $n=-3$ the equation is as follows, producing a curve with 3 components, which looks like some sort of "trivalent hyperbola", called Humbert cubic:

$$r^3\cos 3\theta=a^3\iff z^3+\bar z^3=2a^3\iff x^3-3xy^2=a^3$$

(6) As for the other curves, this follows from our various formulae above. □

Let us study now more in detail the sinusoidal spirals. We first have:

PROPOSITION 15.14. *The sinusoidal spirals, which with $z=x+iy$ are*

$$z^n+\bar z^n=2\left(\frac{z\bar z}{a}\right)^n$$

with $a\neq 0$ and $n\in\mathbb Q-\{0\}$, are as follows:

(1) *With $n=-m$, $m\in\mathbb N$, the equation is $z^m+\bar z^m=2a^m$, degree m.*
(2) *With $n=m$, $m\in\mathbb N$, the equation is $z^m+\bar z^m=2(z\bar z/a)^m$, degree $2m$.*
(3) *With $n=-1/m$, $m\in\mathbb N$, the equation is $(z^{1/m}+\bar z^{1/m})^m=2^ma$.*
(4) *With $n=1/m$, $m\in\mathbb N$, the equation is $(z^{1/m}+\bar z^{1/m})^m=2^mz\bar z/a$.*

PROOF. This is something self-explanatory, the details being as follows:

(1) With $n=-m$ and $m\in\mathbb N$ as in the statement, the equation is, as claimed:

$$z^{-m}+\bar z^{-m}=2\left(\frac{z\bar z}{a}\right)^{-m}\iff z^m+\bar z^m=2a^m$$

(2) This is an empty statement, just a matter of using the new variable $m=n$.

(3) With $n=-1/m$ and $m\in\mathbb N$ as in the statement, the equation is, as claimed:

$$z^{-1/m}+\bar{z}^{-1/m}=2\left(\frac{z\bar{z}}{a}\right)^{-1/m}\iff z^{1/m}+\bar{z}^{1/m}=2a^{1/m}$$
$$\iff (z^{1/m}+\bar{z}^{1/m})^m=2^ma$$

(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 15.15. *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 15.16. *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:

$$|z^n - c^n| = c^n \iff \prod_{k=1}^{n} |z - cw^k| = c^n \iff \sqrt[n]{\prod_{k=1}^{n} |z - cw^k|} = c$$

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 15.17. *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.

15e. Exercises

Exercises:

Exercise 15.18.

Exercise 15.19.

Exercise 15.20.

Exercise 15.21.

Exercise 15.22.

Exercise 15.23.

Exercise 15.24.

Exercise 15.25.

Bonus exercise.

CHAPTER 16

Angle of attack

16a. Angle of incidence

Angle of incidence.

16b. Angle of refraction

Angle of refraction.

16c. Angle of bombing

Angle of bombing.

16d. Angle of attack

Angle of attack.

16e. Exercises

Congratulations for having read this book, and no exercises for this final chapter.

Bibliography

[1] A.A. Abrikosov, Fundamentals of the theory of metals, Dover (1988).

[2] V.I. Arnold, Ordinary differential equations, Springer (1973).

[3] V.I. Arnold, Lectures on partial differential equations, Springer (1997).

[4] V.I. Arnold, Catastrophe theory, Springer (1984).

[5] N.W. Ashcroft and N.D. Mermin, Solid state physics, Saunders College Publ. (1976).

[6] T. Banica, Calculus and applications (2024).

[7] T. Banica, Linear algebra and group theory (2024).

[8] T. Banica, Introduction to modern physics (2024).

[9] G.K. Batchelor, An introduction to fluid dynamics, Cambridge Univ. Press (1967).

[10] M.J. Benton, Vertebrate paleontology, Wiley (1990).

[11] M.J. Benton and D.A.T. Harper, Introduction to paleobiology and the fossil record, Wiley (2009).

[12] S.J. Blundell and K.M. Blundell, Concepts in thermal physics, Oxford Univ. Press (2006).

[13] B. Bollobás, Modern graph theory, Springer (1998).

[14] S.M. Carroll, Spacetime and geometry, Cambridge Univ. Press (2004).

[15] P.M. Chaikin and T.C. Lubensky, Principles of condensed matter physics, Cambridge Univ. Press (1995).

[16] A.R. Choudhuri, Astrophysics for physicists, Cambridge Univ. Press (2012).

[17] J. Clayden, S. Warren and N. Greeves, Organic chemistry, Oxford Univ. Press (2012).

[18] D.D. Clayton, Principles of stellar evolution and nucleosynthesis, Univ. of Chicago Press (1968).

[19] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).

[20] A. Cottrell, An introduction to metallurgy, CRC Press (1997).

[21] C. Darwin, On the origin of species (1859).

[22] P.A. Davidson, Introduction to magnetohydrodynamics, Cambridge Univ. Press (2001).

[23] P.A.M. Dirac, Principles of quantum mechanics, Oxford Univ. Press (1930).

[24] S. Dodelson, Modern cosmology, Academic Press (2003).

[25] S.T. Dougherty, Combinatorics and finite geometry, Springer (2020).

[26] M. Dresher, The mathematics of games of strategy, Dover (1981).

[27] R. Durrett, Probability: theory and examples, Cambridge Univ. Press (1990).

[28] F. Dyson, Origins of life, Cambridge Univ. Press (1984).

[29] A. Einstein, Relativity: the special and the general theory, Dover (1916).

[30] L.C. Evans, Partial differential equations, AMS (1998).

[31] W. Feller, An introduction to probability theory and its applications, Wiley (1950).

[32] E. Fermi, Thermodynamics, Dover (1937).

[33] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics I: mainly mechanics, radiation and heat, Caltech (1963).

[34] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics II: mainly electromagnetism and matter, Caltech (1964).

[35] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics III: quantum mechanics, Caltech (1966).

[36] R.P. Feynman and A.R. Hibbs, Quantum mechanics and path integrals, Dover (1965).

[37] P. Flajolet and R. Sedgewick, Analytic combinatorics, Cambridge Univ. Press (2009).

[38] A.P. French, Special relativity, Taylor and Francis (1968).

[39] J.H. Gillespie, Population genetics, Johns Hopkins Univ. Press (1998).

[40] C. Godsil and G. Royle, Algebraic graph theory, Springer (2001).

[41] H. Goldstein, C. Safko and J. Poole, Classical mechanics, Addison-Wesley (1980).

[42] D.L. Goodstein, States of matter, Dover (1975).

[43] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).

[44] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).

[45] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).

[46] D.J. Griffiths, Revolutions in twentieth-century physics, Cambridge Univ. Press (2012).

[47] V.P. Gupta, Principles and applications of quantum chemistry, Elsevier (2016).

[48] W.A. Harrison, Solid state theory, Dover (1970).

[49] W.A. Harrison, Electronic structure and the properties of solids, Dover (1980).

[50] R.A. Horn and C.R. Johnson, Matrix analysis, Cambridge Univ. Press (1985).

[51] C.E. Housecroft and A.G. Sharpe, Inorganic chemistry, Pearson (2018).

[52] K. Huang, Introduction to statistical physics, CRC Press (2001).

[53] K. Huang, Fundamental forces of nature, World Scientific (2007).

[54] S. Huskey, The skeleton revealed, Johns Hopkins Univ. Press (2017).

[55] L. Hyman, Comparative vertebrate anatomy, Univ. of Chicago Press (1942).

[56] L.P. Kadanoff, Statistical physics: statics, dynamics and renormalization, World Scientific (2000).

[57] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).

[58] C. Kittel, Introduction to solid state physics, Wiley (1953).

[59] D.E. Knuth, The art of computer programming, Addison-Wesley (1968).

[60] M. Kumar, Quantum: Einstein, Bohr, and the great debate about the nature of reality, Norton (2009).

[61] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).

[62] L.D. Landau and E.M. Lifshitz, Mechanics, Pergamon Press (1960).

[63] L.D. Landau and E.M. Lifshitz, The classical theory of fields, Addison-Wesley (1951).

[64] L.D. Landau and E.M. Lifshitz, Quantum mechanics: non-relativistic theory, Pergamon Press (1959).

[65] S. Lang, Algebra, Addison-Wesley (1993).

[66] P. Lax, Linear algebra and its applications, Wiley (2007).

[67] P. Lax, Functional analysis, Wiley (2002).

[68] P. Lax and M.S. Terrell, Calculus with applications, Springer (2013).

[69] P. Lax and M.S. Terrell, Multivariable calculus with applications, Springer (2018).

[70] S. Ling and C. Xing, Coding theory: a first course, Cambridge Univ. Press (2004).

[71] J.P. Lowe and K. Peterson, Quantum chemistry, Elsevier (2005).

[72] S.J. Marshall, The story of the computer: a technical and business history, Create Space Publ. (2022).

[73] M.L. Mehta, Random matrices, Elsevier (2004).

[74] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge Univ. Press (2000).

[75] R.K. Pathria and and P.D. Beale, Statistical mechanics, Elsevier (1972).

[76] T.D. Pollard, W.C. Earnshaw, J. Lippincott-Schwartz and G. Johnson, Cell biology, Elsevier (2022).

[77] J. Preskill, Quantum information and computation, Caltech (1998).

[78] R. Rojas and U. Hashagen, The first computers: history and architectures, MIT Press (2000).

[79] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[80] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[81] W. Rudin, Functional analysis, McGraw-Hill (1973).

[82] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[83] B. Ryden and B.M. Peterson, Foundations of astrophysics, Cambridge Univ. Press (2010).

[84] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).

[85] R. Shankar, Fundamentals of physics I: mechanics, relativity, and thermodynamics, Yale Univ. Press (2014).

[86] R. Shankar, Fundamentals of physics II: electromagnetism, optics, and quantum mechanics, Yale Univ. Press (2016).

[87] N.J.A. Sloane and S. Plouffe, Encyclopedia of integer sequences, Academic Press (1995).

[88] A.M. Steane, Thermodynamics, Oxford Univ. Press (2016).

[89] S. Sternberg, Dynamical systems, Dover (2010).

[90] D.R. Stinson, Combinatorial designs: constructions and analysis, Springer (2006).

[91] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[92] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[93] J. von Neumann and O. Morgenstern, Theory of games and economic behavior, Princeton Univ. Press (1944).

[94] J. Watrous, The theory of quantum information, Cambridge Univ. Press (2018).

[95] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[96] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[97] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).

[98] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).

[99] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).

[100] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).

Index