Forces and vectors

Teo Banica

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

2010 *Mathematics Subject Classification.* 97M50

Key words and phrases. Force, Vector

Abstract. This is a joint introduction to basic mechanics and vector mathematics. We start with a discussion on what happens in 1 dimension, namely the basics of calculus, and some physics too, followed by a similar discussion in 2 dimensions, featuring this time plane vectors, complex numbers, and more advanced physics. Then we go on a lengthy discussion on what happens in 3 dimensions, namely standard vector calculus, standard classical mechanics, and the basics of electromagnetism too. Following Einstein, we discuss then 4 dimensions, and curved spacetime. Finally, we go beyond classical mechanics and small dimensions, with a brief introduction to quantum mechanics.

Preface

How many dimensions do we live in? Good question, and if you ask about this a kid, a good physics graduate, a bad physics graduate, Pac-Man, a Green little man from Mars, an artist, an old senator, an atomic bomb, or a black hole, the answer might differ.

For us, mathematicians, this question is of particular importance, because depending on the number of dimensions $N \in \mathbb{N}$, and I'm using here $\mathbb{N}$ as something approximate, we will have to adapt our mathematics, as for that to be something truly useful.

You would say, for most of basic math, $N = 1$, or perhaps $N = 2$, motivated by angles and trigonometry, which are certainly something very useful, will do. And this is indeed the case, with the bulk of basic math being indeed developed by using $N = 1, 2$. And with the story with $N = 1, 2$ being not over here, because, a bit as Pac-Man teaches us, there are many levels of knowledge here, and you can even spend your whole life, as a mathematican, working on $N = 1, 2$. In fact, for the story, and agreeing this time with old senator's answer, we even have colleagues in math spending their lives on $N = 0$.

But let us get now to what the kid says, $N = 3$. Very reasonable answer, and at the mathematical level, things start getting quite complicated here. A good knowledge of the vector formalism, which was not really needed at $N = 2$, say due to the complex numbers, which do the job at $N = 2$, while technically counting as $N = 1$, is now required.

Among others, vectors in 3 dimensions are subject to the vector product $x \times y$, best understood and computed by using some determinants and the right-hand rule. And with this, we are most likely into some complicated mixture of linear algebra, and physics.

Things however do not stop here, because when asking a physics graduate, the answer will most likely be a certain $N \in \{4, 5, \ldots, \infty\}$, which can vary with the graduate in question. To be more precise, $N = 4$ or higher is something well-established, coming from Einstein, who taught us that space $\mathbb{R}^3$ and time $\mathbb{R}$ are related, and so that we have to look at spacetime, which is a curved version of $\mathbb{R}^4$. Then, the quite plausible possibility of $N = \infty$, when looking at very small scales, came from quantum mechanics, as developed by Heisenberg and others. And finally, $4 < N < \infty$ is more complicated and modern physics business, typically with the bigger the $N < \infty$, the fancier the theory.

So, this was for the story of dimensionality of mathematics, in relation with physics, and although I never personally checked with the Green little men from Mars, I am pretty much sure that they use $N \in \mathbb{N} \cup \{\infty\}$ too. As for the artists, atomic bombs and black holes, their respective answers when asked were "love", "hey" and "yumm", most likely some advanced versions of our usual $N = \infty$ from mathematics.

We will discuss here, in this book, such things, with a joint introduction to basic mechanics, and physics and forces in general, and vector mathematics. We will discuss as well, at the end, some more tricky forces, and their mathematical modeling.

Many thanks to my cats, for precious help with the preparation of the present book, projecting from meow to various $N \leq \infty$ values being a quite easy task, for them.

Cergy, May 2025
Teo Banica

Contents

Part I

One dimension

So ya thought ya
Might like to go to the show
To feel the warm thrill of confusion
That space cadet glow

CHAPTER 1

Real numbers

1a. Numbers

We denote by $\mathbb{N}$ the set of positive integers, $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$, with $\mathbb{N}$ standing for "natural". Quite often in computations we will need negative numbers too, and we denote by $\mathbb{Z}$ the set of all integers, $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$, with $\mathbb{Z}$ standing from "zahlen", which is German for "numbers". Finally, there are many questions in mathematics involving fractions, or quotients, which are called rational numbers:

DEFINITION 1.1. *The rational numbers are the quotients of type*

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

with $a, b \in \mathbb{Z}$, and $b \neq 0$, identified according to the usual rule for quotients, namely:

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

We denote the set of rational numbers by $\mathbb{Q}$, standing for "quotients".

Observe that we have inclusions $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q}$. The integers add and multiply according to the rules that you know well. As for the rational numbers, these add according to the usual rule for quotients, which is as follows, and death penalty for forgetting it:

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

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

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

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

THEOREM 1.2. *The number of possibilities of choosing k objects among n objects is*

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

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

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

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

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

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

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

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

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

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

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

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

ADVICE 1.3. *Always doublecheck what you're doing, regularly, and definitely at the end, either with an alternative proof, or with some numerics.*

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

Which brings us into the question on whether mathematics is an exact science or not. Not clear. Chemistry for instance is an exact science, because findings of type "a mixture of water and salt cannot explode" look rock-solid. Same for biology, with findings of type "crocodiles eat fish" being rock-solid too. In what regards mathematics however, and theoretical physics too, things are always prone to human mistake.

And for ending this discussion, you might ask then, what about engineering? After all, this is mathematics and physics, which is usually 100% correct, because most of the bridges, buildings and other things built by engineers don't collapse. Well, this is because engineers follow, and in a truly maniac way, the above Advice 1.3. You won't declare a project for a bridge, building, engine and so on final and correct, ready for production, until you checked and doublechecked it with 10 different methods or so, won't you.

Back to work now, as an important adding to Theorem 1.2, we have:

CONVENTION 1.4. *By definition,* $0! = 1$.

This convention comes, and no surprise here, from Advice 1.3. Indeed, we obviously have $\binom{n}{n} = 1$, but if we want to recover this formula via Theorem 1.2 we are a bit in trouble, and so we must declare that $0! = 1$, as for the following computation to work:

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

Going ahead now with more mathematics and less philosophy, with Theorem 1.2 complemented by Convention 1.4 being in final form (trust me), we have:

THEOREM 1.5. *We have the binomial formula*

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

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

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

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

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

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

In order to finish, it remains to compute the coefficients C_k. But, according to our product formula, C_k is the number of choices for the k needed a variables among the n available a variables. Thus, according to Theorem 1.2, we have:

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

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

Theorem 1.5 is something quite interesting, so let us doublecheck it with some numerics. At small values of n we obtain the following formulae, which are all correct:

$$(a+b)^0 = 1$$

$$(a+b)^1 = a+b$$

$$(a+b)^2 = a^2+2ab+b^2$$

$$(a+b)^3 = a^3+3a^2b+3ab^2+b^3$$

$$(a+b)^4 = a^4+4a^3b+6a^2b^2+4ab^3+b^4$$

$$(a+b)^5 = a^5+5a^4b+10a^3b^2+10a^2b^3+5a^4b+b^5$$

$$\vdots$$

Now observe that in these formulae, say for memorization purposes, the powers of the a, b variables are something very simple, that can be recovered right away. What matters are the coefficients, which are the binomial coefficients $\binom{n}{k}$, which form a triangle. So, it is enough to memorize this triangle, and this can be done by using:

THEOREM 1.6. *The Pascal triangle, formed by the binomial coefficients* $\binom{n}{k}$*,*

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

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

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

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

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

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

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

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

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

By using the binomial formula, this formula becomes:

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

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

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

There are many more things that can be said about binomial coefficients, with all sorts of interesting formulae, but the idea is always the same, namely that in order to find such formulae you have a choice between algebra and combinatorics, and that when it comes to proofs, the brute-force computation method is useful too. In practice, the best is to master all 3 techniques. Among others, because of Advice 1.3. You will have in this way 3 different methods, for making sure that your formulae are correct indeed.

1b. Real numbers

All the above was very nice, but remember that we are here for doing science and physics, and more specifically for mathematically understanding the numeric variables $x, y, z, \ldots$ coming from real life. Such variables can be lengths, volumes, pressures and so on, which vary continuously with time, and common sense dictates that there is little to no chance for our variables to be rational, $x, y, z, \ldots \notin \mathbb{Q}$. In fact, we will even see soon a theorem, stating that the probability for such a variable to be rational is exactly 0. Or, to put it in a dramatic way, "rational numbers don't exist in real life".

You are certainly familiar with the real numbers, but let us review now their definition, which is something quite tricky. As a first goal, we would like to construct a number $x = \sqrt{2}$ having the property $x^2 = 2$. But how to do this? Let us start with:

PROPOSITION 1.7. *There is no number $r \in \mathbb{Q}_+$ satisfying $r^2 = 2$. In fact, we have*

$$\mathbb{Q}_+ = \left\{p \in \mathbb{Q}_+ \middle| p^2 < 2\right\} \bigsqcup \left\{q \in \mathbb{Q}_+ \middle| q^2 > 2\right\}$$

with this being a disjoint union.

PROOF. In what regards the first assertion, assuming that $r = a/b$ with $a, b \in \mathbb{N}$ prime to each other satisfies $r^2 = 2$, we have $a^2 = 2b^2$, so $a \in 2\mathbb{N}$. But by using again $a^2 = 2b^2$ we obtain $b \in 2\mathbb{N}$, contradiction. As for the second assertion, this is obvious. □

It looks like we are a bit stuck. We can't really tell who $\sqrt{2}$ is, and the only piece of information about $\sqrt{2}$ that we have comes from the knowledge of the rational numbers satisfying $p^2 < 2$ or $q^2 > 2$. To be more precise, the picture that emerges is:

CONCLUSION 1.8. *The number $\sqrt{2}$ is the abstract beast which is bigger than all rationals satisfying $p^2 < 2$, and smaller than all positive rationals satisfying $q^2 > 2$.*

This does not look very good, but you know what, instead of looking for more clever solutions to our problem, what about relaxing, or being lazy, or coward, or you name it, and taking Conclusion 1.8 as a definition for $\sqrt{2}$. This is actually something not that bad, and leads to the following "lazy" definition for the real numbers:

DEFINITION 1.9. *The real numbers $x \in \mathbb{R}$ are formal cuts in the set of rationals,*

$$\mathbb{Q} = \mathbb{Q}_{\leq x} \sqcup \mathbb{Q}_{>x}$$

with such a cut being by definition subject to the following condition:

$$p \in \mathbb{Q}_{\leq x} \ , \ q \in \mathbb{Q}_{>x} \implies p < q$$

These numbers add and multiply by adding and multiplying the corresponding cuts.

This might look quite original, but believe me, there is some genius behind this definition. As a first observation, we have an inclusion $\mathbb{Q} \subset \mathbb{R}$, obtained by identifying each rational number $r \in \mathbb{Q}$ with the obvious cut that it produces, namely:

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

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

$$\mathbb{Q}_{\leq x+y} = \mathbb{Q}_{\leq x} + \mathbb{Q}_{\leq y}$$

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

$$x \leq y \iff \mathbb{Q}_{\leq x} \subset \mathbb{Q}_{\leq y}$$

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

THEOREM 1.10. *The equation $x^2 = 2$ has two solutions over the real numbers, namely the positive solution, denoted $\sqrt{2}$, and its negative counterpart, which is $-\sqrt{2}$.*

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

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

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

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

THEOREM 1.11. *The solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{R}$ are*

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

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

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

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

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

Summarizing, we have a nice definition for the real numbers, that we can certainly do some math with. However, for anything more advanced we are in need of the decimal writing for the real numbers. The result here is as follows:

THEOREM 1.12. *The real numbers $x\in\mathbb{R}$ can be written in decimal form,*

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

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

PROOF. This is something quite non-trivial, assuming that you already have some familiarity with such things, for the rational numbers. The idea is as follows:

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\mathbb{Q}_{\leq \ldots b999\ldots} = \mathbb{Q}_{\leq \ldots (b+1)000\ldots}$$

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

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

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

The above theorem was of course quite difficult, but this is how things are. You might perhaps say why bothering with cuts, and not taking $x = \pm a_1 \ldots a_n.b_1b_2b_3\ldots\ldots$ as definition for the real numbers. Well, this is certainly possible, but when it comes to summing such numbers, or making products, or proving basic things such as the existence of $\sqrt{2}$, things become fairly complicated with the decimal writing picture. So, all the above is not as stupid as it seems. And we will come back anyway to all this later, with a 3rd picture for the real numbers, involving scary things like ε and δ, and it will be up to you to decide, at that time, which picture is the one that you prefer.

Moving on, we made the claim in the beginning of this chapter that "in real life, real numbers are never rational". Here is a theorem, justifying this claim:

THEOREM 1.13. *The probability for a real number $x \in \mathbb{R}$ to be rational is* 0.

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

(1) Before starting, let us point out the fact that probability theory is something quite tricky, with probability 0 not necessarily meaning that the event cannot happen, but rather meaning that "better not count on that". For instance according to my computations the probability of you winning 1 billion at the lottery is 0, but you are of course free to disagree, and prove me wrong, by playing every day at the lottery.

(2) With this discussion made, and extrapolating now from finance and lottery to our question regarding real numbers, your possible argument of type "yes, but if I pick $x \in \mathbb{R}$ to be $x = 3/2$, I have proof that the probability for $x \in \mathbb{Q}$ is nonzero" is therefore dismissed. Thus, our claim as stated makes sense, so let us try now to prove it.

(3) By translation, it is enough to prove that the probability for a real number $x \in [0, 1]$ to be rational is 0. For this purpose, let us write the rational numbers $r \in [0, 1]$ in the form of a sequence $r_1, r_2, r_3 \ldots$, with this being possible say by ordering our rationals $r = a/b$ according to the lexicographic order on the pairs (a, b):

$$\mathbb{Q} \cap [0, 1] = \left\{r_1, r_2, r_3, \ldots\right\}$$

Let us also pick a number $c > 0$. Since the probability of having $x = r_1$ is certainly smaller than $c/2$, then the probability of having $x = r_2$ is certainly smaller than $c/4$, then the probability of having $x = r_3$ is certainly smaller than $c/8$ and so on, the probability for x to be rational satisfies the following inequality:

$$\begin{aligned} P &\leq \frac{c}{2} + \frac{c}{4} + \frac{c}{8} + \ldots \\ &= c\left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots\right) \\ &= c \end{aligned}$$

Here we have used the well-known formula $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = 1$, which comes by dividing $[0, 1]$ into half, and then one of the halves into half again, and so on, and then saying in the end that the pieces that we have must sum up to 1. Thus, we have indeed $P \leq c$, and since the number $c > 0$ was arbitrary, we obtain $P = 0$, as desired. □

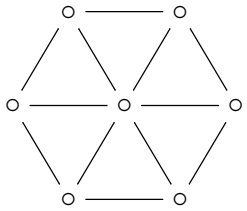
As a comment here, all the above is of course quite tricky, and a bit bordeline in respect to what can be called "rigorous mathematics". But we will be back to this, namely general probability theory, and in particular meaning of the mysterious formula $P = 0$, countable sets, infinite sums and so on, on several occasions, throughout this book.

Moving ahead now, let us construct now some more real numbers. We already know about $\sqrt{2}$ and other numbers of the same type, namely roots of polynomials, and our knowledge here being quite decent, no hurry with this, we will be back to it later. So, let us get now into π and trigonometry. To start with, we have the following result:

THEOREM 1.14. *The following two definitions of π are equivalent:*

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

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



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

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

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

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

Let us end this discussion about real numbers with some trigonometry. There are many things that can be said, that you certainly know, the basics being as follows:

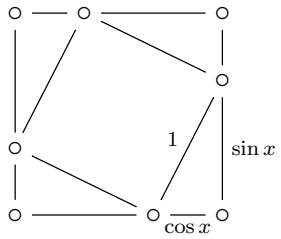
THEOREM 1.15. *The following happen:*

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

PROOF. There are certainly things that you know, the idea being as follows:

(1) The formula $L = 2\pi$ from Theorem 1.14 shows that the length of a quarter of the unit circle is $l = \pi/2$, and so the right angle has indeed this value, $\pi/2$.

(2) As for $\sin^2 x+\cos^2 x = 1$, called Pythagoras' theorem, this comes from the following picture, consisting of two squares and four identical triangles, as indicated:



Indeed, when computing the area of the outer square, we obtain:

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

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

It is possible to say many more things about angles and $\sin x$, $\cos x$, and also talk about some supplementary quantities, such as $\tan x = \sin x/\cos x$. But more on this later, once we will have some appropriate tools, beyond basic geometry, in order to discuss this.

1c. Convergence

We already met, on several occasions, infinite sequences or sums, and their limits. Time now to clarify all this. Let us start with the following definition:

DEFINITION 1.16. *We say that a sequence $\{x_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ converges to $x \in \mathbb{R}$ when:*

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N, |x_n - x| < \varepsilon$$

In this case, we write $\lim_{n\to\infty} x_n = x$, or simply $x_n \to x$.

This might look quite scary, at a first glance, but when thinking a bit, there is nothing scary about it. Indeed, let us try to understand, how shall we translate $x_n \to x$ into mathematical language. The condition $x_n \to x$ tells us that "when n is big, x_n is close to x", and to be more precise, it tells us that "when n is big enough, x_n gets arbitrarily close to x". But n big enough means $n \geq N$, for some $N \in \mathbb{N}$, and x_n arbitrarily close to x means $|x_n - x| < \varepsilon$, for some $\varepsilon > 0$. Thus, we are led to the above definition.

As a basic example for all this, we have:

PROPOSITION 1.17. *We have $1/n \to 0$.*

PROOF. This is obvious, but let us prove it by using Definition 1.16. We have:

$$\left|\frac{1}{n} - 0\right| < \varepsilon \iff \frac{1}{n} < \varepsilon \iff \frac{1}{\varepsilon} < n$$

Thus we can take $N = [1/\varepsilon] + 1$ in Definition 1.16, and we are done. □

There are many other examples, and more on this in a moment. Going ahead with more theory, let us complement Definition 1.16 with:

DEFINITION 1.18. *We write $x_n \to \infty$ when the following condition is satisfied:*

$$\forall K > 0, \exists N \in \mathbb{N}, \forall n \geq N, x_n > K$$

Similarly, we write $x_n \to -\infty$ when the same happens, with $x_n < -K$ at the end.

Again, this is something very intuitive, coming from the fact that $x_n \to \infty$ can only mean that x_n is arbitrarily big, for n big enough. As a basic illustration, we have:

PROPOSITION 1.19. *We have $n^2 \to \infty$.*

PROOF. As before, this is obvious, but let us prove it using Definition 1.18. We have:

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

Thus we can take $N = [\sqrt{K}] + 1$ in Definition 1.18, and we are done. □

We can unify and generalize Proposition 1.17 and Proposition 1.19, as follows:

PROPOSITION 1.20. *We have the following convergence, with $n \to \infty$:*

$$n^a \to \begin{cases} 0 & (a < 0) \\ 1 & (a = 0) \\ \infty & (a > 0) \end{cases}$$

PROOF. This follows indeed by using the same method as in the proof of Proposition 1.17 and Proposition 1.19, first for a rational, and then for a real as well. □

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

THEOREM 1.21. *The following happen:*

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

PROOF. All this is elementary, coming from definitions:

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

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

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

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

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

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

(4) This is clear from definitions. □

Here are as well some general rules for computing limits:

THEOREM 1.22. *The following happen, with the conventions $\infty+\infty = \infty$, $\infty\cdot\infty = \infty$, $1/\infty = 0$, and with the conventions that $\infty - \infty$ and $\infty \cdot 0$ are undefined:*

(1) $x_n \to x$ *implies* $\lambda x_n \to \lambda x$.
(2) $x_n \to x$, $y_n \to y$ *implies* $x_n + y_n \to x + y$.
(3) $x_n \to x$, $y_n \to y$ *implies* $x_n y_n \to xy$.
(4) $x_n \to x$ *with* $x \neq 0$ *implies* $1/x_n \to 1/x$.

PROOF. All this is again elementary, coming from definitions:

(1) This is something which is obvious from definitions.

(2) This follows indeed from the following estimate:

$$|x_n + y_n - x - y| \leq |x_n - x| + |y_n - y|$$

(3) This follows indeed from the following estimate:

$$\begin{aligned} |x_n y_n - xy| &= |(x_n - x)y_n + x(y_n - y)| \\ &\leq |x_n - x| \cdot |y_n| + |x| \cdot |y_n - y| \end{aligned}$$

(4) This is again clear, by estimating $1/x_n - 1/x$, in the obvious way. □

As an application of the above rules, we have the following useful result:

PROPOSITION 1.23. *The $n \to \infty$ limits of quotients of polynomials are given by*

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

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

PROOF. The first assertion comes from the following computation:

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

As for the second assertion, this comes from Proposition 1.20. □

Getting back now to theory, some sequences which obviously do not converge, like for instance $x_n = (-1)^n$, have however "2 limits instead of 1". So let us formulate:

DEFINITION 1.24. *Given a sequence $\{x_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$, we let*

$$\liminf_{n\to\infty} x_n \in [-\infty, \infty] \quad , \quad \limsup_{n\to\infty} x_n \in [-\infty, \infty]$$

to be the smallest and biggest limit of a subsequence of (x_n).

Observe that the above quantities are defined indeed for any sequence x_n. For instance, for $x_n = (-1)^n$ we obtain -1 and 1. Also, for $x_n = n$ we obtain ∞ and ∞. And so on. Of course, and generalizing the $x_n = n$ example, if $x_n \to x$ we obtain x and x.

Going ahead with more theory, here is a key result:

THEOREM 1.25. *A sequence x_n converges, with finite limit $x \in \mathbb{R}$, precisely when*

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall m, n \geq N, |x_m - x_n| < \varepsilon$$

called Cauchy condition.

PROOF. In one sense, this is clear. In the other sense, we can say for instance that the Cauchy condition forces the decimal writings of our numbers x_n to coincide more and more, with $n \to \infty$, and so we can construct a limit $x = \lim_{n\to\infty} x_n$, as desired. □

The above result is quite interesting, and as an application, we have:

THEOREM 1.26. *$\mathbb{R}$ is the completion of $\mathbb{Q}$, in the sense that it is the space of Cauchy sequences over $\mathbb{Q}$, identified when the virtual limit is the same, in the sense that:*

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

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

PROOF. Let us denote the completion operation by $X \to \bar{X} = C_X/\sim$, where C_X is the space of Cauchy sequences over X, and $\sim$ is the above equivalence relation. Since by Theorem 1.25 any Cauchy sequence $(x_n) \in C_{\mathbb{Q}}$ has a limit $x \in \mathbb{R}$, we obtain $\bar{\mathbb{Q}} = \mathbb{R}$. As for the equality $\bar{\mathbb{R}} = \mathbb{R}$, this is clear again by using Theorem 1.25. □

1d. Sums, series

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

DEFINITION 1.27. *Given numbers $x_0, x_1, x_2, \ldots \in \mathbb{R}$, we write*

$$\sum_{n=0}^{\infty} x_n = x$$

with $x \in [-\infty, \infty]$ when $\lim_{k\to\infty} \sum_{n=0}^{k} x_n = x$.

As before with the sequences, there is some general theory that can be developed for the series, and more on this in a moment. As a first, basic example, we have:

THEOREM 1.28. *We have the "geometric series" formula*

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

valid for any $|x| < 1$. For $|x| \geq 1$, the series diverges.

PROOF. Our first claim, which comes by multiplying and simplifying, is that:

$$\sum_{n=0}^{k} x^n = \frac{1-x^{k+1}}{1-x}$$

But this proves the first assertion, because with $k \to \infty$ we get:

$$\sum_{n=0}^{k} x^n \to \frac{1}{1-x}$$

As for the second assertion, this is clear as well from our formula above. □

Less trivial now is the following result, due to Riemann:

THEOREM 1.29. *We have the following formula:*

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots = \infty$$

In fact, $\sum_n 1/n^a$ converges for $a > 1$, and diverges for $a \leq 1$.

PROOF. We have to prove several things, the idea being as follows:

(1) The first assertion comes from the following computation:

$$\begin{aligned}
1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots &= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) + \ldots \\
&\geq 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) + \ldots \\
&= 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \ldots \\
&= \infty
\end{aligned}$$

(2) Regarding now the second assertion, we have that at $a = 1$, and so at any $a \leq 1$. Thus, it remains to prove that at $a > 1$ the series converges. Let us first discuss the case

$a=2$, which will prove the convergence at any $a\geq 2$. The trick here is as follows:

$$\begin{aligned}1+\frac{1}{4}+\frac{1}{9}+\frac{1}{16}+\ldots &\leq 1+\frac{1}{3}+\frac{1}{6}+\frac{1}{10}+\ldots\\ &= 2\left(\frac{1}{2}+\frac{1}{6}+\frac{1}{12}+\frac{1}{20}+\ldots\right)\\ &= 2\left[\left(1-\frac{1}{2}\right)+\left(\frac{1}{2}-\frac{1}{3}\right)+\left(\frac{1}{3}-\frac{1}{4}\right)+\left(\frac{1}{4}-\frac{1}{5}\right)\ldots\right]\\ &= 2\end{aligned}$$

(3) It remains to prove that the series converges at $a\in(1,2)$, and here it is enough to deal with the case of the exponents $a=1+1/p$ with $p\in\mathbb{N}$. We already know how to do this at $p=1$, and the proof at $p\in\mathbb{N}$ will be based on a similar trick. We have:

$$\sum_{n=0}^{\infty}\frac{1}{n^{1/p}}-\frac{1}{(n+1)^{1/p}}=1$$

Let us compute, or rather estimate, the generic term of this series. By using the formula $a^p-b^p=(a-b)(a^{p-1}+a^{p-2}b+\ldots+ab^{p-2}+b^{p-1})$, we have:

$$\begin{aligned}\frac{1}{n^{1/p}}-\frac{1}{(n+1)^{1/p}} &= \frac{(n+1)^{1/p}-n^{1/p}}{n^{1/p}(n+1)^{1/p}}\\ &= \frac{1}{n^{1/p}(n+1)^{1/p}[(n+1)^{1-1/p}+\ldots+n^{1-1/p}]}\\ &\geq \frac{1}{n^{1/p}(n+1)^{1/p}\cdot p(n+1)^{1-1/p}}\\ &= \frac{1}{pn^{1/p}(n+1)}\\ &\geq \frac{1}{p(n+1)^{1+1/p}}\end{aligned}$$

We therefore obtain the following estimate for the Riemann sum:

$$\begin{aligned}\sum_{n=0}^{\infty}\frac{1}{n^{1+1/p}} &= 1+\sum_{n=0}^{\infty}\frac{1}{(n+1)^{1+1/p}}\\ &\leq 1+p\sum_{n=0}^{\infty}\left(\frac{1}{n^{1/p}}-\frac{1}{(n+1)^{1/p}}\right)\\ &= 1+p\end{aligned}$$

Thus, we are done with the case $a=1+1/p$, which finishes the proof. □

Here is another tricky result, this time about alternating sums:

THEOREM 1.30. *We have the following convergence result:*

$$1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\ldots<\infty$$

However, when rearranging terms, we can obtain any $x\in[-\infty,\infty]$ as limit.

PROOF. Both the assertions follow from Theorem 1.29, as follows:

(1) We have the following computation, using the Riemann criterion at $a=2$:

$$\begin{aligned}1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\ldots&=\left(1-\frac{1}{2}\right)+\left(\frac{1}{3}-\frac{1}{4}\right)+\ldots\\&=\frac{1}{2}+\frac{1}{12}+\frac{1}{30}+\ldots\\&<\frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}+\ldots\\&<\infty\end{aligned}$$

(2) We have the following formulae, coming from the Riemann criterion at $a=1$:

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{6}+\frac{1}{8}+\ldots=\frac{1}{2}\left(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots\right)=\infty$$

$$1+\frac{1}{3}+\frac{1}{5}+\frac{1}{7}+\ldots\geq\frac{1}{2}+\frac{1}{4}+\frac{1}{6}+\frac{1}{8}+\ldots=\infty$$

Thus, both these series diverge. The point now is that, by using this, when rearranging terms in the alternating series in the statement, we can arrange for the partial sums to go arbitrarily high, or arbitrarily low, and we can obtain any $x\in[-\infty,\infty]$ as limit. □

Back now to the general case, we first have the following statement:

THEOREM 1.31. *The following hold, with the converses of* (1) *and* (2) *being wrong, and with* (3) *not holding when the assumption $x_n\geq 0$ is removed:*

(1) *If $\sum_n x_n$ converges then $x_n\to 0$.*
(2) *If $\sum_n|x_n|$ converges then $\sum_n x_n$ converges.*
(3) *If $\sum_n x_n$ converges, $x_n\geq 0$ and $x_n/y_n\to 1$ then $\sum_n y_n$ converges.*

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

(1) We know that $\sum_n x_n$ converges when $S_k=\sum_{n=0}^k x_n$ converges. Thus by Cauchy we have $x_k=S_k-S_{k-1}\to 0$, and this gives the result. As for the simplest counterexample for the converse, this is $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots=\infty$, coming from Theorem 1.29.

(2) This follows again from the Cauchy criterion, by using:

$$|x_n+x_{n+1}+\ldots+x_{n+k}|\leq|x_n|+|x_{n+1}|+\ldots+|x_{n+k}|$$

As for the simplest counterexample for the converse, this is $1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\ldots<\infty$, coming from Theorem 1.30, coupled with $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots=\infty$ from (1).

(3) Again, the main assertion here is clear, coming from, for n big:

$$(1-\varepsilon)x_n \leq y_n \leq (1+\varepsilon)x_n$$

In what regards now the failure of the result, when the assumption $x_n \geq 0$ is removed, this is something quite tricky, the simplest counterexample being as follows:

$$x_n = \frac{(-1)^n}{\sqrt{n}} \quad , \quad y_n = \frac{1}{n} + \frac{(-1)^n}{\sqrt{n}}$$

To be more precise, we have $y_n/x_n \to 1$, so $x_n/y_n \to 1$ too, but according to the above-mentioned results from (1,2), modified a bit, $\sum_n x_n$ converges, while $\sum_n y_n$ diverges. □

Summarizing, we have some useful positive results about series, which are however quite trivial, along with various counterexamples to their possible modifications, which are non-trivial. Staying positive, here are some more positive results:

THEOREM 1.32. *The following happen, and in all cases, the situtation where $c = 1$ is indeterminate, in the sense that the series can converge or diverge:*

(1) *If $|x_{n+1}/x_n| \to c$, the series $\sum_n x_n$ converges if $c < 1$, and diverges if $c > 1$.*
(2) *If $\sqrt[n]{|x_n|} \to c$, the series $\sum_n x_n$ converges if $c < 1$, and diverges if $c > 1$.*
(3) *With $c = \limsup_{n\to\infty} \sqrt[n]{|x_n|}$, $\sum_n x_n$ converges if $c < 1$, and diverges if $c > 1$.*

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

(1) Here the main assertions, regarding the cases $c < 1$ and $c > 1$, are both clear by comparing with the geometric series $\sum_n c^n$. As for the case $c = 1$, this is what happens for the Riemann series $\sum_n 1/n^a$, so we can have both convergent and divergent series.

(2) Again, the main assertions, where $c < 1$ or $c > 1$, are clear by comparing with the geometric series $\sum_n c^n$, and the $c = 1$ examples come from the Riemann series.

(3) Here the case $c < 1$ is dealt with as in (2), and the same goes for the examples at $c = 1$. As for the case $c > 1$, this is clear too, because here $x_n \to 0$ fails. □

Finally, generalizing the first assertion in Theorem 1.30, we have:

THEOREM 1.33. *If $x_n \searrow 0$ then $\sum_n (-1)^n x_n$ converges.*

PROOF. We have the $\sum_n (-1)^n x_n = \sum_k y_k$, where:

$$y_k = x_{2k} - x_{2k+1}$$

But, by drawing for instance the numbers x_i on the real line, we see that y_k are positive numbers, and that $\sum_k y_k$ is the sum of lengths of certain disjoint intervals, included in the interval $[0, x_0]$. Thus we have $\sum_k y_k \leq x_0$, and this gives the result. □

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

THEOREM 1.34. *We have the following convergence*

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

where $e=2.71828\ldots$ *is a certain number.*

PROOF. This is something quite tricky, as follows:

(1) Our first claim is that the following sequence is increasing:

$$x_n=\left(1+\frac{1}{n}\right)^n$$

In order to prove this, we use the following arithmetic-geometric inequality:

$$\frac{1+\sum_{i=1}^n\left(1+\frac{1}{n}\right)}{n+1}\geq\sqrt[n+1]{1\cdot\prod_{i=1}^n\left(1+\frac{1}{n}\right)}$$

In practice, this gives the following inequality:

$$1+\frac{1}{n+1}\geq\left(1+\frac{1}{n}\right)^{n/(n+1)}$$

Now by raising to the power $n+1$ we obtain, as desired:

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

(2) Normally we are left with proving that x_n is bounded from above, but this is non-trivial, and we have to use a trick. Consider the following sequence:

$$y_n=\left(1+\frac{1}{n}\right)^{n+1}$$

We will prove that this sequence y_n is decreasing, and together with the fact that we have $x_n/y_n\to 1$, this will give the result. So, this will be our plan.

(3) In order to prove now that y_n is decreasing, we use, a bit as before:

$$\frac{1+\sum_{i=1}^n\left(1-\frac{1}{n}\right)}{n+1}\geq\sqrt[n+1]{1\cdot\prod_{i=1}^n\left(1-\frac{1}{n}\right)}$$

In practice, this gives the following inequality:

$$1-\frac{1}{n+1}\geq\left(1-\frac{1}{n}\right)^{n/(n+1)}$$

Now by raising to the power $n+1$ we obtain from this:

$$\left(1-\frac{1}{n+1}\right)^{n+1}\geq\left(1-\frac{1}{n}\right)^{n}$$

The point now is that we have the following inversion formulae:

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

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

Thus by inverting the inequality that we found, we obtain, as desired:

$$\left(1+\frac{1}{n}\right)^{n+1}\leq\left(1+\frac{1}{n-1}\right)^{n}$$

(4) But with this, we can now finish. Indeed, the sequence x_n is increasing, the sequence y_n is decreasing, and we have $x_n<y_n$, as well as:

$$\frac{y_n}{x_n}=1+\frac{1}{n}\to 1$$

Thus, both sequences x_n,y_n converge to a certain number e, as desired.

(5) Finally, regarding the numerics for our limiting number e, we know from the above that we have $x_n<e<y_n$ for any $n\in\mathbb{N}$, which reads:

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

Thus $e\in[2,3]$, and with a bit of patience, or a computer, we obtain $e=2.71828\ldots$ We will actually come back to this question later, with better methods. □

We should mention that there are many other ways of getting into e. For instance it is possible to prove that we have the following formula, which is a bit more conceptual than the formula in Theorem 1.34, and also with the convergence being very quick:

$$\sum_{n=0}^{\infty}\frac{1}{n!}=e$$

Importantly, all this not the end of the story with e. For instance, in relation with the first formula that we found, from Theorem 1.34, we have, more generally:

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

Also, in relation with the second formula, from above, we have, more generally:

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

To be more precise, these latter two formulae are something that we know at $x = 1$. The case $x = 0$ is trivial, the case $x = -1$ follows from the case $x = 1$, via some simple manipulations, and with a bit more work, we can get these formulae for any $x \in \mathbb{N}$, and then for any $x \in \mathbb{Z}$. However, the general case $x \in \mathbb{R}$ is quite tricky, requiring a good knowledge of the theory of real functions. We will be back to this later.

1e. Exercises

Exercises:

Exercise 1.35.

Exercise 1.36.

Exercise 1.37.

Exercise 1.38.

Exercise 1.39.

Exercise 1.40.

Exercise 1.41.

Exercise 1.42.

Bonus exercise.

CHAPTER 2

Motion basics

2a. Collisions

Good news, with what that we know we can do some physics, in 1 dimension. Let us begin with a discussion of the usual motion, in the absence of forces. That is something very simple, with the motion being linear, the bodies traveling at constant speed, namely their initial speed. And there is no acceleration to worry about.

However, interesting things happen when such objects collide, and we have:

FACT 2.1. *In the context of general linear motion, in the case of a collision between two bodies, m_1, m_2 travelling at speeds v_1, v_2, the total momentum of the system*

$$p = m_1 v_1 + m_2 v_2$$

is conserved. The same happens of course without collision either, and also for systems of N bodies, with $N \in \mathbb{N}$ arbitrary, with all sorts of collisions allowed between them.

As a first comment, is this really physics, or just some abstraction? We know that gravity is everywhere, and that the very existence of m_1, m_2 leads to their gravity, and so to the negation of the general linear motion setting above. However, two trains colliding is certainly physics, and even scary physics, and this has nothing to do with gravity. Thus, what we have here is a true physics principle, dealing with real-life situations.

In order to understand now what is going on, consider two objects as in Fact 2.1, bound for collision:

$$\circ_{m_1} \to_{v_1} \qquad \leftarrow_{v_2} \circ_{m_2}$$

We know from real life that two things can happen, in this situation. The first case is that of an inelastic, also called plastic collision, where m_1, m_2 decide when meeting that they love each other, and pursue their journey as a couple, $m = m_1 + m_2$:

$$\bullet_m \to_v$$

Of course, who really knows what really happens during a plastic collision, at the microscopic level, but assuming somehow that no energy or something is dissipated, during that hot encounter, Fact 2.1 holds indeed, and allows us to do the math.

And the math, coming from the conservation of momentum, is as follows:

PROPOSITION 2.2. *In the context of a plastic collision between two bodies,*

$$m = m_1 + m_2 \quad , \quad v = \frac{m_1 v_1 + m_2 v_2}{m_1 + m_2}$$

are the mass and speed of the resulting body.

PROOF. This follows straight from Fact 2.1, because the momentum of $m = m_1 + m_2$ equals the sum of the initial momenta of m_1, m_2, and is therefore given by:

$$mv = m_1 v_1 + m_2 v_2$$

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

The second case now, that can happen as well, is that of an elastic collision. So, consider as before two objects bound for collision:

$$\circ_{m_1} \to_{v_1} \qquad \leftarrow_{v_2} \circ_{m_2}$$

The elastic collision is then the case opposed to love, with our two bodies meeting, comparing their m_i, v_i, then exchanging some speed depending on that, via a few quick fists, and then either keeping traveling forward, but slower, or going backwards:

$$\bullet_{m_1} \to_{v_1'} \qquad \circ_{m_2} \to_{v_2'}$$

$$\leftarrow_{v_1'} \circ_{m_1} \qquad \leftarrow_{v_2'} \bullet_{m_2}$$

$$\leftarrow_{v_1'} \circ_{m_1} \qquad \circ_{m_2} \to_{v_2'}$$

In the above pictures, the winner, which was m_1 in the first case, and m_2 in the second case, was awarded a black belt. As for the third case, that is some sort of draw.

Getting back now to the conservation of momentum, from Fact 2.1, it is pretty much clear that what we have there won't allow us to do the math. To be more precise, we can get from there only 1 equation, which is not enough for computing the output data. Fortunately, in the case of elastic collisions, Fact 2.1 can be complemented with:

FACT 2.3. *In the context of general linear motion, in the case of an elastic collision between two bodies, m_1, m_2 traveling at speeds v_1, v_2, the total energy of the system*

$$E = \frac{m_1 v_1^2}{2} + \frac{m_2 v_2^2}{2}$$

is conserved. The same happens of course without collision either, and also for systems of N bodies, with $N \in \mathbb{N}$ arbitrary, with multi-elastic collisions allowed between them.

Again, as in the case of the plastic collisions, who really knows what really happens during an elastic collision, at the microscopic level, but again, assuming that no things are lost, during that event, Fact 2.3 holds indeed, and allows us to do the math.

As another comment, while the formula of the momentum $p = mv$ from Fact 2.3 was something quite simple and intuitive, the above formula of the energy $E = mv^2/2$ is obviously something more subtle. We will be back to this, later.

Going ahead now, let us first investigate, just out of curiosity, what happens to the energy during a plastic collision. The result here, contradicting our previous guess that the moment conservation comes somehow from "no energy lost", is as follows:

THEOREM 2.4. *In the context of a plastic collision between two bodies, we have:*

$$E < E_1 + E_2$$

That is, some of the initial energy gets dissipated during the collision.

PROOF. We use the equations found in Proposition 2.2, namely:

$$m = m_1 + m_2 \quad , \quad v = \frac{m_1 v_1 + m_2 v_2}{m_1 + m_2}$$

According to our definition of energy, from Fact 2.3, the initial energy is:

$$E_1 + E_2 = \frac{m_1 v_1^2 + m_2 v_2^2}{2}$$

As for the final energy, this is given by the following formula:

$$E = \frac{mv^2}{2} = \frac{(m_1 v_1 + m_2 v_2)^2}{2(m_1 + m_2)}$$

So, let us compute now the difference between these two quantities. We obtain:

$$\begin{aligned} E_1 + E_2 - E &= \frac{(m_1 + m_2)(m_1 v_1^2 + m_2 v_2^2) - (m_1 v_1 + m_2 v_2)^2}{2(m_1 + m_2)} \\ &= \frac{m_1 m_2 (v_1^2 + v_2^2 - 2v_1 v_2)}{2(m_1 + m_2)} \\ &= \frac{m_1 m_2 (v_1 - v_2)^2}{2(m_1 + m_2)} \\ &\geq 0 \end{aligned}$$

Thus $E_1 + E_2 \geq E$, and since a collision cannot happen when the inital speeds are the same, $v_1 = v_2$, the equality case cannot happen, and so $E_1 + E_2 > E$, as stated. □

Moving ahead now, and back to the elastic collisions, the two conservation principles that we have, namely Fact 2.1 and Fact 2.3, allow us to do the math. In order to discuss this, consider our usual picture of an elastic collision, as follows:

$$\circ_{m_1} \to_{v_1} \qquad \leftarrow_{v_2} \circ_{m_2}$$

Depending on the resulting fight, we can have either a win or m_1 or m_2, or a draw. Abstractly however, we can simply say that we are in a draw situation, the picture being as follows, with the convention that we do not know yet the directions of v_1', v_2':

$$\leftarrow_{v_1'} \circ_{m_1} \qquad \circ_{m_2} \rightarrow_{v_2'}$$

With these conventions made, the precise result is as follows:

PROPOSITION 2.5. *In the context of an elastic collision between two bodies,*

$$v_1' = \frac{(m_1 - m_2)v_1 + 2m_2v_2}{m_1 + m_2}$$

$$v_2' = \frac{(m_2 - m_1)v_2 + 2m_1v_1}{m_1 + m_2}$$

are the resulting speeds of the two bodies.

PROOF. According to our momentum and energy conservation principles from Fact 2.1 and Fact 2.3, the resulting speeds v_1', v_2' satisfy the following two equations:

$$m_1v_1 + m_2v_2 = m_1v_1' + m_2v_2'$$

$$m_1v_1^2 + m_2v_2^2 = m_1v_1'^2 + m_2v_2'^2$$

Now observe that these equations can be written as follows:

$$m_1(v_1 - v_1') = m_2(v_2' - v_2)$$

$$m_1(v_1^2 - v_1'^2) = m_2(v_2'^2 - v_2^2)$$

By dividing the second equation by the first one, our system becomes:

$$m_1(v_1 - v_1') = m_2(v_2' - v_2)$$

$$v_1 + v_1' = v_2' + v_2$$

And by doing now the math, we are led to the formulae in the statement. □

We have in fact a better formulation of Proposition 2.5, as follows:

THEOREM 2.6. *In the context of an elastic collision between two bodies, the resulting speeds of the two bodies are*

$$v_1' = v_1 + \frac{q}{m_1} \quad , \quad v_2' = v_2 - \frac{q}{m_2}$$

where $q \in \mathbb{R}$ is the individual change of momentum, given by

$$\left(\frac{1}{m_1} + \frac{1}{m_2}\right) q = 2(v_2 - v_1)$$

from the perspective of m_1, and from the opposite perspective of m_2.

PROOF. From the perspective of Proposition 2.5, we have done some quick algebra there, without really knowing what we're doing, leading to the following formulae:

$$v_1' = \frac{(m_1 - m_2)v_1 + 2m_2v_2}{m_1 + m_2}$$

$$v_2' = \frac{(m_2 - m_1)v_2 + 2m_1v_1}{m_1 + m_2}$$

Now observe that these two formulae can be alternatively written as follows:

$$v_1' = v_1 + \frac{2m_2(v_2 - v_1)}{m_1 + m_2}$$

$$v_2' = v_2 + \frac{2m_1(v_1 - v_2)}{m_1 + m_2}$$

But this leads to the formulae in the statement, and to that conclusion about q. □

We will be back to collisions in chapter 4, when talking about basic thermodynamics, in 1 dimension, and then later on several occasions, in 2 or 3 dimensions.

2b. Rockets

As a main application now of the general theory developed above, and in relation with gravity as well, we can use momentum for beating gravity, as follows:

THEOREM 2.7. *We can build rockets, by ejecting mass from a body*

$$\cdots\cdots \quad \bullet_M \rightarrow$$

with the body moving in the opposite direction to the ejection direction.

PROOF. The functioning principle of rockets is clear indeed from the conservation of the momentum principle, because ejecting mass to the left will move us to the right. As for the precise math of this, this can be worked out too, the idea being as follows:

(1) Let us first study the case of a single ejection. We begin with M at rest:

$$\bullet_M$$

Now let us eject to the left a mass m, with speed s. The situation becomes:

$$\leftarrow_s \circ_m \qquad\qquad \bullet_{M-m} \rightarrow_v$$

By conservation of momentum we have $(M - m)v = ms$, and so:

$$v = \frac{ms}{M - m}$$

(2) Let us study now a double ejection. At the first stage, we have as above, by labelling now the ejection data with a 1 index, standing for stage 1 of the ejection:

$$\leftarrow_{s_1} \circ_{m_1} \qquad\qquad \bullet_{M-m_1} \rightarrow_{v_1}$$

At the second stage now, that of ejecting a mass m_2, with speed s_2, with the observation that the ejection speed s_2 is only relative to M, the situation becomes:

$$\leftarrow_{s_1} \circ_{m_1} \quad \leftarrow_{s_2-v_1} \circ_{m_2} \qquad\qquad \bullet_{M-m_1-m_2} \rightarrow_{v_2}$$

Neglecting the first ejection, the conservation of momentum tells us that:

$$(M-m_1-m_2)v_2-m_2(s_2-v_1)=(M-m_1)v_1$$

But this equation can be written in the following way:

$$(M-m_1-m_2)v_2=(M-m_1-m_2)v_1+m_2s_2$$

By using now (1) for the value of v_1, the speed after the second ejection is given by:

$$\begin{aligned} v_2 &= v_1+\frac{m_2s_2}{M-m_1-m_2}\\ &= \frac{m_1s_1}{M-m_1}+\frac{m_2s_2}{M-m_1-m_2}\end{aligned}$$

(3) In the general case now, that of a multiple ejection, of masses $m_1,\ldots,m_k$ with respective speeds $s_1,\ldots,s_k$, the same idea applies, and gives as eventual speed:

$$v_k=\frac{m_1s_1}{M-m_1}+\frac{m_2s_2}{M-m_1-m_2}+\ldots+\frac{m_ks_k}{M-m_1-\ldots-m_k}$$

In the particular case where the ejection mass m is constant, we obtain:

$$v_k=\frac{ms_1}{M-m}+\frac{ms_2}{M-2m}+\ldots+\frac{ms_k}{M-km}$$

Also, in the particular case where the ejection speed s is constant, we obtain:

$$v_k=\left(\frac{m_1}{M-m_1}+\frac{m_2}{M-m_1-m_2}+\ldots+\frac{m_k}{M-m_1-\ldots-m_k}\right)s$$

And in the case where both the mass m and speed s are constant, we obtain:

$$v_k=\left(\frac{m}{M-m}+\frac{m}{M-2m}+\ldots+\frac{m}{M-km}\right)s$$

(4) Let us work out now the asymptotics. For simplifying we will assume that we are in the last case, that of a constant ejection mass m and speed s, although modifications of our argument will apply as well more generally. With $m=\varepsilon M$, we have:

$$v_k=\left(\frac{\varepsilon}{1-\varepsilon}+\frac{\varepsilon}{1-2\varepsilon}+\ldots+\frac{\varepsilon}{1-k\varepsilon}\right)s$$

We will assume that ε is small, and that $k\in\mathbb{N}$ is such that the total ejection mass, or rather the fraction $k\varepsilon=r\in(0,1)$ of this total ejection mass, compared to the initial mass M of our rocket, is fixed. Thus, we want to compute v_k in the following regime:

$$\varepsilon=\frac{r}{k}\quad,\quad k\to\infty$$

Now remember the definition of the integral, as the area below the graph of the function, which is approximable by the Riemann method by usual rectangles. In the particular case of the function $1/x$, this picture gives us the following formula:

$$\int_{1-r}^{1}\frac{1}{x}\simeq\frac{1}{k}\left(\frac{1}{1-\varepsilon}+\frac{1}{1-2\varepsilon}+\ldots+\frac{1}{1-k\varepsilon}\right)$$

Thus, the final velocity we are interested in is given by the following formula:

$$\begin{aligned}v&=\varepsilon s\left(\frac{1}{1-\varepsilon}+\frac{1}{1-2\varepsilon}+\ldots+\frac{1}{1-k\varepsilon}\right)\\&=\frac{rs}{k}\left(\frac{1}{1-\varepsilon}+\frac{1}{1-2\varepsilon}+\ldots+\frac{1}{1-k\varepsilon}\right)\\&\simeq rs\int_{1-r}^{1}\frac{1}{x}\\&=-r\log(1-r)s\end{aligned}$$

(5) As an illustration here, assume that our rocket has shrinked, from a continuous ejection process at speed s, up to mass M/e, with $e\simeq 2.718$ being the usual constant from analysis. In this case we have $r=1-1/e$, and the velocity reached is given by:

$$v=-\left(1-\frac{1}{e}\right)\log\left(\frac{1}{e}\right)s=\left(1-\frac{1}{e}\right)s\simeq 0.632s$$

There are of course many other things that can be said here, and in particular we have some interesting questions related to the best strategy to be followed, in order to beat a given force F, such as gravity, or several such forces. More on this later. □

We have done some math in the above, and for future reference, let us record:

THEOREM 2.8. *For a rocket having initial mass M, and functioning by ejecting pieces of mass m at a constant speed s, the speed reached after k ejections is:*

$$v=\left(\frac{m}{M-m}+\frac{m}{M-2m}+\ldots+\frac{m}{M-km}\right)s$$

In the $m=\varepsilon M$, $k\varepsilon=r\in(0,1)$ and $k\to\infty$ regime we have

$$v\simeq-r\log(1-r)s$$

which represents the velocity after the rocket has shrinked to mass $(1-r)M$. Moreover, this latter conclusion holds under the sole assumption that s is constant.

PROOF. Here the two formulae in the statement are our two main formulae, selected from the proof of Theorem 2.7. As for the last assertion, recall also from the proof of

Theorem 2.7 that, assuming only that s is constant, the formula of the velocity is:

$$v=\left(\frac{m_1}{M-m_1}+\frac{m_2}{M-m_1-m_2}+\ldots+\frac{m_k}{M-m_1-\ldots-m_k}\right)s$$

The point now is that, assuming that the ejection pieces $m_1,\ldots,m_k$, instead of being all equal to a certain m, are at least of comparable size, the Riemann integration arguments from the end of the proof of Theorem 2.7 will apply as well, and give the result. □

Let us record as well the continuous version of the above result:

THEOREM 2.9. *For a rocket with initial mass M, ejecting at speed $s=s(x)$, with x being the fraction of the already ejected mass, the speed reached at $x=r$ is:*

$$v=r\int_{1-r}^{1}\frac{s}{x}\,dx$$

In particular, when the ejection speed is constant $s\in\mathbb{R}$, we have $v=-r\log(1-r)s$.

PROOF. Again, this is something which follows from the above. To be more precise, the last assertion follows from Theorem 2.8, or from the first assertion. Regarding now the first assertion, recall from the proof of Theorem 2.7 that the discrete formula is:

$$v_k=\frac{ms_1}{M-m}+\frac{ms_2}{M-2m}+\ldots+\frac{ms_k}{M-km}$$

We can now proceed as in the proof of Theorem 2.7, and with $m=\varepsilon M$ as there, and then with $k\varepsilon=r\in(0,1)$ fixed and $k\to\infty$ as there as well, we obtain:

$$\begin{aligned}v_k&=\varepsilon\left(\frac{s_1}{1-\varepsilon}+\frac{s_2}{1-2\varepsilon}+\ldots+\frac{s_k}{1-k\varepsilon}\right)\\&=\frac{r}{k}\left(\frac{s_1}{1-\varepsilon}+\frac{s_2}{1-2\varepsilon}+\ldots+\frac{s_k}{1-k\varepsilon}\right)\\&\simeq r\int_{1-r}^{1}\frac{s}{x}\,dx\end{aligned}$$

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

As already mentioned, more on this later, when talking about gravity, and trying to beat it with such devices. In particular, we will be back to the above computations, and fine-tune the choice of the ejection method, depending on the problem to be solved.

Finally, let us mention that the above result remains a bit theoretical, because if you are a space engineer, one of your main concerns, besides of course beating gravity, is that of beating atmospheric drag too. More on this, later in this book.

2c. Free falls

Let us start with something immensely important, in the history of science:

FACT 2.10. *Newton invented calculus for formulating the laws of motion as*

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

where x, v, a are the position, speed and acceleration, and the dots are time derivatives.

To be more precise, the variable in Newton's physics is time $t \in \mathbb{R}$, playing the role of the variable $x \in \mathbb{R}$ that we have used in the above. And we are looking at a particle whose position is described by a function $x = x(t)$. Then, it is quite clear that the speed of this particle should be described by the first derivative $v = x'(t)$, and that the acceleration of the particle should be described by the second derivative $a = v'(t) = x''(t)$.

Getting now to the real thing, forces, we will first talk about gravity. We have:

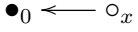
THEOREM 2.11. *The equation of a gravitational free fall, in 1 dimension, is*

$$\ddot{x} = -\frac{GM}{x^2}$$

with M being the attracting mass, and $G \simeq 6.674 \times 10^{-11}$ being a constant.

PROOF. Assume indeed that we have a free falling object, in 1 dimension:

In order to reach to calculus as we know it, we must peform a rotation, as to have all this happening on the Ox axis. By doing this, and assuming that M is fixed at 0, our picture becomes as follows, with the attached numbers being now the coordinates:



Now comes the physics. The gravitational force exterted by M, which is fixed in our formalism, on the object m which moves, is subject to the following equations:

$$F = -G \cdot \frac{Mm}{x^2} \quad , \quad F = ma \quad , \quad a = \dot{v} \quad , \quad v = \dot{x}$$

To be more precise, in the first equation $G \simeq 6.674 \times 10^{-11}$ is the gravitational constant, in usual SI units, and the sign is $-$ because F is attractive. The second equation is something standard and very intuitive, and the last two equations are those from Fact 2.10. Now observe that, with the above data for F, the equation $F = ma$ reads:

$$-G \cdot \frac{Mm}{x^2} = m\ddot{x}$$

Thus, by simplifying, we are led to the equation in the statement. □

2d. N bodies

At a more advanced level, that of several bodies, let us start with:

DEFINITION 2.12. *Associated to a system of bodies $M_1, \ldots, M_k$, located at positions $c_1, \ldots, c_k \in \mathbb{R}$ is their center of mass, located at the following position:*

$$c = \frac{\sum_i c_i M_i}{\sum_i M_i}$$

A single body of mass $\sum_i M_i$ located there, at the center of mass, and with $M_1, \ldots, M_k$ being erased, will be called average of the system formed by $M_1, \ldots, M_k$.

Let us start with some basic mathematics of the center of mass, as constructed above. To be kept in mind first is:

PROPOSITION 2.13. *The center of mass is not a center of gravity, in the sense that the gravity there is not necessarily 0. For instance the center of mass of a dumbell is*

$$\bullet_{M_1} \underbrace{\text{———}}_{\frac{M_2 d}{M_1+M_2}} \star_{cm} \underbrace{\text{———}}_{\frac{M_1 d}{M_1+M_2}} \circ_{M_2}$$

while the center of gravity, which is the unique point where the gravity is 0, is:

$$\bullet_{M_1} \underbrace{\text{———}}_{\frac{\sqrt{M_1} d}{\sqrt{M_1}+\sqrt{M_2}}} \star_{cg} \underbrace{\text{———}}_{\frac{\sqrt{M_2} d}{\sqrt{M_1}+\sqrt{M_2}}} \circ_{M_2}$$

PROOF. There are several assertions here, the idea being as follows:

(1) Regarding the dumbell, pictured above with $M_1 > M_2$, the formula for the center of mass is clear from definitions. Regarding now the center of gravity, the formula there can be found by doing the math, and it works, because the acceleration there is:

$$a = -\frac{GM_1}{\left(\frac{\sqrt{M_1} d}{\sqrt{M_1}+\sqrt{M_2}}\right)^2} + \frac{GM_2}{\left(\frac{\sqrt{M_2} d}{\sqrt{M_1}+\sqrt{M_2}}\right)^2} = 0$$

(2) Getting now to systems $M_1, \ldots, M_k$ with $k \geq 3$, things here are more complicated. Let us first look at the simplest case, that of 3 bodies on a line, at distinct positions. Here, by obvious reasons, we have 2 centers of gravity, as follows:

$$\bullet_{M_1} \text{—} \star_{x_1} \text{—} \bullet_{M_2} \text{—} \star_{x_2} \text{—} \bullet_{M_3}$$

More generally, again by obvious reasons, a system of aligned bodies $M_1, \ldots, M_k$ has $k-1$ centers of gravity, one in between each pair of consecutive bodies. □

Moving ahead, and looking for an easier question, let us still examine the gravity of a rigid object, formed by fixed bodies $M_1, \ldots, M_k$, but at a distance. We have here:

THEOREM 2.14. *Consider a rigid object, consisting of fixed bodies $M_1, \ldots, M_k$, located at positions $c_1, \ldots, c_k \in \mathbb{R}^3$. The corresponding gravitation force, $F = -\nabla V$ with*

$$V = -\sum_i \frac{GmM_i}{|x - c_i|}$$

can be approximated by the force coming from the center of mass, $F_c = -\nabla V$ with

$$V_c = -\frac{Gm \sum_i M_i}{|x - c|}$$

at order zero, when $x >> c_i$. The correction term can be computed as well.

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

(1) The first assertion, $F \simeq F_c$ when $x >> c_i$, is something clear, and with this not even needing c to be the center of mass. Indeed, with V, V_c as above, we have:

$$V = -\sum_i \frac{GmM_i}{|x - c_i|} \simeq -\sum_i \frac{GmM_i}{|x - c|} = V_c$$

(2) Regarding now the correction term, the error to be estimated is:

$$\begin{aligned} V - V_c &= -\sum_i \frac{GmM_i}{|x - c_i|} + \frac{Gm \sum_i M_i}{|x - c|} \\ &= \sum_i GmM_i \left(\frac{1}{|x - c|} - \frac{1}{|x - c_i|} \right) \end{aligned}$$

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

Before going ahead and leaving this subject, let us mention that an interesting generalization of the above comes when considering a "true" rigid body, made of matter arranged according to a certain density function ρ inside it. We will not go into details here, and instead let us just formulate a basic statement, as follows:

THEOREM 2.15. *Consider a rigid body, made of matter arranged according to a certain density function ρ inside it. Its gravitational force is then $F = -\nabla V$ with*

$$V = -\int \frac{Gm\rho(z)}{|x - z|}\, dz$$

and can be approximated by the force coming from the center of mass, $F_c = -\nabla V$ with

$$V_c = -\frac{Gm \int \rho(z) dz}{\left| x - \int u\rho(u) du \right|}\, dz$$

at order zero, when m is far away. The correction term can be computed as well.

PROOF. Here the formulae in the statement, which are perfectly similar to those in Theorem 2.14, can be obtained via the usual philosophy "replace sums by integrals". Observe in particular the formula of the center of mass, producing V_c, namely:

$$c = \int u\rho(u)du$$

As for the last assertion, this can only hold too, by proceeding as in the proof of Theorem 2.14, and replacing everywhere at the end the sums by integrals. □

Finally, let us discuss energy conservation questions. Let us formulate:

DEFINITION 2.16. *An inertial frame is a frame where all basic formulae, namely*

$$|F| = \frac{Gm_1m_2}{(x_1 - x_2)^2} \quad , \quad F = ma \quad , \quad a = \dot{v} \quad , \quad v = \dot{x} \quad , \quad F_{12} = -F_{21}$$

hold, with the last formula standing for Newton's action-reaction principle.

To be more precise here, the first 4 formulae are something that we have been heavily using, so far in this book. As for the last formula, also called Newton's third law, this expresses the fact that when an object 1 acts on an object 2, say via gravity, with force F_{12}, then object 2 acts as well on object 1, with force $F_{21} = -F_{12}$.

In relation with our present considerations, we have the following basic examples:

PROPOSITION 2.17. *In the context of the 2-body problem, the basic frames of type*

$$\lambda_1 M_1 + \lambda_2 M_2$$

are all non-inertial, including the center of mass frame.

PROOF. Since our definition of an inertial frame was something quite informal, so will be this proof. We want to check whether the forces between M_1, M_2 satisfy:

$$F_{12} = -F_{21} = \frac{GM_1M_2(x_1 - x_2)}{|x_1 - x_2|^3}$$

(1) In the case of the frame centered at M_1, the formula $F_{12} = -F_{21}$ certainly does not hold, because the acceleration of M_1 is in this case $\ddot{0} = 0$, and so no force acting upon it, at least from our calculus viewpoint. The same holds for the frame centered at M_2.

(2) In the case now where we have parameters λ_1, λ_2 satisfying $\lambda_1 + \lambda_2 = 0$, the positions of M_1, M_2 are given by the following formulae:

$$z_1 = -\lambda_1 x \quad , \quad z_2 = \lambda_2 x$$

Thus the forces acting upon M_1, M_2, computed according to calculus, are:

$$F_{21} = -M_1\lambda_1\ddot{x} \quad , \quad F_{12} = -M_2\lambda_2\ddot{x}$$

Thus, in order to have $F_{12} = -F_{21}$, the parameters λ_1, λ_2 satisfy $M_1\lambda_1 = M_2\lambda_2$. But these are exactly the parameters of the center of mass.

(3) But the center of mass frame is not inertial either, because due to the fact that we performed a dilation, the magnitude of $F_{12} = -F_{21}$ is not the correct one. □

In what regards the conservation of energy, here we have the following result:

THEOREM 2.18. *With a suitable potential formalism, the total energy*

$$E = \sum_i T_i + V_i$$

of a system of bodies $M_1, \ldots, M_k$ is conserved. Also, the individual energy

$$E' = T' + \sum_i V_i'$$

of an extra body m added is conserved as well, again with a suitable formalism.

PROOF. There are several questions here, the idea being as follows:

(1) In what regards $T = \sum_i T_i$ we have, exactly as in the 2-body problem:

$$\begin{aligned} \dot{T} &= \sum_{i \neq j} < v_i, F_{ji} > \\ &= \sum_{i \neq j} < v_i, -\nabla V_{ji} > \\ &= -\sum_{i \neq j} \dot{V}_{ji} \end{aligned}$$

(2) With this in hand, we can group pairs of terms, and we are led to the conclusion in the statement, with the remark that all the potentials appearing are time-dependent.

(3) In what regards now the second assertion, this is not exactly something of the same nature as the first assertion, becase assuming that by some kind of miracle we would have a theory where all the bodies conserve their energy, the total energy of the system would be trivially conserved too, just by summing, and this does not look normal. So, getting now to the second assertion as formulated, we have, by computing as in (1) above:

$$\dot{T}' = -\sum_i \dot{V}_i'$$

(4) Thus, we are led to the conclusion in the statement, with the problem however that all the potentials appearing there are now time-dependent. □

2e. Exercises

Exercises:

Exercise 2.19.

Exercise 2.20.

Exercise 2.21.

Exercise 2.22.

Exercise 2.23.

Exercise 2.24.

Exercise 2.25.

Exercise 2.26.

Bonus exercise.

CHAPTER 3

Functions, calculus

3a. 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 3.1. *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 3.2. *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 3.1. 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 3.1 does not converge, so we have our counterexample. □

Generally speaking, the last assertion in Proposition 3.2 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 3.3. *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 3.4. *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 3.5. *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 3.6. *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 3.7. *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 3.8. *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 3.7 (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 3.7 (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 3.9. *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 3.3, we have:

THEOREM 3.10. *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 3.3, 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 3.10, we have:

THEOREM 3.11. *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 3.10. 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 3.10 can be used in order to find the maximum and minimum of any differentiable function, and this method is best recalled as follows:

ALGORITHM 3.12. *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.

3b. 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 3.13. *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 3.14. *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 3.15. *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 3.14, 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 3.16. *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 3.3 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 3.3. 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 3.3, and there are many applications of it. As a first such application, justifying Interpretation 3.14 (3), we have the following statement, which is a bit heuristic, but we will call it however Proposition:

PROPOSITION 3.17. *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 3.17, as such, and with $t\to -t$ too:

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

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

By making the average, we obtain the following formula:

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

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

Back to rigorous mathematics, we can improve as well Theorem 3.10, as follows:

THEOREM 3.18. *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 3.16, 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 3.10, 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 3.12 as well. Normally that algorithm stays strong, because Theorem 3.18 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 3.18 is good to know, when applying that algorithm.

3c. 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 3.19. *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 3.16. 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 3.20. *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. □

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

THEOREM 3.21. *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. □

As another 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 3.22. *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. □

3d. Integrals

Let us discuss now the integration of functions, and its relation with differentiability. To start with, we have something very simple, as follows:

DEFINITION 3.23. *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.

Getting now to the computation of integrals, we can use here:

THEOREM 3.24. *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 3.23, because area means side times average height. Thus, we can formulate:

THEOREM 3.25. *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 3.23, via some geometric thinking. Alternatively, this is something which certainly comes from Theorem 3.24. □

Going ahead with more interpretations of the integral, we have:

THEOREM 3.26. *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 3.24 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. □

Finally, here is one more useful interpretation of the integral:

THEOREM 3.27. *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 3.26, 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. □

Still at the general level, let us record as well the following result:

THEOREM 3.28. *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:

$$(b-a)\min(f) \leq \int_a^b f(x)dx \leq (b-a)\max(f)$$

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. □

All this was theory, and getting now to the real thing, explicit computation of the integrals, we have here the following result, called fundamental theorem of calculus:

THEOREM 3.29. *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. 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, 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, also called fundamental theorem of calculus:

THEOREM 3.30. *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 3.29, and is in fact equivalent to it. Indeed, consider the following function:

$$G(s) = \int_a^s F'(x)dx$$

By using Theorem 3.29 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 an illustration for this, solving some concrete integration problems, we have:

THEOREM 3.31. *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 3.32. *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 the standard fact 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 3.31 in a more digest form, as follows:

THEOREM 3.33. *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 3.31, and the last assertion comes from the integration formula given after Definition 3.32. □

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

THEOREM 3.34. *We have the formula*

$$\int f'g + \int fg' = fg$$

called integration by parts.

PROOF. This follows by integrating the Leibnitz formula, namely:

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

Indeed, with our convention for primitives, this gives the formula in the statement. □

It is then possible to pass to usual integrals, and we obtain a formula here as well, as follows, also called integration by parts, with the convention $[\varphi]_a^b = \varphi(b) - \varphi(a)$:

$$\int_a^b f'g + \int_a^b fg' = \Big[fg\Big]_a^b$$

In practice, the most interesting case is that when fg vanishes on the boundary $\{a, b\}$ of our interval, leading to the following formula:

$$\int_a^b f'g = -\int_a^b fg'$$

Examples of this usually come with $[a, b] = [-\infty, \infty]$, and more on this later. Now still at the theoretical level, we have as well the following result:

THEOREM 3.35. *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 before, and whose proof is something elementary:

$$(F\varphi)'(t) = F'(\varphi(t))\varphi'(t)$$

Indeed, by integrating between c and d, we obtain the result. □

As a main application now of our theory, in relation with advanced calculus, and more specifically with the Taylor formula, we have:

THEOREM 3.36. *Given a function* $f : \mathbb{R} \to \mathbb{R}$, *we have the formula*

$$f(x+t) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!}\, t^k + \int_x^{x+t} \frac{f^{(n+1)}(s)}{n!}(x+t-s)^n\, ds$$

called Taylor formula with integral formula for the remainder.

PROOF. This is something which looks a bit complicated, so we will first do some verifications, and then we will go for the proof in general:

(1) At $n = 0$ the formula in the statement is as follows, and certainly holds, due to the fundamental theorem of calculus, which gives $\int_x^{x+t} f'(s)ds = f(x+t) - f(x)$:

$$f(x+t) = f(x) + \int_x^{x+t} f'(s)ds$$

(2) At $n = 1$, the formula in the statement becomes more complicated, as follows:

$$f(x+t) = f(x) + f'(x)t + \int_x^{x+t} f''(s)(x+t-s)ds$$

As a first observation, this formula holds indeed for the linear functions, where we have $f(x+t) = f(x) + f'(x)t$, and $f'' = 0$. So, let us try $f(x) = x^2$. Here we have:

$$f(x+t) - f(x) - f'(x)t = (x+t)^2 - x^2 - 2xt = t^2$$

On the other hand, the integral remainder is given by the same formula, namely:

$$\begin{aligned}\int_x^{x+t} f''(s)(x+t-s)ds &= 2\int_x^{x+t}(x+t-s)ds\\ &= 2t(x+t)-2\int_x^{x+t}sds\\ &= 2t(x+t)-((x+t)^2-x^2)\\ &= 2tx+2t^2-2tx-t^2\\ &= t^2\end{aligned}$$

(3) Still at $n=1$, let us try now to prove the formula in the statement, in general. Since what we have to prove is an equality, this cannot be that hard, and the first thought goes towards differentiating. But this method works indeed, and we obtain the result.

(4) In general, the proof is similar, by differentiating, the computations being similar to those at $n=1$, and we will leave this as an instructive exercise. □

3e. Exercises

Exercises:

Exercise 3.37.

Exercise 3.38.

Exercise 3.39.

Exercise 3.40.

Exercise 3.41.

Exercise 3.42.

Exercise 3.43.

Exercise 3.44.

Bonus exercise.

CHAPTER 4

Waves and heat

4a. Elasticity, waves

Back to physics, we would like to talk now about waves, which are something fundamental. In fact, each major branch of physics is guided by its own wave equation. In practice, and coming a bit in advance, the truth about waves is as follows:

FACT 4.1. *Waves can be of many types, and basically fall into two classes:*

(1) *Mechanical waves, such as the usual water waves, but also the sound waves, or the seismic waves. In all these cases, the wave propagates mechanically, via a certain medium, which can be solid, liquid or gaseous.*
(2) *Electromagnetic waves, coming via a more complicated mechanism, namely an accelerating charge in the context of electromagnetism. These are the radio waves, microwaves, IR, visible light, UV, X-rays and γ-rays.*

Quite remarkably, the behavior of all the above waves is basically described by the same wave equation, which looks as follows, and details on this later:

$$\ddot{\varphi} = v^2\Delta\varphi$$

Getting to work now, in 1 dimension, to start with, the situation is as follows:

THEOREM 4.2. *The wave equation in* 1 *dimension is*

$$\ddot{\varphi} = v^2\varphi''$$

with the dot denoting time derivatives, and $v > 0$ being the propagation speed.

PROOF. We can reach to the above equation via a suitable model, as follows:

(1) In order to understand the propagation of the waves, let us model the space, which is $\mathbb{R}$ for us, as a network of balls, with springs between them, as follows:

$$\cdots \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \cdots$$

(2) Now let us send an impulse, and see how balls will be moving. For this purpose, we zoom on one ball. The situation here is as follows, l being the spring length:

$$\cdots\cdots \bullet_{\varphi(x-l)} \times\!\!\times\!\!\times \bullet_{\varphi(x)} \times\!\!\times\!\!\times \bullet_{\varphi(x+l)} \cdots\cdots$$

We have two forces acting at x. First is the Newton motion force, mass times acceleration, which is as follows, with m being the mass of each ball:

$$F_n = m \cdot \ddot{\varphi}(x)$$

And second is the Hooke force, displacement of the spring, times spring constant. Since we have two springs at x, this is as follows, k being the spring constant:

$$\begin{aligned} F_h &= F_h^r - F_h^l \\ &= k(\varphi(x+l) - \varphi(x)) - k(\varphi(x) - \varphi(x-l)) \\ &= k(\varphi(x+l) - 2\varphi(x) + \varphi(x-l)) \end{aligned}$$

We conclude that the equation of motion, in our model, is as follows:

$$m \cdot \ddot{\varphi}(x) = k(\varphi(x+l) - 2\varphi(x) + \varphi(x-l))$$

(3) 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 (2) reads:

$$\ddot{\varphi}(x) = \frac{KN^2}{M}(\varphi(x+l) - 2\varphi(x) + \varphi(x-l))$$

(4) Now observe that this equation can be written, more conveniently, as follows:

$$\ddot{\varphi}(x) = \frac{KL^2}{M} \cdot \frac{\varphi(x+l) - 2\varphi(x) + \varphi(x-l)}{l^2}$$

With $N \to \infty$, and therefore $l \to 0$, we obtain in this way:

$$\ddot{\varphi}(x) = \frac{KL^2}{M} \cdot \frac{d^2\varphi}{dx^2}(x)$$

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

In order to reach now to some further insight into our spring models above, we must get deeper into elasticity. Indeed, the Hooke law that we used has behind it some trivial elasticity, of "linear" type, and understanding all this, and further modifying our models, according to what elasticity theory exactly says, is certainly an interesting question.

Observe that all this can only lead us too into a better understanding of the fact that the propagation speed is finite, $v < c$. Indeed, the Hooke law is something static, and for better understanding what happens dynamically, we must certainly go into elasticity.

As a starting point for all this, we have the following result:

THEOREM 4.3. *The wave equation can be understood as well directly, as a wave propagating through a linear elastic medium, via stress.*

PROOF. Assume indeed that we have a bar of length L, made of linear elastic material. The stiffness of the bar is then the following quantity, with A being the cross-sectional area, and with E being the Young modulus of the material:

$$K = \frac{EA}{L}$$

Now when sending a pulse, this propagates as follows, M being the total mass:

$$\ddot{\varphi} = \frac{EAL}{M} \cdot \varphi''(x)$$

Bur since $V = AL$ is the volume, with $\rho = M/V$ being the density, we have:

$$\ddot{\varphi} = \frac{E}{\rho} \cdot \varphi''(x)$$

Thus, as a conclusion, the wave propagates with speed $v = \sqrt{E/\rho}$. □

As mentioned in the beginning of this chapter, the next question which appears is that of understanding how exactly the various mechanical waves propagate through solids, liquids and gases, and what corrections to the wave equation are needed, in each case. We will discuss this later in this book, after learning some thermodynamics.

4b. D'Alembert formula

With a bit of mathematical work, we can in fact fully solve the 1D wave equation. In order to explain this, we will need a standard calculus result, as follows:

PROPOSITION 4.4. *The derivative of a function of type*

$$\varphi(x) = \int_{g(x)}^{h(x)} f(s)ds$$

is given by the formula $\varphi'(x) = f(h(x))h'(x) - f(g(x))g'(x)$.

PROOF. Consider a primitive of the function that we integrate, $F' = f$. We have:

$$\begin{aligned}
\varphi(x) &= \int_{g(x)}^{h(x)} f(s)ds \\
&= \int_{g(x)}^{h(x)} F'(s)ds \\
&= F(h(x)) - F(g(x))
\end{aligned}$$

By using now the chain rule for derivatives, we obtain from this:

$$\begin{aligned}\varphi'(x) &= F'(h(x))h'(x) - F'(g(x))g'(x)\\ &= f(h(x))h'(x) - f(g(x))g'(x)\end{aligned}$$

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

Now back to the 1D waves, the result here, due to d'Alembert, is as follows:

THEOREM 4.5. *The solution of the 1D wave equation $\ddot{\varphi} = v^2\varphi''$ with initial value conditions $\varphi(x,0) = f(x)$ and $\dot{\varphi}(x,0) = g(x)$ is given by the d'Alembert formula:*

$$\varphi(x,t) = \frac{f(x-vt) + f(x+vt)}{2} + \frac{1}{2v}\int_{x-vt}^{x+vt} g(s)ds$$

Moreover, in the context of our previous lattice model discretizations, what happens is more or less that the above d'Alembert integral gets computed via Riemann sums.

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

(1) Let us first check that the d'Alembert solution is indeed a solution of the wave equation $\ddot{\varphi} = v^2\varphi''$. The first time derivative is computed as follows:

$$\dot{\varphi}(x,t) = \frac{-vf'(x-vt) + vf'(x+vt)}{2} + \frac{1}{2v}(vg(x+vt) + vg(x-vt))$$

The second time derivative is computed as follows:

$$\ddot{\varphi}(x,t) = \frac{v^2f''(x-vt) + v^2f(x+vt)}{2} + \frac{vg'(x+vt) - vg'(x-vt)}{2}$$

Regarding now space derivatives, the first one is computed as follows:

$$\varphi'(x,t) = \frac{f'(x-vt) + f'(x+vt)}{2} + \frac{1}{2v}(g'(x+vt) - g'(x-vt))$$

As for the second space derivative, this is computed as follows:

$$\varphi''(x,t) = \frac{f''(x-vt) + f''(x+vt)}{2} + \frac{g''(x+vt) - g''(x-vt)}{2v}$$

Thus we have indeed $\ddot{\varphi} = v^2\varphi''$. As for the initial conditions, $\varphi(x,0) = f(x)$ is clear from our definition of φ, and $\dot{\varphi}(x,0) = g(x)$ is clear from our above formula of $\dot{\varphi}$.

(2) Conversely now, we can simply solve our equation, which among others will doublecheck the computations in (1). Let us make the following change of variables:

$$\xi = x - vt \quad , \quad \eta = x + vt$$

With this change of variables, which is quite tricky, mixing space and time variables, our wave equation $\ddot{\varphi} = v^2\varphi''$ reformulates in a very simple way, as follows:

$$\frac{d^2\varphi}{d\xi d\eta} = 0$$

But this latter equation tells us that our new ξ, η variables get separated, and we conclude from this that the solution must be of the following special form:

$$\varphi(x,t) = F(\xi) + G(\eta) = F(x - vt) + G(x + vt)$$

Now by taking into account the intial conditions $\varphi(x,0) = f(x)$ and $\dot{\varphi}(x,0) = g(x)$, and then integrating, we are led to the d'Alembert formula. Finally, in what regards the last assertion, we will leave the study here as an instructive exercise. □

4c. Gases, pressure

We must first talk about gases. Let us start with the following basic fact, which was the beginning of everything, going back to work of Boyle, Charles, Avogardo, Gay-Lussac, Clapeyron and others from the 17th, 18th and 19th centuries, and with final touches from Maxwell, Boltzmann, Gibbs and others, in the late 19th and early 20th centuries:

FACT 4.6. *The ideal gases satisfy the equation $PV = kT$, where:*

(1) *V is the volume of the gas, independently of the shape of the container used.*
(2) *P is the pressure of the gas, measured with a manometer.*
(3) *T is the temperature of the has, measured with a thermometer.*
(4) *k is a constant, depending on the gas.*

That is, $PV = kT$ basically tells us that "pressure and temperature are the same thing".

At the first glance, for instance if you are a mathematician not used to this, this looks more like a joke. Why not defining then $P = T$ or vice-versa, you would say, and what is the point with that long list of distinguished gentlemen having worked hard on this.

Error. The point indeed comes from the following:

EXPLANATION 4.7. *In the equation of state $PV = kT$, as formulated above, the pressure P and the temperature T appear more precisely as follows,*

(1) *The manometer read comes from the gas molecules pushing a piston, so P is a statistical quantity, coming from the statistics of the molecular speeds,*
(2) *The thermometer read is something even more complicated, and T is as well a statistical quantity, coming from the statistics of the molecular speeds,*

so $PV = kT$ is something non-trivial, telling us that the mathematical machinery producing P, T, via manometer and thermometer, out of the molecular speeds, is the same.

Hope you got my point, and getting back now to historical details, Boyle, Charles, Avogardo, Gay-Lussac, Clapeyron, joined by Clausius, Carnot, Joule, Lord Kelvin and others, first observed $PV = kT$, and then reached to a good understanding of what this means, via an axiomatization of P and T. Later Maxwell started to look into the molecular speeds and their statistics, then Boltzmann came with a tough mathematical

computation, proving $PV = kT$, and then, even later, Gibbs and others further built on all this, by formalizing modern thermodynamics, in the form that is still used today.

But probably too much talking, let us get to work. As a first result now, dealing with pressure only, and for the gases without collisions between molecules, we have:

THEOREM 4.8. *The pressure P and volume V of a gas having point molecules, with no collisions between them, satisfy*

$$PV = 2K$$

where K is the total kinetic energy K of the gas.

PROOF. We want to compute the pressure P on the right wall. Since there are no collisions, we can assume by linearity that our gas has 1 molecule, having mass m and traveling at speed v. Our molecule hits the right wall at every $\Delta t = 2L/v$ interval, with its change of momentum being $\Delta p = 2mv$. We obtain:

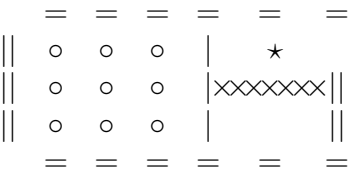
$$P = \frac{F}{L^2} = \frac{\Delta p}{L^2 \Delta t} = \frac{2mv}{L^2 \cdot 2L/v} = \frac{mv^2}{L^3} = \frac{2K}{V}$$

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

Going ahead now with the real problem, namely finding models for the piston, and then doing some math afterwards, we have several choices here. First we have:

MODEL 4.9 (Spring model). *The piston has a spring on its back, with the energy $E = mv^2/2$ of each incoming molecule being converted, over a certain period of time $\Delta t > 0$, into internal energy E_s of the spring, until the molecule comes to a stop, and then released back as identical kinetic energy $E = mv^2/2$, over the same period of time $\Delta t > 0$, of that molecule bouncing back, with speed of same magnitude $||v||$.*

In other words, we are proposing here a model for the piston which is similar to the one which can be found inside the usual, real-life manometers. The functioning is as follows, with $\star$ standing for our displacement measuring devices:



This model surely stands, and certainly brings some fresh air into our physics. Indeed, what we have been doing so far assumes that the collisions with the piston are elastic and instantaneous, $\Delta t = 0$, and the problem now is about fine-tuning our theory using collision times $\Delta t > 0$, as above. In addition, the model can be further complicated afterwards, say allowing for some friction on the vertical, which amounts in allowing heat diffusion at the piston, having something to do with the temperature T of the gas.

As a variation of this model, again inspired by usual manometers, the spring used above is just a flexible solid, so why not using instead a liquid, or even a gas. We are led in this way to the following scheme, with $\bullet$ standing our favorite lab fluid, and with $\star$ standing as usual for our measurement devices, now floating inside this fluid:

$$\begin{array}{cccccccc}
 & = & = & = & = & = & = & \\
\| & \circ & \circ & \circ & | & \star & \bullet & \| \\
\| & \circ & \circ & \circ & | & \bullet & \bullet & \| \\
\| & \circ & \circ & \circ & | & \bullet & \bullet & \| \\
 & = & = & = & = & = & = &
\end{array}$$

To be more precise, assuming for instance that $\bullet$ is another gas, initially at lower pressure than the gas to be studied $\circ$, the piston will certainly start moving to the right, and then after some time, start to stabilize, with an interesting jiggling to be studied.

But do we really have some lab gas $\bullet$, that we know well. Not really, at this point of our story. So we are led into liquids, which are a bit more similar to the solid springs in Model 4.5, but do we really know about compression of liquids, and the answer here is not either. So, we will not use such fluid models, and keep them in mind for later.

As a second main model now, which is intuitive and viable as well, we have:

MODEL 4.10 (Cooking pot model). *The cylinder and piston, functioning now vertically, work as a cooking pot with cover. That is, the gas is cooking inside the pot, the cover has a certain weight $M >> m$ and is subject to an acceleration $g > 0$, not affecting the gas itself, and the molecules m collide elastically with the cover M, assumed to travel frictionless on the vertical only, making it jiggle, around an average height L.*

This model looks quite interesting, and here is the picture, with handles attached, for easy transportation inside the lab, and with $\star$ standing for our measuring devices:

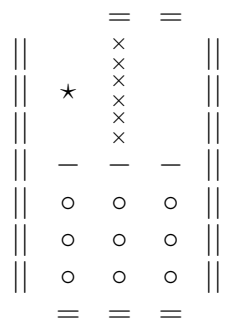
$$\begin{array}{ccccccc}
 & & & \star & & & \\
 & \| & - & - & - & \| & \\
= & \| & \circ & \circ & \circ & \| & = \\
 & \| & \circ & \circ & \circ & \| & \\
 & \| & \circ & \circ & \circ & \| & \\
 & & = & = & = & &
\end{array}$$

The same general comments as for the spring model apply, with this model being something preliminary, which can be subject to further improvements. However, there are two notable differences with the spring model. First, in this cooking model the collisions are still assumed to be instantaneous, $\Delta t = 0$, and so we have less physics to care about. And also, speaking simplicity, our cooking pot model is purely gravitational, and so no need to go into springs and their functioning, we're just ready to go.

Finally, as a third main model, we have something hybrid, as follows:

MODEL 4.11 (Oscillator model). *The cylinder and piston, functioning vertically, work as a cooking pot with harmonic oscillator cover. That is, the gas is cooking inside the pot, the cover has a certain weight $M >> m$ and is subject to an acceleration $g > 0$, not affecting the gas itself, and is attached as well to a spring, freely moving as a harmonic oscillator, and the molecules m collide elastically with the cover M, assumed to travel frictionless on the vertical only, making it jiggle, around an average height L.*

This model looks again interesting, and here is the picture, somewhat hybrid between our previous 2 models, with $\star$ standing as usual for our measuring devices:



The same general comments as for the spring and cooking pot models apply, with this model being something preliminary, which can be subject to further improvements.

Here is now our first result, regarding the cooking pot model:

THEOREM 4.12. *The following happen, for a gas having N point molecules, with no collisions between them, cooking in a pot with cover, as in Model 4.10:*

(1) *In the usual regime, $N >> 0$, the cover mass M and the acceleration g must be subject to the formula $dLMg = 2K$, with d, L, K being as before.*
(2) *At $N = 1$, that is to say, when cooking a single molecule, the cover will bounce up and fall, perfectly in tune with the molecule, which keeps its speed $||v||$.*
(3) *At $N = 2$ however, when cooking two molecules, the initial speeds v_1, v_2 of these molecules, even when taken equal, will change over time, due to the cover.*
(4) *Even worse, at $N = 2$ the system will exhibit chaotic behavior, and this for all choices of the initial data. And the same will happen at any $N \geq 2$.*

PROOF. There are many things to be done here, the idea being as follows:

(1) The molecular force acting on the cover, upwards, is given by the following formula, A being the area of the cover:

$$F = PA = \frac{2KA}{V} = \frac{2K}{L}$$

But this force must cancel $F' = Mg$, pointing downwards, so we have:

$$\frac{2K}{L} = Mg \implies LMg = 2K$$

(2) Let us cook now a single molecule, $N = 1$. The process here will take place in 4 steps, as follows:

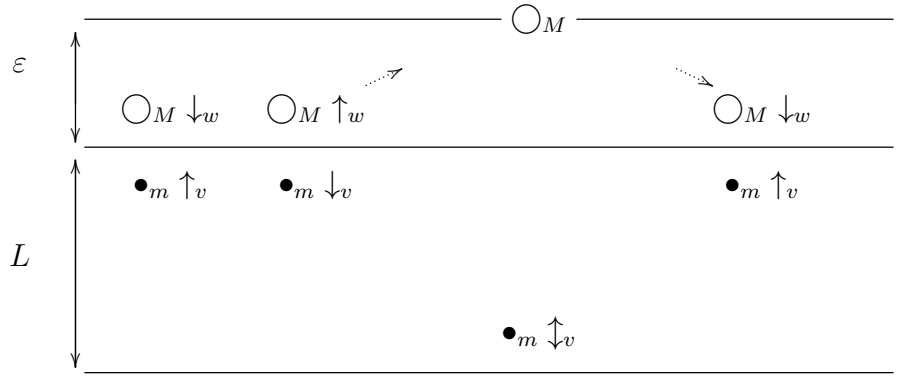
– The molecule, with mass m and upwards speed v, meets the cover, with mass M and downwards speed w, and has an elastic collision with it. Since we want our molecule to simply switch its speed after the collision, $v \to -v$, we must assume $Mw = mv$.

– In the second step, which is also infinitesimal, our molecule is now travelling downwards with speed v, and the cover is now traveling upwards with speed w.

– In the third step, the cover travels upwards during some time t_c, until getting to a halt, under the influence of g. As for the molecule, this travels downwards, during some time t_m, until reaching the bottom of the pot, for an elastic collision there.

– Finally, in the fourth step, the cover falls during time t_c, under the influence of g, until reaching its initial height L, with its intial downwards speed w. As for the molecule, this reaches to the initial height L, with its upwards speed v, in time t_m.

(3) In order now to have a cycle, we must have $t_c = t_m$, as for the whole picture of our cycle to look as follows, over this common traveling time $t_c = t_m$:



(4) But traveling times are easy to compute. In what regards the molecule, its travelling time during half of the full cycle is given by:

$$t_m = \frac{L}{v}$$

As for the cover, its equation of movement, with respect to the origin taken at height L, is $x = wt - gt^2/2$. We have $x = 0$ at $t = 0$, of course, and then again at $t = 2w/g$, and so the travelling time of the cover during half of the full cycle is given by:

$$t_c = \frac{w}{g}$$

Thus, our cycle condition $t_c = t_m$ amounts in saying that we must have $gL = vw$, and so to conclude, our machinery works well when the following conditions are satisfied:

$$Mw = mv \quad , \quad Lg = vw$$

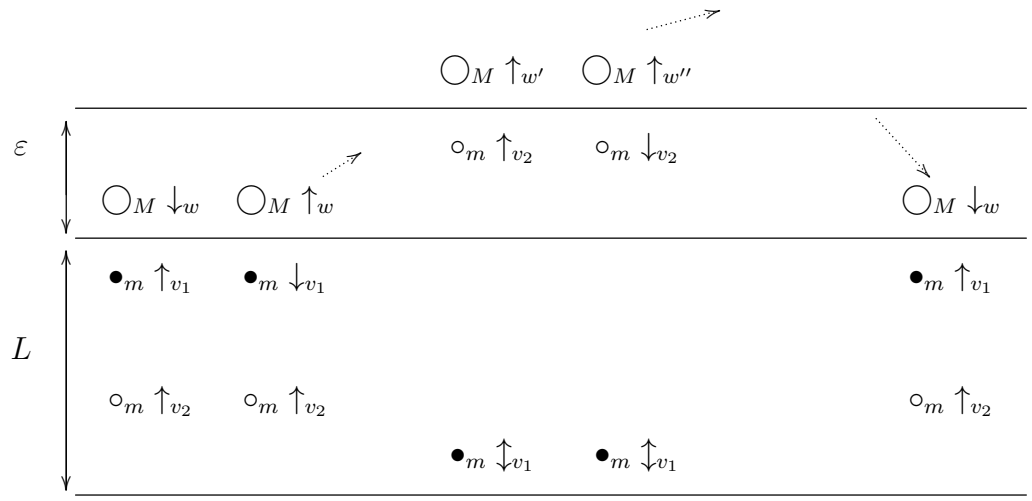
Observe that when multiplying these two equations, as to get rid of the initial cover speed w, we obtain the following equation, which is the one found in (1) above:

$$LMg = mv^2 = 2K$$

(5) At $N = 2$ now, when cooking two molecules, some interesting things happen. To be more precise, our claim is that the initial speeds v_1, v_2 of these two molecules, even when taken equal in magnitude initially, will change over time, due to the cover.

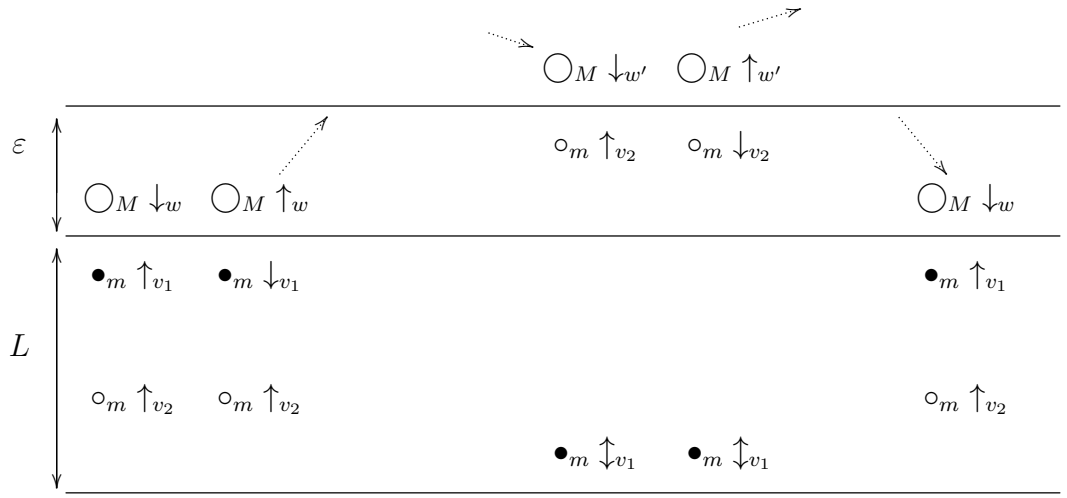
Indeed, in the context of the analysis done in (2-4), a second molecule, hitting the cover after the first one, will hit this cover travelling either upwards or downwards, and in both cases at a speed of different magnitude, $w' \neq w$. Thus, when assuming for instance $v_1 = v_2$ initially, this second collision will be no longer between objects having equal, opposite momenta, and so the speed v_2, instead of simply getting reversed, $v_2 \to -v_2$, will get modified, into something of type $v_2 \to -v_2'$ with $v_2' < v_2$. And so on.

(6) To be more precise, let us show now that there is no possible configuration of the initial parameters as to have a perfect cycle. There are two possible cases. The first case is where the second coming molecule hits the cover during its upwards travel:



But this certainly won't work, because the collision between the cover and the second molecule cannot happen as indicated, on top, due to obvious momentum reasons.

(7) The other case, which is perhaps more realistic, is that when the second coming molecule hits the cover during its downwards travel, as follows:



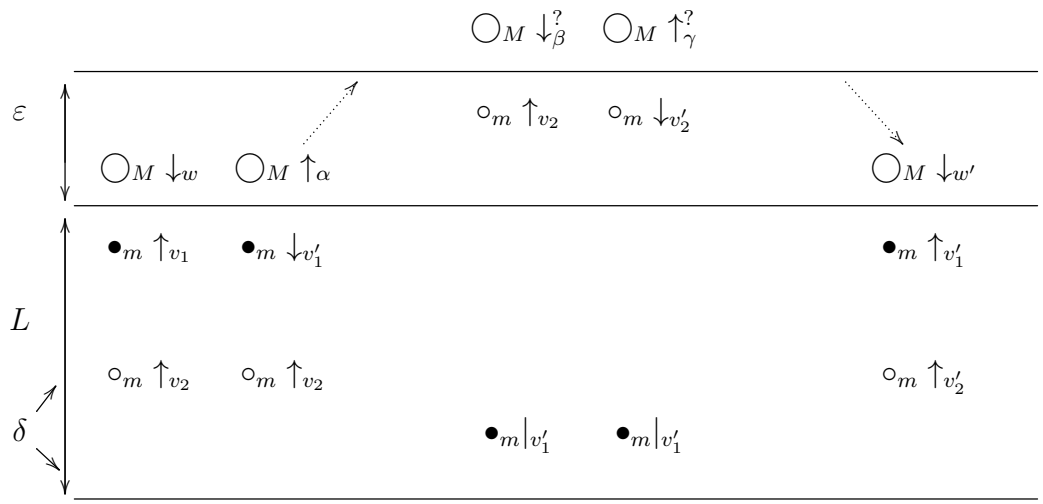
Here both the collisions will perform fine, as indicated, provided that the equal and opposite momenta conditions for them are satisfied, namely:

$$Mw = mv_1 \quad , \quad Mw' = mv_2$$

However, there is a bug at the level of time. On one hand we must have $v_2 > v_1$, since the second molecule has to travel 2ε more than the first one, during the whole cycle. And on the other hand we must have $v_2 < v_1$ due to the above collision equations, since $w' < w$. Thus, contradiction, and this second configuration is ruled out too.

(8) Moving ahead now, the next problem is that of understanding how the speeds w, v_1, v_2 will modify over the time. Assuming, to start with, that we still want to have some sort of cycle, with the positions of the two molecules and of the cover being unchanged

after the cycle, but with the speeds modified, the picture of the problem as follows:



To be more precise, things evolve as indicated, with the upper question marks standing for the fact that we want to deal with all possible orientations there, but we have chosen some orientations, as indicated, for doing our computations, with the convention $\beta, \gamma \in \mathbb{R}$. As for the $|$ signs on the bottom, near the speeds v_1', these stand for the orientations of these speeds $v_1' > 0$, which are irrelevant at that exact moment. And finally, as new parameter we have the distance $\delta > 0$ between the second molecule and the bottom.

(9) Getting now to equations, there are many of them. First we have the two collision equations, momentum and energy, which after simplification for energy are:

$$M(w - \alpha) = m(v_1' - v_1) \quad , \quad M(\beta - \gamma) = m(v_2' - v_2)$$

$$w + \alpha = v_1 + v_1' \quad , \quad \beta + \gamma = v_2 + v_2'$$

We have then two equations relating the speeds of M, left to middle, and middle to right, which can be obtained by conservation of energy, and are as follows:

$$\alpha^2 - \beta^2 = w'^2 - \gamma^2 = 2g\varepsilon$$

We have then equations regarding the partial times t_1, t_2 of our two-step cycle, viewed from the perspective of the second molecule, and of the first molecule, as follows:

$$t_1 = \frac{L - \delta + \varepsilon}{v_2} \quad , \quad t_2 = \frac{L + \delta + \varepsilon}{v_2'} \quad , \quad t_1 + t_2 = \frac{2L}{v_1'}$$

And finally we have degree 2 equations for t_1, t_2 from the perspective of the cover, which are as follows, with the $\pm$ sign standing for upwards vs downwards collision:

$$t_1 = \frac{\alpha \pm \sqrt{\alpha^2 - 2\varepsilon g}}{g} \quad , \quad t_2 = \frac{\gamma + \sqrt{\gamma^2 + 2\varepsilon g}}{g}$$

Looking at these equations, they don't look that bad. The first 6 equations, all regarding speeds, can be used in order to compute α, β, γ and w', v_1', v_2' in terms of w, v_1, v_2. And then we have 5 equations for t_1, t_2, which can be used for computing t_1, t_2, and then for finding what exact conditions must the intial data $m, M, g, L, \delta, w, v_1, v_2$ satisfy, as for the positions of our 3 objects to be the same in the end as in the beginning.

(10) We can only conclude from all this that things are quite chaotic at $N = 2$, and consequently, at $N \geq 3$ too. With the comment however that with $N >> 0$ something interesting must certainly happen, because after all at $N = \infty$ we have equilibrium, as explained in (1). But we are here with our math on the thin edge between equilibrium and chaos, and such things are reputed to be difficult, so we will just stop here. □

Let us discuss now the spring model. Our result here, quite modest, is as follows:

THEOREM 4.13. *For the spring model, the statistics is basically that of*

$$PV = 2K$$

that we already know. More, of chaotic type, can be said via advanced elasticity.

PROOF. Here the first assertion follows from the above discussion, by recycling the computations from the proof of Theorem 4.8, using the Hooke law, namely:

$$F = kx$$

Indeed, we can incorporate this law into our previous computations either directly, or by using a potential energy argument. As for the second assertion, this is something quite plausible in view of Theorem 4.12, and we will not get into details here. □

Let us turn now into our third model, the oscillator one. We have here:

THEOREM 4.14. *For the oscillator model, the statistics is basically that of*

$$PV = 2K$$

but some $N < \infty$ phenomena, of rather chaotic type, can be observed as well.

PROOF. This follows indeed by doing some computations. □

4d. Heat diffusion

We can talk about the heat equation in 1D, as follows:

THEOREM 4.15. *The heat equation in 1 dimension is*

$$\dot{\varphi} = \alpha\varphi''$$

where $\alpha > 0$ is the thermal diffusivity of the medium.

PROOF. As before with the wave equation, this is not exactly a theorem, but rather what comes out of experiments, but we can justify this mathematically, as follows:

(1) As an intuitive explanation for this equation, since the second derivative φ'' computes the average value of a function φ around a point, minus the value of φ at that point, as we know from chapter 1, the heat equation as formulated above tells us that the rate of change $\dot{\varphi}$ of the temperature of the material at any given point must be proportional, with proportionality factor $\alpha > 0$, to the average difference of temperature between that given point and the surrounding material. Which sounds reasonable.

(2) In practice now, we can use, a bit like before for the wave equation, a lattice model as follows, with distance $l > 0$ between the neighbors:

$$\text{---} \circ_{x-l} \overset{l}{\text{---}} \circ_x \overset{l}{\text{---}} \circ_{x+l} \text{---}$$

In order to model now heat diffusion, we have to implement the intuitive mechanism explained above, and in practice, this leads to a condition as follows, expressing the change of the temperature φ, over a small period of time $\delta > 0$:

$$\varphi(x, t+\delta) = \varphi(x,t) + \frac{\alpha\delta}{l^2} \sum_{x \sim y} [\varphi(y,t) - \varphi(x,t)]$$

But this leads, via manipulations as before, to $\dot{\varphi}(x,t) = \alpha \cdot \varphi''(x,t)$, as claimed. □

4e. Exercises

Exercises:

EXERCISE 4.16.

EXERCISE 4.17.

EXERCISE 4.18.

EXERCISE 4.19.

EXERCISE 4.20.

EXERCISE 4.21.

EXERCISE 4.22.

EXERCISE 4.23.

Bonus exercise.

Part II

Two dimensions

Farewell the ashtray girl
Forbidden snowflake
Beware this troubled world
Watch out for earthquakes

CHAPTER 5

Vector calculus

5a. The plane

Vectors in the plane. Many things can be said here.

5b. Linear maps

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:*

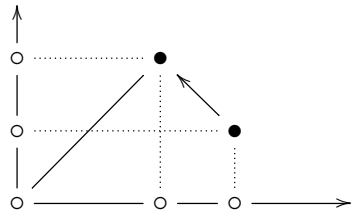
$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} x \\ 0 \end{pmatrix} \quad , \quad \begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} 0 \\ y \end{pmatrix}$$

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 $x \in \mathbb{R}^2$.*

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 5.8. *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 5.8 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 5.9. *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 5.8. Indeed, with the above multiplication convention for matrices and vectors, we can turn Theorem 5.8 into something much simpler, and better-looking, as follows:

THEOREM 5.10. *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 5.8 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 5.11. *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 $\begin{pmatrix}x\\y\end{pmatrix}$ 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 5.12. *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 $\begin{pmatrix}x\\y\end{pmatrix}$ 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 5.13. *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:

$$\binom{x}{0} \quad , \quad \binom{0}{y} \quad , \quad \frac{1}{2}\binom{x+y}{x+y} \quad , \quad \frac{1}{2}\binom{x-y}{y-x}$$

But this gives the formulae in the statement, as before by guessing the matrix. □

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 5.14. *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 5.15. *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 5.16. *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 5.17. *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:

$$\begin{aligned}
R_sR_t &= \begin{pmatrix}\cos s & -\sin s\\ \sin s & \cos s\end{pmatrix}\begin{pmatrix}\cos t & -\sin t\\ \sin t & \cos t\end{pmatrix}\\
&= \begin{pmatrix}\cos s\cos t-\sin s\sin t & -\cos s\sin t-\sin t\cos s\\ \sin s\cos t+\cos s\sin t & -\sin s\sin t+\cos s\cos t\end{pmatrix}\\
&= \begin{pmatrix}\cos(s+t) & -\sin(s+t)\\ \sin(s+t) & \cos(s+t)\end{pmatrix}\\
&= R_{s+t}
\end{aligned}$$

We will be back to this, with many applications, in what follows.

5c. Higher dimensions

We are now ready to discuss 3 and more dimensions. Before doing so, let us point out however that the maps of type $f:\mathbb{R}^3\to\mathbb{R}^2$, or $f:\mathbb{R}\to\mathbb{R}^2$, and so on, are not covered by our results. Since there are many interesting such maps, say obtained by projecting and then rotating, and so on, we will be interested here in the maps $f:\mathbb{R}^N\to\mathbb{R}^M$.

A bit of thinking suggests that such maps should come from the $M\times N$ matrices. Indeed, this is what happens at $M=N=2$, of course, and $M=N=3$ too. But this happens as well at $N=1$, because a linear map $f:\mathbb{R}\to\mathbb{R}^M$ can only be something of the form $f(\lambda)=\lambda v$, with $v\in\mathbb{R}^M$. But $v\in\mathbb{R}^M$ means that v is a $M\times 1$ matrix. So, let us start with the product rule for such matrices, which is as follows:

DEFINITION 5.18. *We can multiply the $M\times N$ matrices with $N\times K$ matrices,*

$$\begin{pmatrix}a_{11} & \ldots & a_{1N}\\ \vdots & & \vdots\\ a_{M1} & \ldots & a_{MN}\end{pmatrix}\begin{pmatrix}b_{11} & \ldots & b_{1K}\\ \vdots & & \vdots\\ b_{N1} & \ldots & b_{NK}\end{pmatrix}$$

the product being the $M\times K$ matrix given by the following formula,

$$\begin{pmatrix}a_{11}b_{11}+\ldots+a_{1N}b_{N1} & \ldots\ldots & a_{11}b_{1K}+\ldots+a_{1N}b_{NK}\\ \vdots & & \vdots\\ \vdots & & \vdots\\ a_{M1}b_{11}+\ldots+a_{MN}b_{N1} & \ldots\ldots & a_{M1}b_{1K}+\ldots+a_{MN}b_{NK}\end{pmatrix}$$

obtained via the usual rule "multiply rows by columns".

Observe that this formula generalizes all the multiplication rules that we have been using so far, between various types of matrices and vectors. Thus, in practice, we can simply forget all the previous multiplication rules, and simply memorize this one.

In case the above formula looks hard to memorize, here is an alternative formulation of it, which is simpler and more powerful, by using the standard algebraic notation for the matrices, $A = (A_{ij})$, that we will heavily use, in what follows:

PROPOSITION 5.19. *The matrix multiplication is given by formula*

$$(AB)_{ij} = \sum_k A_{ik}B_{kj}$$

with A_{ij} standing for the entry of A at row i and column j.

PROOF. This is indeed just a shorthand for the formula in Definition 5.18, by following the rule there, namely "multiply the rows of A by the columns of B". □

As an illustration for the power of the convention in Proposition 5.19, we have:

PROPOSITION 5.20. *We have the following formula, valid for any matrices A, B, C,*

$$(AB)C = A(BC)$$

provided that the sizes of our matrices A, B, C fit.

PROOF. We have the following computation, using indices as above:

$$((AB)C)_{ij} = \sum_k (AB)_{ik}C_{kj} = \sum_{kl} A_{il}B_{lk}C_{kj}$$

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

$$(A(BC))_{ij} = \sum_l A_{il}(BC)_{lj} = \sum_{kl} A_{il}B_{lk}C_{kj}$$

Thus we have $(AB)C = A(BC)$, and we have proved our result. □

We can now talk about linear maps between spaces of arbitrary dimension, generalizing what we have been doing so far. The main result here is as follows:

THEOREM 5.21. *Consider a map $f : \mathbb{R}^N \to \mathbb{R}^M$.*

(1) *f is linear when it is of the form $f(v) = Av$, with $A \in M_{M\times N}(\mathbb{R})$.*
(2) *f is affine when $f(v) = Av + w$, with $A \in M_{M\times N}(\mathbb{R})$ and $w \in \mathbb{R}^M$.*
(3) *We have the composition formula $f_A f_B = f_{AB}$, whenever the sizes fit.*

PROOF. We already know that this happens at $M = N = 2$. In general, the proof is similar, by doing some elementary computations. □

As a first example here, we have the identity matrix, acting as the identity:

$$\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}$$

We have as well the null matrix, acting as the null map:

$$\begin{pmatrix} 0 & \ldots & 0 \\ \vdots & & \vdots \\ 0 & \ldots & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

Here is now an important result, providing us with many examples:

PROPOSITION 5.22. *The diagonal matrices act as follows:*

$$\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} \lambda_1 x_1 \\ \vdots \\ \lambda_N x_N \end{pmatrix}$$

PROOF. This is clear, indeed, from definitions. □

As a more specialized example now, we have:

PROPOSITION 5.23. *The flat matrix, which is as follows,*

$$\mathbb{I}_N = \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix}$$

acts via N times the projection on the all-one vector.

PROOF. The flat matrix acts in the following way:

$$\begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} x_1 + \ldots + x_N \\ \vdots \\ x_1 + \ldots + x_N \end{pmatrix}$$

Thus, in terms of the matrix $P = \mathbb{I}_N/N$, we have the following formula:

$$P \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \frac{x_1 + \ldots + x_N}{N} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Since the linear map $f(x) = Px$ satisfies $f^2 = f$, and since $Im(f)$ consists of the scalar multiples of the all-one vector $\xi \in \mathbb{R}^N$, we conclude that f is a projection on $\mathbb{R}\xi$. Also, with the standard scalar product convention $< x, y >= \sum x_i y_i$, we have:

$$\begin{aligned} < f(x) - x, \xi > &= < f(x), \xi > - < x, \xi > \\ &= \frac{\sum x_i}{N} \times N - \sum x_i \\ &= 0 \end{aligned}$$

Thus, our projection is an orthogonal projection, and we are done. □

Let us develop now some general theory for the square matrices. We will need the following standard result, regarding the changes of coordinates in $\mathbb{R}^N$:

THEOREM 5.24. *For a system $\{v_1,\ldots,v_N\}\subset\mathbb{R}^N$, the following are equivalent:*

(1) *The vectors v_i form a basis of $\mathbb{R}^N$, in the sense that each vector $x\in\mathbb{R}^N$ can be written in a unique way as a linear combination of these vectors:*

$$x=\sum\lambda_i v_i$$

(2) *The following linear map associated to these vectors is bijective:*

$$f:\mathbb{R}^N\to\mathbb{R}^N\quad,\quad\lambda\to\sum\lambda_i v_i$$

(3) *The matrix formed by these vectors, regarded as usual as column vectors,*

$$P=[v_1,\ldots,v_N]\in M_N(\mathbb{R})$$

is invertible, with respect to the usual multiplication of the matrices.

PROOF. Here the equivalence (1) $\iff$ (2) is clear from definitions, and the equivalence (2) $\iff$ (3) is clear as well, because we have $f(x)=Px$. □

Getting back now to the matrices, as an important definition, we have:

DEFINITION 5.25. *Let $A\in M_N(\mathbb{R})$ be a square matrix. We say that $v\in\mathbb{R}^N$ is an eigenvector of A, with corresponding eigenvalue $\lambda\in\mathbb{R}^N$, when:*

$$Av=\lambda v$$

Also, we say that A is diagonalizable when $\mathbb{R}^N$ has a basis formed by eigenvectors of A.

We will see in a moment examples of eigenvectors and eigenvalues, and of diagonalizable matrices. However, even before seeing the examples, it is quite clear that these are key notions. Indeed, for a matrix $A\in M_N(\mathbb{R})$, being diagonalizable is the best thing that can happen, because in this case, once the basis changed, A becomes diagonal.

To be more precise here, we have the following result:

PROPOSITION 5.26. *Assuming that $A\in M_N(\mathbb{R})$ is diagonalizable, we have the formula*

$$A=\begin{pmatrix}\lambda_1&&\\&\ddots&\\&&\lambda_N\end{pmatrix}$$

with respect to the basis $\{v_1,\ldots,v_N\}$ of $\mathbb{R}^N$ consisting of eigenvectors of A.

PROOF. This is clear from the definition of eigenvalues and eigenvectors, and from the formula of linear maps associated to diagonal matrices, from Proposition 5.22. □

Here is an equivalent form of the above result, which is often used in practice, when we prefer not to change the basis, and stay with the usual basis of $\mathbb{R}^N$:

THEOREM 5.27. *Assuming that $A \in M_N(\mathbb{R})$ is diagonalizable, with*

$$v_1, \ldots, v_N \in \mathbb{R}^N \quad , \quad \lambda_1, \ldots, \lambda_N \in \mathbb{R}$$

as eigenvectors and corresponding eigenvalues, we have the formula

$$A = PDP^{-1}$$

with the matrices $P, D \in M_N(\mathbb{R})$ being given by the formulae

$$P = [v_1, \ldots, v_N] \quad , \quad D = diag(\lambda_1, \ldots, \lambda_N)$$

and respectively called passage matrix, and diagonal form of A.

PROOF. This can be viewed in two possible ways, as follows:

(1) As already mentioned, with respect to the basis $v_1, \ldots, v_N \in \mathbb{R}^N$ formed by the eigenvectors, our matrix A is given by:

$$A = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix}$$

But this corresponds precisely to the formula $A = PDP^{-1}$ from the statement, with P and its inverse appearing there due to our change of basis.

(2) We can equally establish the formula in the statement by a direct computation. Indeed, we have $Pe_i = v_i$, where $\{e_1, \ldots, e_N\}$ is the standard basis of $\mathbb{R}^N$, and so:

$$APe_i = Av_i = \lambda_i v_i$$

On the other hand, once again by using $Pe_i = v_i$, we have as well:

$$PDe_i = P\lambda_i e_i = \lambda_i Pe_i = \lambda_i v_i$$

Thus we have $AP = PD$, and so $A = PDP^{-1}$, as claimed. □

Let us discuss now some basic examples, namely the rotations, symmetries and projections in 2 dimensions. The situation is very simple for the projections, as follows:

PROPOSITION 5.28. *The projection on the Ox axis rotated by an angle $t/2 \in \mathbb{R}$,*

$$P_t = \frac{1}{2}\begin{pmatrix} 1 + \cos t & \sin t \\ \sin t & 1 - \cos t \end{pmatrix}$$

is diagonalizable, its diagonal form being as follows:

$$P_t \sim \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

PROOF. This is clear, because if we denote by L the line where our projection projects, we can pick any vector $v \in L$, and this will be an eigenvector with eigenvalue 1, and then pick any vector $w \in L^\perp$, and this will be an eigenvector with eigenvalue 0. Thus, even without computations, we are led to the conclusion in the statement. □

The computation for the symmetries is similar, as follows:

PROPOSITION 5.29. *The symmetry with respect to the Ox axis rotated by $t/2 \in \mathbb{R}$,*

$$S_t = \begin{pmatrix} \cos t & \sin t \\ \sin t & -\cos t \end{pmatrix}$$

is diagonalizable, its diagonal form being as follows:

$$S_t \sim \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

PROOF. This is once again clear, because if we denote by L the line with respect to which our symmetry symmetrizes, we can pick any vector $v \in L$, and this will be an eigenvector with eigenvalue 1, and then pick any vector $w \in L^\perp$, and this will be an eigenvector with eigenvalue -1. Thus, we are led to the conclusion in the statement. □

Regarding now the rotations, here the situation is different, as follows:

PROPOSITION 5.30. *The rotation of angle $t \in [0, 2\pi)$, given by the formula*

$$R_t = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}$$

is diagonal at $t = 0, \pi$, and is not diagonalizable at $t \neq 0, \pi$.

PROOF. The first assertion is clear, because at $t = 0, \pi$ the rotations are:

$$R_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad , \quad R_\pi = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$

As for the rotations of angle $t \neq 0, \pi$, these clearly cannot have eigenvectors. □

Finally, here is one more example, which is the most important of them all:

THEOREM 5.31. *The following matrix is not diagonalizable,*

$$J = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

because it has only 1 eigenvector.

PROOF. The above matrix, called J en hommage to Jordan, acts as follows:

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}$$

Thus the eigenvector/eigenvalue equation $Jv = \lambda v$ reads:

$$\begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{pmatrix} \lambda x \\ \lambda y \end{pmatrix}$$

We have then two cases, depending on λ, as follows, which give the result:

(1) For $\lambda \neq 0$ we must have $y = 0$, coming from the second row, and so $x = 0$ as well, coming from the first row, so we have no nontrivial eigenvectors.

(2) As for the case $\lambda = 0$, here we must have $y = 0$, coming from the first row, and so the eigenvectors here are the vectors of the form $\binom{x}{0}$. □

5d. Scalar products

In order to discuss some interesting examples of matrices, and their diagonalization, in arbitrary dimensions, we will need the following standard fact:

PROPOSITION 5.32. *Consider the scalar product on $\mathbb{R}^N$, given by:*

$$< x, y >= \sum_i x_i y_i$$

We have then the following formula, valid for any vectors x, y and any matrix A,

$$< Ax, y >=< x, A^t y >$$

with A^t being the transpose matrix.

PROOF. By linearity, it is enough to prove the above formula on the standard basis vectors $e_1, \ldots, e_N$ of $\mathbb{R}^N$. Thus, we want to prove that for any i, j we have:

$$< Ae_j, e_i >=< e_j, A^t e_i >$$

The scalar product being symmetric, this is the same as proving that:

$$< Ae_j, e_i >=< A^t e_i, e_j >$$

On the other hand, for any matrix M we have the following formula:

$$M_{ij} =< Me_j, e_i >$$

Thus, the formula to be proved simply reads:

$$A_{ij} = (A^t)_{ji}$$

But this precisely the definition of A^t, and we are done. □

With this, we can develop some theory. We first have:

THEOREM 5.33. *The orthogonal projections are the matrices satisfying:*

$$P^2 = P^t = P$$

These projections are diagonalizable, with eigenvalues $0, 1$.

PROOF. It is obvious that a linear map $f(x) = Px$ is a projection precisely when:

$$P^2 = P$$

In order now for this projection to be an orthogonal projection, the condition to be satisfied can be written and then processed as follows:

$$\begin{aligned} < Px - Py, Px - x >= 0 &\iff < x - y, P^tPx - P^tx >= 0 \\ &\iff P^tPx - P^tx = 0 \\ &\iff P^tP - P^t = 0 \end{aligned}$$

Thus we must have $P^t = P^tP$. Now observe that by transposing, we have as well:

$$P = (P^tP)^t = P^t(P^t)^t = P^tP$$

Thus we must have $P = P^t$, as claimed. Finally, regarding the diagonalization assertion, this is clear by taking a basis of $Im(f)$, which consists of 1-eigenvectors, and then completing with 0-eigenvectors, which can be found inside the orthogonal of $Im(f)$. □

Here is now a key computation of such projections:

THEOREM 5.34. *The rank* 1 *projections are given by the formula*

$$P_x = \frac{1}{||x||^2}(x_ix_j)_{ij}$$

where the constant, namely

$$||x|| = \sqrt{\sum_i x_i^2}$$

is the length of the vector.

PROOF. Consider a vector $y \in \mathbb{R}^N$. Its projection on $\mathbb{R}x$ must be a certain multiple of x, and we are led in this way to the following formula:

$$P_xy = \frac{< y, x >}{< x, x >} x = \frac{1}{||x||^2} < y, x > x$$

With this in hand, we can now compute the entries of P_x, as follows:

$$\begin{aligned} (P_x)_{ij} &= < P_xe_j, e_i > \\ &= \frac{1}{||x||^2} < e_j, x >< x, e_i > \\ &= \frac{x_jx_i}{||x||^2} \end{aligned}$$

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

As an application, we can recover a result that we already know, namely:

PROPOSITION 5.35. *In 2 dimensions, the rank 1 projections, which are the projections on the Ox axis rotated by an angle $t/2 \in [0, \pi)$, are given by the following formula:*

$$P_t = \frac{1}{2}\begin{pmatrix} 1+\cos t & \sin t \\ \sin t & 1-\cos t \end{pmatrix}$$

Together with the following two matrices, which are the rank 0 and 2 projections in $\mathbb{R}^2$,

$$0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \quad , \quad 1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

these are all the projections in 2 dimensions.

PROOF. The first assertion follows from the general formula in Theorem 5.34, by plugging in the following vector, depending on a parameter $s \in [0, \pi)$:

$$x = \begin{pmatrix} \cos s \\ \sin s \end{pmatrix}$$

We obtain in this way the following matrix, which with $t = 2s$ is the one in the statement, via some trigonometry:

$$P_{2s} = \begin{pmatrix} \cos^2 s & \cos s \sin s \\ \cos s \sin s & \sin^2 s \end{pmatrix}$$

As for the second assertion, this is clear from the first one, because outside rank 1 we can only have rank 0 or rank 2, corresponding to the matrices in the statement. □

Here is another interesting application, this time in N dimensions:

PROPOSITION 5.36. *The projection on the all-1 vector $\xi \in \mathbb{R}^N$ is*

$$P_\xi = \frac{1}{N}\begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix}$$

with the all-1 matrix on the right being called the flat matrix.

PROOF. As already pointed out in the proof of Proposition 5.23, the matrix in the statement acts in the following way:

$$P_\xi \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} = \frac{x_1 + \ldots + x_N}{N} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Thus P_ξ is indeed a projection onto $\mathbb{R}\xi$, and the fact that this projection is indeed the orthogonal one follows either by a direct orthogonality computation, or by using the general formula in Theorem 5.34, by plugging in the all-1 vector ξ. □

Let us discuss now, as a final topic of this chapter, the isometries of $\mathbb{R}^N$. We have here the following general result:

THEOREM 5.37. *The linear maps $f : \mathbb{R}^N \to \mathbb{R}^N$ which are isometries, in the sense that they preserve the distances, are those coming from the matrices satisfying:*

$$U^t = U^{-1}$$

These latter matrices are called orthogonal, and they form a set $O_N \subset M_N(\mathbb{R})$ which is stable under taking compositions, and inverses.

PROOF. We have several things to be proved, the idea being as follows:

(1) We recall that we can pass from scalar products to distances, as follows:

$$||x|| = \sqrt{< x, x >}$$

Conversely, we can compute the scalar products in terms of distances, by using the parallelogram identity, which is as follows:

$$\begin{aligned} ||x+y||^2 - ||x-y||^2 &= ||x||^2 + ||y||^2 + 2 < x, y > -||x||^2 - ||y||^2 + 2 < x, y > \\ &= 4 < x, y > \end{aligned}$$

Now given a matrix $U \in M_N(\mathbb{R})$, we have the following equivalences, with the first one coming from the above identities, and with the other ones being clear:

$$\begin{aligned} ||Ux|| = ||x|| &\iff < Ux, Uy > = < x, y > \\ &\iff < x, U^t U y > = < x, y > \\ &\iff U^t U y = y \\ &\iff U^t U = 1 \\ &\iff U^t = U^{-1} \end{aligned}$$

(2) The second assertion is clear from the definition of the isometries, and can be established as well by using matrices, and the $U^t = U^{-1}$ criterion. □

As a basic illustration here, we have:

THEOREM 5.38. *The rotations and symmetries in the plane, given by*

$$R_t = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \quad , \quad S_t = \begin{pmatrix} \cos t & \sin t \\ \sin t & -\cos t \end{pmatrix}$$

are isometries. These are all the isometries in 2 dimensions.

PROOF. We already know that R_t is the rotation of angle t. As for S_t, this is the symmetry with respect to the Ox axis rotated by $t/2 \in \mathbb{R}$. But this gives the result, since the isometries in 2 dimensions are either rotations, or symmetries. □

5e. Exercises

Exercises:

Exercise 5.39.

Exercise 5.40.

Exercise 5.41.

Exercise 5.42.

Exercise 5.43.

Exercise 5.44.

Exercise 5.45.

Exercise 5.46.

Bonus exercise.

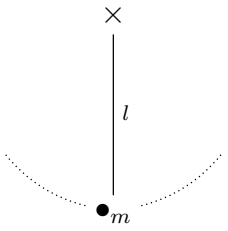
CHAPTER 6

Basic mechanics

6a. The pendulum

Let us start our discussion with something very basic, namely:

DEFINITION 6.1. *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:

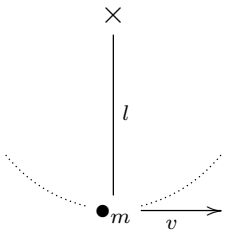
$$V' = -F$$

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 6.1, we can have this understood with not many computations involved, as follows:

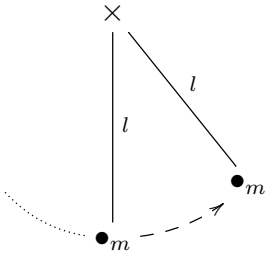
THEOREM 6.2. *For a pendulum starting with speed v from the equilibrium position,*



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:



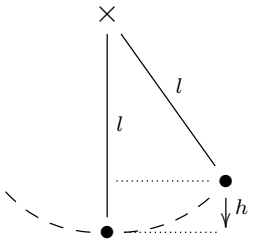
The distance traveled is then $x = l\theta$. As for the force acting, this is $F_{total} = mg$ oriented downwards, with the component alongside x being given by:

$$\begin{aligned} F &= -||F_{total}|| \sin\theta \\ &= -mg\sin\theta \\ &= -mg\sin\left(\frac{x}{l}\right) \end{aligned}$$

(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:

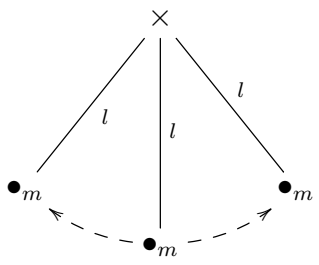
$$\begin{aligned} V' = -\mathcal{F} &\implies V' = mg \\ &\implies V = mgh \\ &\implies V = mgl(1 - \cos\theta) \end{aligned}$$

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:

$$V = E \iff mgl(1 - \cos\theta) = \frac{mv^2}{2}$$

$$\iff 1 - \cos\theta = \frac{v^2}{2gl}$$

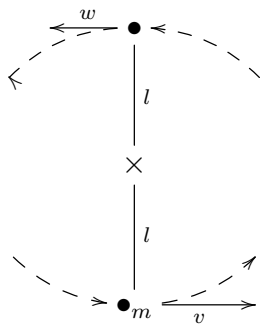
(5) Regarding now the case $v^2 > 4gl$, here the bob will certainly reach the upwards position, with the speed $w > 0$ there being given by the following formula:

$$\frac{mw^2}{2} = E - 2mgl \implies \frac{mw^2}{2} = \frac{mv^2}{2} - 2mgl$$

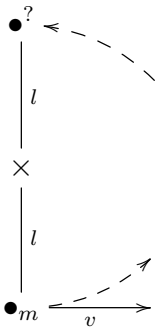
$$\implies w^2 = v^2 - 4gl$$

$$\implies w = \sqrt{v^2 - 4gl}$$

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. □

6b. Harmonic oscillators

Let us discuss now the motion of a particle near an equilibrium point. We have two basic examples of such points provided by the pendulum, namely the downwards one, which is stable, and the upwards one, which is unstable. But our discussion here will be valid for any other types of particles moving, under the influence of forces.

As a first observation, our generalities about motion and energy provide us with:

THEOREM 6.3. *For a particle moving near an equilibrium point $x = 0$, the following equivalent conditions must be satisfied, infinitesimally:*

(1) *The potential energy is $V = kx^2/2$, when assuming $V(0) = 0$.*
(2) *The force acting on our particle is $F = -kx$.*
(3) *The equation of motion is $m\ddot{x} + kx = 0$, with m being the mass.*

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

(1) Let us start with some generalities regarding the potential energy V. Around any given point, that we can choose by translation to be $x = 0$, we can write:

$$V(x) = V(0) + V'(0)x + \frac{V''(0)x^2}{2} + \frac{V'''(0)x^3}{6} + \ldots$$

By definition of V, we can assume $V(0) = 0$. Regarding now the second term, this vanishes too, because our condition of equilibrium reads:

$$V'(0) = -F(0) = 0$$

Thus, with the above normalizations $x = 0$ and $V(0) = 0$ made, our general formula above for V takes at equilibrium the following form, with $k = V''(0)$:

$$V(x) = \frac{kx^2}{2} + \ldots$$

Thus, we are led to the conclusion in the statement, provided that we are indeed in the non-degenerate case, where $k \neq 0$, which amounts in saying that $F'(0) \neq 0$.

(2) This follows indeed from (1), and from $V' = -F$.

(3) This follows indeed from (2), and from $F = ma = m\ddot{x}$. □

The above result suggests formulating the following definition:

DEFINITION 6.4. *A harmonic oscillator is a particle moving as above, following*

$$m\ddot{x} + kx = 0$$

with $k \neq 0$. In the case $k > 0$, we say that we have a simple harmonic oscillator.

There the last convention comes from the fact that our oscillator is unstable when $k < 0$, and stable $k > 0$, and it is in this latter case that we are mostly interested in. And with this, stability depending on the sign of k, coming either from some abstract reasoning along the lines of Theorem 6.3, or from the explicit formulae below.

Very nice, so let us solve now the equation of motion. We have here:

THEOREM 6.5. *The solutions of the equation of motion $m\ddot{x} + kx = 0$ for the harmonic oscillators are as follows:*

(1) $x = ae^{pt} + be^{-pt}$ *with* $p = \sqrt{-k/m}$, *when* $k < 0$.
(2) $x = c\cos wt + d\sin wt$ *with* $w = \sqrt{k/m}$, *when* $k > 0$.

PROOF. This is standard mathematics, as follows:

(1) Assume first that we are in the case $k < 0$. Here, with $p = \sqrt{-k/m}$ as in the statement, the equation of motion takes the following form:

$$\ddot{x} = p^2 x$$

But the functions e^{pt}, e^{-pt} being solutions of this equation, by linearity we obtain that the solutions are exactly the linear combinations of these two functions, as claimed.

(2) Assume now that we are in the case $k > 0$. Here, with $w = \sqrt{k/m}$ as in the statement, the equation of motion takes the following form:

$$\ddot{x} = -w^2 x$$

But the functions $\cos wt, \sin wt$ being solutions, by linearity we obtain that the solutions are exactly the linear combinations of these two functions, as claimed. □

Observe that, as already mentioned above, the formulae that we obtained make it clear that our oscillator is unstable when $k < 0$, and stable when $k > 0$. In fact, we have the following simple consequences of the general formulae obtained above:

PROPOSITION 6.6. *The short and long time behavior of a harmonic oscillator, moving according to $m\ddot{x} + kx = 0$, are as follows:*

(1) *In the case $k < 0$, with $x = ae^{pt} + be^{-pt}$ as above, we have $x \simeq (a+b) + p(a-b)t$ for $t > 0$ small, and $x \simeq ae^{pt}$ for $t >> 0$.*
(2) *In the case $k > 0$, with $x = c\cos wt + d\sin wt$ as above, we have $x \simeq c + dwt$ for $t > 0$ small, and there is no asymptotics for $t >> 0$.*

PROOF. This is indeed standard mathematics based on Theorem 6.5, as follows:

(1) In the case $k < 0$, with $x = ae^{pt} + be^{-pt}$ as in Theorem 6.5, in the $t > 0$ small regime we have indeed the following estimate, coming from $e^z \simeq 1 + z$:

$$\begin{aligned} x &= ae^{pt} + be^{-pt} \\ &\simeq a(1 + pt) + b(1 - pt) \\ &= (a + b) + p(a - b)t \end{aligned}$$

As for the other estimate, namely $x \simeq ae^{pt}$ for $t >> 0$, this is clear.

(2) In the case $k > 0$, with $x = c\cos wt + d\sin wt$ as in Theorem 6.5, in the $t > 0$ small regime we have indeed the following estimate, coming from standard calculus:

$$\begin{aligned} x &= c\cos wt + d\sin wt \\ &\simeq c(1 + o(t)) + dwt \\ &\simeq c + dwt \end{aligned}$$

As for the last assertion, regarding the lack of asymptotics at $k > 0$ in the $t >> 0$ regime, this is clear, because neither cos, nor sin have such asymptotics, and the same happens for any linear combination of them, with non-trivial coefficients. Of course, interesting exercice for you to figure out all this, abstractly, this being not hard. □

As a last piece of mathematics, using this time complex numbers, we have:

THEOREM 6.7. *The solutions of the equation $m\ddot{x} + kx = 0$ are as follows, regardless of the sign of k, and with $a, b, c, d \in \mathbb{C}$ chosen as to have $x \in \mathbb{R}$:*

(1) $x = ae^{pt} + be^{-pt}$, *with* $p = \sqrt{-k/m}$.
(2) $x = c\cos wt + d\sin wt$, *with* $w = \sqrt{k/m}$.

PROOF. This is standard complex number business, the idea being as follows:

(1) As before in the proof of Theorem 6.5 (1), but by using this time complex numbers, we are led to the conclusion in the statement. With two remarks, namely:

– In the case $k < 0$ we have $p \in \mathbb{R}$, and so $a, b \in \mathbb{R}$, and we recover in this way Theorem 6.5 (1) itself.

– As for the case $k > 0$, here we can write $p = iw$ with $w = \sqrt{k/m} \in \mathbb{R}$, and the formula that we get, according to the above, is as follows:

$$x = ae^{iwt} + be^{-iwt}$$

Now in order to have $x \in \mathbb{R}$, which is the same as saying that $x = \bar{x}$, we need:

$$a = \bar{b}$$

Thus we can write $a = c - id$, $b = c + id$ with $c, d \in \mathbb{R}$, and with these substitutions made, the solution found above takes the following form:

$$\begin{aligned} x &= ae^{iwt} + be^{-iwt} \\ &= (c - id)(\cos wt + i \sin wt) + (c + id)(\cos wt - i \sin wt) \\ &= 2(c \cos wt + d \sin wt) \end{aligned}$$

Thus at $k > 0$, up to a 2 factor, we obtain the formula from Theorem 6.5 (2).

(2) Things are similar here. Indeed, as before in the proof of Theorem 6.5 (2), we are led to the conclusion in the statement, and with two remarks to be made, namely:

– In the case $k > 0$ we have $w \in \mathbb{R}$, and so $c, d \in \mathbb{R}$, and we recover in this way Theorem 6.5 (2) itself.

– As for the case $k < 0$, here we can write $w = -ip$ with $p = \sqrt{-k/m} \in \mathbb{R}$, and the formula that we get, according to the above, is as follows:

$$\begin{aligned} x &= c \cos wt + d \sin wt \\ &= c \cos(-ipt) + d \sin(-ipt) \\ &= c \cos(ipt) - d \sin(ipt) \\ &= c \cdot \frac{e^{i(ipt)} + e^{-i(ipt)}}{2} - d \cdot \frac{e^{i(ipt)} - e^{-i(ipt)}}{2i} \\ &= c \cdot \frac{e^{-pt} + e^{pt}}{2} - d \cdot \frac{e^{-pt} - e^{pt}}{2i} \\ &= \frac{1}{2}\left(\left(c + \frac{d}{i}\right)e^{pt} + \left(c - \frac{d}{i}\right)e^{-pt}\right) \end{aligned}$$

Now observe that in order to have $x \in \mathbb{R}$, we must have $c \pm d/i \in \mathbb{R}$. Thus $c \in \mathbb{R}$, and $d = if$ with $f \in \mathbb{R}$, and with this latter substitution made, and then aftwerwards with the notations $a = (c + f)/2$ and $b = (c - f)/2$, we obtain:

$$\begin{aligned} x &= \frac{1}{2}\left((c + f)e^{pt} + (c - f)e^{-pt}\right) \\ &= ae^{pt} + be^{-pt} \end{aligned}$$

Thus at $k < 0$, we obtain the formula from Theorem 6.5 (1). □

Many other things can be said about the harmonic oscillators, in complement to what was said above, and we will be back to this, on a regular basis, in what follows.

6c. Kepler and Newton

Back to gravity, in two dimensions now, we first have the following result:

THEOREM 6.8. *In the context of a free fall from distance $x_0 = R >> 0$, with initial velocity $v_0 = 0$, the equation of the trajectory is*

$$x \simeq R - \frac{gt^2}{2}$$

with the constant being $g = GM/R^2$, called gravity of M, at distance R from it.

PROOF. We can use here the field equation for the gravity, namely:

$$f = \frac{K}{d^2}$$

This equation, with $d = ||x||$, describes the magnitude f of the acceleration a of our moving object m. Now since a points towards 0, which is opposite to x, we have:

$$a = -\frac{K}{d^2} \cdot \frac{x}{||x||} = -\frac{Kx}{||x||^3}$$

Moreover, since the acceleration a is by definition the second derivative of the position vector x, the equation of motion of our object m is as follows:

$$\ddot{x} = -\frac{Kx}{||x||^3}$$

In one dimension now, things get simpler, and the equation of motion reads:

$$\ddot{x} = -\frac{K}{x^2}$$

Since we assumed $R >> 0$, we must look for a solution of type $x \simeq R + ct^2$, with the lack of the t term coming from $v_0 = 0$. But with $x \simeq R + ct^2$, our equation reads:

$$2c \simeq -\frac{K}{R^2}$$

Now by multiplying by $t^2/2$, and adding R, we obtain as solution:

$$x \simeq R - \frac{Kt^2}{2R^2}$$

Thus, we have indeed $x \simeq R - gt^2/2$, with g being the following number:

$$g = \frac{K}{R^2} = \frac{GM}{R^2}$$

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

As an illustration for the above basic result, let us do a numeric terrestrial check, based on it. The gravitational constant, the mass of the Earth, and the average radius of the Earth are as follows, expressed as usual in meters and kilograms:

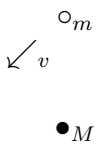
$$G = 6.674 \times 10^{-11} \quad , \quad M = 5.972 \times 10^{24} \quad , \quad R = 6.371 \times 10^{6}$$

We obtain the following value for the number g computed above:

$$g = \frac{6.674 \times 5.972}{6.371 \times 6.371} \times 10 = 9.819$$

Which is quite decent, when compared to the observed value, $g = 9.806$.

As a second toy example now for our 3D gravitation theory, which is more advanced, lying somewhere between 1D and 2D, let us add an arbitrary initial speed $v_0 = v$ to the above situation, which in addition is allowed to be a vector in $\mathbb{R}^2$, as follows:



We obtain in this way the following generalization of Theorem 6.8:

THEOREM 6.9. *In the context of a free fall from distance $x_0 = R >> 0$, with initial plane velocity vector $v_0 = v$, the equation of the trajectory is*

$$x \simeq R + vt - \frac{gt^2}{2}$$

where $g = GM/R^2$ as usual, and with the quantities R, g in the above being regarded now as vectors, pointing upwards. The approximate trajectory is a parabola.

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

(1) Let us first discuss the simpler case where we are still in 1D, as in Theorem 6.8, but with an initial velocity $v_0 = v$ added. In order to find the equation of motion, we can just redo the computations from the proof of Theorem 6.8, with now looking for a general solution of type $x \simeq R + vt + ct^2$, and we get, as stated above:

$$x \simeq R + vt - \frac{gt^2}{2}$$

Alternatively, we can simply argue that, by linearity, what we have to do is to take the solution $x \simeq R - gt^2/2$ found in Theorem 6.8, and add an extra vt term to it.

(2) In the general 2D case now, where the initial velocity $v_0 = v$ is a vector in $\mathbb{R}^2$, the same arguments apply, either by redoing the computations from the proof of Theorem 6.8, or simply by arguing that by linearity we can just take the solution $x \simeq R - gt^2/2$ found there, and add an extra vt term to it. Thus, we have our solution.

(3) Let us study now the solution that we found. In standard (x, y) coordinates, with $v = (p, q)$, and with R, g being now back scalars, our solution looks as follows:

$$x = pt \quad , \quad y \simeq R + qt - \frac{gt^2}{2}$$

From the first equation we get $t = x/p$, and by substituting into the second:

$$y \simeq R + \frac{qx}{p} - \frac{gx^2}{2p^2}$$

We recognize here the approximate equation of a parabola, and we are done. □

Getting now to the real thing, astronomy, the result here, which is the pride of mathematics, physics, and human knowledge in general, is the following theorem:

THEOREM 6.10 (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) 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:

$$a\cos\theta - b\sin\theta = -\frac{K}{r^2}\cos\theta \quad , \quad a\sin\theta + b\cos\theta = -\frac{K}{r^2}\sin\theta$$

But these two formulae can be written in the following way:

$$\left(a+\frac{K}{r^2}\right)\cos\theta=b\sin\theta\quad,\quad\left(a+\frac{K}{r^2}\right)\sin\theta=-b\cos\theta$$

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:

$$r(t) = \frac{1}{f(\theta(t))}$$

With the convention that dots mean as usual derivatives with respect to t, and that the primes will denote derivatives with respect to $\theta = \theta(t)$, we have:

$$\dot{r} = -\frac{f'\dot{\theta}}{f^2} = -\frac{f'}{f^2}\cdot\frac{\lambda}{r^2} = -\lambda f'$$

By differentiating one more time with respect to t, we obtain:

$$\ddot{r} = -\lambda f''\dot{\theta} = -\lambda f''\cdot\frac{\lambda}{r^2} = -\frac{\lambda^2}{r^2}f''$$

On the other hand, our equation for $\ddot{r}$ found in (7) reads:

$$\ddot{r} = \frac{\lambda^2}{r^2}\left(\frac{1}{r} - \frac{1}{c}\right) = \frac{\lambda^2}{r^2}\left(f - \frac{1}{c}\right)$$

Thus, in terms of $f = 1/r$ as above, our equation for $\ddot{r}$ simply reads:

$$f'' + f = \frac{1}{c}$$

But this latter equation is elementary to solve. Indeed, both functions $\cos t, \sin t$ satisfy $g" + g = 0$, so any linear combination of them satisfies as well this equation. But the solutions of $f'' + f = 1/c$ being those of $g'' + g = 0$ shifted by $1/c$, we obtain:

$$f = \frac{1 + \varepsilon\cos\theta + \delta\sin\theta}{c}$$

Now by inverting, we obtain the formula announced in (9), namely:

$$r = \frac{c}{1 + \varepsilon\cos\theta + \delta\sin\theta}$$

(11) But this leads to the conclusion that the trajectory is a conic. Indeed, in terms of the parameter θ, the formulae of the coordinates are:

$$x = \frac{c\cos\theta}{1 + \varepsilon\cos\theta + \delta\sin\theta}$$

$$y = \frac{c\sin\theta}{1 + \varepsilon\cos\theta + \delta\sin\theta}$$

But these are precisely the equations of conics in polar coordinates.

(12) To be more precise, in order to find the precise equation of the conic, observe that the two functions x, y that we found above satisfy the following formula:

$$\begin{aligned} x^2 + y^2 &= \frac{c^2(\cos^2\theta + \sin^2\theta)}{(1 + \varepsilon\cos\theta + \delta\sin\theta)^2} \\ &= \frac{c^2}{(1 + \varepsilon\cos\theta + \delta\sin\theta)^2} \end{aligned}$$

On the other hand, these two functions satisfy as well the following formula:

$$\begin{aligned} (\varepsilon x + \delta y - c)^2 &= \frac{c^2\big(\varepsilon\cos\theta + \delta\sin\theta - (1 + \varepsilon\cos\theta + \delta\sin\theta)\big)^2}{(1 + \varepsilon\cos\theta + \delta\sin\theta)^2} \\ &= \frac{c^2}{(1 + \varepsilon\cos\theta + \delta\sin\theta)^2} \end{aligned}$$

We conclude that our coordinates x, y satisfy the following equation:

$$x^2 + y^2 = (\varepsilon x + \delta y - c)^2$$

But what we have here is an equation of a conic, as claimed. □

The above was theory, and for further applications, here is a sort of "best of" the formulae found in the proof of Theorem 6.10, which are all very useful in practice:

THEOREM 6.11 (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 6.10. 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 6.10. □

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 6.12. *In the context of Theorem 6.11, 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 6.11, the object m begins its movement on Ox. Thus we have $\theta_0 = 0$, and from this we get the formula of r_0 in the statement.

(2) Regarding the initial speed now, the formula of $\dot{\theta}_0$ follows from:

$$\dot{\theta} = \frac{\lambda}{r^2} = \frac{\sqrt{Kc}}{r^2}$$

Also, in what concerns the radial speed, the formula of $\dot{r}_0$ follows from:

$$\begin{aligned}
\dot{r} &= \frac{c(\varepsilon \sin\theta - \delta\cos\theta)\dot{\theta}}{(1 + \varepsilon\cos\theta + \delta\sin\theta)^2} \\
&= \frac{c(\varepsilon \sin\theta - \delta\cos\theta)}{c^2/r^2} \cdot \frac{\sqrt{Kc}}{r^2} \\
&= \frac{\sqrt{K}(\varepsilon \sin\theta - \delta\cos\theta)}{\sqrt{c}}
\end{aligned}$$

(3) Regarding now the initial acceleration, by using $\dot{\theta} = \sqrt{Kc}/r^2$ we find:

$$\ddot{\theta} = -2\sqrt{Kc} \cdot \frac{2r\dot{r}}{r^3} = -\frac{4\sqrt{Kc} \cdot \dot{r}}{r^2}$$

In particular at $t = 0$ we obtain the formula in the statement, namely:

$$\ddot{\theta}_0 = -\frac{4\sqrt{Kc} \cdot \dot{r}_0}{R^2} = \frac{4\sqrt{Kc}}{R^2} \cdot \frac{\delta\sqrt{K}}{\sqrt{c}} = \frac{4\delta K}{R^2}$$

(4) Also regarding acceleration, with $\lambda = \sqrt{Kc}$ our main motion formula reads:

$$\ddot{r} = \frac{Kc}{r^2}\left(\frac{1}{r} - \frac{1}{c}\right)$$

In particular at $t = 0$ we obtain the formula in the statement, namely:

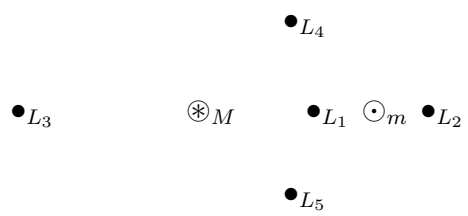
$$\ddot{r}_0 = \frac{Kc}{R^2}\left(\frac{1}{R} - \frac{1}{c}\right) = \frac{Kc}{R^2} \cdot \frac{\varepsilon}{c} = \frac{\varepsilon K}{R^2}$$

(5) Finally, the last assertion is clear, and since the formulae look better anyway in polar coordinates than in affine coordinates, we will not get into details here. □

With the above formulae in hand, which are a precious complement to Theorem 6.11, we can do some reverse engineering at the level of parameters, and work out how various inital speeds and accelerations lead to various types of conics. There are many things that can be said here, and we refer here to any standard mechanics book.

Finally, a word about the 3-body problem. An interesting question here is how to position a specialized scientific satellite, deep in space, and away from the dust and

radiation of the usual orbits around the Earth, as to stay there, under the joint influence of the gravity of the Sun M and of the Earth m. And there are 5 possible solutions here, called Lagrange points L1-L5, whose positions with respect to M, m are as follows:



Moreover, and here comes another interesting point, L4, L5 are stable, in the sense that a satellite installed there will really stay there, regardless of the various tiny little things that might happen, like an asteroid passing by, while L1, L2, L3 are unstable, in the sense that a satellite installed there will need constant tiny adjustments, in order to really stay there. So, which one would you choose for installing your satellite?

You would probably say L4, L5, but this is precisely the wrong answer, because due to their stability, these points attract a lot of asteroids and space garbage, and our satellite will certainly not perform well there, in that crowd. So, with L4, L5 ruled out, and with L3 ruled out too, being too far, the correct choices are L1, L2. But here, due to instability, you still need to learn a lot more mechanics, for knowing how to do this, in practice.

6d. Conservative forces

Let us discuss now an important topic, namely the conservation of energy. The simplest situation is that of a free fall with initial velocity $v_0 = 0$, and we have here:

PROPOSITION 6.13. *In the context of a free fall from distance $x_0 = R >> 0$, with initial velocity $v_0 = 0$, if we define the potential energy to be*

$$V = mgx$$

then the total energy $E = T + V$, with $T = mv^2/2$ as usual, is constant, $E \simeq mgR$.

PROOF. We know that the equation of motion is as follows, with $g = GM/R^2$:

$$x \simeq R - \frac{gt^2}{2}$$

The kinetic energy, from now on to be denoted T, is then given by:

$$T \simeq \frac{mv^2}{2} = \frac{mg^2t^2}{2}$$

Thus with $V = mgx$ as in the statement, and then with $E = T + V$, we have:

$$E = T + V \simeq mgR$$

But this is a constant, and so we have our conservation principle, as desired. □

We know that $E \simeq mgR$, but by some kind of miracle, do we actually have $E = mgR$? Also, what is the meaning of V? What about the meaning of E? What about adding a suitable constant to V, and so to E too, will that make these quantities easier to understand? These questions will be answered in due time. As a next result, we have:

THEOREM 6.14. *In the context of a free fall from distance $x_0 = R >> 0$, with initial velocity vector $v_0 \in \mathbb{R}^2$, if we define the potential energy to be*

$$V = m < g, x >$$

with $g = GM/R^2$ being regarded as usual as a vector pointing upwards, then

$$E = T + V$$

with $T = m||v||^2/2$ as usual, is constant, $E \simeq T_0 + mgR$, with g now back scalar.

PROOF. We can do this in two steps, first by adding an extra parameter to the computation in Proposition 6.13, and then by adding another extra parameter:

(1) Let us first examine the 1D case, where $v_0 = s$ is a vector aligned to x, and so a number. Here the equation of motion is as follows, with $g = GM/R^2$ as usual:

$$x \simeq R + st - \frac{gt^2}{2}$$

The speed being $v \simeq s - gt$, with $V = mgx$ and $E = T + V$ as above, we have:

$$\begin{aligned} E &= T + V \\ &\simeq \frac{m(s-gt)^2}{2} + mg\left(R + st - \frac{gt^2}{2}\right) \\ &= \frac{ms^2}{2} + mgR \\ &= T_0 + mgR \end{aligned}$$

(2) In the general case now, with $v_0 = s$, the equation of motion is as before, with R, g being now vectors pointing upwards, and if we write $s = (a, b)$, then we have:

$$\begin{aligned} T &\simeq \frac{m||s-gt||^2}{2} \\ &= \frac{m((a-gt)^2 + b^2)}{2} \\ &= \frac{m(a^2+b^2)}{2} - magt + \frac{mg^2t^2}{2} \\ &= T_0 - mg\left(at - \frac{gt^2}{2}\right) \end{aligned}$$

With g vector pointing upwards, the last quantity is $m < g, x - R >$, so if we add $V = m < g, x >$, we obtain $T_0 + mgR$, with g, R being back scalars, as desired. □

With the above done, let us get back to the real thing, 3D gravity. We are interested in the general 2-body problem, where M is fixed at 0, and m moves under the gravitational force of M. The above computations, coming from our "kinetic energy gets converted into height, and vice versa" principle, suggest defining the potential energy as:

$$V \sim ||x||$$

However, this is wrong, because in our formula $V = mgx$ the quantity $g = GM/R^2$ depends on the average height, which is the parameter R, no longer assumed to satisfy $R >> 0$. In view of this, the correct formula for the potential energy should be:

$$V \sim \frac{1}{||x||}$$

In order now to find the constant, it is enough to rewrite $V = mgx$ by getting rid of the parameter $g = GM/R^2$. We obtain in this way, with $K = GM$ as usual:

$$V = mgx = \frac{mGMx}{R^2} \simeq \frac{mGM}{||x||} = \frac{Km}{||x||}$$

Thus, we have our formula for V, and the question now is if $E = T+V$ is constant. And the answer here is unfortunately no, due to some bizarre reasons, with rather $E = T - V$ appearing to be constant, or at least that's what computations tend to suggest.

So, let us simply change the sign of V, and see what we get. We are led in this way to the following remarkable result, which not only says that E is approximately constant, as in our previous computations, but is actually a plain constant:

THEOREM 6.15. *In the context of the 2-body problem, with M fixed at 0 and with m moving, if we define the kinetic and potential energy of m to be*

$$T = \frac{m||v||^2}{2} \quad , \quad V = -\frac{Km}{||x||}$$

with $K = GM$ as usual, then the total energy $E = T + V$ is constant.

PROOF. The idea will be that of proving $\dot{E} = 0$. We can do this as follows:

(1) In what regards the derivative of T, the computation here is something very simple, coming straight from the formula $||v||^2 =< v, v >$, as follows:

$$\begin{aligned} \dot{T} &= \frac{m(< v, \dot{v} > + < \dot{v}, v >)}{2} \\ &= m < v, \dot{v} > \\ &= m < v, a > \end{aligned}$$

(2) In order to compute now the derivative of V, let us first compute the derivative of the function $f(x) = 1/||x||$. Again by using $||x||^2 =< x, x >$, we obtain:

$$\begin{aligned}\dot{f} &= -\frac{1}{2}\cdot\frac{< x, \dot{x} > + < \dot{x}, x >}{< x, x >^{3/2}}\\ &= -\frac{< x, \dot{x} >}{< x, x >^{3/2}}\\ &= -\frac{< x, v >}{||x||^3}\end{aligned}$$

(3) Thus, getting now to the potential energy V, we have the following formula:

$$\dot{V} = \frac{Km < x, v >}{||x||^3}$$

In order to further process this, remember the equation of motion of m, namely:

$$a = -\frac{Kx}{||x||^3}$$

We will of course jump on this, as to get rid of $||x||^3$, and we finally obtain:

$$\dot{V} = -m < a, v >$$

(4) We are ready now to prove our result. Indeed, we have:

$$\dot{E} = \dot{T} + \dot{V} = m < v, a > -m < a, v >= 0$$

Now since the derivative vanishes, E is constant, as claimed. □

Nice all this, but we still have to understand the relation with Proposition 6.13 and Theorem 6.14, with that sign of V mysteriously switching. And we have here the following result, upgrading Proposition 6.13 and Theorem 6.14, and clarifying the whole thing:

THEOREM 6.16. *In the context of a free fall from distance $x_0 = R >> 0$, with initial velocity $v_0 = 0$, if we define the kinetic and potential energy of m to be*

$$T = \frac{mv^2}{2} \quad , \quad V = -\frac{Km}{x}$$

with $K = GM$ as usual, then the total energy $E = T + V$ is constant. Moreover,

$$V \simeq mgx - 2mgR$$

with $g = GM/R^2$, and so $E' = T + mgx$ is appoximately constant, $E' \simeq mgR$. The same happens for a free fall from $x_0 = R >> 0$, with initial velocity vector $v_0 \in \mathbb{R}^2$.

PROOF. The first assertion is something that we know, coming from Theorem 6.15. In order to clarify now the relation with Proposition 6.13, we first have:

$$V = -\frac{Km}{x} = -\frac{GMm}{x} = -\frac{mgR^2}{x}$$

Now by writing $x = R(1-\varepsilon)$, we obtain the estimate in the statement, namely:

$$\begin{aligned} V &= -\frac{mgR}{1-\varepsilon} \\ &\simeq -mgR(1+\varepsilon) \\ &= mgR[(1-\varepsilon)-2] \\ &= mgx - 2mgR \end{aligned}$$

Thus with $V' = mgx$ we have $V \simeq V' - 2mgR$, and so $E' = T + V'$ satisfies:

$$\begin{aligned} E' &\simeq E + 2mgR \\ &= E_0 + 2mgR \\ &= V_0 + 2mgR \\ &= mgR \end{aligned}$$

Finally, the last assertion, which is a bit more general, follows in the same way. □

6e. Exercises

Exercises:

Exercise 6.17.

Exercise 6.18.

Exercise 6.19.

Exercise 6.20.

Exercise 6.21.

Exercise 6.22.

Exercise 6.23.

Exercise 6.24.

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:

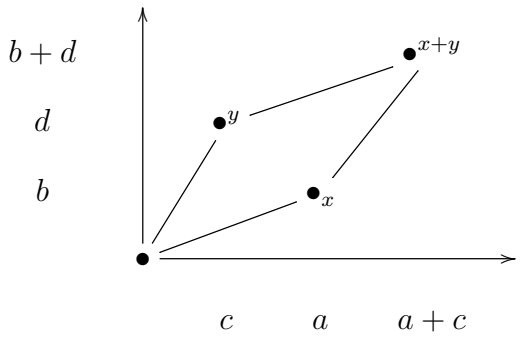
$$x = \begin{pmatrix} a \\ b \end{pmatrix} \quad , \quad y = \begin{pmatrix} c \\ d \end{pmatrix}$$

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. □

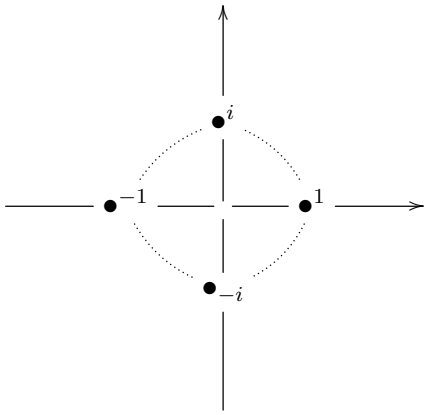
Here we have assumed that you are a bit familiar with vector calculus. If not, no problem, the idea is simply that vectors add by forming a parallelogram, as follows:



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,

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

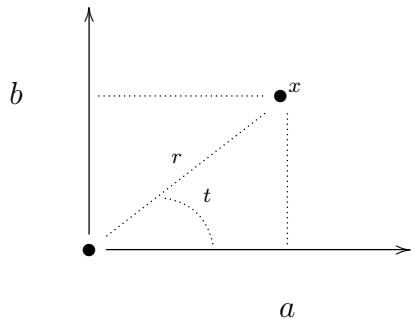
and in the other sense being as follows,

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

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. □

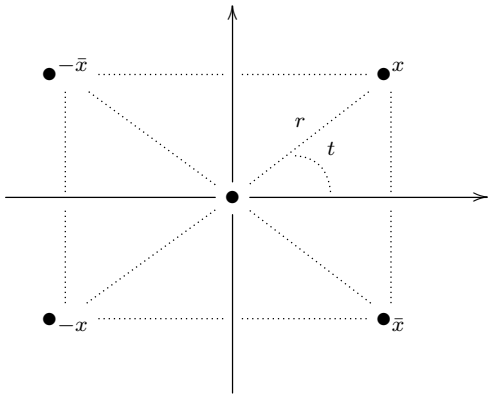
As a last general topic regarding the complex numbers, let us discuss conjugation. This is something quite tricky, complex number specific, as follows:

DEFINITION 7.7. *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 7.8. *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 7.7.

(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. □

7b. Exponential writing

Welcome to complex analysis. Let us start with:

DEFINITION 7.9. *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 book.

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

THEOREM 7.10. *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 7.11. *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. □

Getting back now to less mysterious mathematics, as an application, let us discuss the final and most convenient writing of the complex numbers, which is as follows:

$$x = re^{it}$$

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

THEOREM 7.12. *We have the following formula,*

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

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

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

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

Indeed, let us first recall from Theorem 5.13 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. □

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

THEOREM 7.13. *The complex numbers $x = a + ib$ can be written in polar coordinates,*

$$x = re^{it}$$

with the connecting formulae being

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

and in the other sense being

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

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

PROOF. This is a reformulation of our previous Definition 7.5, by using the formula $e^{it} = \cos t + i\sin t$ from Theorem 7.12, and multiplying everything by r. □

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

THEOREM 7.14. *In polar coordinates, the complex numbers multiply as*

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

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

PROOF. This is something that we already know, from Theorem 7.6, reformulated by using the notations from Theorem 7.13. 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. □

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

THEOREM 7.15. *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, but we can now discuss all this, from a more conceptual viewpoint, the idea being as follows:

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

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

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

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

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

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

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

(4) This comes from the fact, that we know from Theorem 7.8, that the conjugation operation $x \to \bar{x}$ keeps the modulus, and switches the sign of the argument. □

7c. Equations, roots

Getting back to algebra, recall from Theorem 7.2 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 7.16. *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 7.17. *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:

$$\begin{aligned} R(P,Q) &= ad^2(p-r)(q-r) \\ &= ad^2(pq-(p+q)r+r^2) \\ &= cd^2 + bd^2r + ad^2r^2 \\ &= cd^2 - bde + ae^2 \end{aligned}$$

Finally, observe that $R(P,Q) = 0$ corresponds indeed to the fact that P, Q have a common root. Indeed, the root of Q is $r = -e/d$, and we have:

$$\begin{aligned} P(r) &= \frac{ae^2}{d^2} - \frac{be}{d} + c \\ &= \frac{R(P,Q)}{d^2} \end{aligned}$$

Thus $P(r) = 0$ precisely when $R(P,Q) = 0$, as predicted by Theorem 7.17.

Good news, with the above resultant technology in hand, we can now talk about the discriminant of any polynomial, as follows:

THEOREM 7.18. *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 7.17, 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 7.18 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 7.19. *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 7.20. *For a normalized degree 3 equation, namely*

$$x^3 + 3px + 2q = 0$$

the discriminant is $\Delta = -108(p^3 + q^2)$. Assuming $p, q \in \mathbb{R}$ and $\Delta < 0$, the number

$$x = \sqrt[3]{-q + \sqrt{p^3 + q^2}} + \sqrt[3]{-q - \sqrt{p^3 + q^2}}$$

is a real solution of our equation.

PROOF. The formula of Δ is clear from definitions, and with $108 = 4 \times 27$. Now with x as in the statement, by using $(a + b)^3 = a^3 + b^3 + 3ab(a + b)$, we have:

$$\begin{aligned} x^3 &= \left(\sqrt[3]{-q + \sqrt{p^3 + q^2}} + \sqrt[3]{-q - \sqrt{p^3 + q^2}}\right)^3 \\ &= -2q + 3\sqrt[3]{-q + \sqrt{p^3 + q^2}} \cdot \sqrt[3]{-q - \sqrt{p^3 + q^2}} \cdot x \\ &= -2q + 3\sqrt[3]{q^2 - p^3 - q^2} \cdot x \\ &= -2q - 3px \end{aligned}$$

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

There are many more things that can be said about degree 3 equations, along these lines, and we will certainly have an exercise about this, at the end of this chapter.

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

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

which are called roots of unity of order N.

PROOF. This follows from the general multiplication formula for complex numbers from Theorem 7.14. Indeed, with $x = re^{it}$ our equation reads:

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

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

As an illustration here, the roots of unity of small order, 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, from Theorem 7.15, the statement being as follows:

THEOREM 7.22. *Any nonzero complex number, written as*

$$x = re^{it}$$

has exactly N roots of order N, which appear as

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

multiplied by the N roots of unity of order N.

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

$$z^N = y^N$$

Now observe that we can write this equation as follows:

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

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

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

THEOREM 7.23. *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.

This was for our basic presentation of the complex numbers. We will be back to more theory regarding them, and the roots of unity, later on. Among others, we will see later some non-trivial applications of our above solution to $x_1 + \ldots + x_N = 0$.

7d. 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 7.24. *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 7.25. *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 7.26. *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 7.27. *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 7.28. *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, 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 7.29. *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 7.30. *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 7.31. *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 7.32. *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 7.33. *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. □

7e. Exercises

Exercises:

EXERCISE 7.34.

EXERCISE 7.35.

EXERCISE 7.36.

EXERCISE 7.37.

EXERCISE 7.38.

EXERCISE 7.39.

EXERCISE 7.40.

EXERCISE 7.41.

Bonus exercise.

CHAPTER 8

Light and heat

8a. Electrostatics

Let us develop now the basic mathematics for electrostatics. We first have:

DEFINITION 8.1. *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 $F = QE$.

Observe the analogy with gravity, save for the fact that instead of masses $m > 0$ we have now charges $q \in \mathbb{R}$, and that at the level of constants, G gets replaced by K.

More generally, we will be interested in electric fields of various non-discrete configurations of charges, such as charged curves, surfaces and solid bodies. We have already talked about such things in the above, in the gravitational context, but the discussion there, involving the gravitational force of a solid body having non-trivial shape or density, was something rather specialized.

In the electricity context, however, 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 8.2. *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:

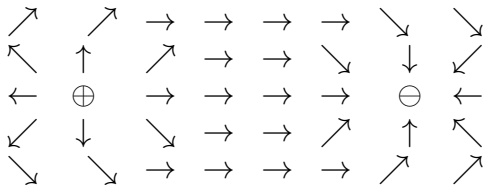
DEFINITION 8.3. *The field lines of $E : \mathbb{R}^3 \to \mathbb{R}^3$ are the oriented curves*

$$\gamma \subset \mathbb{R}^3$$

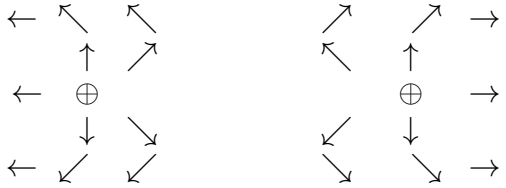
pointing at every point $x \in \mathbb{R}^3$ at the direction of the field, $E(x) \in \mathbb{R}^3$.

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.

In general now, the first thing that can be said about the field lines is that, by definition, they do not cross. Thus, what we have here is some sort of oriented 1D foliation of $\mathbb{R}^3$, in the sense that $\mathbb{R}^3$ is smoothly decomposed into oriented curves $\gamma \subset \mathbb{R}^3$.

The field lines, as constructed in Definition 8.3, 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 8.4. *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. We have two assertions here, the idea being as follows:

(1) The first assertion is clear from definitions, with of course our usual convention that the electric field and its problematics take place outside the locations of the charges, which makes everything in the statement to be indeed well-defined.

(2) Regarding now the last assertion, which is of course a bit informal, this follows from the above discussion. It is possible to be a bit more mathematical here, with a definition, formula and everything, but we will not need this, in what follows. □

Let us introduce now a key definition, as follows:

DEFINITION 8.5. *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.

Regarding the last sentence of Definition 8.5, this is of course something informal, meant to help, coming from the interpretation of the field lines from Proposition 8.4. However, we will see later that this simple interpretation can be of great use.

As a first observation, we could have done of course the same thing with gravity before, but these notions of field lines and flux are not very interesting, in that context.

In the present setting, however, electric fields passing through metal sheets are a common occurence, and all the above is important, for any application.

As a first illustration, let us do a basic computation, as follows:

PROPOSITION 8.6. *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:

$$\begin{aligned}
\Phi_E(S) &= \int_S < E(x), n(x) > dx \\
&= \int_S \left\langle \frac{Kqx}{r^3}, \frac{x}{r} \right\rangle dx \\
&= \int_S \frac{Kq}{r^2}\, dx \\
&= \frac{Kq}{r^2} \times 4\pi r^2 \\
&= 4\pi K q
\end{aligned}$$

Thus with $\varepsilon_0 = 1/(4\pi K)$ as above, we obtain the result. □

As a comment here, the constant $\varepsilon_0 = 1/(4\pi K)$ which appears in the above is the permittivity of free space constant that we talked about before, when discussing units. In what follows we will use this new constant instead of the Coulomb constant K.

More generally now, we have the following result:

THEOREM 8.7. *The flux of a field E through a sphere S is given by*

$$\Phi_E(S) = \frac{Q_{enc}}{\varepsilon_0}$$

where Q_{enc} is the total charge enclosed by S, and $\varepsilon_0 = 1/(4\pi K)$.

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

(1) Before jumping into computations, let us do some manipulations. First, by discretizing the problem, we can assume that we are dealing with a system of point charges. Moreover, by additivity, we can assume that we are dealing with a single charge. And if we denote by $q \in \mathbb{R}$ this charge, located at $v \in \mathbb{R}^3$, we want to prove that we have the following formula, where $B \subset \mathbb{R}^3$ denotes the ball enclosed by S:

$$\Phi_E(S) = \frac{q}{\varepsilon_0}\, \delta_{v \in B}$$

(2) By linearity we can assume that we are dealing with the unit sphere S. Moreover, by rotating we can assume that our charge q lies on the Ox axis, that is, that we have $v = (r, 0, 0)$ with $r \geq 0$, $r \neq 1$. The formula that we want to prove becomes:

$$\Phi_E(S) = \frac{q}{\varepsilon_0}\,\delta_{r<1}$$

(3) Let us start now the computation. With $u = (x, y, z)$, we have:

$$\begin{aligned}
\Phi_E(S) &= \int_S < E(u), u > du \\
&= \int_S \left\langle \frac{Kq(u-v)}{||u-v||^3}, u \right\rangle du \\
&= Kq \int_S \frac{< u-v, u >}{||u-v||^3}\, du \\
&= Kq \int_S \frac{1 - < v, u >}{||u-v||^3}\, du \\
&= Kq \int_S \frac{1 - rx}{(1 - 2xr + r^2)^{3/2}}\, du
\end{aligned}$$

(4) In order to compute the above integral, we will use spherical coordinates for the unit sphere S, which are as follows, with $s \in [0, \pi]$ and $t \in [0, 2\pi]$:

$$\begin{cases}
x &= \cos s \\
y &= \sin s \cos t \\
z &= \sin s \sin t
\end{cases}$$

The corresponding Jacobian is readily computed, as follows:

$$\begin{aligned}
J &= \begin{vmatrix} \cos s & -\sin s & 0 \\ \sin s \cos t & \cos s \cos t & -\sin s \sin t \\ \sin s \sin t & \cos s \sin t & \sin s \cos t \end{vmatrix} \\
&= \sin s \sin t \begin{vmatrix} \cos s & -\sin s \\ \sin s \sin t & \cos s \sin t \end{vmatrix} + \sin s \cos t \begin{vmatrix} \cos s & -\sin s \\ \sin s \cos t & \cos s \cos t \end{vmatrix} \\
&= \sin s(\sin^2 t + \cos^2 t) \begin{vmatrix} \cos s & -\sin s \\ \sin s & \cos s \end{vmatrix} \\
&= \sin s
\end{aligned}$$

(5) With the above change of coordinates, our integral from (3) becomes:

$$\begin{aligned}\Phi_E(S) &= Kq\int_S\frac{1-rx}{(1-2xr+r^2)^{3/2}}\,du\\ &= Kq\int_0^{2\pi}\int_0^{\pi}\frac{1-r\cos s}{(1-2r\cos s+r^2)^{3/2}}\cdot\sin s\,ds\,dt\\ &= 2\pi Kq\int_0^{\pi}\frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}}\,ds\\ &= \frac{q}{2\varepsilon_0}\int_0^{\pi}\frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}}\,ds\end{aligned}$$

(6) The point now is that the integral on the right can be computed with the change of variables $x=\cos s$. Indeed, we have $dx=-\sin s\,ds$, and we obtain:

$$\begin{aligned}\int_0^{\pi}\frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}}\,ds &= \int_{-1}^{1}\frac{1-rx}{(1-2rx+r^2)^{3/2}}\,dx\\ &= \left[\frac{x-r}{\sqrt{1-2rx+r^2}}\right]_{-1}^{1}\\ &= \frac{1-r}{\sqrt{1-2r+r^2}}-\frac{-1-r}{\sqrt{1+2r+r^2}}\\ &= \frac{1-r}{|1-r|}+1\\ &= 2\delta_{r<1}\end{aligned}$$

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

As a comment here, at $r=1$, which is normally avoided by our problematics, the integral I_r computed in (5) above converges too, and can be evaluated as follows:

$$I_1=\left[\frac{x-1}{\sqrt{2-2x}}\right]_{-1}^{1}=\left[-\sqrt{\frac{1-x}{2}}\ \right]_{-1}^{1}=1$$

Thus, we have the correct middle step between the $0,2$ values of the integral I_r, and getting back now to the flux, at $r=1$ we formally have $\Phi_E(S)=q/(2\varepsilon_0)$, which again is the correct middle step between the $0,q/\varepsilon_0$ values of the flux.

Even more generally now, we have the following result, due to Gauss, which is the foundation of advanced electrostatics, and of everything following from it, namely electrodynamics, and then quantum mechanics, and particle physics:

THEOREM 8.8 (Gauss law). *The flux of a field E through a surface S is given by*

$$\Phi_E(S) = \frac{Q_{enc}}{\varepsilon_0}$$

where Q_{enc} is the total charge enclosed by S, and $\varepsilon_0 = 1/(4\pi K)$.

PROOF. This basically follows from Theorem 8.7, or even from Proposition 8.6, by adding to the results there a number of new ingredients, as follows:

(1) Our first claim is that given a closed surface S, with no charges inside, the flux through it of any choice of external charges vanishes:

$$\Phi_E(S) = 0$$

This claim is indeed supported by the intuitive interpretation of the flux, as corresponding to the signed number of field lines crossing S. Indeed, any field line entering as $+$ must exit somewhere as $-$, and vice versa, so when summing we get 0.

(2) In practice now, in order to prove this rigorously, there are several ways. A first argument, which is quite elementary, is the one used by Feynman in [34], based on the fact that, due to $F \sim 1/d^2$, local deformations of S will leave invariant the flux, and so in the end we are left with a rotationally invariant surface, where the result is clear.

(3) A second argument, which basically uses the same idea, but is perhaps a bit more robust, is by redoing the computations in the proof of Theorem 8.7, by assuming this time that the integration takes place on an arbitrary surface as follows:

$$S_\lambda = \left\{ \lambda(u)u \middle| u \in S \right\}$$

To be more precise, here $\lambda : S \to (0, \infty)$ is a certain function, defining the surface, whose derivatives will appear both in the construction of the normal vectors $n(x)$ with $x = \lambda(u)u$, and in the Jacobian of the change of variables $x \to u$, and in the end, when integrating over S as in the proof of Theorem 8.7, this function λ dissapears.

(4) A third argument, used by basically all electrodynamics books at the graduate level, and by some undergraduate books too, is by using heavy calculus, namely partial integration in 3D, and we will discuss this later, more in detail, a bit later.

(5) A fourth argument is by following the idea in (1), namely carefully axiomatizing the field lines, and their relation with the field, and then obtaining $\Phi_E(S) = 0$ by using the in-and-out trick in (1), as explained for instance by Griffiths in [44].

(6) To summmarize, we are led to the conclusion that given a closed surface S, with no charges inside, the flux through it of any choice of external charges vanishes:

$$\Phi_E(S) = 0$$

(7) The point now is that, with this and Proposition 8.6 in hand, we can finish by using a standard math trick. Let us assume indeed, by discretizing, that our system of charges

is discrete, consisting of enclosed charges $q_1, \ldots, q_k \in \mathbb{R}$, and an exterior total charge Q_{ext}. We can surround each of $q_1, \ldots, q_k$ by small disjoint spheres $U_1, \ldots, U_k$, chosen such that their interiors do not touch S, and we have:

$$\begin{aligned}
\Phi_E(S) &= \Phi_E(S - \cup U_i) + \Phi_E(\cup U_i) \\
&= 0 + \Phi_E(\cup U_i) \\
&= \sum_i \Phi_E(U_i) \\
&= \sum_i \frac{q_i}{\varepsilon_0} \\
&= \frac{Q_{enc}}{\varepsilon_0}
\end{aligned}$$

(8) 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 (6). Finally, the end makes use of Proposition 8.6. □

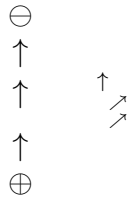
8b. Magnetic fields

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 8.9. *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 8.10 (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 8.9. 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 8.11. *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 8.12 (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 8.10, and in the definition of $\times$. In fact, the Lorentz force law is just a fancy reformulation of Fact 8.10, 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 8.10.

As an interesting application of the Lorentz force law, we have:

THEOREM 8.13. *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 8.14. *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 8.12 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 8.15. *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 8.16. *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 8.17 (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 8.18 (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, and very deep, that will have numerous consequences, in what follows, be that for investigating phenomena like radiation, or for making the link with Einstein's relativity theory, both crucially involving c.

But, first of all, this is certainly an invitation to rediscuss units and constants, as a continuation of our previous discussion on this topic. In what regards the units, we won't be impressed by the ampere, and keep using the coulomb, as a main unit:

CONVENTIONS 8.19. *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 8.18, 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 the speed of light c as constant too, and we are led in this way into the following conventions:

CONVENTIONS 8.20. *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 $K = 1/(4\pi\varepsilon_0)$.

Observe in passing that we are not messing up our figures, which can be quite often the case in this type of situation, because according to our data, and by truncating instead of rounding, as busy theoretical physicists usually do, we have:

$$\varepsilon_0\mu_0c^2 = 8.854 \times 1.256 \times 2.997^2 \times 10^{16-12-6} = 0.998$$

Getting back now to theory and math, the Biot-Savart law has as consequence:

THEOREM 8.21. *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 the magnetic field B of a steady line current I is given by the following formula:

$$B = \frac{\mu_0}{4\pi}\int \frac{I \times x}{||x||^3}$$

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 8.22 (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 the magnetic field B of a steady line current I is given by the following formula:

$$B = \frac{\mu_0}{4\pi} \int \frac{I \times x}{||x||^3}$$

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 8.23. *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 8.21 and Theorem 8.22, which both follow from the Biot-Savart law. □

8c. Light, optics

To start with, we can talk about waves in N dimensions, as follows:

THEOREM 8.24. *The wave equation in $\mathbb{R}^N$ is as follows,*

$$\ddot{\varphi} = v^2 \Delta\varphi$$

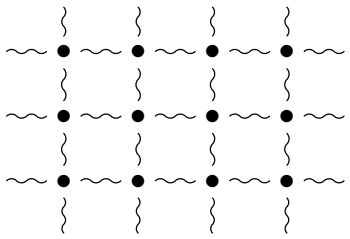
with $v > 0$ being the propagation speed of the wave, and with Δ given by

$$\Delta\varphi = \sum_{i=1}^{N} \frac{d^2\varphi}{dx_i^2}$$

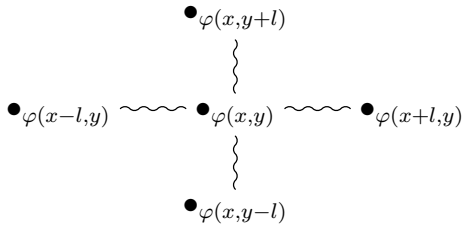
being the Laplace operator, playing the role of a numeric second derivative.

PROOF. We can use here a lattice model as before in 1D, as follows:

(1) In 2 dimensions, to start with, the same argument as before carries on. Indeed, we can use a lattice model as follows, with all the edges standing for small springs:



As before in one dimension, we send an impulse, and we zoom on one ball. The situation here is as follows, with l being the spring length:



We have two forces acting at (x, y). First is the Newton motion force, mass times acceleration, which is as follows, with m being the mass of each ball:

$$F_n = m \cdot \ddot{\varphi}(x, y)$$

And second is the Hooke force, displacement of the spring, times spring constant. Since we have four springs at (x, y), this is as follows, k being the spring constant:

$$\begin{aligned} F_h &= F_h^r - F_h^l + F_h^u - F_h^d \\ &= k(\varphi(x+l,y) - \varphi(x,y)) - k(\varphi(x,y) - \varphi(x-l,y)) \\ &+ k(\varphi(x,y+l) - \varphi(x,y)) - k(\varphi(x,y) - \varphi(x,y-l)) \\ &= k(\varphi(x+l,y) - 2\varphi(x,y) + \varphi(x-l,y)) \\ &+ k(\varphi(x,y+l) - 2\varphi(x,y) + \varphi(x,y-l)) \end{aligned}$$

We conclude that the equation of motion, in our model, is as follows:

$$\begin{aligned} m \cdot \ddot{\varphi}(x, y) &= k(\varphi(x+l,y) - 2\varphi(x,y) + \varphi(x-l,y)) \\ &+ k(\varphi(x,y+l) - 2\varphi(x,y) + \varphi(x,y-l)) \end{aligned}$$

(2) Now let us take the limit of our model, as to reach to continuum. For this purpose we will assume that our system consists of $B^2 >> 0$ balls, having a total mass M, and

spanning a total area L^2. Thus, our previous infinitesimal parameters are as follows, with K being the spring constant of the total system, taken to be equal to k:

$$m = \frac{M}{B^2} \quad , \quad k = K \quad , \quad l = \frac{L}{B}$$

With these changes, our equation of motion found in (3) reads:

$$\begin{aligned} \ddot{\varphi}(x,y) &= \frac{KB^2}{M}(\varphi(x+l,y) - 2\varphi(x,y) + \varphi(x-l,y)) \\ &+ \frac{KB^2}{M}(\varphi(x,y+l) - 2\varphi(x,y) + \varphi(x,y-l)) \end{aligned}$$

Now observe that this equation can be written, more conveniently, as follows:

$$\begin{aligned} \ddot{\varphi}(x,y) &= \frac{KL^2}{M} \times \frac{\varphi(x+l,y) - 2\varphi(x,y) + \varphi(x-l,y)}{l^2} \\ &+ \frac{KL^2}{M} \times \frac{\varphi(x,y+l) - 2\varphi(x,y) + \varphi(x,y-l)}{l^2} \end{aligned}$$

With $N \to \infty$, and therefore $l \to 0$, we obtain in this way:

$$\ddot{\varphi}(x,y) = \frac{KL^2}{M}\left(\frac{d^2\varphi}{dx^2} + \frac{d^2\varphi}{dy^2}\right)(x,y)$$

As a conclusion to this, we are led to the following wave equation in two dimensions, with $v = \sqrt{K/M} \cdot L$ being the propagation speed of our wave:

$$\ddot{\varphi}(x,y) = v^2\left(\frac{d^2\varphi}{dx^2} + \frac{d^2\varphi}{dy^2}\right)(x,y)$$

But we recognize at right the Laplace operator, and we are done. As before in 1D, there is of course some discussion to be made here, arguing that our spring model in (1) is indeed the correct one. But do not worry, experiments confirm our findings.

(3) In 3 dimensions now, which is the case of the main interest, corresponding to our real-life world, the same argument carries over, and the wave equation is as follows:

$$\ddot{\varphi}(x,y,z) = v^2\left(\frac{d^2\varphi}{dx^2} + \frac{d^2\varphi}{dy^2} + \frac{d^2\varphi}{dz^2}\right)(x,y,z)$$

(4) Finally, the same argument, namely a lattice model, carries on in arbitrary N dimensions, and the wave equation here is as follows:

$$\ddot{\varphi}(x_1,\ldots,x_N) = v^2\sum_{i=1}^{N}\frac{d^2\varphi}{dx_i^2}(x_1,\ldots,x_N)$$

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

Light is the wave predicted by electrodynamics, 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 8.25. *An accelerating or decelerating charge produces electromagnetic radiation, called light, whose frequency and wavelength can be explicitely 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.

Getting back now to Fact 8.25, in its general form, as stated above, this is something which can be deduced via some math, based on the Maxwell equations.

Moving ahead, let us go back to the wave equation $\ddot{\varphi} = v^2\Delta\varphi$ from Theorem 8.24, and try to understand its simplest solutions. In 1D, we know that we have:

THEOREM 8.26. *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.

(3) Finally, the last assertion is clear. We will see later in this book, using Fourier analysis, that any solution of the 1D wave equation appears in fact in this way. □

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:

$$\varphi = Re(\widetilde{\varphi}) \quad , \quad \widetilde{\varphi} = \widetilde{A}e^{i(kx-wt)}$$

In fact, with a bit more work, we can fully solve the 1D wave equation, as follows:

THEOREM 8.27. *The solution of the 1D wave equation with initial value conditions $\varphi(x,0) = f(x)$ and $\dot{\varphi}(x,0) = g(x)$ is given by the d'Alembert formula, namely:*

$$\varphi(x,t) = \frac{f(x-vt) + f(x+vt)}{2} + \frac{1}{2v}\int_{x-vt}^{x+vt} g(s)ds$$

In the context of our previous lattice model discretizations, what happens is more or less that the above d'Alembert integral gets computed via Riemann sums.

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

(1) Let us first check that the d'Alembert solution is indeed a solution of the wave equation $\ddot{\varphi} = v^2\varphi''$. The first time derivative is computed as follows:

$$\dot{\varphi}(x,t) = \frac{-vf'(x-vt) + vf'(x+vt)}{2} + \frac{1}{2v}(vg(x+vt) + vg(x-vt))$$

The second time derivative is computed as follows:

$$\ddot{\varphi}(x,t) = \frac{v^2f''(x-vt) + v^2f(x+vt)}{2} + \frac{vg'(x+vt) - vg'(x-vt)}{2}$$

Regarding now space derivatives, the first one is computed as follows:

$$\varphi'(x,t) = \frac{f'(x-vt) + f'(x+vt)}{2} + \frac{1}{2v}(g'(x+vt) - g'(x-vt))$$

As for the second space derivative, this is computed as follows:

$$\varphi''(x,t) = \frac{f''(x-vt) + f''(x+vt)}{2} + \frac{g''(x+vt) - g''(x-vt)}{2v}$$

Thus we have indeed $\ddot{\varphi} = v^2\varphi''$. As for the initial conditions, $\varphi(x,0) = f(x)$ is clear from our definition of φ, and $\dot{\varphi}(x,0) = g(x)$ is clear from our above formula of $\dot{\varphi}$.

(2) Conversely now, we must show that our solution is unique, but instead of going here into abstract arguments, we will simply solve our equation, which among others will doublecheck the computations in (1). Let us make the following change of variables:

$$\xi = x - vt \quad , \quad \eta = x + vt$$

With this change of variables, which is quite tricky, mixing space and time variables, our wave equation $\ddot{\varphi} = v^2\varphi''$ reformulates in a very simple way, as follows:

$$\frac{d^2\varphi}{d\xi d\eta} = 0$$

But this latter equation tells us that our new ξ, η variables get separated, and we conclude from this that the solution must be of the following special form:

$$\varphi(x,t) = F(\xi) + G(\eta) = F(x - vt) + G(x + vt)$$

Now by taking into account the intial conditions $\varphi(x,0) = f(x)$ and $\dot{\varphi}(x,0) = g(x)$, and then integrating, we are led to the d'Alembert formula in the statement.

(3) In regards now with our discretization questions, by using a 1D lattice model with balls and springs as before, what happens to all the above is more or less that the above d'Alembert integral gets computed via Riemann sums, in our model, as stated. □

Moving ahead now towards electromagnetism and 3D, let us formulate:

DEFINITION 8.28. *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 8.26, 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 8.26. And second is the assumption, in connection with the Fourier decomposition result from the end of Theorem 8.26, 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 8.24 and Fact 8.25, we have:

FACT 8.29. *Physically speaking, a monochromatic plane wave is the electromagnetic radiation appearing as in Theorem 8.24 and Fact 8.25, 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.

In practice, we have various types of light, depending on frequency and wavelength. These are normally referred to as "electromagnetic waves", but for keeping things simple,

we will keep using the 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	$625 - 750$

Outside visible light we have, as you probably know it, UV on higher frequencies, and IR on lower frequencies. At the high frequency end we have X-rays, that you surely know about too, and γ rays, which are usually associated with various bad things, such as thunderstorms, solar flares, and small bugs with our nuclear energy technology.

As for the lower frequency end of the scale, first we have microwaves, but if you love physics and chemistry you should learn some cooking, that's first-class chemistry, that you can practice every day. And then we have all sorts of radio wavelenghts, including FM, followed by AM, and then by several more obscure low-frequency waves.

Importantly, both ends of the table are a bit loose. At the high frequency end there are some restrictions coming from quantum mechanics, and more on them later. As for the low frequency end, what's wave and what's not is a bit of a philosophical question, but which is actually not that philosophical, because waves having huge wavelengths can

easily turn around mountains, full countries and so on, and so are of military interest. Secret research here, more of engineering type of course, is still ongoing.

Back now to our business, with all the above 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:

THEOREM 8.30. *Inside a linear, homogeneous medium, where there is no free charge or free current present, both the electric and magnetic fields E and B are subject to*

$$\ddot{\varphi} = v^2 \Delta\varphi$$

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. □

As a first observation here, while the above is something quite trivial, mathematically speaking, from the physical viewpoint we are here into complicated things. Materials can be transparent or opaque, with the distinction between them being something very subtle, and advanced, and Theorem 8.30 obviously deals with the transparent case.

In short, we are here inside advanced materials theory, that we cannot really understand, with our knowledge so far. In what follows we will be interested in transparent materials only, such as glass. Regarding the other materials, such as rock, let us just mention that light dissapears inside them, converted into heat. Of course glass heats too when light crosses it, with this being related to $v < c$ inside it. More on this later.

Next in line, and of interest for us, we have:

FACT 8.31. *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.*

Again, this is something deep, and very old as well, and there are many things that can be said here, ranging from various computations based on the Maxwell equations, to all sorts of considerations belonging to advanced materials theory.

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 [44].

As a simple consequence of the above, we have:

THEOREM 8.32. *Light can be decomposed, by using a prism.*

PROOF. This follows from Fact 8.31. 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 8.33. *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.*

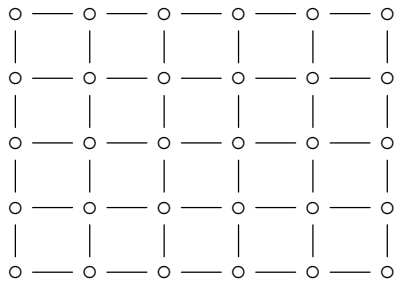
This is the main principle of spectroscopy, and applications, of all kinds, abound. In practice, the mathematical tool needed for doing the "reverse engineering" mentioned above is the Fourier transform, which allows the decomposition of packets of waves, into monochromatic components. Finally, let us mention too that, needless to say, the event can be reconstructed only partially out of its spectral signature.

8d. Heat, revised

Let us discuss now heat, in analogy with what we did in the above, for the waves.

The simplest heat diffusion question, studied and understood since long, concerns a container containing two gases, having initial different temperatures $T_1<T_2$, separated

by a membrane. Heat transfer goes on, in this setting, and obviously, we can model this by focusing on the membrane, with a basic grid model for it:



There is some sort of "game" played by the two gases, over this grid, and we can model this, and then recover the known results about heat diffusion, in this setting.

At a more advanced level, we can remove the membrane. Again, there is some sort of "game" here, played by the two gases, which can be 2D or 3D, depending on modeling. Also, in this setting, we can actually keep the membrane, but allow it to inflate.

Let us go now into heavier, fully powerful models and equations for the heat diffusion mechanism, involving this time more advanced mathematics and physics. The general equation here is quite similar to the one for the waves, as follows:

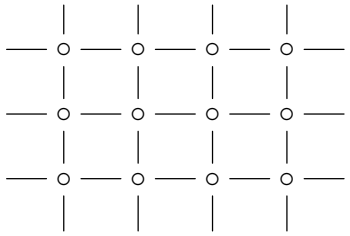
THEOREM 8.34. *Heat diffusion in $\mathbb{R}^N$ is described by the heat equation*

$$\dot{\varphi} = \alpha\Delta\varphi$$

where $\alpha > 0$ is the thermal diffusivity of the medium, and Δ is the Laplace operator.

PROOF. The study here is quite similar to the study of waves, as follows:

(1) Let us first discuss 2 dimensions. Here, as before for the waves, we can use a lattice model as follows, with all lengths being $l > 0$, for simplifying:



(2) We have to implement now the physical heat diffusion mechanism, namely "the rate of change of the temperature of the material at any given point must be proportional, with proportionality factor $\alpha > 0$, to the average difference of temperature between that

given point and the surrounding material". In practice, this leads to a condition as follows, expressing the change of the temperature φ, over a small period of time $\delta > 0$:

$$\varphi(x,y,t+\delta) = \varphi(x,y,t) + \frac{\alpha\delta}{l^2} \sum_{(x,y)\sim(u,v)} [\varphi(u,v,t) - \varphi(x,y,t)]$$

In fact, we can rewrite our equation as follows, making it clear that we have here an equation regarding the rate of change of temperature at x:

$$\frac{\varphi(x,y,t+\delta) - \varphi(x,y,t)}{\delta} = \frac{\alpha}{l^2} \sum_{(x,y)\sim(u,v)} [\varphi(u,v,t) - \varphi(x,y,t)]$$

(3) So, let us do the math. In the context of our 2D model the neighbors of x are the points $(x \pm l, y \pm l)$, so the equation above takes the following form:

$$\begin{aligned} & \frac{\varphi(x,y,t+\delta) - \varphi(x,y,t)}{\delta} \\ = \ & \frac{\alpha}{l^2}\Big[(\varphi(x+l,y,t) - \varphi(x,y,t)) + (\varphi(x-l,y,t) - \varphi(x,y,t))\Big] \\ + \ & \frac{\alpha}{l^2}\Big[(\varphi(x,y+l,t) - \varphi(x,y,t)) + (\varphi(x,y-l,t) - \varphi(x,y,t))\Big] \end{aligned}$$

Now observe that we can write this equation as follows:

$$\begin{aligned} \frac{\varphi(x,y,t+\delta) - \varphi(x,y,t)}{\delta} &= \alpha \cdot \frac{\varphi(x+l,y,t) - 2\varphi(x,y,t) + \varphi(x-l,y,t)}{l^2} \\ &+ \alpha \cdot \frac{\varphi(x,y+l,t) - 2\varphi(x,y,t) + \varphi(x,y-l,t)}{l^2} \end{aligned}$$

(4) As it was the case when modeling the wave equation before, we recognize on the right the usual approximation of the second derivative, coming from calculus. Thus, when taking the continuous limit of our model, $l \to 0$, we obtain the following equation:

$$\frac{\varphi(x,y,t+\delta) - \varphi(x,y,t)}{\delta} = \alpha \left(\frac{d^2\varphi}{dx^2} + \frac{d^2\varphi}{dy^2}\right)(x,y,t)$$

Now with $t \to 0$, we are led in this way to the heat equation, namely:

$$\dot{\varphi}(x,y,t) = \alpha \cdot \Delta\varphi(x,y,t)$$

Finally, in arbitrary N dimensions the same argument carries over, namely a straightforward lattice model, and gives the heat equation, as formulated in the statement. □

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

8e. Exercises

Exercises:

Exercise 8.35.

Exercise 8.36.

Exercise 8.37.

Exercise 8.38.

Exercise 8.39.

Exercise 8.40.

Exercise 8.41.

Exercise 8.42.

Bonus exercise.

Part III

Three dimensions

I am milk
I am red hot kitchen
And I am cool
Cool as the deep blue ocean

CHAPTER 9

Space geometry

9a. Space geometry

Space geometry, in that usual 3 dimensions that we live in. Many interesting things can be said here, in analogy with what we know from chapter 5 about triangles.

9b. Curves, surfaces

At a more advanced level, we can do some algebraic geometry in $\mathbb{R}^3$, in continuation to what we did before in $\mathbb{R}^2$. Here we are right away into a dillema, because the plane curves have two possible generalizations. First we have the algebraic curves in $\mathbb{R}^3$:

DEFINITION 9.1. *An algebraic curve in $\mathbb{R}^3$ is a curve as follows,*

$$C = \left\{(x, y, z) \in \mathbb{R}^3 \middle| P(x, y, z) = 0,\ Q(x, y, z) = 0\right\}$$

appearing as the joint zeroes of two polynomials P, Q.

These curves look of course like the usual plane curves, and at the level of the phenomena that can appear, these are similar to those in the plane, involving singularities and so on, but also knotting, which is a new phenomenon. However, it is hard to say something with bare hands about knots. We will be back to this, later in this book.

On the other hand, as another natural generalization of the plane curves, and this might sound a bit surprising, we have the surfaces in $\mathbb{R}^3$, constructed as follows:

DEFINITION 9.2. *An algebraic surface in $\mathbb{R}^3$ is a surface as follows,*

$$S = \left\{(x, y, z) \in \mathbb{R}^3 \middle| P(x, y, z) = 0\right\}$$

appearing as the zeroes of a polynomial P.

The point indeed is that, as it was the case with the plane curves, what we have here is something defined by a single equation. And with respect to many questions, having a single equation matters a lot, and this is why surfaces in $\mathbb{R}^3$ are "simpler" than curves in $\mathbb{R}^3$. In fact, believe me, they are even the correct generalization of the curves in $\mathbb{R}^2$.

As an example of what can be done with surfaces, which is very similar to what we did with the conics $C \subset \mathbb{R}^2$ in chapter 8, we have the following result:

THEOREM 9.3. *The degree 2 surfaces $S \subset \mathbb{R}^3$, called quadrics, are the ellipsoid*

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

which is the only compact one, plus 16 more, which can be explicitely listed.

PROOF. We will be quite brief here, because we intend to rediscuss all this in a moment, with full details, in arbitrary N dimensions, the idea being as follows:

(1) The equations for a quadric $S \subset \mathbb{R}^2$ are best written as follows, with $A \in M_3(\mathbb{R})$ being a matrix, $B \in M_{1\times 3}(\mathbb{R})$ being a row vector, and $C \in \mathbb{R}$ being a constant:

$$< Au, u > +Bu + C = 0$$

(2) By doing now the linear algebra, and we will come back to this in a moment, with details, or by invoking the theorem of Sylvester on quadratic forms, we are left, modulo degeneracy and linear transformations, with signed sums of squares, as follows:

$$\pm x^2 \pm y^2 \pm z^2 = 0, 1$$

(3) Thus the sphere is the only compact quadric, up to linear transformations, and by applying now linear transformations to it, we are led to the ellipsoids in the statement.

(4) As for the other quadrics, there are many of them, a bit similar to the parabolas and hyperbolas in 2 dimensions, and some work here leads to a 16 item list. □

With this done, instead of further insisting on the surfaces $S \subset \mathbb{R}^3$, or getting into their rivals, the curves $C \subset \mathbb{R}^3$, which appear as intersections of such surfaces, $C = S \cap S'$, let us get instead to arbitrary N dimensions, see what the axiomatics looks like there, with the hope that this will clarify our dimensionality dillema, curves vs surfaces.

So, moving to N dimensions, we have here the following definition, to start with:

DEFINITION 9.4. *An algebraic hypersurface in $\mathbb{R}^N$ is a space of the form*

$$S = \left\{(x_1, \ldots, x_N) \in \mathbb{R}^N \middle| P(x_1, \ldots, x_N) = 0, \forall i\right\}$$

appearing as the zeroes of a polynomial $P \in \mathbb{R}[x_1, \ldots, x_N]$.

Again, this is a quite general definition, covering both the plane curves $C \subset \mathbb{R}$ and the surfaces $S \subset \mathbb{R}^2$, which is certainly worth a systematic exploration. But, no hurry with this, for the moment we are here for talking definitons and axiomatics.

In order to have now a full collection of beasts, in all possible dimensions $N \in \mathbb{N}$, and of all possible dimensions $k \in \mathbb{N}$, we must intersect such algebraic hypersurfaces. We are led in this way to the zeroes of families of polynomials, as follows:

DEFINITION 9.5. *An algebraic manifold in $\mathbb{R}^N$ is a space of the form*

$$X = \left\{(x_1, \ldots, x_N) \in \mathbb{R}^N \middle| P_i(x_1, \ldots, x_N) = 0, \forall i\right\}$$

with $P_i \in \mathbb{R}[x_1, \ldots, x_N]$ being a family of polynomials.

As a first observation, as already mentioned, such a manifold appears as an intersection of hypersurfaces S_i, those associated to the various polynomials P_i:

$$X = S_1 \cap \ldots \cap S_r$$

There is actually a bit of a discussion needed here, regarding the parameter $r \in \mathbb{N}$, shall we allow this parameter to be $r = \infty$ too, or not. We will discuss this later, with some algebra helping, the idea being that allowing $r = \infty$ forces in fact $r < \infty$.

Let us look now more in detail at the hypersurfaces. We have here:

THEOREM 9.6. *The degree 2 hypersurfaces $S \subset \mathbb{R}^N$, called quadrics, are up to degeneracy and to linear transformations the hypersurfaces of the following form,*

$$\pm x_1^2 \pm \ldots \pm x_N^2 = 0, 1$$

and with the sphere being the only compact one.

PROOF. We have two statements here, the idea being as follows:

(1) The equations for a quadric $S \subset \mathbb{R}^N$ are best written as follows, with $A \in M_N(\mathbb{R})$ being a matrix, $B \in M_{1\times N}(\mathbb{R})$ being a row vector, and $C \in \mathbb{R}$ being a constant:

$$< Ax, x > +Bx + C = 0$$

(2) By doing the linear algebra, or by invoking the theorem of Sylvester on quadratic forms, we are left, modulo linear transformations, with signed sums of squares:

$$\pm x_1^2 \pm \ldots \pm x_N^2 = 0, 1$$

(3) To be more precise, with linear algebra, by evenly distributing the terms $x_i x_j$ above and below the diagonal, we can assume that our matrix $A \in M_N(\mathbb{R})$ is symmetric. Thus A must be diagonalizable, and by changing the basis of $\mathbb{R}^N$, as to have it diagonal, our equation becomes as follows, with $D \in M_N(\mathbb{R})$ being now diagonal:

$$< Dx, x > +Ex + F = 0$$

(4) But now, by making squares in the obvious way, which amounts in applying yet another linear transformation to our quadric, the equation takes the following form, with $G \in M_N(-1, 0, 1)$ being diagonal, and with $H \in \{0, 1\}$ being a constant:

$$< Gx, x >= H$$

(5) Now barring the degenerate cases, we can further assume $G \in M_N(-1,1)$, and we are led in this way to the equation claimed in (2) above, namely:

$$\pm x_1^2 \pm \ldots \pm x_N^2 = 0, 1$$

(6) In particular we see that, up to some degenerate cases, namely emptyset and point, the only compact quadric, up to linear transformations, is the one given by:

$$x_1^2 + \ldots + x_N^2 = 1$$

(7) But this is the unit sphere, so are led to the conclusions in the statement. □

Regarding now the examples of hypersurfaces $S \subset \mathbb{R}^N$, or of more general algebraic manifolds $X \subset \mathbb{R}^N$, there are countless of them, and it is impossible to have some discussion started here, without being subjective. The unit sphere $S_{\mathbb{R}}^{N-1} \subset \mathbb{R}^N$ gets of course the crown from everyone, as being the most important manifold after $\mathbb{R}^N$ itself. But then, passed this sphere, things ramify, depending on what exact applications of algebraic geometry you have in mind. In what concerns me, here is my next favorite example:

THEOREM 9.7. *The invertible matrices $A \in M_N(\mathbb{R})$ lie outside the hypersurface*

$$\det A = 0$$

and are therefore dense, in the space of all matrices $M_N(\mathbb{R})$.

PROOF. This is something self-explanatory, but with this result being some key in linear algebra, all this is worth a detailed discussion, as follows:

(1) We certainly know from basic linear algebra that a matrix $A \in M_N(\mathbb{R})$ is invertible precisely when it has nonzero determinant, $\det A \neq 0$. Thus, the invertible matrices $A \in M_N(\mathbb{R})$ are located precisely in the complement of the following space:

$$S = \left\{ A \in M_N(\mathbb{R}) \middle| \det A = 0 \right\}$$

(2) We also know from basic linear algebra, or perhaps not so basic linear algebra, that the determinant $\det A$ is a certain polynomial in the entries of A, of degree N:

$$\det \in \mathbb{R}[X_{11}, \ldots, X_{NN}]$$

(3) We conclude from this that the above set S is a degree N algebraic hypersurface in our sense, in the Euclidean space $M_N(\mathbb{R}) \simeq \mathbb{R}^n$, with $n = N^2$.

(4) Now since the complements of non-trivial hypersurfaces $S \subset \mathbb{R}^n$ are obviously dense, and if needing a formal proof here, for our above hypersurface S this is clear, simply by suitably perturbing the matrix, and in general do not worry, we will be back to this, with full details, we are led to the conclusions in the statement. □

As an illustration for the power of our density result, we have:

THEOREM 9.8. *Given two matrices $A, B \in M_N(\mathbb{R})$, their products*

$$AB, BA \in M_N(\mathbb{R})$$

have the same characteristic polynomial, $P_{AB} = P_{BA}$.

PROOF. This is something quite hard to prove with bare hands, but we can trick by using Theorem 9.7. Indeed, it follows from definitions that the characteristic polynomial of a matrix is invariant under conjugation, in the sense that we have:

$$P_C = P_{ACA^{-1}}$$

Now observe that, when assuming that A is invertible, we have:

$$AB = A(BA)A^{-1}$$

Thus, we obtain the following formula, in the case where A is invertible:

$$P_{AB} = P_{BA}$$

Now by using the density result from Theorem 9.7, we conclude that this formula holds in fact for any matrix A, by continuity, as desired. □

Summarizing, we have some algebraic geometry theory going on, with applications, at least to questions in linear algebra, and presumably in calculus too. Getting back now to the basics, it is in fact possible to do even more generally, as follows:

DEFINITION 9.9. *An algebraic manifold over a field F is a space of the form*

$$X = \left\{(x_1, \ldots, x_N) \in F^N \middle| P_i(x_1, \ldots, x_N) = 0, \forall i\right\}$$

with $P_i \in F[x_1, \ldots, x_N]$ being a family of polynomials.

This might seem a bit abstract, but as a first observation, recall that $F = \mathbb{C}$ is a field too, on par with $F = \mathbb{R}$, and even better than it, in certain contexts. For instance quantum mechanics naturally lives over $F = \mathbb{C}$, instead of our usual $F = \mathbb{R}$. Also, in relation with questions in linear algebra, a matrix $A \in M_N(\mathbb{R})$ is much better viewed as matrix $A \in M_N(\mathbb{C})$, because here it has all N eigenvalues, when counted with multiplicities.

In fact, based on this linear algebra observation, and as our first result in complex algebraic geometry, we can improve Theorem 9.8, as follows:

THEOREM 9.10. *Given two matrices $A, B \in M_N(\mathbb{C})$, their products*

$$AB, BA \in M_N(\mathbb{C})$$

have the same eigenvalues, with the same multiplicities.

PROOF. To start with, Theorem 9.7 holds over $\mathbb{C}$ too, with the invertible matrices $A \in M_N(\mathbb{C})$ being dense, as being complementary to the following hypersurface:

$$\det A = 0$$

But with this in hand, the trick from the proof of Theorem 9.8 applies, and gives:

$$P_{AB} = P_{BA}$$

But this gives the result, because in the complex matrix setting the characteristic polynomial P encodes the eigenvalues, with multiplicities. □

This was for a first result in complex algebraic geometry, perhaps a bit advanced. At the level of more elementary things, the first thought goes to the plane algebraic curves, in a complex sense. But, surprise here, these are the spaces as follows:

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

Now when looking at this formula, we realize that our curve $C \subset \mathbb{C}^2$ is in fact something quite complicated, corresponding to a 2-dimensional surface $X \subset \mathbb{R}^4$. But, no worries, we will come back to this regularly. In fact, in what follows, we will be jointly developing our theory over both $F = \mathbb{R}$ and $F = \mathbb{C}$, with such questions in mind.

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

9c. Regular polyhedra

Switching topics now, let us first discuss, still in relation with space geometry questions, the graphs. As a fundamental result about them, we have:

THEOREM 9.11. *For a connected planar graph we have the Euler formula*

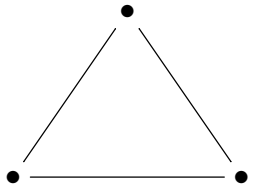
$$v - e + f = 2$$

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.

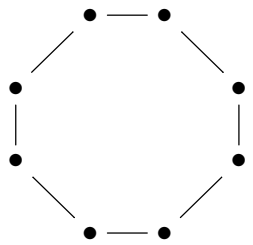
(2) As a first illustration for how this formula works, consider a triangle:



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:



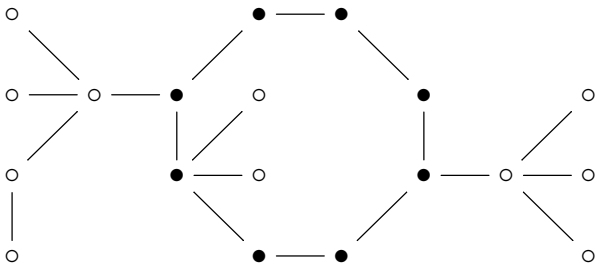
Then, for this graph, the Euler formula holds indeed, as follows:

$$N - N + 2 = 2$$

(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:

$$N - (N - 1) + 1 = 2$$

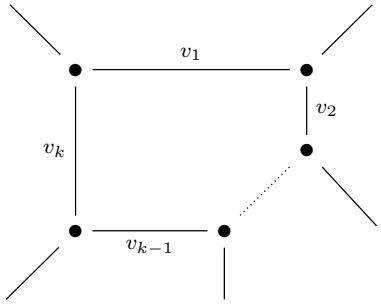
(5) At $f = 2$ now, our graph must be an N-gon as above, but with some trees allowed to grow from the vertices, with an illustrating example here being as follows:



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:

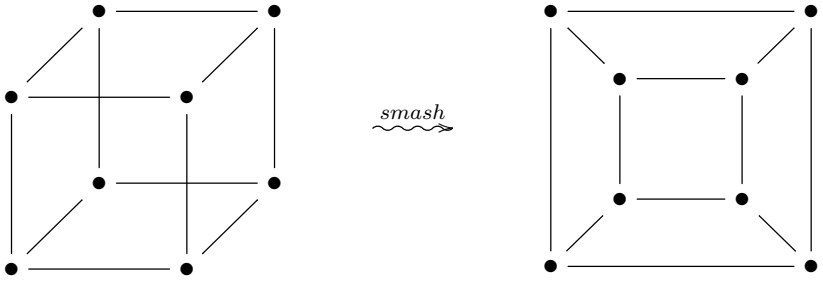
PROPOSITION 9.12. *For a convex polyhedron we have the Euler formula*

$$v - e + f = 2$$

with v, e, f being the number of vertices, edges and faces.

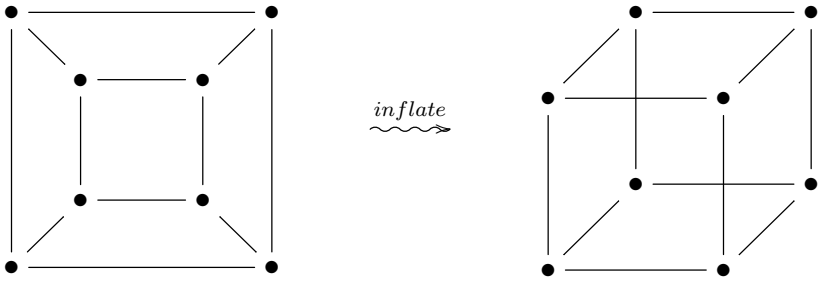
PROOF. This is more or less the same thing as Theorem 9.11, save for getting rid of the internal trees of the planar graph there, the idea being as follows:

(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 9.11.

(2) Conversely, consider a connected planar graph X. Then, save for getting rid of the internal trees, as explained in the proof of Theorem 9.11, we can assume that we are dealing with an agglomeration of N-gons, again as explained in the proof of Theorem 9.11. But now, we can inflate our graph as to obtain a convex polyhedron P:



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:

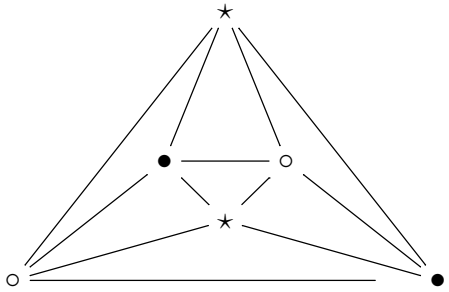
PROPOSITION 9.13. *The Euler formula $v - e + f = 2$ holds indeed for the five possible regular polyhedra, as follows:*

(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:* $12 - 30 + 20 = 2$.

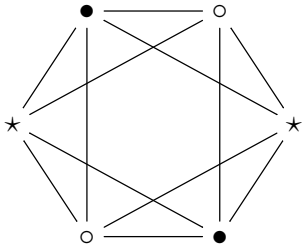
PROOF. The figures in the statement are certainly the good ones for the tetrahedron and the cube. Regarding now the octahedron, again the figures are the good ones, by

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 9.12 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:

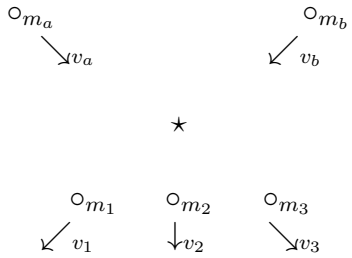


In short, you get the point, quite subjective all this, and as a conclusion, drawing graphs in an appropriate way remains an art. As for the dodecahedron and isocahedron, exercise here for you, and if failing, take some drawing classes. Math is not everything. □

The Euler formula $v - e + f = 2$, in both its above formulations, the graph one from Theorem 9.11, and the polyhedron one from Proposition 9.12, 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.

9d. Solid angles

Let us talk now about interactions between particles. But here, we have some experience from classical mechanics, with the typical picture of what can happen being:



This was for basic interactions in classical mechanics. In our present setting, particle physics, things are a bit more complicated than this, due to a variety of reasons, and experimental physics suggests looking at two main types of interactions, as follows:

FACT 9.14. *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 9.15. *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:

$$\begin{aligned}
\tau &= < t > \\
&= \int_0^\infty t f(t) dt \\
&= \int_0^\infty \lambda t e^{-\lambda t} dt \\
&= \int_0^\infty t(-e^{-\lambda t})' dt \\
&= \int_0^\infty e^{-\lambda t} dt \\
&= \frac{1}{\lambda}
\end{aligned}$$

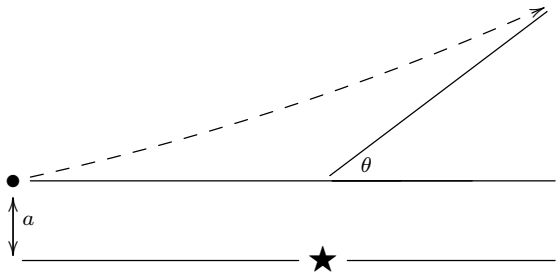
(3) Finally, regarding the half-life, this is by definition the time $t_{1/2}$ required for the decaying quantity to fall to one-half of its initial value. Mathematically, this means:

$$N_t = 2^{-\frac{t}{t_{1/2}}} N_0$$

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 9.16. *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:

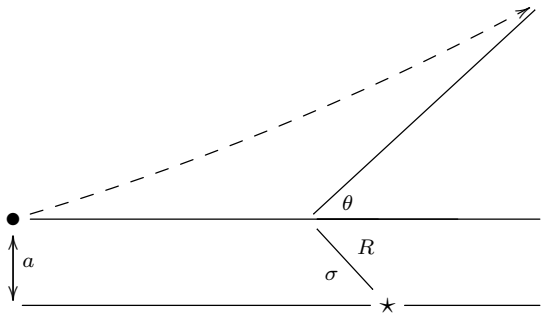
PROPOSITION 9.17. *In the context of classical particle colliding elastically with a hard sphere of radius $R > 0$, we have the formula*

$$a = R \cos \frac{\theta}{2}$$

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 9.16, as follows:

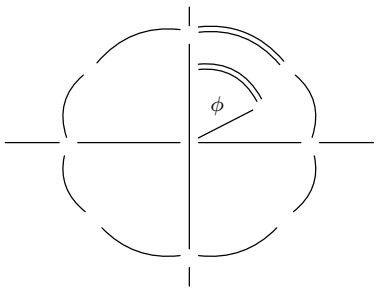
DEFINITION 9.18. *In the general context of scattering, we can:*

(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 9.19. *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 9.18, 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 9.20. *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 9.17 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. □

9e. Exercises

Exercises:

EXERCISE 9.21.

EXERCISE 9.22.

EXERCISE 9.23.

EXERCISE 9.24.

EXERCISE 9.25.

EXERCISE 9.26.

EXERCISE 9.27.

EXERCISE 9.28.

Bonus exercise.

CHAPTER 10

Rotating bodies

10a. Vector products

We will be talking here about all sorts of advanced mechanics, all taking place in 3D. We will need one more mathematical notion, which is something 3D specific, namely:

DEFINITION 10.1. *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's how it works. We are looking for a vector $x \times y$ whose length is proportional to those of x, y. But the second formula tells us that the angle θ between x, y must be involved via $0 \to 0$, and so the factor can only be $\sin\theta$. And with this we are almost there, it's just a matter of choosing the orientation, and this comes from $e_1 \times e_2 = e_3$.

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

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

Many things can be said about vector products, mathematically speaking.

10b. Angular momentum

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

$$J = x \times p$$

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

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

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

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

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

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

10c. Rotating bodies

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

THEOREM 10.3. *Assume that a 3D body rotates along an axis, with angular speed w. For a fixed point of the body, with position vector x, the usual 3D speed is*

$$v = \omega \times x$$

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

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

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

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

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

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

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

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

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

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

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

THEOREM 10.4. *Assuming as before that a 3D body rotates along an axis, the acceleration of a moving point on the body, computed by O as before, is given by*

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

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

PROOF. This comes by using twice the formulae in Theorem 10.3, as follows:

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

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

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

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

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

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

THEOREM 10.5. *The acceleration of an object m subject to a force F is given by*

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

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

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

We refer to any standard undergraduate mechanics book, such as Feynman [33], Kibble [58] or Taylor [86] for more on the above, including various numerics on what happens here on Earth, the Foucault pendulum, history of all this, and many other things. Let us just mention here, as a basic illustration for all this, that a rock dropped from 100m deviates about 1cm from its intended target, due to the formula in Theorem 10.5.

10d. Further results

Further results.

10e. Exercises

Exercises:

EXERCISE 10.6.

EXERCISE 10.7.

EXERCISE 10.8.

EXERCISE 10.9.

EXERCISE 10.10.

EXERCISE 10.11.

EXERCISE 10.12.

EXERCISE 10.13.

Bonus exercise.

CHAPTER 11

Advanced calculus

11a. Partial derivatives

Moving now to several variables, $N \geq 2$, as a first job, given a function $\varphi : \mathbb{R}^N \to \mathbb{R}$, we would like to find a quantity $\varphi'(x)$ making the following formula work:

$$\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h$$

But here, as in 1 variable, there are not so many choices, and the solution is that of defining $\varphi'(x)$ as being the row vector formed by the partial derivatives at x:

$$\varphi'(x) = \begin{pmatrix} \frac{d\varphi}{dx_1} & \ldots & \frac{d\varphi}{dx_N} \end{pmatrix}$$

To be more precise, with this value for $\varphi'(x)$, our approximation formula $\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h$ makes sense indeed, as an equality of real numbers, with $\varphi'(x)h \in \mathbb{R}$ being obtained as the matrix multiplication of the row vector $\varphi'(x)$, and the column vector h. As for the fact that our formula holds indeed, this follows by putting together the approximation properties of each of the partial derivatives $d\varphi/dx_i$, which give:

$$\varphi(x+h) \simeq \varphi(x) + \sum_{i=1}^{N} \frac{d\varphi}{dx_i} \cdot h_i = \varphi(x) + \varphi'(x)h$$

Before moving forward, you might say, why bothering with horizontal vectors, when it is so simple and convenient to have all vectors vertical, by definition. Good point, and in answer, we can indeed talk about the gradient of φ, constructed as follows:

$$\nabla\varphi = \begin{pmatrix} \frac{d\varphi}{dx_1} \\ \vdots \\ \frac{d\varphi}{dx_N} \end{pmatrix}$$

With this convention, $\nabla\varphi$ geometrically describes the slope of φ at the point x, in the obvious way. However, the approximation formula must be rewritten as follows:

$$\varphi(x+h) \simeq \varphi(x) + < \nabla\varphi(x), h >$$

In what follows we will use both φ' and $\nabla\varphi$, depending on the context. Moving now to second derivatives, the main result here is as follows:

THEOREM 11.1. *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 basic calculus, namely:

$$\varphi(x+h) \simeq \varphi(x) + \varphi'(x)h + \frac{\varphi''(x)h^2}{2}$$

(2) At $N = 2$ now, we obviously need to differentiate φ twice, and the point is that we come in this way upon the following formula, called Clairaut formula:

$$\frac{d^2\varphi}{dxdy} = \frac{d^2\varphi}{dydx}$$

But, is this formula correct or not? As an intuitive justification for it, let us consider a product of power functions, $\varphi(z) = x^p y^q$. We have then our formula, due to:

$$\frac{d^2\varphi}{dxdy} = \frac{d}{dx}\left(\frac{dx^p y^q}{dy} \right) = \frac{d}{dx}\left(qx^p y^{q-1} \right) = pqx^{p-1}y^{q-1}$$

$$\frac{d^2\varphi}{dydx} = \frac{d}{dy}\left(\frac{dx^p y^q}{dx} \right) = \frac{d}{dy}\left(px^{p-1} y^q \right) = pqx^{p-1}y^{q-1}$$

Next, let us consider a linear combination of power functions, $\varphi(z) = \sum_{pq} c_{pq} x^p y^q$, which can be finite or not. We have then, by using the above computation:

$$\frac{d^2\varphi}{dxdy} = \frac{d^2\varphi}{dydx} = \sum_{pq} c_{pq} pq x^{p-1} y^{q-1}$$

Thus, we can see that our commutation formula for derivatives holds indeed, due to the fact that the functions in x, y commute. Of course, all this does not fully prove our formula, in general. But exercise for you, to have this idea fully working, or to look up the standard proof of the Clairaut formula, using the mean value theorem.

(3) Moving now to $N = 3$ and higher, we can use here the Clairaut formula with respect to any pair of coordinates, which gives the Schwarz formula, namely:

$$\frac{d^2\varphi}{dx_i dx_j} = \frac{d^2\varphi}{dx_j dx_i}$$

Thus, the second derivative, or Hessian matrix, is symmetric, as claimed.

(4) Getting now to the main topic, namely approximation formula in the statement, in arbitrary N dimensions, this is in fact something which does not need a new proof, because it follows from the one-variable formula in (1), applied to the restriction of φ to the following segment in $\mathbb{R}^N$, which can be regarded as being a one-variable interval:

$$I = [x, x + h]$$

To be more precise, let $y \in \mathbb{R}^N$, and consider the following function, with $r \in \mathbb{R}$:

$$f(r) = \varphi(x + ry)$$

We know from (1) that the Taylor formula for f, at the point $r = 0$, reads:

$$f(r) \simeq f(0) + f'(0)r + \frac{f''(0)r^2}{2}$$

And our claim is that, with $h = ry$, this is precisely the formula in the statement.

(5) So, let us see if our claim is correct. By using the chain rule, we have the following formula, with on the right, as usual, a row vector multiplied by a column vector:

$$f'(r) = \varphi'(x + ry) \cdot y$$

By using again the chain rule, we can compute the second derivative as well:

$$\begin{aligned}
f''(r) &= (\varphi'(x + ry) \cdot y)' \\
&= \left(\sum_i \frac{d\varphi}{dx_i}(x + ry) \cdot y_i \right)' \\
&= \sum_i \sum_j \frac{d^2\varphi}{dx_i dx_j}(x + ry) \cdot \frac{d(x + ry)_j}{dr} \cdot y_i \\
&= \sum_i \sum_j \frac{d^2\varphi}{dx_i dx_j}(x + ry) \cdot y_i y_j \\
&= < \varphi''(x + ry)y, y >
\end{aligned}$$

(6) Time now to conclude. We know that we have $f(r) = \varphi(x + ry)$, and according to our various computations above, we have the following formulae:

$$f(0) = \varphi(x) \quad , \quad f'(0) = \varphi'(x) \quad , \quad f''(0) =< \varphi''(x)y, y >$$

Buit with this data in hand, the usual Taylor formula for our one variable function f, at order 2, at the point $r = 0$, takes the following form, with $h = ry$:

$$\begin{aligned}\varphi(x + ry) &\simeq \varphi(x) + \varphi'(x)ry + \frac{< \varphi''(x)y, y > r^2}{2}\\ &= \varphi(x) + \varphi'(x)t + \frac{< \varphi''(x)h, h >}{2}\end{aligned}$$

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

As before in the one variable case, many more things can be said, as a continuation of the above. For instance the local minima and maxima of $\varphi : \mathbb{R}^N \to \mathbb{R}$ appear at the points $x \in \mathbb{R}^N$ where the derivative vanishes, $\varphi'(x) = 0$, and where the second derivative $\varphi''(x) \in M_N(\mathbb{R})$ is positive, respectively negative. But, you surely know all this.

As a key observation now, generalizing what we know in 1 variable, we have:

PROPOSITION 11.2. *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 11.1, 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. □

With this understood, the problem is now, what can we say about the mathematics of Δ? As a first observation, which is a bit speculative, the Laplace operator appears by

applying twice the gradient operator, in a somewhat formal sense, as follows:

$$\begin{aligned}
\Delta\varphi &= \sum_{i=1}^{N}\frac{d^2\varphi}{dx_i^2}\\
&= \sum_{i=1}^{N}\frac{d}{dx_i}\cdot\frac{d\varphi}{dx_i}\\
&= \left\langle\begin{pmatrix}\frac{d}{dx_1}\\ \vdots\\ \frac{d}{dx_N}\end{pmatrix},\begin{pmatrix}\frac{d\varphi}{dx_1}\\ \vdots\\ \frac{d\varphi}{dx_N}\end{pmatrix}\right\rangle\\
&= <\nabla,\nabla\varphi>
\end{aligned}$$

Thus, it is possible to write a formula of type $\Delta=\nabla^2$, with the convention that the square of the gradient ∇ is taken in a scalar product sense, as above. However, this can be a bit confusing, and in what follows, we will not use this notation.

Instead of further thinking at this, and at double derivatives in general, let us formulate a more straightforward question, inspired by linear algebra, as follows:

QUESTION 11.3. *The Laplace operator being linear,*

$$\Delta(a\varphi+b\psi)=a\Delta\varphi+b\Delta\psi$$

what can we say about it, inspired by usual linear algebra?

In answer now, the space of functions $\varphi:\mathbb{R}^N\to\mathbb{R}$, on which Δ acts, being infinite dimensional, the usual tools from linear algebra do not apply as such, and we must be extremely careful. For instance, we cannot really expect to diagonalize Δ, via some sort of explicit procedure, as we usually do in linear algebra, for the usual matrices.

Thinking some more, there is actually a real bug too with our problem, because at $N=1$ this problem becomes "what can we say about the second derivatives $\varphi'':\mathbb{R}\to\mathbb{R}$ of the functions $\varphi:\mathbb{R}\to\mathbb{R}$, inspired by linear algebra", with answer "not much".

And by thinking even more, still at $N=1$, there is a second bug too, because if $\varphi:\mathbb{R}\to\mathbb{R}$ is twice differentiable, nothing will guarantee that its second derivative $\varphi'':\mathbb{R}\to\mathbb{R}$ is twice differentiable too. Thus, we have some issues with the domain and range of Δ, regarded as linear operator, and these problems will persist at higher N.

So, shall we trash Question 11.3? Not so quick, because, very remarkably, some magic comes at $N=2$ and higher in relation with complex analysis, according to:

PRINCIPLE 11.4. *The functions $\varphi : \mathbb{R}^N \to \mathbb{R}$ which are 0-eigenvectors of Δ,*

$$\Delta\varphi = 0$$

called harmonic functions, have the following properties:

(1) *At $N = 1$, nothing spectacular, these are just the linear functions.*
(2) *At $N = 2$, these are, locally, the real parts of holomorphic functions.*
(3) *At $N \geq 3$, these still share many properties with the holomorphic functions.*

In order to understand this, or at least get introduced to it, let us first look at the case $N = 2$. Here, any function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ can be regarded as function $\varphi : \mathbb{C} \to \mathbb{R}$, depending on $z = x + iy$. But, in view of this, it is natural to enlarge the attention to the functions $\varphi : \mathbb{C} \to \mathbb{C}$, and ask which of these functions are harmonic, $\Delta\varphi = 0$. And here, we have the following remarkable result, making the link with complex analysis:

THEOREM 11.5. *Any holomorphic function $\varphi : \mathbb{C} \to \mathbb{C}$, when regarded as function*

$$\varphi : \mathbb{R}^2 \to \mathbb{C}$$

is harmonic. Moreover, the conjugates $\bar{\varphi}$ of holomorphic functions are harmonic too.

PROOF. The first assertion comes from the following computation, with $z = x + iy$:

$$\begin{aligned}
\Delta z^n &= \frac{d^2 z^n}{dx^2} + \frac{d^2 z^n}{dy^2} \\
&= \frac{d(nz^{n-1})}{dx} + \frac{d(inz^{n-1})}{dy} \\
&= n(n-1)z^{n-2} - n(n-1)z^{n-2} \\
&= 0
\end{aligned}$$

As for the second assertion, this follows from $\Delta\bar{\varphi} = \overline{\Delta\varphi}$, which is clear from definitions, and which shows that if φ is harmonic, then so is its conjugate $\bar{\varphi}$. □

Many more things can be said, along these lines.

11b. Multiple integrals

We can talk about multiple integrals, in the obvious way. Getting now to the general theory and rules, for computing such integrals, the key result here is the change of variable formula. In order to discuss this, let us start with something that we know well, in 1D:

PROPOSITION 11.6. *We have the change of variable formula*

$$\int_a^b f(x)dx = \int_c^d f(\varphi(t))\varphi'(t)dt$$

where $c = \varphi^{-1}(a)$ and $d = \varphi^{-1}(b)$.

PROOF. This follows with $f = F'$, via the following differentiation rule:

$$(F\varphi)'(t) = F'(\varphi(t))\varphi'(t)$$

Indeed, by integrating between c and d, we obtain the result. □

In several variables now, we can only expect the above $\varphi'(t)$ factor to be replaced by something similar, a sort of "derivative of φ, arising as a real number". But this can only be the Jacobian $\det(\varphi'(t))$, and with this in mind, we are led to:

THEOREM 11.7. *Given a transformation $\varphi = (\varphi_1, \ldots, \varphi_N)$, we have*

$$\int_E f(x)dx = \int_{\varphi^{-1}(E)} f(\varphi(t))|J_\varphi(t)|dt$$

with the J_φ quantity, called Jacobian, being given by

$$J_\varphi(t) = \det\left[\left(\frac{d\varphi_i}{dx_j}(x)\right)_{ij}\right]$$

and with this generalizing the formula from Proposition 11.6.

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

(1) Observe first that this generalizes indeed the change of variable formula in 1 dimension, from Proposition 11.6, the point here being that the absolute value on the derivative appears as to compensate for the lack of explicit bounds for the integral.

(2) In general now, we can first argue that, the formula in the statement being linear in f, we can assume $f = 1$. Thus we want to prove $vol(E) = \int_{\varphi^{-1}(E)} |J_\varphi(t)|dt$, and with $D = \varphi^{-1}(E)$, this amounts in proving $vol(\varphi(D)) = \int_D |J_\varphi(t)|dt$.

(3) Now since this latter formula is additive with respect to D, it is enough to prove that $vol(\varphi(D)) = \int_D J_\varphi(t)dt$, for small cubes D, and assuming $J_\varphi > 0$. But this follows by using the usual definition of the determinant, as a volume.

(4) The details and computations however are quite non-trivial, and can be found for instance in Rudin [72]. 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.

11c. Spherical coordinates

Time now do some exciting computations, with the technology that we have. In what regards the applications of Theorem 11.7, these often come via:

PROPOSITION 11.8. *We have polar coordinates in 2 dimensions,*

$$\begin{cases} x &= r\cos t \\ y &= r\sin t \end{cases}$$

the corresponding Jacobian being $J = r$.

PROOF. This is elementary, the Jacobian being:

$$\begin{aligned} J &= \begin{vmatrix} \frac{d(r\cos t)}{dr} & \frac{d(r\cos t)}{dt} \\ \frac{d(r\sin t)}{dr} & \frac{d(r\sin t)}{dt} \end{vmatrix} \\ &= \begin{vmatrix} \cos t & -r\sin t \\ \sin t & r\cos t \end{vmatrix} \\ &= r\cos^2 t + r\sin^2 t \\ &= r \end{aligned}$$

Thus, we have indeed the formula in the statement. □

We can now compute the Gauss integral, which is the best calculus formula ever:

THEOREM 11.9. *We have the following formula,*

$$\int_{\mathbb{R}} e^{-x^2}dx = \sqrt{\pi}$$

called Gauss integral formula.

PROOF. Let I be the above integral. By using polar coordinates, we obtain:

$$\begin{aligned} I^2 &= \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-x^2-y^2}dxdy \\ &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2}rdrdt \\ &= 2\pi\int_0^{\infty}\left(-\frac{e^{-r^2}}{2}\right)'dr \\ &= 2\pi\left[0-\left(-\frac{1}{2}\right)\right] \\ &= \pi \end{aligned}$$

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

Moving now to 3 dimensions, we have here the following result:

PROPOSITION 11.10. *We have spherical coordinates in 3 dimensions,*

$$\begin{cases} x &= r\cos s \\ y &= r\sin s\cos t \\ z &= r\sin s\sin t \end{cases}$$

the corresponding Jacobian being $J(r,s,t) = r^2\sin s$.

PROOF. The fact that we have indeed spherical coordinates is clear. Regarding now the Jacobian, this is given by the following formula:

$$\begin{aligned} & J(r,s,t) \\ = & \begin{vmatrix} \cos s & -r\sin s & 0 \\ \sin s\cos t & r\cos s\cos t & -r\sin s\sin t \\ \sin s\sin t & r\cos s\sin t & r\sin s\cos t \end{vmatrix} \\ = & r^2\sin s\sin t\begin{vmatrix} \cos s & -r\sin s \\ \sin s\sin t & r\cos s\sin t \end{vmatrix} + r\sin s\cos t\begin{vmatrix} \cos s & -r\sin s \\ \sin s\cos t & r\cos s\cos t \end{vmatrix} \\ = & r\sin s\sin^2 t\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} + r\sin s\cos^2 t\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} \\ = & r\sin s(\sin^2 t+\cos^2 t)\begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} \\ = & r\sin s\times 1\times r \\ = & r^2\sin s \end{aligned}$$

Thus, we have indeed the formula in the statement. □

Let us work out now the general spherical coordinate formula, in arbitrary N dimensions. The formula here, which generalizes those at $N = 2, 3$, is as follows:

THEOREM 11.11. *We have spherical coordinates in N dimensions,*

$$\begin{cases} x_1 &= r\cos t_1 \\ x_2 &= r\sin t_1\cos t_2 \\ \vdots & \\ x_{N-1} &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\cos t_{N-1} \\ x_N &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\sin t_{N-1} \end{cases}$$

the corresponding Jacobian being given by the following formula,

$$J(r,t) = r^{N-1}\sin^{N-2}t_1\sin^{N-3}t_2\ \ldots\ \sin^2 t_{N-3}\sin t_{N-2}$$

and with this generalizing the known formulae at $N = 2, 3$.

PROOF. As before, the fact that we have spherical coordinates is clear. Regarding now the Jacobian, also as before, by developing over the last column, we have:

$$\begin{aligned} J_N &= r\sin t_1\ldots\sin t_{N-2}\sin t_{N-1}\times\sin t_{N-1}J_{N-1}\\ &+ r\sin t_1\ldots\sin t_{N-2}\cos t_{N-1}\times\cos t_{N-1}J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}(\sin^2 t_{N-1}+\cos^2 t_{N-1})J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}J_{N-1}\end{aligned}$$

Thus, we obtain the formula in the statement, by recurrence. □

As a comment here, the above convention for spherical coordinates is one among many, designed to best work in arbitrary N dimensions. Also, in what regards the precise range of the angles $t_1,\ldots,t_{N-1}$, we will leave this to you, as an instructive exercise.

As an application, let us compute the volumes of spheres. For this purpose, we must understand how the products of coordinates integrate over spheres. Let us start with the case $N=2$. Here the sphere is the unit circle $\mathbb{T}$, and with $z=e^{it}$ the coordinates are $\cos t,\sin t$. We can first integrate arbitrary powers of these coordinates, as follows:

PROPOSITION 11.12. *We have the following formulae,*

$$\int_0^{\pi/2}\cos^p t\,dt=\int_0^{\pi/2}\sin^p t\,dt=\left(\frac{\pi}{2}\right)^{\varepsilon(p)}\frac{p!!}{(p+1)!!}$$

where $\varepsilon(p)=1$ if p is even, and $\varepsilon(p)=0$ if p is odd, and where

$$m!!=(m-1)(m-3)(m-5)\ldots$$

with the product ending at 2 if m is odd, and ending at 1 if m is even.

PROOF. Let us first compute the integral on the left in the statement:

$$I_p=\int_0^{\pi/2}\cos^p t\,dt$$

We do this by partial integration. We have the following formula:

$$\begin{aligned}(\cos^p t\sin t)' &= p\cos^{p-1}t(-\sin t)\sin t+\cos^p t\cos t\\ &= p\cos^{p+1}t-p\cos^{p-1}t+\cos^{p+1}t\\ &= (p+1)\cos^{p+1}t-p\cos^{p-1}t\end{aligned}$$

By integrating between 0 and $\pi/2$, we obtain the following formula:

$$(p+1)I_{p+1}=pI_{p-1}$$

Thus we can compute I_p by recurrence, and we obtain:

$$\begin{aligned} I_p &= \frac{p-1}{p} I_{p-2} \\ &= \frac{p-1}{p} \cdot \frac{p-3}{p-2} I_{p-4} \\ &= \frac{p-1}{p} \cdot \frac{p-3}{p-2} \cdot \frac{p-5}{p-4} I_{p-6} \\ &\vdots \\ &= \frac{p!!}{(p+1)!!} I_{1-\varepsilon(p)} \end{aligned}$$

But $I_0 = \frac{\pi}{2}$ and $I_1 = 1$, so we get the result. As for the second formula, this follows from the first one, with $t = \frac{\pi}{2} - s$. Thus, we have proved both formulae in the statement. □

We can now compute the volume of the sphere, as follows:

THEOREM 11.13. *The volume of the unit sphere in $\mathbb{R}^N$ is given by*

$$V = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!}$$

with our usual convention $N!! = (N-1)(N-3)(N-5)\ldots$

PROOF. Let us denote by B^+ the positive part of the unit sphere, or rather unit ball B, obtained by cutting this unit ball in 2^N parts. At the level of volumes, we have:

$$V = 2^N V^+$$

We have the following computation, using spherical coordinates:

$$\begin{aligned} V^+ &= \int_{B^+} 1 \\ &= \int_0^1 \int_0^{\pi/2} \ldots \int_0^{\pi/2} r^{N-1} \sin^{N-2} t_1 \ldots \sin t_{N-2}\, dr dt_1 \ldots dt_{N-1} \\ &= \int_0^1 r^{N-1}\, dr \int_0^{\pi/2} \sin^{N-2} t_1\, dt_1 \ldots \int_0^{\pi/2} \sin t_{N-2} dt_{N-2} \int_0^{\pi/2} 1 dt_{N-1} \\ &= \frac{1}{N} \times \left(\frac{\pi}{2}\right)^{[N/2]} \times \frac{(N-2)!!}{(N-1)!!} \cdot \frac{(N-3)!!}{(N-2)!!} \ldots \frac{2!!}{3!!} \cdot \frac{1!!}{2!!} \cdot 1 \\ &= \frac{1}{N} \times \left(\frac{\pi}{2}\right)^{[N/2]} \times \frac{1}{(N-1)!!} \\ &= \left(\frac{\pi}{2}\right)^{[N/2]} \frac{1}{(N+1)!!} \end{aligned}$$

Here we have used the following formula, for computing the exponent of $\pi/2$:

$$\begin{aligned}\varepsilon(0)+\varepsilon(1)+\varepsilon(2)+\ldots+\varepsilon(N-2) &= 1+0+1+\ldots+\varepsilon(N-2)\\ &= \left[\frac{N-2}{2}\right]+1\\ &= \left[\frac{N}{2}\right]\end{aligned}$$

Thus, we obtain the formula in the statement. □

As main particular cases of the above formula, we have:

THEOREM 11.14. *The volumes of the low-dimensional spheres are as follows:*

(1) *At $N=1$, the length of the unit interval is $V=2$.*
(2) *At $N=2$, the area of the unit disk is $V=\pi$.*
(3) *At $N=3$, the volume of the unit sphere is $V=\frac{4\pi}{3}$*
(4) *At $N=4$, the volume of the corresponding unit sphere is $V=\frac{\pi^2}{2}$.*

PROOF. Some of these results are well-known, but we can obtain all of them as particular cases of the general formula in Theorem 11.13, as follows:

(1) At $N=1$ we obtain $V=1\cdot\frac{2}{1}=2$.

(2) At $N=2$ we obtain $V=\frac{\pi}{2}\cdot\frac{4}{2}=\pi$.

(3) At $N=3$ we obtain $V=\frac{\pi}{2}\cdot\frac{8}{3}=\frac{4\pi}{3}$.

(4) At $N=4$ we obtain $V=\frac{\pi^2}{4}\cdot\frac{16}{8}=\frac{\pi^2}{2}$. □

The formula in Theorem 11.13 is certainly nice, but in practice, we would like to have estimates for that sphere volumes too. For this purpose, we will need:

THEOREM 11.15. *We have the Stirling formula*

$$N!\simeq\left(\frac{N}{e}\right)^N\sqrt{2\pi N}$$

valid in the $N\to\infty$ limit.

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

(1) Let us first see what we can get with Riemann sums. We have:

$$\begin{aligned}\log(N!) &= \sum_{k=1}^{N}\log k\\ &\approx \int_1^N\log x\,dx\\ &= N\log N-N+1\end{aligned}$$

By exponentiating, this gives the following estimate, which is not bad:

$$N! \approx \left(\frac{N}{e}\right)^N \cdot e$$

(2) We can improve our estimate by replacing the rectangles from the Riemann sum approach to the integrals by trapezoids. In practice, this gives the following estimate:

$$\begin{aligned}
\log(N!) &= \sum_{k=1}^{N} \log k \\
&\approx \int_1^N \log x\, dx + \frac{\log 1 + \log N}{2} \\
&= N \log N - N + 1 + \frac{\log N}{2}
\end{aligned}$$

By exponentiating, this gives the following estimate, which gets us closer:

$$N! \approx \left(\frac{N}{e}\right)^N \cdot e \cdot \sqrt{N}$$

(3) In order to conclude, we must take some kind of mathematical magnifier, and carefully estimate the error made in (2). Fortunately, this mathematical magnifier exists, called Euler-Maclaurin formula, and after some computations, this leads to:

$$N! \simeq \left(\frac{N}{e}\right)^N \sqrt{2\pi N}$$

(4) However, all this remains a bit complicated, so we would like to present now an alternative approach to (3), which also misses some details, but better does the job, explaining where the $\sqrt{2\pi}$ factor comes from. First, by partial integration we have:

$$N! = \int_0^\infty x^N e^{-x} dx$$

Since the integrand is sharply peaked at $x = N$, as you can see by computing the derivative of $\log(x^N e^{-x})$, this suggests writing $x = N + y$, and we obtain:

$$\begin{aligned}
\log(x^N e^{-x}) &= N \log x - x \\
&= N \log(N + y) - (N + y) \\
&= N \log N + N \log\left(1 + \frac{y}{N}\right) - (N + y) \\
&\simeq N \log N + N\left(\frac{y}{N} - \frac{y^2}{2N^2}\right) - (N + y) \\
&= N \log N - N - \frac{y^2}{2N}
\end{aligned}$$

By exponentiating, we obtain from this the following estimate:

$$x^N e^{-x} \simeq \left(\frac{N}{e}\right)^N e^{-y^2/2N}$$

Now by integrating, and using the Gauss formula, we obtain from this:

$$\begin{aligned} N! &= \int_0^\infty x^N e^{-x} dx \\ &\simeq \int_{-N}^N \left(\frac{N}{e}\right)^N e^{-y^2/2N}\, dy \\ &\simeq \left(\frac{N}{e}\right)^N \int_{\mathbb{R}} e^{-y^2/2N}\, dy \\ &= \left(\frac{N}{e}\right)^N \sqrt{2N} \int_{\mathbb{R}} e^{-z^2}\, dz \\ &= \left(\frac{N}{e}\right)^N \sqrt{2\pi N} \end{aligned}$$

Thus, we have proved the Stirling formula, as formulated in the statement. □

With the above formula in hand, we have many useful applications, such as:

PROPOSITION 11.16. *We have the following estimate for binomial coefficients,*

$$\binom{N}{K} \simeq \left(\frac{1}{t^t(1-t)^{1-t}}\right)^N \frac{1}{\sqrt{2\pi t(1-t)N}}$$

in the $K \simeq tN \to \infty$ limit, with $t \in (0,1]$. In particular we have

$$\binom{2N}{N} \simeq \frac{4^N}{\sqrt{\pi N}}$$

in the $N \to \infty$ limit, for the central binomial coefficients.

PROOF. All this is very standard, by using the Stirling formula etablished above, for the various factorials which appear, the idea being as follows:

(1) This follows from the definition of the binomial coefficients, namely:

$$\begin{aligned}\binom{N}{K} &= \frac{N!}{K!(N-K)!}\\ &\simeq \left(\frac{N}{e}\right)^N\sqrt{2\pi N}\left(\frac{e}{K}\right)^K\frac{1}{\sqrt{2\pi K}}\left(\frac{e}{N-K}\right)^{N-K}\frac{1}{\sqrt{2\pi(N-K)}}\\ &= \frac{N^N}{K^K(N-K)^{N-K}}\sqrt{\frac{N}{2\pi K(N-K)}}\\ &\simeq \frac{N^N}{(tN)^{tN}((1-t)N)^{(1-t)N}}\sqrt{\frac{N}{2\pi tN(1-t)N}}\\ &= \left(\frac{1}{t^t(1-t)^{1-t}}\right)^N\frac{1}{\sqrt{2\pi t(1-t)N}}\end{aligned}$$

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

(2) This estimate follows from a similar computation, as follows:

$$\begin{aligned}\binom{2N}{N} &= \frac{(2N)!}{N!N!}\\ &\simeq \left(\frac{2N}{e}\right)^{2N}\sqrt{4\pi N}\left(\frac{e}{N}\right)^{2N}\frac{1}{2\pi N}\\ &= \frac{4^N}{\sqrt{\pi N}}\end{aligned}$$

Alternatively, we can take $t=1/2$ in (1), then rescale. Indeed, we have:

$$\begin{aligned}\binom{N}{[N/2]} &\simeq \left(\frac{1}{(\frac{1}{2})^{1/2}(\frac{1}{2})^{1/2}}\right)^N\frac{1}{\sqrt{2\pi\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot N}}\\ &= 2^N\sqrt{\frac{2}{\pi N}}\end{aligned}$$

Thus with the change $N\to 2N$ we obtain the formula in the statement. □

We can now estimate the volumes of the spheres, as follows:

THEOREM 11.17. *The volume of the unit sphere in $\mathbb{R}^N$ is given by*

$$V\simeq\left(\frac{2\pi e}{N}\right)^{N/2}\frac{1}{\sqrt{\pi N}}$$

in the $N\to\infty$ limit.

PROOF. We use the formula for V found in Theorem 11.13, namely:

$$V = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!}$$

In the case where N is even, the estimate goes as follows:

$$\begin{aligned} V &= \left(\frac{\pi}{2}\right)^{N/2} \frac{2^N}{(N+1)!!} \\ &\simeq \left(\frac{\pi}{2}\right)^{N/2} 2^N \left(\frac{e}{N}\right)^{N/2} \frac{1}{\sqrt{\pi N}} \\ &= \left(\frac{2\pi e}{N}\right)^{N/2} \frac{1}{\sqrt{\pi N}} \end{aligned}$$

In the case where N is odd, the estimate goes as follows:

$$\begin{aligned} V &= \left(\frac{\pi}{2}\right)^{(N-1)/2} \frac{2^N}{(N+1)!!} \\ &\simeq \left(\frac{\pi}{2}\right)^{(N-1)/2} 2^N \left(\frac{e}{N}\right)^{N/2} \frac{1}{\sqrt{2N}} \\ &= \sqrt{\frac{2}{\pi}} \left(\frac{2\pi e}{N}\right)^{N/2} \frac{1}{\sqrt{2N}} \\ &= \left(\frac{2\pi e}{N}\right)^{N/2} \frac{1}{\sqrt{\pi N}} \end{aligned}$$

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

Getting back now to our main result so far, Theorem 11.13, we can compute in the same way the area of the sphere, the result being as follows:

THEOREM 11.18. *The area of the unit sphere in $\mathbb{R}^N$ is given by*

$$A = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N-1)!!}$$

with the our usual convention for double factorials, namely:

$$N!! = (N-1)(N-3)(N-5)\ldots$$

In particular, at $N = 2, 3, 4$ we obtain respectively $A = 2\pi, 4\pi, 2\pi^2$.

PROOF. Regarding the first assertion, there is no need to compute again, because the formula in the statement can be deduced from Theorem 11.13, as follows:

(1) We can either use the "pizza" argument from plane geometry, which shows that the area and volume of the sphere in $\mathbb{R}^N$ are related by the following formula:

$$A = N \cdot V$$

Together with the formula in Theorem 11.13 for V, this gives the result.

(2) Or, we can start the computation in the same way as we started the proof of Theorem 11.13, the beginning of this computation being as follows:

$$vol(S^+) = \int_0^{\pi/2} \ldots \int_0^{\pi/2} \sin^{N-2} t_1 \ldots \sin t_{N-2}\, dt_1 \ldots dt_{N-1}$$

Now by comparing with the beginning of the proof of Theorem 11.13, the only thing that changes is the following quantity, which now dissapears:

$$\int_0^1 r^{N-1}\, dr = \frac{1}{N}$$

Thus, we have $vol(S^+) = N \cdot vol(B^+)$, and so we obtain the following formula:

$$vol(S) = N \cdot vol(B)$$

But this means $A = N \cdot V$, and together with the formula in Theorem 11.13 for V, this gives the result. As for the last assertion, this can be either worked out directly, or deduced from the results for volumes that we have so far, by multiplying by N. □

11d. Normal variables

We have kept the best for the end. As a starting point, we have:

DEFINITION 11.19. *Let X be a probability space, that is, a space with a probability measure, and with the corresponding integration denoted E, and called expectation.*

(1) *The random variables are the real functions $f \in L^\infty(X)$.*
(2) *The moments of such a variable are the numbers $M_k(f) = E(f^k)$.*
(3) *The law of such a variable is the measure given by $M_k(f) = \int_{\mathbb{R}} x^k d\mu_f(x)$.*

Here the fact that a measure μ_f as above exists indeed is not exactly trivial. But we can do this by looking at formulae of the following type:

$$E(\varphi(f)) = \int_{\mathbb{R}} \varphi(x) d\mu_f(x)$$

Indeed, having this for monomials $\varphi(x) = x^n$, as above, is the same as having it for polynomials $\varphi \in \mathbb{R}[X]$, which in turn is the same as having it for the characteristic functions $\varphi = \chi_I$ of measurable sets $I \subset \mathbb{R}$. Thus, in the end, what we need is:

$$P(f \in I) = \mu_f(I)$$

But this formula can serve as a definition for μ_f, and we are done.

Regarding now independence, we can formulate here the following definition:

DEFINITION 11.20. *Two variables $f, g \in L^\infty(X)$ are called independent when*

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

happens, for any $k, l \in \mathbb{N}$.

Again, this definition hides some non-trivial things, the idea being a bit as before, namely that of looking at formulae of the following type:

$$E[\varphi(f)\psi(g)] = E[\varphi(f)]\, E[\psi(g)]$$

To be more precise, passing as before from monomials to polynomials, then to characteristic functions, we are led to the usual definition of independence, namely:

$$P(f \in I, g \in J) = P(f \in I)\, P(g \in J)$$

As a first result now, which is something very standard, we have:

THEOREM 11.21. *Assuming that $f, g \in L^\infty(X)$ are independent, we have*

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

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

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

$$\int_{\mathbb{R}} x^k d\mu_{f+g}(x) = E((f+g)^k) = \sum_r \binom{k}{r} M_r(f) M_{k-r}(g)$$

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

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

Thus μ_{f+g} and $\mu_f * \mu_g$ have the same moments, so they coincide, as claimed. □

As a second result on independence, which is more advanced, we have:

THEOREM 11.22. *Assuming that $f, g \in L^\infty(X)$ are independent, we have*

$$F_{f+g} = F_f F_g$$

where $F_f(x) = E(e^{ixf})$ is the Fourier transform.

PROOF. This is something which is very standard too, coming from:

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

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

Let us introduce now the normal laws. This can be done as follows:

DEFINITION 11.23. *The normal law of parameter* 1 *is the following measure:*

$$g_1 = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

More generally, the normal law of parameter $t > 0$ *is the following measure:*

$$g_t = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t} dx$$

These are also called Gaussian distributions, with "g" standing for Gauss.

Observe that the above laws have indeed mass 1, as they should. This follows indeed from the Gauss formula, which gives, with $x = \sqrt{2t}\, y$:

$$\begin{aligned} \int_{\mathbb{R}} e^{-x^2/2t} dx &= \int_{\mathbb{R}} e^{-y^2} \sqrt{2t}\, dy \\ &= \sqrt{2t} \int_{\mathbb{R}} e^{-y^2} dy \\ &= \sqrt{2t} \times \sqrt{\pi} \\ &= \sqrt{2\pi t} \end{aligned}$$

Generally speaking, the normal laws appear as bit everywhere, in real life. The reasons behind this phenomenon come from the Central Limit Theorem (CLT), that we will explain in a moment, after developing some general theory. As a first result, we have:

PROPOSITION 11.24. *We have the variance formula*

$$V(g_t) = t$$

valid for any $t > 0$.

PROOF. The first moment is 0, because our normal law g_t is centered. As for the second moment, this can be computed as follows:

$$\begin{aligned} M_2 &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} x^2 e^{-x^2/2t}dx \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} (tx)\left(-e^{-x^2/2t}\right)' dx \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} te^{-x^2/2t}dx \\ &= t \end{aligned}$$

We conclude from this that the variance is $V = M_2 = t$. □

Here is another result, which is the key one for the study of the normal laws:

THEOREM 11.25. *We have the following formula, valid for any $t > 0$:*

$$F_{g_t}(x) = e^{-tx^2/2}$$

In particular, the normal laws satisfy $g_s * g_t = g_{s+t}$, for any $s, t > 0$.

PROOF. The Fourier transform formula can be established as follows:

$$\begin{aligned} F_{g_t}(x) &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} e^{-y^2/2t+ixy}dy \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} e^{-(y/\sqrt{2t}-\sqrt{t/2}ix)^2-tx^2/2}dy \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} e^{-z^2-tx^2/2}\sqrt{2t}dz \\ &= \frac{1}{\sqrt{\pi}}e^{-tx^2/2}\int_{\mathbb{R}} e^{-z^2}dz \\ &= \frac{1}{\sqrt{\pi}}e^{-tx^2/2}\cdot\sqrt{\pi} \\ &= e^{-tx^2/2} \end{aligned}$$

As for the last assertion, this follows from the fact that $\log F_{g_t}$ is linear in t. □

We are now ready to state and prove the CLT, as follows:

THEOREM 11.26 (CLT). *Given random variables $f_1, f_2, f_3, \ldots \in L^\infty(X)$ which are i.i.d., centered, and with variance $t > 0$, we have, with $n \to \infty$, in moments,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f_i \sim g_t$$

where g_t is the Gaussian law of parameter t, having as density $\frac{1}{\sqrt{2\pi t}}e^{-y^2/2t}dy$.

PROOF. We use the Fourier transform, which is by definition given by:

$$F_f(x) = E(e^{ixf})$$

In terms of moments, we have the following formula:

$$\begin{aligned} F_f(x) &= E\left(\sum_{k=0}^{\infty}\frac{(ixf)^k}{k!}\right)\\ &= \sum_{k=0}^{\infty}\frac{(ix)^k E(f^k)}{k!}\\ &= \sum_{k=0}^{\infty}\frac{i^k M_k(f)}{k!}\,x^k \end{aligned}$$

Thus, the Fourier transform of the variable in the statement is:

$$\begin{aligned} F(x) &= \left[F_f\left(\frac{x}{\sqrt{n}}\right)\right]^n\\ &= \left[1-\frac{tx^2}{2n}+O(n^{-2})\right]^n\\ &\simeq \left[1-\frac{tx^2}{2n}\right]^n\\ &\simeq e^{-tx^2/2} \end{aligned}$$

But this latter function being the Fourier transform of g_t, we obtain the result. □

Let us discuss now some further properties of the normal law. We first have:

PROPOSITION 11.27. *The even moments of the normal law are the numbers*

$$M_k(g_t) = t^{k/2}\times k!!$$

where $k!! = (k-1)(k-3)(k-5)\ldots$, *and the odd moments vanish.*

PROOF. We have the following computation, valid for any integer $k\in\mathbb{N}$:

$$\begin{aligned} M_k &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} y^k e^{-y^2/2t}dy\\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} (ty^{k-1})\left(-e^{-y^2/2t}\right)'dy\\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} t(k-1)y^{k-2}e^{-y^2/2t}dy\\ &= t(k-1)\times\frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} y^{k-2}e^{-y^2/2t}dy\\ &= t(k-1)M_{k-2} \end{aligned}$$

Now recall from the proof of Proposition 11.24 that we have $M_0 = 1$, $M_1 = 0$. Thus by recurrence, we are led to the formula in the statement. □

We have the following alternative formulation of the above result:

PROPOSITION 11.28. *The moments of the normal law are the numbers*

$$M_k(g_t) = t^{k/2}|P_2(k)|$$

where $P_2(k)$ is the set of pairings of $\{1, \ldots, k\}$.

PROOF. Let us count the pairings of $\{1, \ldots, k\}$. In order to have such a pairing, we must pair 1 with one of the numbers $2, \ldots, k$, and then use a pairing of the remaining $k - 2$ numbers. Thus, we have the following recurrence formula:

$$|P_2(k)| = (k - 1)|P_2(k - 2)|$$

As for the initial data, this is $P_1 = 0$, $P_2 = 1$. Thus, we are led to the result. □

We are not done yet, and here is one more improvement of the above:

THEOREM 11.29. *The moments of the normal law are the numbers*

$$M_k(g_t) = \sum_{\pi \in P_2(k)} t^{|\pi|}$$

where $P_2(k)$ is the set of pairings of $\{1, \ldots, k\}$, and $|.|$ is the number of blocks.

PROOF. This follows indeed from Proposition 11.28, because the number of blocks of a pairing of $\{1, \ldots, k\}$ is trivially $k/2$, independently of the pairing. □

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

11e. Exercises

Exercises:

EXERCISE 11.30.

EXERCISE 11.31.

EXERCISE 11.32.

EXERCISE 11.33.

EXERCISE 11.34.

EXERCISE 11.35.

EXERCISE 11.36.

EXERCISE 11.37.

Bonus exercise.

CHAPTER 12

Charges, matter

12a. Electrons, charges

Welcome to advanced physics. That needs electrons, and here, you don't necessarily need a power outlet for having them, a basic Van de Graaff machine, or just rubbing some suitable materials together, will do. Let us record this finding as follows:

FACT 12.1. *Each piece of matter has a charge $q \in \mathbb{R}$, which is normally neutral, $q = 0$, but that we can make positive or negative, by using various methods. We say that responsible for the charge is the amount of electrons present, as follows:*

(1) *When the matter lacks electrons, the charge is positive, $q > 0$.*
(2) *When there are more electrons than needed, the charge is negative, $q < 0$.*

And, good news, this will be the starting point for the considerations in this book, the electrons, as defined above. Of course you might say, for instance if you are a math student used to a fair amount of exactness, in your learning, that what we say in Fact 12.1 is a bit borderline, for something to be labeled as axiomatic. But well, physics is not mathematics, it's sort of harder, when it comes to having things started, and that's what we have. Of course we will be back to it, with axioms, later, that is promised.

Moving ahead now, 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 12.1 is indeed more than borderline, as axiomatics for a theory, we have:

FACT 12.2 (Coulomb law). *Any pair of charges $q_1, q_2 \in \mathbb{R}$ is subject to a force as follows, which is attractive if $q_1 q_2 < 0$ and repulsive if $q_1 q_2 > 0$,*

$$||F|| = K \cdot \frac{|q_1 q_2|}{d^2}$$

where $d > 0$ is the distance between the charges, and $K > 0$ is a certain constant.

Observe the amazing similarity with the Newton law for gravity. However, as we will discover soon, passed a few simple facts, things will be far more complicated here.

As in the gravity case, the force F appearing above is understood to be parallel to the vector $x_2 - x_1 \in \mathbb{R}^3$ joining as $x_1 \to x_2$ the locations $x_1, x_2 \in \mathbb{R}^3$ of our charges, and by taking into account the attraction/repulsion rules above, we have:

PROPOSITION 12.3. *The Coulomb force of q_1 at x_1 acting on q_2 at x_2 is*

$$F = K \cdot \frac{q_1 q_2 (x_2 - x_1)}{||x_2 - x_1||^3}$$

with $K > 0$ being the Coulomb constant, as above.

PROOF. We have indeed the following computation:

$$\begin{aligned} F &= sgn(q_1 q_2) \cdot ||F|| \cdot \frac{x_2 - x_1}{||x_2 - x_1||} \\ &= sgn(q_1 q_2) \cdot K \cdot \frac{|q_1 q_2|}{||x_2 - x_1||^2} \cdot \frac{x_2 - x_1}{||x_2 - x_1||} \\ &= K \cdot \frac{q_1 q_2 (x_2 - x_1)}{||x_2 - x_1||^3} \end{aligned}$$

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

In relation with the value of the constant K appearing in the above, called Coulomb constant, things are a bit tricky, as follows:

FACT 12.4. *The Coulomb constant K is given by the formula*

$$K = 8.987\ 551\ 7923(14) \times 10^9$$

in standard units, with the charges being measured in coulombs C, given by

$$1C \simeq 6.241\ 509 \times 10^{18}\, e$$

where e is the elementary charge, namely minus that of an electron.

There are in fact several interesting things going on here. First, at the end you would say why not simply saying that e is the charge of the proton $+$, but the thing is that the proton $+$ and the electron $-$ do not have in fact the same exact charge, with sign switched, and the electron was preferred, as always, over the proton for formulating things.

Which takes us into the question of why the charge of the electron is $-$, instead of $+$. And there is a long story here, involving debates among the 18th century greats, and with a little bit of confusion being involved too, because the electrons $-$ are attracted by positive charges $q > 0$, and so observed around these positive charges $q > 0$, which might lead to the idea that they might have themselves a positive charge $+$, contributing to $q > 0$. Benjamin Franklin is generally credited for the $-$ convention.

Things were later restored in the early 20th century, with the atomic theory of Bohr and others, where electrons $-$ spin around a proton and neutron core $q > 0$, and with this picture, including the signs, looking like something very reasonable.

Passed all this, another peculiarity of Fact 12.4 comes in relation with the definition of the coulomb, which is in fact given by definition by an exact formula, namely:

$$1C = \frac{5 \times 10^{18}}{0.801\ 088\ 317}\ e$$

This in practice gives the following more precise formula for the coulomb, which shows that a charge of $1C$ is something fractionary, that cannot be realized in real life:

$$1C = 6241\ 509\ 074\ 460\ 762\ 607.776\, e$$

The problem comes from the following alternative definition of the coulomb, in terms of the ampere, which is something more complicated, that we will talk about later:

$$1C = 1A \cdot 1s$$

Hang on, we are not done yet. Adding to the confusion, the Coulomb constant is usually denoted K, but also k, or most often k_e, but in fact the most often is written in the following form, with ε_0 being the so-called permittivity of free space:

$$K = \frac{1}{4\pi\varepsilon_0}$$

And the story is not over here, because ε_0 itself is given by the following formula, with μ_0 being the magnetic permeability of free space, and c being the speed of light:

$$\varepsilon_0 = \frac{1}{\mu_0 c^2}$$

And we are surely still not done, because all the above discussion assumes that the other units that are used are standard, namely meter and second, and this is not always standard, due to the about 50 orders of magnitude physics has to deal with.

In any case, let us end this interesting discussion about units with something concrete, useful, and very illustrating, in relation with gravity, as follows:

THEOREM 12.5. *The electrical repulsion between two electrons is about*

$$R = 10^{42}$$

times bigger than their gravitational attraction.

PROOF. Consider indeed two electrons, having masses m, m and charges $-e, -e$. The magnitudes of the electric repulsion F_e and gravity attraction F_g are given by:

$$||F_e|| = \frac{Ke^2}{d^2} \quad , \quad ||F_g|| = \frac{Gm^2}{d^2}$$

Thus the ratio of forces R that we want to measure is given by:

$$R = \frac{||F_e||}{||F_g||} = \frac{Ke^2}{Gm^2}$$

Regarding now the data, this is as follows, with m at rest, and in standard units, namely meters and seconds, also kilograms, and including now coulombs too:

$$K = 8.897 \times 10^9 \quad , \quad G = 6.674 \times 10^{-11}$$

$$e = 1.602 \times 10^{-19} \quad , \quad m = 9.109 \times 10^{-31}$$

We obtain the following approximation for the ratio R considered above:

$$\begin{aligned} R &= \frac{8.897 \times 1.602^2}{6.674 \times 9.109^2} \times \frac{10^9 \times 10^{-38}}{10^{-11} \times 10^{-62}} \\ &= (4.123 \times 10^{-2}) \times 10^{44} \\ &\simeq 10^{42} \end{aligned}$$

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

For adding to the picture, and in order to fully understand what that $R = 10^{42}$ number that we found truly means, let us complement the above result with:

PROPOSITION 12.6. *The universe, or at least the known universe, is about*

$$r = 10^{37}$$

bigger than a hydrogen atom, with this ratio being $10,000$ *smaller than* R.

PROOF. The radius of the hydrogen atom can be anywhere between $25-120$ pm, with 1 pm $= 10^{-12}$ m, depending on the convention used, with a commonly accepted figure being 53 pm, representing the mean distance between the proton and the electron. As for the radius of the known universe, again there is a story here, with a commonly accepted figure being 4.4×10^{26} m. Thus the ratio that we are interested in is:

$$r = \frac{4.4 \times 10^{26}}{53 \times 10^{-12}} \simeq 10^{37}$$

And this is 10,000 smaller than 10^{42}, as claimed. □

As a side comment, however, when speaking masses instead of sizes, the number $R = 10^{42}$ pales when compared to the mass of the known universe, counting ordinary mass only, accounting for 4.9%, divided by the mass of a hydrogen atom, which is:

$$\mathfrak{R} = \frac{1.5 \times 10^{53}}{1.8 \times 10^{-30}} \simeq 10^{83}$$

Getting back now to Theorem 12.5 as it is, let us point out that this is something not at all anecdotical, even in the context of the most abstract theoretical physics that you can ever imagine, not to say pure mathematics, because of the following rule of thumb, which is something widely agreed upon, by most of the scientists:

RULE 12.7. *Don't ever expect the mathematics and physics to be the same, over* 10 *orders of magnitude or so.*

In other words, with this in hand, Theorem 12.5 tells us an interesting thing, namely that the mathematics and physics of the Coulomb force $F_e \sim 1/d^2$ will be in fact very different from the mathematics and physics of the Newton force $F_g \sim 1/d^2$. We will see in what follows that indeed it is so, but it is of course far better to be warned in advance of the potential difficulties on the way. So, Theorem 12.5 is something very smart.

12b. The Gauss law

Let us develop now the basic mathematics for electrostatics. We first have:

DEFINITION 12.8. *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 $F = QE$.

Observe the analogy with gravity, save for the fact that instead of masses $m > 0$ we have now charges $q \in \mathbb{R}$, and that at the level of constants, G gets replaced by K.

More generally, we will be interested in electric fields of various non-discrete configurations of charges, such as charged curves, surfaces and solid bodies. We have already talked about such things in the above, in the gravitational context, but the discussion there, involving the gravitational force of a solid body having non-trivial shape or density, was something rather specialized.

In the electricity context, however, 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 12.9. *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:

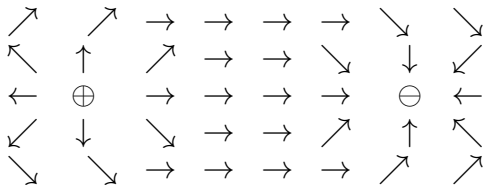
DEFINITION 12.10. *The field lines of $E : \mathbb{R}^3 \to \mathbb{R}^3$ are the oriented curves*

$$\gamma \subset \mathbb{R}^3$$

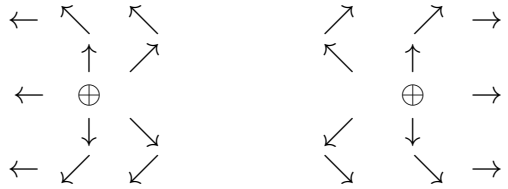
pointing at every point $x \in \mathbb{R}^3$ at the direction of the field, $E(x) \in \mathbb{R}^3$.

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.

In general now, the first thing that can be said about the field lines is that, by definition, they do not cross. Thus, what we have here is some sort of oriented 1D foliation of $\mathbb{R}^3$, in the sense that $\mathbb{R}^3$ is smoothly decomposed into oriented curves $\gamma \subset \mathbb{R}^3$.

The field lines, as constructed in Definition 12.10, 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 12.11. *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. We have two assertions here, the idea being as follows:

(1) The first assertion is clear from definitions, with of course our usual convention that the electric field and its problematics take place outside the locations of the charges, which makes everything in the statement to be indeed well-defined.

(2) Regarding now the last assertion, which is of course a bit informal, this follows from the above discussion. It is possible to be a bit more mathematical here, with a definition, formula and everything, but we will not need this, in what follows. □

Let us introduce now a key definition, as follows:

DEFINITION 12.12. *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.

Regarding the last sentence of Definition 12.12, this is of course something informal, meant to help, coming from the interpretation of the field lines from Proposition 12.11. However, we will see later that this simple interpretation can be of great use.

As a first observation, we could have done of course the same thing with gravity before, but these notions of field lines and flux are not very interesting, in that context.

In the present setting, however, electric fields passing through metal sheets are a common occurence, and all the above is important, for any application.

As a first illustration, let us do a basic computation, as follows:

PROPOSITION 12.13. *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:

$$\begin{aligned}
\Phi_E(S) &= \int_S < E(x), n(x) > dx \\
&= \int_S \left\langle \frac{Kqx}{r^3}, \frac{x}{r} \right\rangle dx \\
&= \int_S \frac{Kq}{r^2}\, dx \\
&= \frac{Kq}{r^2} \times 4\pi r^2 \\
&= 4\pi K q
\end{aligned}$$

Thus with $\varepsilon_0 = 1/(4\pi K)$ as above, we obtain the result. □

As a comment here, the constant $\varepsilon_0 = 1/(4\pi K)$ which appears in the above is the permittivity of free space constant that we talked about before, when discussing units. In what follows we will use this new constant instead of the Coulomb constant K.

More generally now, we have the following result:

THEOREM 12.14. *The flux of a field E through a sphere S is given by*

$$\Phi_E(S) = \frac{Q_{enc}}{\varepsilon_0}$$

where Q_{enc} is the total charge enclosed by S, and $\varepsilon_0 = 1/(4\pi K)$.

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

(1) Before jumping into computations, let us do some manipulations. First, by discretizing the problem, we can assume that we are dealing with a system of point charges. Moreover, by additivity, we can assume that we are dealing with a single charge. And if we denote by $q \in \mathbb{R}$ this charge, located at $v \in \mathbb{R}^3$, we want to prove that we have the following formula, where $B \subset \mathbb{R}^3$ denotes the ball enclosed by S:

$$\Phi_E(S) = \frac{q}{\varepsilon_0}\, \delta_{v \in B}$$

(2) By linearity we can assume that we are dealing with the unit sphere S. Moreover, by rotating we can assume that our charge q lies on the Ox axis, that is, that we have $v = (r, 0, 0)$ with $r \geq 0$, $r \neq 1$. The formula that we want to prove becomes:

$$\Phi_E(S) = \frac{q}{\varepsilon_0}\,\delta_{r<1}$$

(3) Let us start now the computation. With $u = (x, y, z)$, we have:

$$\begin{aligned}
\Phi_E(S) &= \int_S < E(u), u > du \\
&= \int_S \left\langle \frac{Kq(u-v)}{||u-v||^3}, u \right\rangle du \\
&= Kq \int_S \frac{< u-v, u >}{||u-v||^3}\, du \\
&= Kq \int_S \frac{1 - < v, u >}{||u-v||^3}\, du \\
&= Kq \int_S \frac{1 - rx}{(1 - 2xr + r^2)^{3/2}}\, du
\end{aligned}$$

(4) In order to compute the above integral, we will use spherical coordinates for the unit sphere S, which are as follows, with $s \in [0, \pi]$ and $t \in [0, 2\pi]$:

$$\begin{cases}
x &= \cos s \\
y &= \sin s \cos t \\
z &= \sin s \sin t
\end{cases}$$

The corresponding Jacobian is readily computed, as follows:

$$\begin{aligned}
J &= \begin{vmatrix} \cos s & -\sin s & 0 \\ \sin s \cos t & \cos s \cos t & -\sin s \sin t \\ \sin s \sin t & \cos s \sin t & \sin s \cos t \end{vmatrix} \\
&= \sin s \sin t \begin{vmatrix} \cos s & -\sin s \\ \sin s \sin t & \cos s \sin t \end{vmatrix} + \sin s \cos t \begin{vmatrix} \cos s & -\sin s \\ \sin s \cos t & \cos s \cos t \end{vmatrix} \\
&= \sin s(\sin^2 t + \cos^2 t) \begin{vmatrix} \cos s & -\sin s \\ \sin s & \cos s \end{vmatrix} \\
&= \sin s
\end{aligned}$$

(5) With the above change of coordinates, our integral from (3) becomes:

$$\begin{aligned}\Phi_E(S) &= Kq\int_S \frac{1-rx}{(1-2xr+r^2)^{3/2}}\,du\\ &= Kq\int_0^{2\pi}\int_0^{\pi}\frac{1-r\cos s}{(1-2r\cos s+r^2)^{3/2}}\cdot\sin s\,ds\,dt\\ &= 2\pi Kq\int_0^{\pi}\frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}}\,ds\\ &= \frac{q}{2\varepsilon_0}\int_0^{\pi}\frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}}\,ds\end{aligned}$$

(6) The point now is that the integral on the right can be computed with the change of variables $x=\cos s$. Indeed, we have $dx=-\sin s\,ds$, and we obtain:

$$\begin{aligned}\int_0^{\pi}\frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}}\,ds &= \int_{-1}^{1}\frac{1-rx}{(1-2rx+r^2)^{3/2}}\,dx\\ &= \left[\frac{x-r}{\sqrt{1-2rx+r^2}}\right]_{-1}^{1}\\ &= \frac{1-r}{\sqrt{1-2r+r^2}}-\frac{-1-r}{\sqrt{1+2r+r^2}}\\ &= \frac{1-r}{|1-r|}+1\\ &= 2\delta_{r<1}\end{aligned}$$

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

As a comment here, at $r=1$, which is normally avoided by our problematics, the integral I_r computed in (5) above converges too, and can be evaluated as follows:

$$I_1=\left[\frac{x-1}{\sqrt{2-2x}}\right]_{-1}^{1}=\left[-\sqrt{\frac{1-x}{2}}\ \right]_{-1}^{1}=1$$

Thus, we have the correct middle step between the $0,2$ values of the integral I_r, and getting back now to the flux, at $r=1$ we formally have $\Phi_E(S)=q/(2\varepsilon_0)$, which again is the correct middle step between the $0,q/\varepsilon_0$ values of the flux.

Even more generally now, we have the following result, due to Gauss, which is the foundation of advanced electrostatics, and of everything following from it, namely electrodynamics, and then quantum mechanics, and particle physics:

THEOREM 12.15 (Gauss law). *The flux of a field E through a surface S is given by*

$$\Phi_E(S) = \frac{Q_{enc}}{\varepsilon_0}$$

where Q_{enc} is the total charge enclosed by S, and $\varepsilon_0 = 1/(4\pi K)$.

PROOF. This basically follows from Theorem 12.14, or even from Proposition 12.13, by adding to the results there a number of new ingredients, as follows:

(1) Our first claim is that given a closed surface S, with no charges inside, the flux through it of any choice of external charges vanishes:

$$\Phi_E(S) = 0$$

This claim is indeed supported by the intuitive interpretation of the flux, as corresponding to the signed number of field lines crossing S. Indeed, any field line entering as $+$ must exit somewhere as $-$, and vice versa, so when summing we get 0.

(2) In practice now, in order to prove this rigorously, there are several ways. A first argument, which is quite elementary, is the one used by Feynman in [34], based on the fact that, due to $F \sim 1/d^2$, local deformations of S will leave invariant the flux, and so in the end we are left with a rotationally invariant surface, where the result is clear.

(3) A second argument, which basically uses the same idea, but is perhaps a bit more robust, is by redoing the computations in the proof of Theorem 12.14, by assuming this time that the integration takes place on an arbitrary surface as follows:

$$S_\lambda = \left\{ \lambda(u)u \middle| u \in S \right\}$$

To be more precise, here $\lambda : S \to (0, \infty)$ is a certain function, defining the surface, whose derivatives will appear both in the construction of the normal vectors $n(x)$ with $x = \lambda(u)u$, and in the Jacobian of the change of variables $x \to u$, and in the end, when integrating over S as in the proof of Theorem 12.14, this function λ dissapears.

(4) A third argument, used by basically all electrodynamics books at the graduate level, and by some undergraduate books too, is by using heavy calculus, namely partial integration in 3D, and we will discuss this later, more in detail, a bit later.

(5) A fourth argument is by following the idea in (1), namely carefully axiomatizing the field lines, and their relation with the field, and then obtaining $\Phi_E(S) = 0$ by using the in-and-out trick in (1), as explained for instance by Griffiths in [44].

(6) To summmarize, we are led to the conclusion that given a closed surface S, with no charges inside, the flux through it of any choice of external charges vanishes:

$$\Phi_E(S) = 0$$

(7) The point now is that, with this and Proposition 12.13 in hand, we can finish by using a standard math trick. Let us assume indeed, by discretizing, that our system of

charges is discrete, consisting of enclosed charges $q_1, \ldots, q_k \in \mathbb{R}$, and an exterior total charge Q_{ext}. We can surround each of $q_1, \ldots, q_k$ by small disjoint spheres $U_1, \ldots, U_k$, chosen such that their interiors do not touch S, and we have:

$$\begin{aligned}
\Phi_E(S) &= \Phi_E(S - \cup U_i) + \Phi_E(\cup U_i) \\
&= 0 + \Phi_E(\cup U_i) \\
&= \sum_i \Phi_E(U_i) \\
&= \sum_i \frac{q_i}{\varepsilon_0} \\
&= \frac{Q_{enc}}{\varepsilon_0}
\end{aligned}$$

(8) 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 (6). Finally, the end makes use of Proposition 12.13. □

In order to reach to a better understanding of the Gauss law, mathematically speaking, let us start with a standard definition, immersing us into 3D problematics, as follows:

DEFINITION 12.16. *Given a function $f : \mathbb{R}^3 \to \mathbb{R}$, its usual derivative $f'(u) \in \mathbb{R}^3$ can be written as $f'(u) = \nabla f(u)$, where the gradient operator ∇ is given by:*

$$\nabla = \begin{pmatrix} \frac{d}{dx} \\ \frac{d}{dy} \\ \frac{d}{dz} \end{pmatrix}$$

By using ∇, we can talk about the divergence of a function $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$, as being

$$< \nabla, \varphi >= \left\langle \begin{pmatrix} \frac{d}{dx} \\ \frac{d}{dy} \\ \frac{d}{dz} \end{pmatrix}, \begin{pmatrix} \varphi_x \\ \varphi_y \\ \varphi_z \end{pmatrix} \right\rangle = \frac{d\varphi_x}{dx} + \frac{d\varphi_y}{dy} + \frac{d\varphi_z}{dz}$$

as well as about the curl of the same function $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$, as being

$$\nabla \times \varphi = \begin{vmatrix} u_x & \frac{d}{dx} & \varphi_x \\ u_y & \frac{d}{dy} & \varphi_y \\ u_z & \frac{d}{dz} & \varphi_z \end{vmatrix} = \begin{pmatrix} \frac{d\varphi_z}{dy} - \frac{d\varphi_y}{dz} \\ \frac{d\varphi_x}{dz} - \frac{d\varphi_z}{dx} \\ \frac{d\varphi_y}{dx} - \frac{d\varphi_x}{dy} \end{pmatrix}$$

where u_x, u_y, u_z are the unit vectors along the coordinate directions x, y, z.

All this might seem a bit abstract, but is in fact very intuitive. The gradient ∇f points in the direction of the maximal increase of f, with $|\nabla f|$ giving you the rate of increase

of f, in that direction. As for the divergence and curl, these measure the divergence and curl of the vectors $\varphi(u+v)$ around a given point $u \in \mathbb{R}^3$, in a usual, real-life sense.

Getting back now to calculus tools, what was missing from our picture was the higher dimensional analogue of the fundamental theorem of calculus, and more generally of the partial integration formula. In 3 dimensions, we have the following result:

THEOREM 12.17. *The following results hold, in* 3 *dimensions:*

(1) *Fundamental theorem for gradients, namely*

$$\int_a^b < \nabla f, dx >= f(b) - f(a)$$

(2) *Fundamental theorem for divergences, or Gauss or Green formula,*

$$\int_B < \nabla, \varphi >= \int_S < \varphi(x), n(x) > dx$$

(3) *Fundamental theorem for curls, or Stokes formula,*

$$\int_A < (\nabla \times \varphi)(x), n(x) > dx = \int_P < \varphi(x), dx >$$

where S is the boundary of the body B, and P is the boundary of the area A.

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

(1) This is something that we know well in 1D, namely the fundamental theorem of calculus, and the general, N-dimensional formula follows from that.

(2) This is something more subtle, and we had a taste of it when dealing with the Gauss law, and its various proofs. In general, the proof is similar, by using the various ideas from the proof of the Gauss law, and this can be found in any calculus book.

(3) This is again something subtle, and again with a flavor of things that we know, from the proof of the Gauss law, and which can be found in any calculus book. □

All the above was of course quite short, and at this point of reading this book, we can only recommend if needed a short break, for a brief calculus Navy Seals training camp. Such facilities are provided by basically any undergraduate electrodynamics book, in the opening chapter, and a particlarly enjoyable read here is Griffiths [44].

As for further details on all this, including mathematical proofs, generalizations, and more, you can go first to Lax' books for some good linear algebra, then to Rudin [72], [73] for advanced calculus, and then to do Carmo for differential geometry.

Getting back now to electrostatics, as a first application of the above, we have the following new point of view on the Gauss formula, which is more conceptual:

THEOREM 12.18 (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. We have several assertions here, the idea being as follows:

(1) The first formula, called Gauss law in differential form, follows from:

$$\begin{aligned}
\int_B < \nabla, E > &= \int_S < E(x), n(x) > dx \\
&= \Phi_E(S) \\
&= \frac{Q_{enc}}{\varepsilon_0} \\
&= \int_B \frac{\rho}{\varepsilon_0}
\end{aligned}$$

Now since this must hold for any B, this gives the formula in the statement.

(2) As a side remark, the Gauss law in differential form can be established as well directly, with the computation, involving a Dirac mass, being as follows:

$$\begin{aligned}
< \nabla, E > (x) &= \left\langle \nabla, K \int_{\mathbb{R}^3} \frac{\rho(z)(x-z)}{||x-z||^3}\, dz \right\rangle \\
&= K \int_{\mathbb{R}^3} \left\langle \nabla, \frac{x-z}{||x-z||^3} \right\rangle \rho(z)\, dz \\
&= K \int_{\mathbb{R}^3} 4\pi\delta_x \cdot \rho(z) dz \\
&= 4\pi K \int_{\mathbb{R}^3} \delta_x\, \rho(z) dz \\
&= \frac{\rho(x)}{\varepsilon_0}
\end{aligned}$$

And with this in hand, we have via (1) a new proof of the usual Gauss law.

(3) 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 Stokes' theorem, we conclude that the curl of E vanishes, as stated.

(4) Finally, as a side remark, both the formula of the divergence and the vanishing of the curl are somewhat clear by looking at the field lines of E. However, as all the above mathematics shows, there is certainly something to be understood, in all this. □

Moving ahead now, the question appears, what happens to the Gauss equations for the electric field E, as formulated above in Theorem 12.18, when written in terms of the associated potential V. And the answer here, which is remarkable, is as follows:

THEOREM 12.19 (Poisson). *In terms of the electric potential V, the Gauss formula becomes the Poisson equation, namely*

$$\Delta V = -\frac{\rho}{\varepsilon_0}$$

with $\Delta =< \nabla, \nabla >$ being the Laplace operator, given by the formula

$$\Delta f = \sum_i \frac{d^2 f}{dx_i^2}$$

and the curl equation dissapears, being automatic for gradients.

PROOF. Here both the assertions are elementary, as follows:

(1) With $E = -\nabla V$ the Gauss equation $< \nabla, E >= \rho/\varepsilon_0$ becomes:

$$< \nabla, \nabla V >= -\frac{\rho}{\varepsilon_0}$$

Thus we must have $\Delta V = -\rho/\varepsilon_0$, with the operator Δ being given by:

$$\begin{aligned}\Delta f &= <\nabla, \nabla f> \\ &= \left\langle \begin{pmatrix} \frac{d}{dx} \\ \frac{d}{dy} \\ \frac{d}{dz} \end{pmatrix}, \begin{pmatrix} \frac{df}{dx} \\ \frac{df}{dy} \\ \frac{df}{dz} \end{pmatrix} \right\rangle \\ &= \frac{d^2f}{dx^2} + \frac{d^2f}{dy^2} + \frac{d^2f}{dz^2}\end{aligned}$$

Thus, we are led to the Poisson equation in the statement.

(2) Regarding now the curl, our claim is that the equation $\nabla \times E = 0$ simply dissapears, this type of vanishing being automatic for gradients. Indeed, for any f we have:

$$\begin{aligned}\nabla \times \nabla f &= \begin{vmatrix} u_x & \frac{d}{dx} & \frac{df}{dx} \\ u_y & \frac{d}{dy} & \frac{df}{dy} \\ u_z & \frac{d}{dz} & \frac{df}{dz} \end{vmatrix} \\ &= \begin{pmatrix} \frac{d^2f}{dydz} - \frac{d^2f}{dzdy} \\ \frac{d^2f}{dzdx} - \frac{d^2f}{dxdz} \\ \frac{d^2f}{dxdy} - \frac{d^2f}{dydx} \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\end{aligned}$$

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

As an interesting feature of the potential approach, the Poisson equation makes sense, and is in fact very interesting, even when no charge is present, and we have here:

THEOREM 12.20 (Laplace). *In the case where no charges are present, the Poisson equation, and so the Gauss and even Coulomb laws too, in a certain sense, become*

$$\Delta V = 0$$

called Laplace equation, whose solutions are called harmonic functions. These functions have an interesting mathematics, reminding that of the holomorphic functions in 2D.

PROOF. There are many things that can be said here, First, the Laplace equation and its physical meaning come from the Poisson equation, and from the various potential considerations in the above. And mathematically, the idea is that various remarkable results about the holomorphic functions in 2D, such as the mean formula, extend to the harmonic functions. Thus, many things to be discussed, and we refer to Rudin [73] for the mathematics of harmonic functions, and to Griffiths [44] for their physics. □

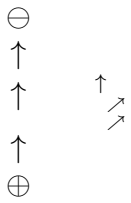
12c. Magnetic fields

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 12.21. *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 12.22 (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 12.21. 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 12.23. *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 12.24 (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 12.22, and in the definition of $\times$. In fact, the Lorentz force law is just a fancy reformulation of Fact 12.22, 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 12.22.

As an interesting application of the Lorentz force law, we have:

THEOREM 12.25. *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 12.26. *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 12.24 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 12.27. *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 12.28. *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 12.29 (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 12.30 (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, and very deep, that will have numerous consequences, in what follows, be that for investigating phenomena like radiation, or for making the link with Einstein's relativity theory, both crucially involving c.

But, first of all, this is certainly an invitation to rediscuss units and constants, as a continuation of our previous discussion on this topic. In what regards the units, we won't be impressed by the ampere, and keep using the coulomb, as a main unit:

CONVENTIONS 12.31. *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 12.30, 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 the speed of light c as constant too, and we are led in this way into the following conventions:

CONVENTIONS 12.32. *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 $K = 1/(4\pi\varepsilon_0)$.

Observe in passing that we are not messing up our figures, which can be quite often the case in this type of situation, because according to our data, and by truncating instead of rounding, as busy theoretical physicists usually do, we have:

$$\varepsilon_0\mu_0c^2 = 8.854 \times 1.256 \times 2.997^2 \times 10^{16-12-6} = 0.998$$

Getting back now to theory and math, the Biot-Savart law has as consequence:

THEOREM 12.33. *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 the magnetic field B of a steady line current I is given by the following formula:

$$B = \frac{\mu_0}{4\pi}\int \frac{I \times x}{||x||^3}$$

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 12.34 (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 the magnetic field B of a steady line current I is given by the following formula:

$$B = \frac{\mu_0}{4\pi}\int \frac{I \times x}{||x||^3}$$

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 12.35. *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 12.33 and Theorem 12.34, which both follow from the Biot-Savart law. □

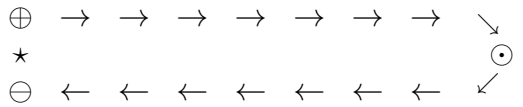
12d. Maxwell equations

Quite remarkably, and at the origin of all modern theory of electromagnetism, and of any type of modern electrical engineering too, we have:

FACT 12.36 (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 12.37. *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 12.36, these can be fine-tuned, and reformulated in terms of the emf, in the following way:

FACT 12.38 (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 12.39 (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 12.40 (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 12.41 (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.

12e. Exercises

Exercises:

EXERCISE 12.42.

EXERCISE 12.43.

EXERCISE 12.44.

EXERCISE 12.45.

EXERCISE 12.46.

EXERCISE 12.47.

EXERCISE 12.48.

EXERCISE 12.49.

Bonus exercise.

Part IV

Four dimensions

There was Slugger O'Toole who was drunk as a rule
And Fighting Bill Treacy from Dover
And your man Mick MacCann from the banks of the Bann
Was the skipper of the Irish Rover

CHAPTER 13

Linear algebra

13a. Diagonalization

Let us discuss now the diagonalization question for linear maps and matrices. The basic diagonalization theory, formulated in terms of matrices, is as follows:

THEOREM 13.1. *A vector $v \in \mathbb{C}^N$ is called eigenvector of $A \in M_N(\mathbb{C})$, with corresponding eigenvalue λ, when A multiplies by λ in the direction of v:*

$$Av = \lambda v$$

In the case where $\mathbb{C}^N$ has a basis $v_1, \ldots, v_N$ formed by eigenvectors of A, with corresponding eigenvalues $\lambda_1, \ldots, \lambda_N$, in this new basis A becomes diagonal, as follows:

$$A \sim \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix}$$

Equivalently, if we denote by $D = diag(\lambda_1, \ldots, \lambda_N)$ the above diagonal matrix, and by $P = [v_1 \ldots v_N]$ the square matrix formed by the eigenvectors of A, we have:

$$A = PDP^{-1}$$

In this case we say that the matrix A is diagonalizable.

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

(1) The first assertion is clear, because the matrix which multiplies each basis element v_i by a number λ_i is precisely the diagonal matrix $D = diag(\lambda_1, \ldots, \lambda_N)$.

(2) The second assertion follows from the first one, by changing the basis. We can prove this by a direct computation as well, because we have $Pe_i = v_i$, and so:

$$PDP^{-1}v_i = PDe_i = P\lambda_i e_i = \lambda_i P e_i = \lambda_i v_i$$

Thus, the matrices A and PDP^{-1} coincide, as stated. □

In order to study the diagonalization problem, the idea is that the eigenvectors can be grouped into linear spaces, called eigenspaces, as follows:

THEOREM 13.2. *Let $A \in M_N(\mathbb{C})$, and for any eigenvalue $\lambda \in \mathbb{C}$ define the corresponding eigenspace as being the vector space formed by the corresponding eigenvectors:*

$$E_\lambda = \left\{ v \in \mathbb{C}^N \middle| Av = \lambda v \right\}$$

These eigenspaces E_λ are then in a direct sum position, in the sense that given vectors $v_1 \in E_{\lambda_1}, \ldots, v_k \in E_{\lambda_k}$ corresponding to different eigenvalues $\lambda_1, \ldots, \lambda_k$, we have:

$$\sum_i c_i v_i = 0 \implies c_i = 0$$

In particular, we have $\sum_\lambda \dim(E_\lambda) \leq N$, with the sum being over all the eigenvalues, and our matrix is diagonalizable precisely when we have equality.

PROOF. We prove the first assertion by recurrence on $k \in \mathbb{N}$. Assume by contradiction that we have a formula as follows, with the scalars $c_1, \ldots, c_k$ being not all zero:

$$c_1 v_1 + \ldots + c_k v_k = 0$$

By dividing by one of these scalars, we can assume that our formula is:

$$v_k = c_1 v_1 + \ldots + c_{k-1} v_{k-1}$$

Now let us apply A to this vector. On the left we obtain:

$$Av_k = \lambda_k v_k = \lambda_k c_1 v_1 + \ldots + \lambda_k c_{k-1} v_{k-1}$$

On the right we obtain something different, as follows:

$$\begin{aligned} A(c_1 v_1 + \ldots + c_{k-1} v_{k-1}) &= c_1 A v_1 + \ldots + c_{k-1} A v_{k-1} \\ &= c_1 \lambda_1 v_1 + \ldots + c_{k-1} \lambda_{k-1} v_{k-1} \end{aligned}$$

We conclude from this that the following equality must hold:

$$\lambda_k c_1 v_1 + \ldots + \lambda_k c_{k-1} v_{k-1} = c_1 \lambda_1 v_1 + \ldots + c_{k-1} \lambda_{k-1} v_{k-1}$$

On the other hand, we know by recurrence that the vectors $v_1, \ldots, v_{k-1}$ must be linearly independent. Thus, the coefficients must be equal, at right and at left:

$$\lambda_k c_1 = c_1 \lambda_1$$

$$\vdots$$

$$\lambda_k c_{k-1} = c_{k-1} \lambda_{k-1}$$

Now since at least one of the numbers c_i must be nonzero, from $\lambda_k c_i = c_i \lambda_i$ we obtain $\lambda_k = \lambda_i$, which is a contradiction. Thus our proof by recurrence of the first assertion is complete. As for the second assertion, this follows from the first one. □

In order to reach now to more advanced results, we can use the following key fact:

THEOREM 13.3. *Given a matrix $A \in M_N(\mathbb{C})$, consider its characteristic polynomial:*

$$P(x) = \det(A - x1_N)$$

The eigenvalues of A are then the roots of P. Also, we have the inequality

$$\dim(E_\lambda) \leq m_\lambda$$

where m_λ is the multiplicity of λ, as root of P.

PROOF. The first assertion follows from the following computation, using the fact that a linear map is bijective when the determinant of the associated matrix is nonzero:

$$\begin{aligned} \exists v, Av = \lambda v &\iff \exists v, (A - \lambda 1_N)v = 0 \\ &\iff \det(A - \lambda 1_N) = 0 \end{aligned}$$

Regarding now the second assertion, given an eigenvalue λ of our matrix A, consider the dimension $d_\lambda = \dim(E_\lambda)$ of the corresponding eigenspace. By changing the basis of $\mathbb{C}^N$, as for the eigenspace E_λ to be spanned by the first d_λ basis elements, our matrix becomes as follows, with B being a certain smaller matrix:

$$A \sim \begin{pmatrix} \lambda 1_{d_\lambda} & 0 \\ 0 & B \end{pmatrix}$$

We conclude that the characteristic polynomial of A is of the following form:

$$P_A = P_{\lambda 1_{d_\lambda}} P_B = (\lambda - x)^{d_\lambda} P_B$$

Thus the multiplicity m_λ of our eigenvalue λ, viewed as a root of P, is subject to the estimate $m_\lambda \geq d_\lambda$, and this leads to the conclusion in the statement. □

Now recall that we are over $\mathbb{C}$, where any polynomial equation of degree $N \in \mathbb{N}$ has exactly N solutions, counted with multiplicities. By using this, we are led to:

THEOREM 13.4. *Given a matrix $A \in M_N(\mathbb{C})$, consider its characteristic polynomial*

$$P(X) = \det(A - X1_N)$$

then factorize this polynomial, by computing the complex roots, with multiplicities,

$$P(X) = (-1)^N (X - \lambda_1)^{n_1} \ldots (X - \lambda_k)^{n_k}$$

and finally compute the corresponding eigenspaces, for each eigenvalue found:

$$E_i = \left\{ v \in \mathbb{C}^N \middle| Av = \lambda_i v \right\}$$

The dimensions of these eigenspaces satisfy then the following inequalities,

$$\dim(E_i) \leq n_i$$

and A is diagonalizable precisely when we have equality for any i.

PROOF. This follows by combining the above results. By summing the inequalities $\dim(E_\lambda) \leq m_\lambda$ from Theorem 13.3, we obtain an inequality as follows:

$$\sum_\lambda \dim(E_\lambda) \leq \sum_\lambda m_\lambda \leq N$$

On the other hand, we know from Theorem 13.2 that our matrix is diagonalizable when we have global equality. Thus, we are led to the conclusion in the statement. □

This was for the main result of linear algebra. There are countless applications of this, and generally speaking, advanced linear algebra consists in building on Theorem 13.4. Let us record as well a useful algorithmic version of the above result:

THEOREM 13.5. *The square matrices $A \in M_N(\mathbb{C})$ can be diagonalized as follows:*

(1) *Compute the characteristic polynomial.*
(2) *Factorize the characteristic polynomial.*
(3) *Compute the eigenvectors, for each eigenvalue found.*
(4) *If there are no N eigenvectors, A is not diagonalizable.*
(5) *Otherwise, A is diagonalizable, $A = PDP^{-1}$.*

PROOF. This is an informal reformulation of Theorem 13.4, with (4) referring to the total number of linearly independent eigenvectors found in (3), and with $A = PDP^{-1}$ in (5) being the usual diagonalization formula, with P, D being as before. □

As a remark here, in step (3) it is always better to start with the eigenvalues having big multiplicity. Indeed, a multiplicity 1 eigenvalue, for instance, can never lead to the end of the computation, via (4), simply because the eigenvectors always exist.

13b. Spectral theorems

Let us go back to the diagonalization question, discussed in the previous section. We have in fact diagonalization results which are far more powerful. We first have:

THEOREM 13.6. *Any matrix $A \in M_N(\mathbb{C})$ which is self-adjoint, $A = A^*$, is diagonalizable, with the diagonalization being of the following type,*

$$A = UDU^*$$

with $U \in U_N$, and with $D \in M_N(\mathbb{R})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because if we take a matrix of the form $A = UDU^*$, with U unitary and D diagonal and real, then we have:

$$A^* = (UDU^*)^* = UD^*U^* = UDU^* = A$$

In the other sense now, assume that A is self-adjoint, $A = A^*$. Our first claim is that the eigenvalues are real. Indeed, assuming $Av = \lambda v$, we have:

$$\begin{aligned}\lambda < v, v > &= < \lambda v, v > \\ &= < Av, v > \\ &= < v, Av > \\ &= < v, \lambda v > \\ &= \bar{\lambda} < v, v >\end{aligned}$$

Thus we obtain $\lambda \in \mathbb{R}$, as claimed. Our next claim now is that the eigenspaces corresponding to different eigenvalues are pairwise orthogonal. Assume indeed that:

$$Av = \lambda v \quad , \quad Aw = \mu w$$

We have then the following computation, using $\lambda, \mu \in \mathbb{R}$:

$$\begin{aligned}\lambda < v, w > &= < \lambda v, w > \\ &= < Av, w > \\ &= < v, Aw > \\ &= < v, \mu w > \\ &= \mu < v, w >\end{aligned}$$

Thus $\lambda \neq \mu$ implies $v \perp w$, as claimed. In order now to finish the proof, it remains to prove that the eigenspaces of A span the whole space $\mathbb{C}^N$. For this purpose, we will use a recurrence method. Let us pick an eigenvector of our matrix:

$$Av = \lambda v$$

Assuming now that we have a vector w orthogonal to it, $v \perp w$, we have:

$$\begin{aligned}< Aw, v > &= < w, Av > \\ &= < w, \lambda v > \\ &= \lambda < w, v > \\ &= 0\end{aligned}$$

Thus, if v is an eigenvector, then the vector space $v^\perp$ is invariant under A. Moreover, since a matrix A is self-adjoint precisely when $< Av, v > \in \mathbb{R}$ for any vector $v \in \mathbb{C}^N$, as one can see by expanding the scalar product, the restriction of A to the subspace $v^\perp$ is self-adjoint. Thus, we can proceed by recurrence, and we obtain the result. □

Observe that, as a consequence of the above result, that you certainly might have heard of, any symmetric matrix $A \in M_N(\mathbb{R})$ is diagonalizable. In fact, we have:

THEOREM 13.7. *Any matrix $A \in M_N(\mathbb{R})$ which is symmetric, $A = A^t$, is diagonalizable, with the diagonalization being of the following type,*

$$A = UDU^t$$

with $U \in O_N$, and with $D \in M_N(\mathbb{R})$ diagonal. The converse holds too.

PROOF. As before, the converse trivially holds, because if we take a matrix of the form $A = UDU^t$, with U orthogonal and D diagonal and real, then we have $A^t = A$. In the other sense now, this follows from Theorem 13.6, and its proof. □

As basic examples of self-adjoint matrices, we have the orthogonal projections:

PROPOSITION 13.8. *The matrices $P \in M_N(\mathbb{C})$ which are projections, $P^2 = P^* = P$, are precisely those which diagonalize as follows,*

$$P = UDU^*$$

with $U \in U_N$, and with $D \in M_N(0,1)$ being diagonal.

PROOF. Since we have $P^* = P$, by using Theorem 13.6, the eigenvalues must be real. Then, by using $P^2 = P$, assuming that we have $Pv = \lambda v$, we obtain:

$$\begin{aligned}
\lambda < v, v > &= < \lambda v, v > \\
&= < Pv, v > \\
&= < P^2 v, v > \\
&= < Pv, Pv > \\
&= < \lambda v, \lambda v > \\
&= \lambda^2 < v, v >
\end{aligned}$$

We therefore have $\lambda \in \{0, 1\}$, as claimed, and as a final conclusion here, the diagonalization of the self-adjoint matrices is as follows, with $e_i \in \{0, 1\}$:

$$P \sim \begin{pmatrix} e_1 & & \\ & \ddots & \\ & & e_N \end{pmatrix}$$

To be more precise, the number of 1 values is the dimension of the image of P. □

In the real case, the result regarding the projections is as follows:

PROPOSITION 13.9. *The matrices $P \in M_N(\mathbb{R})$ which are projections, $P^2 = P^t = P$, are precisely those which diagonalize as follows,*

$$P = UDU^t$$

with $U \in O_N$, and with $D \in M_N(0,1)$ being diagonal.

PROOF. This follows indeed from Proposition 13.8, and its proof. □

An important class of self-adjoint matrices, which includes for instance all the projections, are the positive matrices. The theory here is as follows:

THEOREM 13.10. *For a matrix $A \in M_N(\mathbb{C})$ the following conditions are equivalent, and if they are satisfied, we say that A is positive:*

(1) $A = B^2$, *with* $B = B^*$.
(2) $A = CC^*$, *for some* $C \in M_N(\mathbb{C})$.
(3) $< Ax, x > \geq 0$, *for any vector* $x \in \mathbb{C}^N$.
(4) $A = A^*$, *and the eigenvalues are positive,* $\lambda_i \geq 0$.
(5) $A = UDU^*$, *with* $U \in U_N$ *and with* $D \in M_N(\mathbb{R}_+)$ *diagonal.*

PROOF. The idea is that the equivalences in the statement basically follow from some elementary computations, with only Theorem 13.6 needed, at some point:

(1) $\implies$ (2) This is clear, because we can take $C = B$.

(2) $\implies$ (3) This follows from the following computation:

$$\begin{aligned} < Ax, x > &= < CC^*x, x > \\ &= < C^*x, C^*x > \\ &\geq 0 \end{aligned}$$

(3) $\implies$ (4) By using the fact that $< Ax, x >$ is real, we have:

$$\begin{aligned} < Ax, x > &= < x, A^*x > \\ &= < A^*x, x > \end{aligned}$$

Thus we have $A = A^*$, and the remaining assertion, regarding the eigenvalues, follows from the following computation, assuming $Ax = \lambda x$:

$$\begin{aligned} < Ax, x > &= < \lambda x, x > \\ &= \lambda < x, x > \\ &\geq 0 \end{aligned}$$

(4) $\implies$ (5) This follows indeed by using Theorem 13.6.

(5) $\implies$ (1) Assuming $A = UDU^*$ with $U \in U_N$, and with $D \in M_N(\mathbb{R}_+)$ diagonal, we can set $B = U\sqrt{D}U^*$. Then B is self-adjoint, and its square is given by:

$$\begin{aligned} B^2 &= U\sqrt{D}U^* \cdot U\sqrt{D}U^* \\ &= UDU^* \\ &= A \end{aligned}$$

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

Let us record as well the following technical version of the above result:

THEOREM 13.11. *For a matrix $A \in M_N(\mathbb{C})$ the following conditions are equivalent, and if they are satisfied, we say that A is strictly positive:*

(1) $A = B^2$, *with* $B = B^*$, *invertible.*
(2) $A = CC^*$, *for some* $C \in M_N(\mathbb{C})$ *invertible.*
(3) $< Ax, x >> 0$, *for any nonzero vector* $x \in \mathbb{C}^N$.
(4) $A = A^*$, *and the eigenvalues are strictly positive,* $\lambda_i > 0$.
(5) $A = UDU^*$, *with* $U \in U_N$ *and with* $D \in M_N(\mathbb{R}_+^*)$ *diagonal.*

PROOF. This follows either from Theorem 13.10, by adding the above various extra assumptions, or from the proof of Theorem 13.10, by modifying where needed. □

Let us discuss now the case of the unitary matrices. We have here:

THEOREM 13.12. *Any matrix $U \in M_N(\mathbb{C})$ which is unitary, $U^* = U^{-1}$, is diagonalizable, with the eigenvalues on $\mathbb{T}$. More precisely we have*

$$U = VDV^*$$

with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because given a matrix of type $U = VDV^*$, with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ being diagonal, we have:

$$\begin{aligned} U^* &= (VDV^*)^* \\ &= VD^*V^* \\ &= VD^{-1}V^{-1} \\ &= (V^*)^{-1}D^{-1}V^{-1} \\ &= (VDV^*)^{-1} \\ &= U^{-1} \end{aligned}$$

Let us prove now the first assertion, stating that the eigenvalues of a unitary matrix $U \in U_N$ belong to $\mathbb{T}$. Indeed, assuming $Uv = \lambda v$, we have:

$$\begin{aligned} < v, v > &= < U^*Uv, v > \\ &= < Uv, Uv > \\ &= < \lambda v, \lambda v > \\ &= |\lambda|^2 < v, v > \end{aligned}$$

Thus we obtain $\lambda \in \mathbb{T}$, as claimed. Our next claim now is that the eigenspaces corresponding to different eigenvalues are pairwise orthogonal. Assume indeed that:

$$Uv = \lambda v \quad , \quad Uw = \mu w$$

We have then the following computation, using $U^* = U^{-1}$ and $\lambda, \mu \in \mathbb{T}$:

$$\begin{aligned}
\lambda < v, w > &= < \lambda v, w > \\
&= < Uv, w > \\
&= < v, U^* w > \\
&= < v, U^{-1} w > \\
&= < v, \mu^{-1} w > \\
&= \mu < v, w >
\end{aligned}$$

Thus $\lambda \neq \mu$ implies $v \perp w$, as claimed. In order now to finish the proof, it remains to prove that the eigenspaces of U span the whole space $\mathbb{C}^N$. For this purpose, we will use a recurrence method. Let us pick an eigenvector of our matrix:

$$Uv = \lambda v$$

Assuming that we have a vector w orthogonal to it, $v \perp w$, we have:

$$\begin{aligned}
< Uw, v > &= < w, U^* v > \\
&= < w, U^{-1} v > \\
&= < w, \lambda^{-1} v > \\
&= \lambda < w, v > \\
&= 0
\end{aligned}$$

Thus, if v is an eigenvector, then the vector space $v^\perp$ is invariant under U. Now since U is an isometry, so is its restriction to this space $v^\perp$. Thus this restriction is a unitary, and so we can proceed by recurrence, and we obtain the result. □

Let us record as well the real version of the above result, in a weak form:

PROPOSITION 13.13. *Any matrix $U \in M_N(\mathbb{R})$ which is orthogonal, $U^t = U^{-1}$, is diagonalizable, with the eigenvalues on $\mathbb{T}$. More precisely we have*

$$U = VDV^*$$

with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ being diagonal.

PROOF. This follows indeed from Theorem 13.12. □

Observe that the above result does not provide us with a complete characterization of the matrices $U \in M_N(\mathbb{R})$ which are orthogonal. To be more precise, the question left is that of understanding when the matrices of type $U = VDV^*$, with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ being diagonal, are real, and this is something non-trivial.

As an illustration for the above, for the simplest unitaries that we know, namely the rotations in the real plane, we have the following result:

THEOREM 13.14. *The rotation of angle $t \in \mathbb{R}$ in the real plane, namely*

$$R_t = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}$$

can be diagonalized over the complex numbers, as follows:

$$R_t = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} \begin{pmatrix} e^{-it} & 0 \\ 0 & e^{it} \end{pmatrix} \begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix}$$

Over the real numbers this is impossible, unless $t = 0, \pi$.

PROOF. The last assertion is something clear, that we already know, coming from the fact that at $t \neq 0, \pi$ our rotation is a "true" rotation, having no eigenvectors in the plane. Regarding the first assertion, the point is that we have the following computation:

$$R_t \begin{pmatrix} 1 \\ i \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} 1 \\ i \end{pmatrix} = e^{-it} \begin{pmatrix} 1 \\ i \end{pmatrix}$$

We have as well a second eigenvector, as follows:

$$R_t \begin{pmatrix} 1 \\ -i \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} 1 \\ -i \end{pmatrix} = e^{it} \begin{pmatrix} 1 \\ -i \end{pmatrix}$$

Thus our matrix R_t is diagonalizable over $\mathbb{C}$, with the diagonal form being:

$$R_t \sim \begin{pmatrix} e^{-it} & 0 \\ 0 & e^{it} \end{pmatrix}$$

As for the passage matrix, obtained by putting together the eigenvectors, this is:

$$P = \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}$$

In order to invert now P, we can use the standard inversion formula for the 2×2 matrices, which is the same as the one in the real case, and which gives:

$$P^{-1} = \frac{1}{-2i} \begin{pmatrix} -i & -1 \\ -i & 1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix}$$

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

13c. Normal matrices

Back to generalities, the self-adjoint matrices and the unitary matrices are particular cases of the general notion of a "normal matrix", and we have here:

THEOREM 13.15. *Any matrix $A \in M_N(\mathbb{C})$ which is normal, $AA^* = A^*A$, is diagonalizable, with the diagonalization being of the following type,*

$$A = UDU^*$$

with $U \in U_N$, and with $D \in M_N(\mathbb{C})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because if we take a matrix of the form $A = UDU^*$, with U unitary and D diagonal, then we have:

$$\begin{aligned} AA^* &= UDU^* \cdot UD^*U^* \\ &= UDD^*U^* \\ &= UD^*DU^* \\ &= UD^*U^* \cdot UDU^* \\ &= A^*A \end{aligned}$$

In the other sense now, this is something more technical. Our first claim is that a matrix A is normal precisely when the following happens, for any vector v:

$$||Av|| = ||A^*v||$$

Indeed, the above equality can be written as follows:

$$< AA^*v, v > = < A^*Av, v >$$

But this is equivalent to $AA^* = A^*A$, by expanding the scalar products. Our claim now is that A, A^* have the same eigenvectors, with conjugate eigenvalues:

$$Av = \lambda v \implies A^*v = \bar{\lambda}v$$

Indeed, this follows from the following computation, and from the trivial fact that if A is normal, then so is any matrix of type $A - \lambda 1_N$:

$$\begin{aligned} ||(A^* - \bar{\lambda}1_N)v|| &= ||(A - \lambda 1_N)^*v|| \\ &= ||(A - \lambda 1_N)v|| \\ &= 0 \end{aligned}$$

Let us prove now, by using this, that the eigenspaces of A are pairwise orthogonal. Assume that we have two eigenvectors, corresponding to different eigenvalues, $\lambda \neq \mu$:

$$Av = \lambda v \quad , \quad Aw = \mu w$$

We have the following computation, which shows that $\lambda \neq \mu$ implies $v \perp w$:

$$\begin{aligned} \lambda < v, w > &= < \lambda v, w > \\ &= < Av, w > \\ &= < v, A^*w > \\ &= < v, \bar{\mu}w > \\ &= \mu < v, w > \end{aligned}$$

In order to finish, it remains to prove that the eigenspaces of A span the whole $\mathbb{C}^N$. This is something that we have already seen for the self-adjoint matrices, and for unitaries,

and we will use here these results, in order to deal with the general normal case. As a first observation, given an arbitrary matrix A, the matrix AA^* is self-adjoint:

$$(AA^*)^* = AA^*$$

Thus, we can diagonalize this matrix AA^*, as follows, with the passage matrix being a unitary, $V \in U_N$, and with the diagonal form being real, $E \in M_N(\mathbb{R})$:

$$AA^* = VEV^*$$

Now observe that, for matrices of type $A = UDU^*$, which are those that we supposed to deal with, we have the following formulae:

$$V = U \quad , \quad E = D\bar{D}$$

In particular, the matrices A and AA^* have the same eigenspaces. So, this will be our idea, proving that the eigenspaces of AA^* are eigenspaces of A. In order to do so, let us pick two eigenvectors v, w of the matrix AA^*, corresponding to different eigenvalues, $\lambda \neq \mu$. The eigenvalue equations are then as follows:

$$AA^*v = \lambda v \quad , \quad AA^*w = \mu w$$

We have the following computation, using the normality condition $AA^* = A^*A$, and the fact that the eigenvalues of AA^*, and in particular μ, are real:

$$\begin{aligned}
\lambda < Av, w > &= < \lambda Av, w > \\
&= < A\lambda v, w > \\
&= < AAA^*v, w > \\
&= < AA^*Av, w > \\
&= < Av, AA^*w > \\
&= < Av, \mu w > \\
&= \mu < Av, w >
\end{aligned}$$

We conclude that we have $< Av, w >= 0$. But this reformulates as follows:

$$\lambda \neq \mu \implies A(E_\lambda) \perp E_\mu$$

Now since the eigenspaces of AA^* are pairwise orthogonal, and span the whole $\mathbb{C}^N$, we deduce from this that these eigenspaces are invariant under A:

$$A(E_\lambda) \subset E_\lambda$$

But with this result in hand, we can finish the proof of the theorem. Indeed, we can decompose the problem, and the matrix A itself, following these eigenspaces of AA^*, which in practice amounts in saying that we can assume that we only have 1 eigenspace. By rescaling, this is the same as assuming that we have $AA^* = 1$, and so we are now into the unitary case, that we know how to solve, as explained in Theorem 13.12. □

As a first application of our latest spectral theorem, we have the following result:

THEOREM 13.16. *Given a matrix $A \in M_N(\mathbb{C})$, we can construct a matrix $|A|$ as follows, by using the fact that A^*A is diagonalizable, with positive eigenvalues:*

$$|A| = \sqrt{A^*A}$$

This matrix $|A|$ is then positive, and its square is $|A|^2 = A$. In the case $N = 1$, we obtain in this way the usual absolute value of the complex numbers.

PROOF. Consider indeed the matrix A^*A, which is normal. According to Theorem 13.15, we can diagonalize this matrix as follows, with $U \in U_N$, and with D diagonal:

$$A = UDU^*$$

From $A^*A \geq 0$ we obtain $D \geq 0$. But this means that the entries of D are real, and positive. Thus we can extract the square root $\sqrt{D}$, and then set:

$$\sqrt{A^*A} = U\sqrt{D}U^*$$

Thus, we are basically done. Indeed, if we call this latter matrix $|A|$, then we are led to the conclusions in the statement. Finally, the last assertion is clear from definitions. □

We can now formulate a first polar decomposition result, as follows:

THEOREM 13.17. *Any invertible matrix $A \in M_N(\mathbb{C})$ decomposes as*

$$A = U|A|$$

with $U \in U_N$, and with $|A| = \sqrt{A^*A}$ as above.

PROOF. This is routine, and follows by comparing the actions of $A, |A|$ on the vectors $v \in \mathbb{C}^N$, and deducing from this the existence of a unitary $U \in U_N$ as above. □

Observe that at $N = 1$ we obtain in this way the usual polar decomposition of the nonzero complex numbers. More generally now, we have the following result:

THEOREM 13.18. *Any square matrix $A \in M_N(\mathbb{C})$ decomposes as*

$$A = U|A|$$

with U being a partial isometry, and with $|A| = \sqrt{A^*A}$ as above.

PROOF. Again, this follows by comparing the actions of $A, |A|$ on the vectors $v \in \mathbb{C}^N$, and deducing from this the existence of a partial isometry U as above. Alternatively, we can get this from Theorem 13.17, applied on the complement of the 0-eigenvectors. □

13d. Spectral measures

We would like to discuss now some interesting applications of our various spectral theorems to probability theory. Let us start with something basic, as follows:

DEFINITION 13.19. *Let X be a probability space, that is, a space with a probability measure, and with the corresponding integration denoted E, and called expectation.*

(1) *The random variables are the real functions $f \in L^\infty(X)$.*
(2) *The moments of such a variable are the numbers $M_k(f) = E(f^k)$.*
(3) *The law of such a variable is the measure given by $M_k(f) = \int_{\mathbb{R}} x^k d\mu_f(x)$.*

Here, and in what follows, we use the term "law" for "probability distribution", which means exactly the same thing, and is more convenient. Regarding now the fact that the law μ_f exists indeed, this is true, but not exactly trivial. By linearity, we would like to have a probability measure making hold the following formula, for any $P \in \mathbb{C}[X]$:

$$E(P(f)) = \int_{\mathbb{R}} P(x) d\mu_f(x)$$

By using a standard continuity argument, it is enough to have this formula for the characteristic functions χ_I of the arbitrary measurable sets of real numbers $I \subset \mathbb{R}$:

$$E(\chi_I(f)) = \int_{\mathbb{R}} \chi_I(x) d\mu_f(x)$$

But this latter formula, which reads $P(f \in I) = \mu_f(I)$, can serve as a definition for μ_f, and we are done. Alternatively, assuming some familiarity with measure theory, μ_f is the push-forward of the probability measure on X, via the function $f : X \to \mathbb{R}$.

Let us summarize this discussion in the form of a theorem, as follows:

THEOREM 13.20. *The law μ_f of a random variable f exists indeed, and we have*

$$E(\varphi(f)) = \int_{\mathbb{R}} \varphi(x) d\mu_f(x)$$

for any integrable function $\varphi : \mathbb{R} \to \mathbb{C}$.

PROOF. This follows from the above discussion, and with the precise assumption on $\varphi : \mathbb{R} \to \mathbb{C}$, which is its integrability, in the abstract mathematical sense, being in fact something that we will not really need, in what follows. In fact, for most purposes we will get away with polynomials $\varphi \in \mathbb{C}[X]$, and by linearity this means that we can get away with monomials $\varphi(x) = x^k$, which brings us back to Definition 13.19 (3), as stated. □

Getting now to the case of the matrices $A \in M_N(\mathbb{C})$, here it is quite tricky to figure out what the law of A should mean, based on intuition only. So, in the lack of a bright idea, let us just reproduce Definition 13.19, with a few modifications, as follows:

DEFINITION 13.21. *Let $N \in \mathbb{N}$, and consider the algebra $M_N(\mathbb{C})$ of complex $N \times N$ matrices, with its normalized trace $tr : M_N(\mathbb{C}) \to \mathbb{C}$, given by $tr(A) = Tr(A)/N$.*

(1) *We call random variables the self-adjoint matrices $A \in M_N(\mathbb{C})$.*
(2) *The moments of such a variable are the numbers $M_k(A) = tr(A^k)$.*
(3) *The law of such a variable is the measure given by $M_k(A) = \int_{\mathbb{R}} x^k d\mu_A(x)$.*

Here we have normalized the trace, as to have $tr(1) = 1$, in analogy with the formula $E(1) = 1$ from usual probability. By the way, as a piece of advice here, many confusions appear from messing up tr and Tr, and it is better ot forget about Tr, and always use tr. With the drawback that if you're a physicist, tr might get messed up in quick handwriting with the reduced Planck constant $\hbar = h/2\pi$. However, shall you ever face this problem, I have an advice here too, namely forgetting about h, and using h instead of $\hbar$.

Another comment is that we assumed in (1) that our matrix is self-adjoint, $A = A^*$, with the adjoint matrix being given, as usual, by the formula $(A^*)_{ij} = \bar{A}_{ji}$. Why this, because for instance at $N = 1$ we would like our matrix, which in the case $N = 1$ is a number, to be real, and so we must assume $A = A^*$. Of course there is still some discussion here, for instance because you might argue that why not assuming instead that the entries of A are real. But let us leave this for later, and in the meantime, just trust me. Or perhaps, let us both trust Heisenberg, who was the first intensive user of complex matrices, and who declared that such matrices must be self-adjoint. More later.

Back to work now, what we have in Definition 13.21 looks quite reasonable, but as before with the usual random variables $f \in L^\infty(X)$, some discussion is needed, in order to understand if the law μ_A exists indeed, and by which mechanism. And, good news here, in the case of the simplest matrices, the real diagonal ones, we have:

THEOREM 13.22. *For any diagonal matrix $A \in M_N(\mathbb{R})$ we have the formula*

$$tr(P(A)) = \frac{1}{N}(P(\lambda_1) + \ldots + P(\lambda_N))$$

where $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ are the diagonal entries of A. Thus the measure

$$\mu_A = \frac{1}{N}(\delta_{\lambda_1} + \ldots + \delta_{\lambda_N})$$

can be regarded as being the law of A, in the sense of Definition 13.21.

PROOF. Assume indeed that we have a real diagonal matrix, as follows, with the convention that the matrix entries which are missing are by definition 0 entries:

$$A = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix}$$

The powers of A are then diagonal too, given by the following formula:

$$A^k = \begin{pmatrix} \lambda_1^k & & \\ & \ddots & \\ & & \lambda_N^k \end{pmatrix}$$

In fact, given any polynomial $P \in \mathbb{C}[X]$, we have the following formula:

$$P(A) = \begin{pmatrix} P(\lambda_1) & & \\ & \ddots & \\ & & P(\lambda_N) \end{pmatrix}$$

Thus, the first formula in the statement holds indeed. In particular, we conclude that the moments of A are given by the following formula:

$$M_k(A) = tr(A^k) = \frac{1}{N}\sum_i \lambda_i^k$$

On the other hand, with $\mu_A = \frac{1}{N}(\delta_{\lambda_1} + \ldots + \delta_{\lambda_N})$ as in the statement, we have:

$$\begin{aligned} \int_{\mathbb{R}} x^k d\mu_A(x) &= \frac{1}{N}\sum_i \int_{\mathbb{R}} x^k d\delta_{\lambda_i}(x) \\ &= \frac{1}{N}\sum_i \lambda_i^k \end{aligned}$$

Thus that the law of A exists indeed, and is the measure μ_A, as claimed. □

The point now is that, by using the spectral theorem for self-adjoint matrices, we have the following generalization of Theorem 13.22, dealing with the general case:

THEOREM 13.23. *For a self-adjoint matrix $A \in M_N(\mathbb{C})$ we have the formula*

$$tr(P(A)) = \frac{1}{N}(P(\lambda_1) + \ldots + P(\lambda_N))$$

where $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ are the eigenvalues of A. Thus the measure

$$\mu_A = \frac{1}{N}(\delta_{\lambda_1} + \ldots + \delta_{\lambda_N})$$

can be regarded as being the law of A, in the sense of Definition 13.21.

PROOF. We already know, from Theorem 13.22, that the result holds indeed for the diagonal matrices. In the general case now, that of an arbitrary self-adjoint matrix, we know from before that our matrix is diagonalizable, as follows:

$$A = UDU^*$$

Now observe that the moments of A are given by the following formula:

$$\begin{aligned} tr(A^k) &= tr(UDU^* \cdot UDU^* \ldots UDU^*) \\ &= tr(UD^kU^*) \\ &= tr(D^k) \end{aligned}$$

We conclude from this, by reasoning by linearity, that the matrices A, D have the same law, $\mu_A = \mu_D$, and this gives all the assertions in the statement. □

The above theory is not the end of the story, because we can talk about complex random variables, $f : X \to \mathbb{C}$, and about non-self-adjoint matrices too, $A \neq A^*$. We will see that, with a bit of know-how, we can have some law technology going on, for both.

Let us start with the complex variables $f \in L^\infty(X)$. The main difference with respect to the real case comes from the fact that we have now a pair of variables instead of one, namely $f : X \to \mathbb{C}$ itself, and its conjugate $\bar{f} : X \to \mathbb{C}$. Thus, we are led to:

Definition 13.24. *The moments a complex variable $f \in L^\infty(X)$ are the numbers*

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

depending on colored integers $k = \circ \bullet \bullet \circ \ldots$, with the conventions

$$f^\emptyset = 1 \quad , \quad f^\circ = f \quad , \quad f^\bullet = \bar{f}$$

and multiplicativity, in order to define the colored powers f^k.

Observe that, since $f, \bar{f}$ commute, we can permute terms, and restrict the attention to exponents of type $k = \ldots \circ \circ \circ \bullet \bullet \bullet \bullet \ldots$, if we want to. However, our various results below will look better without doing this, so we will use Definition 13.24 as stated.

Regarding now the notion of law, this extends too, the result being as follows:

Theorem 13.25. *Each complex variable $f \in L^\infty(X)$ has a law, which is by definition a complex probability measure μ_f making the following formula hold,*

$$M_k(f) = \int_{\mathbb{C}} z^k d\mu_f(z)$$

for any colored integer k. Moreover, we have in fact the formula

$$E(\varphi(f)) = \int_{\mathbb{C}} \varphi(x) d\mu_f(x)$$

valid for any integrable function $\varphi : \mathbb{C} \to \mathbb{C}$.

PROOF. The first assertion follows exactly as in the real case, and with z^k being defined exactly as f^k, namely by the following formulae, and multiplicativity:

$$z^\emptyset = 1 \quad , \quad z^\circ = z \quad , \quad z^\bullet = \bar{z}$$

As for the second assertion, this basically follows from this by linearity and continuity, by using standard measure theory, again as in the real case. □

Moving ahead towards matrices, all this leads to a mixture of easy and complicated problems. First, Definition 13.24 has the following straightforward analogue:

DEFINITION 13.26. *The moments a matrix $A \in M_N(\mathbb{C})$ are the numbers*

$$M_k(A) = tr(A^k)$$

depending on colored integers $k = \circ\bullet\bullet\circ\ldots$, with the usual conventions

$$A^\emptyset = 1 \quad , \quad A^\circ = A \quad , \quad A^\bullet = A^*$$

and multiplicativity, in order to define the colored powers A^k.

As a first observation about this, unless the matrix is normal, $AA^* = A^*A$, we cannot switch to exponents of type $k = \ldots\circ\circ\circ\bullet\bullet\bullet\bullet\ldots$, as it was theoretically possible for the complex variables $f \in L^\infty(X)$. Here is an explicit counterexample for this:

PROPOSITION 13.27. *The following matrix, which is not normal,*

$$J = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

has the property $tr(JJ^*JJ^*) \neq tr(JJJ^*J^*)$.

PROOF. We have the following formulae, which show that J is not normal:

$$JJ^* = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

$$J^*J = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Let us compute now the quantities in the statement. We first have:

$$tr(JJ^*JJ^*) = tr((JJ^*)^2) = tr\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \frac{1}{2}$$

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

$$tr(JJJ^*J^*) = tr\left(\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\right) = tr\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = 0$$

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

The above counterexample makes it quite clear that things will be complicated, when attempting to talk about the law of an arbitrary matrix $A \in M_N(\mathbb{C})$. But, there is solution to everything. By being a bit smart, we can formulate things as follows:

DEFINITION 13.28. *The law of a complex matrix $A \in M_N(\mathbb{C})$ is the following functional, on the algebra of polynomials in two noncommuting variables X, X^*:*

$$\mu_A : \mathbb{C} < X, X^* > \to \mathbb{C} \quad , \quad P \to tr(P(A))$$

In the case where we have a complex probability measure $\mu_A \in \mathcal{P}(\mathbb{C})$ such that

$$tr(P(A)) = \int_{\mathbb{C}} P(x)\, d\mu_A(x)$$

we identify this complex measure with the law of A.

As mentioned above, this is something smart, that will take us some time to understand. As a first observation, knowing the law is the same as knowing the moments, because if we write our polynomial as $P = \sum_k c_k X^k$, then we have:

$$tr(P(A)) = tr\left(\sum_k c_k A^k\right) = \sum_k c_k M_k(A)$$

Let us try now to compute some matrix laws, and see what we get. We already did some computations in the real case, and then for the basic 2×2 Jordan block J too, and based on all this, we can formulate the following result, with mixed conclusions:

THEOREM 13.29. *The following happen:*

(1) *If $A = A^*$ then $\mu_A = \frac{1}{N}(\lambda_1 + \ldots + \lambda_N)$, with $\lambda_i \in \mathbb{R}$ being the eigenvalues.*
(2) *If A is diagonal, $\mu_A = \frac{1}{N}(\lambda_1 + \ldots + \lambda_N)$, with $\lambda_i \in \mathbb{C}$ being the eigenvalues.*
(3) *For the basic Jordan block J, the law μ_J is not a complex measure.*
(4) *In fact, assuming $AA^* \neq A^*A$, the law μ_A is not a complex measure.*

PROOF. This follows from the above, with only (4) being new. Assuming $AA^* \neq A^*A$, in order to show that μ_A is not a measure, we can use a positivity trick, as follows:

$$\begin{aligned}
AA^* - A^*A \neq 0 &\implies (AA^* - A^*A)^2 > 0\\
&\implies AA^*AA^* - AA^*A^*A - A^*AAA^* + A^*AA^*A > 0\\
&\implies tr(AA^*AA^* - AA^*A^*A - A^*AAA^* + A^*AA^*A) > 0\\
&\implies tr(AA^*AA^* + A^*AA^*A) > tr(AA^*A^*A + A^*AAA^*)\\
&\implies tr(AA^*AA^*) > tr(AAA^*A^*)
\end{aligned}$$

Thus, we can conclude as in the proof for J, the point being that we cannot obtain both the above numbers by integrating $|z|^2$ with respect to a measure $\mu_A \in \mathcal{P}(\mathbb{C})$. □

Fortunately, by using the spectral theorem for normal matrices, we have:

THEOREM 13.30. *Given a matrix $A \in M_N(\mathbb{C})$ which is normal, $AA^* = A^*A$, we have the following formula, valid for any polynomial $P \in \mathbb{C} < X, X^* >$,*

$$tr(P(A)) = \frac{1}{N}(P(\lambda_1) + \ldots + P(\lambda_N))$$

where $\lambda_1, \ldots, \lambda_N \in \mathbb{C}$ are the eigenvalues of A. Thus the complex measure

$$\mu_A = \frac{1}{N}(\delta_{\lambda_1} + \ldots + \delta_{\lambda_N})$$

is the law of A. In the non-normal case, the law μ_A is not a measure.

PROOF. As before in the diagonal case, since our matrix is normal, $AA^* = A^*A$, knowing its law in the abstract sense of generalized probability is the same as knowing the restriction of this abstract distribution to the usual polynomials in two variables:

$$\mu_A : \mathbb{C}[X, X^*] \to \mathbb{C} \quad , \quad P \to tr(P(A))$$

In order now to compute this functional, we can write $A = UDU^*$, as in Theorem 13.15, and then change the basis via U, which in practice means that we can simply assume $U = 1$. Thus if we denote by $\lambda_1, \ldots, \lambda_N$ the diagonal entries of D, which are the eigenvalues of A, the law that we are looking for is the following functional:

$$\mu_A : \mathbb{C}[X, X^*] \to \mathbb{C} \quad , \quad P \to \frac{1}{N}(P(\lambda_1) + \ldots + P(\lambda_N))$$

But this functional corresponds to integrating P with respect to the following complex measure, that we agree to still denote by μ_A, and call distribution of A:

$$\mu_A = \frac{1}{N}(\delta_{\lambda_1} + \ldots + \delta_{\lambda_N})$$

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

13e. Exercises

Exercises:

EXERCISE 13.31.

EXERCISE 13.32.

EXERCISE 13.33.

EXERCISE 13.34.

EXERCISE 13.35.

EXERCISE 13.36.

EXERCISE 13.37.

EXERCISE 13.38.

Bonus exercise.

CHAPTER 14

Relativity theory

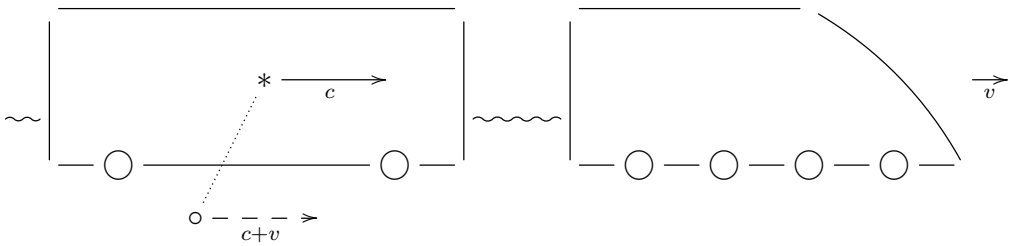
14a. Speed addition

Based on experiments by Fizeau, then Michelson-Morley and others, and some physics by Maxwell and Lorentz too, Einstein came upon the following principles:

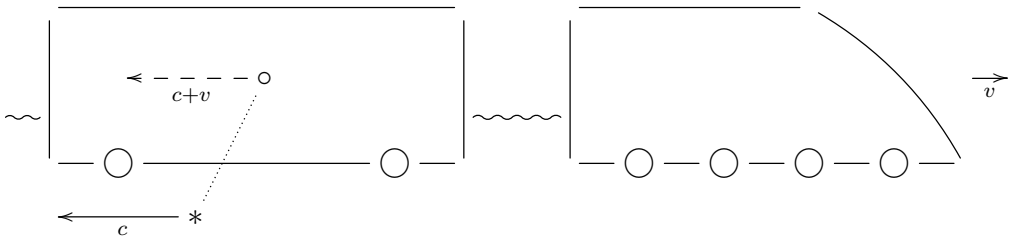
FACT 14.1 (Einstein principles). *The following happen:*

(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 14.1 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 14.2. *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 14.3. *The map wrapping* $[-1, 1]$ *into the unit circle, and then stereographically projecting on* $\mathbb{R}$ *is given by the formula*

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

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

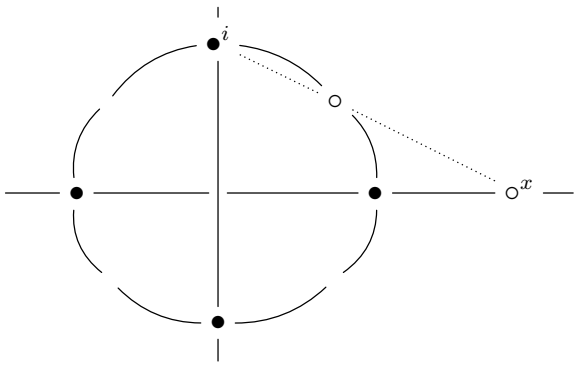
PROOF. Regarding the wrapping, as indicated, this is given by:

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

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

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

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



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

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

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

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

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

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 14.3, and start all over again. Getting back now to Puzzle 14.2, 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 14.4. *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 14.2, 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 14.4. So, we have our idea, and we are led to the following result:

THEOREM 14.5. *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:

$$\tan \to \tanh$$

Which sound quite good and conceptual, and we will stop here our 1D study.

14b. Three dimensions

Let us attempt now 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 14.6. *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 14.7. *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 $||v|| = 1$. Thus, certainly wrong choice, and we must invent something more complicated.

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 14.8. *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:

$$u \times (u \times v) = < u, v > u - < u, u > v$$

As for the formula of $v \times w$, that I forgot to record, we can recover it from the one above of $u \times w$, by using the basic properties of the vector products, as follows:

$$\begin{aligned} v \times (u \times v) &= -v \times (v \times u) \\ &= -(< v, u > v - < v, v > u) \\ &= < v, v > u - < u, v > v \end{aligned}$$

With these formulae in hand, we can now compute the correction term, with the result here, that we will need several times in what comes next, being as follows:

PROPOSITION 14.9. *The correction term* $\gamma_{uvw} = (\alpha u + \beta v) \times w$ *is given by*

$$\gamma_{uvw} = (\alpha < u, v > + \beta < v, v >)u - (\alpha < u, u > + \beta < u, v >)v$$

for any values of the scalars $\alpha, \beta \in \mathbb{R}$*.*

PROOF. According to our vector product formulae above, we have:

$$\begin{aligned}\gamma_{uvw} &= (\alpha u+\beta v)\times w\\ &= \alpha(<u,v>u-<u,u>v)+\beta(<v,v>u-<u,v>v)\\ &= (\alpha<u,v>+\beta<v,v>)u-(\alpha<u,u>+\beta<u,v>)v\end{aligned}$$

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

Time now to get into the real thing, see what happens when $||u||=1$ and $||v||=1$, if we can get indeed $u+_e v=u$ and $u+_e v=v$. It is convenient here to do some reverse engineering. Regarding the first desired formula, namely $u+_e v=u$, we have:

$$\begin{aligned}u+_e v=u &\iff u+v+\gamma_{uvw}=(1+<u,v>)u\\ &\iff \gamma_{uvw}=<u,v>u-v\\ &\iff \alpha=1,\ \beta=0,\ ||u||=1\end{aligned}$$

Thus, with the parameter choice $\alpha=1,\beta=0$, we will have, as desired:

$$||u||=1\implies u+_e v=u$$

In what regards now the second desired formula, namely $u+_e v=v$, here the computation is almost identical, save for a sign switch, which after some thinking comes from our choice $w=u\times v$ instead of $w=v\times u$, clearly favoring u, as follows:

$$\begin{aligned}u+_e v=v &\iff u+v+\gamma_{uvw}=(1+<u,v>)v\\ &\iff \gamma_{uvw}=-u+<u,v>v\\ &\iff \alpha=0,\ \beta=-1,\ ||v||=1\end{aligned}$$

Thus, with the parameter choice $\alpha=0,\beta=-1$, we will have, as desired:

$$||v||=1\implies u+_e v=v$$

All this is mixed news, because we managed to solve both our problems, at $||u||=1$ and at $||v||=1$, but our solutions are different. So, time to breathe, decide that we did enough interesting work for the day, and formulate our conclusion as follows:

PROPOSITION 14.10. *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 14.10 from an abstract, mathematical perspective, there are still many things missing from there, which can be summarized as follows:

QUESTION 14.11. *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 14.10, 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 14.12. *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 14.10.

(2) In order to simplify notation, let us set $\delta=\sqrt{1-||u||^2}$, which is the inverse of the quantity $\gamma=1/\sqrt{1-||u||^2}$. With this convention, we have:

$$\begin{aligned} u+_e v &= \frac{1}{1+<u,v>}\left(u+v+\frac{<u,v>u-||u||^2v}{1+\delta}\right)\\ &= \frac{(1+\delta+<u,v>)u+(1+\delta-||u||^2)v}{(1+<u,v>)(1+\delta)}\end{aligned}$$

Taking now the squared norm and computing gives the following formula:

$$||u+_e v||^2=\frac{(1+\delta)^2||u+v||^2+(||u||^2-2(1+\delta))(||u||^2||v||^2-<u,v>^2)}{(1+<u,v>)^2(1+\delta)^2}$$

But this formula can be further processed by using $\delta=\sqrt{1-||u||^2}$, and by navigating through the various quantities which appear, we obtain, as a final product:

$$||u+_e v||^2=\frac{||u+v||^2-||u||^2||v||^2+<u,v>^2}{(1+<u,v>)^2}$$

But this type of formula is exactly what we need, for what we want to do. Indeed, by assuming $||u||, ||v|| < 1$, we have the following estimate:

$$\begin{aligned} ||u +_e v||^2 < 1 &\iff ||u+v||^2 - ||u||^2||v||^2 + < u, v >^2 < (1 + < u, v >)^2 \\ &\iff ||u+v||^2 - ||u||^2||v||^2 < 1 + 2 < u, v > \\ &\iff ||u||^2 + ||v||^2 - ||u||^2||v||^2 < 1 \\ &\iff (1 - ||u||^2)(1 - ||v||^2) > 0 \end{aligned}$$

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

(3) This is something that we know from Proposition 14.10.

(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 14.12 is the good one, confirmed by experimental physics.

14c. Relativity theory

Time now to draw some concrete conclusions, from the above speed computations. Since speed $v = d/t$ is distance over time, we must fine-tune distance d, or time t, or both. Let us first discuss, following as usual Einstein, what happens to time t. Here the result, which might seem quite surprising, at a first glance, is as follows:

THEOREM 14.13. *Relativistic time is subject to Lorentz dilation*

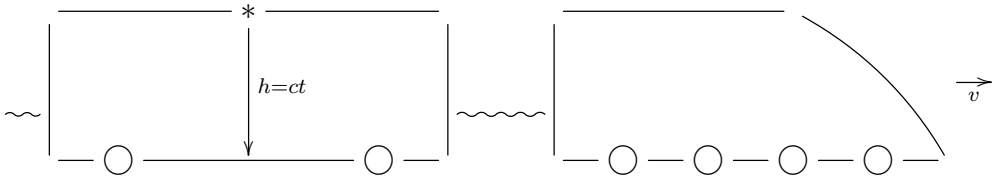
$$t \to \gamma t$$

where the number $\gamma \geq 1$, called Lorentz factor, is given by the formula

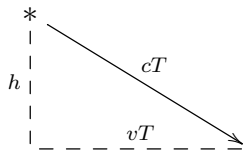
$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}$$

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 travelling 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 travelled, 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. Intuitively, since speed is distance/time, and since time gets dilated, we can somehow expect distance to get dilated too.

However, and a bit surprisingly, this is wrong, and after due thinking and computations, what we have is in fact the following result:

THEOREM 14.14. *Relativistic length is subject to Lorentz contraction*

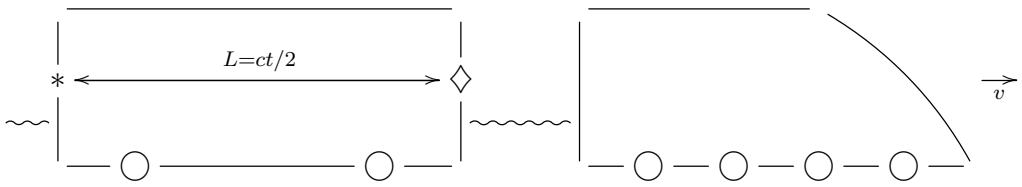
$$L \to L/\gamma$$

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 14.13, 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. □

With this discussed, time now to get to the real thing, see what happens to our usual $\mathbb{R}^4$. Let us start our discussion with a look at the non-relativistic case. Assuming that the object moves with speed v in the x direction, the frame change is given by:

$$x' = x - vt$$

$$y' = y$$

$$z' = z$$

$$t' = t$$

To be more precise, here the first 3 equations come from the law of motion, and $t' = t$ is the old $t' = t$. In the relativistic setting now, the result is more tricky, as follows:

THEOREM 14.15. *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 we obtain as desired:

$$x' = \gamma(x - vt)$$

Regarding y, z, these are obviously unchanged, so done with these too. Finally, regarding time t, a naive thought would suggest that this is subject to a Lorentz contraction by $1/\gamma$, but this is not true, and more thinking leads to the conclusion that we must use the reverse Lorentz transformation, given by the following formulae:

$$x = \gamma(x' + vt')$$

$$y = y'$$

$$z = z'$$

By using the formula of x' we can compute t', and we obtain the following formula:

$$\begin{aligned} t' &= \frac{x - \gamma x'}{\gamma v} \\ &= \frac{x - \gamma^2(x - vt)}{\gamma v} \\ &= \frac{\gamma^2 vt + (1 - \gamma^2)x}{\gamma v} \end{aligned}$$

On the other hand, we have the following computation:

$$\gamma^2 = \frac{c^2}{c^2 - v^2} \implies \gamma^2(c^2 - v^2) = c^2 \implies (\gamma^2 - 1)c^2 = \gamma^2 v^2$$

Thus we can finish the computation of t' as follows:

$$\begin{aligned} t' &= \frac{\gamma^2 vt + (1 - \gamma^2)x}{\gamma v} \\ &= \frac{\gamma^2 vt - \gamma^2 v^2 x/c^2}{\gamma v} \\ &= \gamma\left(t - \frac{vx}{c^2}\right) \end{aligned}$$

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

Now since y, z are irrelevant, we can put them at the end, and put the time t first, as to be close to x. By multiplying as well the time equation by c, our system becomes:

$$ct' = \gamma(ct - vx/c)$$

$$x' = \gamma(x - vt)$$

$$y' = y$$

$$z' = z$$

In linear algebra terms, the result is as follows:

THEOREM 14.16. *The Lorentz transformation is given by*

$$\begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix}$$

where $\gamma = 1/\sqrt{1 - v^2/c^2}$ *as usual, and where* $\beta = v/c$.

PROOF. In terms of $\beta = v/c$, replacing v, the system looks as follows:

$$ct' = \gamma(ct - \beta x)$$

$$x' = \gamma(x - \beta ct)$$

$$y' = y$$

$$z' = z$$

But this gives the formula in the statement. □

As an illustration, let us verify that the inverse Lorentz transformation is indeed given by reversing the speed, $v \to -v$. With notations as in Theorem 14.15, the result is:

THEOREM 14.17. *The inverse of the Lorentz transformation is given by $v \to -v$,*

$$x = \gamma(x' + vt')$$
$$y = y'$$
$$z = z'$$
$$t = \gamma(t' + vx'/c^2)$$

where $\gamma = 1/\sqrt{1 - v^2/c^2}$ is as usual the Lorentz factor, identical for v and $-v$.

PROOF. In terms of the formalism in Theorem 14.16, reversing the speed $v \to -v$ amounts in reversing the $\beta = v/c$ parameter there:

$$\beta \to -\beta$$

What we have to prove, in order to establish the result, is that by doing so, we obtain the inverse of the matrix appearing there, namely:

$$L = \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

That is, we want to prove that the inverse of this matrix is as follows:

$$L^{-1} = \begin{pmatrix} \gamma & \beta\gamma & 0 & 0 \\ \beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

But here, for the verification of the inversion formula $LL^{-1} = 1$, we can restrict the attention to the upper left corner, where the result is clear. □

Let us discuss now what happens to momentum, mass and energy. We would like to fix the momentum conservation equations for the plastic collisions, namely:

$$m = m_1 + m_2$$
$$mv = m_1v_1 + m_2v_2$$

However, this cannot really be done with bare hands, and by this meaning with mathematics only. But with some help from experiments, the conclusion is as follows:

FACT 14.18. *When defining the relativistic mass of an object of rest mass $m > 0$, moving at speed v, by the formula*

$$M = \gamma m \quad : \quad \gamma = \frac{1}{\sqrt{1 - v^2/c^2}}$$

this relativistic mass M, and the corresponding relativistic momentum $P = Mv$, are both conserved during collisions.

In other words, the situation here is a bit similar to that of the Galileo addition vs Einstein addition for speeds. The collision equations given above are in fact low-speed approximations of the correct, relativistic equations, which are as follows:

$$M = M_1 + M_2$$

$$Mv = M_1v_1 + M_2v_2$$

It remains now to discuss kinetic energy. You have certainly heard of the formula $E = mc^2$, which might actually well be on your T-shirt, now as you read this book, and in this case here is the explanation for it, in relation with the above:

THEOREM 14.19. *The relativistic energy of an object of rest mass $m > 0$,*

$$\mathcal{E} = Mc^2 \quad : \quad M = \gamma m$$

which is conserved, as being a multiple of M, can be written as $\mathcal{E} = E + T$, with

$$E = mc^2$$

being its $v = 0$ component, called rest energy of m, and with

$$T = (1 - \gamma)mc^2 \simeq \frac{mv^2}{2}$$

being called relativistic kinetic energy of m.

PROOF. All this is a bit abstract, coming from Fact 14.18, as follows:

(1) Given an object of rest mass $m > 0$, consider its relativistic mass $M = \gamma m$, as appearing in Fact 14.18, and then consider the following quantity:

$$\mathcal{E} = Mc^2$$

We know from Fact 14.18 that the relativistic mass M is conserved, so $\mathcal{E} = Mc^2$ is conserved too. In view of this, is makes somehow sense to call $\mathcal{E}$ energy. There is of course no clear reason for doing that, but let's just do it, and we'll understand later.

(2) Let us compute $\mathcal{E}$. This quantity is by definition given by:

$$\mathcal{E} = Mc^2 = \gamma mc^2 = \frac{mc^2}{\sqrt{1 - v^2/c^2}}$$

Since $1/\sqrt{1-x} \simeq 1 + x/2$ for x small, by calculus, we obtain, for v small:

$$\mathcal{E} \simeq mc^2 \left(1 + \frac{v^2}{2c^2}\right) = mc^2 + \frac{mv^2}{2}$$

And, good news here, we recognize at right the kinetic energy of m.

(3) But this leads to the conclusions in the statement. Indeed, we are certainly dealing with some sort of energies here, and so calling the above quantity $\mathcal{E}$ relativistic energy

is legitimate, and calling $E = mc^2$ rest energy is legitimate too. Finally, the difference between these two energies $T = \mathcal{E} - E$ follows to be given by:

$$T = (1 - \gamma)mc^2 \simeq \frac{mv^2}{2}$$

Thus, calling T relativistic kinetic energy is legitimate too, and we are done. □

14d. Curved spacetime

In the classical case, to begin with, the general frame change is as follows:

THEOREM 14.20. *In the classical case, the general frame change formula is:*

$$x' = x - v_x t$$
$$y' = y - v_y t$$
$$z' = z - v_z t$$
$$t' = t$$

Equivalently, in matrix form, this frame change formula is given by:

$$\begin{pmatrix} x' \\ y' \\ z' \\ t' \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & -v_x \\ 0 & 1 & 0 & -v_y \\ 0 & 0 & 1 & -v_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ t \end{pmatrix}$$

As for the reverse frame change, this is obtained via $v \to -v$.

PROOF. This is indeed clear from definitions. Observe also that the last assertion is verified by the following inversion formula, at the level of the associated matrices:

$$\begin{pmatrix} 1 & 0 & 0 & v_x \\ 0 & 1 & 0 & v_y \\ 0 & 0 & 1 & v_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & -v_x \\ 0 & 1 & 0 & -v_y \\ 0 & 0 & 1 & -v_z \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

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

In the relativistic case, the formulae and computations are more tricky, with some vector calculus involved, and the result is best stated in the following way:

THEOREM 14.21. *In the relativistic case, the general frame change formula is*

$$x' = x + (\gamma - 1)\frac{< v, x > v}{||v||^2} - \gamma t v$$
$$t' = \gamma\left(t - \frac{< v, x >}{c^2}\right)$$

where $\gamma = 1/\sqrt{1 - ||v||^2/c^2}$, and the reverse frame change is obtained via $v \to -v$.

PROOF. As already mentioned, this is something quite tricky, with some vector calculus involved, and we will take our time, and try to understand how this works:

(1) To start with, the formula in the statement looks N-dimensional, and our claim is that, indeed, this formula works in N-dimensional relativity, regardless of $N \in \mathbb{N}$.

(2) As a first illustration, let us first see what happens at $N = 1$. Here both the position variable x and the speed v are usual real numbers, and our first formula becomes:

$$\begin{aligned} x' &= x + (\gamma - 1)\frac{vx \cdot v}{v^2} - \gamma tv \\ &= x + (\gamma - 1)x - \gamma tv \\ &= \gamma x - \gamma tv \\ &= \gamma(x - tv) \end{aligned}$$

Thus, our first formula is correct. As for the second formula, this is correct too:

$$t = \gamma\left(t - \frac{vx}{c^2}\right)$$

(3) As a second illustration, let us move to arbitrary $N \in \mathbb{N}$ dimensions, including the case $N = 3$ that we are mostly interested in, and test our formula in the case where the configuration is standard, that is, where the speed vector is of the following form:

$$v = \begin{pmatrix} \nu \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

In this case, we obtain the correct formula for the position vector, as follows:

$$\begin{aligned} x' &= x + (\gamma - 1)\frac{\nu x_1 \cdot v}{\nu^2} - \gamma tv \\ &= x + (\gamma - 1)x_1 \cdot \frac{v}{\nu} - \gamma tv \\ &= \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} + (\gamma - 1)x_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \gamma t \begin{pmatrix} \nu \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \gamma x_1 - \gamma t\nu \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \end{aligned}$$

As for the resulting time, this is the correct one too, as follows:

$$t' = \gamma\left(t - \frac{\nu x_1}{c^2}\right)$$

(4) Summarizing, the formula in the statement generalizes well everything that we know. In order to prove now this formula, the general idea is that of decomposing the position vectors x, x' as follows, with respect to v and its complement:

$$x = \lambda v + y \quad , \quad y \perp v$$

$$x' = \lambda' v + y' \quad , \quad y' \perp v$$

Indeed, this can only give the result, by using the standard configuration formulae from before, and various abstract or concrete rotation arguments.

(5) In practice now, there are several ways of doing this. As a first observation, the above decomposition argument shows that our time formula is indeed the correct one:

$$t' = \gamma\left(t - \frac{< v, x >}{c^2}\right)$$

But with this in hand, it is possible to trick with an abstract argument, saying on one hand that x' must be linear in x, t, and on the other hand that we must have:

$$||x'||^2 - ct'^2 = ||x||^2 - ct^2$$

To be more precise, this latter formula must hold indeed, with this being something quite subtle, that we will explain later, and together with the above-mentioned linearity requirement, this leads to the formula in the statement for x'. But more on this later.

(6) Going instead on a more pedestrian way, we certainly know that the formula of t' is correct, and it remains to justify the formula of x'. But here, the best is to do first the computation in $N = 2$ dimensions, along the lines suggested in (4). This gives:

$$x' = x + (\gamma - 1)\frac{< v, x > v}{||v||^2} - \gamma t v$$

Thus, we have the formula x' at $N = 2$, and the extension to $N = 3$ and higher is straightforward, either by using a similar computation, or a rotation argument.

(7) Finally, as a matter of making sure that we didn't mess up anything with our reasonings and mathematics, let us verify that the inverse of the general Lorentz transform that we found is indeed given by $v \to -v$, which in practice means:

$$x = x' + (\gamma - 1)\frac{< v, x' > v}{||v||^2} + \gamma t' v$$

$$t = \gamma\left(t' + \frac{< v, x' >}{c^2}\right)$$

(8) For the space variable, the verification goes as follows:

$$\begin{aligned}
& x' + (\gamma-1)\frac{< v, x' > v}{||v||^2} + \gamma t' v \\
= \;& x + (\gamma-1)\frac{< v, x > v}{||v||^2} - \gamma t v \\
& + (\gamma-1)\frac{v}{||v||^2}\left(< v, x > + (\gamma-1) < v, x > - \gamma t ||v||^2\right) \\
& + \gamma^2 \left(t - \frac{< v, x >}{c^2}\right) v \\
= \;& x - (\gamma t + \gamma(\gamma-1)t - \gamma^2 t)v + \left(\frac{\gamma-1}{||v||^2} + \frac{\gamma(\gamma-1)}{||v||^2} - \frac{\gamma^2}{c^2}\right) < v, x > v \\
= \;& x + \left(\frac{\gamma^2-1}{||v||^2} - \frac{\gamma^2}{c^2}\right) < v, x > v \\
= \;& x
\end{aligned}$$

(9) As for the time variable, the verification here goes as follows:

$$\begin{aligned}
& \gamma\left(t' + \frac{< v, x' >}{c^2}\right) \\
= \;& \gamma\left(\gamma\left(t - \frac{< v, x >}{c^2}\right) + \frac{< v, x >}{c^2} + (\gamma-1)\frac{< v, x >}{c^2} - \gamma t \frac{||v||^2}{c^2}\right) \\
= \;& \gamma^2 t\left(1 - \frac{||v||^2}{c^2}\right) + \frac{\gamma < v, x >}{c^2}(-\gamma + 1 + \gamma - 1) \\
= \;& \gamma^2 t\left(1 - \frac{||v||^2}{c^2}\right) \\
= \;& t
\end{aligned}$$

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

What we found in Theorem 14.21 can be reformulated as follows:

THEOREM 14.22. *In the relativistic case, the general frame change formula is*

$$\begin{aligned}
x_1' &= x_1 + (\gamma-1)\frac{< v, x > v_1}{||v||^2} - \gamma t v_1 \\
&\vdots \\
x_N' &= x_N + (\gamma-1)\frac{< v, x > v_N}{||v||^2} - \gamma t v_N \\
ct' &= \gamma\left(ct - \frac{< v, x >}{c}\right)
\end{aligned}$$

where $\gamma = 1/\sqrt{1 - ||v||^2/c^2}$*, and the reverse frame change is obtained via* $v \to -v$.

PROOF. This is indeed a reformulation of what we found in Theorem 14.21, and with the time variable being multiplied, as it is quite standard, by c. □

In standard matrix form, all this does not look that great, and the result here, that we will only formulate at $N = 3$, for simplifying, is as follows:

THEOREM 14.23. *In the relativistic case, the general frame change formula is*

$$\begin{pmatrix} ct' \\ x_1' \\ x_2' \\ x_3' \end{pmatrix} = \begin{pmatrix} \gamma & -\beta_1\gamma & -\beta_2\gamma & -\beta_3\gamma \\ -\beta_1\gamma & 1+\alpha v_1^2 & \alpha v_1v_2 & \alpha v_1v_3 \\ -\beta_2\gamma & \alpha v_1v_2 & 1+\alpha v_2^2 & \alpha v_2v_3 \\ -\beta_3\gamma & \alpha v_1v_3 & \alpha v_2v_3 & 1+\alpha v_3^2 \end{pmatrix} \begin{pmatrix} ct \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

where $\gamma = 1/\sqrt{1-||v||^2/c^2}$ *as usual, and* $\alpha = (\gamma-1)/||v||^2$*, and* $\beta_i = v_i/c$*.*

PROOF. This is indeed a reformulation of what we found in Theorem 14.22, at $N = 3$, and with the rescaled time variable being put in the first position. □

As a second task now, let us recover the speed addition formula, established before, from the Lorentz transform. We can do this now in general, as follows:

THEOREM 14.24. *The speed addition formula in N-dimensional relativity is*

$$u +_e v = \frac{1}{1+<u,v>}\left(u+v+\frac{<u,v>u-<u,u>v}{1+\sqrt{1-||u||^2}}\right)$$

in $c = 1$ units.

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

(1) As before, the idea will be that of differentiating $x_1, \ldots, x_N$ and t in the formulae for the inverse Lorentz transform. With the replacement $v \to u$ for the moving speed, this inverse Lorentz transform is given by the following formula:

$$x_i = x_i' + (\gamma-1)\frac{<u,x'>u_i}{||u||^2} + \gamma t'u_i$$

$$t = \gamma\left(t' + \frac{<u,x'>}{c^2}\right)$$

(2) Now by differentiating, we obtain from this the following formulae:

$$dx_i = dx_i' + (\gamma-1)\frac{<u,dx'>u_i}{||u||^2} + \gamma u_i dt'$$

$$dt = \gamma\left(dt' + \frac{<u,dx'>}{c^2}\right)$$

(3) By dividing now the first formula by the second one, we obtain:

$$\frac{dx_i}{dt}=\frac{1}{\gamma}\cdot\frac{dx_i'+(\gamma-1)<u,dx'>u_i/||u||^2+\gamma u_i dt'}{dt'+<u,dx'>/c^2}$$

(4) Next, by dividing everything on the right by dt', we get from this:

$$\frac{dx_i}{dt}=\frac{1}{\gamma}\cdot\frac{dx_i'/dt'+(\gamma-1)<u,dx'/dt'>u_i/||u||^2+\gamma u_i}{1+<u,dx'/dt'>/c^2}$$

(5) In terms of speeds now, this means that we have, with $w=u+_e v$:

$$w_i=\frac{1}{\gamma}\cdot\frac{v_i+(\gamma-1)<u,v>u_i/||u||^2+\gamma u_i}{1+<u,v>/c^2}$$

(6) Now in $c=1$ units, this formula is as follows, still with $w=u+_e v$:

$$w_i=\frac{1}{\gamma}\cdot\frac{v_i+(\gamma-1)<u,v>u_i/||u||^2+\gamma u_i}{1+<u,v>}$$

(7) In vector notation now, the above formula shows that we have:

$$\begin{aligned}u+_e v&=\frac{1}{1+<u,v>}\cdot\frac{1}{\gamma}\left(v+(\gamma-1)\frac{<u,v>u}{||u||^2}+\gamma u\right)\\&=\frac{1}{1+<u,v>}\left(u+\frac{v}{\gamma}+\left(1-\frac{1}{\gamma}\right)\frac{<u,v>u}{||u||^2}\right)\\&=\frac{1}{1+<u,v>}\left(u+v+\left(1-\frac{1}{\gamma}\right)\left(\frac{<u,v>u}{||u||^2}-v\right)\right)\\&=\frac{1}{1+<u,v>}\left(u+v+\left(1-\frac{1}{\gamma}\right)\frac{<u,v>u-<u,u>v}{||u||^2}\right)\\&=\frac{1}{1+<u,v>}\left(u+v+\frac{<u,v>u-<u,u>v}{1+\sqrt{1-||u||^2}}\right)\end{aligned}$$

(8) Here we have used at the end the following formula, for the Lorentz factor:

$$\begin{aligned}1-\frac{1}{\gamma}&=1-\frac{1}{1/\sqrt{1-||u||^2}}\\&=1-\sqrt{1-||u||^2}\\&=\frac{||u||^2}{1+\sqrt{1-||u||^2}}\end{aligned}$$

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

Getting now to spacetime, in non-relativistic physics two events are separated by space Δx and by time Δt, with these two separation variables being independent. In relativistic physics this is no longer true, and the correct analogue of this comes from:

THEOREM 14.25. *The following quantity, called relativistic spacetime separation*

$$\Delta s^2 = c^2\Delta t^2 - (\Delta x^2 + \Delta y^2 + \Delta z^2)$$

is invariant under relativistic frame changes.

PROOF. We must prove that the quantity $K = c^2t^2 - x^2 - y^2 - z^2$ is invariant under Lorentz transformations. For this purpose, observe that we have:

$$K = \left\langle \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}, \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} \right\rangle$$

Now recall that the Lorentz transformation is given by the following formula, where $\gamma = 1/\sqrt{1 - v^2/c^2}$ as usual, and where $\beta = v/c$:

$$\begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix}$$

Thus, if we denote by L the matrix of the Lorentz transformation, and by E the matrix found before, we must prove that for any vector ξ we have:

$$< E\xi, \xi > = < EL\xi, L\xi >$$

Since L is symmetric we have $< EL\xi, L\xi > = < LEL\xi, \xi >$, so we must prove:

$$E = LEL$$

But this is the same as proving $L^{-1}E = EL$, and by using the fact that $L \to L^{-1}$ is given by $\beta \to -\beta$, what we eventually want to prove is that:

$$L_{-\beta}E = EL_\beta$$

So, let us prove this. As usual we can restrict the attention to the upper left corner, call that NW corner, and here we have the following computations:

$$(L_{-\beta}E)_{NW} = \begin{pmatrix} \gamma & \beta\gamma \\ \beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma \\ \beta\gamma & -\gamma \end{pmatrix}$$

$$(EL_\beta)_{NW} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \gamma & -\beta\gamma \\ -\beta\gamma & \gamma \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma \\ \beta\gamma & -\gamma \end{pmatrix}$$

The matrices on the right being equal, this gives the result. □

Finally, let us discuss gravity. This can be incorporated too, as follows:

THEOREM 14.26 (Einstein). *The theory of gravity can be suitably modified, and merged with relativity, into a theory called general relativity.*

PROOF. All this is a bit complicated, involving some geometry, as follows:

(1) Before anything, we have seen that in the relativistic context, mass m must be replaced by relativistic mass $M = \gamma m$, and momentum $p = mv$ must be replaced by relativistic momentum $P = Mv$. Thus, as with Galileo and many other things, such as the conservation of mass and of momentum, seen above, there is a bug with the Newton formula $F = \dot{p}$, which must be replaced by something of type $F = \dot{P}$.

(2) In practice now, as a starting point, let us go back to the formula $F = -\Delta V$, that we know well. Geometrically, this suggests looking at the gravitational field of k bodies $M_1, \ldots, M_k$ as being represented by $\mathbb{R}^3$ having k holes in it, and with the heavier the M_i, the bigger the hole, and with poor $m \simeq 0$ having to roll on all this.

(3) Of course we are here in 4D, for the full picture, that of the potential V, or rather of its graph, and in order to better understand this, it is of help to first consider the question where our bodies $M_1, \ldots, M_k$ lie in a plane $\mathbb{R}^2$.

(4) Still staying inside classical mechanics, it is possible to further build on the above picture in (2), which was something rather intuitive, now with some precise math formulae, relating the geometry of V to the motion of m under its influence.

(5) The point now is that, with (4) done, the passage to relativity can be understood as well, by modifying a bit the geometry there, as to fit with relativistic spacetime, and by having the $F = \dot{P}$ idea from (1) in mind too. That is the main idea behind general relativity, and in practice, all this needs a bit of technical geometry and formulae. □

This was for the basics of Einstein's relativity theory. For more, we refer to his book [28], which is a must-read, for any mathematician, physicist, scientist, or non-scientist.

14e. Exercises

Exercises:

EXERCISE 14.27.

EXERCISE 14.28.

EXERCISE 14.29.

EXERCISE 14.30.

EXERCISE 14.31.

EXERCISE 14.32.

EXERCISE 14.33.

EXERCISE 14.34.

Bonus exercise.

CHAPTER 15

Infinite matrices

15a. Linear operators

We would like to discuss now the theory of linear operators $T : H \to H$ over a complex Hilbert space H, usually taken separable. Let us start with a basic result, as follows:

THEOREM 15.1. *Given a Hilbert space H, consider the linear operators $T : H \to H$, and for each such operator define its norm by the following formula:*

$$||T|| = \sup_{||x||=1} ||Tx||$$

The operators which are bounded, $||T|| < \infty$, form then a complex algebra $B(H)$, which is complete with respect to $||.||$. When H comes with a basis $\{e_i\}_{i\in I}$, we have

$$B(H) \subset \mathcal{L}(H) \subset M_I(\mathbb{C})$$

where $\mathcal{L}(H)$ is the algebra of all linear operators $T : H \to H$, and $\mathcal{L}(H) \subset M_I(\mathbb{C})$ is the correspondence $T \to M$ obtained via the usual linear algebra formulae, namely:

$$T(x) = Mx \quad , \quad M_{ij} = < Te_j, e_i >$$

In infinite dimensions, none of the above two inclusions is an equality.

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

(1) The fact that we have indeed an algebra, satisfying the product condition in the statement, follows from the following estimates, which are all elementary:

$$||S + T|| \leq ||S|| + ||T|| \quad , \quad ||\lambda T|| = |\lambda| \cdot ||T|| \quad , \quad ||ST|| \leq ||S|| \cdot ||T||$$

(2) Regarding now the completness assertion, if $\{T_n\} \subset B(H)$ is Cauchy then $\{T_n x\}$ is Cauchy for any $x \in H$, so we can define the limit $T = \lim_{n\to\infty} T_n$ by setting:

$$Tx = \lim_{n\to\infty} T_n x$$

Let us first check that the application $x \to Tx$ is linear. We have:

$$\begin{aligned} T(x+y) &= \lim_{n\to\infty} T_n(x+y) \\ &= \lim_{n\to\infty} T_n(x) + T_n(y) \\ &= \lim_{n\to\infty} T_n(x) + \lim_{n\to\infty} T_n(y) \\ &= T(x) + T(y) \end{aligned}$$

Similarly, we have $T(\lambda x) = \lambda T(x)$, and we conclude that $T \in \mathcal{L}(H)$.

(3) With this done, it remains to prove now that we have $T \in B(H)$, and that $T_n \to T$ in norm. For this purpose, observe that we have:

$$\begin{aligned} ||T_n - T_m|| \leq \varepsilon \ , \ \forall n, m \geq N &\implies ||T_n x - T_m x|| \leq \varepsilon \ , \ \forall ||x|| = 1 \ , \ \forall n, m \geq N \\ &\implies ||T_n x - Tx|| \leq \varepsilon \ , \ \forall ||x|| = 1 \ , \ \forall n \geq N \\ &\implies ||T_N x - Tx|| \leq \varepsilon \ , \ \forall ||x|| = 1 \\ &\implies ||T_N - T|| \leq \varepsilon \end{aligned}$$

But this gives both $T \in B(H)$, and $T_N \to T$ in norm, and we are done.

(4) Regarding the embeddings, the correspondence $T \to M$ in the statement is indeed linear, and its kernel is $\{0\}$, so we have indeed an embedding as follows, as claimed:

$$\mathcal{L}(H) \subset M_I(\mathbb{C})$$

In finite dimensions we have an isomorphism, because any $M \in M_N(\mathbb{C})$ determines an operator $T : \mathbb{C}^N \to \mathbb{C}^N$, given by $< Te_j, e_i >= M_{ij}$. However, in infinite dimensions, we have matrices not producing operators, as for instance the all-one matrix.

(5) As for the examples of linear operators which are not bounded, these are more complicated, coming from logic, and we will not really need them in what follows. □

As a second basic result regarding the operators, we will need:

THEOREM 15.2. *Each operator $T \in B(H)$ has an adjoint $T^* \in B(H)$, given by:*

$$< Tx, y >=< x, T^* y >$$

The operation $T \to T^*$ is antilinear, antimultiplicative, involutive, and satisfies:

$$||T|| = ||T^*|| \quad , \quad ||TT^*|| = ||T||^2$$

When H comes with a basis $\{e_i\}_{i\in I}$, the operation $T \to T^*$ corresponds to

$$(M^*)_{ij} = \overline{M}_{ji}$$

at the level of the associated matrices $M \in M_I(\mathbb{C})$.

PROOF. This is standard too, and can be proved in 3 steps, as follows:

(1) The existence of the adjoint operator T^*, given by the formula in the statement, comes from the fact that the function $\varphi(x) =< Tx, y >$ being a linear map $H \to \mathbb{C}$, we must have a formula as follows, for a certain vector $T^*y \in H$:

$$\varphi(x) =< x, T^*y >$$

Moreover, since this vector is unique, T^* is unique too, and we have as well:

$$(S+T)^* = S^* + T^* \quad , \quad (\lambda T)^* = \bar{\lambda} T^* \quad , \quad (ST)^* = T^* S^* \quad , \quad (T^*)^* = T$$

Observe also that we have indeed $T^* \in B(H)$, because:

$$\begin{aligned} ||T|| &= \sup_{||x||=1} \sup_{||y||=1} < Tx, y > \\ &= \sup_{||y||=1} \sup_{||x||=1} < x, T^*y > \\ &= ||T^*|| \end{aligned}$$

(2) Regarding now $||TT^*|| = ||T||^2$, which is a key formula, observe that we have:

$$||TT^*|| \leq ||T|| \cdot ||T^*|| = ||T||^2$$

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

$$\begin{aligned} ||T||^2 &= \sup_{||x||=1} | < Tx, Tx > | \\ &= \sup_{||x||=1} | < x, T^*Tx > | \\ &\leq ||T^*T|| \end{aligned}$$

By replacing $T \to T^*$ we obtain from this $||T||^2 \leq ||TT^*||$, as desired.

(3) Finally, when H comes with a basis, the formula $< Tx, y >=< x, T^*y >$ applied with $x = e_i$, $y = e_j$ translates into the formula $(M^*)_{ij} = \overline{M}_{ji}$, as desired. □

Let us discuss now the diagonalization problem for the operators $T \in B(H)$, in analogy with the diagonalization problem for the usual matrices $A \in M_N(\mathbb{C})$. As a first observation, we can talk about eigenvalues and eigenvectors, as follows:

DEFINITION 15.3. *Given an operator $T \in B(H)$, assuming that we have*

$$Tx = \lambda x$$

we say that $x \in H$ is an eigenvector of T, with eigenvalue $\lambda \in \mathbb{C}$.

We know many things about eigenvalues and eigenvectors, in the finite dimensional case. However, most of these will not extend to the infinite dimensional case, or at least not extend in a straightforward way, due to a number of reasons:

(1) Most of basic linear algebra is based on the fact that $Tx = \lambda x$ is equivalent to $(T - \lambda)x = 0$, so that λ is an eigenvalue when $T - \lambda$ is not invertible. In the infinite dimensional setting $T - \lambda$ might be injective and not surjective, or vice versa, or invertible with $(T - \lambda)^{-1}$ not bounded, and so on.

(2) Also, in linear algebra $T - \lambda$ is not invertible when $\det(T - \lambda) = 0$, and with this leading to most of the advanced results about eigenvalues and eigenvectors. In infinite dimensions, however, it is impossible to construct a determinant function $\det : B(H) \to \mathbb{C}$, and this even for the diagonal operators on $l^2(\mathbb{N})$.

Summarizing, we are in trouble. Forgetting about (2), which obviously leads nowhere, let us focus on the difficulties in (1). In order to cut short the discussion there, regarding the various properties of $T-\lambda$, we can just say that $T-\lambda$ is either invertible with bounded inverse, the "good case", or not. We are led in this way to the following definition:

DEFINITION 15.4. *The spectrum of an operator $T \in B(H)$ is the set*

$$\sigma(T) = \left\{\lambda \in \mathbb{C}\middle| T - \lambda \notin B(H)^{-1}\right\}$$

where $B(H)^{-1} \subset B(H)$ is the set of invertible operators.

As a basic example, in the finite dimensional case, $H = \mathbb{C}^N$, the spectrum of a usual matrix $A \in M_N(\mathbb{C})$ is the collection of its eigenvalues, taken without multiplicities. We will see many other examples. In general, the spectrum has the following properties:

PROPOSITION 15.5. *The spectrum of $T \in B(H)$ contains the eigenvalue set*

$$\varepsilon(T) = \left\{\lambda \in \mathbb{C}\middle| \ker(T - \lambda) \neq \{0\}\right\}$$

and $\varepsilon(T) \subset \sigma(T)$ is an equality in finite dimensions, but not in infinite dimensions.

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

(1) First of all, the eigenvalue set is indeed the one in the statement, because $Tx = \lambda x$ tells us precisely that $T - \lambda$ must be not injective. The fact that we have $\varepsilon(T) \subset \sigma(T)$ is clear as well, because if $T - \lambda$ is not injective, it is not bijective.

(2) In finite dimensions we have $\varepsilon(T) = \sigma(T)$, because $T - \lambda$ is injective if and only if it is bijective, with the boundedness of the inverse being automatic.

(3) In infinite dimensions we can assume $H = l^2(\mathbb{N})$, and the shift operator $S(e_i) = e_{i+1}$ is injective but not surjective. Thus $0 \in \sigma(T) - \varepsilon(T)$. □

Philosophically, the best way of thinking at this is as follows: the numbers $\lambda \notin \sigma(T)$ are good, because we can invert $T - \lambda$, the numbers $\lambda \in \sigma(T) - \varepsilon(T)$ are bad, because so they are, and the eigenvalues $\lambda \in \varepsilon(T)$ are evil. Welcome to operator theory.

Let us develop now some general theory. As a first goal, we would like to prove that the spectra are non-empty. This is something quite tricky, the result being as follows:

THEOREM 15.6. *The spectrum of a bounded operator $T \in B(H)$ is:*

(1) *Compact.*
(2) *Contained in the disc $D_0(||T||)$.*
(3) *Non-empty.*

PROOF. This can be proved by using some complex analysis, as follows:

(1) In view of (2) below, it is enough to prove that $\sigma(T)$ is closed. But this follows from the following computation, with $|\varepsilon|$ being small:

$$\begin{aligned}\lambda \notin \sigma(T) &\implies T - \lambda \in B(H)^{-1}\\ &\implies T - \lambda - \varepsilon \in B(H)^{-1}\\ &\implies \lambda + \varepsilon \notin \sigma(T)\end{aligned}$$

(2) This follows indeed from the following computation:

$$\begin{aligned}\lambda > ||T|| &\implies \left|\left|\frac{T}{\lambda}\right|\right| < 1\\ &\implies 1 - \frac{T}{\lambda} \in B(H)^{-1}\\ &\implies \lambda - T \in B(H)^{-1}\\ &\implies \lambda \notin \sigma(T)\end{aligned}$$

(3) Assume by contradiction $\sigma(T) = \emptyset$. Given a linear form $f \in B(H)^*$, consider the following map, which is well-defined, due to our assumption $\sigma(T) = \emptyset$:

$$\varphi : \mathbb{C} \to \mathbb{C} \quad , \quad \lambda \to f((T - \lambda)^{-1})$$

By using the fact that $T \to T^{-1}$ is differentiable, which is something elementary, we conclude that this map is differentiable, and so holomorphic. Also, we have:

$$\begin{aligned}\lambda \to \infty &\implies T - \lambda \to \infty\\ &\implies (T - \lambda)^{-1} \to 0\\ &\implies f((T - \lambda))^{-1} \to 0\end{aligned}$$

Thus by the Liouville theorem we obtain $\varphi = 0$. But, in view of the definition of φ, this gives $(T - \lambda)^{-1} = 0$, which is a contradiction, as desired. □

Here is now a second basic result regarding the spectra, inspired from what happens in finite dimensions, for the usual complex matrices, and which shows that things do not necessarily extend without troubles to the infinite dimensional setting:

THEOREM 15.7. *We have the following formula, valid for any operators S, T:*

$$\sigma(ST) \cup \{0\} = \sigma(TS) \cup \{0\}$$

In finite dimensions we have $\sigma(ST) = \sigma(TS)$, but this fails in infinite dimensions.

PROOF. There are several assertions here, the idea being as follows:

(1) This is something that we know in finite dimensions, coming from the fact that the characteristic polynomials of the associated matrices A, B coincide:

$$P_{AB} = P_{BA}$$

Thus we obtain $\sigma(ST) = \sigma(TS)$ in this case, as claimed. Observe that this improves twice the general formula in the statement, first because we have no issues at 0, and second because what we obtain is actually an equality of sets with mutiplicities.

(2) In general now, let us first prove the main assertion, stating that $\sigma(ST), \sigma(TS)$ coincide outside 0. We first prove that we have the following implication:

$$1 \notin \sigma(ST) \implies 1 \notin \sigma(TS)$$

Assume indeed that $1 - ST$ is invertible, with inverse denoted R:

$$R = (1 - ST)^{-1}$$

We have then the following formulae, relating our variables R, S, T:

$$RST = STR = R - 1$$

By using $RST = R - 1$, we have the following computation:

$$\begin{aligned}(1 + TRS)(1 - TS) &= 1 + TRS - TS - TRSTS \\ &= 1 + TRS - TS - TRS + TS \\ &= 1\end{aligned}$$

A similar computation, using $STR = R - 1$, shows that we have:

$$(1 - TS)(1 + TRS) = 1$$

Thus $1 - TS$ is invertible, with inverse $1 + TRS$, which proves our claim. Now by multiplying by scalars, we deduce from this that for any $\lambda \in \mathbb{C} - \{0\}$ we have:

$$\lambda \notin \sigma(ST) \implies \lambda \notin \sigma(TS)$$

But this leads to the conclusion in the statement.

(3) Regarding now the counterexample to the formula $\sigma(ST) = \sigma(TS)$, in general, let us take S to be the shift on $H = L^2(\mathbb{N})$, given by the following formula:

$$S(e_i) = e_{i+1}$$

As for T, we can take it to be the adjoint of S, and we have:

$$S^*S = 1 \implies 0 \notin \sigma(SS^*)$$

$$SS^* = Proj(e_0^\perp) \implies 0 \in \sigma(SS^*)$$

Thus, the spectra do not match on 0, and so we have our counterexample. □

15b. Spectral radius

Let us develop now some systematic theory for the computation of the spectra, based on what we know about the eigenvalues of the usual complex matrices. As a first result, which is well-known for the usual matrices, and extends well, we have:

THEOREM 15.8. *We have the "polynomial functional calculus" formula*

$$\sigma(P(T)) = P(\sigma(T))$$

valid for any polynomial $P \in \mathbb{C}[X]$, and any operator $T \in B(H)$.

PROOF. We pick a scalar $\lambda \in \mathbb{C}$, and we decompose the polynomial $P - \lambda$:

$$P(X) - \lambda = c(X - r_1) \ldots (X - r_n)$$

We have then the following equivalences:

$$\begin{aligned}
\lambda \notin \sigma(P(T)) &\iff P(T) - \lambda \in B(H)^{-1} \\
&\iff c(T - r_1) \ldots (T - r_n) \in B(H)^{-1} \\
&\iff T - r_1, \ldots, T - r_n \in B(H)^{-1} \\
&\iff r_1, \ldots, r_n \notin \sigma(T) \\
&\iff \lambda \notin P(\sigma(T))
\end{aligned}$$

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

The above result is something very useful, and generalizing it will be our next task. As a first ingredient here, assuming that $A \in M_N(\mathbb{C})$ is invertible, we have:

$$\sigma(A^{-1}) = \sigma(A)^{-1}$$

It is possible to extend this formula to the arbitrary operators, and we will do this in a moment. Before starting, however, we have to find a class of functions generalizing both the polynomials $P \in \mathbb{C}[X]$ and the inverse function $x \to x^{-1}$. The answer to this question is provided by the rational functions, which are as follows:

DEFINITION 15.9. *A rational function $f \in \mathbb{C}(X)$ is a quotient of polynomials:*

$$f = \frac{P}{Q}$$

Assuming that P, Q are prime to each other, we can regard f as a usual function,

$$f : \mathbb{C} - X \to \mathbb{C}$$

with X being the set of zeros of Q, also called poles of f.

Now that we have our class of functions, the next step consists in applying them to operators. Here we cannot expect $f(T)$ to make sense for any f and any T, for instance because T^{-1} is defined only when T is invertible. We are led in this way to:

DEFINITION 15.10. *Given an operator $T \in B(H)$, and a rational function $f = P/Q$ having poles outside $\sigma(T)$, we can construct the following operator,*

$$f(T) = P(T)Q(T)^{-1}$$

that we can denote as a usual fraction, as follows,

$$f(T) = \frac{P(T)}{Q(T)}$$

due to the fact that $P(T), Q(T)$ commute, so that the order is irrelevant.

To be more precise, $f(T)$ is indeed well-defined, and the fraction notation is justified too. In more formal terms, we can say that we have a morphism of complex algebras as follows, with $\mathbb{C}(X)^T$ standing for the rational functions having poles outside $\sigma(T)$:

$$\mathbb{C}(X)^T \to B(H) \quad , \quad f \to f(T)$$

Summarizing, we have now a good class of functions, generalizing both the polynomials and the inverse map $x \to x^{-1}$. We can now extend Theorem 15.8, as follows:

THEOREM 15.11. *We have the "rational functional calculus" formula*

$$\sigma(f(T)) = f(\sigma(T))$$

valid for any rational function $f \in \mathbb{C}(X)$ having poles outside $\sigma(T)$.

PROOF. We pick a scalar $\lambda \in \mathbb{C}$, we write $f = P/Q$, and we set:

$$F = P - \lambda Q$$

By using now Theorem 15.8, for this polynomial, we obtain:

$$\begin{aligned}
\lambda \in \sigma(f(T)) &\iff F(T) \notin B(H)^{-1} \\
&\iff 0 \in \sigma(F(T)) \\
&\iff 0 \in F(\sigma(T)) \\
&\iff \exists \mu \in \sigma(T), F(\mu) = 0 \\
&\iff \lambda \in f(\sigma(T))
\end{aligned}$$

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

As an application of the above methods, we can investigate certain special classes of operators, such as the self-adjoint ones, and the unitary ones. Let us start with:

PROPOSITION 15.12. *The following happen:*

(1) *We have $\sigma(T^*) = \overline{\sigma(T)}$, for any $T \in B(H)$.*
(2) *If $T = T^*$ then $X = \sigma(T)$ satisfies $X = \overline{X}$.*
(3) *If $U^* = U^{-1}$ then $X = \sigma(U)$ satisfies $X^{-1} = \overline{X}$.*

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

(1) The spectrum of the adjoint operator T^* can be computed as follows:

$$\begin{aligned}\sigma(T^*) &= \left\{\lambda\in\mathbb{C}\Big|T^*-\lambda\notin B(H)^{-1}\right\}\\ &= \left\{\lambda\in\mathbb{C}\Big|T-\bar{\lambda}\notin B(H)^{-1}\right\}\\ &= \overline{\sigma(T)}\end{aligned}$$

(2) This is clear indeed from (1).

(3) For a unitary operator, $U^*=U^{-1}$, Theorem 1.11 and (1) give:

$$\sigma(U)^{-1}=\sigma(U^{-1})=\sigma(U^*)=\overline{\sigma(U)}$$

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

In analogy with what happens for the usual matrices, we would like to improve now (2,3) above, with results stating that the spectrum $X=\sigma(T)$ satisfies $X\subset\mathbb{R}$ for self-adjoints, and $X\subset\mathbb{T}$ for unitaries. This will be tricky. Let us start with:

THEOREM 15.13. *The spectrum of a unitary operator*

$$U^*=U^{-1}$$

is on the unit circle, $\sigma(U)\subset\mathbb{T}$.

PROOF. Assuming $U^*=U^{-1}$, we have the following norm computation:

$$||U||=\sqrt{||UU^*||}=\sqrt{1}=1$$

Now if we denote by D the unit disk, we obtain from this:

$$\sigma(U)\subset D$$

On the other hand, once again by using $U^*=U^{-1}$, we have as well:

$$||U^{-1}||=||U^*||=||U||=1$$

Thus, as before with D being the unit disk in the complex plane, we have:

$$\sigma(U^{-1})\subset D$$

Now by using Theorem 15.11, we obtain $\sigma(U)\subset D\cap D^{-1}=\mathbb{T}$, as desired. □

We have as well a similar result for the self-adjoints, as follows:

THEOREM 15.14. *The spectrum of a self-adjoint operator*

$$T=T^*$$

consists of real numbers, $\sigma(T)\subset\mathbb{R}$.

PROOF. The idea is that we can deduce the result from Theorem 15.13, by using the following remarkable rational function, depending on a parameter $r \in \mathbb{R}$:

$$f(z) = \frac{z + ir}{z - ir}$$

Indeed, for $r >> 0$ the operator $f(T)$ is well-defined, and we have:

$$\left(\frac{T + ir}{T - ir}\right)^* = \frac{T - ir}{T + ir} = \left(\frac{T + ir}{T - ir}\right)^{-1}$$

Thus $f(T)$ is unitary, and by using Theorem 15.13 we obtain:

$$\begin{aligned} \sigma(T) &\subset f^{-1}(f(\sigma(T))) \\ &= f^{-1}(\sigma(f(T))) \\ &\subset f^{-1}(\mathbb{T}) \\ &= \mathbb{R} \end{aligned}$$

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

One key thing that we know about matrices, which is clear for the diagonalizable matrices, and then in general follows by density, is the following formula:

$$\sigma(e^A) = e^{\sigma(A)}$$

We would like to have such formulae for the general operators $T \in B(H)$, but this is something quite technical. Consider the rational calculus morphism from Definition 15.10, which is as follows, with the exponent standing for "having poles outside $\sigma(T)$":

$$\mathbb{C}(X)^T \to B(H) \quad , \quad f \to f(T)$$

As mentioned before, the rational functions are holomorphic outside their poles, and this raises the question of extending this morphism, as follows:

$$Hol(\sigma(T)) \to B(H) \quad , \quad f \to f(T)$$

But for this, we can use the Cauchy formula. Indeed, given a function $f \in \mathbb{C}(X)^T$, the operator $f(T) \in B(H)$ from Definition 15.10 can be recaptured as follows:

$$f(T) = \frac{1}{2\pi i} \int_\gamma \frac{f(z)}{z - T}\, dz$$

Now given an arbitrary function $f \in Hol(\sigma(T))$, we can define $f(T) \in B(H)$ by the exactly same formula, and we obtain in this way the desired correspondence:

$$Hol(\sigma(T)) \to B(H) \quad , \quad f \to f(T)$$

This was for the plan. In practice now, all this needs a bit of care, with many verifications needed, and with the technical remark that a winding number must be added to the above Cauchy formulae, for things to be correct. The result is as follows:

THEOREM 15.15. *Given $T \in B(H)$, we have a morphism of algebras as follows, where $Hol(\sigma(T))$ is the algebra of functions which are holomorphic around $\sigma(T)$,*

$$Hol(\sigma(T)) \to B(H) \quad , \quad f \to f(T)$$

which extends the previous rational functional calculus $f \to f(T)$. We have:

$$\sigma(f(T)) = f(\sigma(T))$$

Moreover, if $\sigma(T)$ is contained in an open set U and $f_n, f : U \to \mathbb{C}$ are holomorphic functions such that $f_n \to f$ uniformly on compact subsets of U then $f_n(T) \to f(T)$.

PROOF. This follows indeed by reasoning along the above lines, by making a heavy use of the Cauchy formula, and for full details here, we refer to any specialized operator theory book. In what follows, we will not really need this result. □

In order to formulate now our next result, we will need the following notion:

DEFINITION 15.16. *Given an operator $T \in B(H)$, its spectral radius*

$$\rho(T) \in \big[0, ||T||\big]$$

is the radius of the smallest disk centered at 0 containing $\sigma(T)$.

Now with this notion in hand, we have the following key result, improving our key theoretical result so far about spectra, namely $\sigma(T) \neq \emptyset$, from Theorem 15.6:

THEOREM 15.17. *The spectral radius of an operator $T \in B(H)$ is given by*

$$\rho(T) = \lim_{n\to\infty} ||T^n||^{1/n}$$

and in this formula, we can replace the limit by an inf.

PROOF. We have several things to be proved, the idea being as follows:

(1) Our first claim is that the numbers $u_n = ||T^n||^{1/n}$ satisfy:

$$(n+m)u_{n+m} \leq nu_n + mu_m$$

Indeed, we have the following estimate, using the Young inequality $ab \leq a^p/p + b^q/q$, with exponents $p = (n+m)/n$ and $q = (n+m)/m$:

$$\begin{aligned} u_{n+m} &= ||T^{n+m}||^{1/(n+m)} \\ &\leq ||T^n||^{1/(n+m)}||T^m||^{1/(n+m)} \\ &\leq ||T^n||^{1/n} \cdot \frac{n}{n+m} + ||T^m||^{1/m} \cdot \frac{m}{n+m} \\ &= \frac{nu_n + mu_m}{n+m} \end{aligned}$$

(2) Our second claim is that the second assertion holds, namely:

$$\lim_{n\to\infty} ||T^n||^{1/n} = \inf_n ||T^n||^{1/n}$$

For this purpose, we just need the inequality found in (1). Indeed, fix $m \geq 1$, let $n \geq 1$, and write $n = lm + r$ with $0 \leq r \leq m - 1$. By using twice $u_{ab} \leq u_b$, we get:

$$\begin{aligned} u_n &\leq \frac{1}{n}(lmu_{lm} + ru_r) \\ &\leq \frac{1}{n}(lmu_m + ru_1) \\ &\leq u_m + \frac{r}{n}\,u_1 \end{aligned}$$

It follows that we have $\limsup_n u_n \leq u_m$, which proves our claim.

(3) Summarizing, we are left with proving the main formula, which is as follows, and with the remark that we already know that the sequence on the right converges:

$$\rho(T) = \lim_{n\to\infty} ||T^n||^{1/n}$$

In one sense, we can use the polynomial calculus formula $\sigma(T^n) = \sigma(T)^n$. Indeed, this gives the following estimate, valid for any n, as desired:

$$\begin{aligned} \rho(T) &= \sup_{\lambda\in\sigma(T)} |\lambda| \\ &= \sup_{\rho\in\sigma(T)^n} |\rho|^{1/n} \\ &= \sup_{\rho\in\sigma(T^n)} |\rho|^{1/n} \\ &= \rho(T^n)^{1/n} \\ &\leq ||T^n||^{1/n} \end{aligned}$$

(4) For the reverse inequality, we fix a number $\rho > \rho(T)$, and we want to prove that we have $\rho \geq \lim_{n\to\infty} ||T^n||^{1/n}$. By using the Cauchy formula, we have:

$$\begin{aligned} \frac{1}{2\pi i}\int_{|z|=\rho} \frac{z^n}{z-T}\,dz &= \frac{1}{2\pi i}\int_{|z|=\rho} \sum_{k=0}^{\infty} z^{n-k-1}T^k\,dz \\ &= \sum_{k=0}^{\infty} \frac{1}{2\pi i}\left(\int_{|z|=\rho} z^{n-k-1}dz\right)T^k \\ &= \sum_{k=0}^{\infty} \delta_{n,k+1}T^k \\ &= T^{n-1} \end{aligned}$$

By applying the norm we obtain from this formula:

$$||T^{n-1}|| \leq \frac{1}{2\pi}\int_{|z|=\rho} \left|\left|\frac{z^n}{z-T}\right|\right| dz \leq \rho^n \cdot \sup_{|z|=\rho} \left|\left|\frac{1}{z-T}\right|\right|$$

Since the sup does not depend on n, by taking n-th roots, we obtain in the limit:

$$\rho \geq \lim_{n\to\infty} ||T^n||^{1/n}$$

Now recall that ρ was by definition an arbitrary number satisfying $\rho > \rho(T)$. Thus, we have obtained the following estimate, valid for any $T \in B(H)$:

$$\rho(T) \geq \lim_{n\to\infty} ||T^n||^{1/n}$$

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

In the case of the normal elements, we have the following finer result:

THEOREM 15.18. *The spectral radius of a normal element,*

$$TT^* = T^*T$$

is equal to its norm.

PROOF. We can proceed in two steps, as follows:

Step 1. In the case $T = T^*$ we have $||T^n|| = ||T||^n$ for any exponent of the form $n = \overline{2^k}$, by using the formula $||TT^*|| = ||T||^2$, and by taking n-th roots we get:

$$\rho(T) \geq ||T||$$

Thus, we are done with the self-adjoint case, with the result $\rho(T) = ||T||$.

Step 2. In the general normal case $TT^* = T^*T$ we have $T^n(T^n)^* = (TT^*)^n$, and by using this, along with the result from Step 1, applied to TT^*, we obtain:

$$\begin{aligned}
\rho(T) &= \lim_{n\to\infty} ||T^n||^{1/n} \\
&= \sqrt{\lim_{n\to\infty} ||T^n(T^n)^*||^{1/n}} \\
&= \sqrt{\lim_{n\to\infty} ||(TT^*)^n||^{1/n}} \\
&= \sqrt{\rho(TT^*)} \\
&= \sqrt{||T||^2} \\
&= ||T||
\end{aligned}$$

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

15c. Normal operators

By using Theorem 15.18 we can say a number of non-trivial things about the normal operators, commonly known as "spectral theorem for normal operators". As a first result here, we can improve the polynomial functional calculus formula, as follows:

THEOREM 15.19. *Given $T \in B(H)$ normal, we have a morphism of algebras*

$$\mathbb{C}[X] \to B(H) \quad , \quad P \to P(T)$$

having the properties $||P(T)|| = ||P_{|\sigma(T)}||$, and $\sigma(P(T)) = P(\sigma(T))$.

PROOF. This is an improvement of Theorem 15.8 in the normal case, with the extra assertion being the norm estimate. But the element $P(T)$ being normal, we can apply to it the spectral radius formula for normal elements, and we obtain:

$$\begin{aligned} ||P(T)|| &= \rho(P(T)) \\ &= \sup_{\lambda \in \sigma(P(T))} |\lambda| \\ &= \sup_{\lambda \in P(\sigma(T))} |\lambda| \\ &= ||P_{|\sigma(T)}|| \end{aligned}$$

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

We can improve as well the rational calculus formula, and the holomorphic calculus formula, in the same way. Importantly now, at a more advanced level, we have:

THEOREM 15.20. *Given $T \in B(H)$ normal, we have a morphism of algebras*

$$C(\sigma(T)) \to B(H) \quad , \quad f \to f(T)$$

which is isometric, $||f(T)|| = ||f||$, and has the property $\sigma(f(T)) = f(\sigma(T))$.

PROOF. The idea here is to "complete" the morphism in Theorem 15.19, namely:

$$\mathbb{C}[X] \to B(H) \quad , \quad P \to P(T)$$

Indeed, we know from Theorem 1.19 that this morphism is continuous, and is in fact isometric, when regarding the polynomials $P \in \mathbb{C}[X]$ as functions on $\sigma(T)$:

$$||P(T)|| = ||P_{|\sigma(T)}||$$

Thus, by Stone-Weierstrass, we have a unique isometric extension, as follows:

$$C(\sigma(T)) \to B(H) \quad , \quad f \to f(T)$$

It remains to prove $\sigma(f(T)) = f(\sigma(T))$, and we can do this by double inclusion:

"$\subset$" Given a continuous function $f \in C(\sigma(T))$, we must prove that we have:

$$\lambda \notin f(\sigma(T)) \implies \lambda \notin \sigma(f(T))$$

For this purpose, consider the following function, which is well-defined:

$$\frac{1}{f - \lambda} \in C(\sigma(T))$$

We can therefore apply this function to T, and we obtain:

$$\left(\frac{1}{f-\lambda}\right)T=\frac{1}{f(T)-\lambda}$$

In particular $f(T)-\lambda$ is invertible, so $\lambda\notin\sigma(f(T))$, as desired.

"$\supset$" Given a continuous function $f\in C(\sigma(T))$, we must prove that we have:

$$\lambda\in f(\sigma(T))\implies\lambda\in\sigma(f(T))$$

But this is the same as proving that we have:

$$\mu\in\sigma(T)\implies f(\mu)\in\sigma(f(T))$$

For this purpose, we approximate our function by polynomials, $P_n\to f$, and we examine the following convergence, which follows from $P_n\to f$:

$$P_n(T)-P_n(\mu)\to f(T)-f(\mu)$$

We know from polynomial functional calculus that we have:

$$P_n(\mu)\in P_n(\sigma(T))=\sigma(P_n(T))$$

Thus, the operators $P_n(T)-P_n(\mu)$ are not invertible. On the other hand, we know that the set formed by the invertible operators is open, so its complement is closed. Thus the limit $f(T)-f(\mu)$ is not invertible either, and so $f(\mu)\in\sigma(f(T))$, as desired. □

As an important comment, Theorem 15.20 is not exactly in final form, because it misses an important point, namely that our correspondence maps:

$$\bar{z}\to T^*$$

However, this is something non-trivial, and we will be back to this later. Observe however that Theorem 15.20 is fully powerful for the self-adjoint operators, $T=T^*$, where the spectrum is real, so where $z=\bar{z}$ on the spectrum. We will be back to this.

As a second result now, along the same lines, we can further extend Theorem 15.20 into a measurable functional calculus theorem, as follows:

THEOREM 15.21. *Given $T\in B(H)$ normal, we have a morphism of algebras as follows, with L^∞ standing for abstract measurable functions, or Borel functions,*

$$L^\infty(\sigma(T))\to B(H)\quad,\quad f\to f(T)$$

which is isometric, $||f(T)||=||f||$, and has the property $\sigma(f(T))=f(\sigma(T))$.

PROOF. As before, the idea will be that of "completing" what we have. To be more precise, we can use the Riesz theorem and a polarization trick, as follows:

(1) Given a vector $x\in H$, consider the following functional:

$$C(\sigma(T))\to\mathbb{C}\quad,\quad g\to<g(T)x,x>$$

By the Riesz theorem, this functional must be the integration with respect to a certain measure μ on the space $\sigma(T)$. Thus, we have a formula as follows:

$$< g(T)x, x >= \int_{\sigma(T)} g(z)d\mu(z)$$

Now given an arbitrary Borel function $f \in L^\infty(\sigma(T))$, as in the statement, we can define a number $< f(T)x, x >\in \mathbb{C}$, by using exactly the same formula, namely:

$$< f(T)x, x >= \int_{\sigma(T)} f(z)d\mu(z)$$

Thus, we have managed to define numbers $< f(T)x, x >\in \mathbb{C}$, for all vectors $x \in H$, and in addition we can recover these numbers as follows, with $g_n \in C(\sigma(T))$:

$$< f(T)x, x >= \lim_{g_n \to f} < g_n(T)x, x >$$

(2) In order to define now numbers $< f(T)x, y >\in \mathbb{C}$, for all vectors $x, y \in H$, we can use a polarization trick. Indeed, for any operator $S \in B(H)$ we have:

$$\begin{aligned} < S(x+y), x+y > &= < Sx, x > + < Sy, y > \\ &\quad + < Sx, y > + < Sy, x > \end{aligned}$$

By replacing $y \to iy$, we have as well the following formula:

$$\begin{aligned} < S(x+iy), x+iy > &= < Sx, x > + < Sy, y > \\ &\quad -i < Sx, y > +i < Sy, x > \end{aligned}$$

By multiplying this latter formula by i, we obtain the following formula:

$$\begin{aligned} i < S(x+iy), x+iy > &= i < Sx, x > +i < Sy, y > \\ &\quad + < Sx, y > - < Sy, x > \end{aligned}$$

Now by summing this latter formula with the first one, we obtain:

$$\begin{aligned} < S(x+y), x+y > +i < S(x+iy), x+iy > &= (1+i)[< Sx, x > + < Sy, y >] \\ &\quad +2 < Sx, y > \end{aligned}$$

(3) But with this, we can now finish. Indeed, by combining (1,2), given a Borel function $f \in L^\infty(\sigma(T))$, we can define numbers $< f(T)x, y >\in \mathbb{C}$ for any $x, y \in H$, and it is routine to check, by using approximation by continuous functions $g_n \to f$ as in (1), that we obtain in this way an operator $f(T) \in B(H)$, having all the desired properties. □

As a comment here, the above result and its proof provide us with more than a Borel functional calculus, because what we got is a certain measure on the spectrum $\sigma(T)$, along with a functional calculus for the L^∞ functions with respect to this measure. We will be back to this later, and for the moment we will only need Theorem 15.21 as formulated, with $L^\infty(\sigma(T))$ standing, a bit abusively, for the Borel functions on $\sigma(T)$.

15d. Diagonalization

Let us discuss now some useful decomposition results for the bounded linear operators $T \in B(H)$, that we can now establish, by using the above measurable calculus technology. We know that any $z \in \mathbb{C}$ can be written as follows, with $a, b \in \mathbb{R}$:

$$z = a + ib$$

Also, we know that both the real and imaginary parts $a, b \in \mathbb{R}$, and more generally any real number $c \in \mathbb{R}$, can be written as follows, with $r, s \geq 0$:

$$c = r - s$$

In order to discuss now the operator theoretic generalizations of these results, which by the way covers the usual matrix case too, let us start with the following basic fact:

THEOREM 15.22. *Any operator $T \in B(H)$ can be written as*

$$T = Re(T) + iIm(T)$$

with $Re(T), Im(T) \in B(H)$ being self-adjoint, and this decomposition is unique.

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

(1) As a first observation, in the case $H = \mathbb{C}$ our operators are usual complex numbers, and the formula in the statement corresponds to the following basic fact:

$$z = Re(z) + iIm(z)$$

(2) In general now, we can use the same formulae for the real and imaginary part as in the complex number case, the decomposition formula being as follows:

$$T = \frac{T + T^*}{2} + i \cdot \frac{T - T^*}{2i}$$

To be more precise, both the operators on the right are self-adjoint, and the summing formula holds indeed, and so we have our decomposition result, as desired.

(3) Regarding now the uniqueness, by linearity it is enough to show that $R + iS = 0$ with R, S both self-adjoint implies $R = S = 0$. But this follows by applying the adjoint to $R + iS = 0$, which gives $R - iS = 0$, and so $R = S = 0$, as desired. □

More generally now, as a continuation of this, and as an answer to some of the questions raised above, in relation with the complex numbers, we have the following result:

THEOREM 15.23. *Given an operator $T \in B(H)$, the following happen:*

(1) *We can write $T = A + iB$, with $A, B \in B(H)$ being self-adjoint.*
(2) *When $T = T^*$, we can write $T = R - S$, with $R, S \in B(H)$ being positive.*
(3) *Thus, we can write any T as a linear combination of 4 positive elements.*

PROOF. All this follows from basic spectral theory, as follows:

(1) This is something that we already know, from Theorem 15.22, with the decomposition formula there being something straightforward, as follows:

$$T=\frac{T+T^*}{2}+i\cdot\frac{T-T^*}{2i}$$

(2) This follows from the measurable functional calculus. Indeed, assuming $T=T^*$ we have $\sigma(T)\subset\mathbb{R}$, so we can use the following decomposition formula on $\mathbb{R}$:

$$1=\chi_{[0,\infty)}+\chi_{(-\infty,0)}$$

To be more precise, let us multiply by z, and rewrite this formula as follows:

$$z=\chi_{[0,\infty)}z-\chi_{(-\infty,0)}(-z)$$

Now by applying these measurable functions to T, we obtain as formula as follows, with both the operators $T_+,T_-\in B(H)$ being positive, as desired:

$$T=T_+-T_-$$

(3) This follows indeed by combining the results in (1) and (2) above. □

Going ahead with our decomposition results, another basic thing that we know about complex numbers is that any $z\in\mathbb{C}$ appears as a real multiple of a unitary:

$$z=re^{it}$$

Finding the correct operator theoretic analogue of this is quite tricky, and this even for the usual matrices $A\in M_N(\mathbb{C})$. As a basic result here, we have:

THEOREM 15.24. *Given an operator $T\in B(H)$, the following happen:*

(1) *When $T=T^*$ and $||T||\leq 1$, we can write T as an average of 2 unitaries:*

$$T=\frac{U+V}{2}$$

(2) *In the general $T=T^*$ case, we can write T as a rescaled sum of unitaries:*

$$T=\lambda(U+V)$$

(3) *Thus, in general, we can write T as a rescaled sum of 4 unitaries.*

PROOF. This follows from the results that we have, as follows:

(1) Assuming $T=T^*$ and $||T||\leq 1$ we have $1-T^2\geq 0$, and the decomposition that we are looking for is as follows, with both the components being unitaries:

$$T=\frac{T+i\sqrt{1-T^2}}{2}+\frac{T-i\sqrt{1-T^2}}{2}$$

To be more precise, the square root can be extracted by using the continuous functional calculus, and the check of the unitarity of the components goes as follows:

$$\begin{aligned}(T+i\sqrt{1-T^2})(T-i\sqrt{1-T^2}) &= T^2+(1-T^2)\\ &= 1\end{aligned}$$

(2) This simply follows by applying (1) to the operator $T/||T||$.

(3) Assuming first that we have $||T|| \leq 1$, we know from Theorem 15.23 (1) that we can write $T = A + iB$, with A, B being self-adjoint, and satisfying $||A||, ||B|| \leq 1$. Now by applying (1) to both A and B, we obtain a decomposition of T as follows:

$$T = \frac{U+V+W+X}{2}$$

In general, we can apply this to the operator $T/||T||$, and we obtain the result. □

Good news, we can now diagonalize the normal operators. We will do this in 3 steps, first for the self-adjoint operators, then for the families of commuting self-adjoint operators, and finally for the general normal operators, by using the following trick:

$$T = Re(T) + iIm(T)$$

However, and coming somehow as bad news, all this will be quite technical. Indeed, the diagonalization in infinite dimensions is more tricky than in finite dimensions, and instead of writing a formula of type $T = UDU^*$, with $U, D \in B(H)$ being respectively unitary and diagonal, we will express our operator as $T = U^*MU$, with $U : H \to K$ being a certain unitary, and $M \in B(K)$ being a certain diagonal operator. The point indeed is that this is how the spectral theorem is used in practice, for concrete applications.

But probably too much talking, let us get to work. We first have:

THEOREM 15.25. *Any self-adjoint operator $T \in B(H)$ can be diagonalized,*

$$T = U^*M_fU$$

with $U : H \to L^2(X)$ being a unitary operator from H to a certain L^2 space associated to T, with $f : X \to \mathbb{R}$ being a certain function, once again associated to T, and with

$$M_f(g) = fg$$

being the usual multiplication operator by f, on the Hilbert space $L^2(X)$.

PROOF. The construction of U, f can be done in several steps, as follows:

(1) We first prove the result in the special case where our operator T has a cyclic vector $x \in H$, with this meaning that the following holds:

$$\overline{span\left(T^kx\middle|n\in\mathbb{N}\right)} = H$$

For this purpose, let us go back to the proof of Theorem 15.21. We will use the following formula from there, with μ being the measure on $X = \sigma(T)$ associated to x:

$$< g(T)x, x >= \int_{\sigma(T)} g(z)d\mu(z)$$

Our claim is that we can define a unitary $U : H \to L^2(X)$, first on the dense part spanned by the vectors $T^k x$, by the following formula, and then by continuity:

$$U[g(T)x] = g$$

Indeed, the following computation shows that U is well-defined, and isometric:

$$\begin{aligned} ||g(T)x||^2 &= < g(T)x, g(T)x > \\ &= < g(T)^* g(T)x, x > \\ &= < |g|^2(T)x, x > \\ &= \int_{\sigma(T)} |g(z)|^2 d\mu(z) \\ &= ||g||_2^2 \end{aligned}$$

We can then extend U by continuity into a unitary $U : H \to L^2(X)$, as claimed. Now observe that we have the following formula:

$$\begin{aligned} UTU^* g &= U[Tg(T)x] \\ &= U[(zg)(T)x] \\ &= zg \end{aligned}$$

Thus our result is proved in the present case, with U as above, and with $f(z) = z$.

(2) We discuss now the general case. Our first claim is that H has a decomposition as follows, with each H_i being invariant under T, and admitting a cyclic vector x_i:

$$H = \bigoplus_i H_i$$

Indeed, this is something elementary, the construction being by recurrence in finite dimensions, in the obvious way, and by using the Zorn lemma in general. Now with this decomposition in hand, we can make a direct sum of the diagonalizations obtained in (1), for each of the restrictions $T_{|H_i}$, and we obtain the formula in the statement. □

The above result is very nice, closing more or less the discussion regarding the self-adjoint operators. At the theoretical level, however, there are still a number of comments that can be made, about this, and we will be back to this, at the end of this chapter.

We have the following technical generalization of the above result:

THEOREM 15.26. *Any family of commuting self-adjoint operators $T_i \in B(H)$ can be jointly diagonalized,*

$$T_i = U^* M_{f_i} U$$

with $U : H \to L^2(X)$ being a unitary operator from H to a certain L^2 space associated to $\{T_i\}$, with $f_i : X \to \mathbb{R}$ being certain functions, once again associated to T_i, and with

$$M_{f_i}(g) = f_i g$$

being the usual multiplication operator by f_i, on the Hilbert space $L^2(X)$.

PROOF. This is similar to the proof of Theorem 15.25, by suitably modifying the measurable calculus formula, and μ itself, as to have this working for all operators T_i. □

We can now discuss the case of the arbitrary normal operators, as follows:

THEOREM 15.27. *Any normal operator $T \in B(H)$ can be diagonalized,*

$$T = U^* M_f U$$

with $U : H \to L^2(X)$ being a unitary operator from H to a certain L^2 space associated to T, with $f : X \to \mathbb{C}$ being a certain function, once again associated to T, and with

$$M_f(g) = fg$$

being the usual multiplication operator by f, on the Hilbert space $L^2(X)$.

PROOF. This is our main diagonalization theorem, the idea being as follows:

(1) Consider the decomposition of T into its real and imaginary parts, namely:

$$T = \frac{T + T^*}{2} + i \cdot \frac{T - T^*}{2i}$$

We know that the real and imaginary parts are self-adjoint operators. Now since T was assumed to be normal, $TT^* = T^*T$, these real and imaginary parts commute:

$$\left[\frac{T + T^*}{2} \, , \, \frac{T - T^*}{2i} \right] = 0$$

Thus Theorem 15.26 applies to these real and imaginary parts, and gives the result. □

This was for our series of diagonalization theorems. There is of course one more result here, regarding the families of commuting normal operators, as follows:

THEOREM 15.28. *Any family of commuting normal operators $T_i \in B(H)$ can be jointly diagonalized,*

$$T_i = U^* M_{f_i} U$$

with $U : H \to L^2(X)$ being a unitary operator from H to a certain L^2 space associated to $\{T_i\}$, with $f_i : X \to \mathbb{C}$ being certain functions, once again associated to T_i, and with

$$M_{f_i}(g) = f_i g$$

being the usual multiplication operator by f_i, on the Hilbert space $L^2(X)$.

PROOF. This is similar to the proof of Theorem 15.26 and Theorem 15.27, by combining the arguments there. To be more precise, this follows as Theorem 15.26, by using the decomposition trick from the proof of Theorem 15.27. □

With the above diagonalization results in hand, we can now "fix" the continuous and measurable functional calculus theorems, with a key complement, as follows:

THEOREM 15.29. *Given a normal operator $T \in B(H)$, the following hold, for both the functional calculus and the measurable calculus morphisms:*

(1) *These morphisms are $*$-morphisms.*
(2) *The function $\bar{z}$ gets mapped to T^*.*
(3) *The functions $Re(z), Im(z)$ get mapped to $Re(T), Im(T)$.*
(4) *The function $|z|^2$ gets mapped to $TT^* = T^*T$.*
(5) *If f is real, then $f(T)$ is self-adjoint.*

PROOF. These assertions are more or less equivalent, with (1) being the main one, which obviously implies everything else. But this assertion (1) follows from the diagonalization result for normal operators, from Theorem 15.27. □

15e. Exercises

Exercises:

EXERCISE 15.30.

EXERCISE 15.31.

EXERCISE 15.32.

EXERCISE 15.33.

EXERCISE 15.34.

EXERCISE 15.35.

EXERCISE 15.36.

EXERCISE 15.37.

Bonus exercise.

CHAPTER 16

Quantum mechanics

16a. Atomic theory

Welcome to quantum mechanics. As a starting point, we have 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:

FACT 16.1 (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.

Here the first part, the Rydberg formula, generalizes the results of Lyman, Balmer, Paschen, which appear at $n_1 = 1, 2, 3$, at least retrospectively. The Rydberg formula predicts further spectral lines, appearing at $n_1 = 4, 5, 6, \ldots$, and these were discovered later, by Brackett in 1922, Pfund in 1924, Humphreys in 1953, and others afterwards, 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
⋮	⋮	⋮	⋮	⋮

Regarding the last assertion, concerning other elements, this was something conjectured and partly verified by Ritz, and fully verified and clarified later, via many experiments, the fine-tuning of R being basically $R \to RZ^2$, where Z is the atomic number.

From a theoretical physics viewpoint, the main result remains the middle assertion, called Ritz-Rydberg combination principle. This is something at the same time extremely simple, and completely puzzling, the informal conclusion being as follows:

THOUGHT 16.2. *The simplest observables of the hydrogen atom, combining via*

$$\frac{1}{\lambda_{n_1n_2}} + \frac{1}{\lambda_{n_2n_3}} = \frac{1}{\lambda_{n_1n_3}}$$

look like quite weird quantities. Why wouldn't they just sum normally.

Fortunately, mathematics comes to the rescue. Indeed, the Ritz-Rydberg combination principle reminds the formula $e_{n_1n_2}e_{n_2n_3} = e_{n_1n_3}$ for the usual matrix units $e_{ij} : e_j \to e_i$. In short, we are in familiar territory here, and we can start dreaming of:

PRINCIPLE 16.3. *Observables in quantum mechanics should be some sort of infinite matrices, generalizing the Lyman, Balmer, Paschen lines of the hydrogen atom, and multiplying between them as the matrices do, as to produce further observables.*

In practice now, all this leads to the following grand conclusion:

CLAIM 16.4 (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, labeled* $1, 2, 3, \ldots$*, 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.*

And the story is not over here. Following now Heisenberg, the next claim is that the underlying mathematics in all the above can lead to a beautiful axiomatization of quantum mechanics, as a "matrix mechanics", along the lines of Principle 16.3.

All this is quite deep, and needs a number of comments, as follows:

(1) First of all, our matrices must be indeed infinite, because so are the series observed by Lyman, Balmer, Paschen, corresponding to $n_1 = 1, 2, 3$ in the Rydberg formula, and making it clear that the range of the second parameter $n_2 > n_1$ is up to ∞.

(2) Although this was not known to Ritz-Rydberg and Heisenberg, let us mention too that some later results of Brackett, Pfund, Humphreys and others, at $n_1 = 4, 5, 6, \ldots$, confirmed the fact that the range of the first parameter n_1 is up to ∞ too.

(3) As a more tricky comment now, going beyond what Principle 16.3 says, our infinite matrices must be in fact complex. This was something known to Heisenberg, and later Schrödinger came with proof that quantum mechanics naturally lives over $\mathbb{C}$.

(4) But all this leads us into some tricky mathematics, because the infinite matrices $A \in M_\infty(\mathbb{C})$ do not act on the vectors $v \in \mathbb{C}^\infty$ just like that. For instance the all-one matrix $A_{ij} = 1$ does not act on the all-one vector $v_i = 1$, for obvious reasons.

16b. Schrödinger equation

Time now to get into the real thing, namely quantum mechanics. However, before getting back to what Heisenberg was saying, based on Lyman, Balmer, Paschen, namely developing some sort of "matrix mechanics", let us hear as well the point of view of Schrödinger, which came a few years later. His idea was to forget about exact things, and try to investigate the hydrogen atom statistically. Let us start with:

QUESTION 16.5. *In the context of the hydrogen atom, assuming that the proton is fixed, what is the probability density $\varphi_t(x)$ of the position of the electron e, at time t,*

$$P_t(e \in V) = \int_V \varphi_t(x)dx$$

as function of an intial probability density $\varphi_0(x)$? Moreover, can the corresponding equation be solved, and will this prove the Bohr claims for hydrogen, statistically?

In order to get familiar with this question, let us first look at examples coming from classical mechanics. In the context of a particle whose position at time t is given by $x_0 + \gamma(t)$, the evolution of the probability density will be given by:

$$\varphi_t(x) = \varphi_0(x) + \gamma(t)$$

However, such examples are somewhat trivial, of course not in relation with the computation of γ, usually a difficult question, but in relation with our questions, and do not apply to the electron. The point indeed is that, in what regards the electron, we have:

FACT 16.6. *In respect with various simple interference experiments:*

(1) *The electron is definitely not a particle in the usual sense.*
(2) *But in most situations it behaves exactly like a wave.*
(3) *But in other situations it behaves like a particle.*

Getting back now to the Schrödinger question, all this suggests to use, as for the waves, an amplitude function $\psi_t(x) \in \mathbb{C}$, related to the density $\varphi_t(x) > 0$ by the formula $\varphi_t(x) = |\psi_t(x)|^2$. Not that a big deal, you would say, because the two are related by simple formulae as follows, with $\theta_t(x)$ being an arbitrary phase function:

$$\varphi_t(x) = |\psi_t(x)|^2 \quad , \quad \psi_t(x) = e^{i\theta_t(x)}\sqrt{\varphi_t(x)}$$

However, such manipulations can be crucial, raising for instance the possibility that the amplitude function satisfies some simple equation, while the density itself, maybe not. And this is what happens indeed. Schrödinger was led in this way to:

CLAIM 16.7 (Schrödinger). *In the context of the hydrogen atom, the amplitude function of the electron $\psi = \psi_t(x)$ is subject to the Schrödinger equation*

$$ih\dot{\psi} = -\frac{h^2}{2m}\Delta\psi + V\psi$$

m being the mass, $h = h_0/2\pi$ the reduced Planck constant, and V the Coulomb potential of the proton. The same holds for movements of the electron under any potential V.

Observe the similarity with the wave equation $\ddot{\varphi} = v^2\Delta\varphi$, and with the heat equation $\dot{\varphi} = \alpha\Delta\varphi$ too. Many things can be said here. Following now Heisenberg and Schrödinger, and then especially Dirac, who did the axiomatization work, we have:

DEFINITION 16.8. *In quantum mechanics the states of the system are vectors of a Hilbert space H, and the observables of the system are linear operators*

$$T : H \to H$$

which can be densely defined, and are taken self-adjoint, $T = T^*$. The average value of such an observable T, evaluated on a state $\xi \in H$, is given by:

$$< T >=< T\xi, \xi >$$

In the context of the Schrödinger mechanics of the hydrogen atom, the Hilbert space is the space $H = L^2(\mathbb{R}^3)$ where the wave function ψ lives, and we have

$$< T >= \int_{\mathbb{R}^3} T(\psi) \cdot \bar{\psi}\, dx$$

which is called "sandwiching" formula, with the operators

$$x \quad , \quad -\frac{ih}{m}\nabla \quad , \quad -ih\nabla \quad , \quad -\frac{h^2\Delta}{2m} \quad , \quad -\frac{h^2\Delta}{2m} + V$$

representing the position, speed, momentum, kinetic energy, and total energy.

In other words, we are doing here two things. First, we are declaring by axiom that various "sandwiching" formulae found before by Heisenberg, involving the operators at the end, that we will not get into in detail here, hold true. And second, we are raising

the possibility for other quantum mechanical systems, more complicated, to be described as well by the mathematics of the operators on a certain Hilbert space H, as above.

So, this was the story of early quantum mechanics, over-simplified as to fit here in a few pages. For more, you can check Feynman [35] for foundations, and everything, including for some nice pictures and explanations regarding Fact 16.6. You have as well Griffiths [45] or Weinberg [91], for further explanations on Definition 16.8, not to forget Dirac's original text [25], and all this is discussed as well in my book [12].

16c. Spherical coordinates

In order to solve now the hydrogen atom, the idea will be that of reformulating the Schrödinger equation in spherical coordinates. And for this purpose, we will need:

THEOREM 16.9. *The Laplace operator in spherical coordinates is*

$$\Delta = \frac{1}{r^2} \cdot \frac{d}{dr}\left(r^2 \cdot \frac{d}{dr}\right) + \frac{1}{r^2 \sin s} \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d}{ds}\right) + \frac{1}{r^2 \sin^2 s} \cdot \frac{d^2}{dt^2}$$

with our standard conventions for these coordinates, in 3D.

PROOF. There are several proofs here, a short, elementary one being as follows:

(1) Let us first see how Δ behaves under a change of coordinates $\{x_i\} \to \{y_i\}$, in arbitrary N dimensions. Our starting point is the chain rule for derivatives:

$$\frac{d}{dx_i} = \sum_j \frac{d}{dy_j} \cdot \frac{dy_j}{dx_i}$$

By using this rule, then Leibnitz for products, then again this rule, we obtain:

$$\begin{aligned}
\frac{d^2 f}{dx_i^2} &= \sum_j \frac{d}{dx_i}\left(\frac{df}{dy_j} \cdot \frac{dy_j}{dx_i}\right) \\
&= \sum_j \frac{d}{dx_i}\left(\frac{df}{dy_j}\right) \cdot \frac{dy_j}{dx_i} + \frac{df}{dy_j} \cdot \frac{d}{dx_i}\left(\frac{dy_j}{dx_i}\right) \\
&= \sum_j \left(\sum_k \frac{d}{dy_k} \cdot \frac{dy_k}{dx_i}\right)\left(\frac{df}{dy_j}\right) \cdot \frac{dy_j}{dx_i} + \frac{df}{dy_j} \cdot \frac{d^2 y_j}{dx_i^2} \\
&= \sum_{jk} \frac{d^2 f}{dy_k dy_j} \cdot \frac{dy_k}{dx_i} \cdot \frac{dy_j}{dx_i} + \sum_j \frac{df}{dy_j} \cdot \frac{d^2 y_j}{dx_i^2}
\end{aligned}$$

(2) Now by summing over i, we obtain the following formula, with A being the derivative of $x \to y$, that is to say, the matrix of partial derivatives dy_i/dx_j:

$$\begin{aligned}\Delta f &= \sum_{ijk} \frac{d^2 f}{dy_k dy_j} \cdot \frac{dy_k}{dx_i} \cdot \frac{dy_j}{dx_i} + \sum_{ij} \frac{df}{dy_j} \cdot \frac{d^2 y_j}{dx_i^2} \\ &= \sum_{ijk} A_{ki} A_{ji} \frac{d^2 f}{dy_k dy_j} + \sum_{ij} \frac{d^2 y_j}{dx_i^2} \cdot \frac{df}{dy_j} \\ &= \sum_{jk} (AA^t)_{jk} \frac{d^2 f}{dy_k dy_j} + \sum_j \Delta(y_j) \frac{df}{dy_j}\end{aligned}$$

(3) So, this will be the formula that we will need. Observe that this formula can be further compacted as follows, with all the notations being self-explanatory:

$$\Delta f = Tr(AA^t H_y(f)) + < \Delta(y), \nabla_y(f) >$$

(4) Getting now to spherical coordinates, $(x, y, z) \to (r, s, t)$, the derivative of the inverse, obtained by differentiating x, y, z with respect to r, s, t, is given by:

$$A^{-1} = \begin{pmatrix} \cos s & -r \sin s & 0 \\ \sin s \cos t & r \cos s \cos t & -r \sin s \sin t \\ \sin s \sin t & r \cos s \sin t & r \sin s \cos t \end{pmatrix}$$

The product $(A^{-1})^t A^{-1}$ of the transpose of this matrix with itself is then:

$$\begin{pmatrix} \cos s & \sin s \cos t & \sin s \sin t \\ -r \sin s & r \cos s \cos t & r \cos s \sin t \\ 0 & -r \sin s \sin t & r \sin s \cos t \end{pmatrix} \begin{pmatrix} \cos s & -r \sin s & 0 \\ \sin s \cos t & r \cos s \cos t & -r \sin s \sin t \\ \sin s \sin t & r \cos s \sin t & r \sin s \cos t \end{pmatrix}$$

But everything simplifies here, and we have the following remarkable formula, which by the way is something very useful, worth to be memorized:

$$(A^{-1})^t A^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2 s \end{pmatrix}$$

Now by inverting, we obtain the following formula, in relation with the above:

$$AA^t = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/r^2 & 0 \\ 0 & 0 & 1/(r^2 \sin^2 s) \end{pmatrix}$$

(5) Let us compute now the Laplacian of r, s, t. We first have the following formula, that we will use many times in what follows, and is worth to be memorized:

$$\begin{aligned}\frac{dr}{dx} &= \frac{d}{dx}\sqrt{x^2+y^2+z^2}\\ &= \frac{1}{2}\cdot\frac{2x}{\sqrt{x^2+y^2+z^2}}\\ &= \frac{x}{r}\end{aligned}$$

Of course the same computation works for y, z too, and we therefore have:

$$\frac{dr}{dx}=\frac{x}{r}\quad,\quad\frac{dr}{dy}=\frac{y}{r}\quad,\quad\frac{dr}{dz}=\frac{z}{r}$$

(6) By using the above formulae, twice, we can compute the Laplacian of r:

$$\begin{aligned}\Delta(r) &= \Delta\left(\sqrt{x^2+y^2+z^2}\right)\\ &= \frac{d}{dx}\left(\frac{x}{r}\right)+\frac{d}{dy}\left(\frac{y}{r}\right)+\frac{d}{dz}\left(\frac{z}{r}\right)\\ &= \frac{r^2-x^2}{r^3}+\frac{r^2-y^2}{r^3}+\frac{r^2-z^2}{r^3}\\ &= \frac{2}{r}\end{aligned}$$

(7) In what regards now s, the computation here goes as follows:

$$\begin{aligned}\Delta(s) &= \Delta\left(\arccos\left(\frac{x}{r}\right)\right)\\ &= \frac{d}{dx}\left(-\frac{\sqrt{r^2-x^2}}{r^2}\right)+\frac{d}{dy}\left(\frac{xy}{r^2\sqrt{r^2-x^2}}\right)+\frac{d}{dz}\left(\frac{xz}{r^2\sqrt{r^2-x^2}}\right)\\ &= \frac{2x\sqrt{r^2-x^2}}{r^4}+\frac{r^2(z^2-2y^2)+2x^2y^2}{r^4\sqrt{r^2-x^2}}+\frac{r^2(y^2-2z^2)+2x^2z^2}{r^4\sqrt{r^2-x^2}}\\ &= \frac{2x\sqrt{r^2-x^2}}{r^4}+\frac{x(2x^2-r^2)}{r^4\sqrt{r^2-x^2}}\\ &= \frac{x}{r^2\sqrt{r^2-x^2}}\\ &= \frac{\cos s}{r^2\sin s}\end{aligned}$$

(8) Finally, in what regards t, the computation here goes as follows:

$$\begin{aligned}\Delta(t) &= \Delta\left(\arctan\left(\frac{z}{y}\right)\right)\\ &= \frac{d}{dx}(0)+\frac{d}{dy}\left(-\frac{z}{y^2+z^2}\right)+\frac{d}{dz}\left(\frac{y}{y^2+z^2}\right)\\ &= 0-\frac{2yz}{(y^2+z^2)^2}+\frac{2yz}{(y^2+z^2)^2}\\ &= 0\end{aligned}$$

(9) We can now plug the data from (4) and (6,7,8) in the general formula that we found in (2) above, and we obtain in this way:

$$\begin{aligned}\Delta f &= \frac{d^2f}{dr^2}+\frac{1}{r^2}\cdot\frac{d^2f}{ds^2}+\frac{1}{r^2\sin^2 s}\cdot\frac{d^2f}{dt^2}+\frac{2}{r}\cdot\frac{df}{dr}+\frac{\cos s}{r^2\sin s}\cdot\frac{df}{ds}\\ &= \frac{2}{r}\cdot\frac{df}{dr}+\frac{d^2f}{dr^2}+\frac{\cos s}{r^2\sin s}\cdot\frac{df}{ds}+\frac{1}{r^2}\cdot\frac{d^2f}{ds^2}+\frac{1}{r^2\sin^2 s}\cdot\frac{d^2f}{dt^2}\\ &= \frac{1}{r^2}\cdot\frac{d}{dr}\left(r^2\cdot\frac{df}{dr}\right)+\frac{1}{r^2\sin s}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{df}{ds}\right)+\frac{1}{r^2\sin^2 s}\cdot\frac{d^2f}{dt^2}\end{aligned}$$

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

Still with me, I hope, and do not worry, one day you will have such computations for breakfast. We can now reformulate the Schrödinger equation in spherical coordinates, and separate the variables, which leads to a radial and angular equation, as follows:

THEOREM 16.10. *The time-independent Schrödinger equation in spherical coordinates separates, for solutions of type $\phi=\rho(r)\alpha(s,t)$, into two equations, as follows,*

$$\frac{d}{dr}\left(r^2\cdot\frac{d\rho}{dr}\right)-\frac{2mr^2}{h^2}(V-E)\rho=K\rho$$

$$\sin s\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{d^2\alpha}{dt^2}=-K\sin^2 s\cdot\alpha$$

with K being a constant, called radial equation, and angular equation.

PROOF. By using the formula in Theorem 16.9, the time-independent Schrödinger equation reformulates in spherical coordinates as follows:

$$(V-E)\phi=\frac{h^2}{2m}\left[\frac{1}{r^2}\cdot\frac{d}{dr}\left(r^2\cdot\frac{d\phi}{dr}\right)+\frac{1}{r^2\sin s}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\phi}{ds}\right)+\frac{1}{r^2\sin^2 s}\cdot\frac{d^2\phi}{dt^2}\right]$$

Let us look now for separable solutions for this latter equation, consisting of a radial part and an angular part, as in the statement, namely:

$$\phi(r,s,t)=\rho(r)\alpha(s,t)$$

By plugging this function into our equation, we obtain:

$$(V-E)\rho\alpha=\frac{h^2}{2m}\left[\frac{\alpha}{r^2}\cdot\frac{d}{dr}\left(r^2\cdot\frac{d\rho}{dr}\right)+\frac{\rho}{r^2\sin s}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{\rho}{r^2\sin^2 s}\cdot\frac{d^2\alpha}{dt^2}\right]$$

In order to solve this equation, we will do two manipulations. First, by multiplying everything by $2mr^2/(h^2\rho\alpha)$, this equation takes the following more convenient form:

$$\frac{2mr^2}{h^2}(V-E)=\frac{1}{\rho}\cdot\frac{d}{dr}\left(r^2\cdot\frac{d\rho}{dr}\right)+\frac{1}{\alpha\sin s}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{1}{\alpha\sin^2 s}\cdot\frac{d^2\alpha}{dt^2}$$

Now observe that by moving the radial terms to the left, and the angular terms to the right, this latter equation can be written as follows:

$$\frac{2mr^2}{h^2}(V-E)-\frac{1}{\rho}\cdot\frac{d}{dr}\left(r^2\cdot\frac{d\rho}{dr}\right)=\frac{1}{\alpha\sin^2 s}\left[\sin s\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{d^2\alpha}{dt^2}\right]$$

Since this latter equation is now separated between radial and angular variables, both sides must be equal to a certain constant $-K$, as follows:

$$\frac{2mr^2}{h^2}(V-E)-\frac{1}{\rho}\cdot\frac{d}{dr}\left(r^2\cdot\frac{d\rho}{dr}\right)=-K$$

$$\frac{1}{\alpha\sin^2 s}\left[\sin s\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{d^2\alpha}{dt^2}\right]=-K$$

But this leads to the conclusion in the statement. □

Let us first study the angular equation. We first have the following result:

PROPOSITION 16.11. *The angular equation that we found before, namely*

$$\sin s\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{d^2\alpha}{dt^2}=-K\sin^2 s\cdot\alpha$$

separates, for solutions of type $\alpha=\sigma(s)\theta(t)$, into two equations, as follows,

$$\frac{1}{\theta}\cdot\frac{d^2\theta}{dt^2}=-m^2$$

$$\frac{\sin s}{\sigma}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\sigma}{ds}\right)+K\sin^2 s=m^2$$

with m being a constant, called azimuthal equation, and polar equation.

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

(1) Let us first recall that $r\in[0,\infty)$ is the radius, $s\in[0,\pi]$ is the polar angle, and $t\in[0,2\pi]$ is the azimuthal angle. Be said in passing, there are several conventions and notations here, and the above ones, that we use here, come from the general ones in N dimensions, because further coordinates can be easily added, in the obvious way.

(2) Getting back now to our question, by plugging $\alpha = \sigma(s)\theta(t)$ into the angular equation, we obtain:

$$\sin s \cdot \theta \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\sigma}{ds}\right) + \sigma \cdot \frac{d^2\theta}{dt^2} = -K \sin^2 s \cdot \sigma\theta$$

By dividing everything by $\sigma\theta$, this equation can be written as follows:

$$-\frac{1}{\theta} \cdot \frac{d^2\theta}{dt^2} = \frac{\sin s}{\sigma} \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\sigma}{ds}\right) + K \sin^2 s$$

Since the variables are separated, we must have, for a certain constant m:

$$\frac{1}{\theta} \cdot \frac{d^2\theta}{dt^2} = -m^2$$

$$\frac{\sin s}{\sigma} \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\sigma}{ds}\right) + K \sin^2 s = m^2$$

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

Regarding the azimuthal equation, things here are quickly settled, as follows:

PROPOSITION 16.12. *The solutions of the azimuthal equation, namely*

$$\frac{1}{\theta} \cdot \frac{d^2\theta}{dt^2} = -m^2$$

are the functions as follows, with $a, b \in \mathbb{C}$ being parameters,

$$\theta(t) = ae^{imt} + be^{-imt}$$

and with only the case $m \in \mathbb{Z}$ being acceptable, on physical grounds.

PROOF. The first assertion is clear, because we have a second order equation, and two obvious solutions for it, $e^{\pm imt}$, and then their linear combinations, and that's all. Regarding the last assertion, the point here is that by using $\theta(t) = \theta(t + 2\pi)$, which is a natural physical assumption on the wave function, we are led to $m \in \mathbb{Z}$, as stated. □

We are now about to solve the angular equation, with only the polar equation remaining to be studied. However, in practice, this polar equation is 10 times more difficult that everything what we did so far, so be patient. We first have:

PROPOSITION 16.13. *The polar equation that we found before, namely*

$$\frac{\sin s}{\sigma} \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\sigma}{ds}\right) + K \sin^2 s = m^2$$

with $m \in \mathbb{Z}$, translates via $\sigma(s) = f(\cos s)$ into the following equation,

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

where $x = \cos s$, called Legendre equation.

PROOF. Let us first do a number of manipulations on our equation, before making the change of variables. By multiplying by σ, our equation becomes:

$$\sin s \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\sigma}{ds}\right) = \left(m^2 - K\sin^2 s\right)\sigma$$

By differentiating at left, this equation becomes:

$$\sin s\left(\cos s \cdot \sigma' + \sin s \cdot \sigma''\right) = \left(m^2 - K\sin^2 s\right)\sigma$$

Finally, by dividing everything by $\sin^2 s$, our equation becomes:

$$\sigma'' + \frac{\cos s}{\sin s} \cdot \sigma' = \left(\frac{m^2}{\sin^2 s} - K\right)\sigma$$

Now let us set $\sigma(s) = f(\cos s)$. With this change of variables, we have:

$$\sigma = f(\cos s)$$

$$\sigma' = -\sin s \cdot f'(\cos s)$$

$$\sigma'' = -\cos s \cdot f'(\cos s) + \sin^2 s \cdot f''(\cos s)$$

By plugging this data, our radial equation becomes:

$$\sin^2 s \cdot f''(\cos s) - 2\cos s \cdot f'(\cos s) = \left(\frac{m^2}{\sin^2 s} - K\right) f(\cos s)$$

Now with $x = \cos s$, which is our new variable, this equation reads:

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

But this is the Legendre equation, as stated. □

Here comes now the difficult point. We have the following non-trivial result:

THEOREM 16.14. *The solutions of the Legendre equation, namely*

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

can be explicitely computed, via complicated math, and only the case

$$K = l(l + 1) \quad : \quad l \in \mathbb{N}$$

is acceptable, on physical grounds.

PROOF. The first part is something quite complicated, involving the hypergeometric functions ${}_2F_1$, that you don't want to hear about, believe me. As for the second part, analysis and physics, this is something not trivial either. See Griffiths [45]. □

In order to construct the solutions, we will need:

THEOREM 16.15. *The orthonormal basis of $L^2[-1,1]$ obtained by starting with the Weierstrass basis $\{x^l\}$, and doing Gram-Schmidt, is the family of polynomials $\{P_l\}$, with each P_l being of degree l, and with positive leading coefficient, subject to:*

$$\int_{-1}^{1} P_k(x)P_l(x)\,dx = \delta_{kl}$$

These polynomials, called Legendre polynomials, satisfy the equation

$$(1-x^2)P_l''(x) - 2xP_l'(x) + l(l+1)P_l(x) = 0$$

which is the Legendre equation at $m = 0$, and with $K = l(l+1)$. Moreover,

$$P_l(x) = \frac{1}{2^l l!}\left(\frac{d}{dx}\right)^l (x^2-1)^l$$

which is called the Rodrigues formula for Legendre polynomials.

PROOF. As a first observation, we are not lost somewhere in abstract math, because of the occurrence of the Legendre equation. As for the proof, this goes as follows:

(1) The first assertion is clear, because the Gram-Schmidt procedure applied to the Weierstrass basis $\{x^l\}$ can only lead to a certain family of polynomials $\{P_l\}$, with each P_l being of degree l, and also unique, if we assume that it has positive leading coefficient, with this $\pm$ choice being needed, as usual, at each step of Gram-Schmidt.

(2) In order to have now an idea about these beasts, here are the first few of them, which can be obtained say via a straightforward application of Gram-Schmidt:

$$\begin{aligned}
P_0 &= 1\\
P_1 &= x\\
P_2 &= (3x^2-1)/2\\
P_3 &= (5x^3-3x)/2\\
P_4 &= (35x^4-30x^2+3)/8\\
P_5 &= (63x^5-70x^3+15x)/8
\end{aligned}$$

(3) Now thinking about what Gram-Schmidt does, this is certainly something by recurrence. And examining the recurrence leads to the Legendre equation, as stated.

(4) As for the Rodrigues formula, by uniqueness no need to try to understand where this formula comes from, and we have two choices here, either by verifying that $\{P_l\}$ is orthonormal, or by verifying the Legendre equation. And both methods work. □

Going ahead now, we can solve in fact the Legendre equation at any m, as follows:

PROPOSITION 16.16. *The general Legendre equation, with parameters $m \in \mathbb{N}$ and $K = l(l+1)$ with $l \in \mathbb{N}$, namely*

$$(1-x^2)f''(x) - 2xf'(x) = \left(\frac{m^2}{1-x^2} - l(l+1)\right) f(x)$$

is solved by the following functions, called Legendre functions,

$$P_l^m(x) = (-1)^m (1-x^2)^{m/2} \left(\frac{d}{dx}\right)^m P_l(x)$$

where P_l are as before the Legendre polynomials. Also, we have

$$P_l^m(x) = (-1)^m \frac{(1-x^2)^{m/2}}{2^l l!} \left(\frac{d}{dx}\right)^{l+m} (x^2-1)^l$$

called Rodrigues formula for Legendre functions.

PROOF. The first assertion is something elementary, coming by differentiating m times the Legendre equation, which leads to the general Legendre equation. As for the second assertion, this follows from the Rodrigues formula for Legendre polynomials. □

And this is the end of our study. Eventually. By putting together all the above results, we are led to the following conclusion:

THEOREM 16.17. *The separated solutions $\alpha = \sigma(s)\theta(t)$ of the angular equation,*

$$\sin s \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\alpha}{ds}\right) + \frac{d^2\alpha}{dt^2} = -K \sin^2 s \cdot \alpha$$

are given by the following formulae, where $l \in \mathbb{N}$ is such that $K = l(l+1)$,

$$\sigma(s) = P_l^m(\cos s) \quad , \quad \theta(t) = e^{imt}$$

and where $m \in \mathbb{Z}$ is a constant, and with P_l^m being the Legendre function,

$$P_l^m(x) = (-1)^m (1-x^2)^{m/2} \left(\frac{d}{dx}\right)^m P_l(x)$$

where P_l are the Legendre polynomials, given by the following formula:

$$P_l(x) = \frac{1}{2^l l!} \left(\frac{d}{dx}\right)^l (x^2-1)^l$$

These solutions $\alpha = \sigma(s)\theta(t)$ are called spherical harmonics.

PROOF. This follows indeed from all the above, and with the comment that everything is taken up to linear combinations. We will normalize the wave function later. □

16d. The hydrogen atom

Hydrogen, eventually. In order now to finish our study, and eventually get to conclusions about hydrogen, it remains to solve the radial equation, for the Coulomb potential V of the proton. Let us begin with some generalities, valid for any time-independent potential V. As a first manipulation on the radial equation, we have:

PROPOSITION 16.18. *The radial equation, written with $K = l(l+1)$,*

$$(r^2\rho')' - \frac{2mr^2}{h^2}(V - E)\rho = l(l+1)\rho$$

takes with $\rho = u/r$ the following form, called modified radial equation,

$$Eu = -\frac{h^2}{2m} \cdot u'' + \left(V + \frac{h^2 l(l+1)}{2mr^2}\right) u$$

which is a time-independent 1D Schrödinger equation.

PROOF. With $\rho = u/r$ as in the statement, we have:

$$\rho = \frac{u}{r} \quad , \quad \rho' = \frac{u'r - u}{r^2} \quad , \quad (r^2\rho')' = u''r$$

By plugging this data into the radial equation, this becomes:

$$u''r - \frac{2mr}{h^2}(V - E)u = \frac{l(l+1)}{r} \cdot u$$

By multiplying everything by $h^2/(2mr)$, this latter equation becomes:

$$\frac{h^2}{2m} \cdot u'' - (V - E)u = \frac{h^2 l(l+1)}{2mr^2} \cdot u$$

But this gives the formula in the statement. As for the interpretation, as time-independent 1D Schrödinger equation, this is clear as well, and with the comment here that the term added to the potential V is some sort of centrifugal term. □

Getting back now to the Coulomb potential of the proton, we have here:

FACT 16.19. *The Coulomb potential of the hydrogen atom proton, acting on the electron by attraction, is given according to the Coulomb law by*

$$V = -\frac{Kep}{r}$$

where p is the charge of the proton, and K is the Coulomb constant. In practice however we have $p \simeq e$ up to order 10^{-7}, and so our formula can be written as

$$V \simeq -\frac{Ke^2}{r}$$

and we will use this latter formula, and with $=$ sign, for simplifying.

Getting back now to math, it remains to solve the modified radial equation, for the above potential V. And we have here the following result, which does not exactly solve this radial equation, but provides us instead with something far better, namely the proof of the original claim by Bohr, which was at the origin of everything:

THEOREM 16.20 (Schrödinger). *In the case of the hydrogen atom, where V is the Coulomb potential of the proton, the modified radial equation, which reads*

$$Eu = -\frac{h^2}{2m} \cdot u'' + \left(-\frac{Ke^2}{r} + \frac{h^2 l(l+1)}{2mr^2} \right) u$$

leads to the Bohr formula for allowed energies,

$$E_n = -\frac{m}{2} \left(\frac{Ke^2}{h} \right)^2 \cdot \frac{1}{n^2}$$

with $n \in \mathbb{N}$, the binding energy being

$$E_1 \simeq -2.177 \times 10^{-18}$$

with means $E_1 \simeq -13.591$ eV.

PROOF. This is again something non-trivial, and we will be following Griffiths [45], with some details missing. The idea is as follows:

(1) By dividing our modified radial equation by E, this becomes:

$$-\frac{h^2}{2mE} \cdot u'' = \left(1 + \frac{Ke^2}{Er} - \frac{h^2 l(l+1)}{2mEr^2} \right) u$$

In terms of $\alpha = \sqrt{-2mE}/h$, this equation takes the following form:

$$\frac{u''}{\alpha^2} = \left(1 + \frac{Ke^2}{Er} + \frac{l(l+1)}{(\alpha r)^2} \right) u$$

In terms of the new variable $p = \alpha r$, this latter equation reads:

$$u'' = \left(1 + \frac{\alpha Ke^2}{Ep} + \frac{l(l+1)}{p^2} \right) u$$

Now let us introduce a new constant S for our problem, as follows:

$$S = -\frac{\alpha Ke^2}{E}$$

In terms of this new constant, our equation reads:

$$u'' = \left(1 - \frac{S}{p} + \frac{l(l+1)}{p^2} \right) u$$

(2) The idea will be that of looking for a solution written as a power series, but before that, we must "peel off" the asymptotic behavior. Which is something that can be done,

of course, heuristically. With $p \to \infty$ we are led to $u'' = u$, and ignoring the solution $u = e^p$ which blows up, our approximate asymptotic solution is:

$$u \sim e^{-p}$$

Similarly, with $p \to 0$ we are led to $u'' = l(l+1)u/p^2$, and ignoring the solution $u = p^{-l}$ which blows up, our approximate asymptotic solution is:

$$u \sim p^{l+1}$$

(3) The above heuristic considerations suggest writing our function u as follows:

$$u = p^{l+1}e^{-p}v$$

So, let us do this. In terms of v, we have the following formula:

$$u' = p^l e^{-p}\left[(l+1-p)v + pv'\right]$$

Differentiating a second time gives the following formula:

$$u'' = p^l e^{-p}\left[\left(\frac{l(l+1)}{p} - 2l - 2 + p\right)v + 2(l+1-p)v' + pv''\right]$$

Thus the radial equation, as modified in (1) above, reads:

$$pv'' + 2(l+1-p)v' + (S - 2(l+1))v = 0$$

(4) We will be looking for a solution v appearing as a power series:

$$v = \sum_{j=0}^{\infty} c_j p^j$$

But our equation leads to the following recurrence formula for the coefficients:

$$c_{j+1} = \frac{2(j+l+1) - S}{(j+1)(j+2l+2)} \cdot c_j$$

(5) We are in principle done, but we still must check that, with this choice for the coefficients c_j, our solution v, or rather our solution u, does not blow up. And the whole point is here. Indeed, at $j >> 0$ our recurrence formula reads, approximately:

$$c_{j+1} \simeq \frac{2c_j}{j}$$

But, surprisingly, this leads to $v \simeq c_0 e^{2p}$, and so to $u \simeq c_0 p^{l+1} e^p$, which blows up.

(6) As a conclusion, the only possibility for u not to blow up is that where the series defining v terminates at some point. Thus, we must have for a certain index j:

$$2(j+l+1) = S$$

In other words, we must have, for a certain integer $n > l$:

$$S = 2n$$

(7) We are almost there. Recall from (1) above that S was defined as follows:

$$S = -\frac{\alpha K e^2}{E} \quad : \quad \alpha = \frac{\sqrt{-2mE}}{h}$$

Thus, we have the following formula for the square of S:

$$S^2 = \frac{\alpha^2 K^2 e^4}{E^2} = -\frac{2mE}{h^2} \cdot \frac{K^2 e^4}{E^2} = -\frac{2mK^2 e^4}{h^2 E}$$

Now by using the formula $S = 2n$ from (6), the energy E must be of the form:

$$E = -\frac{2mK^2 e^4}{h^2 S^2} = -\frac{mK^2 e^4}{2h^2 n^2}$$

Calling this energy E_n, depending on $n \in \mathbb{N}$, we have, as claimed:

$$E_n = -\frac{m}{2}\left(\frac{Ke^2}{h}\right)^2 \cdot \frac{1}{n^2}$$

(8) Thus, we proved the Bohr formula. Regarding numerics, the data is as follows:

$$K = 8.988 \times 10^9 \quad , \quad e = 1.602 \times 10^{-19}$$

$$h = 1.055 \times 10^{-34} \quad , \quad m = 9.109 \times 10^{-31}$$

But this gives the formula of E_1 in the statement. □

As a first remark, all this agrees with the Rydberg formula, due to:

THEOREM 16.21. *The Rydberg constant for hydrogen is given by*

$$R = -\frac{E_1}{h_0 c}$$

where E_1 is the Bohr binding energy, and the Rydberg formula itself, namely

$$\frac{1}{\lambda_{n_1 n_2}} = R\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

simply reads, via the energy formula in Theorem 16.20,

$$\frac{1}{\lambda_{n_1 n_2}} = \frac{E_{n_2} - E_{n_1}}{h_0 c}$$

which is in agreement with the Planck formula $E = h_0 c/\lambda$.

PROOF. Here the first assertion is something numeric, coming from the fact that the formula in the statement gives, when evaluated, the Rydberg constant:

$$R = \frac{-E_1}{h_0 c} = \frac{2.177 \times 10^{-18}}{6.626 \times 10^{-34} \times 2.998 \times 10^8} = 1.096 \times 10^7$$

As a consequence, and passed now what the experiments exactly say, we can define the Rydberg constant of hydrogen abstractly, by the following formula:

$$R = \frac{m}{2h_0c}\left(\frac{Ke^2}{h}\right)^2$$

Regarding now the second assertion, by dividing $R = -E_1/(h_0c)$ by any number of type n^2 we obtain, according to the energy convention in Theorem 16.20:

$$\frac{R}{n^2} = -\frac{E_n}{h_0c}$$

But these are exactly the numbers which are subject to substraction in the Rydberg formula, and so we are led to the conclusion in the statement. □

Let us go back now to our study of the Schrödinger equation. Our conclusions are:

THEOREM 16.22. *The wave functions of the hydrogen atom are the following functions, labelled by three quantum numbers, n, l, m,*

$$\phi_{nlm}(r, s, t) = \rho_{nl}(r)\alpha_l^m(s, t)$$

where $\rho_{nl}(r) = p^{l+1}e^{-p}v(p)/r$ with $p = \alpha r$ as before, with the coefficients of v subject to

$$c_{j+1} = \frac{2(j + l + 1 - n)}{(j + 1)(j + 2l + 2)} \cdot c_j$$

and $\alpha_l^m(s, t)$ being the spherical harmonics found before.

PROOF. This follows indeed by putting together all the results obtained so far, and with the remark that everything is up to the normalization of the wave function. □

In what regards the main wave function, that of the ground state, we have:

THEOREM 16.23. *With the hydrogen atom in its ground state, the wave function is*

$$\phi_{100}(r, s, t) = \frac{1}{\sqrt{\pi a^3}}\, e^{-r/a}$$

where $a = 1/\alpha$ is the inverse of the parameter appearing in our computations above,

$$\alpha = \frac{\sqrt{-2mE}}{h}$$

called Bohr radius of the hydrogen atom. This Bohr radius is the mean distance between the electron and the proton, in the ground state, and is given by the formula

$$a = \frac{h^2}{mKe^2}$$

which numerically means $a \simeq 5.291 \times 10^{-11}$.

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

(1) According to the various formulae in the proof of Theorem 16.20, taken at $n = 1$, the parameter α appearing in the computations there is given by:

$$\alpha = \frac{\sqrt{-2mE}}{h} = \frac{1}{h} \cdot m \cdot \frac{Ke^2}{h} = \frac{mKe^2}{h^2}$$

Thus, the inverse $\alpha = 1/a$ is indeed given by the formula in the statement.

(2) Regarding the wave function, according to Theorem 16.22 this consists of:

$$\rho_{10}(r) = \frac{2e^{-r/a}}{\sqrt{a^3}} \quad , \quad \alpha_0^0(s,t) = \frac{1}{2\sqrt{\pi}}$$

By making the product, we obtain the formula of ϕ_{100} in the statement.

(3) But this formula of ϕ_{100} shows in particular that the Bohr radius a is indeed the mean distance between the electron and the proton, in the ground state.

(4) Finally, in what regards the numerics, these are as follows:

$$a = \frac{1.055^2 \times 10^{-68}}{9.109 \times 10^{-31} \times 8.988 \times 10^9 \times 1.602^2 \times 10^{-38}} = 5.297 \times 10^{-11}$$

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

Getting back now to the general setting of Theorem 16.20, the point is that the polynomials $v(p)$ appearing there are well-known objects in mathematics, as follows:

PROPOSITION 16.24. *The polynomials $v(p)$ are given by the formula*

$$v(p) = L_{n-l-1}^{2l+1}(p)$$

where the polynomials on the right, called associated Laguerre polynomials, are given by

$$L_q^p(x) = (-1)^p \left(\frac{d}{dx}\right)^p L_{p+q}(x)$$

with L_{p+q} being the Laguerre polynomials, given by the following formula:

$$L_q(x) = \frac{e^x}{q!} \left(\frac{d}{dx}\right)^q (e^{-x}x^q)$$

PROOF. The story here is very similar to that of the Legendre polynomials. Consider the Hilbert space $H = L^2[0,\infty)$, with the following scalar product on it:

$$< f, g >= \int_0^\infty f(x)g(x)e^{-x}\,dx$$

(1) The orthogonal basis obtained by applying Gram-Schmidt to the Weierstrass basis $\{x^q\}$ is then the basis formed by the Laguerre polynomials $\{L_q\}$.

(2) We have the explicit formula for L_q in the statement, which is analogous to the Rodrigues formula for the Legendre polynomials.

(3) The first assertion follows from the fact that the coefficients of the associated Laguerre polynomials satisfy the equation for the coefficients of $v(p)$.

(4) Alternatively, the first assertion follows as well by using an equation for the Laguerre polynomials, which is very similar to the Legendre equation. □

With the above result in hand, we can now improve Theorem 16.20, as follows:

THEOREM 16.25. *The wave functions of the hydrogen atom are given by*

$$\phi_{nlm}(r,s,t)=\sqrt{\left(\frac{2}{na}\right)^3\frac{(n-l-1)!}{2n(n+l)!}}e^{-r/na}\left(\frac{2r}{na}\right)^l L_{n-l-1}^{2l+1}\left(\frac{2r}{na}\right)\alpha_l^m(s,t)$$

with $\alpha_l^m(s,t)$ being the spherical harmonics found before.

PROOF. This follows indeed by putting together what we have, namely Theorem 16.20 and Proposition 16.24, and then doing some remaining work, concerning the normalization of the wave function, which leads to the normalization factor appearing above. □

And good news, that is all. The above formula is all you need, in everyday life.

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] A.A. Abrikosov, L.P. Gorkov and I.E. Dzyaloshinski, Methods of quantum field theory in statistical physics, Dover (1963).

[3] G.W. Anderson, A. Guionnet and O. Zeitouni, An introduction to random matrices, Cambridge Univ. Press (2010).

[4] V.I. Arnold, Ordinary differential equations, Springer (1973).

[5] V.I. Arnold, Mathematical methods of classical mechanics, Springer (1974).

[6] V.I. Arnold, Catastrophe theory, Springer (1974).

[7] V.I. Arnold, Lectures on partial differential equations, Springer (1997).

[8] V.I. Arnold and B.A. Khesin, Topological methods in hydrodynamics, Springer (1998).

[9] N.W. Ashcroft and N.D. Mermin, Solid state physics, Saunders College Publ. (1976).

[10] M.F. Atiyah, The geometry and physics of knots, Cambridge Univ. Press (1990).

[11] T. Banica, Calculus and applications (2024).

[12] T. Banica, Introduction to modern physics (2025).

[13] T. Banica, The joys of relativity theory (2024).

[14] G.K. Batchelor, An introduction to fluid dynamics, Cambridge Univ. Press (1967).

[15] R.J. Baxter, Exactly solved models in statistical mechanics, Academic Press (1982).

[16] I. Bengtsson and K. Życzkowski, Geometry of quantum states, Cambridge Univ. Press (2006).

[17] S.M. Carroll, Spacetime and geometry, Cambridge Univ. Press (2004).

[18] P.M. Chaikin and T.C. Lubensky, Principles of condensed matter physics, Cambridge Univ. Press (1995).

[19] A.R. Choudhuri, Astrophysics for physicists, Cambridge Univ. Press (2012).

[20] D.D. Clayton, Principles of stellar evolution and nucleosynthesis, Univ. of Chicago Press (1968).

[21] A. Connes, Noncommutative geometry, Academic Press (1994).

[22] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).

[23] P.A. Davidson, Introduction to magnetohydrodynamics, Cambridge Univ. Press (2001).

[24] P. Di Francesco, P. Mathieu and D. Sénéchal, Conformal field theory, Springer (1996).

[25] P.A.M. Dirac, Principles of quantum mechanics, Oxford Univ. Press (1930).
[26] S. Dodelson, Modern cosmology, Academic Press (2003).
[27] R. Durrett, Probability: theory and examples, Cambridge Univ. Press (1990).
[28] A. Einstein, Relativity: the special and the general theory, Dover (1916).
[29] L.C. Evans, Partial differential equations, AMS (1998).
[30] L.D. Faddeev and L. A. Takhtajan, Hamiltonian methods in the theory of solitons, Springer (2007).
[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] R.P. Feynman, Statistical mechanics: a set of lectures, CRC Press (1988).
[38] A.P. French, Special relativity, Taylor and Francis (1968).
[39] N. Goldenfeld, Lectures on phase transitions and the renormalization group, CRC Press (1992).
[40] H. Goldstein, C. Safko and J. Poole, Classical mechanics, Addison-Wesley (1980).
[41] D.L. Goodstein, States of matter, Dover (1975).
[42] M.B. Green, J.H. Schwarz and E. Witten, Superstring theory, Cambridge Univ. Press (2012).
[43] W. Greiner, L. Neise and H. Stöcker, Thermodynamics and statistical mechanics, Springer (2012).
[44] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).
[45] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).
[46] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).
[47] D.J. Griffiths, Revolutions in twentieth-century physics, Cambridge Univ. Press (2012).
[48] W.A. Harrison, Solid state theory, Dover (1970).
[49] W.A. Harrison, Electronic structure and the properties of solids, Dover (1980).
[50] K. Huang, Introduction to statistical physics, CRC Press (2001).
[51] K. Huang, Quantum field theory, Wiley (1998).
[52] K. Huang, Quarks, leptons and gauge fields, World Scientific (1982).
[53] K. Huang, Fundamental forces of nature, World Scientific (2007).
[54] C. Itzykson and J.B. Zuber, Quantum field theory, Dover (1980).
[55] J.D. Jackson, Classical electrodynamics, Wiley (1962).

[56] V.F.R. Jones, Subfactors and knots, AMS (1991).

[57] L.P. Kadanoff, Statistical physics: statics, dynamics and renormalization, World Scientific (2000).

[58] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).

[59] C. Kittel, Introduction to solid state physics, Wiley (1953).

[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] V.B. Berestetskii, E.M. Lifshitz and L.P. Pitaevskii, Quantum electrodynamics, Butterworth-Heinemann (1982).

[66] M.L. Mehta, Random matrices, Elsevier (2004).

[67] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge Univ. Press (2000).

[68] R.K. Pathria and and P.D. Beale, Statistical mechanics, Elsevier (1972).

[69] A. Peres, Quantum theory: concepts and methods, Kluwer (1993).

[70] M. Peskin and D.V. Schroeder, An introduction to quantum field theory, CRC Press (1995).

[71] B.M. Peterson and B. Ryden, Foundations of astrophysics, Cambridge Univ. Press (2010).

[72] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[73] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[74] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[75] B. Ryden and R.W. Pogge, Interstellar and intergalactic medium, Cambridge Univ. Press (2021).

[76] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).

[77] J. Schwinger, Einstein's legacy: the unity of space and time, Dover (1986).

[78] J. Schwinger, L.L. DeRaad Jr., K.A. Milton and W.Y. Tsai, Classical electrodynamics, CRC Press (1998).

[79] J. Schwinger and B.H. Englert, Quantum mechanics: symbolism of atomic measurements, Springer (2001).

[80] R. Shankar, Fundamentals of physics I: mechanics, relativity, and thermodynamics, Yale Univ. Press (2014).

[81] R. Shankar, Fundamentals of physics II: electromagnetism, optics, and quantum mechanics, Yale Univ. Press (2016).

[82] R. Shankar, Principles of quantum mechanics, Springer (1980).

[83] R. Shankar, Quantum field theory and condensed matter: an introduction, Cambridge Univ. Press (2017).

[84] J. Smit, Introduction to quantum fields on a lattice, Cambridge Univ. Press (2002).

[85] A.M. Steane, Thermodynamics, Oxford Univ. Press (2016).

[86] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[87] D.V. Voiculescu, K.J. Dykema and A. Nica, Free random variables, AMS (1992).

[88] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[89] J. von Neumann and O. Morgenstern, Theory of games and economic behavior, Princeton Univ. Press (1944).

[90] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[91] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[92] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).

[93] S. Weinberg, Cosmology, Oxford Univ. Press (2008).

[94] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).

[95] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).

[96] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).

[97] J.M. Yeomans, Statistical mechanics of phase transitions, Oxford Univ. Press (1992).

[98] J. Zinn-Justin, Path integrals in quantum mechanics, Oxford Univ. Press (2004).

[99] J. Zinn-Justin, Phase transitions and renormalization group, Oxford Univ. Press (2005).

[100] B. Zwiebach, A first course in string theory, Cambridge Univ. Press (2004).

Index