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 first discuss elementary plane geometry, angles and triangles, with the basic properties of the sine and cosine discussed. Then we go on a more advanced discussion, using affine and polar coordinates, with the main results of trigonometry explained. We then get into calculus methods, with an even more advanced study of the trigonometric fuctions, and with some applications discussed too. Finally, we provide an introduction to space geometry, spherical coordinates, solid angles, and related questions and applications.

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, 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, along the lines evoked above. The book is organized in 4 parts, as follows:

Part I - We discuss here elementary plane geometry, angles and triangles, with the sine and cosine introduced, and their basic properties explained.

Part II - Here we go on a more advanced discussion, using affine coordinates, and polar coordinates too, with the main results of trigonometry worked out.

Part III - We get here into calculus methods, with an even more advanced study of the trigonometric fuctions, and with some applications discussed too.

Part IV - We provide here an introduction to space geometry, spherical coordinates, solid angles, and various related questions and applications.

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, that we will discuss too, in this book.

Cergy, May 2025
Teo Banica

Contents

Part I

Geometry, angles

Buffalo Soldier, dreadlock Rasta
Fighting on arrival, fighting for survival
Driven from the mainland
To the heart of the Caribbean

CHAPTER 1

Parallel lines

1a. Parallel lines

Welcome to plane geometry. At the beginner level, which is ours for the moment, this will be a story of points and lines. So, let us try to understand this first, what can be said about points and lines, and in what regards more complicated things like angles, triangles, and of course, trigonometry, we will leave them for later.

So, 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:

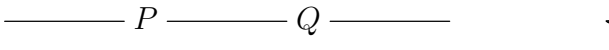
AXIOM 1.1. *Any two distinct points $P \neq Q$ determine a line, denoted PQ.*

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, drawing lines in the real life, for various engineering purposes, we are rather used in practice to draw segments PQ:

This being said, you certainly know from real life that it never hurts to "enhance" your segment, by extending it a bit on both sides, because who knows when you will need that two extra bits, and matter of not getting back to the rule, at that time:

But now in theory, meaning some sort of idealized practice, and going for a big win directly, with no prisoners taken, will having that segment extended to infinity hurt? Certainly not, so this is why our lines PQ in mathematics will be infinite, as above:



Very good all this, so at least we know one thing, why lines instead of segments. And with this being an instance of a general principle that we will heavily use, throughout this book, as mathematicians, namely use our friend ∞, whenever appropriate.

Getting now to point, as already announced, why is Axiom 1.1 an axiom, instead of being a theorem? Not an easy question, the situation being as follows:

(1) You would probably argue here that this theorem can be proved by using a rule, as indicated above.

(2) However, and with my apologies for this, although rock-solid as a scientific proof, this rule thing does not stand for a mathematical proof.

So, 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?

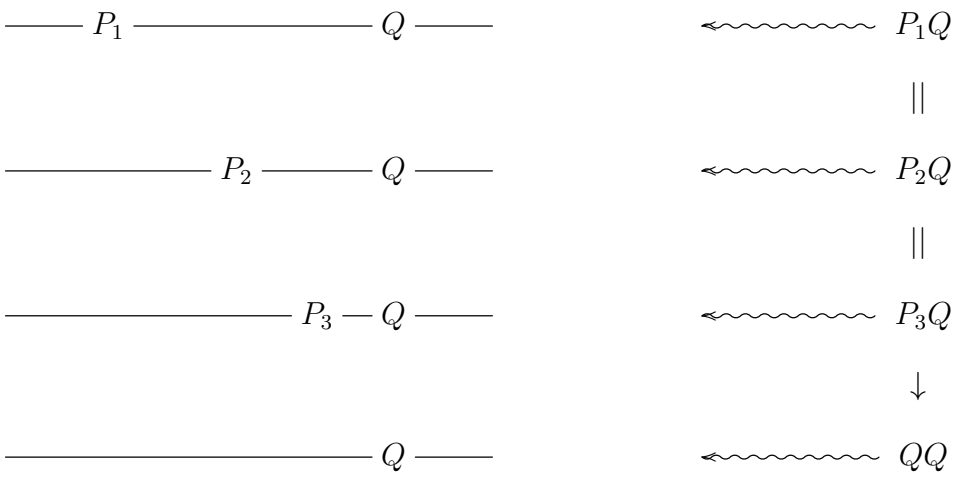
CAT 1.2. *In fact, spacetime itself is bent.*

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.

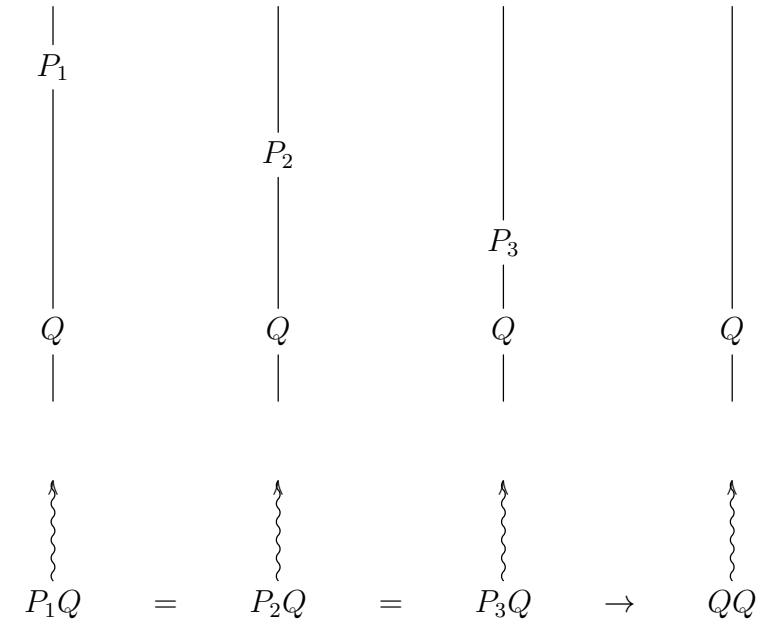
Getting now to more discussion, still around Axiom 1.1, an interesting question appears in connection with our one and only assumption there, namely:

$$P \neq Q$$

Indeed, given a point Q in the plane, we can come up with a sequence of points $P_n \to Q$ horizontally, and in this case the lines P_nQ will all coincide with the horizontal at Q. But then, based on this, we could formally say that the $n \to \infty$ limit of these lines, which makes sense to be denoted QQ, is also, by definition, the horizontal at Q:



However, is this really a good idea, or not. The point indeed is that, when doing exactly the same trick with a series of points $P_n \to Q$ vertically, we will obtain in this way, as our limiting line QQ, the vertical at Q, as shown by the following picture:



Which does not sound very good, so forget about this. However, since we seem to have some sort of valuable idea here, who knows, 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 that the two 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 $P \in K$, $K||L$.*

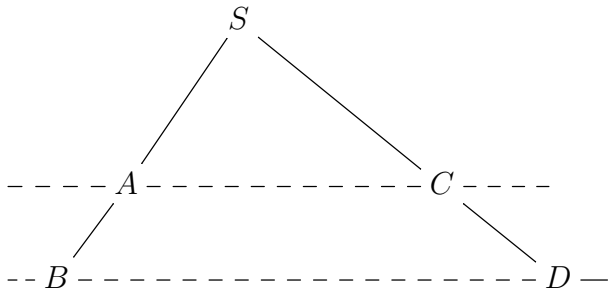
As before, we will leave as an exercise further meditating on all this.

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:

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

To be more precise here, consider a configuration as follows, consisting of two parallel lines, and of two extra lines, which are crossing, and crossing these parallel lines too:



The claim of Thales is then that the following equality holds:

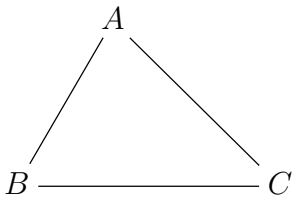
$$\frac{SA}{SB} = \frac{SC}{SD}$$

Moreover, in addition to this, we have some further claims, such as the fact that AC/BD equals the above number, too. And there is more that can be said, along the same lines, this time involving configurations of three parallel lines, and so on.

In what follows the idea will be that of proving the main claim of Thales, which is the equality above, and then deducing from this all sorts of other useful statements, that can be made. But, getting to the point now, how to prove that main claim of Thales?

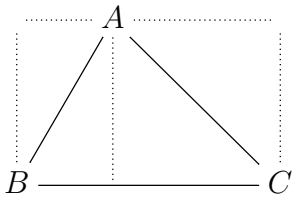
Well, this is actually not obvious, and we will have to trick. Let us start with the following fact, which itself is something quite obvious, and very useful too:

THEOREM 1.11. *The area of a triangle, with a side drawn horizontally,*



is half the product of that side, and of the height.

PROOF. This is clear by completing the picture into a rectangle, as follows:



Indeed, the area of the rectangle is easy to compute, given by:

$$area(\square) = side \times height$$

On the other hand, as it is clear on the above picture, our rectangle appears to be made from two triangles equal to ABC, via some cutting and pasting. Thus:

$$area(\square) = 2 \times area(ABC)$$

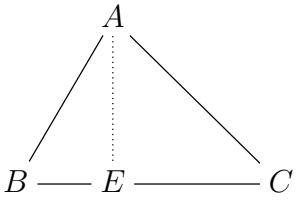
We conclude from this that the area of the triangle is given by:

$$area(ABC) = \frac{1}{2} \times side \times height$$

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

In practice now, it is better to use an equivalent statement, as follows:

THEOREM 1.12. *The area of a triangle, with an altitude drawn,*



is given by the following formula,

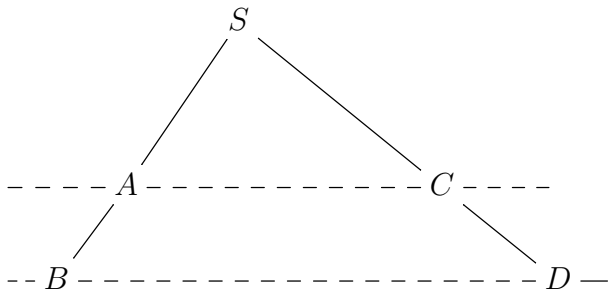
$$area(ABC) = \frac{AE \times BC}{2}$$

and this no matter how our triangle is oriented, in the plane.

PROOF. This follows indeed from Theorem 1.11, by rotating what we found there, or simply by arguing that the method used there in the proof, with constructing that rectangle, works in any direction, with no need for our triangle to lie on the horizontal. □

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

THEOREM 1.13 (Thales). *Proportions are kept, along parallel lines. That is, given a configuration as follows, consisting of two parallel lines, and of two extra lines,*



the following equality holds:

$$\frac{SA}{SB} = \frac{SC}{SD}$$

Moreover, the converse of this holds too, in the sense that, in the context of a picture as above, if this equality is satisfied, then the lines AC and BD must be parallel.

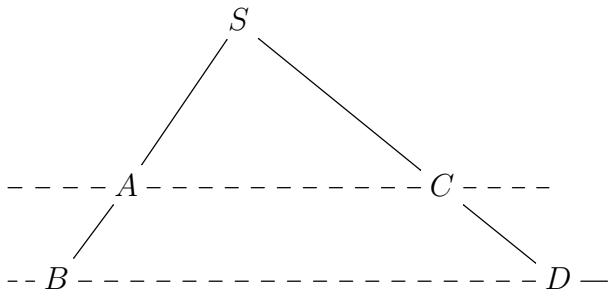
PROOF. We can prove indeed the main assertion via the following computation, based on the area formula in Theorem 1.12, used multiple times:

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

As for the converse, which is actually something quite theoretical, and not that useful in practice, we will leave the proof here as an instructive exercise. □

As already mentioned before, there are many other useful versions of the Thales theorem, which are all good to know. Let us start our discussion here with:

THEOREM 1.14 (Thales 2). *In the context of the Thales theorem configuration,*

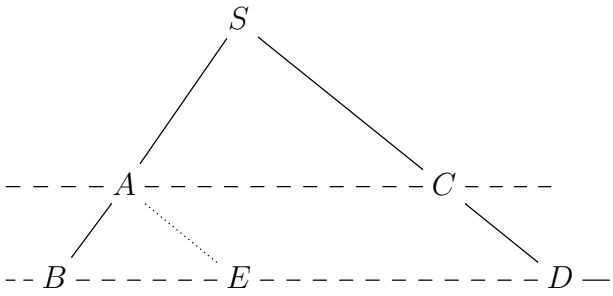


the following equality, involving the same number, holds as well:

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

However, the converse of this does not necessarily hold.

PROOF. In order to prove the formula in the statement, instead of getting lost into some new area computations, let us draw a tricky parallel, as follows:



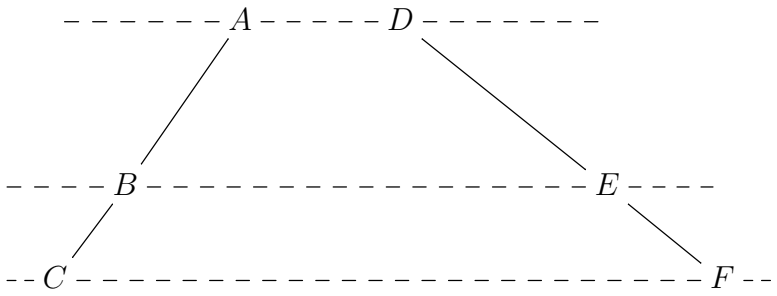
By using Theorem 1.13, we have then the following computation, as desired:

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

As for the converse, as before this is something quite theoretical, and not that useful in practice, we will leave the proof here as an instructive exercise. □

As a third Thales theorem now, which is something beautiful too, we have:

THEOREM 1.15 (Thales 3). *Given a configuration as follows, consisting of three parallel lines, and of two extra lines, which can cross or not,*



the following equality holds:

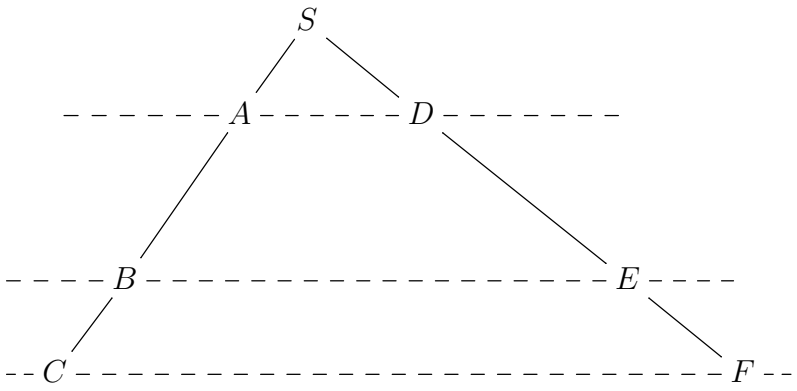
$$\frac{AB}{BC} = \frac{DE}{EF}$$

That is, once again, the proportions are kept, along parallel lines.

PROOF. We have two cases here, as follows:

(1) When the two extra lines are parallel, the result is clear, because we have plenty of parallelograms there, and the fractions in question are plainly equal.

(2) When the two lines cross, let us call S their intersection:



Now by using Theorem 1.13 several times, we obtain:

$$\begin{aligned}\frac{AB}{BC} &= \frac{SB-SA}{SC-SB}\\ &= \frac{1-\frac{SA}{SB}}{\frac{SC}{SB}-1}\\ &= \frac{1-\frac{SD}{SE}}{\frac{SF}{SE}-1}\\ &= \frac{SE-SD}{SF-SE}\\ &= \frac{DE}{EF}\end{aligned}$$

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

Very nice all this, we now master the Thales theorem, in its various formulations, the overall conclusion being that, everything that is clear on pictures, regarding proportions and parallel lines, is true indeed, and we have mathematical proof for that.

As a supplementary conclusion now, still about parallel lines, coming from the proof of Thales 2, which was something quite tricky, with that parallel drawn, we have:

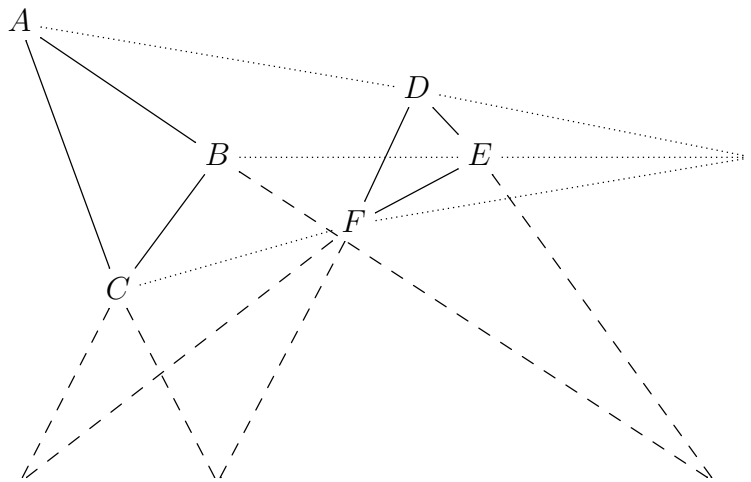
Conclusion 1.16. *Many things can be done with parallel lines, with a suitably drawn such line hopefully solving, by some kind of miracle, your plane geometry problem.*

Which is something good to know. We will see more illustrations for this general principle in the next chapter, when getting more in detail into triangle geometry.

1c. Pappus, Desargues

Moving ahead now, many other things can be said about points and lines, and sometimes parallel lines, as a continuation of the Thales theorem. As a basic statement here, due to Desargues, we have the following fact, that we will prove in what follows:

FACT 1.17 (Desargues). *Two triangles are in perspective centrally if and only if they are in perspective axially. That is, in the context of a configuration of type*



the lines AD, BE, CF cross, so that ABC, DEF are in central perspective, if and only if $AB \cap DE, AC \cap DF, BC \cap EF$ are collinear, so that ABC, DEF are in axial perspective.

Obviously, this is something that can be very useful for various technical computations and drawings, and more on this later. Getting now to the proof of the result, this is something quite tricky. So, with a bit of imagination, we first have:

THEOREM 1.18. *The Desargues claim holds in one sense: central perspectivity implies axial perspectivity.*

PROOF. The trick here is to pass in 3D, as follows:

(1) Assume first that we are in 3D, with our triangles ABC and DEF lying in distinct planes, say $ABC \subset P$ and $DEF \subset Q$. Assuming central perspectivity, the lines AD, BE cross, so the points A, B, D, E are coplanar. But this tells us that the lines AB, DE cross, and that, in addition, their crossing point lies on the intersection of the planes P, Q:

$$(AB \cap DE) \in P \cap Q$$

But a similar argument, again using central perspectivity, shows that we have also:

$$(AC \cap DF) \in P \cap Q \quad , \quad (BC \cap EF) \in P \cap Q$$

Now since the intersection $P \cap Q$ is a certain line in space, we obtain the result.

(2) Thus, almost there, with the theorem proved when the triangles ABC and DEF are both in 3D, in generic position, and the rest is just a matter of finishing. Indeed, when ABC and DEF are still in 3D, but this time lying in the same plane, the result follows too, by perturbing a bit our configuration, as to make it generic. And with this we are done indeed, because we are now in 2D, exactly as in the setting of the theorem. □

In order to prove now to converse, there are several methods and tricks available, and we will choose here to use something quite conceptual. So, temporarily forgetting about Desargues, we have the following result, which is something having its own interest:

THEOREM 1.19. *We have a duality between points and lines, obtained by fixing a circle in the plane, say of center O and radius $r > 0$, and doing the following,*

(1) *Given a point P, construct Q on the line OP, as to have $OP \cdot OQ = r^2$,*
(2) *Draw the perpendicular at Q on the line OQ. This is the dual line p,*

and this duality $P \leftrightarrow p$ transforms collinear points into concurrent lines.

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

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

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

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

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

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

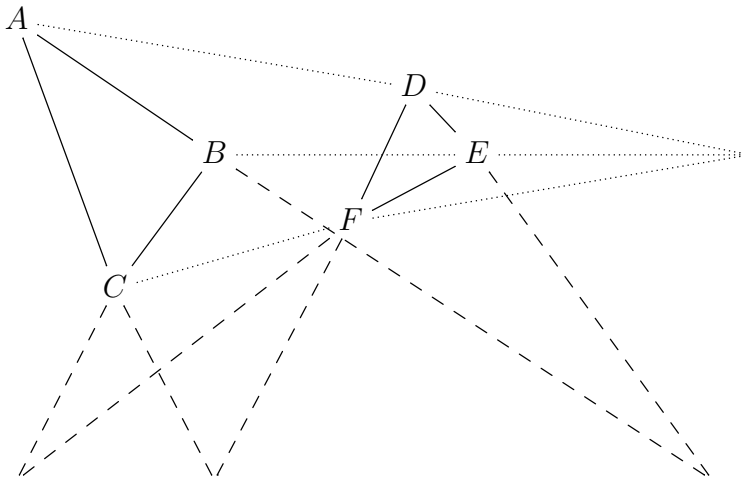
(6) Getting now to the last assertion, this is something which holds indeed. We will be back to this later, with details, once we will know more about circles. □

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

THEOREM 1.20. *The Desargues claim holds in the other sense too: axial perspectivity implies central perspectivity.*

PROOF. As already mentioned, there are several methods and tricks available, in order to prove this, but the simplest is to argue that this is a trivial consequence of Theorem

1.18 and Theorem 1.19. Indeed, let us look at the Desargues configuration, namely:



Let us look now at the dual Desargues configuration, involving triangles abc and def. We have then the following things happening, both coming from Theorem 1.19:

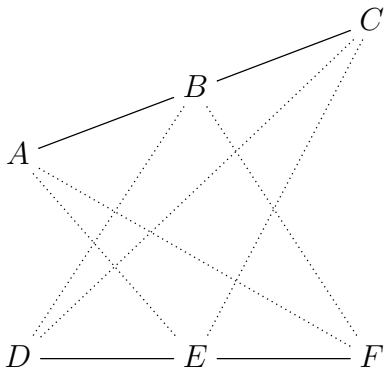
– The original triangles ABC, DEF are in central perspective precisely when the dual triangles abc, def are in axial perspective.

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

But with this, we are done, because Theorem 1.18 applied to the dual triangles abc, def gives the present result, for the original triangles ABC, DEF. □

Summarizing, done with Desargues, and we have learned many interesting things, on this occasion. Next, we have the following fact, going back in time, to Pappus:

FACT 1.21 (Pappus). *Given a configuration as follows,*



the three middle points are collinear.

As before with Desargues, or rather with the tricky implication of Desargues, proving such things will need some preparations. So, temporarily forgetting about Pappus, we have the following result, which is something having its own interest:

THEOREM 1.22. *We can talk about the cross ratio of four collinear points A, B, C, D, as being the following quantity, signed according to our usual sign conventions,*

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

and with this notion in hand, points in central perspective have the same cross ratio:

$$(A, B, C, D) = (A', B', C', D')$$

Moreover, the converse of this fact holds too.

PROOF. As before with Theorem 1.19, there is a lot of mathematics hidden here, and with the formula in the statement coming by drawing a suitable parallel line, and computing both $(A, B, C, D), (A', B', C', D')$ in terms of the new points which appear:

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

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

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

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

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

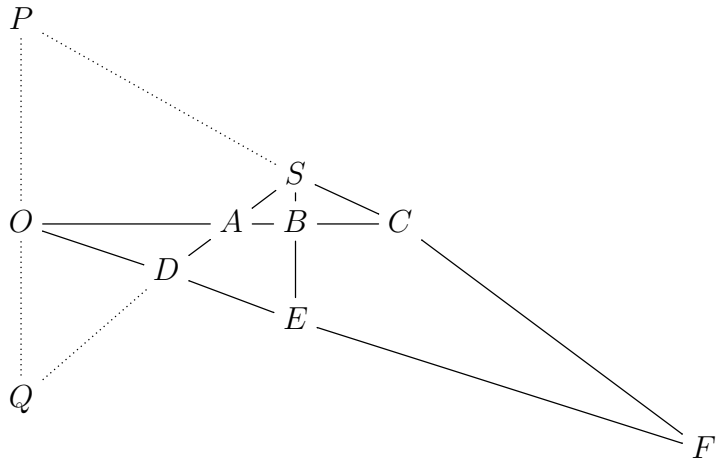
(2) Next, many other things can be said, as for instance being the fact that A, B, C, D are somehow "nicely positioned" on their line when their cross ratio is -1:

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

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

(3) Getting now to what our statement says, in relation with points in central perspective, consider first the following picture, with the points A, B, C, D, E, F and S, O

being as indicated, and with a parallel line to SE drawn on the left, as indicated:



(4) We have then the following equality, obtained by using the Thales theorem:

$$\begin{aligned}(O,B,C,A) &= \frac{OC}{BC}\cdot\frac{BA}{OA}\\ &= \frac{PO}{SB}\cdot\frac{SB}{OQ}\\ &= \frac{PO}{OQ}\end{aligned}$$

On the other hand, again by using the Thales theorem, we have as well:

$$\begin{aligned}(O,E,F,D) &= \frac{OF}{EF}\cdot\frac{ED}{OD}\\ &= \frac{PO}{SE}\cdot\frac{SE}{OQ}\\ &= \frac{PO}{OQ}\end{aligned}$$

We conclude that in the context of the above configuration, we have:

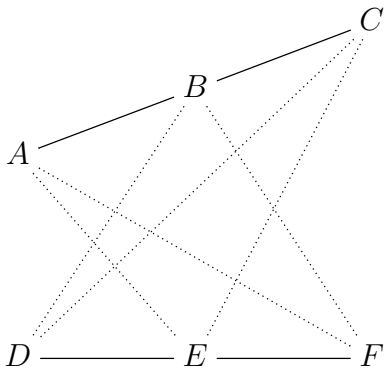
$$(O,B,C,A)=(O,E,F,D)$$

(5) But this gives the equality in statement, by suitably generalizing what we found, somewhat by "blowing up" the point O on the left into a pair of distinct points, and we will leave working out the details here as an instructive exercise.

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

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

THEOREM 1.23 (Pappus). *Given a hexagon $AFBDCE$ with both the odd and the even vertices being collinear*



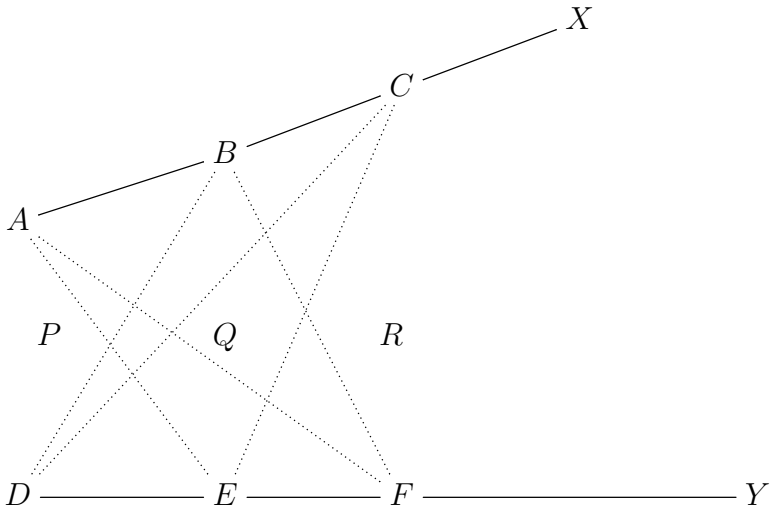
the pairs of opposite sides cross into three collinear points.

PROOF. Observe first the fancier formulation of the statement, with respect to what we had before in Fact 1.21, but this was of course more for fun, of perhaps for some deeper reasons too, these mysterious hexagons sort of rule in plane geometry, and more on this later in this book too, on several occasions. In practice now, what we have to prove remains as in Fact 1.21, and the idea is that can be proved by refining the picture, by adding some extra points, and using the cross ratio technology from Theorem 1.22:

(1) Consider indeed the Pappus configuration in the statement, then let us call P, Q, R the middle points appearing there, and construct points X, Y as follows:

$$X = AC \cap DR \quad , \quad Y = AR \cap DF$$

We obtain in this way an enlarged configuration, which looks as follows:



(2) We have then the following equalities, with the first one coming from Theorem 1.22, via the central perspective coming from the point R, and with the second one being something trivial, valid for any cross ratio, coming from definitions:

$$(A, C, B, X) = (Y, E, F, D) = (D, F, E, Y)$$

(3) But with this equality, we can conclude. Consider indeed the following point, appearing on the left in the picture, that we will need too, in what follows:

$$K = AD \cap PQ$$

Now let us see what happens to the configurations $ACBX$ and $DFEY$, when projected respectively from the points D, A, on the line PQ. Via these projections, we have:

$$ACB \to KQP \quad , \quad DFE \to KQP$$

(4) Now remember the cross ratio formula found in (2), namely:

$$(A, C, B, X) = (D, F, E, Y)$$

In view of this, and by applying again Theorem 1.22, this time in reverse form, we conclude that the images of X, Y via the above projections must coincide:

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

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

$$R \in PQ$$

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

We will be back to such things, points and lines, in the next chapter, with more specialized results, when getting more in detail into triangle geometry. Then we will be back to this in Part II using coordinates, and also by replacing some lines by curves.

As a final comment on all this, quite philosophical, what was more basic, or more fundamental, or just simpler to establish, Pappus or Desargues?

Good and mysterious question, which is a bit beyond our present reach, but we will be back to this too, later in this book, once we will know more.

1d. Projective plane

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 1.24. *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 $P_{\mathbb{R}}^2$, 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 $P_{\mathbb{R}}^2$, and let us also call each plane in $\mathbb{R}^3$ passing through the origin a "line" of $P_{\mathbb{R}}^2$. 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 1.24. More generally now, we have the following construction, in arbitrary dimensions:

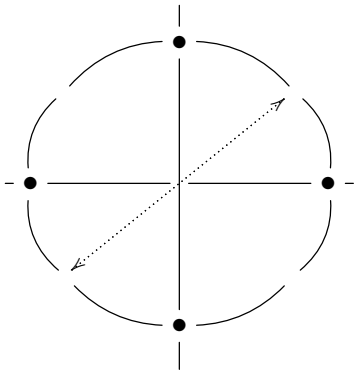
THEOREM 1.25. *We can define the projective space $P_{\mathbb{R}}^{N-1}$ as being the space of lines in $\mathbb{R}^N$ passing through the origin, and in small dimensions:*

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

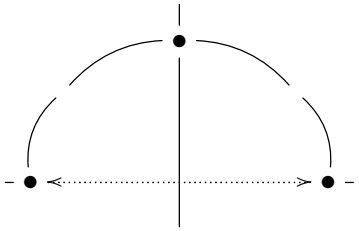
PROOF. We have several assertions here, with all this being of course a bit informal, and self-explanatory, the idea and some further details being as follows:

(1) To start with, the fact that the space $P_{\mathbb{R}}^{N-1}$ constructed in the statement is indeed a projective space in the sense of Definition 1.24 follows from definitions, exactly as in the discussion preceding the statement, regarding the case $N = 3$.

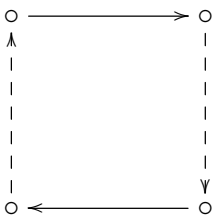
(2) At $N = 2$ now, a line in $\mathbb{R}^2$ passing through the origin corresponds to 2 opposite points on the unit circle $\mathbb{T} \subset \mathbb{R}^2$, according to the following scheme:



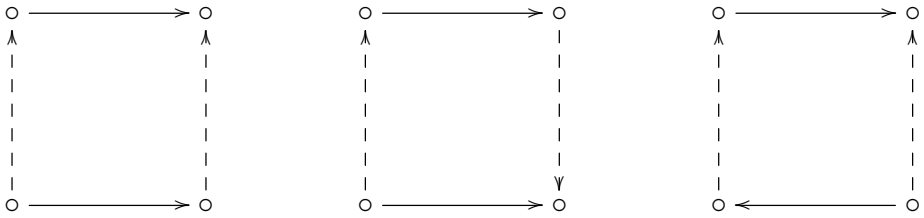
Thus, $P^1_{\mathbb{R}}$ corresponds to the upper semicircle of $\mathbb{T}$, with the endpoints identified, and so we obtain a circle, $P^1_{\mathbb{R}} = \mathbb{T}$, according to the following scheme:



(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.

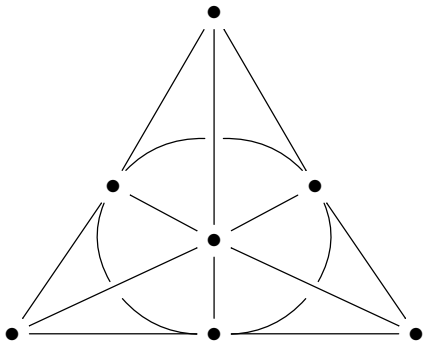
(6) Finally, regarding the third space, the one on the right, we know by symmetry that this must be the Klein bottle too. But we can see this as well via our standard folding method, namely identifying solid edges first, and dotted edges afterwards. Indeed, we first obtain in this way a Möbius strip,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

Getting now to geometry, many things can be said, in the projective setting, as a continuation of the basic fact from Definition 1.24, namely that any two lines cross.

Observe however that, in what regards lengths and areas, we are a bit in trouble. We will be back to this later, once we will have more knowledge of geometry.

Finally, let us mention that Definition 1.24 is something far wider than it might seem. Consider indeed the following configuration of 7 points and 7 lines, called Fano plane:



Here the circle in the middle is by definition a line, and with this convention, the basic axioms in Definition 1.24 are satisfied, in the sense that any two points determine a line, and any two lines determine a point. And isn't this beautiful.

We will be back to this later, with some further explanations, but in the meantime, just a quick discussion. Let us start with something philosophical, as follows:

QUESTION 1.26. *What are the numbers that we are allowed to use, in geometry?*

In answer, you would say, we definitely need $0, 1$. But then once we have these, we have as well $2 = 1 + 1$, then $3 = 2 + 1$, and so on, so we have $\mathbb{N}$. But then once we have $\mathbb{N}$, by looking at $a + b = c$ we are led to $\mathbb{Z}$, and then by looking at $ab = c$ we are led to $\mathbb{Q}$. Thus, on the bottom line, we need for geometry all the rational nunbers, $r \in \mathbb{Q}$.

However, and here comes our point, abstractly speaking, this is not exactly true, because, by a strange twist of fate, the numbers $0, 1$, whose presence in a number system is mandatory, can form themselves a "number system", with addition as follows:

$$1 + 1 = 0$$

Indeed, regarding the operations involving $0, 1$, we certainly must have:

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

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

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

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

As a conclusion, what we get in this way is some sort of a new "number system", $\mathbb{F}_2 = \{0, 1\}$. And the point now is that, by thinking well, the projective plane over $\mathbb{F}_2$, which must be something finite, is the Fano plane. More on this later in this book.

1e. Exercises

Exercises:

Exercise 1.27.

Exercise 1.28.

Exercise 1.29.

Exercise 1.30.

Exercise 1.31.

Exercise 1.32.

Exercise 1.33.

Exercise 1.34.

Bonus exercise.

CHAPTER 2

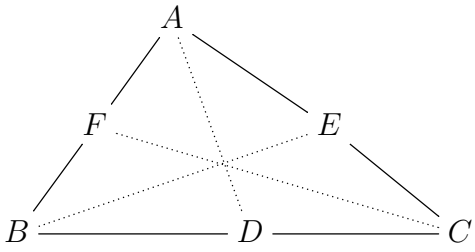
Triangles

2a. Triangles, centers

Welcome to triangle geometry, which is on the route of what we want to do in this book, namely angles and trigonometry. In fact, you can sense this right away, with "triangle" obviously coming from "three angles". And with the point being that, while angles taken alone are quite hard to investigate, angles coming in triplets, that is, in the form of triangles, are relatively easy to get into, via Thales and other techniques.

But let us start our study of triangles with the most important triangle result of them all, which is actually unrelated to angles. This is the barycenter theorem, as follows:

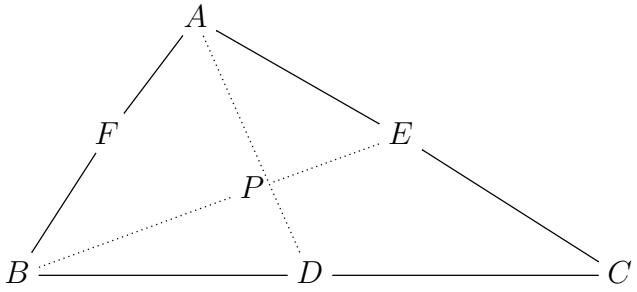
THEOREM 2.1 (Barycenter). *Given a triangle ABC, its medians cross,*



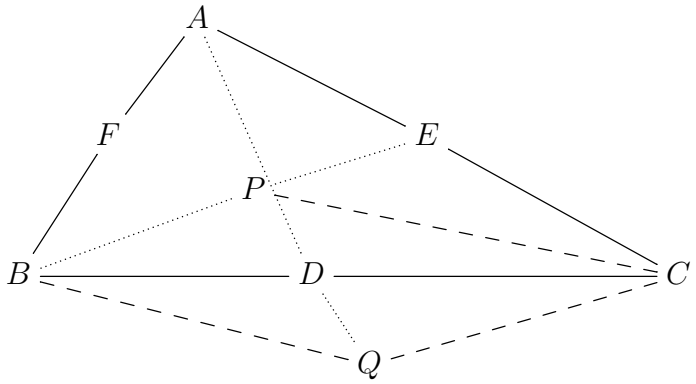
at a point called barycenter, lying at $1/3-2/3$ on each median.

PROOF. The idea is that we can get this from Thales, via some tricks:

(1) Let us draw indeed the medians AD and BE, and call P their intersection:



(2) Now comes the trick. Let us symmetrize P with respect to D, into a point Q:



(3) Since $BD = DC$ and $PD = DQ$, the figure $BPCQ$ is a parallelogram. In particular the sides BP and CQ are parallel, and by Thales, we obtain from this:

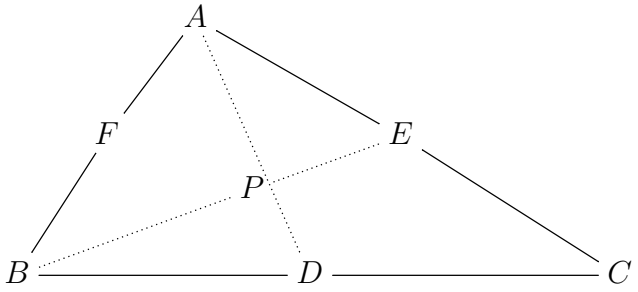
$$AE = EC \implies AP = PQ$$

On the other hand, remember that D was the midpoint of PQ. Thus, we obtain:

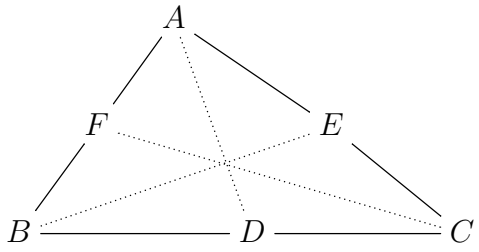
$$AP = 2PD$$

(4) Summarizing, we have proved that when intersecting two medians, the intersection point lies at $1/3 - 2/3$ on one of the two medians. But, by symmetry, this intersection point must lie as well at $1/3 - 2/3$ on the other median, that we have intersected.

(5) So, getting back to our original picture, from (1), we have proved that the $1/3-2/3$ proportions which are obvious on the picture, on both AD, BE, happen indeed:



(6) But with this, we are done. Indeed, if we consider, on each of the 3 medians, the point lying at $1/3 - 2/3$ on that median, then by (5) these 3 points will coincide:



Thus, as a conclusion to this study, the medians of our triangle cross indeed, at a point lying at $1/3 - 2/3$ on each of them, as claimed in the statement. □

The barycenter has many interesting properties, the most important of which, in relation with intuition and physics, can be summarized as follows:

FACT 2.2. *The gravity center of a triangle ABC is as follows:*

(1) *In the 0-dimensional case, that is, when putting equal weigths at the vertices A, B, C, and computing the center, this is the barycenter.*
(2) *In the 1-dimensional case, that is, with the sides AB, BC, AC have weigths proportional with their length, this is, in general, different from the barycenter.*
(3) *In the 2-dimensional case, that is, with the triangle ABC itself, as an area, having a weight, uniformly distributed, this is again the barycenter.*

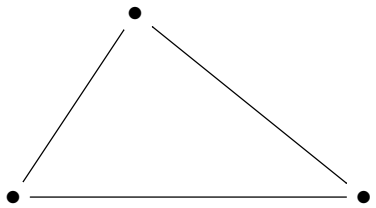
All this looks quite interesting, so let us try now to have some understanding of this. But, we are faced right away with the following question: how to compute, in practice, the barycenter of a configuration of weights, say as in (1), (2), (3) above?

Not an easy question, but based on everyday experience, let us formulate:

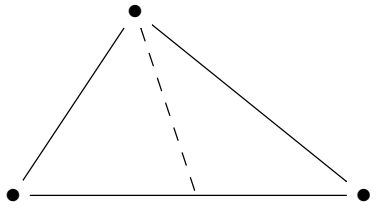
METHOD 2.3. *In order to compute the barycenter of a plane object:*

(1) *We can come up with a blade, put it under the object, and find the correct angle, such as the object lies in equilibrium, on the blade.*
(2) *In this case, with the object lying in equilibrium, we can say, mathematically, that the barycenter lies on the line of the blade edge.*
(3) *And by doing twice this procedure, we can exactly locate the barycenter, as being the interesection of the two lines that we obtain.*

So, let us see how this method works, in relation with Fact 2.2 (1). Consider, as indicated there, a triangle, with equal weights installed at the vertices:



Now let us come with the blade, as indicated in Method 2.3 (1), and try to find the correct angle, as for the upper vertex to lie on the blade, and for the whole triangle to be in equilibrium. Our claim is that, in order for this to happen, the blade must be precisely positioned on the median emanating from the upper vertex, as follows:



Indeed, in this configuration, we have equilibrium, because the upper weight, which is on the blade, will not matter, and the left and right weigths, being equally distanced from the blade, as you can see by drawing two perpendiculars, which will obviously be equal, will cancel each other's effect. Moreover, we can also see that if we move the blade a bit to the right, the triangle will obviously fall to the left, and that if we move the blade a bit to the left, the triangle will obviously fall to the right. Thus, claim proved.

But with this claim proved, we are done, our conclusion being as follows:

Conclusion 2.4. *When computing the physical barycenter as in Method 2.3, a triangle having equal weights installed at the vertices must have its barycenter on each of the three medians. Thus, these three medians cross, at the physical barycenter.*

Which is very nice, not only we have a proof now for what is said Fact 2.2 (1), equality of the mathematical and physical barycenters, but as a bonus, we have as well an alternative proof for Theorem 2.1, using an old-fashioned blade, instead of math.

Getting now to Fact 2.2 (2), that we would like to understand next, that is a negative result, with a degenerate triangle being a counterexample there, as follows:

$$AB \text{ ——————— } C$$

Indeed, the usual barycenter of this degenerate triangle, appearing as in Theorem 2.1, or as in Fact 2.2 (1), obviously lies at $1/3 - 2/3$ on the segment, as follows:

$$AB \text{ —— } P \text{ ———— } C$$

However, in the context of Fact 2.2 (2), the side AB, which is zero, does not matter, and the sides AC, BC both have their centers at the middle of the segment. Thus, the center of gravity of our degenerate triangle is in this case the middle of the segment:

$$AB \text{ ——— } P_1 \text{ ——— } C$$

Getting now to the context of Fact 2.2 (3), this is something a bit more tricky to understand, with a limit involved, and in the end we obtain the usual barycenter:

$$AB \text{ —— } P_2 \text{ ———— } C$$

We will leave some thinking here as an instructive exercise, and this because we will have to come back to Fact 2.2 (3) in a moment, anyway. So, degenerate triangles studied, and as a conclusion to this discussion, around Fact 2.2 (2), let us formulate:

CONCLUSION 2.5. *The centers of a degenerate triangle are as follows, with the subscripts $0, 1, 2$ standing for the dimensionality of the problem, in the sense of Fact 2.2,*

$$AB \text{ —— } P_{0,2} \text{ — } P_1 \text{ ——— } C$$

and with $P_{0,2}$ being the usual, mathematical barycenter, the one from Theorem 2.1.

Summarizing, done with Fact 2.2 (1) and Fact 2.2 (2), via various methods, but Fact 2.3 (3) still remains to be understood, in the case of the arbitrary triangles.

And here, coming as bad news, what we have in Method 2.3 does not apply well to the solid triangles, and their discretizations. So, we must come up with something new. And, a bit of thinking here, again inspired from everyday experience with various objects, leads to the following method, standing as a complement to Method 2.3:

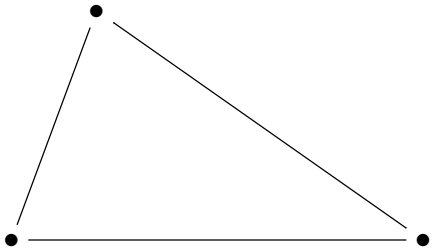
METHOD 2.6. *In order to compute the barycenter of a plane object:*

(1) *We can discretize the object, by approximating chunks of mass ε with point weigths of mass ε, positioned anywhere inside the corresponding chunk.*
(2) *For discrete objects, we can use the rule that a two-point configuration $a - b$ can be replaced with $a + b$ lying at $\frac{b}{a+b} - \frac{a}{a+b}$ on the segment, and recursivity.*
(3) *Thus, we have an algorithm for computing the barycenter of our discretization. And by taking the limit $\varepsilon \to 0$, we reach to the barycenter of the initial object.*

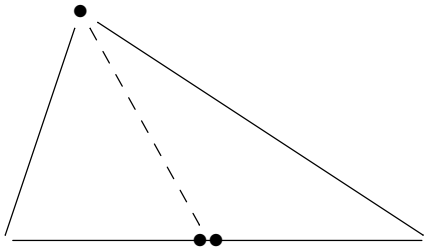
To be more precise here, passed some standard discretization talk, done in (1) and (3), the main point lies in the rule in (2), which itself is something very intuitive, say coming from Method 2.3. Indeed, ignoring the rest of the configuration, a blade passing through the point at $\frac{b}{a+b} - \frac{a}{a+b}$ on the segment will certainly have our $a - b$ configuration lying in

equilibrium, so in practice we can replace if we want this $a-b$ configuration by a point mass of $a+b$ positioned there, at $\frac{b}{a+b}-\frac{a}{a+b}$ on the segment, as indicated above.

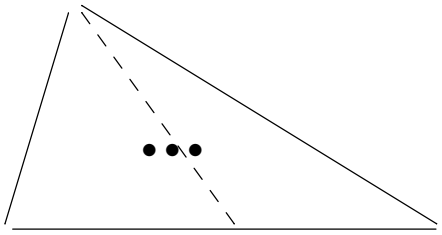
As an illustration, let us reprove Fact 2.2 (1) by using this new method. Consider, as indicated in Fact 2.2 (1), a triangle, with equal weights installed at the vertices:



By using the rule in Method 2.6 (2), we can merge the lower weights, as follows:



But then, by using again this rule, we can further merge our weights, as follows:

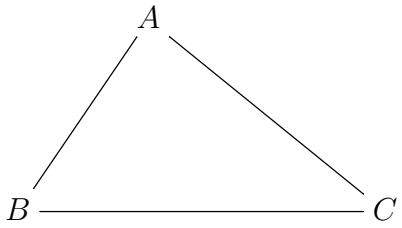


Thus, Fact 2.2 (1) proved again, our conclusion being as follows:

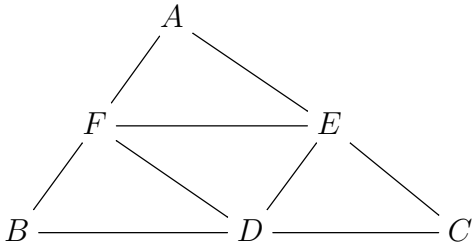
CONCLUSION 2.7. *When computing the physical barycenter as in Method 2.6, a triangle having equal weights installed at the vertices must have its barycenter lying at* $1/3-2/3$ *on each median. Thus, these three medians cross, at the physical barycenter.*

Which is again nice, not only we have now a new proof for what is said Fact 2.2 (1), equality of the mathematical and physical barycenters, but as a bonus, we have as well a full alternative proof for Theorem 2.1, including the $1/3-2/3$ claim there.

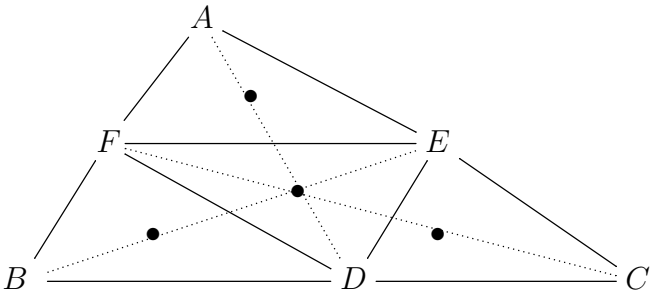
Getting now to what Fact 2.2 (3) says, consider as indicated there a solid triangle, with uniformly distributed weight, all across its surface, of total mass 1:



Now let us discretize this triangle, as in Method 2.6 (1). An easy way of doing so, with $\varepsilon = 1/4$, is by cutting the triangle in 4 obvious equal parts, as follows:



In order to finish our $\varepsilon = 1/4$ discretization, we still have to pick the positions of our 4 point weights, inside the above 4 triangles. As mentioned in Method 2.6 (1), the precise positions of these points will not matter in the end, in relation with our overall $\varepsilon \to 0$ computation, so there are many possible choices here. As a standard choice, however, that we will use here, we have the mathematical barycenters of the above 4 triangles. And with this done, the picture of our $\varepsilon = 1/4$ discretization becomes as follows:



Getting now to the computation of the barycenter of this 4-point configuration, this is clear, because we can see that this 4-point configuration actually consists of a triangle, and its barycenter. Thus, as barycenter, we obtain the point in the middle. Of course, this computation can be done too by using the rules in Method 2.6 (2).

Summarizing, done with $\varepsilon = 1/4$. The next step is $\varepsilon = 1/16$, by cutting each of the small 4 triangles in 4 parts, as before, then $\varepsilon = 1/64$ and so on. We will leave the

computations here as an instructive exercise, and as a conclusion to all this, our method works indeed, and we reach in this way to a proof of Fact 2.2 (3):

CONCLUSION 2.8. *When computing the physical barycenter as in Method 2.6, a solid triangle, with uniformly distributed weight, has a barycenter the usual barycenter.*

Very good all this, so we have now a decent knowledge of the barycenter, and time to talk about something else. However, before doing so, let us listen as well to what cat has to say. Cat indeed is constantly meowing, and this since the beginning of this chapter, when I stated Theorem 2.1. So, what is is, cat, found some mice over there?

CAT 2.9. *Yes, with three mice situated at A, B, C, a cat situated at*

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

can catch them all, and with $1/3 - 2/3$ and everything, without much trouble.

Humm, interesting remark, I can feel here that cat is suggesting that my proof of Theorem 2.1, with that Thales trick and everything, might be actually a terrible complication, hiding both the mathematics and physics of the thing. But, go understand what cat is exactly saying, how come can he sum points of the plane, and then divide by 3, just like this. That is theoretical physics I guess, so let us leave this as homework, for later:

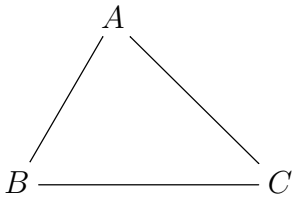
JOB 2.10 (update). *Come back later to the mathematics and physics of the barycenter, with full, conceptual proofs for what is said in both Theorem 2.1, and Fact 2.2.*

Moving ahead now, coming as a complement of Theorem 2.1, we have the following result, making appear 3 more centers of our triangle, which all have their own importance and interest, and which are in general different from the barycenter:

THEOREM 2.11. *Given a triangle ABC, the following happen:*

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

PROOF. 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:

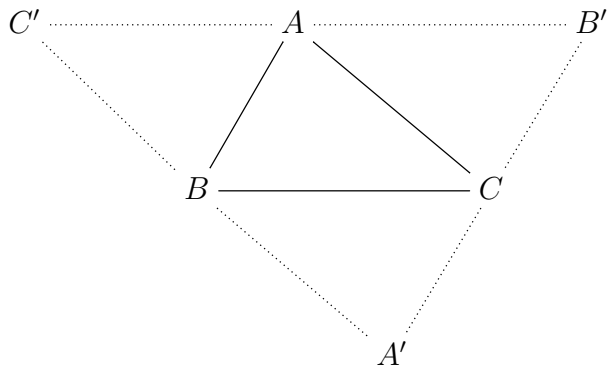


Allowing us the freedom to play with some tricks, as advanced mathematicians, both students and professors, are allowed to, here is how the proof goes:

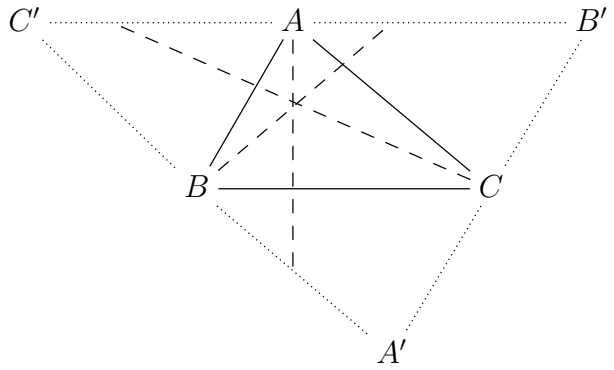
(1) Come with a small circle, inside ABC, and then inflate it, as to touch all 3 edges. The center of the circle will be then at equal distance from all 3 edges, so it will lie on all 3 angle bisectors. Thus, we have constructed the incenter, as required.

(2) 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.

(3) This is something 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, $A'B'C'$, as follows:



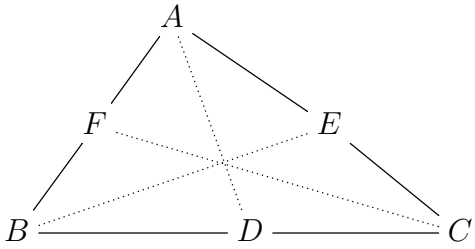
But then, the circumcenter of this bigger triangle $A'B'C'$, that we know to exist from (2), will be the orthocenter of ABC, as shown by the following picture:



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

All this is quite interesting, and as a question emerging from this, we have:

QUESTION 2.12. *Can we have some general theory going, for the various centers of a triangle, notably with results stating that when drawing lines of type AD, BE, CF,*



these lines cross indeed? Also, what about the various centers of a triangle, that we can obtain in this way, what are the exact relations between them?

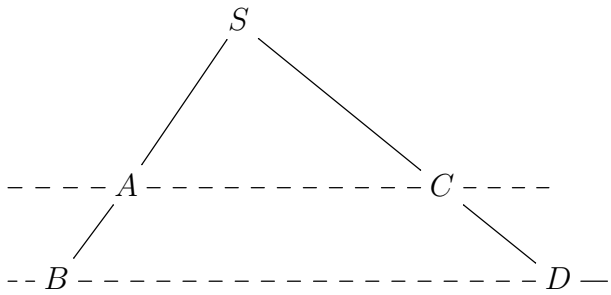
These are all interesting questions, and we will answer them, in due time. The idea in what follows will be that of getting into the study of angles, leading to all sorts of useful results and formulae, and then come back to the above questions, later in this chapter.

2b. 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 2.13. *We can talk about the angle between two crossing lines, and have some basic theory for the angles going, by using triangles, and Thales, in the obvious way.*

To be more precise here, let us go back to the configuration from the Thales theorem, from chapter 1, which was as follows, with two parallel lines, and two other lines:



In this situation, we can say that the two triangles SAC and SBD are similar, and witn an equivalent formulation of similarity being the fact that the angles are equal:

DEFINITION 2.14. *We say that two triangles are similar, and we write*

$$SAC \sim SBD$$

when their respective angles are equal.

The point now is that, in this situation, we can have some mathematics going, for the lengths, coming from the following formula, which is the Thales theorem:

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

Many things can be said here. We will be back to this, with even more.

At the philosophical level now, you might wonder of course what the values of these angles, that we have been heavily using in the above, should be, say as real numbers. But this is something quite tricky, that will take us some time to understand. In the lack of something bright, for the moment, let us formulate the following definition:

DEFINITION 2.15. *We can talk about the numeric value of angles, as follows:*

(1) *The right angle has value* 90°.
(2) *We can double angles, in the obvious way.*
(3) *Thus, the half right angle has value* 45°*, and the flat angle has value* 180°.
(4) *We can also triple, quadruple and so on, again in the obvious way.*
(5) *Thus, we can talk about arbitrary rational multiples of* 90°.
(6) *And, with a bit of analysis helping, we can in fact measure any angle.*

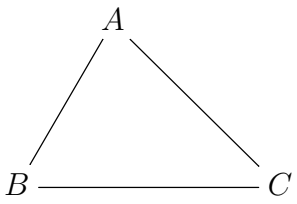
So, this will be our starting definition for the numeric values of the angles. Of course, all this might seem a bit improvized, but do not worry, we will come back later to this, with a better, more advanced definition for these numeric values of the angles.

As another comment, you might wonder what that 90 figure for the right angles stands for. In answer, no one really knows, this is just some convention, old as our modern world and mathematics, say a bit similar to the 10 that we use as numeration basis. Although, with the terrestrial month, based on the movement of the Moon, having about 30 days, we can see here why 10 and its multiples are important to us, humans.

In any case, comment recorded, and we will come back to this later, with a genius new method for rescaling the angles, independently on astronomy and the Moon, with the right angle 90° being destined to be called $\pi/2$, with $\pi=3.1415\ldots$ being a certain very complicated number. And with this being not a joke. More later.

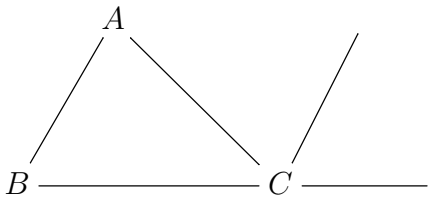
Getting back to work now, theorems and proofs, in relation with the above, here is a key result, which will be our main tool for the study of the angles:

THEOREM 2.16. *In an arbitrary triangle*

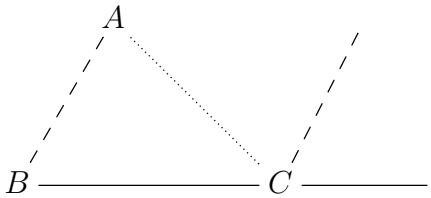


the sum of all three angles is 180°.

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



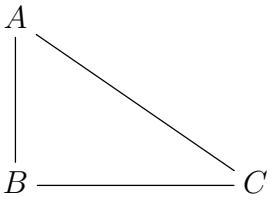
The claim is then that this gives the result. Indeed, in order to better understand what is going on, let us color our picture, as follows:



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

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

DEFINITION 2.17. *We call right triangle a triangle of type*

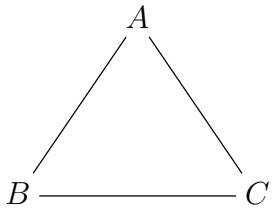


having one of the angles equal to 90°.

Many things can be said about right triangles. We will be back to this.

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

THEOREM 2.18. *In an equilateral triangle, having all sides equal,*

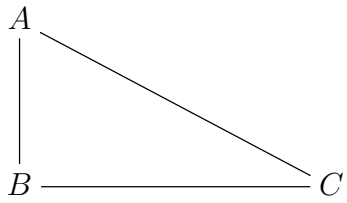


all angles equal 60°*.*

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

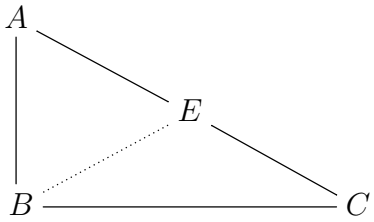
Another interesting angle is the 30° one. About it, we have:

THEOREM 2.19. *In a right triangle having small angles* $30^\circ, 60^\circ$*,*



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

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



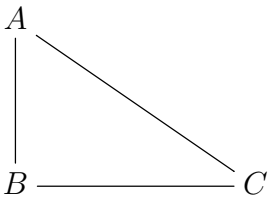
Thus, we are led to the conclusion in the statement. □

We will be back to such things later, when doing trigonometry.

2c. Pythagoras theorem

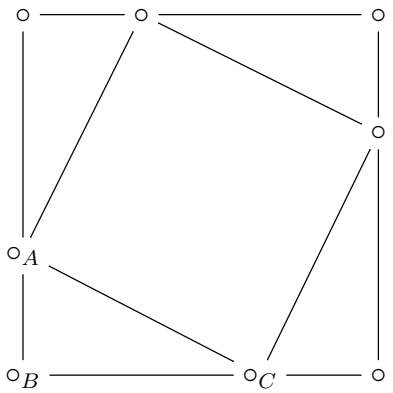
Many other interesting things can be said about the right angle 90°, and about right triangles, in particular with the following key result, due to Pythagoras:

THEOREM 2.20 (Pythagoras). *In a right triangle ABC,*



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

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



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

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

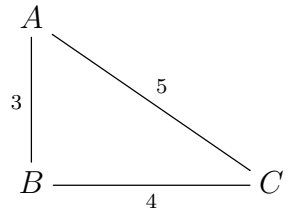
On the other hand, the outer square is made of the smaller square, having side AC, and of four identical right triangles, having sizes AB, BC. Thus:

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

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

As a basic application of the Pythagoras theorem, which is something widely useful in practice, and this since the ancient times, we have:

THEOREM 2.21. *A triangle having sides* $3, 4, 5$ *is a right triangle:*



Thus, for drawing right angles, you only need a loop, with 12 *knots on it.*

PROOF. Here the first assertion comes from the following equality, and from the obvious converse of the Pythagoras theorem, and up to you to check the details here:

$$16 + 9 = 25$$

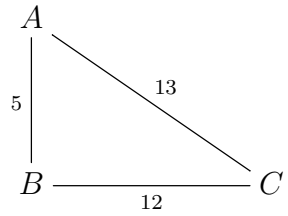
As for the second assertion, what does that mean, and how can that be used in practice, we will leave this as an engineering exercise. □

Still speaking engineering, having 12 knots equally spaced on a loop is certainly possible, and reliable for most tasks, but if we want to improve our tool, it would be desirable to have more knots on out loop. So, here we are looking for integer solutions of:

$$a^2 + b^2 = c^2$$

Which is not exactly obvious, but with a bit of patience, we are led to:

THEOREM 2.22. *A triangle having sides* $5, 12, 13$ *is a right triangle:*



Thus, for drawing right angles, you only need a loop, with 30 *knots on it.*

PROOF. Here the first assertion comes from the following equality, and with the comment that this is the simplest possible one, passed $16 + 9 = 25$:

$$144 + 25 = 169$$

As for the second assertion, we will leave this again as an engineering exercise. As a bonus exercise, try further improving this, say with a solution using 90 knots. □

Along the same lines, at a more advanced level, we have the following result, which fully closes the discussion, regarding the Pythagoras equation over the integers:

THEOREM 2.23. *The Pythagoras equation, namely*

$$a^2+b^2=c^2$$

can be fully solved over the integers, the solutions being

$$a=d(m^2-n^2)\quad,\quad b=2dmn\quad,\quad c=d(m^2+n^2)$$

with $(m,n)=1$*, up to exchanging* a,b*.*

PROOF. This is something standard, due to Euclid, the idea being as follows:

(1) Let us try to solve $a^2+b^2=c^2$. If we divide a,b,c by their greatest common divisor $d=(a,b,c)$, the equation is still satisfied. Thus, we can assume $(a,b,c)=1$, and we want to prove that the solutions are as follows, up to exchanging a,b:

$$a=m^2-n^2\quad,\quad b=2mn\quad,\quad c=m^2+n^2$$

(2) To start with, in one sense our result is clear, because given any two numbers m,n, the above formulae produce a solution to our equation, as shown by:

$$\begin{aligned}(m^2-n^2)^2+(2mn)^2&=m^4+n^4-2m^2n^2+4m^2n^2\\&=m^4+n^4+2m^2n^2\\&=(m^2+n^2)^2\end{aligned}$$

(3) So, we must prove now the converse, stating that if a,b,c satisfying $(a,b,c)=1$ are solutions of $a^2+b^2=c^2$, then we can write them as in (1). For this purpose, the first observation is that, due to $a^2+b^2=c^2$, our assumption $(a,b,c)=1$ implies:

$$(a,b)=(a,c)=(b,c)=1$$

(4) Let us study now the parity of a,b,c. Since $(a,b)=1$, one of these two numbers, say a, is odd. Now assuming that b is odd too, we would get $a^2+b^2=2(4)$, which is impossible, due to $a^2+b^2=c^2$. Thus b must be even, and as a conclusion to this study, up to exchanging a,b, we can assume that the parity of our numbers is as follows:

$$a=\text{odd}\quad,\quad b=\text{even}\quad,\quad c=\text{odd}$$

(5) Now comes the trick. We can rewrite our equation in the following way:

$$\begin{aligned}a^2+b^2=c^2&\iff b^2=c^2-a^2\\&\iff b^2-(c-a)(c+a)\\&\iff\frac{c+a}{b}=\frac{b}{c-a}\end{aligned}$$

(6) With this done, let us look at the fraction on the left. This is a rational number, so we can write it in reduced form, as follows, with $(m,n)=1$:

$$\frac{c+a}{b}=\frac{m}{n}$$

Now observe that our equation, as reformulated in (5), takes the following form:

$$\frac{c+a}{b}=\frac{m}{n}\quad,\quad\frac{c-a}{b}=\frac{n}{m}$$

Equivalently, our equation, as reformulated in (5), takes the following form:

$$\frac{c}{b}+\frac{a}{b}=\frac{m}{n}\quad,\quad\frac{c}{b}-\frac{a}{b}=\frac{n}{m}$$

But this latter system is equivalent to the following two formulae:

$$\frac{a}{b}=\frac{1}{2}\left(\frac{m}{n}-\frac{m}{n}\right)=\frac{m^2-n^2}{2mn}$$

$$\frac{c}{b}=\frac{1}{2}\left(\frac{m}{n}+\frac{m}{n}\right)=\frac{m^2+n^2}{2mn}$$

(7) Good work that we did, and time to breathe, and see what we have. We have proved so far that if a,b,c satisfying $(a,b,c)=1$ are solutions of $a^2+b^2=c^2$, then up to exchanging a,b, we can find numbers m,n satisfying $(m,n)=1$, such that:

$$\frac{a}{b}=\frac{m^2-n^2}{2mn}\quad,\quad\frac{c}{b}=\frac{m^2+n^2}{2mn}$$

Which sounds nice, because due to $(a,b)=(b,c)=1$, as noted in (3), the two fractions on the left are in reduced form. So, if we manage to prove that the two fractions on the right are in reduced form too, this would finish the proof, because we would get:

$$a=m^2-n^2\quad,\quad b=2mn\quad,\quad c=m^2+n^2$$

(8) So, let us look now at the two fractions on the right, appearing above. As a first observation, due to $(m,n)=1$, the following two fractions are in reduced form:

$$\frac{m^2-n^2}{mn}\quad,\quad\frac{m^2+n^2}{mn}$$

The problem, however, is that the fractions in (7) are the halves of these quantities. So, all we need is a study modulo 2, and with this, normally done.

(9) Getting now to the endgame, from $(m,n)=1$, the case where both m,n are even is excluded. But the case where both m,n are odd is excluded too, due to:

$$\frac{a}{b}=\frac{m^2-n^2}{2mn}$$

Indeed, if m,n were both to be odd, we would have $m^2-n^2=0(4)$ and $2mn=2(4)$, so the fraction on the right, when reduced, would have an even denominator. But this would tell us that b must be even, which contradicts our b odd choice from (4).

(10) Summarizing, one of the numbers m,n must be even, and the other must be odd. But this does the job, because it shows that m^2-n^2 and m^2+n^2 are both odd, so when

dividing the reduced fractions from (7) by 2, these fractions remain still reduced. Thus, as a conclusion to our study, the following two fractions are reduced:

$$\frac{m^2-n^2}{2mn} \quad , \quad \frac{m^2+n^2}{2mn}$$

(11) So, theorem proved. Indeed, as indicated in (7), let us look now at:

$$\frac{a}{b}=\frac{m^2-n^2}{2mn} \quad , \quad \frac{c}{b}=\frac{m^2+n^2}{2mn}$$

Since all fractions appearing here are in reduced form, we obtain from this:

$$a=m^2-n^2 \quad , \quad b=2mn \quad , \quad c=m^2+n^2$$

And finally, as indicated in (1), by multiplying a,b,c by an arbitrary number d, we obtain the general solutions from the statement, namely:

$$a=d(m^2-n^2) \quad , \quad b=2dmn \quad , \quad c=d(m^2+n^2)$$

(12) At the level of the interesting examples now, there are of course many of them, and we have for instance a solution as follows:

$$40^2+9^2=1681=41^2$$

Thus, and good news here, we have solved as well a quite difficult exercise left, the one at the end of the proof of Theorem 2.22. □

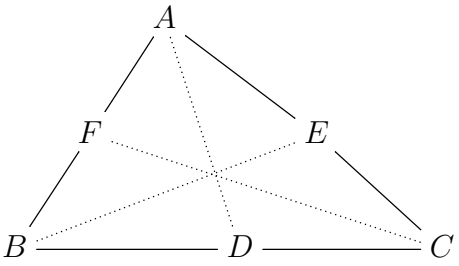
Many other things can be said, as a continuation of the above, notably with the general Fermat equation, which is as follows, involving an arbitary exponent $p\in\mathbb{N}$:

$$a^p+b^p=c^p$$

Also, the Pythagoras theorem has many applications. We will be back to this.

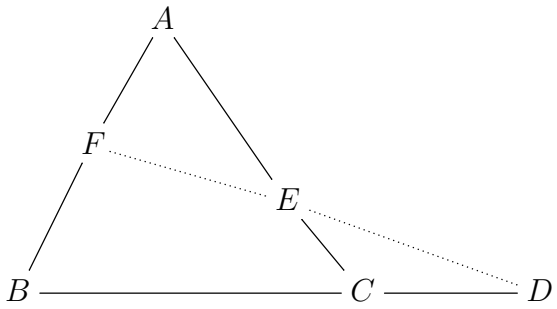
2d. Advanced results

Let us go back now to basic triangle geometry and centers, as developed in the beginning of this chapter. In order to further build on that material, we need to answer Question 2.12, asking for general crossing results, of the following type:



We will discusss this slowly, with several results on this subject, and on related topics. First on our list we have the following key result, due to Menelaus:

THEOREM 2.24 (Menelaus). *In a configuration of the following type, with a triangle ABC cut by a line FED,*



we have the following formula, with all segments being taken oriented:

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

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

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

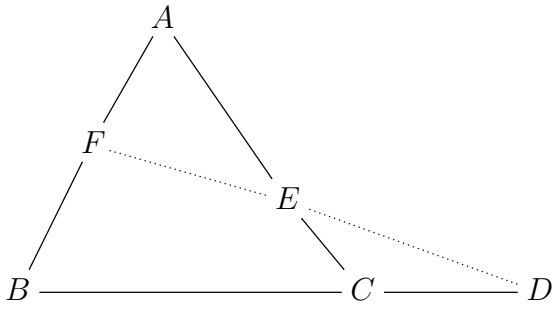
(1) Let us first try to prove the following equality, which is a bit weaker than what the theorem says, with all segments being by definition taken oriented:

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

But this is something clear, because by projecting the vertices A, B, C on the line DEF, into points A', B', C', we have the following computation:

$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=\frac{AA'}{BB'}\cdot\frac{BB'}{CC'}\cdot\frac{CC'}{AA'}=1$$

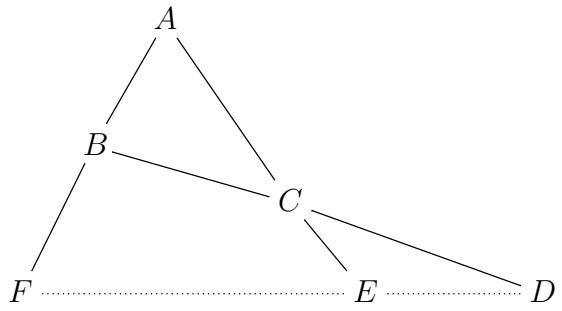
(2) Next, we must see what happens to the above equality, when allowing the segments to be oriented. But here, there are several cases to be considered, depending on whether the line DEF intersects the triangle ABC, a bit as in the picture in the statement, or not. Let us first examine the crossing configuration, as in the statement, namely:



In this case, with all the segments being by definition taken oriented, we are led indeed to the formula in the statement, as follows:

$$\begin{aligned}\frac{AF}{FB}\cdot\frac{BD}{DC}\frac{CE}{EA} &= \frac{|AF|}{|FB|}\left(-\frac{|BD|}{|DC|}\right)\cdot\frac{CE}{EA}\\ &= -\frac{|AF|}{|FB|}\cdot\frac{|BD|}{|DC|}\cdot\frac{|CE|}{|EA|}\\ &= -1\end{aligned}$$

(3) Let us examine now the non-crossing configuration, which is as follows:



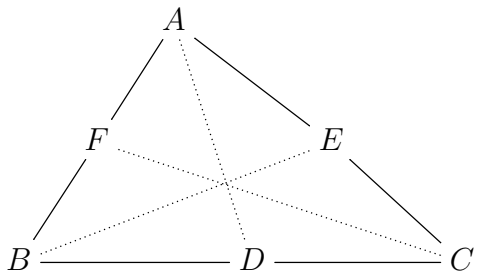
In this case, again with all the segments being by definition taken oriented, we are again led to the formula in the statement, as follows:

$$\begin{aligned}\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA} &= \left(-\frac{|AF|}{|FB|}\right)\left(-\frac{|BD|}{|DC|}\right)\left(-\frac{CE}{EA}\right)\\ &= -\frac{|AF|}{|FB|}\cdot\frac{|BD|}{|DC|}\cdot\frac{|CE|}{|EA|}\\ &= -1\end{aligned}$$

(4) Thus, we have proved the formula in the statement. As for the converse, this follows from the main result, in the obvious way, and as usual with converses of such statements, we will leave the discussion here as an instructive exercise for you. □

We can now answer our original question about crossing lines inside a triangle, drawn from the vertices, with the following remarkable result, due to Ceva:

THEOREM 2.25 (Ceva). *In a configuration of the following type, with a triangle ABC containing inner lines AD, BE, CF which cross,*



we have the following formula:

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

Moreover, the converse holds, with this formula guaranteeing that AD, BE, CF cross.

PROOF. This is indeed something very standard again, the idea being as follows:

(1) A first way of proving this result is by using the Menelaus theorem, applied twice. Indeed, if we denote by O the point in the middle in the above picture, we have the following formula, coming from the line COF cutting the triangle ABD:

$$\frac{AF}{FB}\cdot\frac{BC}{CD}\cdot\frac{DO}{OA}=-1$$

On the other hand, again by using the Menelaus theorem, we have as well the following formula, coming this time from the line BEO cutting the triangle ADC:

$$\frac{AO}{OD}\cdot\frac{DB}{BC}\cdot\frac{CE}{EA}=-1$$

By multiplying now the above two formulae, we obtain, as desired:

$$\begin{aligned}1&=\frac{AF}{FB}\cdot\frac{BC}{CD}\cdot\frac{DO}{OA}\times\frac{AO}{OD}\cdot\frac{DB}{BC}\cdot\frac{CE}{EA}\\&=\frac{AF}{FB}\cdot\frac{BC}{CD}\times\frac{DB}{BC}\cdot\frac{CE}{EA}\\&=\frac{AF}{FB}\cdot\frac{BC}{DC}\times\frac{BD}{BC}\cdot\frac{CE}{EA}\\&=\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}\end{aligned}$$

(2) An alternative proof, which is more elegant, is by using the same idea as for Menelaus, namely some fractions which cancel. Again by denoting by O the point in the middle, we have the following formulae for the quotient AF/FB, in terms of areas:

$$\frac{AF}{FB}=\frac{AFO}{FBO}=\frac{AFC}{FBC}$$

We deduce from this that we have the following extra formula for AF/FB:

$$\frac{AF}{FB}=\frac{AFC-AFO}{FBC-FBO}=\frac{AOC}{BOC}$$

Similarly, we have the following formulae for BD/DC, and for CE/EA:

$$\frac{BD}{DC}=\frac{AOB}{AOC}\quad,\quad\frac{CE}{EA}=\frac{BOC}{AOB}$$

Now by multiplying all these formulae we obtain, as desired:

$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=\frac{AOC}{BOC}\cdot\frac{AOB}{AOC}\cdot\frac{BOC}{AOB}=1$$

(3) As for the converse, this follows from the main result, in the obvious way, and as usual with such converses, we will leave the discussion here as an exercise. □

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

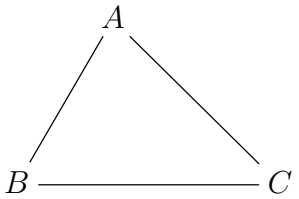
$$\frac{AF}{FB}\cdot\frac{BD}{DC}\cdot\frac{CE}{EA}=1\times1\times1=1$$

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

As further applications of the Ceva theorem, we can try to reprove the incenter and orthocenter theorems too. However, this is something more tricky, involving a bit of trigonometry, and we will defer the discussion here, for the next chapter.

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

THEOREM 2.26. *Besides the* 4 *main centers of a triangle, discussed in the above, many more remarkable points can be associated to a triangle* ABC*,*



and most of these lie on a line, called Euler line of ABC*. In particular, the barycenter* G*, the circumcenter* O *and the orthocenter* H *lie on this line, and* $GH=2GO$*.*

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

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

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

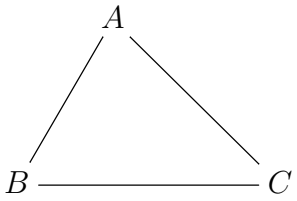
(3) At a more advanced level now, as indicated in the statement, it is possible to construct other interesting centers of a triangle, which usually lie on the Euler line. We will be back to this in the next theorem, when discussing the nine-point circle.

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

(5) So, this was for the idea, and in practice, although proofs for what is said above can be worked out, we would rather prefer to defer the discussion here for later, when we will have more advanced tools, and more specifically vectors, for dealing with this. □

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

THEOREM 2.27. *Associated to any triangle ABC,*



we have a nine-point circle, passing through the following points:

(1) *The midpoints of each side.*
(2) *The feet of each altitude.*
(3) *The midpoint of each segment vertex - orthocenter.*

Moreover, the center of this circle lies on the Euler line, midway between H and O.

PROOF. Again, this is something more technical, which can be proved as well, and we will leave working out the details here as an instructive, advanced exercise. We will be back to this later in this book, with more powerful technology, namely vectors. □

Quite nice all this, with as a philosophical conclusion, any triangle ABC not coming exactly alone, but rather accompanied by an extra line, and by a circle too. Thus, and anticipating here a bit, what we have is a configuration of total degree 6. And we will see later in this book other magical occurences of degree 6 configurations.

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

2e. Exercises

Exercises:

Exercise 2.28.

Exercise 2.29.

Exercise 2.30.

Exercise 2.31.

Exercise 2.32.

Exercise 2.33.

Exercise 2.34.

Exercise 2.35.

Bonus exercise.

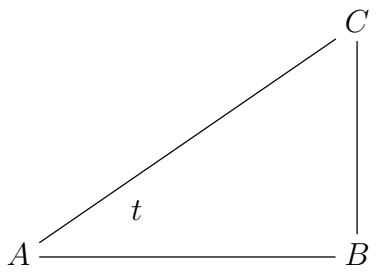
CHAPTER 3

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. *Given a right triangle ABC,*



we define the sine and cosine of the angle at A, denoted t, by the following formulae:

$$\sin t = \frac{BC}{AC} \quad , \quad \cos t = \frac{AB}{AC}$$

We call the sine and cosine basic trigonometric functions.

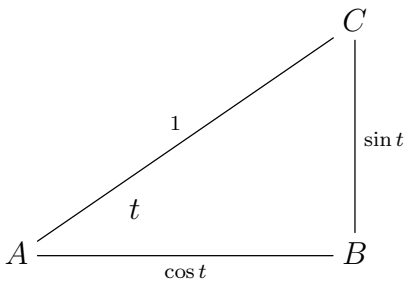
As a first observation, the sine and cosine do not depend on the choice of the given right triangle ABC having an angle t at A, and this due to the Thales theorem. In view of this, whenever possible, we will choose the right triangle ABC as to have:

$$AC = 1$$

In this case, the formulae defining the sine and cosine simplify, as follows:

$$\sin t = BC \quad , \quad \cos t = AB$$

Equivalently, we can encode all this in a single picture, as follows:



As a few basic examples now, for the sine, coming from things that we know well, about right triangles, from the previous chapter, we have:

$$\sin 0^\circ = 0 \quad , \quad \sin 30^\circ = \frac{1}{2} \quad , \quad \sin 45^\circ = \frac{1}{\sqrt{2}} \quad , \quad \sin 60^\circ = \frac{\sqrt{3}}{2} \quad , \quad \sin 90^\circ = 1$$

Let us record as well the list of corresponding cosines. These are as follows:

$$\cos 0^\circ = 1 \quad , \quad \cos 30^\circ = \frac{\sqrt{3}}{2} \quad , \quad \cos 45^\circ = \frac{1}{\sqrt{2}} \quad , \quad \cos 60^\circ = \frac{1}{2} \quad , \quad \cos 90^\circ = 0$$

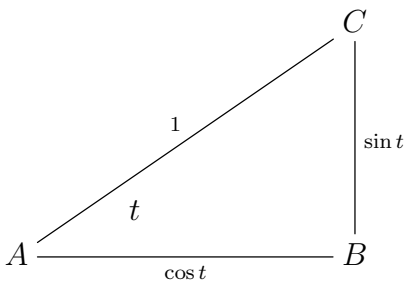
Observe that the numbers in the above two lists are the same, but written backwards in the second list. In fact, we have the following result, regarding this:

THEOREM 3.2. *The sines and cosines are subject to the formulae*

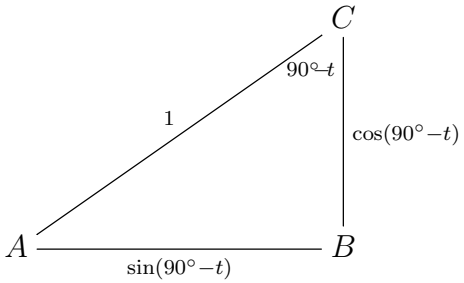
$$\sin(90^\circ - t) = \cos t \quad , \quad \cos(90^\circ - t) = \sin t$$

valid for any angle $t \in [0^\circ, 90^\circ]$.

PROOF. In order to understand this, the best is to choose our right triangle ABC with $AC = 1$. In this case, the picture coming from Definition 3.1 is as follows:



On the other hand, by focusing now at the angle at C, and perhaps twisting a bit our minds too, we have as well the following picture, for the same triangle:



Thus, we are led to the conclusion in the statement, and by the way congratulations, with this being our first trigonometry theorem. Many more to come. □

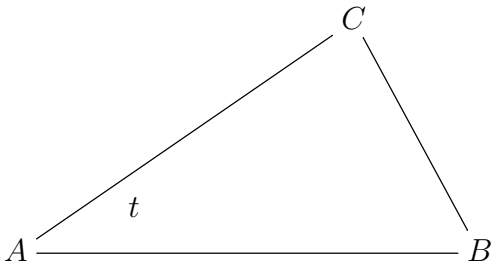
Before going ahead with more trigonometry, all sorts of properties of the sine and cosine that we can surely work out, with bravery, a question that you might have:

QUESTION 3.3. *Why bothering with sine and cosine?*

In answer, good question indeed, and you won't believe me, but when writing this book, at this very point that we are now, I asked this myself too, and could not find any simple answer. So, I went into a tour of my Mathematics Department, here at Cergy, desperately asking colleagues about this, with some of them being actually world class geometers, but no one knew the answer to this question either.

So, what do to. And here, you guessed it right, go back home at full speed, using various driving techniques that I learned as a youngster, in the Bucharest of the early 1990s, good times back then, and ask the cat. And cat looked at me, and declared:

CAT 3.4. *The area of an arbitrary triangle, having an angle t at A,*

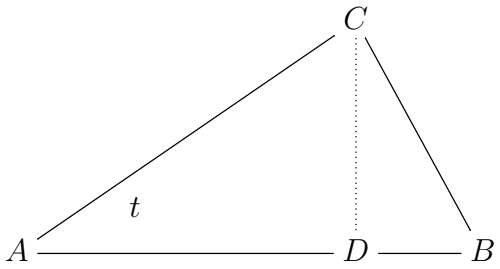


is given by the following formula, making appear the sine:

$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

As for the need for cosines, homework for you buddy.

Thanks cat, interesting all this, let us try to understand it. To start with, the formula of cat looks like some sort of mathematical theorem, that we must prove. But, in order to do so, the simplest is to draw an altitude of our triangle, as follows:



Now with this altitude drawn, we have the following computation:

$$\begin{aligned} area(ABC) &= \frac{basis \times height}{2} \\ &= \frac{AB \times CD}{2} \\ &= \frac{AB \times AC \times \sin t}{2} \end{aligned}$$

Thus, theorem proved, so the sine is definitely a good and useful thing, as cat says. As for the cosine, damn cat has assigned this to us as an exercise, so we will have to think about it, and come back to it, in due time. And no late homework, of course.

Moving forward now, still in relation with Cat 3.4, we have the following question:

QUESTION 3.5. *What happens to the cat formula,*

$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

when the angle at A is obtuse, $t > 90^\circ$?

Which looks like a very good question. In answer now, given a triangle which is obtuse at A, we can simply rotate the AC side to the right, as for that obtuse angle to become acute, $t' = 180^\circ - t$, and the area of the triangle obviously remains the same, and this since both the basis and height remain unchanged. Thus, the correct definition for $\sin t$ for obtuse angles should be the one making the following formula work:

$$\frac{AB \times AC \times \sin t}{2} = \frac{AB \times AC \times \sin(180^\circ - t)}{2}$$

Now by simplifying, we are led to the following formula:

$$\sin t = \sin(180^\circ - t)$$

Thus, Question 3.5 answered, with our conclusions being as follows:

THEOREM 3.6. *We can talk about the sine of any angle $t \in [0°, 180°]$, according to*

$$\sin t = \sin(180° - t)$$

and with this, the cat formula for the area of a triangle, namely

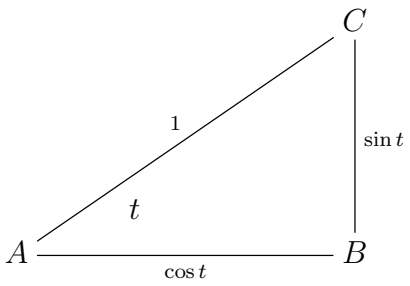
$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

holds for any triangle, without any assumption on it.

PROOF. This follows indeed from the above discussion. □

Moving ahead now, defining sines as in Definition 3.1 for $t \in [0°, 90°]$, and as above for $t \in [90°, 180°]$ certainly does the job, as explained above, but is not very elegant. So, let us try to improve this. We have here the following obvious speculation:

SPECULATION 3.7. *The sine of any angle $t \in [0°, 180°]$ can be defined geometrically, according to the usual picture*



with the convention that for $t > 90°$, the triangle is drawn at the left of A.

Which sounds quite good, but when thinking some more, things fine of course with the sine, but what about the cosine? The problem indeed is that, in the case $t > 90°$, when the triangle is drawn at the left of A, the lower side AB changes orientation:

$$AB \to BA$$

But, as we know well from chapter 2, from various considerations regarding segments and orientation, this would amount in saying that we are replacing:

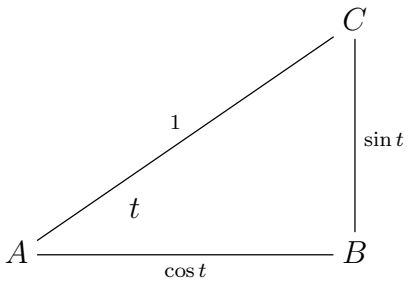
$$AB \to -AB$$

And so, we are led to the following formula for the cosine, in this case:

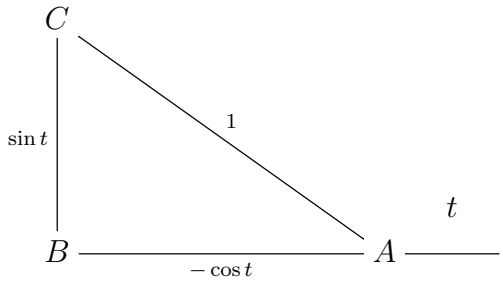
$$\cos t = -\cos(180° - t)$$

Very good all this, so let us update now Theorem 3.6, and by incorporating as well Speculation 3.7, in the form of a grand result, in the following way:

THEOREM 3.8 (update). *We can talk about the sine and cosine of any angle $t \in [0°, 180°]$, according to the following picture,*



which in the case of obtuse angles becomes by definition as follows,



and with this, we have the following formulae, valid for any $t \in [0°, 180°]$:

$$\sin t = \sin(180° - t) \quad , \quad \cos t = -\cos(180° - t)$$

Moreover, the cat formula for the area of a triangle, namely

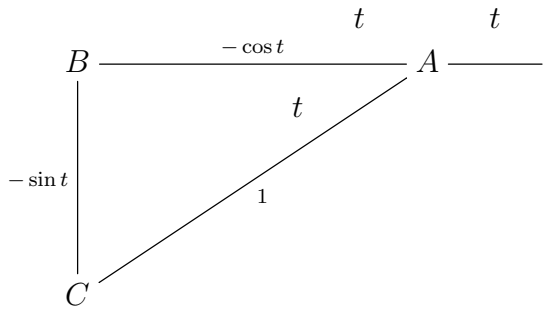
$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

holds for any triangle, without any assumption on it.

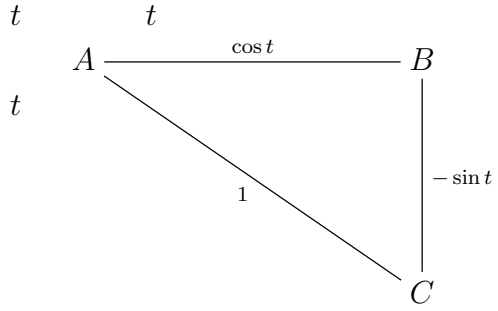
PROOF. This follows indeed by putting together all the above. □

Which sounds quite good, and normally end of the story, but let us be crazy now, and try to talk as well about the sine or cosine of angles $t < 0°$, or $t > 180°$.

Indeed, we know the recipe, namely suitably drawing our right triangle, with attention to positive and negatives. Thus, for $t \in [180°, 270°]$, our picture should be as follows:



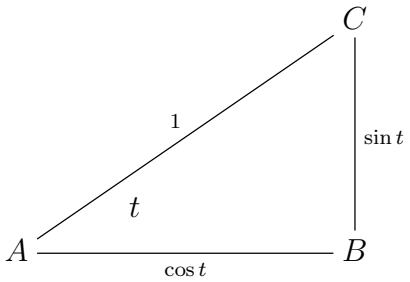
As for the next case, $t \in [270°, 360°]$, here our picture should be as follows:



But with this, we are done, because adding or substracting 360° to our angles won't change the corresponding right triangle, and so won't change the sine and cosine.

Hope you're still with me, after all these wild speculations. Good work that we did, and time now to further improve Theorem 3.8, into something really final, as follows:

THEOREM 3.9 (final update). *We can talk about the sine and cosine of any angle $t \in \mathbb{R}$, according to the following picture,*



suitably drawn for angles $t < 0°$, or $t > 90°$, with attention to positive and negative lengths, as explained above. With this, all the basic formulae still hold, for any $t \in \mathbb{R}$.

PROOF. This follows indeed by putting together all the above, and with the basic formulae in question being as follows, and in the hope that I forgot none:

$$\sin(-t) = -\sin t \quad , \quad \cos(-t) = \cos t$$

$$\sin(90^\circ - t) = \cos t \quad , \quad \cos(90^\circ - t) = \sin t$$

$$\sin(90^\circ + t) = \cos t \quad , \quad \cos(90^\circ + t) = -\sin t$$

$$\sin(180^\circ - t) = \sin t \quad , \quad \cos(180^\circ - t) = -\cos t$$

$$\sin(180^\circ + t) = -\sin t \quad , \quad \cos(180^\circ + t) = -\cos t$$

$$\sin(270^\circ - t) = -\cos t \quad , \quad \cos(270^\circ - t) = -\sin t$$

$$\sin(270^\circ + t) = -\cos t \quad , \quad \cos(270^\circ + t) = \sin t$$

$$\sin(360^\circ - t) = -\sin t \quad , \quad \cos(360^\circ - t) = \cos t$$

$$\sin(360^\circ + t) = \sin t \quad , \quad \cos(360^\circ + t) = \cos t$$

Plus of course, not to forget about this, and thanks cat for meowing and reminding me this, the cat formula for the area of a triangle, which was as follows:

$$area(ABC) = \frac{AB \times AC \times \sin t}{2}$$

Here actually some discussion is needed, in relation with positives and negatives, and we will leave this as an instructive exercise for you, reader. □

And with this, good news, done with definitions and other preliminary material, at least for our present purposes, in this opening chapter on trigonometry.

And job for us now, to study sin and cos defined on $\mathbb{R}$, say motivated by the above cat formula, for the area of an arbitrary triangle.

3b. Pythagoras, again

In order to study sin and cos, let us first update the numerics that we already have, for very simple angles in $[0^\circ, 90^\circ]$, to more angles, in $[0^\circ, 360^\circ]$.

We have here the following statement, which is straightforward:

THEOREM 3.10. *The sines of the basic angles are as follows,*

$$\sin 0^\circ = 0 \quad , \quad \sin 30^\circ = \frac{1}{2} \quad , \quad \sin 45^\circ = \frac{1}{\sqrt{2}} \quad , \quad \sin 60^\circ = \frac{\sqrt{3}}{2} \quad , \quad \sin 90^\circ = 1$$

$$\sin 120^\circ = \frac{\sqrt{3}}{2} \quad , \quad \sin 135^\circ = \frac{1}{\sqrt{2}} \quad , \quad \sin 150^\circ = \frac{1}{2} \quad , \quad \sin 180^\circ = 0$$

$$\sin 210^\circ = -\frac{1}{2} \quad , \quad \sin 225^\circ = -\frac{1}{\sqrt{2}} \quad , \quad \sin 240^\circ = -\frac{\sqrt{3}}{2} \quad , \quad \sin 270^\circ = -1$$

$$\sin 300^\circ = -\frac{\sqrt{3}}{2} \quad , \quad \sin 315^\circ = -\frac{1}{\sqrt{2}} \quad , \quad \sin 330^\circ = -\frac{1}{2} \quad , \quad \sin 360^\circ = 0$$

and the cosines of the basic angles are as follows,

$$\cos 0^\circ = 1 \quad , \quad \cos 30^\circ = \frac{\sqrt{3}}{2} \quad , \quad \cos 45^\circ = \frac{1}{\sqrt{2}} \quad , \quad \cos 60^\circ = \frac{1}{2} \quad , \quad \cos 90^\circ = 0$$

$$\cos 120^\circ = -\frac{1}{2} \quad , \quad \cos 135^\circ = -\frac{1}{\sqrt{2}} \quad , \quad \cos 150^\circ = -\frac{\sqrt{3}}{2} \quad , \quad \cos 180^\circ = -1$$

$$\cos 210^\circ = -\frac{\sqrt{3}}{2} \quad , \quad \cos 225^\circ = -\frac{1}{\sqrt{2}} \quad , \quad \cos 240^\circ = -\frac{1}{2} \quad , \quad \cos 270^\circ = 0$$

$$\cos 300^\circ = \frac{1}{2} \quad , \quad \cos 315^\circ = \frac{1}{\sqrt{2}} \quad , \quad \cos 330^\circ = \frac{\sqrt{3}}{2} \quad , \quad \cos 360^\circ = 1$$

with this coming from the basic geometry of right triangles.

PROOF. This is indeed self-explanatory, with input coming from chapter 2. □

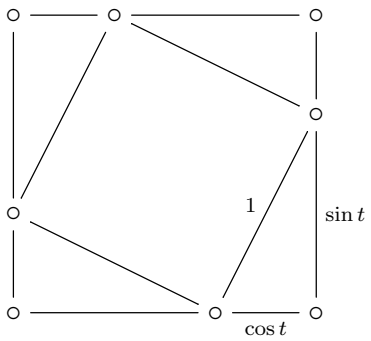
The problem is now, how to get beyond the above formulae? Not an easy question, but do not worry, we will be back to this, in due time. For the moment, as a complement to the above, let us record the following key formula, coming from Pythagoras:

THEOREM 3.11. *The sines and cosines are subject to the formula*

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

coming from Pythagoras' theorem.

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



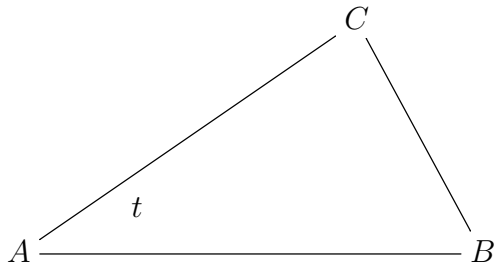
When computing the area of the outer square, we obtain:

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

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

Next, with our knowledge of the sine and cosine, we can now formulate a technical generalization of the Pythagoras theorem, in the following way:

THEOREM 3.12. *Given an arbitrary triangle, as follows,*

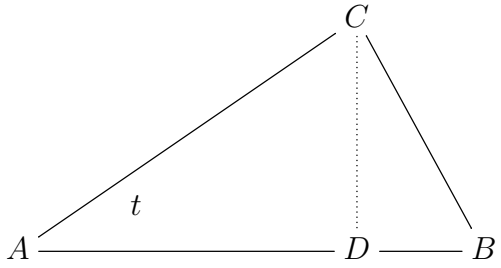


the length of the side which is away from the vertex A is given by the formula

$$BC^2=AB^2+AC^2-2AB\cdot AC\cdot\cos t$$

called law of cosines, and with this generalizing Pythagoras.

PROOF. Let us draw indeed an altitude of our triangle, as follows:



We have then the following computation, coming from Pythagoras, applied twice:

$$\begin{aligned} BC^2 &= CD^2 + BD^2 \\ &= CD^2 + (AB - AD)^2 \\ &= CD^2 + AB^2 + AD^2 - 2AB \cdot AD \\ &= AB^2 + AC^2 - 2AB \cdot AD \\ &= AB^2 + AC^2 - 2AB \cdot AC \cdot \cos t \end{aligned}$$

Finally, the last assertion is clear, because with $\cos t = 0$ we obtain Pythagoras. □

The above result looks quite interesting, for engineering purposes, and we have:

CONCLUSION 3.13. *The law of cosines found above can be effectively used for making money, by computing distances BC over wild land, for various interested customers. In fact, financially speaking, the law of cosines might be well more interesting than the cat formula for the area of triangles, involving the sines.*

Which might sound quite interesting, for us humans, but my cat, who is not into making money, seems unfazed. In fact, here is what he has to say, about this:

CAT 3.14. *That law of cosines is ugly, and no match for my law of sines:*

$$area(ABC) = \frac{AB \cdot AC \cdot \sin t}{2}$$

I would suggest you humans to look into the quantity

$$< AB, AC >= AB \cdot AC \cdot \cos t$$

in order to understand what the cosines are good for. And change your diet, too.

Quite interesting all this, but in practice, this $< AB, AC >$ quantity does not seem to be something very intuitive, at least to my human brain. We will leave this for later.

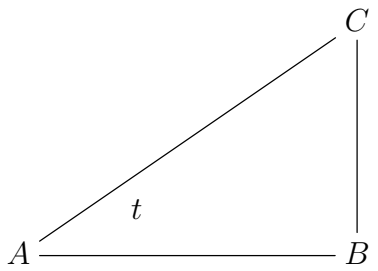
Back now to the basics, it is possible to say many more things about angles and $\sin x$, $\cos x$, and also talk about some supplementary quantities, such as the tangent:

DEFINITION 3.15. *We can talk about the tangent of angles $t \in \mathbb{R}$, as being given by*

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

with $\sin x, \cos x$ being defined as before.

In more geometric terms, consider an arbitrary right triangle, as follows:



We have then the following computation, for the tangent of t:

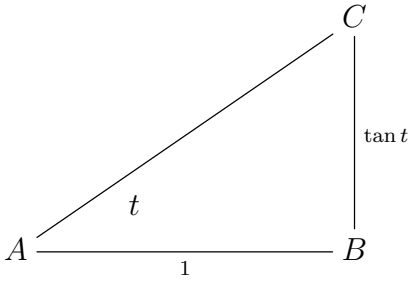
$$\tan t = \frac{\sin t}{\cos t} = \frac{BC}{AC} \Big/ \frac{AB}{AC} = \frac{BC}{AB}$$

Thus, the tangent defined above complements the sine and cosine, because we have:

$$\sin t = \frac{BC}{AC} \quad , \quad \cos t = \frac{AB}{AC} \quad , \quad \tan t = \frac{BC}{AB}$$

A similar interpretation works for obtuse right triangles, and even for right triangles with an arbitrary angle $t \in \mathbb{R}$, and we can formulate, in the spirit of Theorem 3.9:

THEOREM 3.16. *We can talk, geometrically, about the tangent of any angle $t \in \mathbb{R}$, according to the following picture,*



suitably drawn for angles $t < 0°$, or $t > 90°$, with attention to positive and negative lengths, as explained above. With this, all the basic formulae still hold, for any $t \in \mathbb{R}$.

PROOF. Here the first assertion follows by reasoning as in the proof of Theorem 3.9, or simply follows from Theorem 3.9 itself. As for the second assertion, the basic formulae for the tangent, all coming from what we know, are as follows:

$$\tan(-t) = -\tan t$$

$$\tan(90° - t) = \frac{1}{\tan t} \quad , \quad \cos(90° + t) = -\frac{1}{\tan t}$$

$$\tan(180° - t) = -\tan t \quad , \quad \tan(180° + t) = \tan t$$

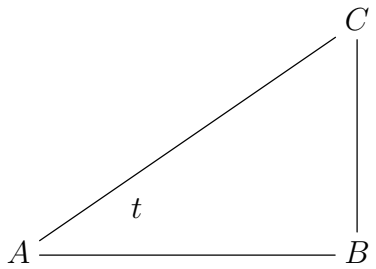
Let us record as well the formulae for the basic angles. These are as follows:

$$\tan 0^\circ = 0 \quad , \quad \tan 30^\circ = \frac{1}{\sqrt{3}} \quad , \quad \sin 45^\circ = 1 \quad , \quad \sin 60^\circ = \sqrt{3}$$

$$\tan 120^\circ = -\sqrt{3} \quad , \quad \tan 135^\circ = -1 \quad , \quad \tan 150^\circ = -\frac{1}{\sqrt{3}} \quad , \quad \tan 180^\circ = 0$$

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

Very nice all this, but are we really done with generalities and definitions? Not yet, because, let us go back to our basic right triangle, with an angle t, as follows:



We know from the above that we have the following formulae:

$$\sin t = \frac{BC}{AC} \quad , \quad \cos t = \frac{AB}{AC} \quad , \quad \tan t = \frac{BC}{AB}$$

However, there are still 3 fractions left, in need of a name, so let us formulate:

DEFINITION 3.17. *We can talk about the secant, cosecant and cotangent, as being*

$$\sec t = \frac{AC}{BC} \quad , \quad \csc t = \frac{AC}{AB} \quad , \quad \cot t = \frac{BC}{AB}$$

in the context of a right triangle, as above, or equivalently, as being

$$\sec t = \frac{1}{\sin t} \quad , \quad \csc t = \frac{1}{\cos t} \quad , \quad \cot t = \frac{1}{\tan t}$$

in terms of the standard trigonometric functions $\sin, \cos, \tan$.

Very nice all this, so shall we study now all these new functions too, in the spirit of what we did in the aboves for $\sin, \cos, \tan$, after all we can potentially have some mathematical fun, with our enlarged collection of 6 trigonometric functions, which looks complete, symmetric and beautiful. Not clear, so time to ask the cat. And cat says:

CAT 3.18. *These* $\sec, \csc, \cot$ *functions sound more like pure mathematics. What about trying instead* $\arcsin, \arccos, \arctan$*? Or* $\sinh, \cosh, \tanh$*? Or* $\text{arcsinh}, \text{arccosh}, \text{arctanh}$*?*

Which sounds quite interesting, not that I fully understand what cat says, but one thing is sure, namely that we won't potentially get to any new mathematics by applying $x \to x^{-1}$, so I kind of agree with the first advice, forget about $\sec, \csc, \cot$.

As for the rest, yes I have this feeling too that many more interesting trigonometric functions are waiting for us, and more on this later in this book, once we will have some appropriate tools, beyond basic triangle geometry, in order to discuss them.

Thus being said, wait. Remember the discussion following the Ceva theorem, from the previous chapter? We had some unfinished business there, in what regards the applications, and we promised to get back to this, once we know some trigonometry. So, time to do this, and as a surprise, we get into something involving secants and cotangents:

THEOREM 3.19. *The barycenter, incenter and orthocenter theorems can be all deduced from the Ceva theorem, with the computations being respectively as follows,*

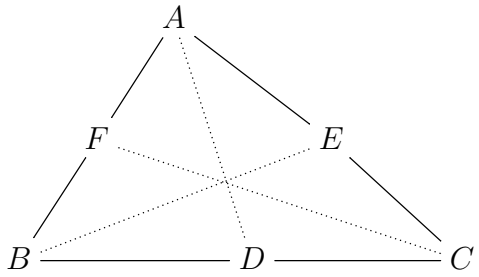
$$1 \times 1 \times 1 = 1$$

$$\frac{\sec A}{\sec B} \cdot \frac{\sec B}{\sec C} \cdot \frac{\sec C}{\sec A} = 1$$

$$\frac{\cot A}{\cot B} \cdot \frac{\cot B}{\cot C} \cdot \frac{\cot C}{\cot A} = 1$$

with A, B, C being the angles of our triangle.

PROOF. Let us first recall from chapter 2 that the Ceva theorem concerns a configuration as follows, with a triangle ABC containing inner lines AD, BE, CF:



The theorem states that AD, BE, CF cross precisely when the following happens:

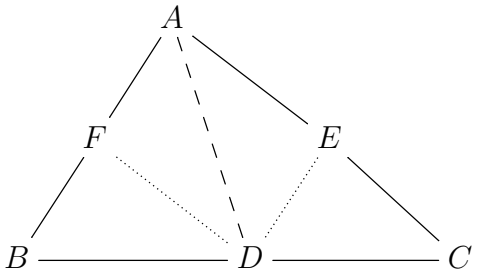
$$\frac{AF}{FB} \cdot \frac{BD}{DC} \cdot \frac{CE}{EA} = 1$$

Regarding now the barycenter, incenter and orthocenter, the situation is as follows:

(1) In what regards the barycenter, the computation is trivial, as follows:

$$1 \times 1 \times 1 = 1$$

(2) In order to deal now with the incenter, consider indeed a triangle, with an angle bisector drawn, and with two perpendiculars drawn as well, as indicated:



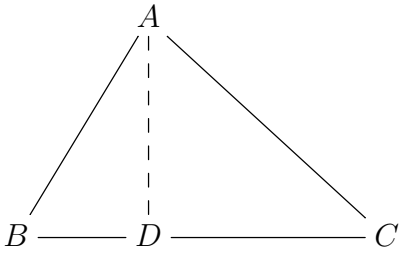
We have then the following computation, using $FD = DE$:

$$\frac{BD}{DC} = \frac{FD \sec B}{DE \sec C} = \frac{\sec B}{\sec C}$$

We conclude that Ceva gives indeed the incenter, with the computation being:

$$\frac{\sec A}{\sec B} \cdot \frac{\sec B}{\sec C} \cdot \frac{\sec C}{\sec A} = 1$$

(3) Finally, in order to deal now with the orthocenter, a bit in a similar way, consider indeed a triangle, with an altitude drawn, as follows:



We have then the following computation, coming from definitions:

$$\frac{BD}{DC} = \frac{BD}{AD} \Big/ \frac{DC}{AD} = \frac{\cot B}{\cot C}$$

Thus Ceva gives as well the orthocenter, with the computation being as follows:

$$\frac{\cot A}{\cot B} \cdot \frac{\cot B}{\cot C} \cdot \frac{\cot C}{\cot A} = 1$$

And with this being something nice, remember the mess with the orthocenter when first proving the theorem, with that trick involved. Gone all that. □

The above is quite nice, and as a byproduct, we seem to have a contradiction here with Cat 3.18. But cat is gone, mumbling something about cosecants.

3c. Sums, duplication

Getting back now to the basics, sine and cosine, how these can be computed, and what can be done with them, we have the following key result:

THEOREM 3.20. *The sines and cosines of sums are given by*

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

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

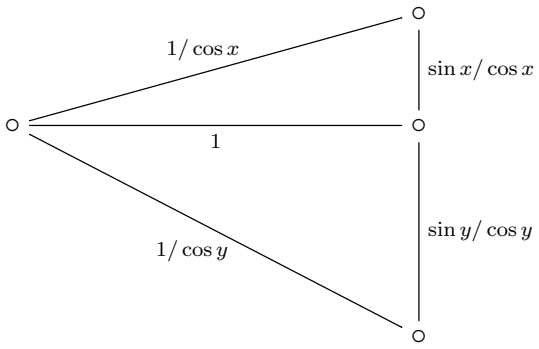
and these formulae give a formula for the tangent too, namely

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

provided of course that the denominator is nonzero.

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

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



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

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

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

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

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

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

$$\begin{aligned}\tan(x+y) &= \frac{\sin x\cos y+\cos x\sin y}{\cos x\cos y-\sin x\sin y}\\ &= \frac{\sin x\cos y/\cos x\cos y+\cos x\sin y/\cos x\cos y}{1-\sin x\sin y/\cos x\cos y}\\ &= \frac{\tan x+\tan y}{1-\tan x\tan y}\end{aligned}$$

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

The above theorem is something very useful, in practice, so let us record as well what happens when replacing sums by substractions. The formulae here are as follows:

THEOREM 3.21. *The sines and cosines of differences are given by*

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

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

and these formulae give a formula for the tangent too, namely

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

provided of course that the denominator is nonzero.

PROOF. These are all consequences of what we have in Theorem 3.20, as follows:

(1) Regarding the sine, we have here the following computation:

$$\begin{aligned}\sin(x-y) &= \sin x\cos(-y)+\cos x\sin(-y)\\ &= \sin x\cos y-\cos x\sin y\end{aligned}$$

(2) Regarding the cosine, the computation here is similar, as follows:

$$\begin{aligned}\cos(x-y) &= \cos x\cos(-y)-\sin x\sin(-y)\\ &= \cos x\cos y+\sin x\sin y\end{aligned}$$

(3) Finally, in what regards the tangent, I would not mess with it, and say instead:

$$\begin{aligned}\tan(x-y) &= \frac{\sin x\cos y-\cos x\sin y}{\cos x\cos y+\sin x\sin y}\\ &= \frac{\sin x\cos y/\cos x\cos y-\cos x\sin y/\cos x\cos y}{1+\sin x\sin y/\cos x\cos y}\\ &= \frac{\tan x-\tan y}{1+\tan x\tan y}\end{aligned}$$

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

As illustrations for the above formulae, we can now compute the sine, cosine and tangent of various interesting new angles, appearing as sums and differences, such as:

$$15^\circ=45^\circ-30^\circ$$

$$75^\circ=45^\circ+30^\circ$$

In fact, thinking well, this is pretty much it, modulo periodicity formulae. So, all in all, we can deal now with all multiples of 15°. Let us record our result here, as follows:

THEOREM 3.22. *The sine, cosine and tangent of multiples of* 15° *are given by*

$$\sin 15^\circ=\frac{\sqrt{3}-1}{2\sqrt{2}}\quad,\quad \sin 30^\circ=\frac{1}{2}\quad,\quad \sin 45^\circ=\frac{1}{\sqrt{2}}\quad,\quad \sin 60^\circ=\frac{\sqrt{3}}{2}\quad,\quad \sin 75^\circ=\frac{\sqrt{3}+1}{2\sqrt{2}}$$

$$\cos 15^\circ=\frac{\sqrt{3}+1}{2\sqrt{2}}\quad,\quad \cos 30^\circ=\frac{\sqrt{3}}{2}\quad,\quad \cos 45^\circ=\frac{1}{\sqrt{2}}\quad,\quad \cos 60^\circ=\frac{1}{2}\quad,\quad \cos 75^\circ=\frac{\sqrt{3}-1}{2\sqrt{2}}$$

$$\tan 15^\circ=\frac{\sqrt{3}-1}{\sqrt{3}+1}\quad,\quad \tan 30^\circ=\frac{1}{\sqrt{3}}\quad,\quad \tan 45^\circ=1\quad,\quad \tan 60^\circ=\sqrt{3}\quad,\quad \tan 75^\circ=\frac{\sqrt{3}+1}{\sqrt{3}-1}$$

plus $\sin 0^\circ=0$*, and various periodicity formulae.*

PROOF. For the quantity $\sin 15^\circ=\cos 75^\circ$, we have the following computation:

$$\begin{aligned}\sin 15^\circ &= \sin(45^\circ-30^\circ)\\ &= \sin 45^\circ\cos 30^\circ-\cos 45^\circ\sin 30^\circ\\ &= \frac{1}{\sqrt{2}}\cdot\frac{\sqrt{3}}{2}-\frac{1}{\sqrt{2}}\cdot\frac{1}{2}\\ &= \frac{\sqrt{3}-1}{2\sqrt{2}}\end{aligned}$$

Also, for the quantity $\cos 15^\circ = \sin 75^\circ$, we have the following computation:

$$\begin{aligned}\sin 15^\circ &= \cos(45^\circ - 30^\circ)\\ &= \cos 45^\circ \cos 30^\circ + \sin 45^\circ \sin 30^\circ\\ &= \frac{1}{\sqrt{2}} \cdot \frac{\sqrt{3}}{2} + \frac{1}{\sqrt{2}} \cdot \frac{1}{2}\\ &= \frac{\sqrt{3}+1}{2\sqrt{2}}\end{aligned}$$

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

Time now for more advanced trigonometry. Indeed, by taking $x = y$ in Theorem 3.20 we obtain some interesting formulae for the duplication of angles, as follows:

THEOREM 3.23. *The sines of the doubles of angles are given by*

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

and the corresponding cosines are given by the following equivalent formulae,

$$\begin{aligned}\cos(2t) &= \cos^2 t - \sin^2 t\\ &= 2\cos^2 t - 1\\ &= 1 - 2\sin^2 t\end{aligned}$$

with all these three formulae being useful, in practice.

PROOF. By taking $x = y = t$ in the formulae from Theorem 3.20, we obtain:

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

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

As for the extra formulae for $\cos(2t)$, these follow by using $\cos^2 + \sin^2 = 1$. □

Let us record as well the formula for the tangents, which is as follows:

THEOREM 3.24. *The tangents of the doubles of angles are given by*

$$\tan(2t) = \frac{2\tan t}{1 - \tan^2 t}$$

provided as usual that the denominator is nonzero.

PROOF. This follows indeed by taking $x = y = t$ in the formula for tangents from Theorem 3.20. Equivalently, you can check, as an easy, instructive exercise, that this is indeed what we get, by dividing the sine and cosine computed in Theorem 3.22. □

The point now is that, with this, we can substantially improve our data from Theorem 3.22, by computing the cosines of the halves of the angles there, using the above formula for $\cos(2t)$, and then computing the sines of these angles too, by using Pythagoras, and finally by computing the tangents too, as quotients. As a result here, let us record:

THEOREM 3.25. *The sine, cosine and tangent of* 7.5° *are given by*

$$\sin 7.5^\circ = \sqrt{\frac{4-\sqrt{2}-\sqrt{6}}{8}}\ ,\ \cos 7.5^\circ = \sqrt{\frac{4+\sqrt{2}+\sqrt{6}}{8}}\ ,\ \tan 7.5^\circ = \sqrt{\frac{4-\sqrt{2}-\sqrt{6}}{4+\sqrt{2}+\sqrt{6}}}$$

and for the odd multiples of 7.5°*, we have similar formulae.*

PROOF. For the cosine we can use $\cos(2t) = 2\cos^2 t - 1$, and we obtain:

$$\begin{aligned}
\cos 7.5^\circ &= \sqrt{\frac{1+\cos 15^\circ}{2}}\\
&= \sqrt{\frac{1+\frac{1+\sqrt{3}}{2\sqrt{2}}}{2}}\\
&= \sqrt{\frac{2\sqrt{2}+1+\sqrt{3}}{4\sqrt{2}}}\\
&= \sqrt{\frac{4+\sqrt{2}+\sqrt{6}}{8}}
\end{aligned}$$

For the sine we can use Pythagoras, $\sin^2 + \cos^2 = 1$, and we obtain:

$$\begin{aligned}
\sin 7.5^\circ &= \sqrt{1-\cos^2 7.5^\circ}\\
&= \sqrt{1-\frac{4+\sqrt{2}+\sqrt{6}}{8}}\\
&= \sqrt{\frac{4-\sqrt{2}-\sqrt{6}}{8}}
\end{aligned}$$

Finally, by taking the quotient we obtain the formula for the tangent. As for the last assertion, it is clear that the same method will work for all multiples of 7.5°, with input from Theorem 3.22, and we will leave the computations here as an instructive exercise. □

As a conclusion to all this, we have quite mixed news, as follows:

(1) On one hand the formulae in Theorem 3.23 are definitely something powerful, allowing us in theory to indefinitely halve the angles that we know, and so to virtually obtain, via some limits if needed, all the sines and cosines in this world.

(2) On the other hand, in practice, all this leads us into the question of extracting square roots, which rather belongs to arithmetics. So, all in all, not that much of a total kill, Theorem 3.22 transferring our questions, from trigonometry to arithmetics.

3d. Higher formulae

We have seen that some interesting mathematics appears in relation with the sines and cosines of sums of angles, $x+y$. This suggests, as a continuation, summing 3 or more angles, and we will explore this here. To start with, we have the following result:

THEOREM 3.26. *The sines and cosines of sums of* 3 *angles are given by*

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

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

and we have a formula for the tangent too, namely

$$\tan(x+y+z) = \frac{\tan x+\tan y+\tan z-\tan x\tan y\tan z}{1-\tan x\tan y-\tan x\tan z-\tan y\tan z}$$

provided of course that the denominator is nonzero.

PROOF. We use the addition formulae from Theorem 3.20, namely:

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

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

In what regards the sine, the computation here is as follows:

$$\begin{aligned}
& \sin(x+y+z)\\
= \ & \sin x\cos(y+z) + \cos x\sin(y+z)\\
= \ & \sin x(\cos y\cos z - \sin y\sin z) + \cos x(\sin y\cos z + \cos y\sin z)\\
= \ & \sin x\cos y\cos z + \cos x\sin y\cos z + \cos x\cos y\sin z - \sin x\sin y\sin z
\end{aligned}$$

In what regards the cosine, the computation here is similar, as follows:

$$\begin{aligned}
& \cos(x+y+z)\\
= \ & \cos x\cos(y+z) - \sin x\sin(y+z)\\
= \ & \cos x(\cos y\cos z - \sin y\sin z) - \sin x(\sin y\cos z + \cos y\sin z)\\
= \ & \cos x\cos y\cos z - \cos x\sin y\sin z - \sin x\cos y\sin z - \sin x\sin y\cos z
\end{aligned}$$

Regarding now the tangent, this follows by taking the quotient. However, since the tangent function seems to be the winner, in all this, it is perhaps instructive to come up as well with a tangent-only proof. According to Theorem 3.20, we have:

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

By using this formula twice, we obtain, for a sum of three angles:

$$\begin{aligned}\tan(x+y+z) &= \frac{\tan x+\tan(y+z)}{1-\tan x\tan(y+z)}\\ &= \frac{\tan x+\frac{\tan y+\tan z}{1-\tan y\tan z}}{1-\tan x\frac{\tan y+\tan z}{1-\tan y\tan z}}\\ &= \frac{\tan x+\tan y+\tan z-\tan x\tan y\tan z}{1-\tan x\tan y-\tan x\tan z-\tan y\tan z}\end{aligned}$$

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

As a consequence of the above result, obtained with $x=y=z=t$, we have:

THEOREM 3.27. *The sines and cosines of sums of triple of angles are given by*

$$\sin(3t)=3\sin t-4\sin^3 t$$

$$\cos(3t)=4\cos^3 t-3\cos t$$

and we have a formula for the tangent too, namely

$$\tan(3t)=\frac{3\tan t-\tan^3 t}{1-3\tan^2 t}$$

provided of course that the denominator is nonzero.

PROOF. With $x=y=z=t$ in the sine formula from Theorem 3.26, we obtain:

$$\begin{aligned}\sin(3t) &= 3\sin t\cos^2 t-\sin^3 t\\ &= 3\sin t(1-\sin^2 t)-\sin^3 t\\ &= 3\sin t-4\sin^3 t\end{aligned}$$

Similarly, with $x=y=z=t$ in the cosine formula from Theorem 3.26, we obtain:

$$\begin{aligned}\cos(3t) &= \cos^3 t-3\cos t\sin^2 t\\ &= \cos^3 t-3\cos t(1-\cos^2 t)\\ &= 4\cos^3 t-3\cos t\end{aligned}$$

Finally, with $x=y=z=t$ in the tangent formula from Theorem 3.26 we obtain the formula for the tangent in the statement, without any further manipulation. □

Getting now to numeric applications, the above formulae raise the possibility of computing the trigonometric functions of 10° and its multiples, by solving the corresponding cubic equations. However, this will not work very well, because do we really know how to solve the cubic equations. So, let us record here something modest, as follows:

THEOREM 3.28. *The quantities* $a = \sin 10^\circ$, $b = \cos 10^\circ$, $c = \tan 10^\circ$ *satisfy*

$$3a - 4a^3 = \frac{1}{2} \quad , \quad 4b^3 - 3b = \frac{\sqrt{3}}{2} \quad , \quad \frac{3c - c^3}{1 - 3c^2} = \frac{1}{\sqrt{3}}$$

and we have similar equations, for the multiples of 10°.

PROOF. By taking $t = 10^\circ$ in the formulae from Theorem 3.27, we obtain:

$$\sin(30^\circ) = 3a - 4a^3$$
$$\cos(30^\circ) = 4b^3 - 3b$$
$$\tan(30^\circ) = \frac{3c - c^3}{1 - 3c^2}$$

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

Moving on now, let us see what happens for a sum of 4 angles. In view of Theorem 3.26, we do not really want to deal with the sine and the cosine, where the formulae will be most likely quite complicated, so we will focus on the tangent instead. We have:

THEOREM 3.29. *The tangents of sums of* 4 *angles are given by*

$$\tan(x+y+z+t) = \frac{\begin{pmatrix} \tan x + \tan y + \tan z + \tan t - \tan x \tan y \tan z \\ - \tan x \tan y \tan t - \tan x \tan z \tan t - \tan y \tan z \tan t \end{pmatrix}}{\begin{pmatrix} 1 - \tan x \tan y - \tan x \tan z - \tan x \tan t - \tan y \tan z \\ - \tan y \tan t - \tan z \tan t + \tan x \tan y \tan z \tan t \end{pmatrix}}$$

provided of course that the denominator is nonzero.

PROOF. We use the formula for the tangents of sums from Theorem 3.20, namely:

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

By using this formula twice we obtain, for a sum of four angles:

$$\begin{aligned} & \tan(x+y+z+t) \\ = \ & \frac{\tan(x+y) + \tan(z+t)}{1 - \tan(x+y)\tan(z+t)} \\ = \ & \frac{\frac{\tan x + \tan y}{1 - \tan x \tan y} + \frac{\tan z + \tan t}{1 - \tan z \tan t}}{1 - \frac{\tan x + \tan y}{1 - \tan x \tan y} \cdot \frac{\tan z + \tan t}{1 - \tan z \tan t}} \\ = \ & \frac{\begin{pmatrix} \tan x + \tan y + \tan z + \tan t - \tan x \tan y \tan z \\ - \tan x \tan y \tan t - \tan x \tan z \tan t - \tan y \tan z \tan t \end{pmatrix}}{\begin{pmatrix} 1 - \tan x \tan y - \tan x \tan z - \tan x \tan t - \tan y \tan z \\ - \tan y \tan t - \tan z \tan t + \tan x \tan y \tan z \tan t \end{pmatrix}} \end{aligned}$$

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

And the problem is now, is what we found in Theorem 3.29 good news, or not? You would probably say, definitely no, that looks like the end of the world, but listen to the old man here, who has seen all sorts of complicated formulae, over his career, what we have in Theorem 3.29 is in fact not that bad. Indeed, we can now formulate:

THEOREM 3.30. *The tangents of the sums of angles are given by*

$$\tan(x+y) = \frac{a+b}{1-ab}$$

$$\tan(x+y+z) = \frac{a+b+c-abc}{1-ab-ac-bc}$$

$$\tan(x+y+z+t) = \frac{a+b+c+d-abc-abd-acd-bcd}{1-ab-ac-ad-bc-bd-cd+abcd}$$

$$\vdots$$

where $a = \tan x$, $b = \tan y$, $c = \tan z$, $d = \tan t$, *...*, *with on top odd symmetric functions of* $a, b, c, d, \ldots$, *and on the bottom even symmetric functions of* $a, b, c, d, \ldots$

PROOF. Here the formulae in the statement are those from Theorems 3.20, 3.26 and 3.29, and the conclusion at the end is something quite self-explanatory. We will leave some thinking here as an exercise, and we will be back to this, later in this book. □

Finally, a word on formulae of type $\cos(nt) = P_n(\cos t)$, and $\sin(nt) = Q_n(\cos t)\sin t$. We have seen that such formulae hold indeed at $n = 2, 3$, and with a bit more work, we can have them in general, and with P_n, Q_n being certain polynomials, called Chebycheff polynomials of the first and second kind. More on these, later in this book.

3e. Exercises

Exercises:

EXERCISE 3.31.

EXERCISE 3.32.

EXERCISE 3.33.

EXERCISE 3.34.

EXERCISE 3.35.

EXERCISE 3.36.

EXERCISE 3.37.

EXERCISE 3.38.

Bonus exercise.

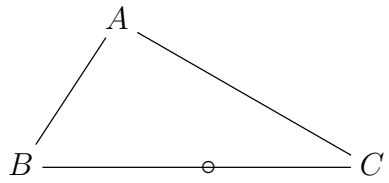
CHAPTER 4

Circles, pi

4a. Circles, chords

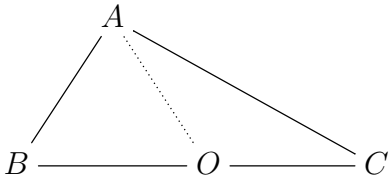
Let us get now into a more advanced study of the angles, by using circles, which are quite advanced technology. We have here the following key result, to start with:

THEOREM 4.1. *Any triangle lying on a circle, with two vertices on a diameter,*



is a right triangle.

PROOF. This is clear, because we have on the full picture of our triangle, with the center of the circle marked, two isosceles triangles appearing, as follows:



Thus, at the level of the corresponding angles, the $180°$ equation for our triangle is as follows, with r, s being respectively the angles at B, C:

$$2r + 2s = 180°$$

Thus, we obtain $r + s = 90°$, as claimed. □

Many other things can be said, as a continuation of this, the idea being that Theorem 4.1 provides us with a number of circle methods for studying the angles.

At a more advanced level, we have many interesting plane geometry results featuring circles, due to Monge, Apollonius and others, somehow in analogy with what we know about points and lines. Again, many interesting things can be said here.

We will be back to this later, at the end of the present chapter.

4b. Pi, numeric angles

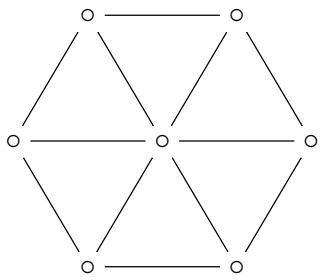
Let us get now into an even more advanced study of the angles. For this purpose, the best is to talk first about circles, more in detail, and about the number π.

But, do we really know what the number π is. So here, to start with, we have the following result, which can be regarded as being something quite axiomatic:

THEOREM 4.2. *The following two definitions of π are equivalent:*

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

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



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

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

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

In what regards now the precise value of π, the above picture at $N = 6$ shows that we have $\pi > 3$, but not by much. More can be said by using some basic trigonometry, for instance by replacing the hexagon used in the above with higher polygons:

THEOREM 4.3. *We can work out approximations of*

$$\pi = 3.14\ldots$$

by using various polygons, and basic trigonometry.

PROOF. This is indeed quite standard, as explained above. In practice, we are led in this way into estimating square roots of integers, and even square roots of reals containing square roots, which is not a simple question, either. We will be back to this, more in detail, at the end of the present chapter, with full computations for the small polygons. □

Getting now to really reliable results and data, which are actually known since long, obtained via lots of efforts, the precise figure for π is as follows:

$$\pi = 3.14159\ldots$$

We will come back to such approximation questions for π later, once we will have appropriate tools for dealing with them, coming from more advanced analysis.

It is also possible to prove that π is irrational, $\pi \notin \mathbb{Q}$, and even transcendental, but this is not trivial either. Again, we will be back to such questions later.

Getting now to what we wanted to do in this chapter, in relation with the angles, and their numeric measuring, let us formulate the following definition:

DEFINITION 4.4. *The value of an angle is obtained by putting that angle on the center of a circle of radius* 1*, and measuring the corresponding arc length.*

And this, which is something quite smart, will replace our previous conventions for the measuring of angles, with the basic conversion formulae being as follows:

$$0^\circ = 0 \quad , \quad 90^\circ = \frac{\pi}{2} \quad , \quad 180^\circ = \pi \quad , \quad 270^\circ = \frac{3\pi}{2}$$

Let us record as well the conversion formulae for the halves of these angles:

$$45^\circ = \frac{\pi}{4} \quad , \quad 135^\circ = \frac{3\pi}{4} \quad , \quad 225^\circ = \frac{5\pi}{4} \quad , \quad 315^\circ = \frac{7\pi}{4}$$

Finally, let us record as well the formulae for the thirds of the basic angles:

$$30^\circ = \frac{\pi}{6} \quad , \quad 60^\circ = \frac{\pi}{3} \quad , \quad 120^\circ = \frac{2\pi}{3} \quad , \quad 150^\circ = \frac{5\pi}{6}$$

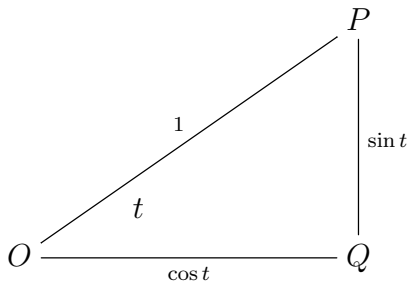
$$210^\circ = \frac{7\pi}{6} \quad , \quad 240^\circ = \frac{4\pi}{3} \quad , \quad 300^\circ = \frac{5\pi}{3} \quad , \quad 330^\circ = \frac{11\pi}{6}$$

In relation now with sin and cos, we are led in this way to the following alternate definitions, which better explain the various sign conventions made in chapter 3:

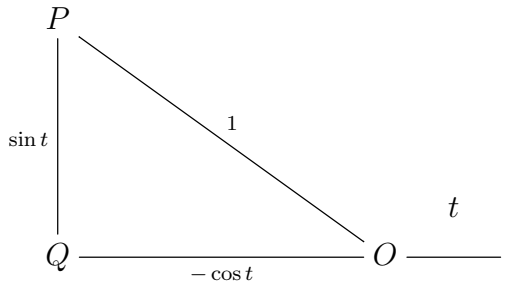
THEOREM 4.5. *The sine and cosine of an angle are obtained by putting the angle on the unit circle, as above, then projecting on the vertical and the horizontal, and then measuring the oriented segments that we get, on the vertical and horizontal.*

PROOF. This is clear from definitions, but for full clarity here, let us review now the detailed construction of the sine and cosine, for the arbitrary angles, from the previous chapter, with attention to signs, in the present setting. We have 4 cases, as follows:

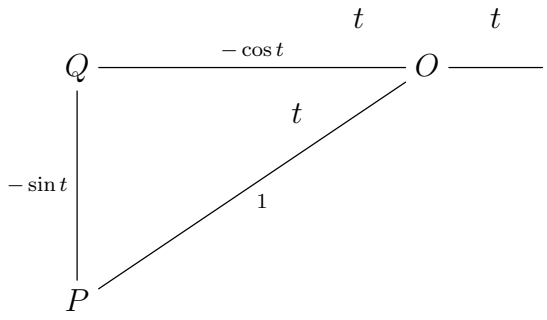
(1) In the simplest case, namely $t \in [0, \pi/2]$, the sine and cosine are indeed computed according to the following picture, which is the one in the statement:



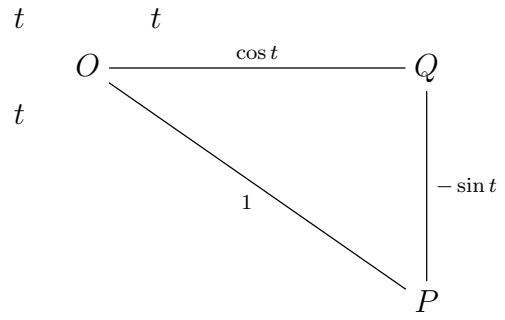
(2) In the case of obtuse angles, $t \in [\pi/2, \pi]$, the picture becomes as follows:



(3) In the next case, namely $t \in [\pi, 3\pi/2]$, the picture becomes as follows:



(4) As for the last case, namely $t \in [3\pi/2, 2\pi]$, here our picture is as follows:

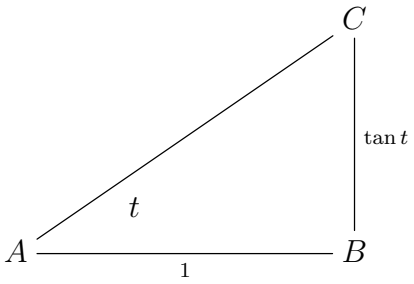


Thus, we are led to the conclusions in the statement. □

As an interesting fact, we can complement Theorem 4.5 with a statement regarding the tangent, the trigonometric function that we often forgot so far, as follows:

THEOREM 4.6. *The tangent of an angle can be obtained by putting the angle on the unit circle, as before, and then measuring the oriented segment that we get, on the vertical, outside the circle, on the vertical tangent at right.*

PROOF. This is, again, something quite self-explanatory, with the picture here being something that we already know from chapter 3, namely:



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

Now that we know well about sine, cosine and tangent, time perhaps to introduce the remaining trigonometric functions too. These are as follows:

DEFINITION 4.7. *Reciprocals of* $\sin, \cos, \tan$.

We will be actually not using much these latter functions, which rather bring confusion into the math formulae, two many definitions being, generally speaking, a bad thing.

On the opposite now, here are some truly interesting functions:

DEFINITION 4.8. *Inverses of* $\sin, \cos, \tan$*, and of their reciprocals too.*

Here a bijectivity discussion is of course needed. There are actually many things can that be said about these inverses. We will be back to this.

4c. Basic estimates

Let us get now into an interesting question, namely estimating $\sin, \cos, \tan$ and the other trigonometric functions. For this purpose, let us first recall the basic formulae for the sums of angles, that we established in chapter 3, which were as follows:

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

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

Obviously, these formulae allow us to transport our approximation questions around $t = 0$, so with this understood, let us get now to what happens around 0.

And here, to start with, we have the following basic estimates:

THEOREM 4.9. *We have the following estimates,*

$$\sin t \leq t \leq \tan t$$

valid for small sngles.

PROOF. The above two estimates are indeed both clear from our circle picture for the angles, and trigonometric functions. One interesting question concerns the exact range of the above estimates, and we will leave the discussion here as an interesting exercise. □

In fact, by using our circle technology, we are led to the following result:

THEOREM 4.10. *The following happen, for small angles:*

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

PROOF. This can be indeed established as follows:

(1) This is clear indeed on the circle.

(2) This comes from (1), and from Pythagoras. Indeed, knowing $\sin t \simeq t$, when looking for a quantity $\cos t$ making the Pythagoras formula $\sin^2 t + \cos^2 t = 1$ hold, we are led, via some quick thinking, to the formula $\cos t \simeq 1 - t^2/2$, as stated. Here is the verification, and with the result itself coming via some reverse engineering, from this:

$$\begin{aligned}\left(1 - \frac{t^2}{2}\right)^2 + t^2 &= \left(1 - t^2 + \frac{t^4}{4}\right) + t^2\\ &\simeq 1 - t^2 + t^2\\ &= 1\end{aligned}$$

(3) This is again clear on the circle. □

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

THEOREM 4.11. *The following happen, for small angles:*

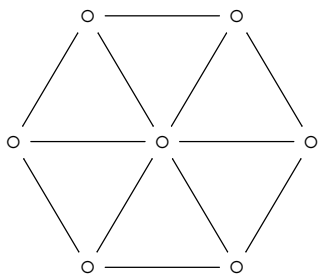
(1) $\sin t \simeq t - t^3/6$.
(2) $\cos t \simeq 1 - t^2/2 + t^4/24$.
(3) $\tan t \simeq t + t^3/3$.

PROOF. This is something which is substantially harder to prove, and with the comment that, as before with the estimates in Theorem 4.10, there are some relations between the above estimates, at various orders, due to Pythagoras, and to the formula for the tangent, in terms of the sine and cosine. Here is for instance the verification for the fact that the above formulae for the sine and cosine are compatible indeed with Pythagoras:

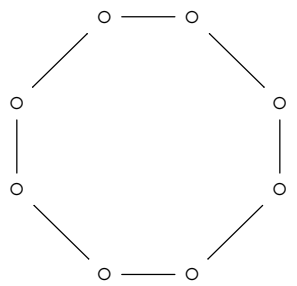
$$\begin{aligned}
\sin^2 t + \cos^2 t &\simeq \left(t - \frac{t^3}{6}\right)^2 + \left(1 - \frac{t^2}{2} + \frac{t^4}{24}\right)^2\\
&= \left(t^2 - \frac{t^4}{3} + \frac{t^6}{36}\right) + \left(1 + \frac{t^4}{4} + \frac{t^8}{576} - t^2 + \frac{t^4}{12} - \frac{t^6}{24}\right)\\
&\simeq \left(t^2 - \frac{t^4}{3} + \frac{t^6}{36}\right) + \left(1 + \frac{t^4}{4} - t^2 + \frac{t^4}{12} - \frac{t^6}{24}\right)\\
&= \left(t^2 - \frac{t^4}{3} + \frac{t^6}{36}\right) + \left(1 - t^2 + \frac{t^4}{3} - \frac{t^6}{24}\right)\\
&\simeq \left(t^2 - \frac{t^4}{3}\right) + \left(1 - t^2 + \frac{t^4}{3}\right)\\
&= 1
\end{aligned}$$

Quite wild all this, hope you agree with me. We will be back to such questions later in this book, once we will have better tools for dealing with them. □

Finally, still talking analysis, we have a lot of interesting estimates, of varying levels of difficuty, regarding π itself. As already mentioned in the beginning of this chapter, we can see right away that we have $\pi > 3$, and not by much, by using a hexagon:



Leaving the heptagon aside, next we have the octagon, which is as follows:



And here, with some trigonometry help from Conor and Khabib, and with GSP helping with the square roots, we can have some approximations for π going. To be more precise, we can use here the following formulae, established in chapter 3:

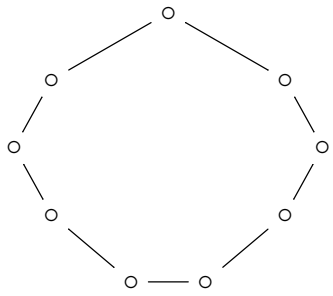
$$\sin(2t) = 2\sin t\cos t$$

$$\begin{aligned}\cos(2t) &= \cos^2 t - \sin^2 t\\ &= 2\cos^2 t - 1\\ &= 1 - 2\sin^2 t\end{aligned}$$

$$\tan(2t) = \frac{2\tan t}{1-\tan^2 t}$$

Thus, we can compute the octagon edge, and then approximate π.

Next, we can have some computations for the nonagon, which looks as follows:



And so on, with increasingly more complex computations, which are all interesting, hiding all sorts of mysterious mathematics, and with all this being quite addictive.

We will be back to such things later in this book, with some better methods for approximating π, when systematically doing analysis.

4d. Polar geometry

As a last topic for the present chapter, and for the present Part I of this book, with our acquired knowledge of circles, and basic trigonometry, we can now go back to some things discussed a bit in a hurry in chapter 1, in relation with the duality between points and lines in the plane, and fully clarify them. We first have the following result:

THEOREM 4.12. *We have a duality between points and lines, obtained by fixing a circle in the plane, say of center O and radius $r > 0$, and doing the following,*

(1) *Given a point P, construct Q on the line OP, as to have $OP \cdot OQ = r^2$,*
(2) *Draw the perpendicular at Q on the line OQ. This is the dual line p,*

and this duality $P \leftrightarrow p$ transforms collinear points into concurrent lines.

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

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

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

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

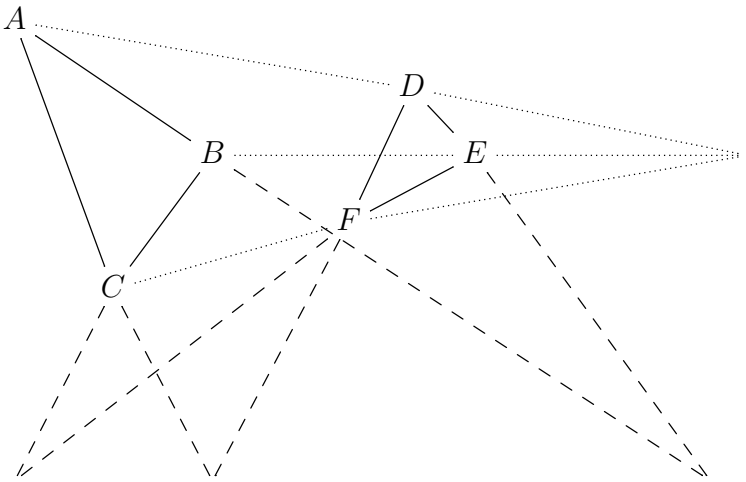
(4) Before getting further, let us make a few simple observations. As a first remark, when P belongs to the circle, p is the tangent to the circle, drawn at that point P.

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

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

As a basic example for this, that we already used in chapter 1, when proving the second implication of the Desargues theorem, starting from the first one, we have:

EXAMPLE 4.13. *The Desargues configuration, namely*

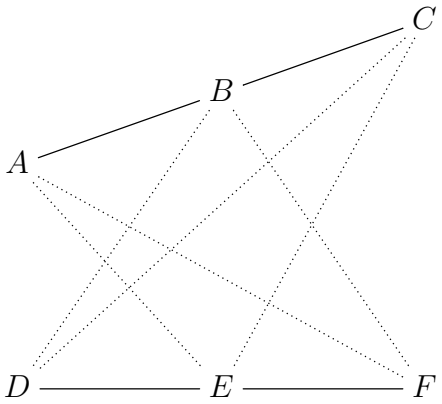


is self-dual.

To be more precise, this is indeed something quite obvious, and we refer to chapter 1 for more about this, both self-duality explanations, and applications.

Next, still in relation with what we did in chapter 1, we have:

EXAMPLE 4.14. *The Pappus configuration, namely*

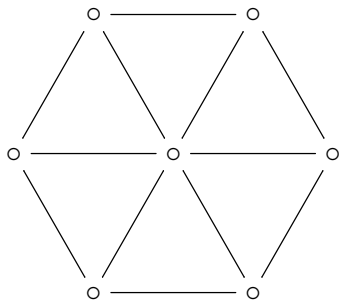


is self-dual too.

Again, this is something quite obvious, and we refer to chapter 1 for more about this, both self-duality explanations, and applications, and with the comment, already made there in chapter 1, that Desargues and Pappus are related to each other, too.

Coming next, at a more complicated level, we have the following theorem:

THEOREM 4.15 (Pascal). *Given a hexagon lying on a circle*

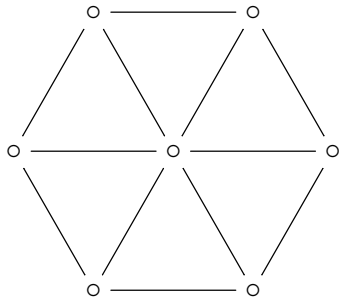


the pairs of opposite sides intersect in points which are collinear.

PROOF. This can be proved indeed, with some tricks. Observe the similarity with Pappus. We will see in fact, later in this book, when talking about conics, that the Pascal theorem generalizes to the case of conics, and with this fully generalizing Pappus. □

And here is now, at a truly advanced level, a quite scary theorem:

THEOREM 4.16 (Brianchon). *Given a hexagon circumscribed around on a circle*



the main diagonals intersect.

PROOF. This is nearly impossible to prove, with bare hands, and ask around kids preparing for Math Olympiads, they will witness for that. But, this follows by duality from Pascal. As before with Pascal, we will see later that this extends to conics. □

Quite interesting all these results about hexagons, and the relation between them, so let us ask the cat, what she thinks about all this. And cat answers:

CAT 4.17. *In hexagrammum mysticum you will trust.*

Okay, and not that I understand what cat says, but the plan for what follows next becomes now clear, keep developing trigonometry, by keeping an eye on hexagons.

4e. Exercises

Exercises:

EXERCISE 4.18.

EXERCISE 4.19.

EXERCISE 4.20.

EXERCISE 4.21.

EXERCISE 4.22.

EXERCISE 4.23.

EXERCISE 4.24.

EXERCISE 4.25.

Bonus exercise.

Part II

Affine coordinates

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:*

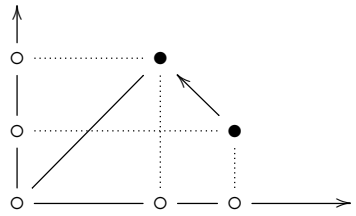
$$\begin{pmatrix}x\\y\end{pmatrix}\to\begin{pmatrix}x\\0\end{pmatrix}\quad,\quad\begin{pmatrix}x\\y\end{pmatrix}\to\begin{pmatrix}0\\y\end{pmatrix}$$

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) $\implies$ (2) We know that f satisfies the following equation, and $f(0) = 0$:

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

By setting $y = 0$, and by using our assumption $f(0) = 0$, we obtain, as desired:

$$f(tx) = tf(x)$$

As for the first condition, regarding sums, this can be established as follows:

$$\begin{aligned} f(x+y) &= f\left(2 \cdot \frac{x+y}{2}\right) \\ &= 2f\left(\frac{x+y}{2}\right) \\ &= 2 \cdot \frac{f(x)+f(y)}{2} \\ &= f(x) + f(y) \end{aligned}$$

(2) $\implies$ (1) Conversely now, assuming that f satisfies $f(x+y) = f(x) + f(y)$ and $f(\lambda x) = \lambda f(x)$, then f must map lines to lines, as shown by:

$$\begin{aligned} f(tx + (1-t)y) &= f(tx) + f((1-t)y) \\ &= tf(x) + (1-t)f(y) \end{aligned}$$

Also, we have $f(0) = f(2 \cdot 0) = 2f(0)$, which gives $f(0) = 0$, as desired. □

The above result is very useful, and in practice, we will often use the condition (2) there, somewhat as a new definition for the linear maps.

Let us record this finding as an upgrade of our formalism, as follows:

DEFINITION 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:

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

Regarding now $f(\lambda x) = \lambda f(x)$, since f maps lines to lines, it must map the line $0 - x$ to the line $0 - f(x)$, so we have a formula as follows, for any λ, x:

$$f(\lambda x) = \varphi_x(\lambda) f(x)$$

But since f maps parallel lines to parallel lines, by Thales the function $\varphi_x : \mathbb{R} \to \mathbb{R}$ does not depend on x. Thus, we have a formula as follows, for any λ, x:

$$f(\lambda x) = \varphi(\lambda) f(x)$$

We know that we have $\varphi(0) = 0$ and $\varphi(1) = 1$, and we must prove that we have $\varphi(\lambda) = \lambda$ for any λ. For this purpose, we use a trick. On one hand, we have:

$$f((\lambda + \mu)x) = \varphi(\lambda + \mu)f(x)$$

On the other hand, since f maps sums to sums, we have as well:

$$\begin{aligned} f((\lambda + \mu)x) &= f(\lambda x) + f(\mu x) \\ &= \varphi(\lambda)f(x) + \varphi(\mu)f(x) \\ &= (\varphi(\lambda) + \varphi(\mu))f(x) \end{aligned}$$

Thus our rescaling function $\varphi : \mathbb{R} \to \mathbb{R}$ satisfies the following conditions:

$$\varphi(0) = 0 \quad , \quad \varphi(1) = 1 \quad , \quad \varphi(\lambda + \mu) = \varphi(\lambda) + \varphi(\mu)$$

But with these conditions in hand, it is clear that we have $\varphi(\lambda) = \lambda$, first for all the inverses of integers, $\lambda = 1/n$ with $n \in \mathbb{N}$, then for all rationals, $\lambda \in \mathbb{Q}$, and finally by continuity for all reals, $\lambda \in \mathbb{R}$. Thus, we have proved the following formula:

$$f(\lambda x) = \lambda f(x)$$

But this finishes the proof of (2) $\implies$ (1), and we are done. □

All this is very nice, and there are some further things that can be said, but getting to business, Definition 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:

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

As already mentioned, all this comes from our findings from Theorem 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* $l \in (0, 2q)$*:*

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

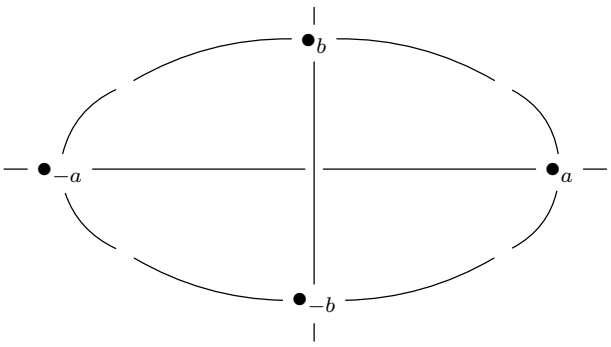
(3) *As the curves appearing when drawing a circle, from various perspectives:*

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

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

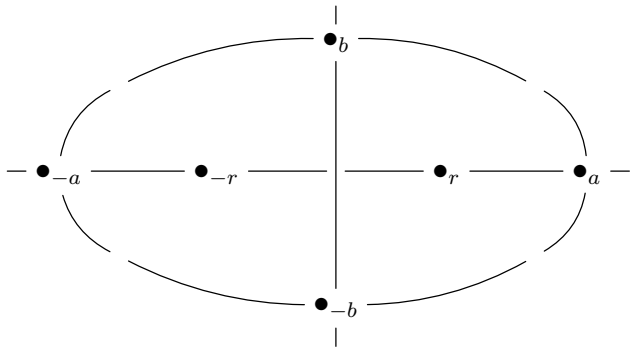
PROOF. This might look a bit confusing, and you might say, what exactly is to be proved here. Good point, and in answer, what is to be proved is that the above constructions (1-4) give rise to the same class of curves. And this can be done as follows:

(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 $z=(x,y)$:

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

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

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

(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. □

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, which actually requires a bit of knowledge of calculus and equations. We will come back to this, with more details, once we will know more about calculus, in Part III below. □

5d. Some arithmetic

Getting back to the basics, all of what we have been doing in the above was based on the basic fact that the points on a line can be indexed by real numbers.

But, based on this, we can develop geometry by using lines indexed by other types of numbers. And, regarding these numbers, that we can use, these usually come from:

DEFINITION 5.22. *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 $a\neq0$ has a multiplicative inverse $a^{-1}\in F$, satisfying $aa^{-1}=1$.*
(3) *The sum and product are compatible via $a(b+c)=ab+ac$.*

As a basic example here, passed the reals that we know well, we have the field of rational numbers $\mathbb{Q}$, with its usual addition and multiplication, namely:

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

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

$$1+1=0$$

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

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

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

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

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

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

THEOREM 5.23. *The following happen:*

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

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

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

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

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

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

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

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

Now, let us do some basic geometry, say over $\mathbb{F}_p$. However, things are a bit bizarre here, and we have for instance the following result, to start with:

PROPOSITION 5.24. *The circle of radius zero $x^2+y^2=0$ over $\mathbb{F}_p$ is as follows:*

(1) *At $p=2$, this has 2 points.*
(2) *At $p=1(4)$, this has $2p-1$ points.*
(3) *At $p=3(4)$, this has 1 point.*

PROOF. Our circle $x^2+y^2=0$ is formed by the point $(0,0)$, and then of the solutions of $x^2+y^2=0$, with $x,y\neq 0$. But this latter equation is equivalent to $(x/y)^2+1=0$, and so to $(x/y)^2=-1$, so the number of points of our circle is:

$$N=1+(p-1)\#\{r|r^2=-1\}$$

But at $p=2$ this gives $N=1+1\times 1=2$, then at $p=1(4)$ this gives $N=1+(p-1)\times 2=2p-1$, and finally at $p=3(4)$ this gives $N=1+(p-1)\times 0=1$. □

When looking at more general conics, still over $\mathbb{F}_p$, things do not necessarily improve, and we have some other bizarre results, along the same lines, such as:

THEOREM 5.25. *Any curve over $\mathbb{F}_2$ is a conic. However, this is not the case for $\mathbb{F}_p$ with $p\geq 3$.*

PROOF. This is again something elementary, as follows:

(1) Let us find the conics over $\mathbb{F}_2$. These are given by equations as follows:

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

Since $x^2=x$ holds in $\mathbb{F}_2$, the first 2 terms dissapear, and we are left with:

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

– The first case, $c=0$, corresponds to the lines over $\mathbb{F}_2$. But there are 8 such lines, all distinct, given by $r=0$, $x=r$, $y=r$, $x+y=r$, with $r=0,1$.

– The second case, $c\neq 0$, corresponds to the non-degenerate conics over $\mathbb{F}_2$. But there are 8 such conics, all distinct, and distinct as well from the 8 lines found above, given by $xy=r$, $x(y+1)=r$, $(x+1)y=r$, $(x+1)(y+1)=r$, with $r=0,1$.

Summarizing, we have $8+8=16$ conics over $\mathbb{Z}_2$. But since the plane $\mathbb{F}_2\times\mathbb{F}_2$ has $2\times 2=4$ points, there are $2^4=16$ possible curves. Thus, all the curves are conics.

(2) Regarding now $\mathbb{F}_p$ with $p\geq 3$, here the plane $\mathbb{F}_p\times\mathbb{F}_p$ has p^2 points, so there are 2^{p^2} curves. Among these curves, the conics are given by equations as follows:

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

Thus, we have at most p^6 conics, and since we have $2^{p^2}>p^6$ for any $p\geq 4$, we are done with the case $p\geq 5$. In the remaining case now, $p=3$, the $3^6=729$ possible conics split into the $2^5=243$ ones with $a=0$, and the $2\times 243=486$ ones with $a\neq 0$. But

these latter conics appear twice, as we can see by dividing everything by a, and so there are only $1 \times 243 = 243$ of them. Thus, we have at most $243 + 243 = 486$ conics, and this is smaller than the number of curves of $\mathbb{F}_3 \times \mathbb{F}_3$, which is $2^9 = 512$, as desired. □

As a conclusion to all this, better stay away from characteristic p.

On the positive side, however, we have the following interesting result:

THEOREM 5.26. *Given a field F, we can talk about the projective space P_F^{N-1}, as being the space of lines in F^N passing through the origin. At $N = 3$ we have*

$$|P_F^2| = q^2 + q + 1$$

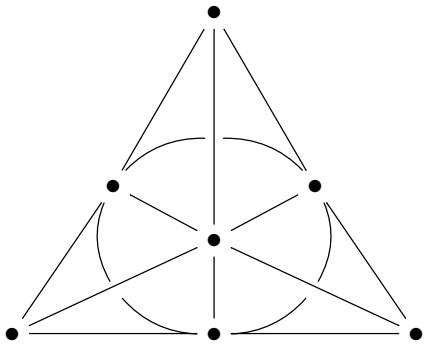
where $q = |F|$, in the case where our field F is finite.

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

$$\begin{aligned} |P_F^2| &= \frac{|F^3 - \{0\}|}{|F - \{0\}|} \\ &= \frac{q^3 - 1}{q - 1} \\ &= q^2 + q + 1 \end{aligned}$$

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

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



Here the circle in the middle is by definition a line, and with this convention, the basic projective geometry axioms from chapter 1 are satisfied, in the sense that any two points determine a line, and any two lines determine a point. And isn't this beautiful.

5e. Exercises

Exercises:

Exercise 5.27.

Exercise 5.28.

Exercise 5.29.

Exercise 5.30.

Exercise 5.31.

Exercise 5.32.

Exercise 5.33.

Exercise 5.34.

Bonus exercise.

CHAPTER 6

Basic trigonometry

6a. Triangles, revised

Time now to see how our coordinate technology works, if that is worth something, or not. We will review here all the triangle and basic geometry material from Part I, with new proofs for everything, using coordinates, no less than that.

So, God bless, and let us get started. As a first good surprise, in what regards the axiomatics from chapter 1, that is literally nuked by coordinates.

We first have, indeed, regarding the first axiom of geometry, that we started this book with, the following theorem, coming along with a trivial proof:

THEOREM 6.1. *Any two distinct points $P \neq Q$ determine a line, denoted PQ.*

PROOF. This is clear indeed, with coordinates, because we have:

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

So, very good news, axiom becoming theorem, what more can we wish for. □

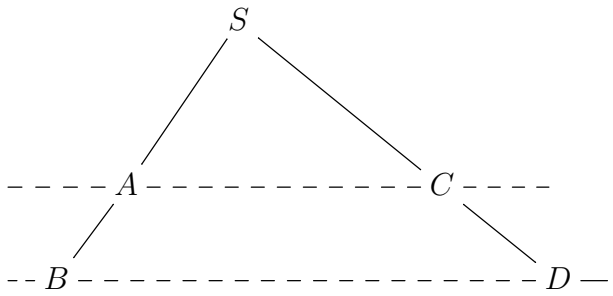
Same situation for the second axiom, which becomes a theorem too:

THEOREM 6.2. *Given a point not lying on a line, $P \notin L$, we can draw through P a unique parallel to L. That is, we can find a line K satisfying $P \in K$, $K||L$.*

PROOF. This is again clear with coordinates. □

Getting now to the next thing that we did in Part I, namely the Thales theorem, and as further good news, that drastically simplifies with coordinates, as follows:

THEOREM 6.3 (Thales). *Proportions are kept, along parallel lines. That is, given a configuration as follows, consisting of two parallel lines, and of two extra lines,*



the following equality holds:

$$\frac{SA}{SB}=\frac{SC}{SD}$$

Moreover, the converse of this holds too, in the sense that, in the context of a picture as above, if this equality is satisfied, then the lines AC and BD must be parallel.

PROOF. Again, this is clear with coordinates, and in fact the other formulations of the Thales theorem, also from Part I, are clear as well too, again with coordinates. To be more precise, for the above configuration, the conclusion is as follows:

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

In addition, we can prove Thales 3 as well, again using coordinates. □

Next, we have the Desargues theorem:

THEOREM 6.4 (Desargues). *Two triangles are in perspective axially if and only if they are in perspective centrally.*

PROOF. Again, this is clear with coordinates. □

Next, we have the Pappus theorem:

THEOREM 6.5 (Pappus). *Given a hexagon with both the odd and the even vertices being colinear, the pairs of opposite sides cross into three colinear points.*

PROOF. Again, this is clear with coordinates. □

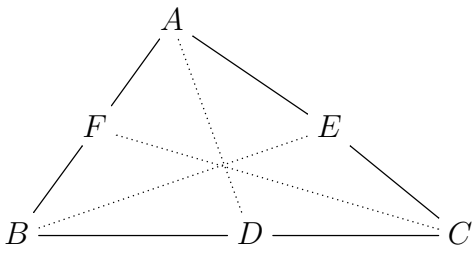
In relation with the projective geometry considerations from the end of chapter 1, coordinates can halp in that setting too, and we have the following result:

THEOREM 6.6. *Projective coordinates.*

PROOF. This is something that can be done too, and many interesting things can be said here. We will be back to this on several occasions, in what follows. □

Getting now to the barycenter theorem, this drastically simplifies, as follows:

THEOREM 6.7 (Barycenter). *Given a triangle ABC, its medians cross,*



at a point called barycenter, lying at $1/3 - 2/3$ on each median.

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

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

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

We can prove now as well some things claimed in chapter 2, as follows:

THEOREM 6.8. *The gravity center of a triangle ABC is as follows:*

(1) *In the 0-dimensional case, that is, when putting equal weigths at the vertices A, B, C, and computing the center, this is the barycenter.*
(2) *In the 1-dimensional case, that is, with the sides AB, BC, AC have weigths proportional with their length, this is, in general, different from the barycenter.*
(3) *In the 2-dimensional case, that is, with the triangle ABC itself, as an area, having a weight, uniformly distributed, this is again the barycenter.*

PROOF. Again, this is clear with coordinates. □

Getting now to the other centers of a triangle, we have here:

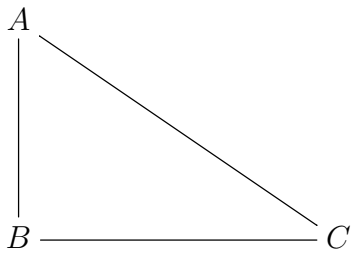
THEOREM 6.9. *Given a triangle ABC, the following happen:*

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

PROOF. Again, such things can be proved with coordinates, and patience. We will actually leave some of the calculations here as an instructive exercise for you, reader. □

Coming next, we have the theorem of Pythagoras:

THEOREM 6.10 (Pythagoras). *In a right triangle ABC,*

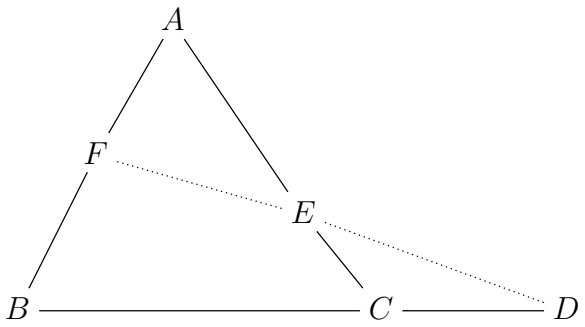


we have $AB^2 + BC^2 = AC^2$.

PROOF. Again, this is clear with coordinates. □

Next, we have the following key result, due to Menelaus:

THEOREM 6.11 (Menelaus). *In a configuration of the following type, with a triangle ABC cut by a line FED,*



we have the following formula, with all segments being taken oriented:

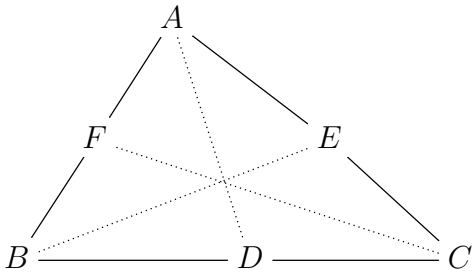
$$\frac{AF}{FB} \cdot \frac{BD}{DC} \cdot \frac{CE}{EA} = -1$$

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

PROOF. Again, this is clear with coordinates. □

Next, we have the following remarkable result, due to Ceva:

THEOREM 6.12 (Ceva). *In a configuration of the following type, with a triangle ABC containing inner lines AD, BE, CF which cross,*



we have the following formula:

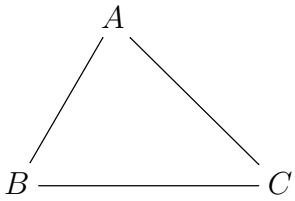
$$\frac{AF}{FB} \cdot \frac{BD}{DC} \cdot \frac{CE}{EA} = 1$$

Moreover, the converse holds, with this formula guaranteeing that AD, BE, CF cross.

PROOF. Again, this is clear with coordinates. □

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

THEOREM 6.13. *Besides the 4 main centers of a triangle, discussed in the above, many more remarkable points can be associated to a triangle ABC,*

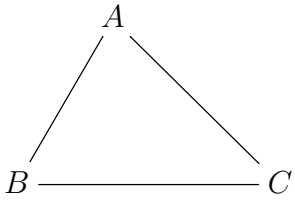


and most of these lie on a line, called Euler line of ABC.

PROOF. Proving this with coordinates is a good exercise for you, reader. □

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

THEOREM 6.14. *Associated to a triangle ABC,*



we have as well a nine-point circle, whose center lies on the Euler line.

PROOF. Again, proving this with coordinates is a good exercise for you, reader. □

As a conclusion to all this, coordinates seem to perfom quite well, and you might probably have this question right now, why not having started the present book with coordinates. In answer, modesty and patience, this is how math is best learned. We will actually see in what follows that our present $\mathbb{R}^2$ coordinates can be beaten themselves by some better coordinates, namely the $\mathbb{C}$ ones. So, long story still to go, and ho hurry.

6b. Polar coordinates

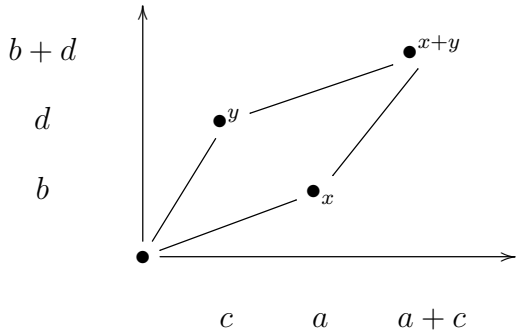
We have seen so far that many things from plane geometry can be understood by using coordinates, with each point $x \in \mathbb{R}^2$ being written as a vector, as follows:

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

Of particular interest was the summing operation for such vectors, which, according to the usual calculus rules for the vectors, was given by the following formula:

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

As explained in chapter 5, in great detail, geometrically, the idea here was simply that the vectors add by forming a parallelogram, as follows:



In practice, the summing operation is usefully complemented by the multiplication by scalars operation, which is given by the following very intuitive formula:

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

Finally, of particular interest too, in relation with the computation of the lengths, was the following formula, allowing us to compute the length of any vector:

$$x = \begin{pmatrix} a \\ b \end{pmatrix} \implies ||x|| = \sqrt{a^2 + b^2}$$

Our idea in what follows will be that of improving part of this vector technology, by using polar coordinates. The idea here is very simple, as follows:

THEOREM 6.15. *The points of the plane $x \in \mathbb{R}^2$, written as vectors*

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

can be written in polar coordinates, as follows,

$$x = \begin{pmatrix} r\cos t \\ r\sin t \end{pmatrix}$$

with the connecting formulae being as follows,

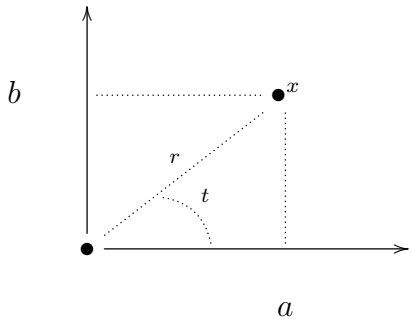
$$a = r\cos t \quad , \quad b = r\sin t$$

and in the other sense being as follows,

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

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

PROOF. This is something self-explanatory and intuitive, with $r = \sqrt{a^2 + b^2}$ being as usual the length of the vector, and with t being the angle made by the vector with the Ox axis. That is, with the picture for what is going on in the above being as follows:



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

Many interesting things can be done with polar coordinates. We will be back to this.

6c. Circles and angles

Circles and angles. Ellipses, revised. Parabolas and hyperbolas.

6d. Basic trigonometry

Basic trigonometry, in the plane. We will be back to this, in the next chapter.

6e. Exercises

Exercises:

Exercise 6.16.

Exercise 6.17.

Exercise 6.18.

Exercise 6.19.

Exercise 6.20.

Exercise 6.21.

Exercise 6.22.

Exercise 6.23.

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:

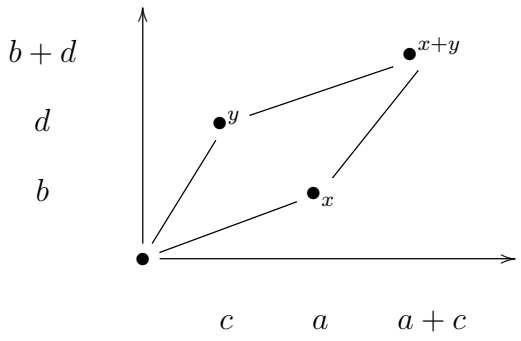
$$x = \begin{pmatrix} a \\ b \end{pmatrix} \quad , \quad y = \begin{pmatrix} c \\ d \end{pmatrix}$$

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. □

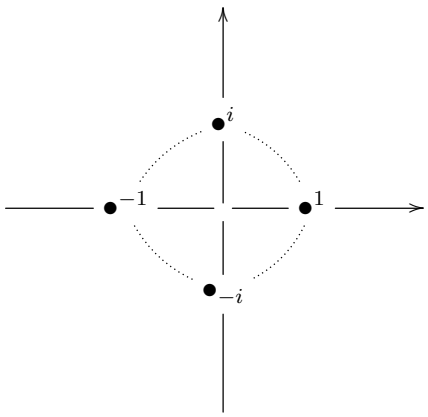
Here we have assumed that you are a bit familiar with vector calculus. If not, no problem, the idea is simply that vectors add by forming a parallelogram, as follows:



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,

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

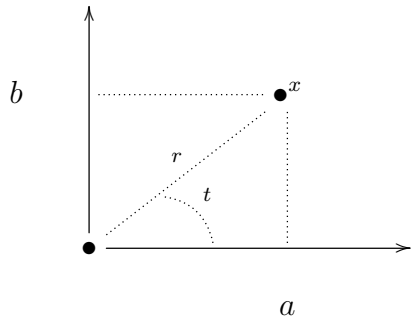
and in the other sense being as follows,

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

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. □

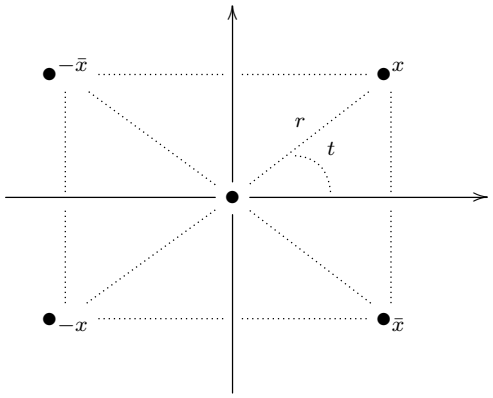
As a last general topic regarding the complex numbers, let us discuss conjugation. This is something quite tricky, complex number specific, as follows:

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. □

Still talking polynomials and their roots, let us try now to understand what the analogue of $\Delta = b^2 - 4ac$ is, for an arbitrary polynomial $P \in \mathbb{C}[X]$. We will need:

THEOREM 7.13. *Given two polynomials $P, Q \in \mathbb{C}[X]$, written as follows,*

$$P = c(X - a_1) \ldots (X - a_k) \quad , \quad Q = d(X - b_1) \ldots (X - b_l)$$

the following quantity, which is called resultant of P, Q,

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

is a polynomial in the coefficients of P, Q, with integer coefficients, and we have

$$R(P, Q) = 0$$

precisely when P, Q have a common root.

PROOF. Given $P, Q \in \mathbb{C}[X]$, we can certainly construct the quantity $R(P, Q)$ in the statement, and we have then $R(P, Q) = 0$ precisely when P, Q have a common root. The whole point is that of proving that $R(P, Q)$ is a polynomial in the coefficients of P, Q, with integer coefficients. But this can be checked as follows:

(1) We can expand the formula of $R(P, Q)$, and in what regards $a_1, \ldots, a_k$, which are the roots of P, we obtain in this way certain symmetric functions in these variables, which will be therefore polynomials in the coefficients of P, with integer coefficients.

(2) We can then look what happens with respect to the remaining variables $b_1, \ldots, b_l$, which are the roots of Q. Once again what we have here are certain symmetric functions, and so polynomials in the coefficients of Q, with integer coefficients.

(3) Thus, we are led to the conclusion in the statement, that $R(P, Q)$ is a polynomial in the coefficients of P, Q, with integer coefficients, and with the remark that the $c^l d^k$ factor is there for these latter coefficients to be indeed integers, instead of rationals. □

All this might seem a bit complicated, and as an illustration, let us work out an example. Consider the case of a polynomial of degree 2, and a polynomial of degree 1:

$$P = ax^2 + bx + c \quad , \quad Q = dx + e$$

In order to compute the resultant, let us factorize our polynomials:

$$P = a(x - p)(x - q) \quad , \quad Q = d(x - r)$$

The resultant can be then computed as follows, by using the method above:

$$\begin{aligned} R(P, Q) &= ad^2(p - r)(q - r) \\ &= ad^2(pq - (p + q)r + r^2) \\ &= cd^2 + bd^2r + ad^2r^2 \\ &= cd^2 - bde + ae^2 \end{aligned}$$

With this, we can now talk about the discriminant of any polynomial, as follows:

THEOREM 7.14. *Given a polynomial $P \in \mathbb{C}[X]$, written as*

$$P(X) = cX^N + dX^{N-1} + \ldots$$

its discriminant, defined as being the following quantity,

$$\Delta(P) = \frac{(-1)^{\binom{N}{2}}}{c} R(P, P')$$

is a polynomial in the coefficients of P, with integer coefficients, and

$$\Delta(P) = 0$$

happens precisely when P has a double root.

PROOF. This follows from Theorem 7.13, applied with $P = Q$, with the division by c being indeed possible, under $\mathbb{Z}$, and with the sign being there for various reasons, including the compatibility with some well-known formulae, at small values of $N \in \mathbb{N}$. □

As an illustration, let us see what happens in degree 2. Here we have:

$$P = aX^2 + bX + c \quad , \quad P' = 2aX + b$$

Thus, the resultant is given by the following formula:

$$\begin{aligned} R(P, P') &= ab^2 - b(2a)b + c(2a)^2 \\ &= 4a^2c - ab^2 \\ &= -a(b^2 - 4ac) \end{aligned}$$

With the normalizations in Theorem 7.14 made, we obtain, as we should:

$$\Delta(P) = b^2 - 4ac$$

As another illustration, let us work out what happens in degree 3. Here the result, which is useful and interesting, and is probably new to you, is as follows:

THEOREM 7.15. *The discriminant of a degree* 3 *polynomial,*

$$P = aX^3 + bX^2 + cX + d$$

is the number $\Delta(P) = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd$.

PROOF. We need to do some tough computations here. Let us first compute resultants. Consider two polynomials, of degree 3 and degree 2, written as follows:

$$P = aX^3 + bX^2 + cX + d = a(X - p)(X - q)(X - r)$$

$$Q = eX^2 + fX + g = e(X - s)(X - t)$$

The resultant of these two polynomials is then given by:

$$\begin{aligned} R(P, Q) &= a^2e^3(p - s)(p - t)(q - s)(q - t)(r - s)(r - t) \\ &= a^2 \cdot e(p - s)(p - t) \cdot e(q - s)(q - t) \cdot e(r - s)(r - t) \\ &= a^2Q(p)Q(q)Q(r) \\ &= a^2(ep^2 + fp + g)(eq^2 + fq + g)(er^2 + fr + g) \end{aligned}$$

By expanding, we obtain the following formula for this resultant:

$$\begin{aligned} \frac{R(P, Q)}{a^2} &= e^3p^2q^2r^2 + e^2f(p^2q^2r + p^2qr^2 + pq^2r^2) \\ &+ e^2g(p^2q^2 + p^2r^2 + q^2r^2) + ef^2(p^2qr + pq^2r + pqr^2) \\ &+ efg(p^2q + pq^2 + p^2r + pr^2 + q^2r + qr^2) + f^3pqr \\ &+ eg^2(p^2 + q^2 + r^2) + f^2g(pq + pr + qr) \\ &+ fg^2(p + q + r) + g^3 \end{aligned}$$

Note in passing that we have 27 terms on the right, as we should, and with this kind of check being mandatory, when doing such computations. Next, we have:

$$p+q+r=-\frac{b}{a}\quad,\quad pq+pr+qr=\frac{c}{a}\quad,\quad pqr=-\frac{d}{a}$$

By using these formulae, we can produce some more, as follows:

$$p^2+q^2+r^2=(p+q+r)^2-2(pq+pr+qr)=\frac{b^2}{a^2}-\frac{2c}{a}$$

$$p^2q+pq^2+p^2r+pr^2+q^2r+qr^2=(p+q+r)(pq+pr+qr)-3pqr=-\frac{bc}{a^2}+\frac{3d}{a}$$

$$p^2q^2+p^2r^2+q^2r^2=(pq+pr+qr)^2-2pqr(p+q+r)=\frac{c^2}{a^2}-\frac{2bd}{a^2}$$

By plugging now this data into the formula of $R(P,Q)$, we obtain:

$$\begin{aligned}R(P,Q)&=a^2e^3\cdot\frac{d^2}{a^2}-a^2e^2f\cdot\frac{cd}{a^2}+a^2e^2g\left(\frac{c^2}{a^2}-\frac{2bd}{a^2}\right)+a^2ef^2\cdot\frac{bd}{a^2}\\&+a^2efg\left(-\frac{bc}{a^2}+\frac{3d}{a}\right)-a^2f^3\cdot\frac{d}{a}\\&+a^2eg^2\left(\frac{b^2}{a^2}-\frac{2c}{a}\right)+a^2f^2g\cdot\frac{c}{a}-a^2fg^2\cdot\frac{b}{a}+a^2g^3\end{aligned}$$

Thus, we have the following formula for the resultant:

$$\begin{aligned}R(P,Q)&=d^2e^3-cde^2f+c^2e^2g-2bde^2g+bdef^2-bcefg+3adefg\\&-adf^3+b^2eg^2-2aceg^2+acf^2g-abfg^2+a^2g^3\end{aligned}$$

Getting back now to our discriminant problem, with $Q=P'$, which corresponds to $e=3a$, $f=2b$, $g=c$, we obtain the following formula:

$$\begin{aligned}R(P,P')&=27a^3d^2-18a^2bcd+9a^2c^3-18a^2bcd+12ab^3d-6ab^2c^2+18a^2bcd\\&-8ab^3d+3ab^2c^2-6a^2c^3+4ab^2c^2-2ab^2c^2+a^2c^3\end{aligned}$$

By simplifying terms, and dividing by a, we obtain the following formula:

$$-\Delta(P)=27a^2d^2-18abcd+4ac^3+4b^3d-b^2c^2$$

But this gives the formula in the statement, and we are done. □

Still talking degree 3 equations, let us try to solve $P=0$, with $P=aX^3+bX^2+cX+d$ as above. By linear transformations we can assume $a=1,b=0$, and then it is convenient to write $c=3p,d=2q$. Thus, our equation becomes $x^3+3px+2q=0$, and regarding such equations, we have the following famous result, due to Cardano:

THEOREM 7.16. *For a normalized degree 3 equation, namely*

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

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

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

with $w=1,e^{2\pi i/3},e^{4\pi i/3}$ *are the solutions of our equation.*

PROOF. With z as above, by using $(a+b)^3=a^3+b^3+3ab(a+b)$, we have:

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

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

In degree 4 now, we have the following result, which is famous as well:

THEOREM 7.17. *The roots of a normalized degree 4 equation, written as*

$$x^4+6px^2+4qx+3r=0$$

are as follows, with y *satisfying the equation* $(y^2-3r)(y-3p)=2q^2$,

$$x_1=\frac{1}{\sqrt{2}}\left(-\sqrt{y-3p}+\sqrt{-y-3p+\frac{4q}{\sqrt{2y-6p}}}\right)$$

$$x_2=\frac{1}{\sqrt{2}}\left(-\sqrt{y-3p}-\sqrt{-y-3p+\frac{4q}{\sqrt{2y-6p}}}\right)$$

$$x_3=\frac{1}{\sqrt{2}}\left(\sqrt{y-3p}+\sqrt{-y-3p-\frac{4q}{\sqrt{2y-6p}}}\right)$$

$$x_4=\frac{1}{\sqrt{2}}\left(\sqrt{y-3p}-\sqrt{-y-3p-\frac{4q}{\sqrt{2y-6p}}}\right)$$

and with y *being computable via the Cardano formula.*

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

(1) To start with, let us write our equation in the following form:

$$x^4=-6px^2-4qx-3r$$

Now assume that we have a number y satisfying the following equation:

$$(y^2-3r)(y-3p)=2q^2$$

With this magic number y in hand, our equation takes the following form:

$$\begin{aligned}(x^2+y)^2&=x^4+2x^2y+y^2\\&=-6px^2-4qx-3r+2x^2y+y^2\\&=(2y-6p)x^2-4qx+y^2-3r\\&=(2y-6p)x^2-4qx+\frac{2q^2}{y-3p}\\&=\left(\sqrt{2y-6p}\cdot x-\frac{2q}{\sqrt{2y-6p}}\right)^2\end{aligned}$$

(2) Which looks very good, leading us to the following degree 2 equations:

$$x^2+y+\sqrt{2y-6p}\cdot x-\frac{2q}{\sqrt{2y-6p}}=0$$

$$x^2+y-\sqrt{2y-6p}\cdot x+\frac{2q}{\sqrt{2y-6p}}=0$$

Now let us write these two degree 2 equations in standard form, as follows:

$$x^2+\sqrt{2y-6p}\cdot x+\left(y-\frac{2q}{\sqrt{2y-6p}}\right)=0$$

$$x^2-\sqrt{2y-6p}\cdot x+\left(y+\frac{2q}{\sqrt{2y-6p}}\right)=0$$

(3) Regarding the first equation, the solutions there are as follows:

$$x_1=\frac{1}{2}\left(-\sqrt{2y-6p}+\sqrt{-2y-6p+\frac{8q}{\sqrt{2y-6p}}}\right)$$

$$x_2=\frac{1}{2}\left(-\sqrt{2y-6p}-\sqrt{-2y-6p+\frac{8q}{\sqrt{2y-6p}}}\right)$$

As for the second equation, the solutions there are as follows:

$$x_3=\frac{1}{2}\left(\sqrt{2y-6p}+\sqrt{-2y-6p-\frac{8q}{\sqrt{2y-6p}}}\right)$$

$$x_4=\frac{1}{2}\left(\sqrt{2y-6p}-\sqrt{-2y-6p-\frac{8q}{\sqrt{2y-6p}}}\right)$$

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

We still have to compute the number y appearing in the above via Cardano, and the result here, adding to what we already have in Theorem 7.17, is as follows:

THEOREM 7.18 (continuation). *The value of y in the previous theorem is*

$$y=t+p+\frac{a}{t}$$

where the number t is given by the formula

$$t=\sqrt[3]{b+\sqrt{b^2-a^3}}$$

with $a=p^2+r$ and $b=2p^2-3pr+q^2$.

PROOF. The legend has it that this is what comes from Cardano, but depressing and normalizing and solving $(y^2-3r)(y-3p)=2q^2$ makes it for too many operations, so the most pragmatic way is to simply check this equation. With y as above, we have:

$$\begin{aligned}y^2-3r&=t^2+2pt+(p^2+2a)+\frac{2pa}{t}+\frac{a^2}{t^2}-3r\\&=t^2+2pt+(3p^2-r)+\frac{2pa}{t}+\frac{a^2}{t^2}\end{aligned}$$

With this in hand, we have the following computation:

$$\begin{aligned}(y^2-3r)(y-3p)&=\left(t^2+2pt+(3p^2-r)+\frac{2pa}{t}+\frac{a^2}{t^2}\right)\left(t-2p+\frac{a}{t}\right)\\&=t^3+(a-4p^2+3p^2-r)t+(2pa-6p^3+2pr+2pa)\\&\quad+(3p^2a-ra-4p^2a+a^2)\frac{1}{t}+\frac{a^3}{t^3}\\&=t^3+(a-p^2-r)t+2p(2a-3p^2+r)+a(a-p^2-r)\frac{1}{t}+\frac{a^3}{t^3}\\&=t^3+2p(-p^2+3r)+\frac{a^3}{t^3}\end{aligned}$$

Now by using the formula of t in the statement, this gives:

$$\begin{aligned}(y^2-3r)(y-3p)&=b+\sqrt{b^2-a^3}-4p^2+6pr+\frac{a^3}{b+\sqrt{b^2-a^3}}\\&=b+\sqrt{b^2-a^3}-4p^2+6pr+b-\sqrt{b^2-a^3}\\&=2b-4p^2+6pr\\&=2(2p^2-3pr+q^2)-4p^2+6pr\\&=2q^2\end{aligned}$$

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

In degree 5 and more, things become fairly complicated, and we have:

THEOREM 7.19. *There is no general formula for the roots of polynomials of degree $N = 5$ and higher, with the reason for this, coming from Galois theory, being that the group S_5 is not solvable. The simplest numeric example is $P = X^5 - X - 1$.*

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

(1) Given a field F, assume that the roots of $P \in F[X]$ can be computed by using iterated roots, a bit as for the degree 2 equation, or the degree 3 and 4 equations. Then, algebrically speaking, this gives rise to a tower of fields as folows, with $F_0 = F$, and each F_{i+1} being obtained from F_i by adding a root, $F_{i+1} = F_i(x_i)$, with $x_i^{n_i} \in F_i$:

$$F_0 \subset F_1 \subset \ldots \subset F_k$$

(2) In order for Galois theory to apply to this situation, we must make all the extensions normal, which amounts in replacing each $F_{i+1} = F_i(x_i)$ by its extension $K_i(x_i)$, with K_i extending F_i by adding a n_i-th root of unity. Thus, with this replacement, we can assume that the tower in (1) in normal, meaning that all Galois groups are cyclic.

(3) Now by Galois theory, at the level of the corresponding Galois groups we obtain a tower of groups as follows as follows, which is a resolution of the last group G_k, the Galois group of P, in the sense of group theory, in the sense that all quotients are cyclic:

$$G_1 \subset G_2 \subset \ldots \subset G_k$$

As a conclusion, Galois theory tells us that if the roots of a polynomial $P \in F[X]$ can be computed by using iterated roots, then its Galois group $G = G_k$ must be solvable.

(4) In the generic case, the conclusion is that Galois theory tells us that, in order for all polynomials of degree 5 to be solvable, via square roots, the group S_5, which appears there as Galois group, must be solvable, in the sense of group theory. But this is wrong, because the alternating subgroup $A_5 \subset S_5$ is simple, and therefore not solvable.

(5) Finally, regarding the polynomial $P = X^5 - X - 1$, some elementary computations here, based on arithmetic over $\mathbb{F}_2, \mathbb{F}_3$, and involving various cycles of length $2, 3, 5$, show that its Galois group is S_5. Thus, we have our counterexample.

(6) Finally, let us mention that all this shows as well that a random polynomial of degree 5 or higher is not solvable by square roots, and with this being an elementary consequence of the main result from (4), via some standard analysis arguments. □

There is a lot of further interesting theory that can be developed here, following Galois and others. For more on all this, we recommend any number theory book.

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 7.20. *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 7.21. *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 from this that the solutions z of our equation 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 7.22. *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 7.23.

Exercise 7.24.

Exercise 7.25.

Exercise 7.26.

Exercise 7.27.

Exercise 7.28.

Exercise 7.29.

Exercise 7.30.

Bonus exercise.

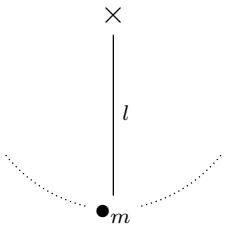
CHAPTER 8

Basic applications

8a. Basic mechanics

Time now for some applications, of the trigonometry that we learned so far. Let us start our discussion with something very basic, namely:

DEFINITION 8.1. *A simple pendulum is a device of type*



consisting of a bob of mass m, attached to a rigid rod of length l.

In order to study the physics of the pendulum, which can easily lead to a lot of complicated computations, when approached with bare hands, the most convenient is to use the notion of energy. For a particle moving under the influence of a force F, the position x, speed v and acceleration a are related by the following formulae:

$$v = \dot{x} \quad , \quad a = \dot{v} = \ddot{x} \quad , \quad F = ma$$

The kinetic energy of our particle is then given by the following formula:

$$T = \frac{mv^2}{2}$$

By differentiating with respect to time t, we obtain the following formula:

$$\dot{T} = mv\dot{v} = mva = Fv$$

Now by integrating, also with respect to t, this gives the following formula:

$$T = \int Fvdt = \int F\dot{x}dt = \int Fdx$$

But this suggests to define the potential energy V by the following formula, up to a constant, with the derivative being with respect to the space variable x:

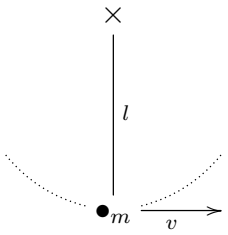
$$V' = -F$$

Indeed, we know from the above that we have $T' = F$, so if we define the total energy to be $E = T + V$, then this total energy is constant, as shown by:

$$E' = T' + V' = 0$$

Very nice all this, and by getting back now to the pendulum from Definition 8.1, we can have this understood with not many computations involved, as follows:

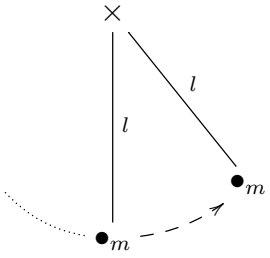
THEOREM 8.2. *For a pendulum starting with speed v from the equilibrium position,*



the motion will be confined if $v^2 < 4gl$, and circular if $v^2 > 4gl$.

PROOF. There are many ways of proving this result, along with working out several other useful related formulae, for which we will refer to the proof below, and with a quite elegant approach to this, using no computations or almost, being as follows:

(1) Let us first examine what happens when the bob has traveled an angular distance $\theta > 0$, with respect to the vertical. The picture here is as follows:



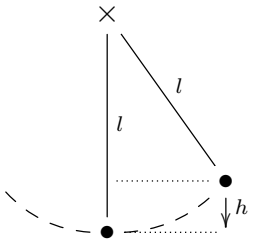
The distance traveled is then $x = l\theta$. As for the force acting, this is $F_{total} = mg$ oriented downwards, with the component alongside x being given by:

$$\begin{aligned} F &= -||F_{total}|| \sin\theta \\ &= -mg\sin\theta \\ &= -mg\sin\left(\frac{x}{l}\right) \end{aligned}$$

(2) But with this, we can compute the potential energy. With the convention that this vanishes at the equilibrium position, $V(0) = 0$, we obtain the following formula:

$$\begin{aligned} V' = -F &\implies V' = mg \sin\left(\frac{x}{l}\right) \\ &\implies V = mgl\left(1 - \cos\left(\frac{x}{l}\right)\right) \\ &\implies V = mgl(1 - \cos\theta) \end{aligned}$$

(3) Alternatively, in case this sounds too wizarding, we can compute the potential energy in the old fashion, by letting the bob fall, the picture being as follows:



The height of the fall is then $h = l - l\cos\theta$, and since for this fall the force is constant, $\mathcal{F} = -mg$, we obtain the following formula for the potential energy:

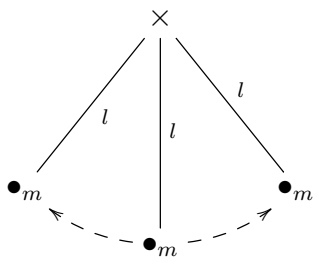
$$\begin{aligned} V' = -\mathcal{F} &\implies V' = mg \\ &\implies V = mgh \\ &\implies V = mgl(1 - \cos\theta) \end{aligned}$$

Summarizing, one way or another we have our formula for the potential energy V.

(4) Now comes the discussion. The motion will be confined when the initial kinetic energy, namely $E = mv^2/2$, satisfies the following condition:

$$\begin{aligned} E < \sup_\theta V = 2mgl &\iff \frac{mv^2}{2} < 2mgl \\ &\iff v^2 < 4gl \end{aligned}$$

In this case, the motion will be confined between two angles $-\theta, \theta$, as follows:



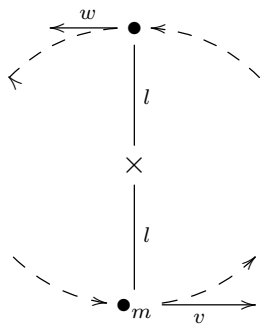
To be more precise here, the two extreme angles $-\theta, \theta \in (-\pi, \pi)$ can be explicitly computed, as being solutions of the following equation:

$$\begin{aligned} V = E &\iff mgl(1-\cos\theta) = \frac{mv^2}{2} \\ &\iff 1-\cos\theta = \frac{v^2}{2gl} \end{aligned}$$

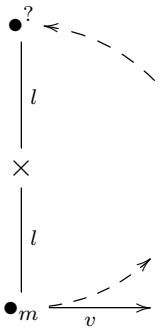
(5) Regarding now the case $v^2 > 4gl$, here the bob will certainly reach the upwards position, with the speed $w > 0$ there being given by the following formula:

$$\begin{aligned} \frac{mw^2}{2} = E - 2mgl &\implies \frac{mw^2}{2} = \frac{mv^2}{2} - 2mgl \\ &\implies w^2 = v^2 - 4gl \\ &\implies w = \sqrt{v^2 - 4gl} \end{aligned}$$

Thus, with the convention in the statement for v, that is, going to the right, the motion of the pendulum will be counterclockwise circular, and perpetual:



(6) Finally, in the case $v^2 = 4gl$, the bob will also reach the upwards position, but with speed $w = 0$ there, and then, at least theoretically, will remain there:



(7) Actually, it is quite interesting in this latter situation, $v^2 = 4gl$, to further speculate on what can happen, when making our problem more realistic. For instance, we can add to our setting the assumption that when the bob is stuck on top, with speed 0, there is a

33% chance for it to keep going, to the left, a 33% chance for it to come back, to the right, and a 33% chance for it to remain stuck. In this case there are infinitely many possible trajectories, which are best investigated by using probability. Welcome to chaos.

(8) As a final comment, yes I know that the figures in (7) don't add up to 100%. This is because there is as well a remaining 1% possibility, where a relativistic black cat appears, with a continuous effect on the bob, via a paw slap, when on top, with speed $w' \in (0.3c, 0.7c)$, with c being the speed of light. In this case, the set of possible trajectories becomes uncountable, and is again best investigated by using probability. □

Many other things can be said, along the above lines.

8b. Electrostatics

Time now for electricity. Let us start with something very basic, namely:

FACT 8.3. *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 8.3 is indeed more than borderline, as axiomatics for a theory, we have:

FACT 8.4 (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 8.5. *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 8.6. *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 8.7. *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 $F=QE$.

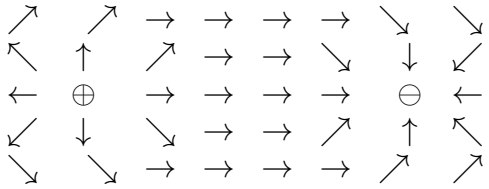
With the above definitions in hand, it is most convenient now to forget about the charges, and focus on the study of the corresponding electric fields E.

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:

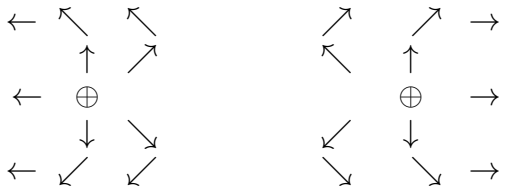
DEFINITION 8.8. *The field lines of an electric field $E:\mathbb{R}^3\to\mathbb{R}^3$ are the oriented curves $\gamma\subset\mathbb{R}^3$ pointing at every point $x\in\mathbb{R}^3$ at the direction of the field, $E(x)\in\mathbb{R}^3$.*

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 8.9. *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. □

Let us introduce now a key definition, as follows:

DEFINITION 8.10. *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.

We have the following key result, due to Gauss:

THEOREM 8.11 (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 is something quite tricky, 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) Let us assume now, 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 then 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 the following computation:

$$\begin{aligned}\Phi_E(S) &= \Phi_E(S-\cup U_i)+\Phi_E(\cup U_i)\\ &= 0+\Phi_E(\cup U_i)\\ &= \sum_i \Phi_E(U_i)\\ &= \sum_i \frac{q_i}{\varepsilon_0}\\ &= \frac{Q_{enc}}{\varepsilon_0}\end{aligned}$$

(3) 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 (1). Finally, the end makes use of some standard calculus. □

8c. 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 8.12. *An algebraic curve in $\mathbb{R}^2$ is the vanishing set*

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

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

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

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

PROPOSITION 8.13. *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 8.14. *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 8.15. *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 8.16. *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 8.17. *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 8.18. *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 8.19. *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 8.20. *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 8.21. *The Cayley sextic, given in polar coordinates by*

$$r = a \cos^3\left(\frac{\theta}{3}\right)$$

makes the dream of everyone come true, by looking like a self-intersecting heart.

PROOF. As before, picture mandatory. With $z = re^{i\theta}$ and $u = z^{1/3}$ we have:

$$\begin{aligned}
r = a \cos^3\left(\frac{\theta}{3}\right) &\iff ar \cos^3\left(\frac{\theta}{3}\right) = r^2 \\
&\iff a\left(\frac{u+\bar{u}}{2}\right)^3 = r^2 \\
&\iff a(u^3 + \bar{u}^3 + 3u\bar{u}(u+\bar{u})) = 8r^2 \\
&\iff 3au\bar{u} \cdot \frac{u+\bar{u}}{2} = 4r^2 - ax \\
&\iff 27a^3r^6 \cdot \frac{r^2}{a} = (4r^2 - ax)^3 \\
&\iff 27a^2(x^2+y^2)^2 = (4x^2 + 4y^2 - ax)^3
\end{aligned}$$

Thus, we have indeed a sextic, as claimed. □

8d. Field lines

In order to further advance, let us go back to the various plane curves discussed above. 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 8.22. *The sinusoidal spirals, which are as follows,*

$$r^n = a^n \cos n\theta$$

with $a \neq 0$ and $n \in \mathbb{Q} - \{0\}$, include the following curves:

(1) $n = -1$ *line.*
(2) $n = 1$ *circle,* $n = -1/2$ *parabola,* $n = -2$ *hyperbola.*
(3) $n = -3$ *Humbert cubic,* $n = -1/3$ *Tschirnhausen curve.*
(4) $n = 1/2$ *cardioid,* $n = 2$ *Bernoulli lemniscate.*
(5) $n = 3$ *Kiepert trefoil,* $n = 1/3$ *Cayley sextic.*

PROOF. We first have to prove that the sinusoidal spirals are indeed algebraic curves. But this is best done by using the complex coordinate $z = re^{i\theta}$, as follows:

$$\begin{aligned} r^n = a^n \cos n\theta &\iff r^n \cos n\theta = \left(\frac{r^2}{a}\right)^n \\ &\iff z^n + \bar{z}^n = 2\left(\frac{z\bar{z}}{a}\right)^n \\ &\iff (x+iy)^n + (x-iy)^n = 2\left(\frac{x^2+y^2}{a}\right)^n \end{aligned}$$

As a first observation now, in the case $n \in \mathbb{N}$ we can simply use the binomial formula, and we get an algebraic equation of degree $2n$, as follows:

$$\sum_{k=0}^{[n/2]} (-1)^k \binom{n}{2k} x^{n-2k} y^{2k} = \left(\frac{x^2+y^2}{a}\right)^n$$

In general, things are a bit more complicated, as shown for instance by our computation for the Cayley sextic. However, the same idea as there applies, and we are led in this way to the equation of an algebraic curve, as claimed. Regarding now the examples:

(1) At $n = -1$ the equation is as follows, producing a line:

$$r\cos\theta = a \iff x = a$$

(2) At $n = 1$ the equation is as follows, producing a circle:

$$r = a\cos\theta \iff r^2 = ax \iff x^2 + y^2 = ax$$

(3) At $n = -1/2$ the equation is as follows, producing a parabola:

$$a = r\cos^2(\theta/2) \iff r + x = 2a \iff y^2 = 4a(a-x)$$

(4) At $n=-2$ the equation is as follows, producing a hyperbola:

$$a^2=r\cos^2 2\theta\iff a^2=2x^2-r^2\iff (x+y)(x-y)=a^2$$

(5) At $n=-3$ the equation is as follows, producing a curve with 3 components, which looks like some sort of "trivalent hyperbola", called Humbert cubic:

$$r^3\cos 3\theta=a^3\iff z^3+\bar{z}^3=2a^3\iff x^3-3xy^2=a^3$$

(6) As for the other curves, this follows from our various formulae above. □

Let us study now more in detail the sinusoidal spirals. We first have:

PROPOSITION 8.23. *The sinusoidal spirals, which with $z=x+iy$ are*

$$z^n+\bar{z}^n=2\left(\frac{z\bar{z}}{a}\right)^n$$

with $a\neq 0$ and $n\in\mathbb{Q}-\{0\}$, are as follows:

(1) *With $n=-m$, $m\in\mathbb{N}$, the equation is $z^m+\bar{z}^m=2a^m$, degree m.*
(2) *With $n=m$, $m\in\mathbb{N}$, the equation is $z^m+\bar{z}^m=2(z\bar{z}/a)^m$, degree $2m$.*
(3) *With $n=-1/m$, $m\in\mathbb{N}$, the equation is $(z^{1/m}+\bar{z}^{1/m})^m=2^m a$.*
(4) *With $n=1/m$, $m\in\mathbb{N}$, the equation is $(z^{1/m}+\bar{z}^{1/m})^m=2^m z\bar{z}/a$.*

PROOF. This is something self-explanatory, the details being as follows:

(1) With $n=-m$ and $m\in\mathbb{N}$ as in the statement, the equation is, as claimed:

$$z^{-m}+\bar{z}^{-m}=2\left(\frac{z\bar{z}}{a}\right)^{-m}\iff z^m+\bar{z}^m=2a^m$$

(2) This is an empty statement, just a matter of using the new variable $m=n$.

(3) With $n=-1/m$ and $m\in\mathbb{N}$ as in the statement, the equation is, as claimed:

$$\begin{aligned}z^{-1/m}+\bar{z}^{-1/m}=2\left(\frac{z\bar{z}}{a}\right)^{-1/m}&\iff z^{1/m}+\bar{z}^{1/m}=2a^{1/m}\\&\iff (z^{1/m}+\bar{z}^{1/m})^m=2^m a\end{aligned}$$

(4) With $n=1/m$ and $m\in\mathbb{N}$ as in the statement, the equation is, as claimed:

$$z^{1/m}+\bar{z}^{1/m}=2\left(\frac{z\bar{z}}{a}\right)^{1/m}\iff (z^{1/m}+\bar{z}^{1/m})^m=2^m\cdot\frac{z\bar{z}}{a}$$

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

Observe that in the fractionary cases, $n=\pm 1/m$, the equations in the above statement are not polynomial in x,y, unless at very small values of m. To be more precise:

(1) In the case $n=-1/m$, we certainly have at $m=1,2,3$ the $d=1$ line, $d=2$ parabola, and $d=3$ Tschirnhausen curve, but at $m=4$ things change, with the equation

$(z^{1/4}+\bar{z}^{1/4})^4=16a$ being no longer polynomial in x,y, and requiring a further square operation to make it polynomial, and therefore leading to a curve of degree $d=8$.

(2) As for the case $n=1/m$, this is more complicated, with the data that we have at $m=1,2,3$, namely the $d=2$ circle, $d=3$ cardioid, and $d=6$ Cayley sextic, being not very good, and with things getting even more complicated at $m=4$ and higher.

In short, things quite complicated, and the general case, $n=\pm p/q$ with $p,q\in\mathbb{N}$, is certainly even more complicated. Instead of insisting on this, let us focus now on the simplest sinusoidal spirals that we have, namely those with $n=\pm m$, with $m\in\mathbb{N}$.

The point indeed is that the sinusoidal spirals with $n\in\mathbb{N}$ are also part of another remarkable family of plane algebraic curves, going back to Cassini, as follows:

THEOREM 8.24. *The polynomial lemniscates, which are as follows,*

$$|P(z)|=b^n$$

with $P\in\mathbb{C}[X]$ having n distinct roots, and $b>0$, include the following curves:

(1) *The sinusoidal spirals with $n\in\mathbb{N}$, including the $n=1$ circle, $n=2$ Bernoulli lemniscate, and $n=3$ Kiepert trefoil.*

(2) *The Cassini ovals, which are the quartics given by $|z+c|\cdot|z-c|=b^2$, covering too the Bernoulli lemniscate, appearing at $b=c$.*

PROOF. This is something quite self-explanatory, the details being as follows:

(1) Regarding the sinusoidal spirals with $n\in\mathbb{N}$, their equation is, with $a^n=2c^n$:

$$\begin{aligned}z^n+\bar{z}^n=2\left(\frac{z\bar{z}}{a}\right)^n&\iff c^n(z^n+\bar{z}^n)=(z\bar{z})^n\\&\iff(z^n-c^n)(\bar{z}^n-c^n)=c^{2n}\\&\iff|z^n-c^n|=c^n\end{aligned}$$

(2) Regarding the Cassini ovals, these correspond to the case where the polynomial $P\in\mathbb{C}[X]$ has degree 2, and we already know from the above that these cover the Bernoulli lemniscate. In general, the equation for the Cassini ovals is:

$$\begin{aligned}|z+c|\cdot|z-c|=b^2&\iff|z^2-c^2|=b^2\\&\iff(z^2-c^2)(\bar{z}^2-c^2)=b^4\\&\iff(z\bar{z})^2-c^2(z^2+\bar{z}^2)+c^4=b^4\\&\iff(x^2+y^2)^2-c^2(x^2-y^2)+c^4=b^4\\&\iff(x^2+y^2)^2=c^2(x^2-y^2)+b^4-c^4\end{aligned}$$

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

The polynomial lemniscates can be geometrically understood as follows:

THEOREM 8.25. *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 8.26. *Given points in the plane $c_1, \ldots, c_n \in \mathbb{C}$ and a number $d \in \mathbb{R}$, construct the associated stelloid as being the set of points $z \in \mathbb{C}$ verifying*

$$\frac{1}{n} \sum_{k=1}^{n} \alpha_v (z - c_i) = d$$

with α_v denoting the angle with respect to a direction v. Then the stelloid is an algebraic curve, not depending on v, and at the level of examples we have the sinusoidal spirals with $n \in -\mathbb{N}$, including the $n = -1$ line, $n = -2$ hyperbola, and $n = -3$ Humbert cubic.

PROOF. All this is quite self-explanatory, and we will leave the verification of the various generalities regarding the stelloids, as well as the verification of the relation with the sinusoidal spirals with $n \in -\mathbb{N}$, as an instructive exercise. As a bonus exercise, try understanding the precise relation between stelloids, and polynomial lemniscates. □

So long for plane algebraic curves. Needless to say, all the above is old-style, first class mathematics, having countless applications. For instance when doing classical mechanics or electrodynamics, you will certainly meet polynomial lemniscates and stelloids, when looking at the field lines. Also, the image of any circle passing though 0 by $z \to z^2$ is a cardioid, and the famous Mandelbrot set is organized around such a cardioid.

8e. Exercises

Exercises:

Exercise 8.27.

Exercise 8.28.

Exercise 8.29.

Exercise 8.30.

Exercise 8.31.

Exercise 8.32.

Exercise 8.33.

Exercise 8.34.

Bonus exercise.

Part III

Calculus methods

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*

$$f'(x) = \lim_{t \to 0} \frac{f(x+t) - f(x)}{t}$$

called derivative of f at that point x, exists.

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

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

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:

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

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

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

Thus, the limit in Definition 9.1 does not converge, 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:

$$f'(x)=\lim_{t\to 0}\frac{f(x+t)-f(x)}{t}$$

By multiplying by t, we obtain that we have, once again in the $t\to 0$ limit:

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

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

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

THEOREM 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,*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$f''(x)\geq 0$$

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

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

THEOREM 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:

$$P(t) = a_0 + a_1 t + \ldots + a_n t^n$$

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

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

In order to prove now that this approximation holds indeed, we can use L'Hôpital's rule, applied several times, as in the proof of Theorem 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:

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

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

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

Thus, Theorem 9.24 at $p = \pm 1/2$ 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 trivially 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 $e^{a+b} = e^a e^b$, that we know from analysis. □

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

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

THEOREM 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:* $(re^{it})^a = r^a e^{ita}$.
(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:

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

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

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

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 $e^{x+y}=e^xe^y$ 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 $e^{x+iy}=e^xe^{iy}$ 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 $e^{x+iy}=e^xe^{iy}$, 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:

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

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:

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

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:

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

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

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

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 $e^x = \sum_k \frac{x^k}{k!}$. □

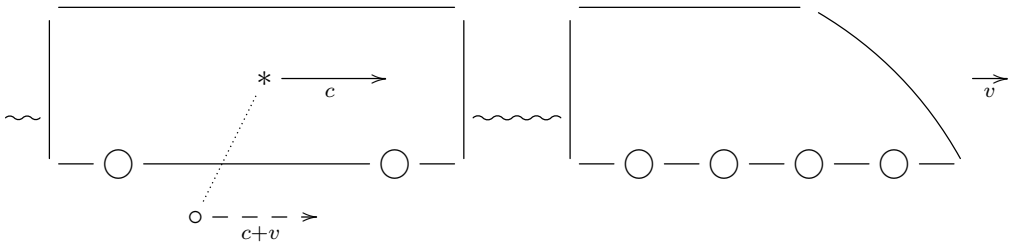
Ready for some physics? Based on experiments by Fizeau, then Michelson-Morley and others, and some physics by Maxwell and Lorentz too, Einstein came upon:

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

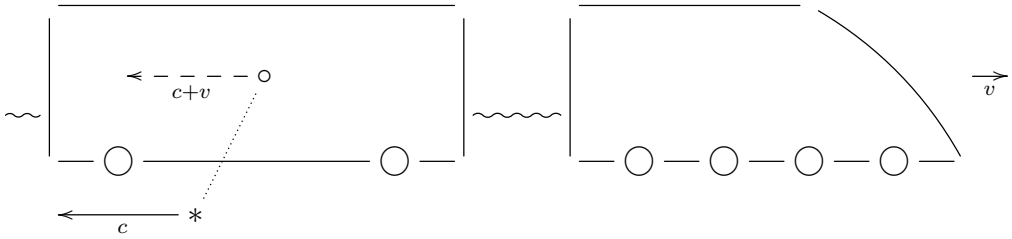
(1) *Light travels in vacuum at a finite speed,* $c < \infty$.
(2) *This speed* c *is the same for all inertial observers.*
(3) *In non-vacuum, the light speed is lower,* $v < c$.
(4) *Nothing can travel faster than light,* $v \not> c$.

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 $\pm1\to i$*, with negatives on the left, and positives on the right.*

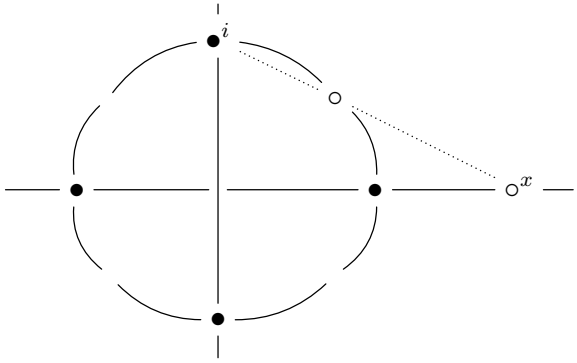
PROOF. Regarding the wrapping, as indicated, this is given by:

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

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

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

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



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

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

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

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

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

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

So, what to do? Obviously, trash Proposition 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. Integration theory

There are several possible viewpoints on the integral, which are all useful, and good to know. To start with, we have something very simple, as follows:

DEFINITION 11.1. *The integral of a continuous function $f : [a, b] \to \mathbb{R}$, denoted*

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

is the area below the graph of f, signed $+$ where $f \geq 0$, and signed $-$ where $f \leq 0$.

Here it is of course understood that the area in question can be computed, and with this being something quite subtle, that we will get into later. In order to compute now integrals, we can use our geometric knowledge. Here are some basic results:

PROPOSITION 11.2. *We have the following results:*

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

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

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

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

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

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

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

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

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

As an interesting observation, (2) in the above result makes it quite clear that f does not necessarily need to be continuous, in order to talk about its integral. Indeed, assuming that f is piecewise linear on $[a = a_1, a_2, \ldots, a_n = b]$, but not necessarily continuous, we can still talk about its integral, in the obvious way, exactly as in Definition 11.1, and we have an explicit formula for this integral, generalizing the one found in (2), namely:

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

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

DEFINITION 11.3. *We say that a function $f : [a, b] \to \mathbb{R}$ is integrable when the area below its graph is computable. In this case we denote by*

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

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

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

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

PROPOSITION 11.4. *We have the following formulae,*

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

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

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

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

Moving ahead now, passed the above results, we must do some analysis, in order to compute integrals. This is something quite tricky, and we have here:

THEOREM 11.5. *We have the Riemann integration formula,*

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

which can serve as a definition for the integral.

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

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

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

Observe that the above formula suggests that $\int_a^b f(x)dx$ is the length of the interval $[a, b]$, namely $b - a$, times the average of f on the interval $[a, b]$. Thinking a bit, this is indeed something true, with no need for Riemann sums, coming directly from Definition 11.1, because area means side times average height. Thus, we can formulate:

THEOREM 11.6. *The integral of a function $f : [a, b] \to \mathbb{R}$ is given by*

$$\int_a^b f(x)dx = (b-a) \times A(f)$$

where $A(f)$ is the average of f over the interval $[a, b]$.

PROOF. As explained above, this is clear from Definition 11.1, via some geometric thinking. Alternatively, this is something which certainly comes from Theorem 11.5. □

The point of view in Theorem 11.6 is something quite useful, and as an illustration for this, let us review the results that we already have, by using this interpretation. First, we have the formula for linear functions from Proposition 11.2, namely:

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

But this formula is totally obvious with our new viewpoint, from Theorem 11.6. The same goes for the results in Proposition 11.4, which become even more obvious with the viewpoint from Theorem 11.6. However, not everything trivializes in this way, and the result which is left, namely the formula $\int_{-1}^{1} \sqrt{1-x^2}\, dx = \pi/2$ from Proposition 11.2 (3), not only does not trivialize, but becomes quite opaque with our new philosophy.

In short, modesty. Integration is a quite delicate business, and we have several equivalent points of view on what an integral means, and all these points of view are useful, and must be learned, with none of them being clearly better than the others.

Going ahead with more interpretations of the integral, we have:

THEOREM 11.7. *We have the Monte Carlo integration formula,*

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

with $x_1, \ldots, x_N \in [a, b]$ being random.

PROOF. We recall from Theorem 11.5 that the idea is that we have a formula as follows, with the points $x_1, \ldots, x_N \in [a, b]$ being uniformly distributed:

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

But this works as well when the points $x_1, \ldots, x_N \in [a, b]$ are randomly distributed, for somewhat obvious reasons, and this gives the result. □

Observe that Monte Carlo integration works better than Riemann, when it comes to computing as usual, by estimating, and refining the estimate. Also, Monte Carlo is smarter than Riemann, because the symmetries of the function can fool Riemann, but not Monte Carlo. All this is good to know, say when integrating by using a computer.

Finally, here is one more useful interpretation of the integral:

THEOREM 11.8. *The integral of a function $f : [a, b] \to \mathbb{R}$ is given by*

$$\int_a^b f(x)dx = (b-a) \times E(f)$$

where $E(f)$ is the expectation of f, regarded as random variable.

PROOF. This is just some sort of fancy reformulation of Theorem 11.7, the idea being that what we can "expect" from a random variable is of course its average. We will be back to this later in this book, when systematically discussing probability theory. □

Our purpose now will be to understand which functions $f : \mathbb{R} \to \mathbb{R}$ are integrable, and how to compute their integrals. For this purpose, the Riemann formula in Theorem 11.5 will be our favorite tool. Let us begin with some theory. We first have:

THEOREM 11.9. *The following functions are integrable:*

(1) *The piecewise continuous functions.*
(2) *The piecewise monotone functions.*

PROOF. This is indeed something quite standard, as follows:

(1) It is enough to prove the first assertion for a function $f : [a, b] \to \mathbb{R}$ which is continuous, and our claim here is that this follows from the uniform continuity of f. To be more precise, given $\varepsilon > 0$, let us choose $\delta > 0$ such that the following happens:

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

In order to prove the result, let us pick two divisions of $[a, b]$, as follows:

$$I = [a = a_1 < a_2 < \ldots < a_n = b]$$

$$I' = [a = a'_1 < a'_2 < \ldots < a'_m = b]$$

Our claim, which will prove the result, is that if these divisions are sharp enough, of resolution $< \delta/2$, then the associated Riemann sums $\Sigma_I(f), \Sigma_{I'}(f)$ are close within ε:

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

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

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

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

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

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

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

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

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

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

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

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

$$|f(a_i) - f(a'_j)| < \varepsilon$$

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

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

Thus our two Riemann sums are close enough, provided that they are both chosen to be fine enough, and this finishes the proof of the first assertion.

(5) Regarding now the second assertion, this is something more technical, that we will not really need in what follows. We will leave the proof here, which uses similar ideas to those in the proof of (1) above, namely subdivisions and estimates, as an exercise. □

Going ahead with more theory, let us establish some abstract properties of the integration operation. We already know from Proposition 11.4 that the integrals behave well with respect to sums and multiplication by scalars. Along the same lines, we have:

THEOREM 11.10. *The integrals behave well with respect to taking limits,*

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

and with respect to taking infinite sums as well,

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

with both these formulae being valid, undwer mild assumptions.

PROOF. This is something quite standard, by using the standard general theory for the sequences and series of functions. To be more precise, (1) follows by using this quite standard material, via Riemann sums, and then (2) follows as a particular case of (1). We will leave the clarification of all this as an instructive exercise. □

Finally, still at the general level, let us record as well the following result:

THEOREM 11.11. *Given a continuous function $f:[a,b]\to\mathbb{R}$, we have*

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

with this being called mean value property.

PROOF. Our claim is that this follows from the following trivial estimate:

$$\min(f) \leq f \leq \max(f)$$

Indeed, by integrating this over $[a, b]$, we obtain the following estimate:

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

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

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

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

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

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

Next, we have the following key result, called fundamental theorem of calculus:

THEOREM 11.12. *Given a continuous function $f : [a, b] \to \mathbb{R}$, if we set*

$$F(x) = \int_a^x f(s)ds$$

then $F' = f$. That is, the derivative of the integral is the function itself.

PROOF. This follows from the Riemann integration picture, and more specifically, from the mean value property from Theorem 11.11. Indeed, we have:

$$\frac{F(x+t) - F(x)}{t} = \frac{1}{t}\int_x^{x+t} f(x)dx$$

On the other hand, our function f being continuous, by using the mean value property from Theorem 11.11, we can find a number $c \in [x, x+t]$ such that:

$$\frac{1}{t}\int_x^{x+t} f(x)dx = f(x)$$

Thus, putting our formulae together, we conclude that we have:

$$\frac{F(x+t) - F(x)}{t} = f(c)$$

Now with $t \to 0$, no matter how the number $c \in [x, x+t]$ varies, one thing that we can be sure about is that we have $c \to x$. Thus, by continuity of f, we obtain:

$$\lim_{t \to 0} \frac{F(x+t) - F(x)}{t} = f(x)$$

But this means exactly that we have $F' = f$, and we are done. □

We have as well the following result, which is something equivalent, and a hair more beautiful, also called fundamental theorem of calculus:

THEOREM 11.13. *Given a function $F : \mathbb{R} \to \mathbb{R}$, we have*

$$\int_a^b F'(x)dx = F(b) - F(a)$$

for any interval $[a, b]$.

PROOF. As already mentioned, this is something which follows from Theorem 11.12, and is in fact equivalent to it. Indeed, consider the following function:

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

By using Theorem 11.12 we have $G' = F'$, and so our functions F, G differ by a constant. But with $s = a$ we have $G(a) = 0$, and so the constant is $F(a)$, and we get:

$$F(s) = G(s) + F(a)$$

Now with $s = b$ this gives $F(b) = G(b) + F(a)$, which reads:

$$F(b) = \int_a^b F'(x)dx + F(a)$$

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

As a first illustration for all this, solving our previous problems, we have:

THEOREM 11.14. *We have the following integration formulae,*

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

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

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

all obtained, in case you ever forget them, via the fundamental theorem of calculus.

PROOF. We already know some of these formulae, but the best is to do everything, using the fundamental theorem of calculus. The computations go as follows:

(1) With $F(x) = x^{p+1}$ we have $F'(x) = px^p$, and we get, as desired:

$$\int_a^b px^p\, dx = b^{p+1} - a^{p+1}$$

(2) Observe first that the formula (1) does not work at $p = -1$. However, here we can use $F(x) = \log x$, having as derivative $F'(x) = 1/x$, which gives, as desired:

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

(3) With $F(x) = \cos x$ we have $F'(x) = -\sin x$, and we get, as desired:

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

(4) With $F(x) = \sin x$ we have $F'(x) = \cos x$, and we get, as desired:

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

(5) With $F(x) = e^x$ we have $F'(x) = e^x$, and we get, as desired:

$$\int_a^b e^x\,dx = e^b - e^a$$

(6) This is something more tricky. We are looking for a function satisfying:

$$F'(x) = \log x$$

This does not look doable, but fortunately the answer to such things can be found on the internet. But, what if the internet connection is down? So, let us think a bit, and try to solve our problem. Speaking logarithm and derivatives, what we know is:

$$(\log x)' = \frac{1}{x}$$

But then, in order to make appear log on the right, the idea is quite clear, namely multiplying on the left by x. We obtain in this way the following formula:

$$(x\log x)' = 1\cdot\log x + x\cdot\frac{1}{x} = \log x + 1$$

We are almost there, all we have to do now is to substract x from the left, as to get:

$$(x\log x - x)' = \log x$$

But this this formula in hand, we can go back to our problem, and we get the result. □

Getting back now to theory, inspired by the above, let us formulate:

Definition 11.15. *Given f, we call primitive of f any function F satisfying:*

$$F' = f$$

We denote such primitives by $\int f$, and also call them indefinite integrals.

Observe that the primitives are unique up to an additive constant, in the sense that if F is a primitive, then so is $F+c$, for any $c\in\mathbb{R}$, and conversely, if F,G are two primitives, then we must have $G=F+c$, for some $c\in\mathbb{R}$, with this latter fact coming from a result from chapter 9, saying that the derivative vanishes when the function is constant.

As for the convention at the end, $F=\int f$, this comes from the fundamental theorem of calculus, which can be written as follows, by using this convention:

$$\int_a^b f(x)dx=\left(\int f\right)(b)-\left(\int f\right)(a)$$

By the way, observe that there is no contradiction here, coming from the indeterminacy of $\int f$. Indeed, when adding a constant $c\in\mathbb{R}$ to the chosen primitive $\int f$, when conputing the above difference the c quantities will cancel, and we will obtain the same result.

We can now reformulate Theorem 11.14 in a more digest form, as follows:

THEOREM 11.16. *We have the following formulae for primitives,*

$$\int x^p=\frac{x^{p+1}}{p+1}\quad,\quad\int\frac{1}{x}=\log x$$

$$\int\sin x=-\cos x\quad,\quad\int\cos x=\sin x$$

$$\int e^x=e^x\quad,\quad\int\log x=x\log x-x$$

allowing us to compute the corresponding definite integrals too.

PROOF. Here the various formulae in the statement follow from Theorem 11.14, or rather from the proof of Theorem 11.14, or even from chapter 9, for most of them, and the last assertion comes from the integration formula given after Definition 11.15. □

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

THEOREM 11.17. *We have the formula*

$$\int f'g+\int fg'=fg$$

called integration by parts.

PROOF. This follows by integrating the Leibnitz formula, namely:

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

Indeed, with our convention for primitives, this gives the above formula. □

It is then possible to pass to usual integrals, and we obtain a formula here as well, as follows, also called integration by parts, with the convention $[\varphi]_a^b = \varphi(b) - \varphi(a)$:

$$\int_a^b f'g + \int_a^b fg' = \Big[fg\Big]_a^b$$

In practice, the most interesting case is that when fg vanishes on the boundary $\{a, b\}$ of our interval, leading to the following formula:

$$\int_a^b f'g = -\int_a^b fg'$$

Examples of this usually come with $[a, b] = [-\infty, \infty]$, and more on this later. Now still at the theoretical level, we have as well the following result:

THEOREM 11.18. *We have the change of variable formula*

$$\int_a^b f(x)dx = \int_c^d f(\varphi(t))\varphi'(t)dt$$

where $c = \varphi^{-1}(a)$ *and* $d = \varphi^{-1}(b)$.

PROOF. This follows with $f = F'$, from the following differentiation rule, that we know from chapter 9, and whose proof is something elementary:

$$(F\varphi)'(t) = F'(\varphi(t))\varphi'(t)$$

Indeed, by integrating between c and d, we obtain the result. □

11b. More about e

s

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

THEOREM 11.19. *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 $\exp(x) = \sum c_k x^k$, 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:

$$\begin{aligned}
N = 2 &\implies 2.5 < e < 2.75 \\
N = 3 &\implies 2.666\ldots < e < 2.722\ldots \\
N = 4 &\implies 2.70833\ldots < e < 2.71875\ldots \\
N = 5 &\implies 2.71666\ldots < e < 2.71833\ldots \\
N = 6 &\implies 2.71805\ldots < e < 2.71828\ldots \\
N = 7 &\implies 2.71825\ldots < e < 2.71828\ldots
\end{aligned}$$

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 11.20. *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:

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

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

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

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

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

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

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

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:

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

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

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

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. □

11c. 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 11.21. *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 11.22. *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. □

11d. Special functions

Special functions.

11e. Exercises

Exercises:

EXERCISE 11.23.

EXERCISE 11.24.

EXERCISE 11.25.

EXERCISE 11.26.

EXERCISE 11.27.

EXERCISE 11.28.

EXERCISE 11.29.

EXERCISE 11.30.

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):

$$Z = (-1)^n \left(\frac{q-1}{2} \right)!$$

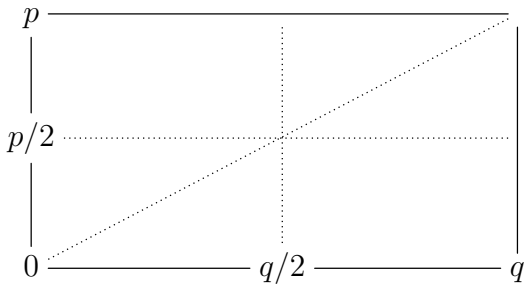
(5) But this is what we need, because when comparing with what we have in (2), we obtain the following formula, which is exactly the one claimed by the Gauss lemma:

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

(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:

$$\left(\frac{a}{2}\right)=\begin{cases}1 & \text{if } a=\pm1(8)\\ 0 & \text{if } a=0(2)\\ -1 & \text{if } a=\pm3(8)\end{cases}$$

The answer for the second symbol is as follows:

$$\left(\frac{a}{-1}\right)=\begin{cases}1 & \text{if } a\geq 0\\ -1 & \text{if } a<0\end{cases}$$

As for the answer for the third symbol, this is as follows:

$$\left(\frac{a}{0}\right)=\begin{cases}1 & \text{if } a=\pm1\\ 0 & \text{if } a\neq\pm1\end{cases}$$

And we will leave 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:

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

(6) Now let us look at $w^p + w^{-p}$, as usual complex number. Since $w = e^{\pi i/4}$, this quantity will depend only on p modulo 8, and more precisely, we have:

$$w^p + w^{-p} = \begin{cases} w + w^{-1} & \text{if } p = \pm 1(8) \\ -w - w^{-1} & \text{if } p = \pm 3(8) \end{cases}$$

Thus $w^p + w^{-p} = \pm t$, with the sign depending on p modulo 8, and more specifically:

$$w^p + w^{-p} = (-1)^{\frac{p^2-1}{8}} t$$

(7) Time now to put everything together. By combining (4,5,6) we obtain:

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

By dividing by t, this gives the following formula:

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

But the mod p symbol can now be dropped, because our equality is between two ± 1 quantities, and we obtain the formula in the statement. □

With the same idea, we can prove as well the quadratic reciprocity theorem:

THEOREM 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:

$$\frac{1}{6}\quad,\quad -\frac{1}{30}\quad,\quad \frac{1}{42}\quad,\quad -\frac{1}{30}\quad,\quad \frac{5}{66}\quad,\quad -\frac{691}{2730}\quad,\quad \frac{7}{6}\quad,\quad \ldots$$

(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:

$$\begin{aligned} xg' &= x(-\cot x - x \cot' x) \\ &= x(-\cot x + x \cot^2 x + x) \\ &= g + g^2 + x^2 \end{aligned}$$

(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:

$$\begin{aligned}\int_0^1\int_0^1\frac{1}{1-xy}\,dxdy&=\int_0^1\left[-\frac{\log(1-xy)}{y}\right]_0^1dy\\&=-\int_0^1\frac{\log(1-y)}{y}\,dy\\&=-\int_0^\infty\frac{\log(e^{-t})}{1-e^{-t}}\,e^{-t}dt\\&=\int_0^\infty\frac{t}{e^t-1}\,dt\\&=\zeta(2)\Gamma(2)\\&=\zeta(2)\end{aligned}$$

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, in that usual 3 dimensions that we live in. Many interesting things can be said here, in analogy with what we know from Part I about triangles.

13b. Regular polyhedra

Let us first discuss the graphs. As a fundamental result about them, we have:

THEOREM 13.1. *For a connected planar graph we have the Euler formula*

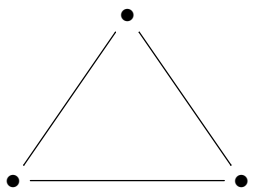
$$v - e + f = 2$$

with v, e, f being the number of vertices, edges and faces.

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

(1) Regarding the precise statement, given a connected planar graph, drawn in a planar way, without crossings, we can certainly talk about the numbers v and e, as for any graph, and also about f, as being the number of faces that our graph has, in our picture, with these including by definition the outer face too, the one going to ∞. With these conventions, the claim is that the Euler formula $v - e + f = 2$ holds indeed.

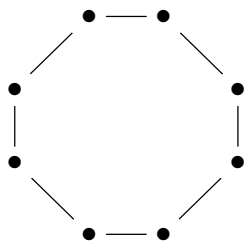
(2) As a first illustration for how this formula works, consider a triangle:



Here we have $v = e = 3$, and $f = 2$, with this accounting for the interior and exterior, and we conclude that the Euler formula holds indeed in this case, as follows:

$$3 - 3 + 2 = 2$$

(3) More generally now, let us look at an arbitrary N-gon graph:



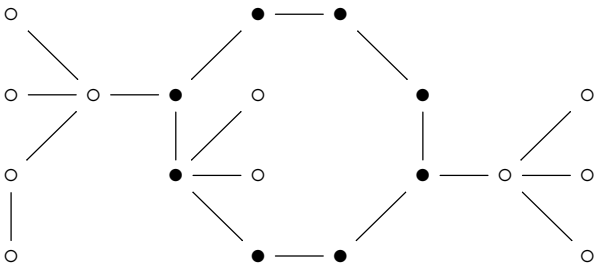
Then, for this graph, the Euler formula holds indeed, as follows:

$$N - N + 2 = 2$$

(4) With these examples discussed, let us look now for a proof. The idea will be to proceed by recurrence on the number of faces f. And here, as a first observation, the result holds at $f = 1$, where our graph must be planar and without cycles, and so must be a tree. Indeed, with N being the number of vertices, the Euler formula holds, as:

$$N - (N - 1) + 1 = 2$$

(5) At $f = 2$ now, our graph must be an N-gon as above, but with some trees allowed to grow from the vertices, with an illustrating example here being as follows:

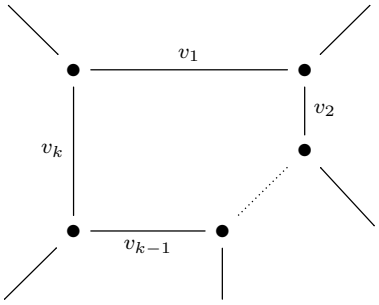


But here we can argue, again based on the fact that for a rooted tree, the non-root vertices are in obvious bijection with the edges, that removing all these trees won't change the problem. So, we are left with the problem for the N-gon, already solved in (3).

(6) And so on, the idea being that we can first remove all the trees, by using the argument in (5), and then we are left with some sort of agglomeration of N-gons, for which we can check the Euler formula directly, a bit as in (3), or by recurrence.

(7) To be more precise, let us try to do the recurrence on the number of faces f. For this purpose, consider one of the faces of our graph, which looks as follows, with v_i

denoting the number of vertices on each side, with the endpoints excluded:



(8) Now let us collapse this chosen face to a single point, in the obvious way. In this process, the total number of vertices v, edges e, and faces f, evolves as follows:

$$v \to v - k + 1 - \sum v_i$$

$$e \to e - \sum (v_i + 1)$$

$$f \to f - 1$$

Thus, in this process, the Euler quantity $v - e + f$ evolves as follows:

$$\begin{aligned} v - e + f &\to v - k + 1 - \sum v_i - e + \sum (v_i + 1) + f - 1 \\ &= v - k + 1 - \sum v_i - e + \sum v_i + k + f - 1 \\ &= v - e + f \end{aligned}$$

So, done with the recurrence, and the Euler formula is proved. □

As a famous application, or rather version, of the Euler formula, let us record:

PROPOSITION 13.2. *For a convex polyhedron we have the Euler formula*

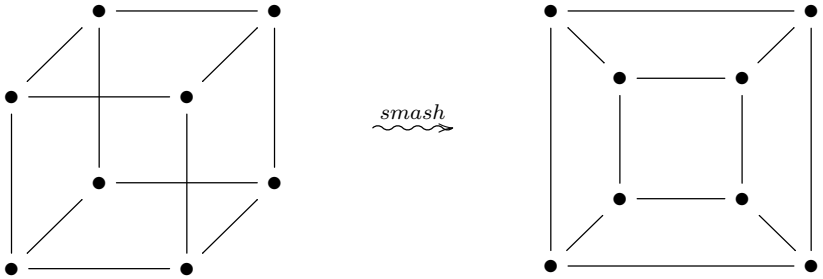
$$v - e + f = 2$$

with v, e, f being the number of vertices, edges and faces.

PROOF. This is more or less the same thing as Theorem 13.1, save for getting rid of the internal trees of the planar graph there, the idea being as follows:

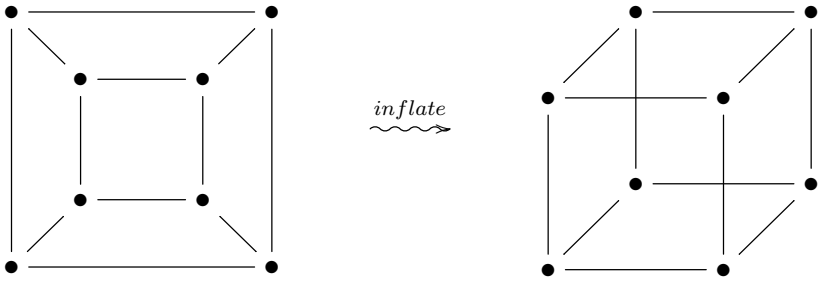
(1) In one sense, consider a convex polyhedron P. We can then enlarge one face, as much as needed, and then smash our polyhedron with a big hammer, as to get a planar

graph X. As an illustration, here is how this method works, for a cube:



But, in this process, each of the numbers v, e, f stays the same, so we get the Euler formula for P, as a consequence of the Euler formula for X, from Theorem 13.1.

(2) Conversely, consider a connected planar graph X. Then, save for getting rid of the internal trees, as explained in the proof of Theorem 13.1, we can assume that we are dealing with an agglomeration of N-gons, again as explained in the proof of Theorem 13.1. But now, we can inflate our graph as to obtain a convex polyhedron P, as follows:



Again, in this process, each of the numbers v, e, f will stay the same, and so we get the Euler formula for X, as a consequence of the Euler formula for P. □

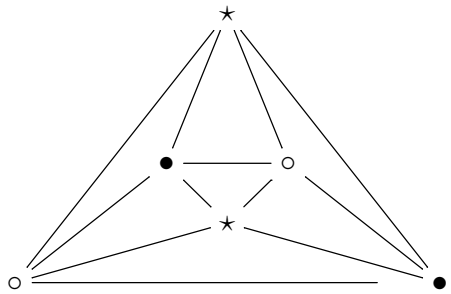
Summarizing, Euler formula understood, but as a matter of making sure that we didn't mess up anything with our mathematics, let us do some direct checks as well:

PROPOSITION 13.3. *The Euler formula $v - e + f = 2$ holds indeed for the five possible regular polyhedra, as follows:*

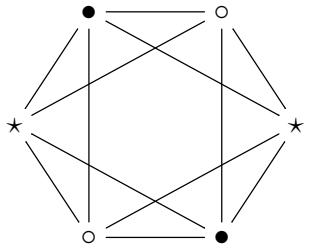
(1) *Tetrahedron:* $4 - 6 + 4 = 2$.
(2) *Cube:* $8 - 12 + 6 = 2$.
(3) *Octahedron:* $6 - 12 + 8 = 2$.
(4) *Dodecahedron:* $20 - 30 + 12 = 2$.
(5) *Isocahedron:* $12 - 30 + 20 = 2$.

PROOF. The figures in the statement are certainly the good ones for the tetrahedron and the cube. Regarding now the octahedron, again the figures are the good ones, by thinking in 3D, but as an interesting exercise for us, which is illustrating for the above, let us attempt to find a nice way of drawing the corresponding graph:

(1) To start with, the "smashing" method from the proof of Proposition 13.2 provides us with a graph which is certainly planar, but which, even worse than before for the cube, sort of misses the whole point with the 3D octahedron, its symmetries, and so on:



(2) Much nicer, instead, is the following picture, which still basically misses the 3D beauty of the octahedron, but at least reveals some of its symmetries:



In short, you get the point, quite subjective all this, and as a conclusion, drawing graphs in an appropriate way remains an art. As for the dodecahedron and isocahedron, exercise here for you, and if failing, take some drawing classes. Math is not everything. □

The Euler formula $v - e + f = 2$, in both its above formulations, the graph one from Theorem 13.1, and the polyhedron one from Proposition 13.2, is something very interesting, at the origin of modern pure mathematics, and having countless other versions and generalizations. We will be back to it on several occasions, in what follows.

13c. Higher dimensions

Higher dimensions. Many things can be said here.

13d. Matrices, rotations

We can talk about matrices and rotations in arbitrary dimensions, in the obvious way. In two complex dimensions now, it is convenient to restrict the attention to the unitaries of determinant 1, which are subject to the following well-known result:

THEOREM 13.4. *We have the following formula,*

$$SU_2 = \left\{ \begin{pmatrix} a & b \\ -\bar{b} & \bar{a} \end{pmatrix} \;\middle|\; |a|^2 + |b|^2 = 1 \right\}$$

which makes SU_2 isomorphic to the unit sphere $S^1_{\mathbb{C}} \subset \mathbb{C}^2$.

PROOF. Consider indeed an arbitrary 2×2 matrix, written as follows:

$$U = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Assuming that we have $\det U = 1$, the inverse must be given by:

$$U^{-1} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

On the other hand, assuming $U \in U_2$, the inverse must be the adjoint:

$$U^{-1} = \begin{pmatrix} \bar{a} & \bar{c} \\ \bar{b} & \bar{d} \end{pmatrix}$$

We are therefore led to the following equations, for the matrix entries:

$$d = \bar{a} \quad , \quad c = -\bar{b}$$

Thus our matrix must be of the following special form in the statement. Moreover, since the determinant of this matrix is 1, we must have, as stated:

$$|a|^2 + |b|^2 = 1$$

Thus, we are done with one inclusion. As for the converse, this is clear, the matrices in the statement being unitaries, and of determinant 1, and so being elements of SU_2. Finally, regarding the last assertion, this is something clear too. □

The matrices in Theorem 13.4 are known to be diagonalizable, and we will leave their explicit diagonalization as an instructive exercise. Moving forward now, in 3 dimensions things are more complicated, and in order to discuss this, we will need:

THEOREM 13.5. *We have the following formula,*

$$SU_2 = \left\{ p\beta_1 + q\beta_2 + r\beta_3 + s\beta_4 \;\middle|\; p^2 + q^2 + r^2 + s^2 = 1 \right\}$$

where $\beta_1, \beta_2, \beta_3, \beta_4$ are the following matrices,

$$\beta_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad , \quad \beta_2 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \quad , \quad \beta_3 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad , \quad \beta_4 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$$

called Pauli spin matrices.

PROOF. We know from Theorem 13.4 that the group SU_2 can be parametrized by the points of the real sphere $S^3_{\mathbb{R}} \subset \mathbb{R}^4$, in the following way:

$$SU_2 = \left\{ \begin{pmatrix} p+iq & r+is \\ -r+is & p-iq \end{pmatrix} \;\middle|\; p^2+q^2+r^2+s^2 = 1 \right\}$$

But this gives the formula in the statement, with the Pauli matrices $\beta_1, \beta_2, \beta_3, \beta_4$ being the coefficients of p, q, r, s, in this parametrization. □

The above result is often the most convenient one, when dealing with SU_2. This is because the Pauli matrices have a number of remarkable properties, as follows:

PROPOSITION 13.6. *The Pauli matrices multiply according to the following formulae,*

$$\beta_2^2 = \beta_3^2 = \beta_4^2 = -1$$

$$\beta_2\beta_3 = -\beta_3\beta_2 = \beta_4$$

$$\beta_3\beta_4 = -\beta_4\beta_3 = \beta_2$$

$$\beta_4\beta_2 = -\beta_2\beta_4 = \beta_3$$

they conjugate according to the following rules,

$$\beta_1^* = \beta_1,\ \beta_2^* = -\beta_2,\ \beta_3^* = -\beta_3,\ \beta_4^* = -\beta_4$$

and they form an orthonormal basis of $M_2(\mathbb{C})$, with respect to the scalar product

$$< x, y >= tr(xy^*)$$

with $tr : M_2(\mathbb{C}) \to \mathbb{C}$ being the normalized trace of 2×2 matrices, $tr = Tr/2$.

PROOF. The first two assertions, regarding the multiplication and conjugation rules for the Pauli matrices, follow from some elementary computations. As for the last assertion, this follows by using these rules. Indeed, the fact that the Pauli matrices are pairwise orthogonal follows from computations of the following type, for $i \neq j$:

$$< \beta_i, \beta_j >= tr(\beta_i\beta_j^*) = tr(\pm\beta_i\beta_j) = tr(\pm\beta_k) = 0$$

As for the fact that the Pauli matrices have norm 1, this follows from:

$$< \beta_i, \beta_i >= tr(\beta_i\beta_i^*) = tr(\pm\beta_i^2) = tr(\beta_1) = 1$$

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

Getting now towards SO_3, we first have the following result:

PROPOSITION 13.7. *The adjoint action $SU_2 \curvearrowright M_2(\mathbb{C})$, given by $T_U(A) = UAU^*$, leaves invariant the following real vector subspace of $M_2(\mathbb{C})$,*

$$\mathbb{R}^4 = span(\beta_1, \beta_2, \beta_3, \beta_4)$$

and we obtain in this way a group morphism $SU_2 \to GL_4(\mathbb{R})$.

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

(1) We must first prove that, with $E \subset M_2(\mathbb{C})$ being the real vector space in the statement, we have the following implication:

$$U \in SU_2, A \in E \implies UAU^* \in E$$

But this is clear from the multiplication rules for the Pauli matrices, from Proposition 13.6. Indeed, let us write our matrices U, A as follows:

$$U = x\beta_1 + y\beta_2 + z\beta_3 + t\beta_4$$

$$A = a\beta_1 + b\beta_2 + c\beta_3 + d\beta_4$$

We know that the coefficients x, y, z, t and a, b, c, d are all real, due to $U \in SU_2$ and $A \in E$. The point now is that when computing UAU^*, by using the various rules from Proposition 13.6, we obtain a matrix of the same type, namely a combination of $\beta_1, \beta_2, \beta_3, \beta_4$, with real coefficients. Thus, we have $UAU^* \in E$, as desired.

(2) In order to conclude, let us identify $E \simeq \mathbb{R}^4$, by using the basis $\beta_1, \beta_2, \beta_3, \beta_4$. The result found in (1) shows that we have a correspondence as follows:

$$SU_2 \to M_4(\mathbb{R}) \quad , \quad U \to (T_U)_{|E}$$

Now observe that for any $U \in SU_2$ and any $A \in M_2(\mathbb{C})$ we have:

$$T_{U^*}T_U(A) = U^*UAU^*U = A$$

Thus $T_{U^*} = T_U^{-1}$, and so the correspondence that we found can be written as:

$$SU_2 \to GL_4(\mathbb{R}) \quad , \quad U \to (T_U)_{|E}$$

But this a group morphism, due to the following computation:

$$T_UT_V(A) = UVAV^*U^* = T_{UV}(A)$$

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

The above result is quite interesting, and as a continuation of it, we have:

PROPOSITION 13.8. *With respect to the standard basis $\beta_1, \beta_2, \beta_3, \beta_4$ of the vector space $\mathbb{R}^4 = span(\beta_1, \beta_2, \beta_3, \beta_4)$, the morphism $T : SU_2 \to GL_4(\mathbb{R})$ is given by:*

$$T_U = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & p^2+q^2-r^2-s^2 & 2(qr-ps) & 2(pr+qs) \\ 0 & 2(ps+qr) & p^2+r^2-q^2-s^2 & 2(rs-pq) \\ 0 & 2(qs-pr) & 2(pq+rs) & p^2+s^2-q^2-r^2 \end{pmatrix}$$

Thus, when looking at T as a group morphism $SU_2 \to O_4$, what we have in fact is a group morphism $SU_2 \to O_3$, and even $SU_2 \to SO_3$.

PROOF. With notations from Proposition 13.7 and its proof, let us first look at the action $L : SU_2 \curvearrowright \mathbb{R}^4$ by left multiplication, $L_U(A) = UA$. We have:

$$L_U = \begin{pmatrix} p & -q & -r & -s \\ q & p & -s & r \\ r & s & p & -q \\ s & -r & q & p \end{pmatrix}$$

Similarly, in what regards now the action $R : SU_2 \curvearrowright \mathbb{R}^4$ by right multiplication, $R_U(A) = AU^*$, the corresponding matrix is given by:

$$R_U = \begin{pmatrix} p & q & r & s \\ -q & p & -s & r \\ -r & s & p & -q \\ -s & -r & q & p \end{pmatrix}$$

Now by composing, the matrix of the adjoint matrix in the statement is:

$$\begin{aligned} T_U &= R_U L_U \\ &= \begin{pmatrix} p & q & r & s \\ -q & p & -s & r \\ -r & s & p & -q \\ -s & -r & q & p \end{pmatrix} \begin{pmatrix} p & -q & -r & -s \\ q & p & -s & r \\ r & s & p & -q \\ s & -r & q & p \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & p^2+q^2-r^2-s^2 & 2(qr-ps) & 2(pr+qs) \\ 0 & 2(ps+qr) & p^2+r^2-q^2-s^2 & 2(rs-pq) \\ 0 & 2(qs-pr) & 2(pq+rs) & p^2+s^2-q^2-r^2 \end{pmatrix} \end{aligned}$$

Thus, we have the formula in the statement, and this gives the result. □

We can now formulate a famous result, due to Euler-Rodrigues, as follows:

THEOREM 13.9. *We have the Euler-Rodrigues formula*

$$U = \begin{pmatrix} p^2+q^2-r^2-s^2 & 2(qr-ps) & 2(pr+qs) \\ 2(ps+qr) & p^2+r^2-q^2-s^2 & 2(rs-pq) \\ 2(qs-pr) & 2(pq+rs) & p^2+s^2-q^2-r^2 \end{pmatrix}$$

with $p^2+q^2+r^2+s^2=1$, for the generic elements of SO_3.

PROOF. We know from the above that we have a group morphism $SU_2 \to SO_3$, given by the formula in the statement, and the problem now is that of proving that this is a double cover map, in the sense that it is surjective, and with kernel $\{\pm 1\}$.

(1) Regarding the kernel, this is elementary to compute, as follows:

$$\begin{aligned}\ker(SU_2\to SO_3) &= \left\{U\in SU_2\middle|T_U(A)=A,\forall A\in E\right\}\\ &= \left\{U\in SU_2\middle|UA=AU,\forall A\in E\right\}\\ &= \left\{U\in SU_2\middle|U\beta_i=\beta_iU,\forall i\right\}\\ &= \{\pm 1\}\end{aligned}$$

(2) Thus, we are left with proving that our group morphism $SU_2\to SO_3$ is surjective. However, this latter fact is something quite non-trivial, and we will defer the discussion here for later in this book, when systematically discussing the Lie groups. □

And with the above result, good news, if looking for some difficult exercises, in relation with the spectral theorem for unitaries, try diagonalizing the above matrices.

13e. Exercises

Exercises:

Exercise 13.10.

Exercise 13.11.

Exercise 13.12.

Exercise 13.13.

Exercise 13.14.

Exercise 13.15.

Exercise 13.16.

Exercise 13.17.

Bonus exercise.

CHAPTER 14

Spherical coordinates

14a. Advanced calculus

Welcome to multivariable calculus. As a first job, given a function $\varphi : \mathbb{R}^N \to \mathbb{R}$, we would like to find a quantity $\varphi'(x)$ making the following formula work:

$$\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h$$

But here, as in 1 variable, there are not so many choices, and the solution is that of defining $\varphi'(x)$ as being the row vector formed by the partial derivatives at x:

$$\varphi'(x) = \begin{pmatrix} \frac{d\varphi}{dx_1} & \ldots & \frac{d\varphi}{dx_N} \end{pmatrix}$$

To be more precise, with this value for $\varphi'(x)$, our approximation formula $\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h$ makes sense indeed, as an equality of real numbers, with $\varphi'(x)h \in \mathbb{R}$ being obtained as the matrix multiplication of the row vector $\varphi'(x)$, and the column vector h. As for the fact that our formula holds indeed, this follows by putting together the approximation properties of each of the partial derivatives $d\varphi/dx_i$, which give:

$$\varphi(x+h) \simeq \varphi(x) + \sum_{i=1}^{N} \frac{d\varphi}{dx_i} \cdot h_i = \varphi(x) + \varphi'(x)h$$

Before moving forward, you might say, why bothering with horizontal vectors, when it is so simple and convenient to have all vectors vertical, by definition. Good point, and in answer, we can indeed talk about the gradient of φ, constructed as follows:

$$\nabla\varphi = \begin{pmatrix} \frac{d\varphi}{dx_1} \\ \vdots \\ \frac{d\varphi}{dx_N} \end{pmatrix}$$

With this convention, $\nabla\varphi$ geometrically describes the slope of φ at the point x, in the obvious way. However, the approximation formula must be rewritten as follows:

$$\varphi(x+h) \simeq \varphi(x) + < \nabla\varphi(x), h >$$

In what follows we will use both φ' and $\nabla\varphi$, depending on the context. Moving now to second derivatives, the main result here is as follows:

THEOREM 14.1. *The second derivative of a function $\varphi : \mathbb{R}^N \to \mathbb{R}$, making the formula*

$$\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h + \frac{<\varphi''(x)h, h>}{2}$$

work, is its Hessian matrix $\varphi''(x) \in M_N(\mathbb{R})$, given by the following formula:

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

Moreover, this Hessian matrix is symmetric, $\varphi''(x)_{ij} = \varphi'(x)_{ji}$.

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

(1) As a first observation, at $N = 1$ the Hessian matrix constructed above is simply the 1×1 matrix having as entry the second derivative $\varphi''(x)$, and the formula in the statement is something that we know well from chapter 9, namely:

$$\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h + \frac{\varphi''(x)h^2}{2}$$

(2) At $N = 2$ now, we obviously need to differentiate φ twice, and the point is that we come in this way upon the following formula, called Clairaut formula:

$$\frac{d^2\varphi}{dxdy} = \frac{d^2\varphi}{dydx}$$

But, is this formula correct or not? As an intuitive justification for it, let us consider a product of power functions, $\varphi(z) = x^p y^q$. We have then our formula, due to:

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

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

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

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

Thus, we can see that our commutation formula for derivatives holds indeed, due to the fact that the functions in x, y commute. Of course, all this does not fully prove our formula, in general. But exercise for you, to have this idea fully working, or to look up the standard proof of the Clairaut formula, using the mean value theorem.

(3) Moving now to $N = 3$ and higher, we can use here the Clairaut formula with respect to any pair of coordinates, which gives the Schwarz formula, namely:

$$\frac{d^2\varphi}{dx_i dx_j} = \frac{d^2\varphi}{dx_j dx_i}$$

Thus, the second derivative, or Hessian matrix, is symmetric, as claimed.

(4) Getting now to the main topic, namely approximation formula in the statement, in arbitrary N dimensions, this is in fact something which does not need a new proof, because it follows from the one-variable formula in (1), applied to the restriction of φ to the following segment in $\mathbb{R}^N$, which can be regarded as being a one-variable interval:

$$I = [x, x + h]$$

To be more precise, let $y \in \mathbb{R}^N$, and consider the following function, with $r \in \mathbb{R}$:

$$f(r) = \varphi(x + ry)$$

We know from (1) that the Taylor formula for f, at the point $r = 0$, reads:

$$f(r) \simeq f(0) + f'(0)r + \frac{f''(0)r^2}{2}$$

And our claim is that, with $h = ry$, this is precisely the formula in the statement.

(5) So, let us see if our claim is correct. By using the chain rule, we have the following formula, with on the right, as usual, a row vector multiplied by a column vector:

$$f'(r) = \varphi'(x + ry) \cdot y$$

By using again the chain rule, we can compute the second derivative as well:

$$\begin{aligned}
f''(r) &= (\varphi'(x + ry) \cdot y)' \\
&= \left(\sum_i \frac{d\varphi}{dx_i}(x + ry) \cdot y_i \right)' \\
&= \sum_i \sum_j \frac{d^2\varphi}{dx_i dx_j}(x + ry) \cdot \frac{d(x + ry)_j}{dr} \cdot y_i \\
&= \sum_i \sum_j \frac{d^2\varphi}{dx_i dx_j}(x + ry) \cdot y_i y_j \\
&= < \varphi''(x + ry)y, y >
\end{aligned}$$

(6) Time now to conclude. We know that we have $f(r) = \varphi(x + ry)$, and according to our various computations above, we have the following formulae:

$$f(0) = \varphi(x) \quad , \quad f'(0) = \varphi'(x) \quad , \quad f''(0) =< \varphi''(x)y, y >$$

Buit with this data in hand, the usual Taylor formula for our one variable function f, at order 2, at the point $r = 0$, takes the following form, with $h = ry$:

$$\begin{aligned}\varphi(x + ry) &\simeq \varphi(x) + \varphi'(x)ry + \frac{< \varphi''(x)y, y > r^2}{2}\\ &= \varphi(x) + \varphi'(x)t + \frac{< \varphi''(x)h, h >}{2}\end{aligned}$$

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

As before in the one variable case, many more things can be said, as a continuation of the above. For instance the local minima and maxima of $\varphi : \mathbb{R}^N \to \mathbb{R}$ appear at the points $x \in \mathbb{R}^N$ where the derivative vanishes, $\varphi'(x) = 0$, and where the second derivative $\varphi''(x) \in M_N(\mathbb{R})$ is positive, respectively negative. But, you surely know all this.

As a key observation now, generalizing what we know in 1 variable, we have:

PROPOSITION 14.2. *Intuitively, the following quantity, called Laplacian of φ,*

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

measures how much different is $\varphi(x)$, compared to the average of $\varphi(y)$, with $y \simeq x$.

PROOF. As before with 1 variable, this is something a bit heuristic, but good to know. Let us write the formula in Theorem 14.1, as such, and with $h \to -h$ too:

$$\varphi(x + h) \simeq \varphi(x) + \varphi'(x)h + \frac{< \varphi''(x)h, h >}{2}$$

$$\varphi(x - h) \simeq \varphi(x) - \varphi'(x)h + \frac{< \varphi''(x)h, h >}{2}$$

By making the average, we obtain the following formula:

$$\frac{\varphi(x + h) + \varphi(x - h)}{2} = \varphi(x) + \frac{< \varphi''(x)h, h >}{2}$$

Thus, thinking a bit, we are led to the conclusion in the statement, modulo some discussion about integrating all this, that we will not really need, in what follows. □

We can talk about multiple integrals, in the obvious way. Getting now to the general theory and rules, for computing such integrals, the key result here is the change of variable formula. In order to discuss this, let us start with something that we know well, in 1D:

PROPOSITION 14.3. *We have the change of variable formula*

$$\int_a^b f(x)dx = \int_c^d f(\varphi(t))\varphi'(t)dt$$

where $c = \varphi^{-1}(a)$ and $d = \varphi^{-1}(b)$.

PROOF. This follows with $f = F'$, via the following differentiation rule:

$$(F\varphi)'(t) = F'(\varphi(t))\varphi'(t)$$

Indeed, by integrating between c and d, we obtain the result. □

In several variables now, we can only expect the above $\varphi'(t)$ factor to be replaced by something similar, a sort of "derivative of φ, arising as a real number". But this can only be the Jacobian $\det(\varphi'(t))$, and with this in mind, we are led to:

THEOREM 14.4. *Given a transformation $\varphi = (\varphi_1, \ldots, \varphi_N)$, we have*

$$\int_E f(x)dx = \int_{\varphi^{-1}(E)} f(\varphi(t))|J_\varphi(t)|dt$$

with the J_φ quantity, called Jacobian, being given by

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

and with this generalizing the formula from Proposition 14.3.

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

(1) Observe first that this generalizes indeed the change of variable formula in 1 dimension, from Proposition 14.3, the point here being that the absolute value on the derivative appears as to compensate for the lack of explicit bounds for the integral.

(2) In general now, we can first argue that, the formula in the statement being linear in f, we can assume $f = 1$. Thus we want to prove $vol(E) = \int_{\varphi^{-1}(E)} |J_\varphi(t)|dt$, and with $D = \varphi^{-1}(E)$, this amounts in proving $vol(\varphi(D)) = \int_D |J_\varphi(t)|dt$.

(3) Now since this latter formula is additive with respect to D, it is enough to prove that $vol(\varphi(D)) = \int_D J_\varphi(t)dt$, for small cubes D, and assuming $J_\varphi > 0$. But this follows by using the usual definition of the determinant, as a volume.

(4) The details and computations however are quite non-trivial, and can be found for instance in Rudin [79]. So, please read that. With this, reading the complete proof of the present theorem from Rudin, being part of the standard math experience. □

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

14b. Spherical coordinates

Time now do some exciting computations, with the technology that we have. In what regards the applications of Theorem 14.4, these often come via:

PROPOSITION 14.5. *We have polar coordinates in 2 dimensions,*

$$\begin{cases} x &= r\cos t \\ y &= r\sin t \end{cases}$$

the corresponding Jacobian being $J = r$.

PROOF. This is elementary, the Jacobian being:

$$\begin{aligned} J &= \begin{vmatrix} \frac{d(r\cos t)}{dr} & \frac{d(r\cos t)}{dt} \\ \frac{d(r\sin t)}{dr} & \frac{d(r\sin t)}{dt} \end{vmatrix} \\ &= \begin{vmatrix} \cos t & -r\sin t \\ \sin t & r\cos t \end{vmatrix} \\ &= r\cos^2 t + r\sin^2 t \\ &= r \end{aligned}$$

Thus, we have indeed the formula in the statement. □

We can now compute the Gauss integral, which is the best calculus formula ever:

THEOREM 14.6. *We have the following formula,*

$$\int_{\mathbb{R}} e^{-x^2} dx = \sqrt{\pi}$$

called Gauss integral formula.

PROOF. Let I be the above integral. By using polar coordinates, we obtain:

$$\begin{aligned} I^2 &= \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-x^2-y^2} dxdy \\ &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2} rdrdt \\ &= 2\pi\int_0^{\infty} \left(-\frac{e^{-r^2}}{2}\right)' dr \\ &= 2\pi\left[0 - \left(-\frac{1}{2}\right)\right] \\ &= \pi \end{aligned}$$

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

Moving now to 3 dimensions, we have here the following result:

PROPOSITION 14.7. *We have spherical coordinates in* 3 *dimensions,*

$$\begin{cases} x &= r\cos s \\ y &= r\sin s\cos t \\ z &= r\sin s\sin t \end{cases}$$

the corresponding Jacobian being $J(r,s,t) = r^2\sin s$.

PROOF. The fact that we have indeed spherical coordinates is clear. Regarding now the Jacobian, this is given by the following formula:

$$\begin{aligned} & J(r,s,t) \\ = & \begin{vmatrix} \cos s & -r\sin s & 0 \\ \sin s\cos t & r\cos s\cos t & -r\sin s\sin t \\ \sin s\sin t & r\cos s\sin t & r\sin s\cos t \end{vmatrix} \\ = & r^2\sin s\sin t\begin{vmatrix} \cos s & -r\sin s \\ \sin s\sin t & r\cos s\sin t \end{vmatrix} + r\sin s\cos t\begin{vmatrix} \cos s & -r\sin s \\ \sin s\cos t & r\cos s\cos t \end{vmatrix} \\ = & r\sin s\sin^2 t\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} + r\sin s\cos^2 t\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} \\ = & r\sin s(\sin^2 t+\cos^2 t)\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} \\ = & r\sin s\times 1\times r \\ = & r^2\sin s \end{aligned}$$

Thus, we have indeed the formula in the statement. □

Let us work out now the general spherical coordinate formula, in arbitrary N dimensions. The formula here, which generalizes those at $N = 2, 3$, is as follows:

THEOREM 14.8. *We have spherical coordinates in* N *dimensions,*

$$\begin{cases} x_1 &= r\cos t_1 \\ x_2 &= r\sin t_1\cos t_2 \\ \vdots & \\ x_{N-1} &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\cos t_{N-1} \\ x_N &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\sin t_{N-1} \end{cases}$$

the corresponding Jacobian being given by the following formula,

$$J(r,t) = r^{N-1}\sin^{N-2}t_1\sin^{N-3}t_2\ \ldots\ \sin^2 t_{N-3}\sin t_{N-2}$$

and with this generalizing the known formulae at $N = 2, 3$.

PROOF. As before, the fact that we have spherical coordinates is clear. Regarding now the Jacobian, also as before, by developing over the last column, we have:

$$\begin{aligned} J_N &= r\sin t_1\ldots\sin t_{N-2}\sin t_{N-1}\times\sin t_{N-1}J_{N-1}\\ &+ r\sin t_1\ldots\sin t_{N-2}\cos t_{N-1}\times\cos t_{N-1}J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}(\sin^2 t_{N-1}+\cos^2 t_{N-1})J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}J_{N-1}\end{aligned}$$

Thus, we obtain the formula in the statement, by recurrence. □

As a comment here, the above convention for spherical coordinates is one among many, designed to best work in arbitrary N dimensions. Also, in what regards the precise range of the angles $t_1,\ldots,t_{N-1}$, we will leave this to you, as an instructive exercise.

As an application, let us compute the volumes of spheres. For this purpose, we must understand how the products of coordinates integrate over spheres. Let us start with the case $N=2$. Here the sphere is the unit circle $\mathbb{T}$, and with $z=e^{it}$ the coordinates are $\cos t,\sin t$. We can first integrate arbitrary powers of these coordinates, as follows:

PROPOSITION 14.9. *We have the following formulae,*

$$\int_0^{\pi/2}\cos^p t\,dt=\int_0^{\pi/2}\sin^p t\,dt=\left(\frac{\pi}{2}\right)^{\varepsilon(p)}\frac{p!!}{(p+1)!!}$$

where $\varepsilon(p)=1$ if p is even, and $\varepsilon(p)=0$ if p is odd, and where

$$m!!=(m-1)(m-3)(m-5)\ldots$$

with the product ending at 2 if m is odd, and ending at 1 if m is even.

PROOF. Let us first compute the integral on the left in the statement:

$$I_p=\int_0^{\pi/2}\cos^p t\,dt$$

We do this by partial integration. We have the following formula:

$$\begin{aligned}(\cos^p t\sin t)' &= p\cos^{p-1}t(-\sin t)\sin t+\cos^p t\cos t\\ &= p\cos^{p+1}t-p\cos^{p-1}t+\cos^{p+1}t\\ &= (p+1)\cos^{p+1}t-p\cos^{p-1}t\end{aligned}$$

By integrating between 0 and $\pi/2$, we obtain the following formula:

$$(p+1)I_{p+1}=pI_{p-1}$$

Thus we can compute I_p by recurrence, and we obtain:

$$\begin{aligned} I_p &= \frac{p-1}{p} I_{p-2} \\ &= \frac{p-1}{p} \cdot \frac{p-3}{p-2} I_{p-4} \\ &= \frac{p-1}{p} \cdot \frac{p-3}{p-2} \cdot \frac{p-5}{p-4} I_{p-6} \\ &\vdots \\ &= \frac{p!!}{(p+1)!!} I_{1-\varepsilon(p)} \end{aligned}$$

But $I_0 = \frac{\pi}{2}$ and $I_1 = 1$, so we get the result. As for the second formula, this follows from the first one, with $t = \frac{\pi}{2} - s$. Thus, we have proved both formulae in the statement. □

We can now compute the volume of the sphere, as follows:

THEOREM 14.10. *The volume of the unit sphere in $\mathbb{R}^N$ is given by*

$$V = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!}$$

with our usual convention $N!! = (N-1)(N-3)(N-5)\ldots$

PROOF. Let us denote by B^+ the positive part of the unit sphere, or rather unit ball B, obtained by cutting this unit ball in 2^N parts. At the level of volumes, we have:

$$V = 2^N V^+$$

We have the following computation, using spherical coordinates:

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

Here we have used the following formula, for computing the exponent of $\pi/2$:

$$\begin{aligned}\varepsilon(0)+\varepsilon(1)+\varepsilon(2)+\ldots+\varepsilon(N-2) &= 1+0+1+\ldots+\varepsilon(N-2)\\ &= \left[\frac{N-2}{2}\right]+1\\ &= \left[\frac{N}{2}\right]\end{aligned}$$

Thus, we obtain the formula in the statement. □

As main particular cases of the above formula, we have:

THEOREM 14.11. *The volumes of the low-dimensional spheres are as follows:*

(1) *At $N=1$, the length of the unit interval is $V=2$.*
(2) *At $N=2$, the area of the unit disk is $V=\pi$.*
(3) *At $N=3$, the volume of the unit sphere is $V=\frac{4\pi}{3}$*
(4) *At $N=4$, the volume of the corresponding unit sphere is $V=\frac{\pi^2}{2}$.*

PROOF. Some of these results are well-known, but we can obtain all of them as particular cases of the general formula in Theorem 14.10, as follows:

(1) At $N=1$ we obtain $V=1\cdot\frac{2}{1}=2$.

(2) At $N=2$ we obtain $V=\frac{\pi}{2}\cdot\frac{4}{2}=\pi$.

(3) At $N=3$ we obtain $V=\frac{\pi}{2}\cdot\frac{8}{3}=\frac{4\pi}{3}$.

(4) At $N=4$ we obtain $V=\frac{\pi^2}{4}\cdot\frac{16}{8}=\frac{\pi^2}{2}$. □

14c. Kepler and Newton

You surely know a bit about gravity, including the findings of Kepler and Newton. The result here, and its proof, which is the pride of mathematics, physics, and human knowledge in general, is the following famous theorem:

THEOREM 14.12 (Kepler, Newton). *Planets and other celestial bodies move around the Sun on conics, that is, on curves of type*

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

with $P\in\mathbb{R}[x,y]$ being of degree 2. The same is true for any body moving around another body, provided that we are not in the situation of a free fall.

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

(1) According to observations and calculations performed over the centuries, since the ancient times, and first formalized by Newton, following some groundbreaking work of Kepler, the force of attraction between two bodies of masses M, m is given by:

$$||F|| = G \cdot \frac{Mm}{d^2}$$

Here d is the distance between the two bodies, and $G \simeq 6.674 \times 10^{-11}$ is a constant. Now assuming that M is fixed at $0 \in \mathbb{R}^3$, the force exterted on m positioned at $x \in \mathbb{R}^3$, regarded as a vector $F \in \mathbb{R}^3$, is given by the following formula:

$$F = -||F|| \cdot \frac{x}{||x||} = -\frac{GMm}{||x||^2} \cdot \frac{x}{||x||} = -\frac{GMmx}{||x||^3}$$

But $F = ma = m\ddot{x}$, with $a = \ddot{x}$ being the acceleration, second derivative of the position, so the equation of motion of m, assuming that M is fixed at 0, is:

$$\ddot{x} = -\frac{GMx}{||x||^3}$$

(2) Obviously, the problem happens in 2 dimensions, and you can even find, as an exercise, a formal proof of that, based on the above equation. Now here the most convenient is to use standard x, y coordinates, and denote our point as $z = (x, y)$. With this change made, and by setting $K = GM$, the equation of motion becomes:

$$\ddot{z} = -\frac{Kz}{||z||^3}$$

In other words, in terms of the coordinates x, y, the equations are:

$$\ddot{x} = -\frac{Kx}{(x^2 + y^2)^{3/2}} \quad , \quad \ddot{y} = -\frac{Ky}{(x^2 + y^2)^{3/2}}$$

(3) Let us begin with a simple particular case, that of the circular solutions. To be more precise, we are interested in solutions of the following type:

$$x = r\cos\alpha t \quad , \quad y = r\sin\alpha t$$

In this case we have $||z|| = r$, so our equation of motion becomes:

$$\ddot{z} = -\frac{Kz}{r^3}$$

On the other hand, differentiating x, y leads to the following formula:

$$\ddot{z} = (\ddot{x}, \ddot{y}) = -\alpha^2(x, y) = -\alpha^2 z$$

Thus, we have a circular solution when the parameters r, α satisfy:

$$r^3\alpha^2 = K$$

(4) In the general case now, the problem can be solved via some calculus. Let us write indeed our vector $z = (x, y)$ in polar coordinates, as follows:

$$x = r\cos\theta \quad , \quad y = r\sin\theta$$

We have then $||z|| = r$, and our equation of motion becomes, as in (3):

$$\ddot{z} = -\frac{Kz}{r^3}$$

Let us differentiate now x, y. By using the standard calculus rules, we have:

$$\dot{x} = \dot{r}\cos\theta - r\sin\theta\cdot\dot{\theta}$$

$$\dot{y} = \dot{r}\sin\theta + r\cos\theta\cdot\dot{\theta}$$

Differentiating one more time gives the following formulae:

$$\ddot{x} = \ddot{r}\cos\theta - 2\dot{r}\sin\theta\cdot\dot{\theta} - r\cos\theta\cdot\dot{\theta}^2 - r\sin\theta\cdot\ddot{\theta}$$

$$\ddot{y} = \ddot{r}\sin\theta + 2\dot{r}\cos\theta\cdot\dot{\theta} - r\sin\theta\cdot\dot{\theta}^2 + r\cos\theta\cdot\ddot{\theta}$$

Consider now the following two quantities, appearing as coefficients in the above:

$$a = \ddot{r} - r\dot{\theta}^2 \quad , \quad b = 2\dot{r}\dot{\theta} + r\ddot{\theta}$$

In terms of these quantities, our second derivative formulae read:

$$\ddot{x} = a\cos\theta - b\sin\theta$$

$$\ddot{y} = a\sin\theta + b\cos\theta$$

(5) We can now solve the equation of motion from (4). Indeed, with the formulae that we found for $\ddot{x}, \ddot{y}$, our equation of motion takes the following form:

$$a\cos\theta - b\sin\theta = -\frac{K}{r^2}\cos\theta$$

$$a\sin\theta + b\cos\theta = -\frac{K}{r^2}\sin\theta$$

But these two formulae can be written in the following way:

$$\left(a + \frac{K}{r^2}\right)\cos\theta = b\sin\theta$$

$$\left(a + \frac{K}{r^2}\right)\sin\theta = -b\cos\theta$$

By making now the product, and assuming that we are in a non-degenerate case, where the angle θ varies indeed, we obtain by positivity that we must have:

$$a + \frac{K}{r^2} = b = 0$$

(6) We are almost there. Let us first examine the second equation, $b = 0$. Remembering who b is, from (4), this equation can be solved as follows:

$$\begin{aligned} b = 0 &\iff 2\dot{r}\dot{\theta} + r\ddot{\theta} = 0 \\ &\iff \frac{\ddot{\theta}}{\dot{\theta}} = -2\frac{\dot{r}}{r} \\ &\iff (\log\dot{\theta})' = (-2\log r)' \\ &\iff \log\dot{\theta} = -2\log r + c \\ &\iff \dot{\theta} = \frac{\lambda}{r^2} \end{aligned}$$

As for the first equation the we found, namely $a + K/r^2 = 0$, remembering from (4) that a was by definition given by $a = \ddot{r} - r\dot{\theta}^2$, this equation now becomes:

$$\ddot{r} - \frac{\lambda^2}{r^3} + \frac{K}{r^2} = 0$$

(7) As a conclusion to all this, in polar coordinates, $x = r\cos\theta$, $y = r\sin\theta$, our equations of motion are as follows, with λ being a constant, not depending on t:

$$\ddot{r} = \frac{\lambda^2}{r^3} - \frac{K}{r^2} \quad , \quad \dot{\theta} = \frac{\lambda}{r^2}$$

Even better now, by writing $K = \lambda^2/c$, these equations read:

$$\ddot{r} = \frac{\lambda^2}{r^2}\left(\frac{1}{r} - \frac{1}{c}\right) \quad , \quad \dot{\theta} = \frac{\lambda}{r^2}$$

(8) As an illustration, let us quickly work out the case of a circular motion, where r is constant. Here $\ddot{r} = 0$, so the first equation gives $c = r$. Also we have $\dot{\theta} = \alpha$, with:

$$\alpha = \frac{\lambda}{r^2}$$

Assuming $\theta = 0$ at $t = 0$, from $\dot{\theta} = \alpha$ we obtain $\theta = \alpha t$, and so, as in (3) above:

$$x = r\cos\alpha t \quad , \quad y = r\sin\alpha t$$

Observe also that the condition found in (3) is indeed satisfied:

$$r^3\alpha^2 = \frac{\lambda^2}{r} = \frac{\lambda^2}{c} = K$$

(9) Back to the general case now, our claim is that we have the following formula, for the distance $r = r(t)$ as function of the angle $\theta = \theta(t)$, for some $\varepsilon, \delta \in \mathbb{R}$:

$$r = \frac{c}{1 + \varepsilon\cos\theta + \delta\sin\theta}$$

Let us first check that this formula works indeed. With r being as above, and by using our second equation found before, $\dot{\theta} = \lambda/r^2$, we have the following computation:

$$\begin{aligned}\dot{r} &= \frac{c(\varepsilon\sin\theta - \delta\cos\theta)\dot{\theta}}{(1+\varepsilon\cos\theta+\delta\sin\theta)^2}\\ &= \frac{\lambda c(\varepsilon\sin\theta - \delta\cos\theta)}{r^2(1+\varepsilon\cos\theta+\delta\sin\theta)^2}\\ &= \frac{\lambda(\varepsilon\sin\theta - \delta\cos\theta)}{c}\end{aligned}$$

Thus, the second derivative of the above function r is given, as desired, by:

$$\begin{aligned}\ddot{r} &= \frac{\lambda(\varepsilon\cos\theta+\delta\sin\theta)\dot{\theta}}{c}\\ &= \frac{\lambda^2(\varepsilon\cos\theta+\delta\sin\theta)}{r^2c}\\ &= \frac{\lambda^2}{r^2}\left(\frac{1}{r}-\frac{1}{c}\right)\end{aligned}$$

(10) The above check was something quite informal, and now we must prove that our formula is indeed the correct one. For this purpose, we use a trick. Let us write:

$$r(t) = \frac{1}{f(\theta(t))}$$

Abbreviated, and by always reminding that f takes $\theta = \theta(t)$ as variable, this reads:

$$r = \frac{1}{f}$$

With the convention that dots mean as usual derivatives with respect to t, and that the primes will denote derivatives with respect to $\theta = \theta(t)$, we have:

$$\dot{r} = -\frac{f'\dot{\theta}}{f^2} = -\frac{f'}{f^2}\cdot\frac{\lambda}{r^2} = -\lambda f'$$

By differentiating one more time with respect to t, we obtain:

$$\ddot{r} = -\lambda f''\dot{\theta} = -\lambda f''\cdot\frac{\lambda}{r^2} = -\frac{\lambda^2}{r^2}f''$$

On the other hand, our equation for $\ddot{r}$ found in (7) reads:

$$\ddot{r} = \frac{\lambda^2}{r^2}\left(\frac{1}{r}-\frac{1}{c}\right) = \frac{\lambda^2}{r^2}\left(f-\frac{1}{c}\right)$$

Thus, in terms of $f = 1/r$ as above, our equation for $\ddot{r}$ simply reads:

$$f'' + f = \frac{1}{c}$$

But this latter equation is elementary to solve. Indeed, both functions $\cos t, \sin t$ satisfy $g" + g = 0$, so any linear combination of them satisfies as well this equation. But the solutions of $f'' + f = 1/c$ being those of $g'' + g = 0$ shifted by $1/c$, we obtain:

$$f = \frac{1 + \varepsilon \cos\theta + \delta \sin\theta}{c}$$

Now by inverting, we obtain the formula announced in (9), namely:

$$r = \frac{c}{1 + \varepsilon \cos\theta + \delta \sin\theta}$$

(11) But this leads to the conclusion that the trajectory is a conic. Indeed, in terms of the parameter θ, the formulae of the coordinates are:

$$x = \frac{c \cos\theta}{1 + \varepsilon \cos\theta + \delta \sin\theta}$$

$$y = \frac{c \sin\theta}{1 + \varepsilon \cos\theta + \delta \sin\theta}$$

But these are precisely the equations of conics in polar coordinates.

(12) To be more precise, in order to find the precise equation of the conic, observe that the two functions x, y that we found above satisfy the following formula:

$$\begin{aligned} x^2 + y^2 &= \frac{c^2(\cos^2\theta + \sin^2\theta)}{(1 + \varepsilon \cos\theta + \delta \sin\theta)^2} \\ &= \frac{c^2}{(1 + \varepsilon \cos\theta + \delta \sin\theta)^2} \end{aligned}$$

On the other hand, these two functions satisfy as well the following formula:

$$\begin{aligned} (\varepsilon x + \delta y - c)^2 &= \frac{c^2\big(\varepsilon \cos\theta + \delta \sin\theta - (1 + \varepsilon \cos\theta + \delta \sin\theta)\big)^2}{(1 + \varepsilon \cos\theta + \delta \sin\theta)^2} \\ &= \frac{c^2}{(1 + \varepsilon \cos\theta + \delta \sin\theta)^2} \end{aligned}$$

We conclude that our coordinates x, y satisfy the following equation:

$$x^2 + y^2 = (\varepsilon x + \delta y - c)^2$$

But what we have here is an equation of a conic, as claimed. □

Still with me, I hope, after all these computations.

In practice now, in view of some further applications, here is a sort of useful "best of" the formulae found in the proof of Theorem 14.12:

THEOREM 14.13 (Kepler, Newton). *In the context of a 2-body problem, with M fixed at 0, and m starting its movement from Ox, the equation of motion of m, namely*

$$\ddot{z} = -\frac{Kz}{||z||^3}$$

with $K = GM$, and $z = (x, y)$, becomes in polar coordinates, $x = r\cos\theta$, $y = r\sin\theta$,

$$\ddot{r} = \frac{\lambda^2}{r^2}\left(\frac{1}{r} - \frac{1}{c}\right) \quad , \quad \dot{\theta} = \frac{\lambda}{r^2}$$

for some $\lambda, c \in \mathbb{R}$, related by $\lambda^2 = Kc$. The value of r in terms of θ is given by

$$r = \frac{c}{1 + \varepsilon\cos\theta + \delta\sin\theta}$$

for some $\varepsilon, \delta \in \mathbb{R}$. At the level of the affine coordinates x, y, this means

$$x = \frac{c\cos\theta}{1 + \varepsilon\cos\theta + \delta\sin\theta} \quad , \quad y = \frac{c\sin\theta}{1 + \varepsilon\cos\theta + \delta\sin\theta}$$

with $\theta = \theta(t)$ being subject to $\dot{\theta} = \lambda^2/r$, as above. Finally, we have

$$x^2 + y^2 = (\varepsilon x + \delta y - c)^2$$

which is a degree 2 equation, and so the resulting trajectory is a conic.

PROOF. As already mentioned, this is a sort of "best of" the formulae found in the proof of Theorem 14.12. And in the hope of course that we have not forgotten anything. Finally, let us mention that the simplest illustration for this is the circular motion, and for details on this, not included in the above, we refer to the proof of Theorem 14.12. □

As a first question, we would like to understand how the various parameters appearing above, namely $\lambda, c, \varepsilon, \delta$, which via some basic math can only tell us more about the shape of the orbit, appear from the initial data. The formulae here are as follows:

THEOREM 14.14. *In the context of Theorem 14.13, and in polar coordinates, $x = r\cos\theta$, $y = r\sin\theta$, the initial data is as follows, with $R = r_0$:*

$$r_0 = \frac{c}{1 + \varepsilon} \quad , \quad \theta_0 = 0$$

$$\dot{r}_0 = -\frac{\delta\sqrt{K}}{\sqrt{c}} \quad , \quad \dot{\theta}_0 = \frac{\sqrt{Kc}}{R^2}$$

$$\ddot{r}_0 = \frac{\varepsilon K}{R^2} \quad , \quad \ddot{\theta}_0 = \frac{4\delta K}{R^2}$$

The corresponding formulae for the affine coordinates x, y can be deduced from this. Also, the various motion parameters c, ε, δ and $\lambda = \sqrt{Kc}$ can be recovered from this data.

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

(1) As mentioned in Theorem 14.13, the object m begins its movement on Ox. Thus we have $\theta_0 = 0$, and from this we get the formula of r_0 in the statement.

(2) Regarding the initial speed now, the formula of $\dot{\theta}_0$ follows from:

$$\dot{\theta} = \frac{\lambda}{r^2} = \frac{\sqrt{Kc}}{r^2}$$

Also, in what concerns the radial speed, the formula of $\dot{r}_0$ follows from:

$$\begin{aligned}
\dot{r} &= \frac{c(\varepsilon \sin\theta - \delta\cos\theta)\dot{\theta}}{(1 + \varepsilon\cos\theta + \delta\sin\theta)^2} \\
&= \frac{c(\varepsilon \sin\theta - \delta\cos\theta)}{c^2/r^2} \cdot \frac{\sqrt{Kc}}{r^2} \\
&= \frac{\sqrt{K}(\varepsilon \sin\theta - \delta\cos\theta)}{\sqrt{c}}
\end{aligned}$$

(3) Regarding now the initial acceleration, by using $\dot{\theta} = \sqrt{Kc}/r^2$ we find:

$$\ddot{\theta} = -2\sqrt{Kc} \cdot \frac{2r\dot{r}}{r^3} = -\frac{4\sqrt{Kc} \cdot \dot{r}}{r^2}$$

In particular at $t = 0$ we obtain the formula in the statement, namely:

$$\ddot{\theta}_0 = -\frac{4\sqrt{Kc} \cdot \dot{r}_0}{R^2} = \frac{4\sqrt{Kc}}{R^2} \cdot \frac{\delta\sqrt{K}}{\sqrt{c}} = \frac{4\delta K}{R^2}$$

(4) Also regarding acceleration, with $\lambda = \sqrt{Kc}$ our main motion formula reads:

$$\ddot{r} = \frac{Kc}{r^2}\left(\frac{1}{r} - \frac{1}{c}\right)$$

In particular at $t = 0$ we obtain the formula in the statement, namely:

$$\ddot{r}_0 = \frac{Kc}{R^2}\left(\frac{1}{R} - \frac{1}{c}\right) = \frac{Kc}{R^2} \cdot \frac{\varepsilon}{c} = \frac{\varepsilon K}{R^2}$$

(5) Finally, the last assertion is clear, and since the formulae look better anyway in polar coordinates than in affine coordinates, we will not get into details here. □

14d. Gauss law, revised

We already know a bit about the Gauss law, but time now to review this, with full details. Let us start with a basic computation, as follows:

PROPOSITION 14.15. *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:

$$\begin{aligned}
\Phi_E(S) &= \int_S < E(x), n(x) > dx \\
&= \int_S \left\langle \frac{Kqx}{r^3}, \frac{x}{r} \right\rangle dx \\
&= \int_S \frac{Kq}{r^2}\, dx \\
&= \frac{Kq}{r^2} \times 4\pi r^2 \\
&= 4\pi Kq
\end{aligned}$$

Thus with $\varepsilon_0 = 1/(4\pi K)$ as above, we obtain the result. □

More generally now, we have the following result:

THEOREM 14.16. *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 spherical coordinates now, our integral above becomes:

$$\begin{aligned}
\Phi_E(S) &= Kq \int_S \frac{1-rx}{(1-2xr+r^2)^{3/2}} du \\
&= Kq \int_0^{2\pi} \int_0^{\pi} \frac{1-r\cos s}{(1-2r\cos s+r^2)^{3/2}} \cdot \sin s \, ds \, dt \\
&= 2\pi Kq \int_0^{\pi} \frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}} ds \\
&= \frac{q}{2\varepsilon_0} \int_0^{\pi} \frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}} ds
\end{aligned}$$

(5) 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 14.17 (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. 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 the following computation, as desired:

$$\begin{aligned}
\Phi_E(S) &= \Phi_E(S - \cup U_i) + \Phi_E(\cup U_i) \\
&= 0 + \Phi_E(\cup U_i) \\
&= \sum_i \Phi_E(U_i) \\
&= \sum_i \frac{q_i}{\varepsilon_0} \\
&= \frac{Q_{enc}}{\varepsilon_0}
\end{aligned}$$

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 14.15. □

The above is not the end of the story, with the Gauss law. We will be back to it.

14e. Exercises

Exercises:

EXERCISE 14.18.

EXERCISE 14.19.

EXERCISE 14.20.

EXERCISE 14.21.

EXERCISE 14.22.

EXERCISE 14.23.

EXERCISE 14.24.

EXERCISE 14.25.

Bonus exercise.

CHAPTER 15

Vector products

15a. Vector products

Getting started with more applications to physics, both classical mechanics and electrodynamics, and perhaps even beyond, here is the notion what we will need:

DEFINITION 15.1. *The vector product of two vectors in $\mathbb{R}^3$ is given by*

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

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

$$\begin{array}{c} \uparrow_{x\times y} \\ \leftarrow_{x} \\ \swarrow_{y} \end{array}$$

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

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

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

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

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

So, here is how this method works, instantly, or almost:

(1) The first formula, $||x \times y|| \sim ||x||, ||y||$, tells us that we are looking for a certain vector $x \times y$, whose length is proportional to those of x, y.

(2) But the second formula, $x \times x = 0$, tells us that the angle θ between x, y must be involved via $0 \to 0$, and so the proportionality factor can only be $\sin\theta$.

(3) And with this we are almost there, it's just a matter now of choosing the orientation of our vector, and this comes from the third formula, $e_1 \times e_2 = e_3$.

As with the second definition of the vector product, the one with that determinants, that I actually like the most, what I remember here is simply:

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

Indeed, when trying to compute this determinant, by developing over the first column, what you get as coefficients are the entries of $x \times y$. And with the good middle sign.

It is also good to know that $x \times y$ exists only in 3 dimensions, with our only tool in $N \neq 3$ dimensions being the usual $< x, y >$. This is actually quite interesting for us, in relation with our conservation principles for gravity, but more on this later.

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 15.2. *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.

(2) But this suggests precisely that, in order to fix the lack of conservation of the momentum $p = mv$, what we have to do is to make a vector product with the position vector x. Leading to the angular momentum $J = x \times p$, constructed above.

(3) 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. □

While the above principle looks like something quite trivial, the mathematics behind it is quite interesting, and has several notable consequences.

Let us begin with a standard application, in the context of the Kepler and Newton gravitational 2-body problem, that we investigated in chapter 14, as follows:

THEOREM 15.3. *In the context of a 2-body problem, the following happen:*

(1) *The fact that the direction of J is fixed tells us that the trajectory of one body with respect to the other lies in a plane.*
(2) *The fact that the magnitude of J is fixed tells us that the Kepler 2 law holds, namely that we have same areas sweeped by Ox over the same times.*

PROOF. This follows indeed from Theorem 15.2, as follows:

(1) We have by definition $J = m(x \times v)$, and since a vector product is orthogonal on both the vectors it comes from, we deduce from this that we have:

$$J \perp x, v$$

But this can be written as follows, with $J^\perp$ standing for the plane orthogonal to J:

$$x, v \in J^\perp$$

Now since J is fixed by Theorem 15.2, we conclude that both x, v, and in particular the position x, and so the whole trajectory, lie in this fixed plane $J^\perp$, as claimed.

(2) Conversely now, forget about Theorem 15.2, and assume that the trajectory lies in a certain plane E. Thus $x \in E$, and by differentiating we have $v \in E$ too, and so $x, v \in E$. Thus $E = J^\perp$, and so $J = E^\perp$, so the direction of J is fixed, as claimed.

(3) Regarding now the last assertion, we already know from chapter 14, or rather from the exercises there, if you worked hard on them, that Kepler 2 is more or less equivalent to $\dot{\theta} = \lambda/r^2$. However, the derivation of $\dot{\theta} = \lambda/r^2$ was something tricky, and what we want to prove now is that this appears as a simple consequence of $||J|| = \text{constant}$.

(4) In order to to so, let us compute J, according to its definition $J = x \times p$, but in polar coordinates, which will change everything. Since $p = m\dot{x}$, we have:

$$J = r \begin{pmatrix} \cos\theta \\ \sin\theta \\ 0 \end{pmatrix} \times m \begin{pmatrix} \dot{r}\cos\theta - r\sin\theta \cdot \dot{\theta} \\ \dot{r}\sin\theta + r\cos\theta \cdot \dot{\theta} \\ 0 \end{pmatrix}$$

Now recall from the definition of the vector product that we have:

$$\begin{pmatrix} a \\ b \\ 0 \end{pmatrix} \times \begin{pmatrix} c \\ d \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ ad - bc \end{pmatrix}$$

Thus J is a vector of the above form, with its last component being:

$$\begin{aligned} J_z &= rm\begin{vmatrix}\cos\theta & \dot{r}\cos\theta - r\sin\theta\cdot\dot{\theta}\\ \sin\theta & \dot{r}\sin\theta + r\cos\theta\cdot\dot{\theta}\end{vmatrix}\\ &= rm\cdot r(\cos^2\theta+\sin^2\theta)\dot{\theta}\\ &= r^2m\cdot\dot{\theta}\end{aligned}$$

(5) Now with the above formula in hand, our claim is that the magnitude $||J||$ is constant precisely when $\dot{\theta}=\lambda/r^2$, for some $\lambda\in\mathbb{R}$. Indeed, up to the obvious fact that the orientation of J is a binary parameter, who cannot just switch like that, let us just agree on this, knowing J is the same as knowing J_z, and is also the same as knowing $||J||$. Thus, our claim is proved, and this leads to the conclusion in the statement. □

15b. Rotating bodies

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

THEOREM 15.4. *Assume that a 3D body rotates along an axis, with angular speed w. For a fixed point of the body, with position vector x, the usual 3D speed is*

$$v=\omega\times x$$

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

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

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

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

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

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

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

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

$$\begin{aligned} V &= \dot{x}_1u_1+\dot{x}_2u_2+\dot{x}_3u_3+x_1\dot{u}_1+x_2\dot{u}_2+x_3\dot{u}_3\\ &= \dot{x}+(x_1\cdot\omega\times u_1+x_2\cdot\omega\times u_2+x_3\cdot\omega\times u_3)\\ &= \dot{x}+w\times(x_1u_1+x_2u_2+x_3u_3)\\ &= \dot{x}+\omega\times x\end{aligned}$$

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

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

THEOREM 15.5. *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 15.4, 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 15.6. *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 15.6.

15c. Curved spacetime

Still talking basic 3D geometry, let us discuss now curved spacetime, as a continuation of our 1D study from chapter 10. We recall from there that we have:

THEOREM 15.7. *If we sum the 1D 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. This is something that we know well from chapter 10, and we refer to the material there for discussion, examples, and also for some further facts, such as the following equivalent formula, making the link with the hyperbolic trigonometric functions:

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

We refer as well to chapter 10 for more on the physics of this, and with the comment of course that, physically speaking, the Einstein formula is indeed the correct one. □

Getting to work now, in 3D, we would like to solve the following question:

QUESTION 15.8. *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?*

With some inspiration from the 1D case, 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 15.9. *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 15.10. *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 $||v|| \implies u +_e v = v$?

In order to solve this latter puzzle, the first observation is that $\gamma_{uvw} = w$ will not do, and this for several reasons. First, this vector points in the wrong direction, orthogonal to the plane spanned by u, v, and we certainly don't want to leave this plane, with our correction. Also, as a technical remark to be put on top of this, the choice $\gamma_{uvw} = w$ will not bring any simplifications, as required above, in the cases $||u|| = 1$ or $||v|| = 1$.

Moving ahead now, as 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, and so as a reasonable candidate for our correction term, we have:

$$\gamma_{uvw} = (\alpha u + \beta v) \times w$$

Here $\alpha, \beta \in \mathbb{R}$ are some scalars to be determined, but let us take a break, and leave the computations for later. We did some good work, time to update our puzzle:

PUZZLE 15.11. *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 15.12. *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 15.13. *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:

$$\begin{aligned} ||u|| = 1 &\implies \gamma_{uvw} =< u, v > u - v \\ &\implies u + v + \gamma_{uvw} = u + < u, v > u \\ &\implies \frac{u + v + \gamma_{uvw}}{1 + < u, v >} = u \end{aligned}$$

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 15.13 from an abstract, mathematical perspective, there are still many things missing from there, which can be summarized as follows:

QUESTION 15.14. *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 15.13, 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 with this being the end of mathematics, we did a nice job, and we can now formulate our findings as a theorem, as follows:

THEOREM 15.15. *When defining the Einstein speed summation in* 3D *as*

$$u +_e v = \frac{1}{1 + < u, v >} \left(u + v + \frac{u \times (u \times v)}{1 + \sqrt{1 - ||u||^2}} \right)$$

in $c = 1$ *units, the following happen:*

(1) *When* $u \sim v$*, we recover the previous* 1D *formula.*
(2) *We have* $||u||, ||v|| < 1 \implies ||u +_e v|| < 1$.
(3) *When* $||u|| = 1$*, we have* $u +_e v = u$.
(4) *When* $||v|| = 1$*, we have* $||u +_e v|| = 1$.
(5) *However,* $||v|| = 1$ *does not imply* $u +_e v = v$.
(6) *Also, the formula* $u +_e v = v +_e u$ *fails.*

PROOF. This follows from the above discussion, as follows:

(1) This is something that we know from Proposition 15.13, coming from:

$$u \sim v \implies u \times v = 0 \implies u +_e v = \frac{u + v}{1 + < u, v >}$$

(2) This is more tricky. Let us set indeed, as in the statement:

$$u +_e v = \frac{1}{1 + < u, v >} \left(u + v + \frac{u \times (u \times v)}{1 + \sqrt{1 - ||u||^2}} \right)$$

In order to simplify notation, let us set $\delta = \sqrt{1 - ||u||^2}$, which is the inverse of the quantity $\gamma = 1/\sqrt{1 - ||u||^2}$. With this convention, we have:

$$\begin{aligned} u +_e v &= \frac{1}{1+ < u, v >} \left(u + v + \frac{< u, v > u - ||u||^2 v}{1 + \delta} \right) \\ &= \frac{(1 + \delta + < u, v >)u + (1 + \delta - ||u||^2)v}{(1+ < u, v >)(1 + \delta)} \end{aligned}$$

Taking now the squared norm gives the following formula:

$$\begin{aligned} ||u +_e v||^2 &= \frac{1}{(1+ < u, v >)^2(1 + \delta)^2} \times \\ &\quad \Big[(1 + \delta + < u, v >)^2 ||u||^2 \\ &\quad + (1 + \delta - ||u||^2)^2 ||v||^2 \\ &\quad + 2(1 + \delta + < u, v >)(1 + \delta - ||u||^2) < u, v > \Big] \end{aligned}$$

By expanding the two squares and multiplying at the end, and then simplifying everything, we obtain a quite compact formula, as follows:

$$||u +_e v||^2 = \frac{(1 + \delta)^2 ||u + v||^2 + (||u||^2 - 2(1 + \delta))(||u||^2 ||v||^2 - < u, v >^2)}{(1+ < u, v >)^2(1 + \delta)^2}$$

But this formula can be further processed by using $\delta = \sqrt{1 - ||u||^2}$, and by navigating through the various quantities which appear, we obtain, as a final product:

$$||u +_e v||^2 = \frac{||u + v||^2 - ||u||^2 ||v||^2 + < u, v >^2}{(1+ < u, v >)^2}$$

But this type of formula is exactly what we need, for what we want to do. Indeed, by assuming $||u||, ||v|| < 1$, we have the following estimate:

$$\begin{aligned} ||u +_e v||^2 < 1 &\iff ||u + v||^2 - ||u||^2 ||v||^2 + < u, v >^2 < (1+ < u, v >)^2 \\ &\iff ||u + v||^2 - ||u||^2 ||v||^2 < 1 + 2 < u, v > \\ &\iff ||u||^2 + ||v||^2 - ||u||^2 ||v||^2 < 1 \\ &\iff (1 - ||u||^2)(1 - ||v||^2) > 0 \end{aligned}$$

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

(3) This is something that we know from Proposition 15.13, coming from:

$$\begin{aligned}
||u|| = 1 &\implies u \times (u \times v) =< u, v > u - v\\
&\implies \frac{u \times (u \times v)}{1 + \sqrt{1 - ||u||^2}} =< u, v > u - v\\
&\implies u + v + \frac{u \times (u \times v)}{1 + \sqrt{1 - ||u||^2}} = u + < u, v > u\\
&\implies u +_e v = u
\end{aligned}$$

(4) This comes from the squared norm formula established in the proof of (2) above, because when assuming $||v|| = 1$, we obtain:

$$\begin{aligned}
||u +_e v||^2 &= \frac{||u + v||^2 - ||u||^2 + < u, v >^2}{(1 + < u, v >)^2}\\
&= \frac{||u||^2 + 1 + 2 < u, v > - ||u||^2 + < u, v >^2}{(1 + < u, v >)^2}\\
&= \frac{1 + 2 < u, v > + < u, v >^2}{(1 + < u, v >)^2}\\
&= 1
\end{aligned}$$

(5) This is clear, from the obvious lack of symmetry of our formula.

(6) This is again clear, from the obvious lack of symmetry of our formula. □

That was nice, all this mathematics, and hope you're still with me. And good news, the formula in Theorem 15.15 is the good one, confirmed by experimental physics.

We can further build on the above, with the following surprising conclusion:

THEOREM 15.16. *Relativistic time is subject to Lorentz dilation*

$$t \to \gamma t$$

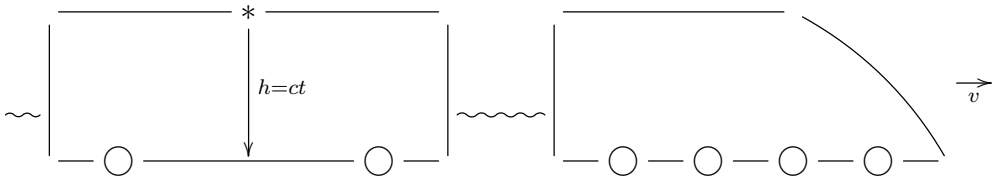
where the number $\gamma \geq 1$, called Lorentz factor, is given by the formula

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}$$

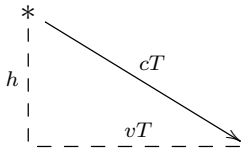
with v being the moving speed, at which time is measured.

PROOF. Assume indeed that we have a train, moving to the right with speed v, through vacuum. In order to compute the height h of the train, the passenger onboard

switches on the ceiling light bulb, measures the time t that the light needs to hit the floor, by traveling at speed c, and concludes that the train height is $h = ct$:



On the other hand, an observer on the ground will see here something different, namely a right triangle, with on the vertical the height of the train h, on the horizontal the distance vT that the train has traveled, and on the hypotenuse the distance cT that light has travelled, with T being the duration of the event, according to his watch:



Now by Pythagoras applied to this triangle, we have:

$$h^2 + (vT)^2 = (cT)^2$$

Thus, the observer on the ground will reach to the following formula for h:

$$h = \sqrt{c^2 - v^2} \cdot T$$

But h must be the same for both observers, so we have the following formula:

$$\sqrt{c^2 - v^2} \cdot T = ct$$

It follows that the two times t and T are indeed not equal, and are related by:

$$T = \frac{ct}{\sqrt{c^2 - v^2}} = \frac{t}{\sqrt{1 - v^2/c^2}} = \gamma t$$

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

Let us discuss now what happens to length. The result here is as follows:

THEOREM 15.17. *Relativistic length is subject to Lorentz contraction*

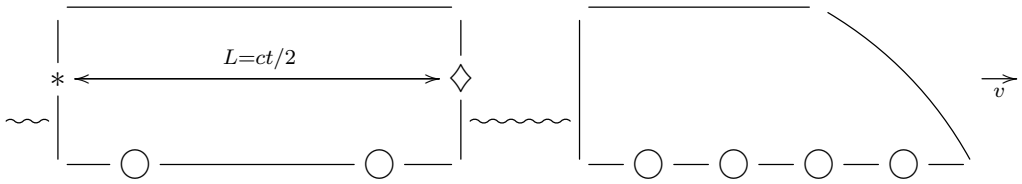
$$L \to L/\gamma$$

where the number $\gamma \geq 1$, called Lorentz factor, is given by the usual formula

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}$$

with v being the moving speed, at which length is measured.

PROOF. As before in the proof of Theorem 15.16, meaning in the same train traveling at speed v, in vacuum, imagine now that the passenger wants to measure the length L of the car. For this purpose he switches on the light bulb, now at the rear of the car, and measures the time t needed for the light to reach the front of the car, and get reflected back by a mirror installed there, according to the following scheme:



He concludes that, as marked above, the length L of the car is given by:

$$L = \frac{ct}{2}$$

Now viewed from the ground, the duration of the event is $T = T_1 + T_2$, where $T_1 > T_2$ are respectively the time needed for the light to travel forward, among others for beating v, and the time for the light to travel back, helped this time by v. More precisely, if l denotes the length of the train car viewed from the ground, the formula of T is:

$$T = T_1 + T_2 = \frac{l}{c-v} + \frac{l}{c+v} = \frac{2lc}{c^2 - v^2}$$

With this data, the formula $T = \gamma t$ of time dilation established before reads:

$$\frac{2lc}{c^2 - v^2} = \gamma t = \frac{2\gamma L}{c}$$

Thus, the two lengths L and l are indeed not equal, and related by:

$$l = \frac{\gamma L(c^2 - v^2)}{c^2} = \gamma L\left(1 - \frac{v^2}{c^2}\right) = \frac{\gamma L}{\gamma^2} = \frac{L}{\gamma}$$

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

As a main consequence of the above, beautiful as they come, we have:

THEOREM 15.18. *In the context of a relativistic object moving with speed v along the x axis, the frame change is given by the Lorentz transformation*

$$x' = \gamma(x - vt)$$

$$y' = y$$

$$z' = z$$

$$t' = \gamma(t - vx/c^2)$$

with $\gamma = 1/\sqrt{1 - v^2/c^2}$ being as usual the Lorentz factor.

PROOF. We know that, with respect to the non-relativistic formulae, x is subject to the Lorentz dilation by γ, and so we obtain, as desired:

$$x' = \gamma(x - vt)$$

Regarding y, z, these are obviously unchanged, so done with these too. Finally, regarding time t, we can use here the reverse Lorentz transformation, given by:

$$x = \gamma(x' + vt')$$

$$y = y'$$

$$z = z'$$

By using the formula of x' we can compute t', and we obtain the following formula:

$$t' = \frac{x - \gamma x'}{\gamma v} = \frac{x - \gamma^2(x - vt)}{\gamma v} = \gamma\left(t - \frac{vx}{c^2}\right)$$

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

Many other things can be said about relativity. We will be back to this.

15d. Maxwell equations

We already know about electrostatics, field lines and the Gauss law from chapter 8. However, that is only the tip of the iceberg, because when our electrical charges start moving, many interesting things happen, in relation with magnetism. So, time to discuss all this, and expect of course a lot of tricky 3D math, including vector products.

We first have the following point of view on the Gauss formula, more conceptual:

THEOREM 15.19 (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, follows from:

$$\begin{aligned}
\int_B < \nabla, E > &= \int_S < E(x), n(x) > dx \\
&= \Phi_E(S) \\
&= \frac{Q_{enc}}{\varepsilon_0} \\
&= \int_B \frac{\rho}{\varepsilon_0}
\end{aligned}$$

Regarding 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:

$$\begin{aligned}
\int_a^b < E(x), dx > &= \int_a^b \left\langle \frac{Kqx}{||x||^3}, dx \right\rangle \\
&= \int_a^b \left\langle \frac{Kq}{r^2} \cdot \frac{x}{||x||}, dx \right\rangle \\
&= \int_a^b \frac{Kq}{r^2}\, dr \\
&= \left[-\frac{Kq}{r} \right]_a^b \\
&= Kq \left(\frac{1}{r_a} - \frac{1}{r_b} \right)
\end{aligned}$$

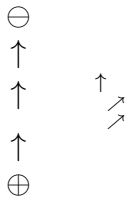
In particular the integral of E over any closed loop vanishes, and by using now the Stokes theorem, we conclude that the curl of E vanishes, as stated. □

Time for electricity. Just by feeding a light bulb with a battery, and looking at the cables, and playing a bit with them, we are led to the following interesting conclusion:

FACT 15.20. *Parallel electric currents in opposite directions repel, and parallel electric currents in the same direction attract.*

We can in fact say even more, by further playing with the cables, armed this time with a compass. The conclusion is that each cable produces some kind of "magnetic field" around it, which interestingly, is not oriented in the direction of the current, but is rather orthogonal to it, given by the right-hand rule, as follows:

FACT 15.21 (Right-hand rule). *An electric current produces a magnetic field B which is orthogonal to it, whose direction is given by the right-hand rule,*



namely wrap your right hand around the cable, with the thumb pointing towards the direction of the current, and the movement of your wrist will give you the direction of B.

This is something even more interesting than Fact 15.20. Indeed, not only moving charges produce something new, that we'll have to investigate, but they know well about 3D, and more specifically about orientation there, left and right, even if living in 1D.

And isn't this amazing. Let us summarize this discussion with:

FACT 15.22. *Charges are smart, they know about* $3D$*, and about left and right.*

With this discussed, let us go ahead and investigate the charge smartness, and more specifically the magnetic fields discovered above. In order to evaluate the properties of the magnetic fields B coming from electric currents, the simplest way is that of making them act on exterior charges Q. And we have here the following formula:

FACT 15.23 (Lorentz force law). *The magnetic force on a charge* Q*, moving with velocity* v *in a magnetic field* B*, is as follows, with* $\times$ *being a vector product:*

$$F_m = (v \times B)Q$$

In the presence of both electric and magnetic fields, the total force on Q *is*

$$F = (E + v \times B)Q$$

where E *is the electric field.*

Here the occurrence of the vector product $\times$ is not surprising, due to the fact that the right-hand rule appears both in Fact 15.21, and in the definition of $\times$. In fact, the Lorentz force law is just a fancy mathematical reformulation of Fact 15.21, telling us that, once the magnetic fields B duly axiomatized, and with this being a remaining problem, their action on exterior charges Q will be proportional to the charge, $F_m \sim Q$, and with the orientation and magnitude coming from the 3D of the right-hand rule in Fact 15.21.

As an interesting application of the Lorentz force law, we have:

THEOREM 15.24. *Magnetic forces do not work.*

PROOF. This might seem quite surprising, but the math is there, as follows:

$$\begin{aligned} dW_m &= < F_m, dx > \\ &= < (v \times B)Q, v\,dt > \\ &= Q < v \times B, v > dt \\ &= 0 \end{aligned}$$

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

Moving ahead now, let us talk axiomatization of electric currents, including units. We have here the following definition, clarifying our previous discussion about coulombs:

DEFINITION 15.25. *The electric currents* I *are measured in amperes, given by:*

$$1A = 1C/s$$

As a consequence, the coulomb is given by $1C = 1A \times 1s$.

With this notion in hand, let us keep building the math and physics of magnetism. So, assume that we are dealing with an electric current I, producing a magnetic field B. In this context, the Lorentz force law from Fact 15.23 takes the following form:

$$F_m = \int (dx \times B)I$$

The current being typically constant along the wire, this reads:

$$F_m = I \int dx \times B$$

We can deduce from this the following result:

THEOREM 15.26. *The volume current density J satisfies*

$$< \nabla, J >= -\dot{\rho}$$

called continuity equation.

PROOF. We have indeed the following computation, for any surface S enclosing a volume V, based on the Lorentz force law, and on the overall chage conservation:

$$\begin{aligned}\int_V < \nabla, J > &= \int_S < J, n(x) > dx\\ &= -\frac{d}{dt}\int_V \rho\\ &= -\int_V \dot{\rho}\end{aligned}$$

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

Moving ahead now, let us formulate the following definition:

DEFINITION 15.27. *The realm of magnetostatics is that of the steady currents,*

$$\dot{\rho} = 0 \quad , \quad \dot{J} = 0$$

in analogy with electrostatics, dealing with fixed charges.

As a first observation, for steady currents the continuity equation reads:

$$< \nabla, J >= 0$$

We have here a bit of analogy between electrostatics and magnetostatics, and with this in mind, let us look for equations for the magnetic field B. We have:

FACT 15.28 (Biot-Savart law). *The magnetic field of a steady line current is given by*

$$B = \frac{\mu_0}{4\pi}\int \frac{I \times x}{||x||^3}$$

where μ_0 is a certain constant, called the magnetic permeability of free space.

This law not only gives us all we need, for studying steady currents, and we will talk about this in a moment, with math and everything, but also makes an amazing link with the Coulomb force law, due to the following fact, which is also part of it:

FACT 15.29 (Biot-Savart, continued). *The electric permittivity of free space ε_0 and the magnetic permeability of free space μ_0 are related by the formula*

$$\varepsilon_0\mu_0 = \frac{1}{c^2}$$

where c is as usual the speed of light.

This is something truly remarkable, that will have numerous consequences, for instance by making the link with Einstein's relativity theory, also crucially involving c.

But, first of all, this is certainly an invitation to rediscuss units and constants:

CONVENTIONS 15.30. *We keep using standard units, namely meters, kilograms, seconds, along with the coulomb, defined by the following exact formula*

$$1C = \frac{5 \times 10^{18}}{0.801\ 088\ 317}\ e$$

with e being minus the charge of the electron, which in practice means:

$$1C \simeq 6.241 \times 10^{18}\, e$$

We will also use the ampere, defined as $1A = 1C/s$, for measuring currents.

In what regards constants, however, time to do some cleanup. We have been boycotting for some time already the Coulomb constant K, and using instead $\varepsilon_0 = 1/(4\pi K)$, due to the ubiquitous 4π factor, first appearing as the area of the unit sphere, $A = 4\pi$, in the computation for the Gauss law for the unit sphere. Together with Fact 15.29, this suggests using the numbers ε_0, μ_0 as our new constants, by always keeping in mind $\varepsilon_0\mu_0 = 1/c^2$, and by having of course c as constant too, and we are led in this way into:

CONVENTIONS 15.31. *We use from now on as constants the electric permittivity of free space ε_0 and the magnetic permeability of free space μ_0, given by*

$$\varepsilon_0 = 8.854\ 187\ 8128(13) \times 10^{-12}$$

$$\mu_0 = 1.256\ 637\ 062\ 12(19) \times 10^{-6}$$

as well as the speed of light, given by the following exact formula,

$$c = 299\,792\,458$$

which are related by $\varepsilon_0\mu_0 = 1/c^2$, and with the Coulomb constant being $K = 1/(4\pi\varepsilon_0)$.

Getting back now to theory and math, the Biot-Savart law has as consequence:

THEOREM 15.32. *We have the following formula:*

$$< \nabla, B >= 0$$

That is, the divergence of the magnetic field vanishes.

PROOF. We recall that the Biot-Savart law tells us that we have:

$$B = \frac{\mu_0}{4\pi}\int \frac{I \times x}{||x||^3}$$

By applying the divergence operator to this formula, we obtain:

$$\begin{aligned} < \nabla, B > &= \frac{\mu_0}{4\pi}\int \left\langle \nabla, \frac{I \times x}{||x||^3} \right\rangle \\ &= \frac{\mu_0}{4\pi}\int \left\langle \nabla \times J, \frac{x}{||x||^3} \right\rangle - \left\langle \nabla \times \frac{x}{||x||^3}, J \right\rangle \\ &= \frac{\mu_0}{4\pi}\int \left\langle 0, \frac{x}{||x||^3} \right\rangle - \langle 0, J \rangle \\ &= 0 \end{aligned}$$

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

Regarding now the curl, we have here a similar result, as follows:

THEOREM 15.33 (Ampère law). *We have the following formula,*

$$\nabla \times B = \mu_0 J$$

computing the curl of the magnetic field.

PROOF. Again, we use the Biot-Savart law, telling us that we have:

$$B = \frac{\mu_0}{4\pi}\int \frac{I \times x}{||x||^3}$$

By applying the curl operator to this formula, we obtain:

$$\begin{aligned} \nabla \times B &= \frac{\mu_0}{4\pi}\int \nabla \times \frac{I \times x}{||x||^3} \\ &= \frac{\mu_0}{4\pi}\int \left\langle \nabla, \frac{x}{||x||^3} \right\rangle J - < \nabla, J > \frac{x}{||x||^3} \\ &= \frac{\mu_0}{4\pi}\int 4\pi\delta_x \cdot J - \frac{\mu_0}{4\pi} \cdot 0 \\ &= \mu_0 \int \delta_x \cdot J \\ &= \mu_0 J \end{aligned}$$

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

As a conclusion to all this, the equations of magnetostatics are as follows:

THEOREM 15.34. *The equations of magnetostatics are*

$$< \nabla, B >= 0 \quad , \quad \nabla \times B = \mu_0 J$$

with the second equation being the Ampère law.

PROOF. This follows indeed from the above discussion, and more specifically from Theorem 15.32 and Theorem 15.33, which both follow from the Biot-Savart law. □

Observe the obvious analogy with the Gauss equations of electrostatics, namely:

$$< \nabla, E >= \frac{\rho}{\varepsilon_0} \quad , \quad \nabla \times E = 0$$

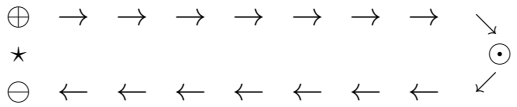
As a conclusion to all this, looks like someone has played here with basic 3D math, vectors, products and so on, and messed them up, as for electrostatics to become magnetostatics, and vice versa. More on this later, when talking about unification.

Quite remarkably now, and at the origin of all modern theory of electromagnetism, and of any type of modern electrical engineering too, we have:

FACT 15.35 (Faraday laws). *The following happen:*

(1) *Moving a wire loop γ through a magnetic field B produces a current through γ.*
(2) *Keeping γ fixed, but changing the strength of B, produces too current through γ.*

In order to understand what is going on here, let us start with the simplest electric loop that we know, namely a battery feeding a light bulb:



Here the star stands for the fact that we don't really know what happens inside the battery, typically a complicated chemical process. Nor we will actually worry about the bulb, let us simply assume that this bulb does not exist at all. We will be interested in the force driving the current around the loop, and we have here:

PROPOSITION 15.36. *When writing the force driving the current through a loop γ as*

$$F = F_\star + F_e$$

with $F_\star$ coming from the source, and F_e coming from the loop, the quantity

$$\mathcal{E} = \int_\gamma < F(x), dx >$$

called electromotive force, or emf of the loop, is simply obtained by integrating $F_\star$.

PROOF. We have indeed the following computation, based on the fact that F_e being an electrostatic force, its integral over the loop vanishes:

$$\begin{aligned}\mathcal{E} &= \int_\gamma < F(x), dx > \\ &= \int_\gamma < F_\star(x), dx > + \int_\gamma < F_e(x), dx > \\ &= \int_\gamma < F_\star(x), dx > +0 \\ &= \int_\gamma < F_\star(x), dx >\end{aligned}$$

Thus, we have our result, and with the remark of course that the emf $\mathcal{E} \in \mathbb{R}$ is not really a force, but this is the standard terminology, and we will use it. □

In relation now with the Faraday principles from Fact 15.35, these can be fine-tuned, and reformulated in terms of the emf, in the following way:

FACT 15.37 (Faraday). *The emf of a loop γ moving through a magnetic field B is*

$$\mathcal{E} = -\dot{\Phi}$$

where Φ is the flux of the field B through the loop γ, given by:

$$\Phi = \int_\gamma < B(x), dx >$$

As for the emf of a fixed loop γ in a changing magnetic field B, this is

$$\mathcal{E} = -\int_\gamma < \dot{B}(x), dx >$$

which by Stokes is equivalent to the Faraday law $\Delta \times E = -\dot{B}$.

All the above is very useful in electromechanics, for construcing electric motors. Getting back now to theory, the above considerations lead to the following conclusion:

FACT 15.38 (Faraday). *In the context of moving chages, the electrostatics law*

$$\nabla \times E = 0$$

must be replaced by the following equation,

$$\nabla \times E = -\dot{B}$$

called Faraday law.

Along the same lines, and following now Maxwell, there is a correction as well to be made to the main law of magnetostatics, namely the Ampère law, as follows:

FACT 15.39 (Maxwell). *In the context of moving chages, the Ampère law*

$$\nabla \times B = \mu_0 J$$

must be replaced by the following equation,

$$\nabla \times B = \mu_0(J + \varepsilon_0 \dot{E})$$

called Ampère law with Maxwell correction term.

Now by putting everything together, and perhaps after doublecheking as well, with all sorts of experiments, that the remaining electrostatics and magnetostatics laws, that we have not modified, work indeed fine in the dynamic setting, we obtain:

THEOREM 15.40 (Maxwell). *Electrodynamics is governed by the formulae*

$$< \nabla, E >= \frac{\rho}{\varepsilon_0} \quad , \quad < \nabla, B >= 0$$

$$\nabla \times E = -\dot{B} \quad , \quad \nabla \times B = \mu_0 J + \mu_0 \varepsilon_0 \dot{E}$$

called Maxwell equations.

PROOF. This follows indeed from the above, the details being as follows:

(1) The first equation is the Gauss law, that we know well.

(2) The second equation is something anonymous, that we know well too.

(3) The third equation is a previously anonymous law, modified into Faraday's law.

(4) And the fourth equation is the Ampère law, as modified by Maxwell. □

The Maxwell equations are in fact not the end of everything, because in the context of the 2-body problem, they must be replaced by quantum mechanics. More later.

We will be back to this later, with more about it. In the meantime, back to abstract electromagnetism, we have here the following key result, due to Lorentz:

THEOREM 15.41. *The Maxwell equations found before, namely*

$$< \nabla, E >= \frac{\rho}{\varepsilon_0} \quad , \quad < \nabla, B >= 0$$

$$\nabla \times E = -\dot{B} \quad , \quad \nabla \times B = \mu_0 J + \mu_0 \varepsilon_0 \dot{E}$$

are invariant under Lorentz transformations.

PROOF. Consider an electromagnetic field (E, B). This is altered by a Lorentz transformation into a field (E', B'), the equations for E' being as follows:

$$E'_x = E_x$$

$$E'_y = \gamma(E_y - vB_z)$$

$$E'_z = \gamma(E_z + vB_y)$$

As for the equations of B', these are quite similar, as follows:

$$B'_x = B_x$$
$$B'_y = \gamma\left(B_y + \frac{v}{c^2}\,E_z\right)$$
$$B'_z = \gamma\left(B_z - \frac{v}{c^2}\,E_y\right)$$

In order to do the math, consider the following matrices, with $\beta = v/c$ as usual:

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \gamma & 0 \\ 0 & 0 & \gamma \end{pmatrix} \quad , \quad M = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\beta\gamma \\ 0 & \beta\gamma & 0 \end{pmatrix}$$

In terms of these matrices, the formulae for the new field (E', B') read:

$$E' = DE + cMB \quad , \quad B' = DB - \frac{M}{c}E$$

But this is already not that bad, and starting from these formulae, it is possible to prove that (E', B') satisfies as well the Maxwell equations, as desired. □

15e. Exercises

Exercises:

Exercise 15.42.

Exercise 15.43.

Exercise 15.44.

Exercise 15.45.

Exercise 15.46.

Exercise 15.47.

Exercise 15.48.

Exercise 15.49.

Bonus exercise.

CHAPTER 16

Solid angles

16a. Solid angles

Solid angles. Many things can be said here.

16b. Waves, optics

Let us discuss now light. We first have the following basic result:

THEOREM 16.1. *The wave equation in $\mathbb{R}^N$ is*

$$\ddot{\varphi} = v^2 \Delta\varphi$$

where Δ is the Laplace operator, given by

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

and $v > 0$ is the propagation speed.

PROOF. As a first disclaimer, the equation in the statement is what comes out of experiments. However, allowing us a bit of imagination, and trust in this imagination, we can mathematically "prove" this equation, by discretizing, as follows:

(1) Let us first consider the 1D case. In order to understand the propagation of waves, we will model $\mathbb{R}$ as a network of balls, with springs between them, as follows:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

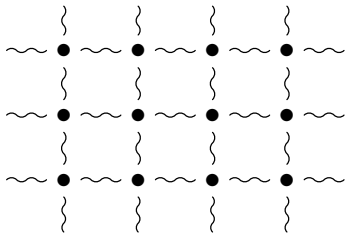
$$\ddot{\varphi}(x) = \frac{KL^2}{M} \cdot \frac{\varphi(x+l) - 2\varphi(x) + \varphi(x-l)}{l^2}$$

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

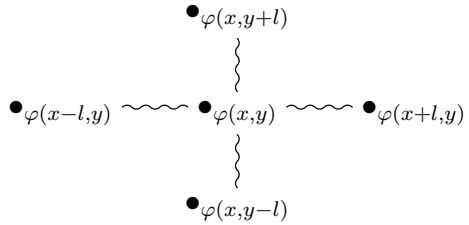
$$\ddot{\varphi}(x) = \frac{KL^2}{M} \cdot \frac{d^2\varphi}{dx^2}(x)$$

We are therefore led to the wave equation in the statement, which is $\ddot{\varphi} = v^2\varphi''$ in our present $N = 1$ dimensional case, the propagation speed being $v = \sqrt{K/M} \cdot L$.

(3) In 2 dimensions now, the same argument carries on. Indeed, we can use here a lattice model as follows, with all the edges standing for small springs:



As before in one dimension, we send an impulse, and we zoom on one ball. The situation here is as follows, with l being the spring length:



By doing the math, as before in 1D, we are led to the following equation:

$$\begin{aligned}\ddot{\varphi}(x,y) &= \frac{KL^2}{M}\times\frac{\varphi(x+l,y)-2\varphi(x,y)+\varphi(x-l,y)}{l^2}\\ &+ \frac{KL^2}{M}\times\frac{\varphi(x,y+l)-2\varphi(x,y)+\varphi(x,y-l)}{l^2}\end{aligned}$$

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

$$\ddot{\varphi}(x,y)=\frac{KL^2}{M}\left(\frac{d^2\varphi}{dx^2}+\frac{d^2\varphi}{dy^2}\right)(x,y)$$

Thus, we are led in this way to the wave equation in two dimensions, as in the statement, with $v=\sqrt{K/M}\cdot L$ being the propagation speed of our wave.

(4) In 3 dimensions now, which is the case of the main interest, corresponding to our real-life world, the same argument carries over, and the wave equation is as follows:

$$\ddot{\varphi}(x,y,z)=v^2\left(\frac{d^2\varphi}{dx^2}+\frac{d^2\varphi}{dy^2}+\frac{d^2\varphi}{dz^2}\right)(x,y,z)$$

Finally, the same argument carries on in arbitrary N dimensions. □

The point now is that, in relation with the Maxwell equations, we have:

THEOREM 16.2. *In regions of space where there is no charge or current present the Maxwell equations for electrodynamics read*

$$<\nabla,E>=<\nabla,B>=0$$

$$\nabla\times E=-\dot{B}\quad,\quad\nabla\times B=\dot{E}/c^2$$

and both the electric field E and magnetic field B are subject to the wave equation

$$\ddot{\varphi}=c^2\Delta\varphi$$

where $\Delta=\sum_i d^2/dx_i^2$ is the Laplace operator, and c is the speed of light.

PROOF. Under the circumstances in the statement, namely no charge or current present, the Maxwell equations from chapter 15 simply read:

$$< \nabla, E >=< \nabla, B >= 0$$

$$\nabla \times E = -\dot{B} \quad , \quad \nabla \times B = \dot{E}/c^2$$

Now by applying the curl operator to the last two equations, we obtain:

$$\nabla \times (\nabla \times E) = -\nabla \times \dot{B} = -(\nabla \times B)' = -\ddot{E}/c^2$$

$$\nabla \times (\nabla \times B) = \nabla \times \dot{E}/c^2 = (\nabla \times E)'/c^2 = -\ddot{B}/c^2$$

But the double curl operator is subject to the following formula:

$$\nabla \times (\nabla \times \varphi) = \nabla < \nabla, \varphi > -\Delta\varphi$$

Now by using the first two equations, we are led to the conclusion in the statement. □

So, what is light? Light is the wave predicted by Theorem 16.2, traveling in vacuum at the maximum possible speed, c, and with an important extra property being that it depends on a real positive parameter, that can be called, upon taste, frequency, wavelength, or color. And in what regards the creation of light, the mechanism here is as follows:

FACT 16.3. *An accelerating or decelerating charge produces electromagnetic radiation, called light, whose frequency and wavelength can be explicitly computed.*

This phenomenon can be observed is a variety of situations, such as the usual light bulbs, where electrons get decelerated by the filament, acting as a resistor, or in usual fire, which is a chemical reaction, with the electrons moving around, as they do in any chemical reaction, or in more complicated machinery like nuclear plants, particle accelerators, and so on, leading there to all sorts of eerie glows, of various colors.

Moving ahead, let us go back to the wave equation $\ddot{\varphi} = v^2\Delta\varphi$ from Theorem 16.1, and try to understand its simplest solutions. In 1D, the situation is as follows:

THEOREM 16.4. *The 1D wave equation has as basic solutions the functions*

$$\varphi(x) = A\cos(kx - wt + \delta)$$

with A being called amplitude, $kx-wt+\delta$ being called the phase, k being the wave number, w being the angular frequency, and δ being the phase constant. We have

$$\lambda = \frac{2\pi}{k} \quad , \quad T = \frac{2\pi}{kv} \quad , \quad \nu = \frac{1}{T} \quad , \quad w = 2\pi\nu$$

relating the wavelength λ, period T, frequency ν, and angular frequency w. Moreover, any solution of the wave equation appears as a linear combination of such basic solutions.

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

(1) Our first claim is that the function φ in the statement satisfies indeed the wave equation, with speed $v = w/k$. For this purpose, observe that we have:

$$\ddot{\varphi} = -w^2\varphi \quad , \quad \frac{d^2\varphi}{dx^2} = -k^2\varphi$$

Thus, the wave equation is indeed satisfied, with speed $v = w/k$:

$$\ddot{\varphi} = \left(\frac{w}{k}\right)^2 \frac{d^2\varphi}{dx^2} = v^2 \frac{d^2\varphi}{dx^2}$$

(2) Regarding now the other things in the statement, all this is basically terminology, which is very natural, when thinking how $\varphi(x) = A\cos(kx - wt + \delta)$ propagates. As for the last assertion, this is something standard, coming from Fourier analysis. □

As a first observation, the above result invites the use of complex numbers. Indeed, we can write the solutions that we found in a more convenient way, as follows:

$$\varphi(x) = Re\left[A\, e^{i(kx-wt+\delta)}\right]$$

And we can in fact do even better, by absorbing the quantity $e^{i\delta}$ into the amplitude A, which becomes now a complex number, and writing our formula as:

$$\varphi = Re(\widetilde{\varphi}) \quad , \quad \widetilde{\varphi} = \widetilde{A}e^{i(kx-wt)}$$

Moving ahead now towards electromagnetism and 3D, let us formulate:

DEFINITION 16.5. *A monochromatic plane wave is a solution of the 3D wave equation which moves in only 1 direction, making it in practice a solution of the 1D wave equation, and which is of the special from found in Theorem 16.4, with no frequencies mixed.*

In other words, we are making here two assumptions on our wave. First is the 1-dimensionality assumption, which gets us into the framework of Theorem 16.4. And second is the assumption, in connection with the Fourier decomposition result from the end of Theorem 16.4, that our solution is of "pure" type, meaning a wave having a well-defined wavelenght and frequency, instead of being a "packet" of such pure waves.

All this is still mathematics, and making now the connection with physics and electromagnetism, and more specifically with Theorem 16.2 and Fact 16.3, we have:

FACT 16.6. *Physically speaking, a monochromatic plane wave is the electromagnetic radiation appearing as in Theorem 16.2 and Fact 16.3, via equations of type*

$$E = Re(\widetilde{E}) \quad : \quad \widetilde{E} = \widetilde{E}_0\, e^{i(<k,x>-wt)}$$

$$B = Re(\widetilde{B}) \quad : \quad \widetilde{B} = \widetilde{B}_0\, e^{i(<k,x>-wt)}$$

with the wave number being now a vector, $k \in \mathbb{R}^3$. Moreover, it is possible to add to this an extra parameter, accounting for the possible polarization of the wave.

To be more precise, what we are doing here is to import the conclusions of our mathematical discussion so far, from Theorem 16.4 and Definition 16.5, into the context of our original physics discussion, from Fact 16.3. And also to add an extra twist coming from physics, and more specifically from the notion of polarization. More on this later.

In any case, we have now a decent intuition about what light is, and more on this later, and let us discuss now the examples. The idea is that we have various types of light, depending on frequency and wavelength. These are normally referred to as "electromagnetic waves", but for keeping things simple and luminous, we will keep using the familiar term "light". The classification, in a rough form, is as follows:

Frequency	Type	Wavelength
	–	
$10^{18} - 10^{20}$	γ rays	$10^{-12} - 10^{-10}$
$10^{16} - 10^{18}$	X – rays	$10^{-10} - 10^{-8}$
$10^{15} - 10^{16}$	UV	$10^{-8} - 10^{-7}$
	–	
$10^{14} - 10^{15}$	blue	$10^{-7} - 10^{-6}$
$10^{14} - 10^{15}$	yellow	$10^{-7} - 10^{-6}$
$10^{14} - 10^{15}$	red	$10^{-7} - 10^{-6}$
	–	
$10^{11} - 10^{14}$	IR	$10^{-6} - 10^{-3}$
$10^{9} - 10^{11}$	microwave	$10^{-3} - 10^{-1}$
$1 - 10^{9}$	radio	$10^{-1} - 10^{8}$

Observe the tiny space occupied by the visible light, all colors there, and the many more missing, being squeezed under the $10^{14} - 10^{15}$ frequency banner. Here is a zoom on that part, with of course the remark that all this, colors, is something subjective:

Frequency THz $= 10^{12}$ Hz	Color	Wavelength nm $= 10^{-9}$ m
	–	
$670 - 790$	violet	$380 - 450$
$620 - 670$	blue	$450 - 485$
$600 - 620$	cyan	$485 - 500$
$530 - 600$	green	$500 - 565$
$510 - 530$	yellow	$565 - 590$
$480 - 510$	orange	$590 - 625$
$400 - 480$	red	$625 - 750$

Back now to our business, with the above theory of light in hand, we can do some optics. Light usually comes in "bundles", with waves of several wavelenghts coming at the same time, from the same source, and the first challenge is that of separating these wavelenghts. In order to discuss this, let us start with the following fact:

FACT 16.7. *Inside a linear, homogeneous medium, where there is no free charge or current present, the Maxwell equations for electrodynamics read*

$$< \nabla, E >=< \nabla, B >= 0$$

$$\nabla \times E = -\dot{B} \quad , \quad \nabla \times B = \varepsilon\mu\dot{E}$$

with E, B being as before the electric and the magnetic field, and with $\varepsilon > \varepsilon_0$ and $\mu > \mu_0$ being the electric permittivity and magnetic permeability of the medium.

Observe that this statement is precisely the first part of Theorem 16.2, with the vacuum constants ε_0, μ_0 being now replaced by their versions ε, μ, concerning the medium in question. In what regards now the second part of Theorem 16.2, we have:

THEOREM 16.8. *Inside a linear, homogeneous medium, where there is no free charge or free current present, both E and B are subject to the wave equation*

$$\ddot{\varphi} = v^2\Delta\varphi$$

with v being the speed of light inside the medium, given by

$$v = \frac{c}{n} \quad : \quad n = \sqrt{\frac{\varepsilon\mu}{\varepsilon_0\mu_0}}$$

with the quantity on the right $n > 1$ being called refraction index of the medium.

PROOF. This is something that we know well in vacuum, from the above, and the proof in general is identical, with the resulting speed being:

$$v = \frac{1}{\sqrt{\varepsilon\mu}}$$

But this formula can be written is a more familiar from, as above. □

Next in line, and of interest for us, as we will soon discover, we have:

FACT 16.9. *When traveling through a material, and hitting a new material, some of the light gets reflected, at the same angle, and some of it gets refracted, at a different angle, depending both on the old and the new material, and on the wavelength.*

As a basic formula here, we have the famous Snell law, which relates the incidence angle θ_1 to the refraction angle θ_2, via the following simple formula:

$$\frac{\sin\theta_2}{\sin\theta_1} = \frac{n_1(\lambda)}{n_2(\lambda)}$$

Here $n_i(\lambda)$ are the refraction indices of the two materials, adjusted for the wavelength, and with this adjustment for wavelength being the whole point, which is something quite complicated. For an introduction to all this, we refer for instance to Griffiths [43].

As a simple consequence of the above, of great practical interest, we have:

THEOREM 16.10. *Light can be decomposed, by using a prism.*

PROOF. This follows from Fact 16.9. Indeed, when hitting a piece of glass, provided that the hitting angle is not 90°, the light will decompose over the wavelenghts present, with the corresponding refraction angles depending on these wavelengths. And we can capture these split components at the exit from the piece of glass, again deviated a bit, provided that the exit surface is not parallel to the entry surface. And the simplest device doing the job, that is, having two non-parallel faces, is a prism. □

With this in hand, we can now talk about spectroscopy:

FACT 16.11. *We can study events via spectroscopy, by capturing the light the event has produced, decomposing it with a prism, carefully recording its "spectral signature", consisting of the wavelenghts present, and their density, and then doing some reverse engineering, consisting in reconstructing the event out of its spectral signature.*

This is the main principle of spectroscopy, and applications of it, of all kinds, abound. In practice, the mathematical tool needed for doing the "reverse engineering" mentioned above is the Fourier transform, which allows the decomposition of packets of waves, into monochromatic components. Finally, let us mention too that, needless to say, the event can be reconstructed only partially out of its spectral signature.

16c. Particle physics

Getting now to some truly exciting applications of light and spectroscopy, let us discuss the beginnings of the atomic theory. There is a long story here, involving many discoveries, around 1890-1900, focusing on hydrogen H. We will present here things a bit retrospectively. First on our list is the following discovery, by Lyman in 1906:

FACT 16.12 (Lyman). *The hydrogen atom has spectral lines given by the formula*

$$\frac{1}{\lambda} = R\left(1 - \frac{1}{n^2}\right)$$

where $R \simeq 1.097 \times 10^7$ *and* $n \geq 2$*, which are as follows,*

n	Name	Wavelength	Color
	–	–	
2	α	121.567	UV
3	β	102.572	UV
4	γ	97.254	UV
⋮	⋮	⋮	⋮
∞	limit	91.175	UV

called Lyman series of the hydrogen atom.

Observe that all the Lyman series lies in UV, which is invisible to the naked eye. Due to this fact, this series, while theoretically being the most important, was discovered only second. The first discovery, which was the big one, and the breakthrough, was by Balmer, the founding father of all this, back in 1885, in the visible range, as follows:

FACT 16.13 (Balmer). *The hydrogen atom has spectral lines given by the formula*

$$\frac{1}{\lambda} = R\left(\frac{1}{4} - \frac{1}{n^2}\right)$$

where $R \simeq 1.097 \times 10^7$ *and* $n \geq 3$*, which are as follows,*

n	Name	Wavelength	Color
	–	–	
3	α	656.279	red
4	β	486.135	aqua
5	γ	434.047	blue
6	δ	410.173	violet
7	ε	397.007	UV
$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	limit	346.600	UV

called Balmer series of the hydrogen atom.

So, this was Balmer's original result, which started everything. As a third main result now, this time in IR, due to Paschen in 1908, we have:

FACT 16.14 (Paschen). *The hydrogen atom has spectral lines given by the formula*

$$\frac{1}{\lambda} = R\left(\frac{1}{9} - \frac{1}{n^2}\right)$$

where $R \simeq 1.097 \times 10^7$ *and* $n \geq 4$*, which are as follows,*

n	Name	Wavelength	Color
	–	–	
4	α	1875	IR
5	β	1282	IR
6	γ	1094	IR
$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	limit	820.4	IR

called Paschen series of the hydrogen atom.

Observe the striking similarity between the above three results. In fact, we have here the following fundamental, grand result, due to Rydberg in 1888, based on the Balmer series, and with later contributions by Ritz in 1908, using the Lyman series as well:

CONCLUSION 16.15 (Rydberg, Ritz). *The spectral lines of the hydrogen atom are given by the Rydberg formula, depending on integer parameters $n_1 < n_2$,*

$$\frac{1}{\lambda_{n_1 n_2}} = R\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

with R being the Rydberg constant for hydrogen, which is as follows:

$$R \simeq 1.096\ 775\ 83 \times 10^7$$

These spectral lines combine according to the Ritz-Rydberg principle, as follows:

$$\frac{1}{\lambda_{n_1 n_2}} + \frac{1}{\lambda_{n_2 n_3}} = \frac{1}{\lambda_{n_1 n_3}}$$

Similar formulae hold for other atoms, with suitable fine-tunings of R.

The spectral lines at $n_1 = 4, 5, 6, \ldots$, predicted by this conclusion, were discovered later, by Brackett in 1922, Pfund in 1924, Humphreys in 1953, and others aftwerwards, with all these extra lines being in far IR. The simplified complete table is as follows:

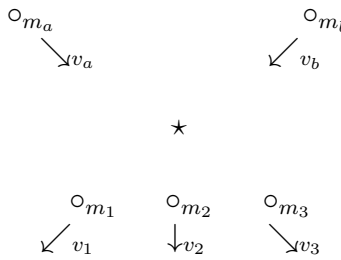
n_1	n_2	Series name	Wavelength $n_2 = \infty$	Color $n_2 = \infty$
		–	–	
1	$2 - \infty$	Lyman	91.13 nm	UV
2	$3 - \infty$	Balmer	364.51 nm	UV
3	$4 - \infty$	Paschen	820.14 nm	IR
		–	–	
4	$5 - \infty$	Brackett	1458.03 nm	far IR
5	$6 - \infty$	Pfund	2278.17 nm	far IR
6	$7 - \infty$	Humphreys	3280.56 nm	far IR
⋮	⋮	⋮	⋮	⋮

The explanation for the above comes from the following claim, by Bohr and others, which was subsequently proved by Heisenberg, and Schrödinger:

CLAIM 16.16 (Bohr and others). *The atoms are formed by a core of protons and neutrons, surrounded by a cloud of electrons, basically obeying to a modified version of electromagnetism. And with a fine mechanism involved, as follows:*

(1) *The electrons are free to move only on certain specified elliptic orbits, labelled $1, 2, 3, \ldots$, situated at certain specific heights.*
(2) *The electrons can jump or fall between orbits $n_1 < n_2$, absorbing or emitting light and heat, that is, electromagnetic waves, as accelerating charges.*
(3) *The energy of such a wave, coming from $n_1 \to n_2$ or $n_2 \to n_1$, is given, via the Planck viewpoint, by the Rydberg formula, applied with $n_1 < n_2$.*
(4) *The simplest such jumps are those observed by Lyman, Balmer, Paschen. And multiple jumps explain the Ritz-Rydberg formula.*

Now that we know a bit about elementary particles, and quantum mechanics, let us talk about interactions between these 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 16.17. *In particle physics, we have two main types of interactions, namely:*

(1) *Decay. This is when a particle decomposes, as a result of whatever internal mechanism, into a sum of other particles,* $*_0 \to *_1 + \ldots + *_n$.
(2) *Scattering. This is when two particles meet, by colliding, or almost, and combine and decompose into a sum of other particles,* $*_a + *_b \to *_1 + \ldots + *_n$.

Obviously, all this departs a bit from our classical mechanics knowledge, as explained above, and several comments are in order here, as follows:

(1) In what regards decay, something that we talked a lot about, when doing thermodynamics, and then quantum mechanics, is an electron of an atom changing its energy level, and emitting a photon. But this can be regarded as being decay.

(2) As for scattering, the simplest example here appears again from an electron of an atom, changing its energy level, but this time by absorbing a photon. Of course, there are many other possible examples, such as the electron-positron annihilation.

Getting to work for good now, decay and its mathematics. Ignoring the physics, this is basically a matter of probability and statistics, and the basics here are as follows:

THEOREM 16.18. *In the context of decay, the quantity to look at is the decay rate* λ*, which is the probability per unit time that the particle will disintegrate. With this:*

(1) *The number of particles remaining at time* $t > 0$ *is* $N_t = e^{-\lambda t} N_0$.
(2) *The mean lifetime of a particle is* $\tau = 1/\lambda$.
(3) *The half-life of the substance is* $t_{1/2} = (\log 2)/\lambda$.

PROOF. As said above, this is basic probability, as follows:

(1) In mathematical terms, our definition of the decay rate reads:

$$\frac{dN}{dt} = -\lambda N$$

By integrating, we are led to the formula in the statement, namely:

$$N_t = e^{-\lambda t} N_0$$

(2) Let us first convert what we have into a probability law. We have:

$$\int_0^\infty N_t dt = \int_0^\infty N_0 e^{-\lambda t} dt = \frac{N_0}{\lambda}$$

Thus, the density of the probability decay function is given by:

$$f(t) = \frac{\lambda}{N_0} \cdot N_0 e^{-\lambda t} = \lambda e^{-\lambda t}$$

We can now compute the mean lifetime, by integrating by parts, as follows:

$$\begin{aligned}
\tau &= < t > \\
&= \int_0^\infty t f(t) dt \\
&= \int_0^\infty \lambda t e^{-\lambda t} dt \\
&= \int_0^\infty t(-e^{-\lambda t})' dt \\
&= \int_0^\infty e^{-\lambda t} dt \\
&= \frac{1}{\lambda}
\end{aligned}$$

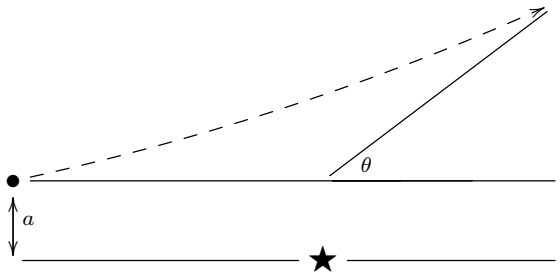
(3) Finally, regarding the half-life, this is by definition the time $t_{1/2}$ required for the decaying quantity to fall to one-half of its initial value. Mathematically, this means:

$$N_t = 2^{-\frac{t}{t_{1/2}}} N_0$$

Now by comparing with $N_t = e^{-\lambda t} N_0$, this gives $t_{1/2} = (\log 2)/\lambda$, as stated. □

Getting now to scattering, this is something far more familiar, because we can fully use here our experience from classical mechanics. Let us start with:

DEFINITION 16.19. *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:

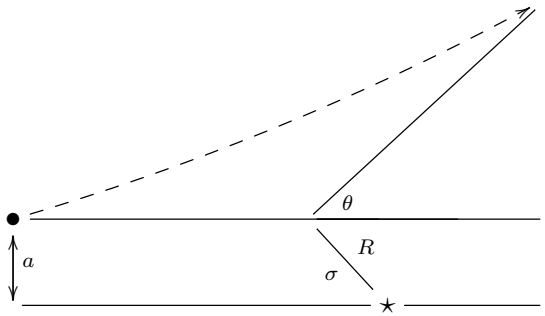
PROPOSITION 16.20. *In the context of classical particle colliding elastically with a hard sphere of radius $R > 0$, we have the formula*

$$a = R \cos \frac{\theta}{2}$$

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 16.19, as follows:

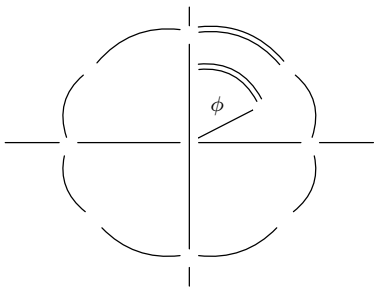
DEFINITION 16.21. *In the general context of scattering, we can:*

(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 16.22. *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 16.21, 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 16.23. *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 16.20 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. □

16d. Decay, scattering

Decay, scattering.

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, Principles of mathematics (2025).

[7] T. Banica, Calculus and applications (2024).

[8] T. Banica, Introduction to modern physics (2025).

[9] G.K. Batchelor, An introduction to fluid dynamics, Cambridge Univ. Press (1967).

[10] M.J. Benton, Vertebrate paleontology, Wiley (1990).

[11] M.J. Benton and D.A.T. Harper, Introduction to paleobiology and the fossil record, Wiley (2009).

[12] S.J. Blundell and K.M. Blundell, Concepts in thermal physics, Oxford Univ. Press (2006).

[13] B. Bollobás, Modern graph theory, Springer (1998).

[14] S.M. Carroll, Spacetime and geometry, Cambridge Univ. Press (2004).

[15] P.M. Chaikin and T.C. Lubensky, Principles of condensed matter physics, Cambridge Univ. Press (1995).

[16] A.R. Choudhuri, Astrophysics for physicists, Cambridge Univ. Press (2012).

[17] J. Clayden, S. Warren and N. Greeves, Organic chemistry, Oxford Univ. Press (2012).

[18] D.D. Clayton, Principles of stellar evolution and nucleosynthesis, Univ. of Chicago Press (1968).

[19] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).

[20] A. Cottrell, An introduction to metallurgy, CRC Press (1997).

[21] C. Darwin, On the origin of species (1859).

[22] P.A. Davidson, Introduction to magnetohydrodynamics, Cambridge Univ. Press (2001).

[23] P.A.M. Dirac, Principles of quantum mechanics, Oxford Univ. Press (1930).

[24] S. Dodelson, Modern cosmology, Academic Press (2003).

[25] S.T. Dougherty, Combinatorics and finite geometry, Springer (2020).

[26] M. Dresher, The mathematics of games of strategy, Dover (1981).

[27] R. Durrett, Probability: theory and examples, Cambridge Univ. Press (1990).

[28] F. Dyson, Origins of life, Cambridge Univ. Press (1984).

[29] A. Einstein, Relativity: the special and the general theory, Dover (1916).

[30] L.C. Evans, Partial differential equations, AMS (1998).

[31] W. Feller, An introduction to probability theory and its applications, Wiley (1950).

[32] E. Fermi, Thermodynamics, Dover (1937).

[33] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics I: mainly mechanics, radiation and heat, Caltech (1963).

[34] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics II: mainly electromagnetism and matter, Caltech (1964).

[35] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics III: quantum mechanics, Caltech (1966).

[36] R.P. Feynman and A.R. Hibbs, Quantum mechanics and path integrals, Dover (1965).

[37] P. Flajolet and R. Sedgewick, Analytic combinatorics, Cambridge Univ. Press (2009).

[38] A.P. French, Special relativity, Taylor and Francis (1968).

[39] J.H. Gillespie, Population genetics, Johns Hopkins Univ. Press (1998).

[40] C. Godsil and G. Royle, Algebraic graph theory, Springer (2001).

[41] H. Goldstein, C. Safko and J. Poole, Classical mechanics, Addison-Wesley (1980).

[42] D.L. Goodstein, States of matter, Dover (1975).

[43] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).

[44] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).

[45] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).

[46] D.J. Griffiths, Revolutions in twentieth-century physics, Cambridge Univ. Press (2012).

[47] V.P. Gupta, Principles and applications of quantum chemistry, Elsevier (2016).

[48] W.A. Harrison, Solid state theory, Dover (1970).

[49] W.A. Harrison, Electronic structure and the properties of solids, Dover (1980).

[50] R.A. Horn and C.R. Johnson, Matrix analysis, Cambridge Univ. Press (1985).

[51] C.E. Housecroft and A.G. Sharpe, Inorganic chemistry, Pearson (2018).

[52] K. Huang, Introduction to statistical physics, CRC Press (2001).

[53] K. Huang, Fundamental forces of nature, World Scientific (2007).

[54] S. Huskey, The skeleton revealed, Johns Hopkins Univ. Press (2017).

[55] L. Hyman, Comparative vertebrate anatomy, Univ. of Chicago Press (1942).

[56] L.P. Kadanoff, Statistical physics: statics, dynamics and renormalization, World Scientific (2000).

[57] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).

[58] C. Kittel, Introduction to solid state physics, Wiley (1953).

[59] D.E. Knuth, The art of computer programming, Addison-Wesley (1968).

[60] M. Kumar, Quantum: Einstein, Bohr, and the great debate about the nature of reality, Norton (2009).

[61] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).

[62] L.D. Landau and E.M. Lifshitz, Mechanics, Pergamon Press (1960).

[63] L.D. Landau and E.M. Lifshitz, The classical theory of fields, Addison-Wesley (1951).

[64] L.D. Landau and E.M. Lifshitz, Quantum mechanics: non-relativistic theory, Pergamon Press (1959).

[65] S. Lang, Algebra, Addison-Wesley (1993).

[66] P. Lax, Linear algebra and its applications, Wiley (2007).

[67] P. Lax, Functional analysis, Wiley (2002).

[68] P. Lax and M.S. Terrell, Calculus with applications, Springer (2013).

[69] P. Lax and M.S. Terrell, Multivariable calculus with applications, Springer (2018).

[70] S. Ling and C. Xing, Coding theory: a first course, Cambridge Univ. Press (2004).

[71] J.P. Lowe and K. Peterson, Quantum chemistry, Elsevier (2005).

[72] S.J. Marshall, The story of the computer: a technical and business history, Create Space Publ. (2022).

[73] M.L. Mehta, Random matrices, Elsevier (2004).

[74] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge Univ. Press (2000).

[75] R.K. Pathria and and P.D. Beale, Statistical mechanics, Elsevier (1972).

[76] T.D. Pollard, W.C. Earnshaw, J. Lippincott-Schwartz and G. Johnson, Cell biology, Elsevier (2022).

[77] J. Preskill, Quantum information and computation, Caltech (1998).

[78] R. Rojas and U. Hashagen, The first computers: history and architectures, MIT Press (2000).

[79] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[80] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[81] W. Rudin, Functional analysis, McGraw-Hill (1973).

[82] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[83] B. Ryden and B.M. Peterson, Foundations of astrophysics, Cambridge Univ. Press (2010).

[84] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).

[85] R. Shankar, Fundamentals of physics I: mechanics, relativity, and thermodynamics, Yale Univ. Press (2014).

[86] R. Shankar, Fundamentals of physics II: electromagnetism, optics, and quantum mechanics, Yale Univ. Press (2016).

[87] N.J.A. Sloane and S. Plouffe, Encyclopedia of integer sequences, Academic Press (1995).

[88] A.M. Steane, Thermodynamics, Oxford Univ. Press (2016).

[89] S. Sternberg, Dynamical systems, Dover (2010).

[90] D.R. Stinson, Combinatorial designs: constructions and analysis, Springer (2006).

[91] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[92] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[93] J. von Neumann and O. Morgenstern, Theory of games and economic behavior, Princeton Univ. Press (1944).

[94] J. Watrous, The theory of quantum information, Cambridge Univ. Press (2018).

[95] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[96] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[97] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).

[98] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).

[99] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).

[100] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).

Index