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 mathematics and physics, guided by the dimensionality of the ambient space, and with the mathematics and physics chapters alternating. We start with a discussion on what happens in 1 dimension, namely the basics of calculus, and some physics too, including 1D collisions, relativity, gravity, waves and heat. Then we go on a similar discussion in 2D, featuring this time plane vectors, complex numbers, and more advanced physics, such as elliptic orbits under gravity. We then discuss what happens in 3 dimensions, namely standard vector calculus, and 3D classical mechanics, and the basics of electromagnetism. Finally, we discuss 4D and curved spacetime, and we end with a brief introduction to quantum mechanics.

Preface

How many dimensions do we live in? Three I guess, hope we agree on this, and please be reassured, we won't be here with this book for contradicting this basic fact. However, and here comes my point, for learning mathematics and physics, as a student, or for fine-tuning your knowledge in mathematics and physics, as an amateur or a professional, a look at the whole spectrum of possible dimensions, $N = 0, 1, 2, 3, 4, \ldots, \infty$, not to talk about some more exotic dimensions that can be invented too, such as fractionary, or mystical, can be something instructive. Here is the situation, and judge yourself:

$\underline{N = 0}$. This is something a bit philosophical, with some mathematics going on here, which can be quite interesting, but remains something quite specialized, and with no physics, at least in some reasonable sense, going on. So, we should better skip this.

$\underline{N = 1}$. This is something interesting, because most of the basic mathematics that must be learned, namely real numbers, sequences, series, and then functions and basic calculus, takes place here. As for the 1D physics, this can be serious business too.

$\underline{N = 2}$. This is definitely interesting, because many of the 3-dimensional questions that we would like to solve, be them in mathematics or physics, live in fact in a suitable plane, and so, in 2D. Many things, both mathematics and physics, to be learned here.

$\underline{N = 3}$. This needs no presentation I guess, and with respect to the previous learning from lower dimensions, one key new thing in mathematics is the vector product $x \times y$. As for physics, of genuine 3D nature are thermodynamics and electromagnetism.

$\underline{N = 4}$. This is something more modern, coming from Einstein, who discovered not that long ago that space $\mathbb{R}^3$ and time $\mathbb{R}$ are not independent, as previously thought. So, the true space that we live in is in fact a curved version of $\mathbb{R}^3 \times \mathbb{R} = \mathbb{R}^4$.

$\underline{N \geq 5}$. This is something even more modern, with $N = \infty$ coming from quantum mechanics, as developed by Heisenberg and others, and with $4 < N < \infty$ being more recent physics business, typically with the bigger the $N < \infty$, the fancier the theory.

So, this was for the quick story of dimensionality in mathematics and physics. Might sound a bit like psychedelia, and yes I confirm, artists are not the only ones not knowing in how many dimensions they live in, we mathematicians and physicists do the same.

This book will be an introduction to this, basic mathematics and physics, or theory of forces and vectors, if you prefer, guided by the dimensionality of the ambient space. We won't particularly favor mathematics over physics, or vice versa, the book being written $50-50$, with the mathematics and physics chapters alternating, the odd-numbered ones $1-3-\ldots-15$ being mathematics, and the even-numbered ones $2-4-\ldots-16$ being physics. The organization is in 4 parts, following a $N=1,2,3,4$ scheme, as follows:

(1) We start with a discussion on what happens in 1 dimension, namely the basics of calculus, and 1D physics, regarding collisions, relativity, gravity, waves and heat.

(2) Then we go on a similar discussion in 2D, featuring this time plane vectors, complex numbers, and more advanced physics, such as elliptic orbits under gravity.

(3) We then discuss what happens in 3 dimensions, namely standard vector calculus, followed by 3D classical mechanics, and the basics of electromagnetism.

(4) Finally, motivated by the findings of Einstein, we discuss 4D and curved spacetime, and we end with a brief introduction to quantum mechanics.

In the hope that you will find this book useful, whether here for learning mathematics and physics, or for fine-tuning some previous knowledge. Personally, I learned many interesting things when writing it, with the organization of the book, which is something quite original, forcing me to rethink all sorts of things that I previously knew.

This book will be, of course, just an introduction to the subject. As a continuation, for more mathematics you have my book [11], for more physics and quantum mechanics you have my book [12], and for more relativity theory you have my book [13].

As mentioned, the preparation of this book, according to the above guidelines, was a quite challenging task. And many thanks here to my cats, for some help with this, projecting from meow to various $N\leq\infty$ values being a quite easy matter, for them.

Cergy, July 2026
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

Welcome to mathematics and physics, and whether you know none or some, this will be an enjoyable trip, and we will learn interesting things together. The plan will be to investigate mathematics and physics in 1D in the present Part I, then switch to 2D, which is more complicated, in Part II, then to 3D in Part III, and then to 4D in Part IV.

Which looks like something quite mathematical, so as our first theorem I guess, save for the conversion between Arabic and Roman numerals, which is the nightmare of humanity for centuries and counting, we have the following equation, for this book:

$$\text{Part X} = \text{XD science}$$

This being said, you might probably wonder, with us starting now Part I, are there enough things to be learned there? I mean, we certainly live in 3D, or perhaps 4D according to Einstein, so why bothering with 1D, we should learn, as a quick warm-up, some 2D stuff, and then go with the 3D, or perhaps 4D, that we are interested in.

In answer, not so quick, there are many interesting things to be learned in 1D, trust me here. In fact, there are so many interesting things going on there, both math and physics, that I won't be surprised, if one day we face an advanced alien 1D civilization, that these guys will be more advanced than us, and we will have to ask for mercy.

By the way, talking aliens living in 1D, no need to go to space for meeting them, because here on Earth we have ants. At my house at least, plenty of them from time to time, coming doing their business, in 1D formation, walking along the walls. And having some sort of status of apex predators, the other animals dwelling at the house, namely cats, rats, spiders and myself, being quite peaceful, and not interested in them.

So, let's see what one of these ants has to say, about living in 1D:

Ant 1.1. *There is life in 1D, and we are quite advanced at math and physics. And one day, after WW3 and the nuclear fallout, we will grow 100 times bigger.*

Okay, thanks ant, so my intuition was right, certainly worth spending the first 100 pages in this book in understanding 1D phenomena, both math and physics.

Getting started now, what exactly is 1D? Well, 1D is the real line $\mathbb{R}$, and in order to get familiar with it, we first have to talk about usual numbers, $n \in \mathbb{N}$, then about integers, $m \in \mathbb{Z}$, then about fractions, $r \in \mathbb{Q}$, and finally about reals, $x \in \mathbb{R}$.

In what regards the usual numbers $n \in \mathbb{N}$, you are certainly familiar with them, and notably with their summing operation $+$, and product operation $\times$. So, I won't bother you much with this, but as something that you might have noticed or not, never too late for learning new things, here is an interesting observation about these numbers:

FACT 1.2. *There is no need for digits* $0, 1, \ldots, 9$ *in order to talk about the numbers* $n \in \mathbb{N}$*, because these can be designated by repeating any symbol of your choice:*

$$5 = \heartsuit\heartsuit\heartsuit\heartsuit\heartsuit$$

However, the digits $a, b, c, \ldots$ *and the usual notation* $n = abc\ldots$ *for the numbers* $n \in \mathbb{N}$ *are useful when dealing with large numbers, as shown for instance by*

$$47 = \begin{matrix} \heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit \\ \heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit \\ \heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit \\ \heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit \end{matrix}$$

and for doing arithmetic with these numbers, via the well-known algorithms for $+$ *and* $\times$*. Also, the choice of the numeration basis* 10 *is something human, and modern.*

To be more precise, and I insist here, there are things that you certainly know from school, save however for the last assertion, regarding the numeration basis 10, which is something traditionally not taught, and what a pity, because many interesting things can be said here. So, exercise for you to learn more about such things, numeration bases.

Getting now to the arbitrary integers, $m \in \mathbb{Z}$, again these are beasts that you know well from school, and I won't bother you with this. Let us just record, about them:

FACT 1.3. *The arbitrary integers* $m \in \mathbb{Z}$ *are there as for the following equation, with* $a, b \in \mathbb{N}$*, to always have solutions, over the integers:*

$$a + x = b$$

In practice, the integers fall into two classes, namely positive, $m \in \mathbb{N}$*, and negative,* $m = -n$ *with* $n \in \mathbb{N}$*. Also,* $\mathbb{Z}$ *comes from "zahlen", which is German for "numbers".*

As before, these are things that you know well from school, save perhaps for the last assertion, regarding the notation $\mathbb{Z}$. And beware of Germans I guess, these guys call abstract beasts like -15 numbers. By the way, our previous $\mathbb{N}$ comes from "natural".

Time now to upgrade to fractions, or quotients, or, in fancier mathematical parlance, to rational numbers? You surely know about these too, and with these beasts being something more subtle, worth a mathematical Definition I guess, as follows:

DEFINITION 1.4. *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 an inclusion $\mathbb{Z} \subset \mathbb{Q}$, given by $m = m/1$. The rational numbers add and substract according to the usual rules for quotients, which are as follows:

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

Also, they multiply and divide according to the usual rules for quotients, namely:

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

Finally, in analogy with Fact 1.3, observe that the rational numbers are there as for the following equation, with $a, b \in \mathbb{N}$, to have solutions, over the rationals:

$$ax = b$$

Which is something quite interesting, so as our first theorem, let us formulate:

THEOREM 1.5. *The passage $\mathbb{N} \to \mathbb{Q}$ appears as for the equations*

$$a + x = b \quad , \quad ax = b$$

with $a, b \in \mathbb{N}$ to have solutions, over the rationals.

PROOF. This is indeed obvious, and of course with $a \neq 0$ for the products. However, this theorem can be useful for instance for bragging about mathematics, because with the convention that a mathematical field is a set where $a + x = b$ and $ax = b$ have solutions, what we just proved is that "$\mathbb{Q}$ is the simplest field containing $\mathbb{N}$". Nice. □

As our second theorem now, which is something of true usefulness, we have:

THEOREM 1.6. *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. This is something more subtle, the idea being as follows:

(1) 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:

$$N = n(n-1)\ldots(n-k+1) = \frac{n!}{(n-k)!}$$

(2) However, thinking a bit, this 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:

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

(3) 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, as claimed:

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

(4) Finally, as a complement to what is said in the statement, let us mention that we have to make the convention $0! = 1$, as for the following computation to work:

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

And with this, we have now everything that we need to know, on this subject. □

As our third theorem now, further building on Theorem 1.6, we have:

THEOREM 1.7. *We have the binomial formula*

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

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

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

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

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

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

In order now to finish, it remains to compute the above coefficients C_k. But, according to our initial 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 1.6, we have:

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

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

Theorem 1.7 is something quite interesting, so let us doublecheck it with some numerics. At small values of n we obtain the following formulae, which are all correct:

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

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

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

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

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

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

$$(a+b)^6 = a^6+6a^5b+15a^4b^2+20a^3b^3+15a^2b^4+6ab^5+b^6$$

$$\vdots$$

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

$$1$$

$$1 \quad , \quad 1$$

$$1 \quad , \quad 2 \quad , \quad 1$$

$$1 \quad , \quad 3 \quad , \quad 3 \quad , \quad 1$$

$$1 \quad , \quad 4 \quad , \quad 6 \quad , \quad 4 \quad , \quad 1$$

$$1 \quad , \quad 5 \quad , \quad 10 \quad , \quad 10 \quad , \quad 5 \quad , \quad 1$$

$$1 \quad , \quad 6 \quad , \quad 15 \quad , \quad 20 \quad , \quad 15 \quad , \quad 6 \quad , \quad 1$$

$$\vdots$$

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

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

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

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

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

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

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

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

By using the binomial formula, this formula becomes:

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

Now let us perform the multiplication on the right. We obtain a certain sum of terms of type $a^k b^{n-k}$, and to be more precise, each such $a^k b^{n-k}$ 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 $a^k b^{n-1-k}$ multiplied by b. Thus, the coefficient of $a^k b^{n-k}$ 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. □

1b. Real numbers

Getting now to where we wanted to get, the real numbers $x \in \mathbb{R}$, you are certainly familiar with them, but after all, what exactly are these beasts?

Which is not an easy question, imagine for instance our Stone Age ancestors, these gentlemen certainly knew about integers and fractions, but I bet that they didn't bother

with reals. In short, the real numbers are a relatively recent discovery. Mathematically speaking, we have 3 possible definitions for them, all interesting, as follows:

THEOREM 1.9. *The real numbers $x \in \mathbb{R}$ appear in 3 possible ways, as follows:*

(1) *As cuts in the set of rationals, $\mathbb{Q} = A_x \sqcup B_x$, with $p \in A_x, q \in B_x \implies p < q$.*
(2) *As limits of rational numbers, $x = \lim_{n\to\infty} r_n$ with $r_n \in \mathbb{Q}$, $r_n - r_m \to 0$.*
(3) *As decimal expressions, $x = \pm a_1 \ldots a_n.b_1b_2b_3 \ldots$, with $a_i, b_i \in \{0, \ldots, 9\}$*

PROOF. This is something not exactly trivial, the idea being as follows:

(1) In what regards the first definition of the reals, as cuts $\mathbb{Q} = A_x \sqcup B_x$, this is something quite intuitive, with $\sqrt{2}$ for instance coming from the following cut:

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

Be said in passing, observe that what we say in the statement must be complemented with the following condition, in order to avoid some troubles when $x \in \mathbb{Q}$:

$$\inf B_x \notin B_x$$

In any case, what we have here is a viable approach to the reals, and as an interesting feature of this approach, the addition and multiplication of reals is something quite simple, obtained by adding and multiplying the corresponding cuts, in the obvious way. Up to some mess with the positives and negatives, which is quite easy to untangle.

(2) This is again something quite intuitive, up to some clarifications regarding the precise meaning of what we say in the statement, and more on this in a moment. On the upside, when compared to (1), the addition and multiplication become a bit simpler:

$$\lim_{n\to\infty} r_n + \lim_{n\to\infty} p_n = \lim_{n\to\infty} r_n + p_n$$

$$\lim_{n\to\infty} r_n \times \lim_{n\to\infty} p_n = \lim_{n\to\infty} r_n p_n$$

However, on the downside, we have for instance the following question:

$$\sqrt{2} = ?$$

And by this I mean, is it really obvious that $\sqrt{2}$ is a limit of rationals? Not really.

(3) This is something that you surely know, and that everyone uses, in practice. However, theoretically speaking, this is the worst approach, because besides the above $\sqrt{2} = ?$ problem from (2), which persists with this approach, we have, as questions:

$$a_1 \ldots a_n.b_1b_2b_3 \ldots + c_1 \ldots c_m.d_1d_2d_3 \ldots = ?$$

$$a_1 \ldots a_n.b_1b_2b_3 \ldots \times c_1 \ldots c_m.d_1d_2d_3 \ldots = ?$$

And, you should agree with me here, solving these 2 problems is no easy business.

(4) So, this was for our general comments regarding the constructions (1,2,3), these are all interesting, with each one having is own advantages and disadvantages.

(5) Regarding now the proof, things here are quite long, the standard way being that of using (1) for building the theory of real numbers, which is certainly something that can be done, with some patience, and then developing some analysis theory for these real numbers, and proving the equivalence with (2) and with (3). For details here, you can check for instance my calculus book [11], where the whole story is explained. □

Back now to more concrete things, in the same topics, we have as well:

THEOREM 1.10. *The following happen, in relation with the passage* $\mathbb{Q} \to \mathbb{R}$*:*

(1) $\mathbb{Q}$ *is countable, while* $\mathbb{R}$ *is not countable.*
(2) $r \in \mathbb{R}$ *is rational precisely when its decimal writing is periodic.*
(3) *The probability for a randomly picked* $x \in \mathbb{R}$ *to be rational is* 0*.*

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

(1) We can count the positive rationals, with some redundancies, as follows:

$$\begin{array}{ccccc}
1/1 \to 1/2 & & 1/3 \to 1/4 & & \ldots \\
\swarrow & \nearrow & \swarrow & & \\
2/1 & 2/2 & 2/3 & 2/4 & \ldots \\
\downarrow & \nearrow & \swarrow & & \\
3/1 & 3/2 & 3/3 & 3/4 & \ldots \\
\swarrow & & & & \\
4/1 & 4/2 & 4/3 & 4/4 & \ldots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

Now after eliminating the redundancies, and then adding the negatives, which must be countable too, say via an alternating $+/-$ scheme, countability of $\mathbb{Q}$ proved.

(2) Regarding now the reals, assume by contradiction that $[0,1]$ is countable, listed as follows, and with the convention that the writings of type $\ldots 999 \ldots$ are avoided:

$$x_1 = 0.a_1a_2a_3\ldots$$
$$x_2 = 0.b_1b_2b_3\ldots$$
$$x_3 = 0.c_1c_2c_3\ldots$$
$$\ldots$$

Now pick digits $\sigma_1 \neq a_1$, $\sigma_2 \neq b_2$, $\sigma_3 \neq c_3$ and so on, again with a technical convention here, that these are different from 9, and define $x \in \mathbb{R}$ as follows:

$$x = 0.\sigma_1\sigma_2\sigma_3\ldots$$

We have then $x \in [0,1]$, and since x is obviously not on the above list, this is a contradiction. Thus $[0,1]$ is not countable, and it follows that $\mathbb{R}$ is not countable either.

(3) Assuming that $r \in \mathbb{R}$ has periodic decimal writing, the following computation, based on $(10^p-1)(\sum_k 10^{-kp})=1$, obtained by multiplying, gives $r \in \mathbb{Q}$:

$$\begin{aligned} r &= \pm a_1 \ldots a_m . b_1 \ldots b_n c_1 \ldots c_p c_1 \ldots c_p \ldots \\ &= \pm \frac{1}{10^n}\left(a_1 \ldots a_m b_1 \ldots b_n + c_1 \ldots c_p \left(\frac{1}{10^p}+\frac{1}{10^{2p}}+\ldots\right)\right) \\ &= \pm \frac{1}{10^n}\left(a_1 \ldots a_m b_1 \ldots b_n + \frac{c_1 \ldots c_p}{10^p-1}\right) \end{aligned}$$

(4) Conversely, given a rational number $r=k/l$, we can find its decimal writing by performing the usual division algorithm, k divided by l. But this algorithm will be surely periodic, after some time, so the decimal writing of r is indeed periodic.

(5) Finally, let us prove that the probability for a real number $x \in [0,1]$ to be rational is 0. So, let us write the rationals $r \in [0,1]$ in the form of a sequence, as follows:

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

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

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

Thus $P \leq c$, and since $c>0$ was arbitrary, we obtain $P=0$, as desired. □

With this discussed, time now for the real thing, analysis, which among others will clarify some of the considerations above? At the beginning of everything, we have:

DEFINITION 1.11. *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.

Examples in a moment, but before that, let us complement Definition 1.11 with:

DEFINITION 1.12. *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 now, we have:

PROPOSITION 1.13. *We have the following convergence,*

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

with $n \to \infty$.

PROOF. This is quite obvious, but let us prove it using Definitions 1.11 and 1.12:

(1) Assuming $a < 0$, pick $m \in \mathbb{N}$ such that $a < -1/m$. Since $n^a < n^{-1/m}$, it is enough to prove that $n^{-1/m} \to 0$. But for this latter purpose, observe that we have:

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

Thus we can take $N = [1/\varepsilon^m] + 1$ in Definition 1.11, and we are done.

(2) Assuming $a > 0$, pick $m \in \mathbb{N}$ such that $a > 1/m$. Since $n^a > n^{1/m}$, it is enough to prove that $n^{1/m} \to \infty$. But for this latter purpose, observe that we have:

$$n^{1/m} > K \iff n > K^m$$

Thus we can take $N = [K^m] + 1$ in Definition 1.12, and we are done. □

Next, we have some useful general results about limits, summarized as follows:

THEOREM 1.14. *The following happen:*

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

PROOF. All this is elementary, coming from definitions:

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

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

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

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

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

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

(4) This is clear from definitions. □

Here are as well some general rules for computing limits:

THEOREM 1.15. *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* $x_ny_n\to xy$.
(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:

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

(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 1.16. *The $n\to\infty$ limits of quotients of polynomials are given by*

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

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

PROOF. The first assertion comes from the following computation:

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

As for the second assertion, this comes from Proposition 1.13. □

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 1.17. *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.

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

THEOREM 1.18. *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 1.19. *$\mathbb{R}$ 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. This is related to what we were previously saying, in Theorem 1.9 (2), and I will leave the details here, using Theorem 1.18, to you, as an exercise. □

Moving on, towards some more exciting mathematics, let us start with:

DEFINITION 1.20. *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 illustration, we have:

PROPOSITION 1.21. *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. The first assertion can be established indeed as follows:

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

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 1.22. *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 $p\in\mathbb{N}$. But this comes from the following computation, with exercise for you, or check [11], for the estimate used:

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

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 1.23. *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 1.22, 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 1.24. *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 1.22.

(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 1.23, 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 $x_n \geq 0$ is removed, consider:

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

We have then $x_n/y_n \to 1$, but $\sum_n x_n$ converges, while $\sum_n y_n$ diverges. □

Here are some more useful positive results about the convergence of series:

THEOREM 1.25. *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 1.25, we have:

THEOREM 1.26. *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. □

1c. The number e

In order to formulate our next result, we will need the following key fact:

THEOREM 1.27. *We have the following inequality, for any $a_1, \ldots, a_n \geq 0$,*

$$\frac{a_1 + \ldots + a_n}{n} \geq \sqrt[n]{a_1 \ldots a_n}$$

telling us that the arithmetic mean is bigger than the geometric mean.

PROOF. This can be established in two steps, as follows:

(1) To start with, the result holds indeed at $n = 2$, with this coming from:

$$\frac{a+b}{2} \geq \sqrt{ab} \iff (\sqrt{a} - \sqrt{b})^2 \geq 0$$

But with this, we can prove our inequality at $n = 4$ too, then at $n = 8$, then at $n = 16$, and so on. Thus, as a conclusion, we know how to prove the result at any $n = 2^s$.

(2) In general now, given numbers $a_1, \ldots, a_n \geq 0$, consider their arithmetic mean:

$$m = \frac{a_1 + \ldots + a_n}{n}$$

Now pick $s \in \mathbb{N}$ such that $n \leq 2^s$, and let us complete our series $a_1, \ldots, a_n$ with $2^s - n$ copies of m. The arithmetic mean stays the same, and by using (1) we obtain:

$$\begin{aligned} m \geq \sqrt[2^s]{a_1 \ldots a_n m^{2^s - n}} &\implies m^{2^s} \geq a_1 \ldots a_n m^{2^s - n} \\ &\implies m^n \geq a_1 \ldots a_n \\ &\implies m \geq \sqrt[n]{a_1 \ldots a_n} \end{aligned}$$

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

Good news, we can talk now about a very interesting convergence, as follows:

THEOREM 1.28. *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 claim, we can use the arithmetic-geometric inequality, applied to the number 1, and to n copies of the number $1 + 1/n$. Indeed, this gives:

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

By rearranging a bit, and 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 will 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, as before for x_n, this gives the following inequality:

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

But now, by inverting this 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 in a moment, with better methods. □

More generally now, we have the following result, featuring a variable $x \in \mathbb{R}$:

THEOREM 1.29. *We have the following formula,*

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

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

PROOF. We know that this holds at $x = 1$, by definition of e, and by inverting, we have it at $x = -1$ too. But then, when $x \in \mathbb{R}$ is arbitrary, we can proceed as follows:

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

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

All this is very nice, but we have as well an alternative approach to e, as follows:

THEOREM 1.30. *We have the following formula,*

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

which can stand as an alternative definition for $e = 2.71828\ldots$

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

(1) In practice, we want to prove that we have the following equality:

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

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

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

In order to prove now that this limit is indeed e, observe that we have:

$$\left(1 + \frac{1}{n}\right)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot \frac{1}{n^k} \leq \sum_{k=0}^{n} \frac{1}{k!}$$

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

(2) For the reverse inequality, we have the following computation:

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

Next, we can use the following trivial inequality, valid for any number $x \in (0, 1)$:

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

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

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

Thus, we have our needed estimate, and this finishing the proof. □

As a last result now about e, which is something quite far-reaching, we have:

THEOREM 1.31. *We have the following formula,*

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

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

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

(1) In order to prove the result, consider the following function, whose convergence is clear by using our various criteria for series from Theorem 1.25:

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

(2) By using the binomial formula, we have the following computation:

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

(3) Next, observe that this shows that f is continuous. Indeed, at $x = 0$ we have:

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

But from this, we get $f(x+t) = f(x)f(t) \to f(x)$ with $t \to 0$, at any x. Thus, as a conclusion, our function f is continuous, and satisfies the following conditions:

$$f(x+y) = f(x)f(y) \quad , \quad f(1) = e$$

(4) But with this, we can finish. Indeed, by iterating, we have $f(nx) = f(x)^n$ for any $n \in \mathbb{N}$. Then, by extracting roots, we have $f(rx) = f(x)^r$ for any $r \in \mathbb{Q}$. Thus $f(r) = e^r$ for any $r \in \mathbb{Q}$, and by continuity we obtain $f(x) = e^x$ for any $x \in \mathbb{R}$, as desired. □

1d. Logarithms

As a last topic for this chapter, which is of key importance for the physics to come next, in chapter 2 below, we can talk about the logarithm function, as follows:

THEOREM 1.32. *The exponential function, as constructed before,*

$$\exp : \mathbb{R} \to (0, \infty) \quad , \quad x \to e^x$$

is invertible, with its inverse being the logarithm function, denoted as follows:

$$\log : (0, \infty) \to \mathbb{R} \quad , \quad \log = \exp^{-1}$$

This logarithm function is continuous, increasing, and bijective.

PROOF. This is indeed something self-explanatory, based on our theory of e. However, for fully understanding this, do not hesitate to cheat a bit, and draw some pictures, with these 2D pictures being of course officially forbidden, in the present Part I. □

In practice, the logarithm is a very useful function, and more on this later. In the meantime, how to compute it? And here, we have the following collection of results:

PROPOSITION 1.33. *The logarithm can be computed via the equivalent formulae*

$$e^{\log x} = x \quad , \quad \log(e^x) = x$$

and in practice, we have the following useful rules:

(1) $\log(xy) = \log x + \log y$.
(2) $\log(1/y) = -\log y$.
(3) $\log(x/y) = \log x - \log y$.
(4) $\log(x^p) = p \log x$.

PROOF. The first two formulae come from the fact that the inverse of a function f can be defined either via $f(f^{-1}(x)) = x$, or via $f^{-1}(f(x)) = x$. As for the rest:

(1) This comes indeed from the following computation:

$$e^{\log(xy)} = xy = e^{\log x} e^{\log y} = e^{\log x + \log y}$$

(2) This comes from (1), by setting $x = 1/y$, and using $\log 1 = 0$.

(3) This comes also from (1), by replacing $y \to 1/y$, and using (2).

(4) This formula, generalizing (1) with $x = y$, and also (2), comes as follows:

$$e^{\log(x^p)} = x^p = (e^{\log x})^p = e^{p \log x}$$

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

Time now for some tough mathematics? Here is a main result about log:

THEOREM 1.34. *We have the following formula, valid for $|x| < 1$:*

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

Moreover, this holds as well at $x = 1$, with the formula here being:

$$\log 2 = \sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}$$

As for remaining cases, $|x| > 1$ and $x = -1$, here the series diverges.

PROOF. We have several things going on here, the idea being as follows:

(1) To start with, the series in the statement converges at $|x| < 1$, due to:

$$\frac{x^{k+1}}{k+1} : \frac{x^k}{k} = x \cdot \frac{k+1}{k} \to x$$

The series converges as well at $x = 1$, say as an alternating series. At $x = -1$ however we obtain the Riemann divergent sum. As for the case $|x| > 1$, here the series grossly diverges, in the sense that its terms do not go to 0, as required by convergence.

(2) Getting now to the proof, we will prove that our series has the main required abstract property of $\log(1+x)$, without reference to exp, namely:

$$\log((1+x)(1+y)) = \log(1+x) + \log(1+y)$$

So, let us get into this. We have the following computation, to start with:

$$\begin{aligned}\sum_{k=1}^{\infty}(-1)^{k+1}\frac{(x+y+xy)^k}{k} &= \sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}\sum_{r=0}^{k}\binom{k}{r}(x+xy)^r y^{k-r}\\ &= \sum_{r+s\geq 1}\frac{(-1)^{r+s+1}}{r+s}\binom{r+s}{r}x^r(1+y)^r y^s\end{aligned}$$

(3) The contribution of $r = 0$ to this sum is the following quantity:

$$C_0 = \sum_{s\geq 1}(-1)^{s+1}\frac{y^s}{s}$$

(4) In order to compute now the contribution coming from $r \geq 1$, we will need the following standard inversion formula for $(1-t)^r$, which comes by recurrence, as a consequence of the Pascal formula for binomials, that I will leave as an exercise for you:

$$\frac{1}{(1-t)^r} = \sum_{s\geq 0}\frac{1}{r+s}\binom{r+s-1}{r-1}t^s$$

(5) Indeed, with this in hand, the contribution coming from $r \geq 1$ is given by:

$$\begin{aligned} C_+ &= \sum_{r\geq 1}(-1)^{r+1}x^r(1+y)^r\sum_{s\geq 0}\frac{(-1)^s}{r+s}\binom{r+s}{r}y^s \\ &= \sum_{r\geq 1}(-1)^{r+1}\frac{x^r(1+y)^r}{r}\sum_{s\geq 0}\frac{(-1)^s}{r+s}\binom{r+s-1}{r-1}y^s \\ &= \sum_{r\geq 1}(-1)^{r+1}\frac{x^r(1+y)^r}{r}\cdot\frac{1}{(1+y)^r} \\ &= \sum_{r\geq 1}(-1)^{r+1}\frac{x^r}{r} \end{aligned}$$

(6) Summarizing, we have proved the desired formula, namely:

$$\sum_{k=1}^{\infty}(-1)^{k+1}\frac{(x+y+xy)^k}{k} = \sum_{r=1}^{\infty}(-1)^{r+1}\frac{x^r}{r} + \sum_{s=1}^{\infty}(-1)^{s+1}\frac{y^s}{s}$$

(7) But with this, we can now finish, a bit as we did before for exp, in the proof of Theorem 1.31, and I will leave the technical details here to you, as an exercise. □

1e. Exercises

This was a standard introduction to the real numbers, and as exercises, we have:

EXERCISE 1.35. *Learn values of binomial coefficients, as many as you can.*

EXERCISE 1.36. *Clarify what we said, regarding the real numbers appearing as cuts.*

EXERCISE 1.37. *Clarify as well everything in relation with the decimal expansion.*

EXERCISE 1.38. *Clarify also the fact that* $\mathbb{R}$ *appears as the completion of* $\mathbb{Q}$.

EXERCISE 1.39. *Do the remaining work, in the proof of the Riemann theorem.*

EXERCISE 1.40. *Clarify the various details, in the proof of* $e^x = \sum_{k=0}^{\infty} x^k/k!$.

EXERCISE 1.41. *Do the same for the proof of* $\log(1+x) = \sum_{k=1}^{\infty}(-1)^{k+1}x^k/k$.

EXERCISE 1.42. *Learn also about logarithms in other bases.*

As bonus exercise, have a look as well at the basics of probability theory.

CHAPTER 2

Motion basics

2a. Motion, collisions

Good news, with the mathematics that we know we can do some physics, in 1D. Let us begin with a discussion of the usual motion, in the absence of forces. This 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 traveling at speeds v_1, v_2, the total momentum of the system*

$$p = m_1v_1 + m_2v_2$$

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 exactly is going on, in relation with the above, consider two objects as in Fact 2.1, which are 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 will allow us to do the math. And, more on this, explicit computations based on Fact 2.1 and Fact 2.3, in a moment.

As a second 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. It is possible to come up with some heuristics here, with the knowledge that we have, but it is better to defer the discussion for later, when talking forces, Newton and calculus. So, more on this later, and trust me in the meantime.

Going ahead now, let us first investigate, just out of curiosity, what happens to the energy during a plastic collision. And things here are quite interesting, with the result, which might seem quite surprising, contradicting our previous guess that the moment conservation comes somehow from a "no energy lost" principle, being 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:

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

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. □

As already mentioned, the above result might seem quite surprising, contradicting our previous guess that the moment conservation principle comes from something of type "no energy lost". Let us record this finding in the form of an informal statement:

CONCLUSION 2.5. *Momentum ain't the same thing as energy.*

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':

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

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

PROPOSITION 2.6. *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 2.6, as follows:

THEOREM 2.7. *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 2.6, we have done some quick algebra there, without really knowing what we're doing, leading to the following formulae:

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

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

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. Rocket science

Now that we learned some physics, time for some engineering? I bet that you would love doing so. So, as a main application of the general theory developed above, and in relation with gravity as well, we can use momentum for beating gravity, as follows:

THEOREM 2.8. *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. Many things can be said here, the idea being as follows:

(1) To start with, the functioning principle of the rockets, as formulated in the statement, is clear indeed from the conservation of the momentum principle, because ejecting mass to the left will move us to the right. Thus, victory, we have our rockets.

(2) Getting now to the mathematics of the rockets, as a warm-up computation, let us first study the case of a single ejection. We begin with our rocket M at rest:

$$\bullet_M$$

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

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

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

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

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

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

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:

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

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 (2) for the value of v_1, the speed after the second ejection is given by:

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

(4) 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 finally, in the case where both the mass m and speed s are constant, we get:

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

Summarizing, we have our rockets, along with some useful math for their speed. □

We learned many things in the above proof, and for future reference, let us record:

THEOREM 2.9 (addendum). *In the above context of rocket engineering,*

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

is the speed, for a rocket of rest mass M, after ejecting k times mass m, at speed s.

PROOF. This is indeed the main formula that we obtained in the previous proof, at the end, which captures the essentials, and that we will use in what follows. □

Let us investigate now the asymptotics. We will assume that we are in the context of Theorem 2.9, 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:

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

Mathematically, this leads us into the following question, which is non-trivial:

QUESTION 2.10. *What is the asymptotics of the rocket speed*

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

in the regime where $\varepsilon > 0$ is small, and $k\varepsilon = r \in (0,1)$ is constant?

In order to deal now with this latter question, we can use the geometric series formula from chapter 1, which leads us to the following formula, for the rocket speed:

$$\begin{aligned}\frac{v_k}{s} &= \frac{\varepsilon}{1-\varepsilon}+\frac{\varepsilon}{1-2\varepsilon}+\frac{\varepsilon}{1-3\varepsilon}+\frac{\varepsilon}{1-4\varepsilon}+\ldots\ldots+\frac{\varepsilon}{1-k\varepsilon}\\ &= \varepsilon+\varepsilon^2+\varepsilon^3+\varepsilon^4+\varepsilon^5+\ldots\\ &\quad +\varepsilon+2\varepsilon^2+4\varepsilon^3+8\varepsilon^4+16\varepsilon^5+\ldots\\ &\quad +\varepsilon+3\varepsilon^2+9\varepsilon^3+27\varepsilon^4+81\varepsilon^5+\ldots\\ &\quad +\varepsilon+4\varepsilon^2+16\varepsilon^3+64\varepsilon^4+256\varepsilon^5+\ldots\\ &\quad \vdots\\ &\quad +\varepsilon+k\varepsilon^2+k^2\varepsilon^3+k^3\varepsilon^4+k^4\varepsilon^5+\ldots\end{aligned}$$

Now by summing over the columns, we obtain from this the following formula:

$$\begin{aligned}\frac{v_k}{s} &= (1+1+1+1+\ldots+1)\varepsilon\\ &\quad +(1+2+3+4+\ldots+k)\varepsilon^2\\ &\quad +(1+4+9+16+\ldots+k^2)\varepsilon^3\\ &\quad +(1+8+27+64+\ldots+k^3)\varepsilon^4\\ &\quad +(1+16+81+256+\ldots+k^4)\varepsilon^5\\ &\quad +\ldots\end{aligned}$$

Summarizing, at the level of mathematics, we are led into the following question:

$$1^p+\ldots+k^p=?$$

Now browsing through what we did in chapter 1, all sorts of interesting things there, no doubt about this, but with none truly answering our question above. Nevermind. This is actually commonplace when doing theoretical physics, you work and work, struggling with phenomenology, as to reach to some neat, concrete mathematical question, and then you ask your mathematics collegues, and no one knows. Although mathematicians can complain too about physicists, for some other reasons, and more on this later.

Well, this is the situation, and time I guess for the two of us to do some mathematics, matter of showing that folks what two physicists can do? Here is our theorem:

THEOREM 2.11. *We have the following estimate,*

$$1^p+2^p+\ldots+k^p\simeq\frac{k^{p+1}}{p+1}$$

valid for any $p\in\mathbb{N}$, in the $k\to\infty$ limit.

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

(1) At $p = 0$, there is obviously nothing to do, because we have the following formula, which is exact, and which proves our estimate:

$$1^0 + 2^0 + \ldots + k^0 = k$$

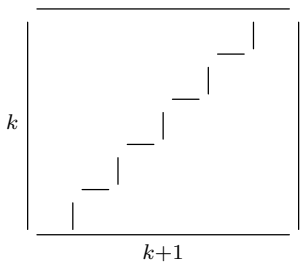
(2) At $p = 1$ now, we are confronted with a well-known question, namely the computation of $1 + 2 + \ldots + k$. But this is simplest done by arguing that the average of the numbers $1, 2, \ldots, k$ being the number in the middle, we have:

$$\frac{1 + 2 + \ldots + k}{k} = \frac{k + 1}{2}$$

Thus, we obtain the following formula, which again solves our question:

$$1 + 2 + \ldots + k = \frac{k(k + 1)}{2} \simeq \frac{k^2}{2}$$

(3) At $p = 2$ now, go compute $1^2 + 2^2 + \ldots + k^2$, this does not look obvious. So as a preliminary, let us go back to the case $p = 1$, and try to find a new proof there, which might have some chances to extend at $p = 2$. And here, and coming with my apologies for using 2 dimensions, officially forbidden in this Part I, we have the following trick:



Now the point is that this trick works at $p = 2$ too. Indeed, if we consider the 3D shape P formed by a succession of solid squares, having sizes 1×1, 2×2, 3×3, and so on up to $k \times k$, if we stack 6 copies of P we get a parallelepiped, which gives:

$$1^2 + 2^2 + \ldots + k^2 = \frac{k(k + 1)(2k + 1)}{6} \simeq \frac{k^3}{3}$$

Alternatively, and avoding any higher dimensional reasoning, as per our usual 1D guidelines for the present Part I, you can get this by computing $1^2 + 2^2 + \ldots + k^2$ for small values of k, then based on this, conjecturing the above formula, and then proving your conjecture by recurrence. We will leave this as a nice, relaxing exercise.

(4) At $p = 3$ now, the legend has it that by deeply thinking in 4D we are led to the following formula, a bit as in the cases $p = 1, 2$, explained above:

$$1^3 + 2^3 + \ldots + k^3 = \frac{k^2(k + 1)^2}{4} \simeq \frac{k^4}{4}$$

Alternatively, assuming that the gods of combinatorics are with us, we can see right away the following formula, which gives the result:

$$1^3+2^3+\ldots+k^3=(1+2+\ldots+k)^2$$

In any case, in practice, the above formula holds indeed, and you can of course come upon it via some numerics too, and then check it by recurrence. Exercise for you.

(5) All this is very nice, but coming now as bad news, at $p=4,5,6,\ldots$ the situation is more complicated, with the formulae, provable by recurrence, being as follows:

$$1^4+2^4+\ldots+k^4=\frac{k(k+1)(2k+1)(3k^2+3k-1)}{30}\simeq\frac{k^5}{5}$$

$$1^5+2^5+\ldots+k^5=\frac{k^2(k+1)^2(2k^2+2k+1)}{12}\simeq\frac{k^6}{6}$$

$$1^6+2^6+\ldots+k^6=\frac{k(k+1)(2k+1)(3k^4+6k^3-3k+1)}{42}\simeq\frac{k^7}{7}$$

$$\vdots$$

(6) In order to discuss now the general case, based somehow on the above formulae, let us introduce the Bernoulli numbers $B_n\in\mathbb{Q}$, via the following recurrence:

$$B_1=1\qquad,\qquad\sum_{n=0}^{n}\binom{m+1}{n}B_n=\delta_{m0}$$

In practice, using the recurrence, the data for the first such numbers is as follows:

$$B_0=1\ ,\quad B_1=-\frac{1}{2}\ ,\quad B_2=\frac{1}{6}\ ,\quad B_3=0\ ,\quad B_4=-\frac{1}{30}\ ,\quad B_5=0\ ,\quad B_6=\frac{1}{42}$$

(7) Now back to our questions, the point is that we have the following formula, valid at any $p\in\mathbb{N}$, and which is provable by a straightforward recurrence:

$$1^p+2^p+\ldots+k^p=\frac{1}{p+1}\sum_{n=0}^{p}(-1)^n\binom{p+1}{n}B_nk^{p+1-n}$$

To be more precise, this is compatible with our previous findings at $p=0,\ldots,6$, by using the first Bernoulli numbers from (6). As for the proof, when trying to establish this by recurrence, we are led into the standard recurrence formula for the Bernoulli numbers, from (6). And we will leave clarifying this, and some further learning, as an exercise.

(8) Now the point is that the above formula does the job for our asymptotic purposes, because we obtain right away the following estimate, exactly as needed:

$$1^p+2^p+\ldots+k^p\simeq\frac{k^{p+1}}{p+1}$$

Summarizing, theorem proved, with the cases $p=0,1,2,3$ being elementary, then $p=4,5,6,\ldots$ being doable too, and $p\in\mathbb{N}$ general requiring Bernoulli numbers. □

Still with me I hope, after all this mathematics, and with the reiterated comment that, even when planning to do a research career in physics, or engineering, you will have to know how to do some, when needed. We can go back now to the rockets, and we have:

THEOREM 2.12. *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

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

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. This follows by putting together all the above, as follows:

(1) We know the first formula of v from Theorem 2.9, and the asymptotic study, using the computations made after Question 2.10, recalled here for convenience, written in a more compact form, and then the key estimate from Theorem 2.11, is as follows:

$$\begin{aligned}
v&=\left(\frac{m}{M-m}+\frac{m}{M-2m}+\ldots+\frac{m}{M-km}\right)s\\
&=\left(\frac{\varepsilon}{1-\varepsilon}+\frac{\varepsilon}{1-2\varepsilon}+\ldots+\frac{\varepsilon}{1-k\varepsilon}\right)s\\
&=\left(\sum_{n=1}^{k}\sum_{p=0}^{\infty}n^p\varepsilon^{p+1}\right)s\\
&=\left(\sum_{p=0}^{\infty}\varepsilon^{p+1}\sum_{n=1}^{k}n^p\right)s\\
&\simeq\left(\sum_{p=0}^{\infty}\varepsilon^{p+1}\cdot\frac{k^{p+1}}{p+1}\right)s\\
&=\left(\sum_{p=0}^{\infty}\frac{r^{p+1}}{p+1}\right)s\\
&=-\log(1-r)s
\end{aligned}$$

(2) As an illustration, assume that our rocket has shrinked, from a continuous ejection process at speed s, up to mass $M/\sqrt{e}$. In this case we have $r=1-1/\sqrt{e}$, and according

to our general formula, which now becomes exact, the velocity reached is:

$$v = -\log\left(\frac{1}{\sqrt{e}}\right)s = \frac{s}{2}$$

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.

(3) Regarding now the last assertion, recall from the proof of Theorem 2.8 that, assuming only that s is constant, we have the following formula, for the rocket speed:

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

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 various mathematical arguments from the proof of (1) above will apply as well, and give the result.

(4) Finally, as a question that you might have, is that crazy mathematics from the proof of Theorem 2.11 really needed, in order to establish such a simple formula, like the one in the statement? In answer, we can kill our problem with calculus, as follows:

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

And more on such things, integrals, their meaning and computation, in chapter 3. So, come back here later, after reading chapter 3, when we will discuss calculus. □

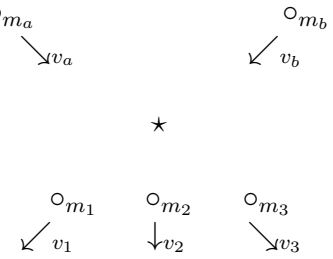
Finally, let us mention that the above results remain a bit theoretical, because if you are a space engineer, one of your main concerns, besides of course beating gravity, is that of beating atmospheric drag too. More on this, hopefully, later in this book.

2c. Particle decay

As a continuation of our discussion about collisions, ejections and rockets, and still with civil engineering ideas in mind, let us discuss now radioactivity. I mean, now that we have rockets, we are in need of some warheads for them, right.

In order to do so, we must talk about the interactions between the various types of particles, appearing from nuclear physics. But here, we already have some experience

from classical physics, with the typical picture of what can happen, represented here in 2D just for the sake of having a better picture, being as follows:



To be more precise, this diagram describes a collision between two particles, but we can of course allow further particles entering the collision, and then the several particles emerging from this collision. The data of each particle, which in classical mechanics means mass m and speed v, is carefully recorded, with of course the aim of recovering the output data from the input data. Finally, all this can take place in arbitrary dimensions, and also, importantly, the collision can be elastic, or plastic, or any mixture of these.

This was for basic interactions in classical mechanics. In the particle physics setting, 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 2.13. *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,* $*_a + *_b \to *_1 + \ldots + *_n$.

Getting to work now, with our present knowledge of mathematics and physics, we can look at the mathematics of decay. Indeed, ignoring the physics, which is certainly complicated business, this is basically a matter of probability and statistics:

THEOREM 2.14. *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 before with Theorem 2.12, this sounds a bit like calculus, but we can make our way through all this with bare hands, or almost, the idea being as follows:

(1) As a first remark, in mathematical terms, with N_t being the number of particles remaining at time $t > 0$, our definition of the decay rate λ reads:

$$\lim_{h \to 0} \frac{N_{t+h} - N_t}{h} = -\lambda N_t$$

We also know that at time $t = 0$, the number of particles is a given N_0. Now let us try to find an expression for our function N_t, as a series in the variable t:

$$N_t = N_0 + c_1 t + c_2 t^2 + c_3 t^3 + \ldots$$

In order to do so, observe that, according to the binomial formula, we have the following estimate, for any $t \in \mathbb{R}$, and any exponent $p \in \mathbb{N}$, in the $h \to 0$ limit:

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

Equivalently, we have the following estimate, for any $t \in \mathbb{R}$, and any $p \in \mathbb{N}$:

$$\lim_{h \to 0} \frac{(t+h)^p - t^p}{h} = pt^{p-1}$$

But with this, we conclude that the definition of the decay rate λ, in terms of the function $N_t = N_0 + c_1 t + c_2 t^2 + c_3 t^3 + \ldots$ that we are looking for, reads:

$$c_1 + 2c_2 t + 3c_3 t^2 + 4c_4 t^3 + \ldots = -\lambda(N_0 + c_1 t + c_2 t^2 + c_3 t^3 + \ldots)$$

Thus the coefficients c_k can be computed by recurrence, and we obtain:

$$c_1 = -\lambda N_0 \quad , \quad c_2 = \frac{\lambda^2}{2} N_0 \quad , \quad c_3 = -\frac{\lambda^3}{6} N_0 \quad , \quad c_4 = \frac{\lambda^4}{24} N_0 \quad , \quad \ldots$$

We are therefore led to the formula in the statement, namely:

$$\begin{aligned}N_t &= N_0 + c_1 t + c_2 t^2 + c_3 t^3 + \ldots \\ &= N_0 - \lambda N_0 t + \frac{\lambda^2}{2} N_0 t^2 - \frac{\lambda^3}{6} N_0 t^3 + \frac{\lambda^4}{24} N_0 t^4 - \ldots \\ &= \left(1 - \lambda t + \frac{\lambda^2 t^2}{2} - \frac{\lambda^3 t^3}{6} + \frac{\lambda^4 t^4}{24} - \ldots\right) N_0 \\ &= e^{-\lambda t} N_0\end{aligned}$$

Finally, in relation with this, you might argue that the power series assumption that we made, $N_t = N_0 + c_1 t + c_2 t^2 + c_3 t^3 + \ldots$, was something quite strong, and who knows, maybe there are some other solutions of our decay rate equation, not of this type. Good point, and in answer, calculus is there for clarifying such things, the idea being:

$$\begin{aligned}N_t' = -\lambda N_t &\implies (e^{\lambda t} N_t)' = \lambda e^{\lambda t} \cdot N_t + e^{\lambda t} \cdot (-\lambda N_t) = 0 \\ &\implies N_t = e^{-\lambda t} N_0\end{aligned}$$

In short, come back here after reading chapter 3, for this quick, complete proof.

(2) Getting now to the mean lifetime question, the starting point will be the formula $N_t = e^{-\lambda t}N_0$ found above, for the number N_t of particles remaining at time t. In probabilistic terms, we density describing the number of particles left at time t is:

$$D_t = \frac{N_t}{c} \quad , \quad c = \lim_{T\to\infty}\lim_{k\to\infty}\frac{T}{k}\left(N_{T/k} + N_{2T/k} + \ldots + N_{(k-1)T/k}\right)$$

So, let us compute the limit on the right. This is quite easy business, as follows:

$$\begin{aligned}
c &= \lim_{T\to\infty}\lim_{k\to\infty}\frac{T}{k}\left(N_{T/k} + N_{2T/k} + \ldots + N_{(k-1)T/k}\right)\\
&= N_0\lim_{T\to\infty}\lim_{k\to\infty}\frac{T}{k}\left(1 + e^{-\lambda T/k} + e^{-2\lambda T/k} + \ldots + e^{-(k-1)\lambda T/k}\right)\\
&= N_0\lim_{T\to\infty}\lim_{k\to\infty}\frac{T}{k}\cdot\frac{1-e^{-\lambda T}}{1-e^{-\lambda T/k}}\\
&= \frac{N_0}{\lambda}\lim_{T\to\infty}(1-e^{-\lambda T})\lim_{k\to\infty}\frac{\lambda T/k}{1-e^{-\lambda T/k}}\\
&= \frac{N_0}{\lambda}\lim_{T\to\infty}(1-e^{-\lambda T})\\
&= \frac{N_0}{\lambda}
\end{aligned}$$

We conclude that the density function that we were looking for is given by:

$$D_t = \frac{N_t}{c} = \frac{e^{-\lambda t}N_0}{N_0/\lambda} = \lambda e^{-\lambda t}$$

Now let us compute the mean lifetime τ. A bit of probabilistic thinking shows that, in terms of the density function, this is given by the following formula:

$$\tau = \lim_{T\to\infty}\lim_{k\to\infty}\frac{T}{k}\left(\frac{T}{k}D_{T/k} + \frac{2T}{k}D_{2T/k} + \ldots + \frac{(k-1)T}{k}D_{(k-1)T/k}\right)$$

So, here we go again with computing limits. Using the formula of D_t, we have:

$$\begin{aligned}
\tau &= \lim_{T\to\infty}\lim_{k\to\infty}\frac{T}{k}\left(\frac{T}{k}D_{T/k} + \frac{2T}{k}D_{2T/k} + \ldots + \frac{(k-1)T}{k}D_{(k-1)T/k}\right)\\
&= \lim_{T\to\infty}\lim_{k\to\infty}\frac{T}{k}\left(\frac{\lambda T}{k}e^{-\lambda T/k} + \frac{2\lambda T}{k}e^{-2\lambda T/k} + \ldots + \frac{(k-1)\lambda T}{k}e^{-(k-1)\lambda T/k}\right)\\
&= \frac{1}{\lambda}\lim_{T\to\infty}\lim_{k\to\infty}\left(\frac{\lambda T}{k}\right)^2\left(e^{-\lambda T/k} + 2e^{-2\lambda T/k} + \ldots + (k-1)e^{-(k-1)\lambda T/k}\right)
\end{aligned}$$

And with this, we are a bit in trouble, because we need here a formula for sums of type $S = x + 2x^2 + 3x^3 + \ldots + (k-1)x^{k-1}$, and that formula is nowhere to be found in chapter 1. So, shame on those who wrote chapter 1, these mathematicians should keep an

eye on physics, when developing their mathematics, as for their work to be truly useful, and I guess we will have to do this computation, the two of us. We have:

$$\begin{aligned}
S &= x + 2x^2 + 3x^3 + \ldots + (k-1)x^{k-1} \\
&= x + x^2 + x^3 + \ldots + x^{k-1} \\
&\quad + x^2 + x^3 + \ldots + x^{k-1} \\
&\quad \vdots \\
&\quad + x^{k-1} \\
&= \frac{x - x^k}{1-x} + \frac{x^2 - x^k}{1-x} + \ldots + \frac{x^{k-1} - x^k}{1-x} \\
&= \frac{x + x^2 + \ldots + x^{k-1}}{1-x} - \frac{(k-1)x^k}{1-x} \\
&= \frac{x - x^k}{(1-x)^2} - \frac{(k-1)x^k}{1-x}
\end{aligned}$$

Now by getting back to the computation of the mean lifetime, we have:

$$\begin{aligned}
\tau &= \frac{1}{\lambda} \lim_{T\to\infty} \lim_{k\to\infty} \left(\frac{\lambda T}{k}\right)^2 \left(e^{-\lambda T/k} + 2e^{-2\lambda T/k} + \ldots + (k-1)e^{-(k-1)\lambda T/k}\right) \\
&= \frac{1}{\lambda} \lim_{T\to\infty} \lim_{k\to\infty} \left(\frac{\lambda T}{k}\right)^2 \left(\frac{e^{-\lambda T/k} - e^{-\lambda T}}{(1-e^{-\lambda T/k})^2} - \frac{(k-1)e^{-\lambda T}}{1-e^{-\lambda T/k}}\right) \\
&= \frac{1}{\lambda} \lim_{T\to\infty} \left(\lim_{k\to\infty} \left(\frac{\lambda T}{k}\right)^2 \frac{e^{-\lambda T/k} - e^{-\lambda T}}{(1-e^{-\lambda T/k})^2} - \lim_{k\to\infty} \left(\frac{\lambda T}{k}\right)^2 \frac{(k-1)e^{-\lambda T}}{1-e^{-\lambda T/k}}\right)
\end{aligned}$$

So, let us compute the above two limits. The first one is easily done, as follows:

$$\begin{aligned}
\lim_{k\to\infty} \left(\frac{\lambda T}{k}\right)^2 \frac{e^{-\lambda T/k} - e^{-\lambda T}}{(1-e^{-\lambda T/k})^2} &= \lim_{k\to\infty} \left(\frac{\lambda T}{k}\right)^2 \frac{1 - \lambda T/k - e^{-\lambda T}}{(\lambda T/k)^2} \\
&= 1 - e^{-\lambda T}
\end{aligned}$$

As for second limit appearing above, this is easy to compute too, as follows:

$$\begin{aligned}
\lim_{k\to\infty} \left(\frac{\lambda T}{k}\right)^2 \frac{(k-1)e^{-\lambda T}}{1-e^{-\lambda T/k}} &= \lim_{k\to\infty} \left(\frac{\lambda T}{k}\right)^2 \frac{(k-1)e^{-\lambda T}}{\lambda T/k} \\
&= \lambda T e^{-\lambda T}
\end{aligned}$$

We can now finish the computation of the mean lifetime, and we obtain:

$$\begin{aligned}\tau &= \frac{1}{\lambda}\lim_{T\to\infty} 1-e^{-\lambda T}-\lambda T e^{-\lambda T}\\ &= \frac{1}{\lambda}(1-0-0)\\ &= \frac{1}{\lambda}\end{aligned}$$

Good work that we did. For the rest, as usual with such computations, the comment is that some knowledge of calculus can drastically simplify all this. Indeed, in what concerns our first computation, that of the density function D_t, that can be done as follows:

$$D_t = N_0e^{-\lambda t}\Big/\int_0^\infty N_0e^{-\lambda t}dt = N_0e^{-\lambda t}\Big/\frac{N_0}{\lambda} = \lambda e^{-\lambda t}$$

As for the computation of the mean lifetime τ, that can be done as follows:

$$\tau = \int_0^\infty \lambda t e^{-\lambda t}dt = \int_0^\infty t(-e^{-\lambda t})'dt = \int_0^\infty e^{-\lambda t}dt = \frac{1}{\lambda}$$

In short, come back here after reading chapter 3, for this quick, complete proof.

(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:

$$N_t = 2^{-t/t_{1/2}}N_0$$

Now by comparing with $N_t = e^{-\lambda t}N_0$, this gives $t_{1/2} = (\log 2)/\lambda$, as stated. □

And with this, end of our discussion regarding particle decay. We will come back to such questions later in this book, when knowing more about the particles in question.

2d. Einstein, relativity

We have seen so far that we can have some interesting 1D physics going on, on several nice topics. However, at the mathematical level, all the above makes it quite clear that, for more, we are in need of calculus. So, it is probably wiser to stop here with physics, and defer the discussion for later, in chapter 4, after learning calculus in chapter 3.

So, no hurry and no worries, and please be sure that we will do a lot of interesting physics in chapter 4, using the calculus that we will soon learn in chapter 3, dealing with basic notions such as forces, speed, acceleration, gravity and free falls, then elasticity and waves, and then pressure and heat. All these happening in 1D, of course.

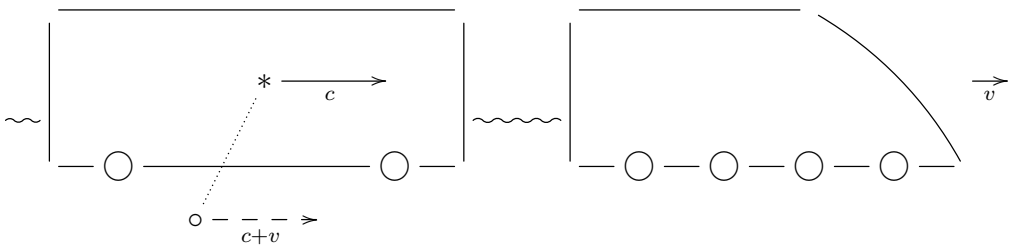
This being said, as a quite amazing fact, we can still talk here, at the end of this chapter, about Einstein's relativity theory, in 1D. To be more precise, by a strange twist of fate, while being certainly something very modern, and of great deepness, the basics of Einstein's theory do not need any calculus, or advanced mathematics in general.

Getting started, based on various experiments, and a bit of abstract thinking too, Einstein came upon the following principles, which contradict classical mechanics:

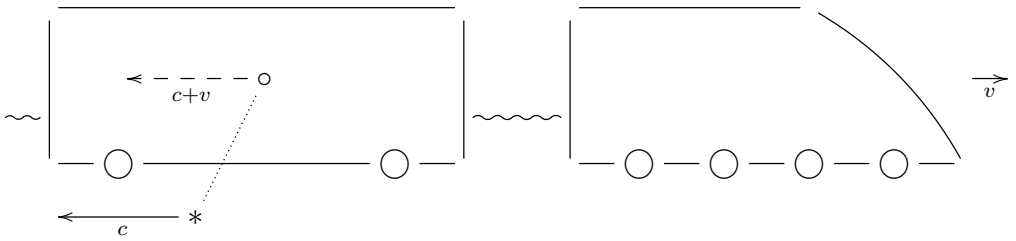
FACT 2.15 (Einstein principles). *The following happen:*

(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, $v \not> c$.*

All this needs of course some details and explanations, and we will come to this, in a moment. However, before that, let us see what the contradiction with classical mechanics is. This basically comes from (2), which is the tricky point in the above, because if we assume 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 traveling at speed $c + v > c$, which is a contradiction:



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



Obviously, in view of this, there are many things to be done. Since speed is distance over time, $v = d/t$, maybe something is wrong with distance d, or with time t, or with both. Or maybe something is wrong with the formula $v = d/t$ itself, who knows. In addition, even when assuming that we manage to fix all this, speeds, then comes a review of the momentum p, and of energy E, which crucially depend on speed. And so on.

Summarizing, many things to be done. As a first task, however, before going into computations and everything, let us remain a bit philosophers, and think some more at Fact 2.15. With the question being, can that damn thing be 1-dimensional?

The problem indeed comes from light, because when assuming a bit of knowledge of advanced physics, and more on this later in this book, light is in fact an electromagnetic phenomenon, and with electromagnetism itself being something notoriously 3D.

So, can we have light in 1D, the framework for the present chapter, and book part? Not an easy question, and in the lack of an answer here, I will have I guess to ask an expert. And, picking one on the ground, no idea if that's the same guy as the one from chapter 1, these little fellows all look the same to me, here is what our expert says:

ANT 2.16. *Dude, look me in the eye, we ants are not blind, sure we have light in 1D. But I won't tell you by which precise mechanism.*

Okay, interesting all this, not that I have a full answer to my question, but the first part will do, so we are good for some 1D relativity. By the way, I must admit that I find these little ants more and more annoying, now that I know them better. But with the other big animals at the house, namely cats and rats, being totally cool about them, I will keep my cool too. So thanks little ant, and see you later, for more discussions.

Getting to work now, we need to invent some new mathematics, as for $c + v = c$ to hold. And here, let us first recall that, in the classical case, we have:

PROPOSITION 2.17. *The classical speeds add according to the Galileo formula*

$$v_{AC} = v_{AB} + v_{BC}$$

where v_{AB} denotes the relative speed of A with respect to B.

PROOF. This is clear indeed from the definition of speed, and very intuitive. □

In order to find now the fix, we will use two tricks. First, we will forget about absolute speeds, with respect to a given frame, and talk about relative speeds only. In this case we are allowed to sum only quantities of type v_{AB}, v_{BC}, and we denote by $v_{AB} +_g v_{BC}$ the corresponding sum v_{AC}. With this convention, the Galileo formula becomes:

$$u +_g v = u + v$$

As a second trick, observe that this Galileo formula holds in any system of units. In order now to deal with our problems, basically involving high speeds, it is convenient to change the system of units, as to have $c = 1$. With this convention our $c + v = c$ problem becomes $1 + v = 1$, and more specifically, we have the following math puzzle:

PUZZLE 2.18. *How to define a new sum for speeds in $c = 1$ units, that is, $u, v \leq 1$,*

$$(u, v) \to u +_e v$$

such that $u +_e v \simeq u + v$ at low speeds, $u, v \simeq 0$, and such that $1 +_e v = 1$?

Which does not look very difficult, and by doing the mathematics, and do not hesitate here to stop the reading, and solve this puzzle by yourself, we are led to:

THEOREM 2.19. *If we define the Einstein sum $+_e$ of relative speeds by*

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

in $c = 1$ units, then we have the formula $1 +_e v = 1$, valid for any v.

PROOF. Indeed, if we plug in $u = 1$ in the above formula, we obtain as result:

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

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

Summarizing, we solved our problem. In order now to formulate a final result, we must do some reverse engineering, by waiving the two tricks that we used. First, by getting back to usual units, $v \to v/c$, our new addition formula becomes:

$$\frac{u}{c} +_e \frac{v}{c} = \frac{\frac{u}{c} + \frac{v}{c}}{1 + \frac{u}{c} \cdot \frac{v}{c}}$$

By multiplying by c, we can write this formula in a better way, as follows:

$$u +_e v = \frac{u + v}{1 + uv/c^2}$$

In order now to finish, it remains to get back to absolute speeds, as in Proposition 2.17. And by doing so, we are led to the following result:

THEOREM 2.20. *If we sum the speeds according to the Einstein formula*

$$v_{AC} = \frac{v_{AB} + v_{BC}}{1 + v_{AB}v_{BC}/c^2}$$

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

$$v_{AC} \simeq v_{AB} + v_{BC}$$

and if we have $v_{AB} = c$ or $v_{BC} = c$, the resulting sum is $v_{AC} = c$.

PROOF. We have two assertions here, which are both clear, as follows:

(1) Regarding the first assertion, if we are at low speeds, $v_{AB}, v_{BC} << c$, the correction term $v_{AB}v_{BC}/c^2$ dissapears, and we are left with the Galileo formula, as claimed:

$$v_{AC} = \frac{v_{AB} + v_{BC}}{1 + v_{AB}v_{BC}/c^2} \simeq \frac{v_{AB} + v_{BC}}{1 + 0} = v_{AB} + v_{BC}$$

(2) As for the second assertion, this follows from the above discussion, but let us doublecheck this, to make sure that we have made no mistake. With $v_{AB} = c$ we get:

$$v_{AC} = \frac{c + v_{BC}}{1 + v_{BC}/c} = c$$

Similarly, assuming $v_{BC} = c$, the resulting speed that we get is:

$$v_{AC} = \frac{v_{AB} + c}{1 + v_{AC}/c} = c$$

Thus, no mistakes in our mathematics, and we are done. □

As a conclusion, we solved the problem, mathematically. All this is very nice, and fits with the Einstein principles from Fact 2.15, and to be more precise, fixes the laws of motion, as to make them fit with that principles. But the problem is now, does all this math really correspond to physics? And the answer here is fortunately yes, due to:

FACT 2.21. *The Einstein speed addition formula*

$$v_{AC} = \frac{v_{AB} + v_{BC}}{1 + v_{AB}v_{BC}/c^2}$$

is correct, physically speaking.

This is of course a physics fact, based on many things, and more on this later. Now with this discussed, let us do some further mathematics. We first have:

PROPOSITION 2.22. *The error term in the Galileo formula, $\varepsilon = u +_g v - u +_e v$, is*

$$\varepsilon \simeq \frac{(u+v)uv}{c^2}$$

which at small speeds, from everyday life, is truly tiny, as follows:

(1) *At $u, v \simeq 10$ m/s, running speed, we have $\varepsilon \simeq 2 \times 10^{-16}$.*
(2) *At $u, v \simeq 100$ m/s, driving speed, we have $\varepsilon \simeq 2 \times 10^{-13}$.*
(3) *At $u, v \simeq 1000$ m/s, bullet speed, we have $\varepsilon \simeq 2 \times 10^{-10}$.*

PROOF. Regarding the general estimate in the statement, this comes from the following computation, based on the formulae in Theorem 2.20:

$$\begin{aligned}
\varepsilon &= u +_g v - u +_e v \\
&= u + v - \frac{u+v}{1 + uv/c^2} \\
&= (u+v) \cdot \frac{uv}{c^2 + uv} \\
&\simeq \frac{(u+v)uv}{c^2}
\end{aligned}$$

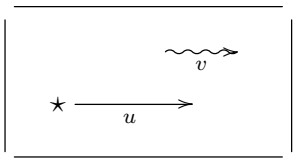
As for the numerics, at $u, v \simeq 10$ m/s, running speed, we have indeed:

$$\varepsilon \simeq \frac{20 \times 100}{300,000,000^2} = \frac{20 \times 100}{9 \times 10^{18}} \simeq 2 \times 10^{-16}$$

As for the cases $u, v \simeq 100$ m/s, driving speed, and $u, v \simeq 1000$ m/s, bullet speed, no need for new computations here, ε being of degree 3 with respect to $u \simeq v$. □

Next, we would like to say a word about physics and phenomenology. Let us recall the following key experiment of Fizeau, which according to Einstein himself [28] was the main inspiration and verification of his relativity theory, at the early stages:

EXPERIMENT 2.23 (Fizeau, 1851). *Assume that light moves through a liquid at speed $u < c$. Then, when this liquid moves through a tube at speed $v > 0$,*



the observed speed of light is not the Galilean $u +_g v = u + v$, but rather

$$u +_f v = u + v\left(1 - \frac{1}{n^2}\right)$$

where $n = c/u$ is the index of refraction of the liquid.

You must agree with me that this looks very good, especially in what regards the first part, with the observed speed by Fizeau being not the Galilean one, $u +_f v \neq u +_g v$, but then with the second part too, with Fizeau's sum $u +_f v$ looking quite similar to the Einstein sum $u +_e v$. So, let us do now the math, and compare what Fizeau and Einstein say. The result here, which is certainly a success, if you are a bit familiar with the difficulties of experimental physics, say via daily cooking at home, is as follows:

THEOREM 2.24. *The Fizeau speed summation, which in $c = 1$ units is*

$$u +_f v = u + v - u^2 v \simeq (u + v)(1 - uv)$$

is compatible with the Einstein speed summation, which in $c = 1$ units is

$$u +_e v = \frac{u + v}{1 + uv} \simeq (u + v)(1 - uv)$$

with the approximations coming from $u >> v$, and from $1/(1 + x) \simeq 1 - x$.

PROOF. This is something rather self-explanatory, but let us work out the details. In $c = 1$ units the index of refraction of the liquid is $n = 1/u$, and we have:

$$\begin{aligned} u +_f v &= u + v\left(1 - \frac{1}{(1/u)^2}\right) \\ &= u + v(1 - u^2) \\ &= u + v - u^2 v \\ &\simeq u + v - u^2 v - uv^2 \\ &= (u + v)(1 - uv) \end{aligned}$$

To be more precise here, we have used, for the approximation at the end:

$$v << u \quad \implies \quad uv^2 << u, v, u^2v$$

As for the processing of the Einstein formula, this simply uses $1/(1+x) \simeq 1-x$. □

Getting back now to the Einstein summation formula from Theorem 2.20, or rather to its more compact formulation from Theorem 2.19, that formula, while looking very simple, is in fact quite subtle, and must be handled with care. Indeed, we have:

THEOREM 2.25. *The Einstein speed summation, written in $c = 1$ units as*

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

has the following properties:

(1) $u, v < 1$ *implies* $u +_e v < 1$.
(2) $u +_e v = v +_e u$.
(3) $(u +_e v) +_e w = u +_e (v +_e w)$.
(4) *However,* $\lambda u +_e \lambda v = \lambda(u +_e v)$ *fails.*

PROOF. All these assertions are elementary, as follows:

(1) This follows from the following formula, valid for any speeds u, v:

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

Indeed, we deduce from this that $u, v < 1$ implies $u +_e v < 1$, as claimed.

(2) This is clear, coming from the following observation:

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

(3) We have indeed the following computation, for the first sum:

$$\begin{aligned}
(u +_e v) +_e w &= \frac{u+v}{1+uv} +_e w \\
&= \frac{\frac{u+v}{1+uv} + w}{1 + \frac{u+v}{1+uv} \cdot w} \\
&= \frac{u+v+w+uvw}{1+uv+uw+vw}
\end{aligned}$$

As for the second sum, this is as follows, given by the same formula:

$$\begin{aligned} u+_e(v+_e w) &= u+_e\frac{v+w}{1+vw}\\ &= \frac{u+\frac{v+w}{1+vw}}{1+u\cdot\frac{v+w}{1+vw}}\\ &= \frac{u+v+w+uvw}{1+uv+uw+vw} \end{aligned}$$

(4) This is clear, with the remark however that it works at $\lambda=-1,0,1$, or when $u=0$, $v=0$, or $u+v=0$. To be more precise, we have the following computation:

$$\begin{aligned} \lambda u+_e\lambda v=\lambda(u+_e v) &\iff \frac{\lambda u+\lambda v}{1+\lambda^2uv}=\lambda\cdot\frac{u+v}{1+uv}\\ &\iff \lambda(u+v)=0,\ \text{or}\ \lambda^2uv=uv\\ &\iff \lambda=0,\ \text{or}\ u+v=0,\ \text{or}\ \lambda^2=1,\ \text{or}\ uv=0 \end{aligned}$$

Thus, we are led to the above conclusions. □

Summarizing, the Einstein summation in 1D must be used with care, and this due to Theorem 2.25 (4). We will see later that things get even worse in higher dimensions, with $u+_e v=v+_e u$ and $(u+_e v)+_e w=u+_e(v+_e w)$ both failing there.

2e. Exercises

Welcome to physics, in 1 dimension, and as exercises, we have:

Exercise 2.26. *Meditate on the energy formula $E=mv^2/2$.*

Exercise 2.27. *Memorize the formulae of $1^p+\ldots+k^p$, at small p.*

Exercise 2.28. *Complete our proof of $1^p+\ldots+k^p\simeq k^{p+1}/(p+1)$.*

Exercise 2.29. *Learn more about Bernoulli numbers, and their properties.*

Exercise 2.30. *Learn a bit about particle physics, decay and scattering.*

Exercise 2.31. *Clarify all the details, in our study of the decay.*

Exercise 2.32. *Meditate on the origins and speed of light in 1D.*

Exercise 2.33. *Read about the experiments of Fizeau and others.*

As bonus exercise, explore a bit gravity and thermodynamics, in 1D.

CHAPTER 3

Functions, calculus

3a. Newton, derivatives

Back now to mathematics, according to the general policy for this book, the odd-numbered chapters $1-3-\ldots-15$ being mathematics, and the even-numbered ones $2-4-\ldots-16$ being physics, in view of what we did in chapter 2, with our physics in 1D, we definitely need to talk about calculus. In fact, save for some standard phenomenology discussion, and some easy computations, what we basically did in chapter 2 was to compute derivatives and integrals, which are the business of calculus.

So, what are the derivatives, integrals, and calculus itself? In answer, we have:

PRINCIPLE 3.1. *Calculus is the modern study of functions, the idea being:*

(1) *The derivative of a function is the rate of change of the function.*
(2) *The integral of a function is its average, up to a normalization.*
(3) *The differentiation and integration operations are inverse to each other.*
(4) *For learning, it is better to differentiate first, and integrate after.*

To be more precise here, (1) and (2) are obviously definitions, and more on them in a moment. Then (3) is a theorem, called fundamental theorem of calculus. And finally (4) is something more technical, the point being that, while according to (3) the differentiation and integration are interchangeable, for learning purposes, in practice things are a bit asymmetric, with the theory of differentiation being a bit easier to develop.

As a further comment on this, looking at what we did in chapter 2, we were mostly computing integrals there, instead of derivatives, somehow contradicting (4). Well, this is because the choice of the topics for chapter 2 was something biased, with me author preferring to discuss there, for various reasons, physics involving easy integrals, instead of physics involving easy derivatives. So, no issue here, (4) still stands, trust me.

But probably too much talking. Getting started, with derivatives, we have:

FACT 3.2. *Newton invented the derivatives for formulating the laws of motion as*

$$v = x' \quad , \quad a = v''$$

where x, v, a are the position, speed, acceleration, and prime is the time derivative.

To be more precise, the variable in Newton's physics is time $t \in \mathbb{R}$, playing the role of the variable $x \in \mathbb{R}$ generally used in mathematics. 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 our regular mathematical business, in this odd-numbered chapter 3, do we have any purely mathematical reasons for introducing the derivatives? In answer, sure yes. Indeed, in mathematics we are generally interested in the 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:

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

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

DEFINITION 3.3. *A function $f : \mathbb{R} \to \mathbb{R}$ is called differentiable at x when*

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}$$

called derivative of f at that point x, exists.

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

$$\lim_{h \to 0} [f(x + h) - f(x)] = 0$$

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 3.4. *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 $f(x) = |x|$, we first have:

$$\lim_{h \searrow 0} \frac{|0 + h| - |0|}{h} = \lim_{h \searrow 0} \frac{h - 0}{h} = 1$$

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

$$\lim_{h \nearrow 0} \frac{|0 + h| - |0|}{h} = \lim_{h \nearrow 0} \frac{-h - 0}{h} = -1$$

Thus, the limit in Definition 3.3 does not converge, so we have our counterexample. □

Generally speaking, the last assertion in Proposition 3.4 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 $f(x+h) \simeq f(x)$. We can now do this, using the derivative, as follows:

THEOREM 3.5. *Assuming that f is differentiable at x, we have:*

$$f(x+h) \simeq f(x) + f'(x)h$$

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:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

By multiplying by h, we obtain that we have, once again in the $h \to 0$ limit:

$$f(x+h) - f(x) \simeq f'(x)h$$

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 3.6. *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 $p \in \mathbb{N}$, that we already met in chapter 2, in the context of our decay computations there, we can use the binomial formula, which gives, as desired:

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

(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 $a=(x+h)^{m/n}$ and $b=x^{m/n}$. We obtain, as desired:

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

(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:

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

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

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

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:

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

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

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

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 3.7. *We have the following results:*

(1) $(e^x)'=e^x$.
(2) $(\log x)'=1/x$.

PROOF. The computations here are simpler, the idea being as follows:

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

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

(2) 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 (1), by taking the logarithm:

$$\begin{aligned}(\log x)' &= \lim_{h\to 0}\frac{\log(x+h)-\log x}{h}\\ &= \lim_{h\to 0}\frac{\log(1+h/x)}{h}\\ &= \frac{1}{x}\end{aligned}$$

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

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

THEOREM 3.8. *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. This is something that we already met, and in a slightly more general form, involving a parameter λ, in the context of our decay computations from chapter 2, but always good to talk about it again. Consider a power series satisfying $f'=f$ and $f(0)=1$. Due to $f(0)=1$, the first term must be 1, so our function must look as follows:

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

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

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

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

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

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

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

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

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

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

We will be back to this, with an even better characterization of exp, later in this chapter. Now let us develop some theory. We have the following key statement:

THEOREM 3.9. *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:

$$\begin{aligned}
(f + g)'(x) &= \lim_{h \to 0} \frac{(f + g)(x + h) - (f + g)(x)}{h} \\
&= \lim_{h \to 0} \left(\frac{f(x + h) - f(x)}{h} + \frac{g(x + h) - g(x)}{h} \right) \\
&= \lim_{h \to 0} \frac{f(x + h) - f(x)}{h} + \lim_{h \to 0} \frac{g(x + h) - g(x)}{h} \\
&= f'(x) + g'(x)
\end{aligned}$$

(2) This follows from definitions too, the computation, by using the more convenient formula $f(x + h) \simeq f(x) + f'(x)h$ as a definition for the derivative, being as follows:

$$\begin{aligned}
(fg)(x + h) &= f(x + h)g(x + h) \\
&\simeq (f(x) + f'(x)h)(g(x) + g'(x)h) \\
&\simeq f(x)g(x) + (f'(x)g(x) + f(x)g'(x))h
\end{aligned}$$

Indeed, we obtain from this that the derivative is the coefficient of h, 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 $f(x+h) \simeq f(x) + f'(x)h$ as a definition for the derivative:

$$\begin{aligned}(f \circ g)(x+h) &= f(g(x+h)) \\ &\simeq f(g(x) + g'(x)h) \\ &\simeq f(g(x)) + f'(g(x))g'(x)h\end{aligned}$$

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 3.10. *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 3.9 (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 3.9 (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. □

At the theoretical level now, further building on Theorem 3.5, we have:

THEOREM 3.11. *The local minima and maxima of a differentiable function $f : \mathbb{R} \to \mathbb{R}$ appear at the points $x \in \mathbb{R}$ where:*

$$f'(x) = 0$$

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

PROOF. The first assertion follows from the formula in Theorem 3.5, namely:

$$f(x + h) \simeq f(x) + f'(x)h$$

Indeed, 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:

$$f(x + h) \geq f(x) \quad , \quad \forall h \in [-\varepsilon, \varepsilon]$$

Thus we must have $f'(x)h \geq 0$ for sufficiently small h, and since this small h can be both positive or negative, this gives the following condition, as desired:

$$f'(x) = 0$$

The discussion for the local minima is similar. Finally, in what regards the converse, the simplest counterexample here is the function $f(x) = x^3$, at $x = 0$. □

In practice, Theorem 3.11 can be used in order to find the maximum and minimum of any differentiable function, and this method is best recalled as follows:

ALGORITHM 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.

Moving on, as another important consequence of Theorem 3.11, we have:

THEOREM 3.13. *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 3.11. Now in what regards our statement, due to Lagrange, this follows from Rolle, applied to the following function:

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

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

As a key consequence of Theorem 3.13, of great practical interest, we have:

THEOREM 3.14. *For a differentiable function we have*

$$f' = 0 \quad \implies \quad f = \text{constant}$$

and with the converse of this being of course true too.

PROOF. This is indeed something self-explanatory, coming from Theorem 3.13. □

Time for some applications? As something truly remarkable, we have:

THEOREM 3.15. *We have the generalized binomial formula*

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

with the generalized binomial coefficients being given by

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

valid for any exponent $p \in \mathbb{R}$, and any $|x| < 1$.

PROOF. The series in the statement f converges indeed at $|x| < 1$, thanks to our convergence results from chapter 1, which apply. Also, we have the following formula:

$$(1 + x)f'(x) = pf(x)$$

Now by using this formula, we have the following computation:

$$\big((1 + x)^{-p} f(x)\big)' = -p(1 + x)^{-p-1} f(x) + (1 + x)^{-p} f'(x) = 0$$

Thus we have $f(x) = c(1 + x)^p$, with $c = f(0) = 1$, as desired. □

As another application of Theorem 3.14, we can improve Theorem 3.8, as follows:

THEOREM 3.16. *The exponential function is the unique solution of*

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

and as a consequence, $e = f(1)$, with f being this unique solution.

PROOF. Since we have $f(0) = 1$ and $f' = f$ we conclude that we have $f \geq 1$ for $x \geq 0$, and a similar backwards argument shows that we have as well $f > 0$, for $x < 0$. In short, we have $f > 0$ over the whole $\mathbb{R}$, and in particular $f \neq 0$. But with this, we have:

$$\begin{aligned} f' = f &\implies \frac{f'}{f} = 1 \\ &\implies (\log f)' = 1 \\ &\implies \log f = x + c \\ &\implies f = \lambda e^x \end{aligned}$$

Now by using $f(0) = 1$ we conclude that we have $f(x) = e^x$, as desired. □

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 3.17. *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 3.18. *The second derivative $f''(x) \in \mathbb{R}$ is the number which:*

(1) *Expresses the growth rate of the slope $f'(y)$ 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 $f(y)$ with $y \simeq x$.*

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, with respect to time t, being as follows:

$$v = x' \quad , \quad a = v''$$

As in what regards (3) in the above, this is something more subtle, of statistical nature, that we will 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 3.19. *The second derivatives of the basic functions are as follows:*

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

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:

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

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

$$(\log x)'' = (-1/x)' = -1/x^2$$

Of course, this is not the end of the story, because these formulae remain quite opaque, and must be examined in view of Interpretation 3.18, in order to see what exactly is going on. Also, we have various compositions, and inverse 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 3.20. *Any twice differentiable function $f : \mathbb{R} \to \mathbb{R}$ is locally quadratic,*

$$f(x+h) \simeq f(x) + f'(x)h + \frac{f''(x)}{2}\,h^2$$

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:

$$f(x+h) \simeq a + bh + ch^2$$

We must have $a = f(x)$, and we also know from Theorem 3.5 that $b = f'(x)$ is the correct choice for the coefficient of h. Thus, our approximation must be as follows:

$$f(x+h) \simeq f(x) + f'(x)h + ch^2$$

In order to find the correct choice for $c \in \mathbb{R}$, observe that the function $t \to f(x+h)$ matches with $t \to f(x)+f'(x)h+ch^2$ in what regards the value at $h = 0$, and also in what regards the value of the derivative at $h = 0$. Thus, the correct choice of $c \in \mathbb{R}$ should be the one making match the second derivatives at $h = 0$, and this gives:

$$f''(x) = 2c$$

We are therefore led to the formula in the statement, namely:

$$f(x+h) \simeq f(x) + f'(x)h + \frac{f''(x)}{2}\,h^2$$

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 3.5. Now by using this, if we denote by $\varphi(h) \simeq P(h)$ the formula to be proved, we have:

$$\begin{aligned}
\frac{\varphi(h) - P(h)}{h^2} &\simeq \frac{\varphi'(h) - P'(h)}{2h}\\
&\simeq \frac{\varphi''(h) - P''(h)}{2}\\
&= \frac{f''(x) - f''(x)}{2}\\
&= 0
\end{aligned}$$

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

The above result substantially improves Theorem 3.5, and there are many applications of it. As a first such application, justifying Interpretation 3.18 (3), we have the following statement, which is a bit heuristic, but we will call it however Proposition:

PROPOSITION 3.21. *Intuitively speaking, the second derivative $f''(x) \in \mathbb{R}$ computes how much different is $f(x)$, compared to the average of $f(y)$, with $y \simeq x$.*

PROOF. As already mentioned, this is something a bit heuristic, but which is good to know. Let us write the formula in Theorem 3.20, as such, and with $h \to -h$ too:

$$f(x+h) \simeq f(x) + f'(x)h + \frac{f''(x)}{2}\,h^2$$

$$f(x-h) \simeq f(x) - f'(x)h + \frac{f''(x)}{2}\,h^2$$

By making the average, we obtain the following formula:

$$\frac{f(x+h) + f(x-h)}{2} \simeq f(x) + \frac{f''(x)}{2}\,h^2$$

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

Back to rigorous mathematics, we can improve as well Theorem 3.11, as follows:

THEOREM 3.22. *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 3.20, which in the case $f'(x) = 0$ reads:

$$f(x+h) \simeq f(x) + \frac{f''(x)}{2}\,h^2$$

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

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

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

There are some comments to be made in relation with Algorithm 3.12 as well. Normally that algorithm stays strong, because Theorem 3.22 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 3.22 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 3.23. *Assuming that $f : \mathbb{R} \to \mathbb{R}$ is differentiable n times, we have*

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

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

PROOF. Consider the function to be approximated, namely $\varphi(h) = f(x+h)$, and 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 h, of the following type, approximating our function:

$$P(h) = a_0 + a_1 h + \ldots + a_n h^n$$

The natural conditions to be imposed are those stating that P and φ should match at $h = 0$, 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:

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

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 3.20. To be more precise, if we denote by $\varphi(h) \simeq P(h)$ the approximation to be proved, we have:

$$\begin{aligned}
\frac{\varphi(h) - P(h)}{h^n} &\simeq \frac{\varphi'(h) - P'(h)}{nh^{n-1}} \\
&\simeq \frac{\varphi''(h) - P''(h)}{n(n-1)h^{n-2}} \\
&\vdots \\
&\simeq \frac{\varphi^{(n)}(h) - P^{(n)}(h)}{n!} \\
&= \frac{f^{(n)}(x) - f^{(n)}(x)}{n!} \\
&= 0
\end{aligned}$$

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

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

PROPOSITION 3.24. *For a polynomial of degree n, the Taylor approximation*

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

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:

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

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. □

In order to further comment now on Theorem 3.23, which remains something quite subtle, it is perhaps time to clarify our meaning of $\simeq$. We have been using this sign, since the beginning of this book, for approximation in a general, intuitive sense. However, since in Theorem 3.23 we have all kinds of infinitesimals appearing, namely $h,h^2,\ldots,h^n$, we must invent something better. And here, the best is to state things as follows:

THEOREM 3.25. *Assuming that $f:\mathbb{R}\to\mathbb{R}$ is differentiable n times we have*

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

with o being the Landau symbol, meaning $o(s)/s\to 0$, with $s\to 0$.

PROOF. This is indeed something self-explanatory, based on Theorem 3.23. □

As a last comment about Theorem 3.23, an interesting situation, which appears quite often, is that when f is infinitely differentiable. Here the result is as follows:

THEOREM 3.26. *Assuming that $f:\mathbb{R}\to\mathbb{R}$ is infinitely differentiable, we have*

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

for any $n\in\mathbb{N}$, according to the Taylor theorem. However, the asymptotic formula

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

might hold or not, depending on f, and generically, does not hold.

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

(1) To start with, the first assertion is something that we know well.

(2) The second assertion is something more subtle, with the examples there abounding. However, we have as well counterexamples, with a standard counterexample being:

$$f(x)=\begin{cases}e^{-1/x^2} & (x\neq 0)\\ 0 & (x=0)\end{cases}$$

Indeed, for this function we have the following estimate, valid for any $n\in\mathbb{N}$:

$$f(h)=(e^{1/h^2})^{-1}\leq\left(\frac{1/h^{2n}}{n!}\right)^{-1}=n!h^{2n}=o(h^n)$$

Thus f is infinitely differentiable at 0, with all its derivatives vanishing there, and so its Taylor series at 0 is the null series, which cannot be equal to f itself. That is, f is designed not to take off from 0, but it manages however to take off, very slowly.

(3) In what regards now the very last claim, this is something more technical, an intuitive explanation here being that there should be more functions $f:\mathbb{R}\to\mathbb{R}$, even taken infinitely differentiable, than series $\psi=\sum c_kh^k$. And in practice, up to you to learn here how to count such beasts, as an exercise, and reach to the above conclusion. □

In relation now with the local extrema, and getting back to our usual, informal $\simeq$ convention, in order to quickly explain what happens, we have the following result:

THEOREM 3.27. *Assuming that $f:\mathbb{R}\to\mathbb{R}$ is n times differentiable, and*

$$f(x+h)\simeq f(x)+\frac{f^{(n)}(x)}{n!}\,h^n$$

with $f^{(n)}(x)\neq 0$, this tells us if x is a local minimum or maximum of f.

PROOF. This is a quite compact statement, coming from the Taylor formula, the idea in practice being that we have an algorithm here, as follows:

(1) We can start with $n=1$, and with the following formula, that we know well:

$$f(x+h)\simeq f(x)+f'(x)h$$

Indeed, this formula tells us that when $f'(x)\neq 0$, the point x cannot be a local minimum or maximum, due to the fact that $h\to -h$ will invert the growth.

(2) In the case left, $f'(x)=0$, we switch to $n=2$, where the Taylor formula is:

$$f(x+h)\simeq f(x)+\frac{f''(x)}{2}\,h^2$$

And here, when $f''(x)<0$ we have a local maximum, and when $f''(x)>0$ we have a local minimum. As for the remaining case, $f''(x)=0$, things here remain open.

(3) In the case left, $f''(x) = 0$, we switch to $n = 3$, where the Taylor formula is:

$$f(x + h) \simeq f(x) + \frac{f'''(x)}{6}\, h^3$$

But this solves the problem in the case $f'''(x) \neq 0$, because here we cannot have a local minimum or maximum, due to $h \to -h$, which switches growth. As for the remaining case, $f'''(x) = 0$, things here remain open, and we have to go at higher order.

(4) Summarizing, we have a recurrence method for solving our problem. In order to comment now on what happens at the n-th step, let us write, as in the statement:

$$f(x + h) \simeq f(x) + \frac{f^{(n)}(x)}{n!}\, h^n$$

Then, when n is even, if $f^{(n)}(x) < 0$ we have a local maximum, and if $f^{(n)}(x) > 0$ we have a local minimum. As for the case where n is odd, here with $f^{(n)}(x) \neq 0$ we cannot have a local minimum or maximum, due to $h \to -h$ which switches growth.

(5) And so on, until the algorithm stops, either due to $f^{(n)}(x) \neq 0$, as it would be ideal, solving our problem, or due to the fact that $f^{(n)}$ is no longer differentiable at x, telling us that we have to use some alternative methods, or due the fact that we are facing a tricky function, resisting our algorithm until the very end, after $n = \infty$ steps, such as the function $f(x) = e^{-1/x^2}$, that we met in the proof of Theorem 3.30. □

Finally, and getting now to more concrete things, let us compute the Taylor series of the basic functions that we know. We have here the following result:

THEOREM 3.28. *We have the following formulae,*

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

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

PROOF. There are two assertions here, the proofs being as follows:

(1) Regarding the first assertion, here the needed formulae, which lead to the formulae in the statement for the corresponding Taylor series, are as follows:

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

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

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

(2) As for the fact that the formulae in the statement extend beyond the small x setting, coming from Taylor series, this is something that we know from chapter 1. □

And with this, end of our learning regarding derivatives. The above material was of course quite theoretical, yes I know, but do not worry, we will do some heavy physics in chapter 4 and afterwards, with complicated functions and computations, using this.

3d. Riemann, integrals

Following Riemann, let us discuss now the integration of functions, and its relation with differentiability. The definition of integrals is something very simple, as follows:

DEFINITION 3.29. *The integral of a function f on $[a,b]$ is its normalized average*

$$\int_a^b f(x)dx = (b-a) \times A(f)$$

that is, average $A(f)$ times the length of the interval $b-a$. Equivalently,

$$\int_a^b f(x)dx = (b-a) \times \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f\left(a + \frac{b-a}{n} \cdot k\right)$$

which can stand as a formal definition for the integral.

As a first comment on this, this is rather the 1D perspective on the Riemann integral, suitable for this book, at the present point of writing, which is a bit probabilistic in nature, and which would probably make Riemann himself, who was a geometer, turn in his grave. So, before going forward, a bit of discussion must be made:

(1) To start with, our definition is very beautiful, and fits with our physics computations from chapter 2. Indeed, we got there into computing all sorts of averages of functions, and Definition 3.29 brings an abstract framework, for such computations.

(2) Next, you might wonder what that $b-a$ normalization factor stands for. In answer, this is something quite subtle, the idea being that the $b-a$ factor is there as for the fundamental theorem of calculus, that we will learn soon, to work fine.

(3) This being said, doing mathematics without that $b-a$ factor makes sense too, and this is what professional probabilists do, and with the integral, which is now the plain average $A(f)$, being more poetically denoted $E(f)$, and called "expectation".

(4) Finally, getting to Riemann, from a 2D perspective the integral of f is the signed area below its graph, with this coming by drawing a picture, and approximating the area with rectangles. And with this explaining, again, the need for the $b-a$ factor.

So these were our comments, hope you got it, the integral is something quite subtle, having several possible interpretations, which all must be known. In what follows we will rather use the probabilistic approach, but with the Riemann $b-a$ normalization, as in Definition 3.29, with this best fitting with what we need in this chapter, and Part I.

By the way, talking interpretations of the integral, here is one more, coming by further building on our probabilistic approach from Definition 3.29:

PROPOSITION 3.30. *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 the points $x_1, \ldots, x_n \in [a, b]$ being chosen at random.

PROOF. We recall from Definition 3.29 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. □

With this discussed, time perhaps for some theory and computations? At the theoretical level we first have the following result, which is something very useful:

PROPOSITION 3.31. *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. □

Next, we have the following result, called fundamental theorem of calculus:

THEOREM 3.32. *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 mean value property. Indeed, we have:

$$\frac{F(x+h)-F(x)}{h}=\frac{1}{h}\int_x^{x+h}f(x)dx$$

On the other hand, our function f being continuous, by using the mean value property from Proposition 3.31, we can find a number $c\in[x,x+h]$ such that:

$$\frac{1}{h}\int_x^{x+h}f(x)dx=f(c)$$

Thus, putting our formulae together, we conclude that we have:

$$\frac{F(x+h)-F(x)}{h}=f(c)$$

Now with $h\to 0$, no matter how the number $c\in[x,x+h]$ varies, one thing that we can be sure about is that we have $c\to x$. Thus, by continuity of f, we obtain:

$$\lim_{h\to 0}\frac{F(x+h)-F(x)}{h}=f(x)$$

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 3.33. *Given a differentiable 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. This is something which follows from Theorem 3.32, and is in fact equivalent to it. Indeed, given $F:\mathbb{R}\to\mathbb{R}$ as above, consider the following function:

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

By using Theorem 3.32 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 3.34. *We have the following integration formulae,*

$$\int_a^b x^p dx = \frac{b^{p+1} - a^{p+1}}{p+1} \quad , \quad \int_a^b \frac{1}{x}\, dx = \log\left(\frac{b}{a}\right)$$

$$\int_a^b e^x dx = e^b - e^a \quad , \quad \int_a^b \log x\, dx = b\log b - a\log a - b + a$$

all obtained, in case you ever forget them, via the fundamental theorem of calculus.

PROOF. We already know most of these formulae, obtained via hard work in chapter 2, sorry for that, but the best is to redo everything, using Theorem 3.33:

(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 $F(x) = e^x$ we have $F'(x) = e^x$, and we get, as desired:

$$\int_a^b e^x\, dx = e^b - e^a$$

(4) 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 3.35. *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 3.34 in a more digest form, as follows:

THEOREM 3.36. *We have the following formulae for primitives,*

$$\int x^p = \frac{x^{p+1}}{p+1} \quad , \quad \int \frac{1}{x} = \log x$$

$$\int e^x = e^x \quad , \quad \int \log x = x \log x - x$$

allowing us to compute the corresponding definite integrals too.

PROOF. Here the various formulae in the statement follow from Theorem 3.34, and the last assertion comes from the integration formula given after Definition 3.35. □

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

THEOREM 3.37. *We have the formula*

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

called integration by parts.

PROOF. This follows by integrating the Leibnitz formula for derivatives, 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 the product function 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 3.38. *We have the change of variable formula*

$$\int_a^b f(x)dx = \int_c^d f(\varphi(y))\varphi'(y)dy$$

where $c = \varphi^{-1}(a)$ and $d = \varphi^{-1}(b)$.

PROOF. This follows with $f = F'$, from the following differentiation rule:

$$(F\varphi)'(y) = F'(\varphi(y))\varphi'(y)$$

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

As a theoretical application now of our integration theory, we have:

THEOREM 3.39. *Given a n times differentiable function $f : \mathbb{R} \to \mathbb{R}$, we have*

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

called Taylor formula with integral formula for the remainder.

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

(1) At $n = 1$ the formula in the statement is as follows, and certainly holds, due to the fundamental theorem of calculus, which gives $\int_x^{x+h} f'(s)ds = f(x+h) - f(x)$:

$$f(x+h) = f(x) + \int_x^{x+h} f'(s)ds$$

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

$$f(x+h) = f(x) + f'(x)h + \int_x^{x+h} f''(s)(x+h-s)ds$$

As a first observation, this formula holds indeed for the linear functions, where we have $f(x+h)=f(x)+f'(x)h$, and $f''=0$. So, let us try $f(x)=x^2$. Here we have:

$$f(x+h)-f(x)-f'(x)h=(x+h)^2-x^2-2xh=h^2$$

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

$$\begin{aligned}\int_x^{x+h}f''(s)(x+h-s)ds&=2\int_x^{x+h}(x+h-s)ds\\&=2h(x+h)-2\int_x^{x+h}sds\\&=2h(x+h)-((x+h)^2-x^2)\\&=2hx+2h^2-2hx-h^2\\&=h^2\end{aligned}$$

(3) Still at $n=2$, 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 $n=2$, and we will leave this as an instructive exercise. □

3e. Exercises

We have learned many interesting things here, and as exercises, we have:

EXERCISE 3.40. *Learn a bit about calculus history, Netwon, Leibnitz and the others.*

EXERCISE 3.41. *Compute the first derivatives of some functions of your choice.*

EXERCISE 3.42. *Compute as well minima and maxima, of functions of your choice.*

EXERCISE 3.43. *Meditate on the binomial formula, at negative integer exponents.*

EXERCISE 3.44. *Clarify what we said, regarding the interpretations of $f''(x)$.*

EXERCISE 3.45. *Try to find some similar possible interpretations of $f'''(x)$.*

EXERCISE 3.46. *Learn more about Monte Carlo integration, and its use.*

EXERCISE 3.47. *Compute more Riemann sums, with enthusiasm and pleasure.*

As bonus exercise, review what we did in this chapter, from a 2D perspective.

CHAPTER 4

Basic mechanics

4a. Gravity basics

Time now to get back to physics, with a substantial improvement of what we did in chapter 2, and with an introduction to modern physics in general, by using the mathematics that we just learned in chapter 3, namely functions and calculus.

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

FACT 4.1. *Newton invented calculus for formulating the laws of mechanics as*

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

with F being the force, m the mass, and x, v, a the position, speed and acceleration.

So, this will be our starting point for the considerations here. Regarding now the explanations for the above, all this is quite intuitive, the idea being as follows:

(1) To start with, the variable in Newton's physics is time $t \in \mathbb{R}$, playing the role of the variable $x \in \mathbb{R}$ that we use in general in mathematics. And we are looking at a particle whose position is described by a function $x = x(t)$.

(2) But then, it is quite clear that the speed of this particle should be described by the first derivative $v = \dot{x}(t)$, rate of change of the position, with the standard convention, in Newtonian physics, that the dots denote the time derivatives.

(3) Similarly, the acceleration of the particle should be described by the derivative $a = \dot{v}(t)$, rate of change of the speed. In relation with this, observe that by combining this with (2), the acceleration is the second derivative of the position, $a = \ddot{x}(t)$.

(4) Getting now to mechanics, appearing when our particle is subject to a force F, it is quite clear that we should have $F \sim m$ and $F \sim a$, and so $F \sim ma$, and nothing else. Thus, by suitably normalizing the force, we can say that we have $F = ma$.

Summarizing, everything in Fact 4.1 is quite intuitive, but with all this being of course a physics Fact, and not a mathematical Theorem. Getting now to the consequences of this, as a first application, we can talk about energy, the result being as follows:

THEOREM 4.2. *For a particle of mass m, subject to a force F, if we define its kinetic and potential energy by the following formulae, the prime being a space derivative,*

$$T = \frac{mv^2}{2} \quad , \quad V' = -F$$

and so with V being unique up to an additive constant, the total energy

$$E = T + V$$

is constant over time. The same happens for a system of N non-colliding particles.

PROOF. Many things can be said here, the idea being as follows:

(1) To start with, we already talked about energy in chapter 2, in the context of the linear motion, in the absence of forces. We were saying at that time that $T = mv^2/2$ must be constant, and now we have a better explanation of this, coming from:

$$\dot{T} = mv\dot{v} = mva = 0$$

By the way, altough this computation yields 0, we can better understand now why the kinetic energy is defined as $T = mv^2/2$, instead of $T = mv^2$, the reason for this coming from the fact that we would like to have $\dot{T} = mv\dot{v}$, as above, instead of $\dot{T} = 2mv\dot{v}$.

(2) Getting now to what the statement says, with $T = mv^2/2$ as before, we have:

$$\dot{T} = mv\dot{v} = mva = Fv$$

Now by integrating, also with respect to t, this gives the following formula:

$$T = \int Fvdt = \int F\dot{x}dt = \int Fdx$$

Next, by taking the derivative with respect to the space variable x, we obtain:

$$T' = F$$

(3) But this suggests to define the potential energy V by the following formula, up to an additive 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:

$$E' = T' + V' = F - F = 0$$

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

All this was quite theoretical, and getting now to the real thing, forces appearing in the real life, and the subsequent mechanics of particles, let us talk about gravity.

So, what is gravity? In answer I guess, as you certainly know, gravity is when you stumble and fall. But with this quick and unpleasant phenomenon, unfortunately, not allowing you to carefully measure the magnitude of the force F acting on you.

However, gravity can be successfully observed by looking at the skies, with a telescope. For instance observing a comet heading to the Sun leads to the conclusion that F grows when the distance d between the comet and the Sun decreases, and a more careful study even gives an exact formula, namely $F \sim 1/d^2$. We are led in this way to:

FACT 4.3 (Newton). *The force of attraction between two bodies of masses m_1, m_2, having distance $d > 0$ between them, is given by*

$$F = -G \cdot \frac{m_1 m_2}{d^2}$$

where $G \simeq 6.674 \times 10^{-11}$ is a constant. This force alters the trajectory of one body with respect to another according to the formula $F = ma$, with a being the acceleration.

To be more precise here, as explained above, various observations from astronomy, performed over the centuries, led to the conclusion that we have $F \sim 1/d^2$. Now since we certainly have $F \sim m_1$ and $F \sim m_2$ too, and also since F is attractive, and so technically $F < 0$, we must have a formula of type $F = -G \cdot m_1 m_2/d^2$, as above.

Summarizing, gravity mechanism understood, thanks to the work of Newton, and many others before him, up the value of the constant G. And, regarding the precise value of G, coming of course from the above-mentioned astronomy observations, we have:

FACT 4.4. *The gravitational constant is given by*

$$G = 6.67430(15) \times 10^{-11}$$

in standard units, with the 15 standing for the standard uncertainty.

To be more precise, we are using here standard units (meters, kilograms, seconds), so our constant is dimensionless, and with the known experiments leading to the above figure. As for the standard uncertainty, this is something of statistical nature.

In practice, in order to get more familiar with the numerics, Fact 4.4 tells us that the force of attraction between two bodies having masses $m_1 = m_2 = 1\,\text{kg}$, having distance $d = 1\,\text{m}$ between them, is $F = 6.67430(15) \times 10^{-11}\,\text{N}$. Which itself means that, assuming that m_1 is fixed, m_2 is subject to an acceleration $a = 6.67430(15) \times 10^{-11}\,\text{m/s}^2$.

Equivalently, for being a bit more reasonable with our figures, Fact 4.4 tells us that given two asteroids having masses $m_1 = m_2 = 10^{15}\,\text{kg}$, with this corresponding to the mass

of a cube of ice having side $r = 10\,\mathrm{km}$, which are about to meet, having distance between them $d = 10^5\,\mathrm{m} = 100\,\mathrm{km}$, the force of attraction on m_1 is $F = 6.67430(15) \times 10^9\,\mathrm{N}$. Which itself means that, when assuming that m_1 is fixed, m_2 is subject to an acceleration $a = 6.67430(15) \times 10^{-6}\,\mathrm{m/s^2}$. Which sounds good, our asteroids will certainly meet.

As a philosophical comment now, I must say that Fact 4.4 looks to me more exciting than Fact 4.3. And by this I mean, Fact 4.3 looks like something quite standard, leading to some mathematics, which does not look that complicated, and that we can certainly solve, with some patience. While Fact 4.4 is something completely alien, what is the formula of that constant G, will us be ever able to say something about it?

Hope you get my point here, these physics constants are not a joke, hiding more mysteries than the mathematics. So, let us formulate this as a conclusion, as follows:

CONCLUSION 4.5 (mathematician's take). *The physics constants are what is most interesting, in physics. The rest being basically mathematics.*

And more such comments later in this book, every time we will meet a new physics constant. I mean, knowing myself, I will certainly make some advertisement for these.

But probably too much talking, let us get to work. Forgetting about Fact 4.4, maybe later about this, once we will know 1000 times more mathematics and physics, what we have in Fact 4.1 and Fact 4.3 leads to the following concise, clear statement:

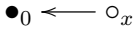
THEOREM 4.6. *The equation of a gravitational free fall, in* 1 *dimension, is*

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

with $K = GM$, where M is the attracting mass, and G is the gravitational constant.

PROOF. Assume indeed that we have a free falling object, in 1 dimension, represented here vertically, for making the link with what we know, from the real life:

In order to reach to calculus as we know it, it is better to perform a rotation, as to have this happening horizontally. By doing this, and assuming that M is fixed at 0, our picture becomes as follows, the attached numbers being now the coordinates:



According now to the physics, Fact 4.1 and Fact 4.3, the gravitational force exterted by M, which is fixed at 0 in our formalism, on the object m which moves, is subject to

the following equations, with x, v, a being the position, speed and acceleration of m:

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

But, forgetting now about F, v, a, this leads to the following equation for x:

$$m\ddot{x} = m\dot{v} = ma = F = -G \cdot \frac{Mm}{x^2} = -\frac{Km}{x^2}$$

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

In practice now, solving the equation in Theorem 4.6 is not exactly an easy task, and more on this later. In the meantime, let us begin with the following approximate result, which is something of certain practical interest, for instance here on Earth:

THEOREM 4.7. *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 know that the equation of motion is as follows, with $K = GM$:

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

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, $g = 9.806$.

At a more advanced level, let us add now an arbitrary initial speed $v_0 = v$ to the above situation. We obtain in this way the following generalization of Theorem 4.7:

THEOREM 4.8 (update). *In the context of a free fall from distance $x_0 = R >> 0$, with initial velocity vector v_0, the equation of the trajectory is*

$$x \simeq R + v_0 t - \frac{gt^2}{2}$$

where $g = GM/R^2$ being, as usual, the gravity of M, at distance R from it.

PROOF. This is something straightforward, because in order to find the equation of motion, we can just redo the computations from the proof of Theorem 4.7, with now looking for a general solution of type $x \simeq R + v_0 t + ct^2$, and we get, as stated above:

$$x \simeq R + v_0 t - \frac{gt^2}{2}$$

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 4.7, and add an extra $v_0 t$ term to it. □

As a comment now on this, the formula in Theorem 4.8 is in fact the exact formula of a free fall, under the assumption that the attractive force F is constant, and with this latter assumption being usually called "uniform gravity case". Let us record this as:

THEOREM 4.9 (version). *In the context of a free fall from distance $x_0 = R >> 0$, under uniform gravity, with initial velocity vector v_0, the equation of the trajectory is*

$$x = R + v_0 t - \frac{gt^2}{2}$$

where $g = GM/R^2$ being, as usual, the gravity of M, at distance R from it

PROOF. We more or less know this from Theorem 4.8, but the best is to redo everything. Under the uniform gravity assumption in the statement, the equations are:

$$F = -gm \quad , \quad F = ma \quad , \quad a = \dot{v} \quad , \quad v = \dot{x}$$

By forgetting now about F, v, a, the equation of motion is as follows:

$$\ddot{x} = \dot{v} = a = \frac{F}{m} = -g$$

But this latter equation, with $x_0 = R$ and $\dot{x}_0 = v_0$, is easy to solve, as follows:

$$\begin{aligned}\ddot{x} = -g &\iff \dot{x} = v_0 - gt\\ &\iff x = R + v_0 t - \frac{gt^2}{2}\end{aligned}$$

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

Finally, still in the uniform gravity context, we can talk about kinetic and potential energy, and the conservation of their sum, as in Theorem 4.2, as follows:

THEOREM 4.10. *In the context of a free fall from distance $x_0 = R >> 0$, up to distance $x_{fin} = R - h$, under uniform gravity, with initial velocity vector v_0, the kinetic energy is*

$$T = \frac{mv^2}{2} = \frac{m(v_0 - gt)^2}{2}$$

and the standard potential energy, normalized as to be 0 on the ground, is:

$$V = mg(x - x_{fin}) = mg\left(h + v_0 t - \frac{gt^2}{2}\right)$$

Also, the total energy $E = T + V$ is indeed constant, $E = mv_0^2/2 + mgh$.

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

(1) We use the trajectory formula found in Theorem 4.9, namely:

$$x = R + v_0 t - \frac{gt^2}{2}$$

In what regards the formula of the kinetic energy, we have indeed:

$$T = \frac{mv^2}{2} = \frac{m\dot{x}^2}{2} = \frac{m(v_0 - gt)^2}{2}$$

(2) Regarding now the potential energy, this must be given by $V' = -F = mg$, up to a constant. Thus $V = mgx - c$, and with the standard choice $c = mg(R - h)$, which is there for having $V = 0$ on the ground, where $x = x_{fin} = R - h$, we obtain:

$$V = mg(x - R + h) = mg\left(h + v_0 t - \frac{gt^2}{2}\right)$$

(3) We can now compute the total energy, which is constant, as it should:

$$\begin{aligned}E &= \frac{m(v_0 - gt)^2}{2} + mg\left(h + v_0 t - \frac{gt^2}{2}\right)\\ &= m\left(\frac{v_0^2}{2} + \frac{g^2t^2}{2} - v_0 gt + gh + v_0 gt - \frac{g^2t^2}{2}\right)\\ &= \frac{mv_0^2}{2} + mgh\end{aligned}$$

(4) Finally, let us mention that this energy technology can be useful for many practical purposes. For instance the terminal speed v_{fin}, which is a bit complicated to compute by using Theorem 4.9, and have a look at this in order to understand what I am talking about, is very easy to compute by using energy conservation, which gives:

$$\frac{mv_{fin}^2}{2} = \frac{mv_0^2}{2} + mgh$$

Thus, we obtain the following formula, for the terminal speed:

$$v_{fin} = \pm\sqrt{v_0^2 + 2gh}$$

As an illustration here, if you throw a rock up in the sky at speed $v_0 = 10\,\mathrm{m/s}$, that rock will fall back on your head at speed $v_{fin} = -10\,\mathrm{m/s}$. In general, I will leave the discussion of the sign in the above formula to you, as an interesting exercise. □

4b. Free falls

Getting now to the real thing, true gravity as in Fact 4.3, and the study of the subsequent equation of motion from Theorem 4.6, we have here the following result:

THEOREM 4.11. *The equation of a gravitational free fall in* 1 *dimension,*

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

can be successfully studied, by computing $t = t(x)$ *instead of* $x = x(t)$*, and we have*

$$t = \sqrt{\frac{x_0^3}{2K}}\left(\sqrt{\frac{x}{x_0}\left(1 - \frac{x}{x_0}\right)} + \varphi\left(\sqrt{\frac{x}{x_0}}\right)\right)$$

with x_0 *being the initial position, and* φ *being such that* $\varphi'(x) = -1/\sqrt{1-x^2}$*.*

PROOF. Many things can be said here, the idea being as follows:

(1) To start with, the equation in the statement, $\ddot{x} = -K/x^2$, is not really solvable. As a first idea here, we can look for a solution as follows, with $T > 0$ and $\lambda > 0$:

$$x = (T - \lambda t)^p$$

In order for our equation to be satisfied, the following quantities must be equal:

$$\ddot{x} = p(p-1)\lambda^2(T - \lambda t)^{p-2} \quad , \quad -\frac{K}{x^2} = -K(T - \lambda t)^{-2p}$$

At the level of the exponent, we must have $p - 2 = -2p$, and so $p = 2/3$. As for the coefficient, the equation here is $(-2/9)\lambda^2 = -K$, so $\lambda^2 = 9K/2$, and $\lambda = 3\sqrt{K/2}$.

Summarizing, we have our particular solution, which is as follows, with $T > 0$:

$$x = \left(T - 3\sqrt{\frac{K}{2}}\, t \right)^{2/3}$$

However, this is not the correct solution for our problem, and any attempt of further perturbing this solution, as to get the correct solution, fails. In fact, there is no explicit formula for the solution of $\ddot{x} = -K/x^2$, and we will have to live with this.

(2) In order to say however something on the subject, we can trick as in the statement, by computing $t = t(x)$ instead of $x = x(t)$. Now in order to do the inversion, we will need the following standard computation, coming from the chain rule, applied twice:

$$\begin{aligned} f(g(x)) = x &\implies f'(g(x))g'(x) = 1 \\ &\implies f'(g(x)) = \frac{1}{g'(x)} \\ &\implies f''(g(x))g'(x) = -\frac{g''(x)}{g'(x)^2} \\ &\implies f''(g(x)) = -\frac{g''(x)}{g'(x)^3} \end{aligned}$$

(3) So, consider our equation, written as follows, with $f(t)$ being the position:

$$f''(t) = -\frac{K}{f(t)^2}$$

When setting $t = g(x)$, with $f(g(x)) = x$ as above, our equation becomes:

$$\begin{aligned} -\frac{g''(x)}{g'(x)^3} = -\frac{K}{x^2} &\implies \left(\frac{1}{g'(x)^2} \right)' = \left(\frac{2K}{x} \right)' \\ &\implies \frac{1}{g'(x)^2} = \frac{2K}{x} + c \\ &\implies g'(x) = -\frac{1}{\sqrt{2K/x + c}} \end{aligned}$$

Here we have chosen the above $-$ sign at the end, when extracting the root, because the function $g = g(x)$, expressing time in terms of position, must be decreasing.

(4) Next, at the initial position x_0 the time must vanish, $g(x_0) = 0$, and we can expect its derivative to explode there, $g'(x_0) = -\infty$. Thus the constant c appearing in the above must be $c = -2K/x_0$, and our equation takes the following form:

$$g'(x) = -\frac{1}{\sqrt{2K}} \cdot \frac{1}{\sqrt{1/x - 1/x_0}}$$

(5) Next, with the change of variables $x = x_0 y$, this equation becomes:

$$g'(x_0 y) = -\frac{x_0}{\sqrt{2K}} \cdot \frac{1}{\sqrt{1/(x_0 y) - 1/x_0}} = -\sqrt{\frac{x_0^3}{2K}} \cdot \frac{1}{\sqrt{1/y - 1}}$$

(6) Now in order to solve this latter equation, observe that, assuming that we have a function φ satisfying $\varphi'(x) = -1/\sqrt{1-x^2}$, as in the statement, we have:

$$\begin{aligned}\left(\sqrt{y-y^2} + \varphi(\sqrt{y})\right)' &= \frac{1-2y}{2\sqrt{y-y^2}} - \frac{1}{2\sqrt{y}} \cdot \frac{1}{\sqrt{1-y}} \\ &= -\frac{y}{\sqrt{y-y^2}} \\ &= -\frac{1}{\sqrt{1/y-1}}\end{aligned}$$

(7) Of course, this might sound a bit odd, but in any case, problem solved, and the solution of our initial equation from (3), with initial data $g(x_0) = 0$, is as follows:

$$g(x_0 y) = \sqrt{\frac{x_0^3}{2K}} \left(\sqrt{y-y^2} + \varphi(\sqrt{y})\right)$$

Now with $x = x_0 y$ we are led to the formula in the statement, namely:

$$g(x) = \sqrt{\frac{x_0^3}{2K}} \left(\sqrt{\frac{x}{x_0}\left(1 - \frac{x}{x_0}\right)} + \varphi\left(\sqrt{\frac{x}{x_0}}\right)\right)$$

Summarizing, we made it, free fall question eventually cracked. Very nice. □

With the above discussed, what is next? Finding I guess the mysterious function φ which is needed in Theorem 4.11, subject to the following formula:

$$\varphi'(x) = -\frac{1}{\sqrt{1-x^2}}$$

However, this does not look obvious, with the calculus knowledge that we have so far, so I will have the ask here our usual 1D expert, the ant. And ant says:

ANT 4.12. *Sure we have functions for everything, in our 1D world. The basic ones come via $i^2 = -1$, that's something that we teach to our babies.*

Okay, thanks ant, this looks quite interesting, you guys are certainly good at math, and all of a sudden, my baby ant apprentice status makes me feel very intelligent.

Indeed, with the help of a formal number satisfying $i^2 = -1$, here is what we can do:

THEOREM 4.13. *By using a formal number satisfying $i^2 = -1$, we can write*

$$e^{ix} = \cos x + i \sin x$$

and the functions cos, sin *and their inverses* $\cos^{-1}$, $\sin^{-1}$ *differentiate as:*

$$\cos(x)' = -\sin x \quad , \quad \sin(x)' = \cos x \ , \ \cos^{-1}(x)' = -\frac{1}{\sqrt{1-x^2}} \ , \ \sin^{-1}(x)' = \frac{1}{\sqrt{1-x^2}}$$

Also, the functions cos, sin *are subject to the formula* $\cos^2 + \sin^2 = 1$.

PROOF. This is tricky ant mathematics, the idea being as follows:

(1) To start with, once we have a formal number satisfying $i^2 = -1$, and we certainly have available such a beast, as the "formal" adjective indicates, we can write:

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

(2) Thus, we have $e^{ix} = \cos x + i \sin x$ as stated, with cos, sin being given by:

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

Observe that both series converge, by using our convergence criteria from chapter 1.

(3) Next, observe that for various purposes, including differentiation, no really need for the series, because we can use the following alternative formulae for cos, sin:

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

(4) As a first consequence now of these two latter formulae, we have the following remarkable identity relating cos and sin, called Pythagoras theorem:

$$\begin{aligned} \cos^2 x + \sin^2 x &= \left(\frac{e^{ix} + e^{-ix}}{2} \right)^2 + \left(\frac{e^{ix} - e^{-ix}}{2i} \right)^2 \\ &= \frac{e^{2ix} + e^{-2ix} + 2}{4} - \frac{e^{2ix} + e^{-2ix} - 2}{4} \\ &= 1 \end{aligned}$$

(5) Getting now to differentiation, by using the formulae in (3) we have, as stated:

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

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

Alternatively, these formulae follow as well by differentiating the series in (2).

(6) Summarizing, done with cos, sin, at least we know one thing. In what regards now the inverse functions $\cos^{-1}$, $\sin^{-1}$, things here are more tricky, because do we really know if cos, sin are invertible, on which exact intervals, and so on. And with the study here not looking obvious at all, with the 1D technology that we have, and with the ant being gone, we will prefer here to defer this discussion for later in this book, when doing 2D.

(7) Getting now to the differentiation of $\cos^{-1}$, $\sin^{-1}$, this can be done by using the chain rule. Indeed, for $\cos^{-1}$ we can use the following computation:

$$\begin{aligned}(\cos\circ\cos^{-1})'(x) &= \cos'(\cos^{-1}x)\cos^{-1}(x)'\\ &= -\sin(\cos^{-1}x)\ \cos^{-1}(x)'\end{aligned}$$

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

$$\cos^{-1}(x)' = -\frac{1}{\sin(\cos^{-1}x)}$$

But with $t = \cos^{-1}x$ we have $\cos t = x$, so the denominator is given by:

$$\sin(\cos^{-1}x) = \sin t = \sqrt{1-\cos^2 t} = \sqrt{1-x^2}$$

We are therefore led to the formula in the statement, namely:

$$\cos^{-1}(x)' = -\frac{1}{\sqrt{1-x^2}}$$

(8) In what regards now $\sin^{-1}$, we can use here the following computation:

$$\begin{aligned}(\sin\circ\sin^{-1})'(x) &= \sin'(\sin^{-1}x)\sin^{-1}(x)'\\ &= \cos(\sin^{-1}x)\ \sin^{-1}(x)'\end{aligned}$$

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

$$\sin^{-1}(x)' = \frac{1}{\cos(\sin^{-1}x)}$$

But with $t = \sin^{-1}x$ we have $\sin t = x$, so the denominator is given by:

$$\cos(\sin^{-1}x) = \cos t = \sqrt{1-\sin^2 t} = \sqrt{1-x^2}$$

We are therefore led to the formula in the statement, namely:

$$\sin^{-1}(x)' = \frac{1}{\sqrt{1-x^2}}$$

(9) Finally, in answer to a question that you might have, regarding the terminology, the story here is that, while I was working on this, and looking for notations for our 2 new functions, a spider came by, and whispered to me cos and sin. More on this in Part II, where we will do 2D mathematics, guided by our 2D expert, that spider. □

Very nice, and we can now upgrade Theorem 4.11, into something final, as follows:

THEOREM 4.14 (upgrade). *The equation of a gravitational free fall, in* 1 *dimension,*

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

can be successfully studied, by computing $t = t(x)$ *instead of* $x = x(t)$*, and we have*

$$t = \sqrt{\frac{x_0^3}{2K}}\left(\sqrt{\frac{x}{x_0}\left(1-\frac{x}{x_0}\right)} + \cos^{-1}\sqrt{\frac{x}{x_0}}\right)$$

with x_0 *being the initial position, at launching.*

PROOF. This comes indeed from Theorem 4.11, via $\varphi = \cos^{-1}$ coming from Theorem 4.13. Of course, many other things can be said, as a continuation of this, and we will be back to this later in this book, in Part II, after learning more about cos and sin. □

As a last topic to be discussed, and in answer to a question that you might have, according to the differentiation formulae in Theorem 4.13, we have:

$$\cos^{-1}(x)' + \sin^{-1}(x)' = -\frac{1}{\sqrt{1-x^2}} + \frac{1}{\sqrt{1-x^2}} = 0$$

Thus, the function $\cos^{-1} + \sin^{-1}$ must be constant, and so, what is this constant? However, this is not an easy question, and more on this later in Part II, when doing 2D. We will learn at that time that the formula is as follows, with $\pi = 3.14159\ldots$ being a certain mysterious number, I mean with no formula for it, naturally appearing in 2D:

$$\cos^{-1} + \sin^{-1} = \frac{\pi}{2}$$

Quite interesting all this, and to be related I guess to my previous Conclusion 4.5, which looks now a bit embarrassing, and I don't even know what to say, about all this. Fortunately the cat, passing by, makes the following interesting comment:

COMMENT 4.15 (cat's take). *The mathematics constants are what is most interesting, in mathematics. The rest being basically philosophy.*

And with this, we will end our discussion about gravity and related topics, in 1 dimension. More on gravity later in this book, in 2 and more dimensions.

4c. Elasticity, waves

We would like to talk now about waves, which are something fundamental in physics. In fact, as we will discover later in this book, each branch of modern physics is guided by its own version of the wave equation, and so, physics is all about waves.

The waves that we will be talking about here are the simplest ones, 1D and mechanical. In order to discuss them, we first need to talk about elasticity, which is behind the propagation of the mechanical waves. And, at the very basic level here, we have:

FACT 4.16 (Hooke law). *When pushing or pulling a spring, with one end fixed,*

$$| \times\!\!\times\!\!\times\!\!\times\!\!\times\!\!\times\!\!\times\!\!\times\!\!\times \circ$$

the force acting on the other end, with orientation opposed to that of our action, is

$$F = kd$$

with d being the displacement of the spring, and k being the spring constant.

Which looks like something quite simple, but believe me, there are tons of advanced mathematics and physics behing this conclusion, that elasticity is indeed linear at the first order, with elasticity itself being a highly non-trivial business. However, for our purposes here, we won't bother much with the meaning of this, where elasticity comes from, and take Fact 4.16 somehow as a definition, for an elastic medium.

Be said in passing, philosophically we will be here, talking about 1D waves without really knowing what 1D elasticity is, in the same kind of trouble as in chapter 2, when we talked about 1D relativity without really knowing what 1D light is. But, as said above, let us just ignore all this, and do the mathematics, see what we get.

Getting now to the mechanical waves, these are impulses traveling through an elastic medium. In the 1D setting, as usual in this chapter, the situation is as follows:

THEOREM 4.17. *The wave equation in 1 dimension is*

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

with the dots 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:

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

Next, observe that this equation can be written, more conveniently, as follows:

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

(4) Which sounds like calculus, so let us have a closer look at the right term. In the $N \to \infty$ limit that we are interested in, and therefore with $l \to 0$, we have:

$$\begin{aligned} & \lim_{l \to 0} \frac{\varphi(x+l) - 2\varphi(x) + \varphi(x-l)}{l^2} \\ = & \lim_{l \to 0} \frac{[\varphi(x) + \varphi'(x)l + \varphi''(x)l^2/2] - 2\varphi(x) + [\varphi(x) - \varphi'(x)l + \varphi''(x)l^2/2]}{l^2} \\ = & \lim_{l \to 0} \frac{\varphi''(x)l^2}{l^2} \\ = & \varphi''(x) \end{aligned}$$

Thus with $N \to \infty$, and therefore with $l \to 0$, our equation in (3) reads:

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

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

Now that we have our equation, let us try to find its solutions. And here, surprise, the simplest solutions are as follows, involving the function cos from Theorem 4.13:

THEOREM 4.18. *The 1D wave equation with propagation speed $v > 0$, namely*

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

has as basic solutions the following functions, with the function cos *being as before,*

$$\varphi(x,t) = A\cos(kx - wt + \delta)$$

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.

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:

$$\ddot{\varphi} = -w^2\varphi \quad , \quad \varphi'' = -k^2\varphi$$

Thus, the wave equation is indeed satisfied, with speed $v = w/k$:

$$\ddot{\varphi} = \left(\frac{w}{k}\right)^2\varphi'' = v^2\varphi''$$

(2) Regarding the rest, all this is basically terminology, which might sound a bit strange in the present 1D context, but that we will discover soon, when doing 2D, to be quite natural, when thinking how $\varphi(x,t) = A\cos(kx - wt + \delta)$ propagates. □

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 4.19. *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 4.20. *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:*

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

Also, 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) As a comment here, consider the basic solutions of the wave equation $\ddot{\varphi} = v^2\varphi''$ that we found in Theorem 4.18, which were as follows, with $w/k = v$:

$$\varphi(x,t) = A\cos(kx - wt + \delta)$$

The initial data for these particular solutions is then as follows:

$$f(x) = \varphi(x,0) = A\cos(kx + \delta)$$

$$g(x) = \dot{\varphi}(x,0) = wA\sin(kx + \delta)$$

Now by plugging this into the d'Alembert formula, and doing the computation, with this being an easy exercise for you, we obtain indeed $\varphi(x,t) = A\cos(kx - wt + \delta)$.

(3) Next, we must show that our solution is unique. And here, instead of going 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:

$$y = x - vt \quad , \quad z = x + vt$$

The point now is that, with this change of variables, which is something quite tricky, mixing space and time variables, and by assuming some multivariable calculus know-how, our wave equation $\ddot{\varphi} = v^2\varphi''$ reformulates in a very simple way, as follows:

$$\frac{d^2\varphi}{dydz} = 0$$

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

$$\varphi(x,t) = F(y) + G(z) = F(x - vt) + G(x + vt)$$

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.

(4) Summarizing, uniqueness proved, modulo some multivariable calculus trickery, sorry for this, that we will learn later in this book. We will be back to this.

(5) 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 d'Alembert integral gets computed via Riemann sums, in our model, as stated.

(6) Finally, as one more thing that can be said, and which is something quite intuitive, when thinking at waves, the solutions appear in fact as superpositions of the basic solutions from (2). And more on this, hopefully later, when discussing Fourier analysis. □

And with this, end of our discussion regarding the waves. As a conclusion, save for all sorts of issues, all related to our low dimensionality in this chapter, namely precise meaning of 1D elasticity, understanding of the function cos in 1D, and missing 1D argument, short-circuiting the standard proof of the uniqueness, in the d'Alembert formula, we managed to have some interesting 1D theory going on, for the mechanical waves.

Which is very nice, remember from the beginning of this section that modern physics means waves. So, with our 1D waves, we have modernity in 1D. Nice.

4d. Pressure, heat

As a last topic of discussion for this chapter, we would like to talk about thermodynamics. However, things here will be quite complicated. The point indeed is that basic thermodynamics, that of the ideal gases, while being formally based only on particles moving, and collisions, is in fact a complicated science, whose foundational results are of genuine 3D nature, appearing when the particles are free to move in 3D.

This being said, we can however do a number of things. Let us start with some simple observations, showing the difficulty of properly talking about 1D gases:

OBSERVATION 4.21. *The most obvious models for the 1D gases are both wrong:*

(1) *With particles subject to collisions, the physics of the gas, enclosed in an interval $[a, b] \subset \mathbb{R}$, will statistically depend only on 2 of its particles, namely the one which is the most at left, and the one which is the most at right. Not good.*
(2) *With particles not subject to collisions, the particles will have to all move with the same speed v. And when enclosing the gas in an interval $[a, b] \subset \mathbb{R}$, these particles will bounce back with speed $-v$, contradiction, unless $v = 0$.*

So, shall we give up? Not so quick, because thinking a bit at 3D, where true thermodynamics happens, we can come up with the following model for the 1D gases:

MODEL 4.22 (ant model). *We call 1D gas a usual 3D gas, without collisions, with all particles moving in the same direction. Equivalently, our gas is in 1D, with the particles avoiding collisions, a bit like ants do, when moving back and forth in their 1D.*

Which sounds quite good, and I guess with this, we are ready for doing some mathematics. Here is our first result, regarding the pressure of our 1D gases:

THEOREM 4.23. *The pressure P and volume V of a 1D gas as above satisfy*

$$PV = 2K$$

where K is the total kinetic energy of the gas.

PROOF. Assume indeed that we have a gas as in Model 4.22, enclosed in a cubic volume, $V = L^3$. 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 $\Delta p = 2mv$. Thus:

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

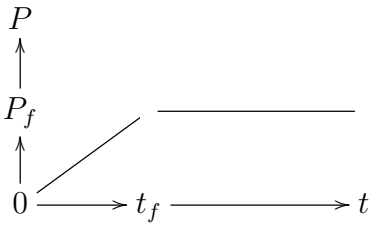
We are therefore led to the conclusion in the statement. □

The above result is quite interesting, so let us keep building on it. We have:

THEOREM 4.24 (addendum). *The correct time for reading the correct pressure is*

$$t_f = \frac{2V^{1/3}}{|v|} \quad : \quad P_f = \frac{2K}{V}$$

with $|v|$ being the average molecular speed, with the precise pressure reading being



and with this being taken in an approximate, statistical sense.

PROOF. Assuming as before that we are in a cubic container, $V = L^3$, we know that each molecule i hits the right wall, where P is measured, at $\Delta t_i = 2L/|v_i|$ intervals. But with this picture in hand, it is quite clear that, on average, the pressure reading process will be linear, starting from $P = 0$, up to time $t_f = 2L/|v|$, with $|v|$ being the average molecular speed, where the correct pressure $P_f = 2K/V$ will be read, and constant at P_f afterwards. Now since $L = V^{1/3}$, this gives the formulae in the statement. □

And with this, end of our discussion about 1D gases, the problem being that for more advanced physics, such as temperature, we would really need collisions between our particles, and this is is barred in 1D, for the reasons explained in Observation 4.21.

Nevermind, what we have in Theorems 4.23 and 4.24 is quite interesting, and more on such things, in higher dimensions, later in this book. Switching topics now, and still talking thermodynamics, we can talk as well about heat diffusion in 1D, as follows:

THEOREM 4.25. *The heat diffusion 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 well from chapter 2, 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) And with all this coming, of course, with the comment that, a bit as before with relativity, waves and gases, we do not really have a 1D phenomenology explaining heat, so what we are doing here is rather a 1D projection of things from 3D, where the phenomenology is known to exist. But let's not talk anymore about such things, you got it I hope, with the present Part I being towards its end, we are less strict with 1D.

(3) 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 $\delta > 0$:

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

(4) Now let us do the math. In the context of our 1D model the neighbors of x are the points $x \pm l$, and so the equation that we wrote above takes the following form:

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

Now observe that we can write this equation as follows:

$$\frac{\varphi(x,t+\delta) - \varphi(x,t)}{\delta} = \alpha \cdot \frac{\varphi(x+l,t) - 2\varphi(x,t) + \varphi(x-l,t)}{l^2}$$

(5) As it was the case with the wave equation before, we recognize on the right the usual approximation of the second derivative, coming from calculus. Thus, when taking the continuous limit of our model, $l \to 0$, we obtain the following equation:

$$\frac{\varphi(x,t+\delta) - \varphi(x,t)}{\delta} = \alpha \cdot \varphi''(x,t)$$

Now with $t \to 0$, we are led in this way to the heat equation in the statement. □

Getting now to the solutions of the heat equation in 1D, we first have:

THEOREM 4.26. *The heat equation $\dot{\varphi} = \alpha\varphi''$ has as solution the function*

$$U_t(x) = \frac{1}{\sqrt{t}}\, e^{-x^2/4\alpha t}$$

which is called fundamental solution, or unnormalized heat kernel.

PROOF. The time derivative of the function in the statement is given by:

$$\dot{U} = -\frac{1}{2t\sqrt{t}}\, e^{-x^2/4\alpha t} + \frac{x^2}{4\alpha t^2\sqrt{t}}\, e^{-x^2/4\alpha t}$$

Regarding the first space derivative, this is given by the following formula:

$$U' = -\frac{x}{2\alpha t\sqrt{t}}\, e^{-x^2/4\alpha t}$$

As for the second space derivative, this is given by the following formula:

$$U'' = -\frac{1}{2\alpha t\sqrt{t}}\, e^{-x^2/4\alpha t} + \frac{x^2}{4\alpha^2 t^2\sqrt{t}}\, e^{-x^2/4\alpha t}$$

We conclude that the heat equation $\dot{U} = \alpha U''$ is indeed satisfied, as claimed. □

With a bit of analytic know-how, we can in fact do better, as follows:

THEOREM 4.27. *The heat equation $\dot{\varphi} = \alpha\varphi''$ has as solution the functions*

$$\varphi(x,t) = \int_{\mathbb{R}} U_t(x-y) f(y) dy$$

where U is its fundamental solution, or unnormalized heat kernel, given by:

$$U_t(x) = \frac{1}{\sqrt{t}}\, e^{-x^2/4\alpha t}$$

Moreover, any solution of the heat equation appears in this way.

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

(1) The first assertion follows from Theorem 4.26, and its proof. The point indeed is that when perturbing the fundamental solution U by an arbitrary function f, according to the procedure in the statement, which is called convolution, all the computations from the proof of Theorem 4.26 will perturb well, according to the following two formulae:

$$\dot{\varphi}(x,t) = \int_{\mathbb{R}} \dot{U}(x-y) f(y) dy$$

$$\varphi''(x,t) = \int_{\mathbb{R}} U''(x-y) f(y) dy$$

Thus, we can see that the heat equation $\dot{\varphi} = \alpha\varphi''$ is indeed satisfied.

(2) As for the second assertion, this is something more advanced, the idea being that the arbitrary solutions can be obtained by convolving the fundamental solution U, suitably normalized, with the initial data. And more on this, in a moment. □

Let us get now to the normalization question for the heat kernel, namely:

$$U_t(x) = \frac{1}{\sqrt{t}}\, e^{-x^2/4\alpha t}$$

This question brings us, mathematically, into the question of integrating the exponentials on the right. And here, we have the following famous result, due to Gauss:

THEOREM 4.28. *We have the following formula,*

$$\int_{\mathbb{R}} e^{-x^2} dx = \sqrt{\pi}$$

called Gauss integral formula.

PROOF. The Gauss integral cannot be computed with bare hands, in 1D. However, this integral can be computed by using 2D, as follows, with $\tan = \sin / \cos$:

$$\begin{aligned}
\int_{\mathbb{R}}\int_{\mathbb{R}} e^{-x^2-y^2} dxdy &= 4\int_0^\infty \int_0^\infty e^{-x^2-y^2} dxdy \\
&= 4\int_0^\infty \int_0^\infty e^{-t^2y^2-y^2} ydtdy \\
&= 4\int_0^\infty \int_0^\infty ye^{-y^2(1+t^2)} dydt \\
&= 2\int_0^\infty \int_0^\infty \left(-\frac{e^{-y^2(1+t^2)}}{1+t^2}\right)' dydt \\
&= 2\int_0^\infty \frac{dt}{1+t^2} \\
&= 2\int_0^\infty (\tan^{-1} t)' dt \\
&= \pi
\end{aligned}$$

To be more precise here, the differentiation formula for $\tan^{-1} t$ comes by using the mathematics from Theorem 4.13, and $\pi = 3.14159\ldots$ is the number mentioned after Theorem 4.14, and I would leave here the various computations needed, as an exercise for you. More on all this, of course, later in this book, when doing 2D. □

Getting back now to the heat equation, we have the following result, about it:

THEOREM 4.29. *The heat equation, $\dot{\varphi} = \alpha\varphi''$ with $\alpha > 0$, with initial condition $\varphi(x, 0) = f(x)$, has as solution the function*

$$\varphi(x, t) = \int_{\mathbb{R}} K_t(x - y) f(y) dy$$

where the function $K_t : \mathbb{R} \to \mathbb{R}$, called heat kernel, given by

$$K_t(x) = \frac{1}{\sqrt{4\pi\alpha t}}\, e^{-x^2/4\alpha t}$$

is the standard solution, coming from $f = \delta_0$, normalized as to have mass 1.

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

(1) To start with, the function K_t in the statement is indeed the standard solution that we found in Theorem 4.26, normalized as to have mass 1, with this coming from:

$$\begin{aligned}\int_{\mathbb{R}} K_t(x)dx &= \frac{1}{\sqrt{4\pi\alpha t}}\int_{\mathbb{R}} e^{-x^2/4\alpha t}dx\\ &= \frac{1}{\sqrt{4\pi\alpha t}}\int_{\mathbb{R}} e^{-y^2}\sqrt{4\alpha t}\,dy\\ &= \frac{1}{\sqrt{4\pi\alpha t}}\cdot\sqrt{4\alpha t}\cdot\sqrt{\pi}\\ &= 1\end{aligned}$$

(2) Next, we can see that by plugging $f=\delta_0$ in the statement, we obtain precisely this standard solution K_t. And with this being something which is not surprising, K_t coming from the simplest situation, that of a radiating point body placed at 0.

(3) Finally, more generally, we can see that by using an arbitrary function f, we obtain the solution of the heat equation with initial condition $\varphi(x,0)=f(x)$. Moreover, it can be shown that this solution is unique, and more on such things, later in this book. □

4e. Exercises

This was a very standard physics chapter, only basics, and as exercises, we have:

EXERCISE 4.30. *What are other forces that you can think of, besides gravity?*

EXERCISE 4.31. *Learn about the story of gravity, Kepler, Newton and the others.*

EXERCISE 4.32. *Do some numerics, for the uniform falls on the Moon, and Sun.*

EXERCISE 4.33. *Try solving the free fall equation $\ddot{x}=-K/x^2$ with bare hands.*

EXERCISE 4.34. *Learn more about* cos, sin *and their inverses* $\cos^{-1}$, $\sin^{-1}$.

EXERCISE 4.35. *Try computing the mathematical constant $c=\cos^{-1}+\sin^{-1}$.*

EXERCISE 4.36. *Try proving the uniqueness, in the d'Alembert theorem.*

EXERCISE 4.37. *Then try the same, for the solutions of the heat equation.*

As bonus exercise, read a bit of thermodynamics. This is the hot subject.

Part II

Two dimensions

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

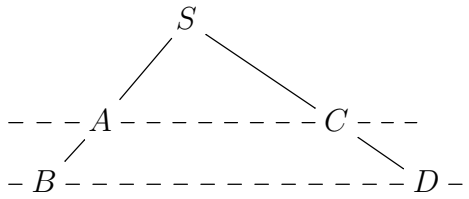
CHAPTER 5

Plane geometry

5a. Angles, triangles

Welcome to 2 dimensions, and many things to be done here, both mathematics and physics, as a continuation of what we did in Part I. As a starting point, we have:

THEOREM 5.1 (Thales). *Proportions are kept, along parallel lines. That is, given a configuration as follows, consisting of two parallel lines, and of two extra lines,*



the following equalities hold:

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

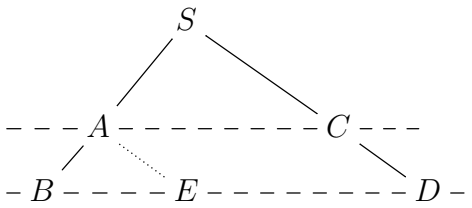
Conversely, the first equality, $SA/SB = SC/SD$, implies $AC||BD$.

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

(1) We use the fact that the area of a triangle is half the side times the corresponding altitude, which is obvious, by completing to a rectangle. With this, we obtain:

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

(2) In order to prove now the second equality in the statement, instead of getting lost into some new area computations, let us draw a tricky parallel, as follows:



By using (1), we have then the following computation, as desired:

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

(3) As for the study of the converses, with $SA/SB=SC/SD$ implying $AC||BD$, but with $SA/SB=AC/DB$ not implying $AC||BD$, we will leave this as an exercise. □

We can now talk about angles, in the obvious way, by using triangles:

FACT 5.2. *We can talk about the angle between two crossing lines, and have some basic theory for the angles going, by using triangles and Thales, in the obvious way.*

To be more precise here, let us go back to the Thales configuration. In that situation, we can say that we have two similar triangles, $SAC\sim SBD$, and with an equivalent formulation of similarity being the fact that the respective angles are equal.

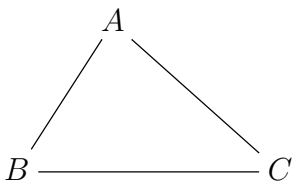
As an important addition to Fact 5.2, making the link with numbers, let us formulate:

DEFINITION 5.3. *We can talk about the numeric value of angles, as follows:*

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

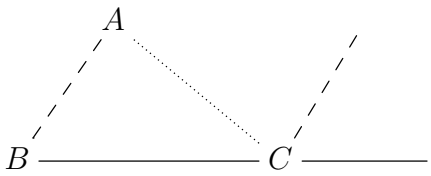
Getting back now to formal work, with theorems and proofs, in relation with the above, here is a key result, which will be our main tool for the study of angles:

THEOREM 5.4. *In an arbitrary triangle in a plane*



the sum of all three angles is 180°*, twice the right angle.*

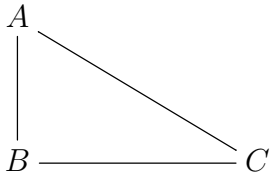
PROOF. We can do this, with a trick. Let us prolong indeed the segment BC a bit, on the C side, and then draw a parallel at C, to the line AB, as follows:



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

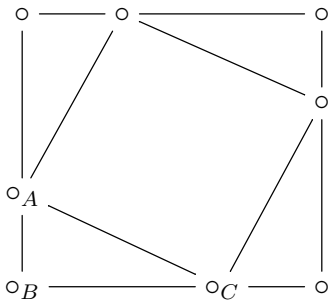
Let us discuss now the simplest angle of them all, which is the right angle $90°$. A triangle having one of the angles equal to $90°$ is called right triangle, and we have:

THEOREM 5.5 (Pythagoras). *In a right triangle ABC,*



we have $AB^2 + BC^2 = AC^2$. Moreover, the converse holds too.

PROOF. The formula in the statement comes from the following magic picture, consisting of two squares, and four triangles which are identical to ABC, as indicated:



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

$$S = (AB + BC)^2 = AB^2 + BC^2 + 2 \times AB \times BC$$

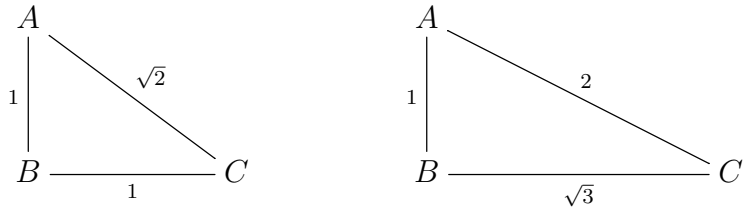
On the other hand, the outer square is made of the smaller square, having side AC, and of four identical right triangles, having sizes AB, BC. Thus:

$$S = AC^2 + 4 \times \frac{AB \times BC}{2} = AC^2 + 2 \times AB \times BC$$

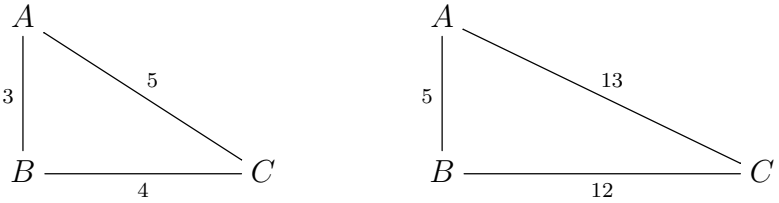
Thus, we have $AB^2 + BC^2 = AC^2$. As for the converse, exercise for you. □

The Pythagoras theorem has many interesting consequences, and here are some:

THEOREM 5.6. *The following are right triangles, with angles* $45° - 45°$ *and* $30° - 60°$*:*



Also, in relation with engineering matters, the following are right triangles,



and so for drawing right angles, you just need a loop, with 12 *or* 30 *knots on it.*

PROOF. Many things going on here, all coming from Pythagoras, as follows:

(1) In what regards the first triangle, having angles $45° - 45°$, things clear.

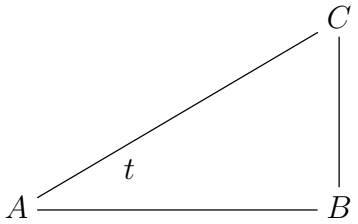
(2) Next, consider a $30° - 60°$ triangle ABC. If we denote by D the middle of AC, then ABD is isosceles, $AD = BD$, and the angle at A is $60°$. By using Theorem 5.4 we conclude that all angles of ABD equal $60°$, and it follows that all sides are equal too. Thus $AB = AD$, and so $AC = 2AB$, and the rest comes from Pythagoras.

(3) Regarding the third triangle, this comes from $9 + 16 = 25$.

(4) Finally, regarding the last triangle, this comes from $25 + 144 = 169$. □

Good news, we can now start talking about trigonometry. Let us begin with:

DEFINITION 5.7. *Given a right triangle* ABC *in the plane,*



we define the sine and cosine of the angle at A *by the following formulae,*

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

with the remark that, by Thales, these functions depend indeed on t *only.*

Observe the link with the functions sin and cos from chapter 4, introduced there via $e^{it} = \cos t + i \sin t$. We will see later in this chapter that, up to a rescaling of the angles, which is something quite natural, our old and new sin and cos functions coincide.

As a few basic illustrations now for this, coming from Theorem 5.6, we have:

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

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

Observe that the numbers in the above two lists have several interesting relations between them. In fact, we have the following general result, regarding this:

THEOREM 5.8. *The sines and cosines are subject to the following formulae:*

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

Also, we have $\sin^2 t + \cos^2 t = 1$*, coming from the Pythagoras theorem.*

PROOF. The first assertion is clear, coming from the picture in Definition 5.7, the angle at C being $90^\circ - t$. Observe that, as an equivalent formulation, we have:

$$\sin(45^\circ - t) = \cos(45^\circ + t) \quad , \quad \cos(45^\circ - t) = \sin(45^\circ + t)$$

As for $\sin^2 t + \cos^2 t = 1$, this comes indeed from the Pythagoras theorem. □

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

THEOREM 5.9. *The sines and cosines of sums are given by the following formulae:*

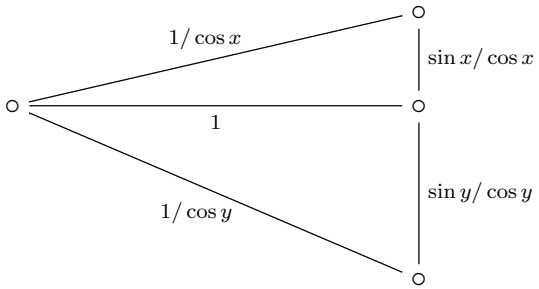
$$\sin(x + y) = \sin x \cos y + \cos x \sin y$$

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

In particular, we can compute $\sin(2t)$*,* $\cos(2t)$*, and* $\sin(t/2)$*,* $\cos(t/2)$ *too.*

PROOF. Many things going on here, the idea being as follows:

(1) Consider the following picture, consisting of a length 1 segment, angles x, y drawn on each side, and everything being completed, and lengths computed, as indicated:



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

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

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

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

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

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

(3) Regarding duplication, as an application of our formula for the sines, we have:

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

For the cosine, by using Pythagoras we have in fact 3 formulae, all useful, namely:

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

(4) Regarding now halving, with $t\to t/2$ in the formulae of $\cos(2t)$, we get:

$$\cos t=2\cos^2\left(\frac{t}{2}\right)-1=1-2\sin^2\left(\frac{t}{2}\right)$$

Now by solving the equations, we are led to the following formulae for halves:

$$\cos\left(\frac{t}{2}\right)=\sqrt{\frac{1+\cos t}{2}}\quad,\quad\sin\left(\frac{t}{2}\right)=\sqrt{\frac{1-\cos t}{2}}$$

(5) Finally, all this was theory, and exercise now for you, to do some computations, say for the sin, cos of all multiples of $15°$, or $7.5°$, or even $3.75°$. Have fun with this. □

Next, as our most pressing task, we would like to establish the link with the material from chapter 4, and more specifically, with the formula $e^{it}=\cos t+i\sin t$ from there.

And here, to start with, I bet that our convention for measuring angles, with 90° coming from the 30 days of the lunar month, is not the good one. So, what to do? In the lack of the bright idea here, I will ask our 2D expert, the spider. Who answers:

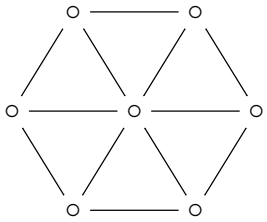
SPIDER 5.10. *In 2 dimensions, it is all about circles and rays. By the way, you should eat some flies too, our intelligence comes from there.*

Okay, thanks spider, so let us attempt now to get into a more advanced study of the angles, by using circles. We have here the following key result, to start with:

THEOREM 5.11. *The following two definitions of $\pi = 3.14159\ldots$ 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 the equivalence, let us cut the unit disk as a pizza, into N slices, and forgetting about gastronomy, leave aside the rounded parts:



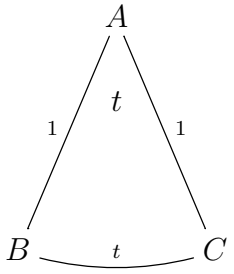
The area to be eaten can be then computed as follows, where H is the height of the slices, S is the length of their sides, and $P = NS$ is the total length of the sides:

$$A = N \times \frac{HS}{2} = \frac{HP}{2} \simeq \frac{1 \times L}{2}$$

Thus with $N \to \infty$ we get $A = L/2$, as desired. As for the numerics, $\pi > 3$, not by much, and a more careful study using higher polygons leads to $\pi = 3.14159\ldots$ □

Getting now to what we wanted to do, in relation with the angles, we have:

THEOREM 5.12. *We can measure angles by putting them in the middle of a circle of radius 1, and assigning to them the corresponding arc lengths:*



Equivalently, we can use twice the area of the disk slice, which equals the arc length. In this way, the multiples of 90° get converted into corresponding multiples of $\pi/2$.

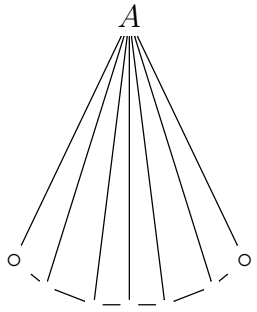
PROOF. We have two things to be proved here, as follows:

(1) First is the fact that our measuring method is indeed good, in the sense that doubling the angles will double their values, tripling the angles will triple their values, and so on. But this is something which is plainly obvious, so done with this.

(2) And then, there is the claim that we have the following formula, with on the left the area of the disk slice ABC, and on the right the arc length BC:

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

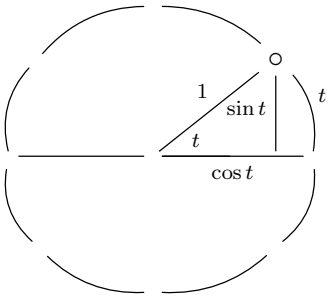
But this is something which is clear for isosceles triangles having altitude 1, and then our disk slice can be approximated by unions of such isosceles triangles, as follows:



Thus, we conclude that our area formula holds indeed, as desired. □

As a question now that you might have, is doing the above, namely replacing our beloved 90° coming from astronomy by that crazy $\pi/2$ number, a good thing? In answer, our new convention shines when it comes to trigonometry. We first have:

THEOREM 5.13. *The sine and cosine of any $t \in \mathbb{R}$ can be computed according to*



with the convention that inverted segments count as negatives.

PROOF. This is something that we basically know, but with the unit circle and that length t arc certainly bringing some new light on the sine and cosine. For instance, we can see on the above picture that the correspondence $t \to \sin t$, which is now something geometric, is such that $\sin t \simeq t$ for t small. And, more on this in a moment. □

Let us get now into an interesting question, namely estimating the trigonometric functions. For this purpose, we can use the formulae for sums, which allow us to transport our approximation questions around $t = 0$. And, around 0, we have:

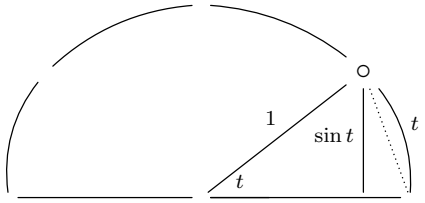
THEOREM 5.14. *We have the following estimates, valid for small angles,*

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

where $\tan t = \sin t / \cos t$*, with coming from our new convention for numeric angles.*

PROOF. Many things can be said here, the idea being as follows:

(1) The general idea is that the estimates are both clear from our circle picture for the angles, and trigonometric functions. Indeed, the picture for the sine is:



Now by using the standard fact that the shortest distance between a point and a line is achieved by constructing the orthogonal projection on that line, we obtain:

$$\sin t \leq t$$

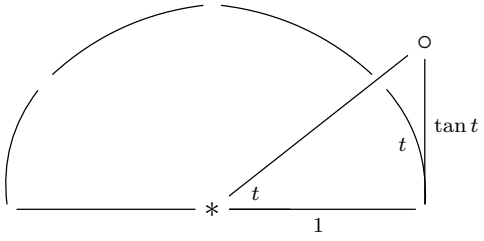
(2) Equivalently, and a bit more rigorously, we can draw the dotted segment above, having length $2\sin(t/2)$, and with Pythagoras on the left, followed by shortest distance between two points being achieved by that dotted segment on the right, we obtain:

$$\sin t \leq 2\sin(t/2) \leq t$$

(3) As yet another proof, we can compare the area of the above isosceles triangle with the area of the disk slice, which gives right away the following estimate, as desired:

$$\frac{\sin t}{2} \leq \frac{t}{2}$$

(4) Regarding now $\tan t = \sin t / \cos t$, again for $t \in [0, \pi/2]$, by using similar triangles we can see that the picture is as follows, justifying the name "tangent":



But then, we can argue that the arc t and segment $\tan t$ are related by a projection from $*$, which lands orthogonally on the arc, and obliquely on the segment, and since orthogonal projections notoriously provide the best view, we obtain, as claimed:

$$t \leq \tan t$$

(5) Equivalently, and more rigorously, by comparing areas we get, as desired:

$$\frac{t}{2} \leq \frac{\tan t}{2}$$

(6) Summarizing, done, one way or another, both the inequalities proved. □

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

THEOREM 5.15. *The following happen, for small angles, again coming from our new convention for numeric angles, and best justifying this convention:*

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

PROOF. This can be indeed established as follows:

(1) This is clear indeed on the circle, by arguing like in the previous proof, and we will leave the various details here as an instructive exercise. Equivalently, this follows from $\sin t \leq t \leq \tan t$, by using $\tan t = \sin t / \cos t \simeq \sin t$, coming from $\cos t \simeq 1$.

(2) This comes from (1), and from Pythagoras. Indeed, knowing $\sin t \simeq t$, when looking for a quantity $\cos t$ making the Pythagoras formula $\sin^2 t + \cos^2 t = 1$ hold, we are led, via some quick thinking, to the formula $\cos t \simeq 1 - t^2/2$, according to:

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

(3) This is again clear on the circle, or simply follows from (1,2), by dividing. □

5b. Affine coordinates

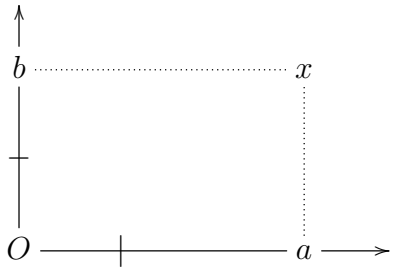
With the above discussed, basic plane geometry and trigonometry, time now for more advanced technology, using affine coordinates. Let us start with:

DEFINITION 5.16. *A vector is a pair of real numbers, written vertically:*

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

We identify the vectors with the points in the plane, in the obvious way.

To be more precise here, let us fix a system of coordinates in the plane. Now given a point x in the plane, we can project it onto the coordinate axes, and call the numbers $a, b \in \mathbb{R}$ describing the positions of these projections, with respect to the origin O, the coordinates of x, with the picture for this being as follows:



Observe now that, conversely, given two real numbers $a, b \in \mathbb{R}$, these will uniquely determine a certain point x is the plane, constructed according to the above picture. That is, we draw a on the horizontal axis, b on the vertical axis, than we draw perpendiculars as above, and x will be then the intersection of these two perpendiculars.

Finally, regarding the vertical notation for vectors, you will have to trust me here:

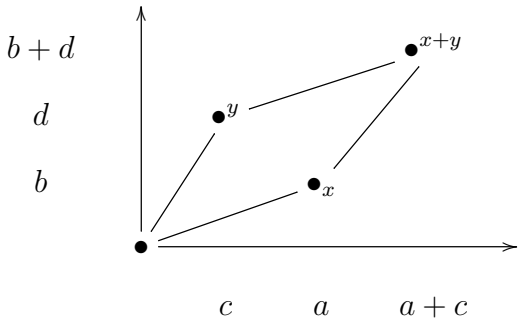
ADVICE 5.17. *Never write the vectors horizontally, because you will regret later, when learning linear algebra, and multivariable calculus, all that needing vertical vectors.*

Many interesting things can be done with vectors, and of particular interest is the summing operation for such vectors, given by the following formula:

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

Geometrically, and coming as a simple application of the Thales theorem, the idea with this operation is that the vectors add by forming a parallelogram, as shown by:

THEOREM 5.18. *The vector addition can be understood geometrically,*

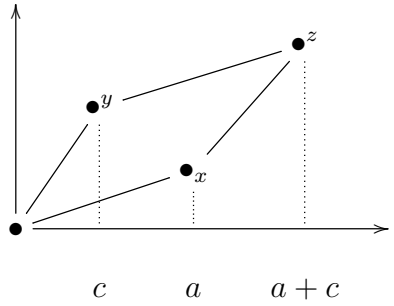


with $x + y$ completing the parallelogram based at O, x, y.

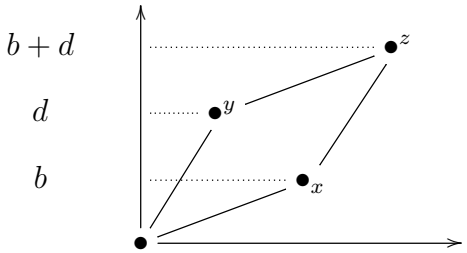
PROOF. This is something quite self-explanatory. Consider indeed a parallelogram in the plane, with three of its vertices being as follows:

$$O = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad , \quad x = \begin{pmatrix} a \\ b \end{pmatrix} \quad , \quad y = \begin{pmatrix} c \\ d \end{pmatrix}$$

Now let us draw verticals from x, y, and from the fourth vertex z too. From Thales we obtain that the first coordinate of z is $a + c$, according to the following picture:



Similarly, if we draw horizontals from x, y, and from z too, from Thales we obtain that the second coordinate of z is $b + d$, according to the following picture:



Thus we are led to the picture in the statement, and with the final conclusion being that the coordinates of the fourth vertex z can be computed according to:

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

But this is exactly the summing formula for the vectors, as desired. □

In practice, the summing operation is usefully complemented by the multiplication by scalars operation, which is given by the following very intuitive formula:

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

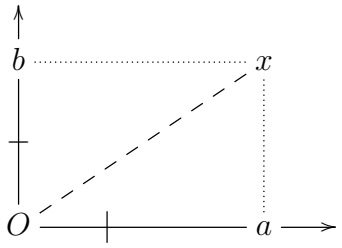
Finally, of particular interest too, in relation with the computation of the lengths, is the following result, allowing us to compute the length of any vector:

THEOREM 5.19. *The length of a vector is given by the following formula:*

$$x = \begin{pmatrix} a \\ b \end{pmatrix} \implies ||x|| = \sqrt{a^2 + b^2}$$

Also, the vector lengths satisfy $||\lambda x|| = |\lambda| \cdot ||x||$*, and* $||x + y|| \leq ||x| + ||y||$.

PROOF. In what regards the first assertion, this follows as a basic application of the theorem of Pythagoras, according to the following picture:



Regarding now the second assertion, with $x = \binom{a}{b}$, we have indeed:

$$\begin{aligned} ||\lambda x|| &= \sqrt{(\lambda a)^2 + (\lambda b)^2} \\ &= |\lambda| \sqrt{a^2 + b^2} \\ &= |\lambda| \cdot ||x|| \end{aligned}$$

Finally, in what regards the last assertion, this is clear geometrically. However, we can prove this algebrically as well. Indeed, with $x = \binom{a}{b}$ and $y = \binom{c}{d}$, we must prove:

$$\sqrt{(a + c)^2 + (b + d)^2} \leq \sqrt{a^2 + b^2} + \sqrt{c^2 + d^2}$$

And I will leave finishing this to you, by raising to the square and so on. □

Many interesting things can be done with the affine coordinates, as introduced above, both theory and computations, but we will prefer here to defer the discussion for later, after learning about complex coordinates, which are something even smarter.

5c. 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 5.20. *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 their 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:

$$xy = (ac - bd) + i(ad + bc)$$

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, we have the following result:

THEOREM 5.21. *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:

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

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 5.22. *The complex numbers, written as usual*

$$x = a + ib$$

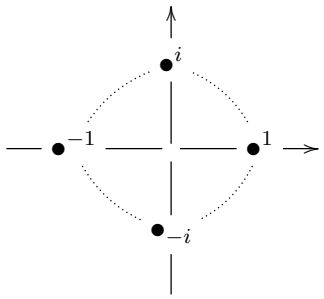
can be represented in the plane, according to the following identification:

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

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

PROOF. This is indeed something self-explanatory, with the last assertion coming by comparing the definitions of the sums of vectors and complex numbers. □

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:



In order to understand now the multiplication operation, we must do something more complicated, namely using polar coordinates. Let us start with:

DEFINITION 5.23. *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,

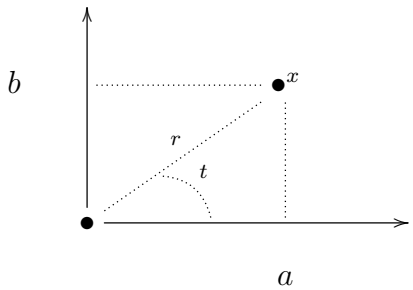
$$a = r \cos t \quad , \quad b = r \sin t$$

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 5.22, 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 5.23 is as follows:



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 5.24. *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. 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. □

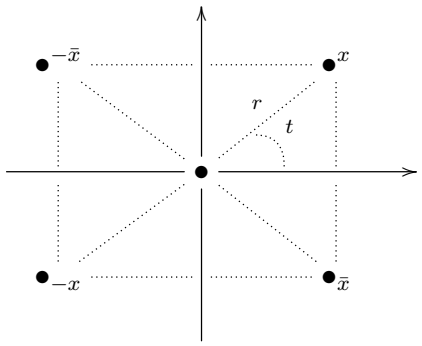
As a last general topic regarding the complex numbers, let us discuss conjugation:

DEFINITION 5.25. *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 operations on complex numbers, a quick picture says it all:



There are many things that can be said about the conjugation, as follows:

THEOREM 5.26. *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 5.25.

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

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

(6) This comes from the formula of the solutions, that we know from Theorem 5.21, 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. □

Getting now to complex functions and analysis, with the aim of establishing the formula $e^{it} = \cos t + i \sin t$, still on our to-do list, we first have the following result:

THEOREM 5.27. *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:

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

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

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

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

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

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

$$\lim_{t\to0}e^{x+t}=\lim_{t\to0}e^xe^t=e^x\lim_{t\to0}e^t=e^x\cdot1=e^x$$

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

Good news, we can now, eventually, close a discussion started some time ago:

THEOREM 5.28. *We have the following formula,*

$$e^{it}=\cos t+i\sin t$$

due to Euler, valid for any $t\in\mathbb{R}$.

PROOF. Many things can be said here, the idea being as follows:

(1) The Euler formula shows that $t\to e^{it}$ maps $\mathbb{R}$ to the unit circle $\mathbb{T}$, which might seem quite surprising, so let us first try to understand this. We have, for any $x\in\mathbb{C}$:

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

Also, we have as well the following computation, again valid for any $x\in\mathbb{C}$:

$$e^xe^{-x}=e^{x-x}=e^0=1\implies(e^x)^{-1}=e^{-x}$$

(2) Now these two formulae, applied with $x=it$, with $t\in\mathbb{R}$, give:

$$e^{-it}=\overline{e^{it}}\quad,\quad(e^{it})^{-1}=e^{-it}$$

We conclude from this that the complex number $z=e^{it}$ has the property $z^{-1}=\bar{z}$, which is exactly the equation of the unit circle $\mathbb{T}\subset\mathbb{C}$, as desired.

(3) Getting now to the proof of the Euler formula, we know that $t \to e^{it}$ maps $\mathbb{R} \to \mathbb{T}$, and also that this is a group morphism, meaning mapping sums in $\mathbb{R}$ into products in $\mathbb{T}$. But in view of this, barring any pathologies, this operation can only appear by "wrapping". That is, we must have a formula as follows, for a certain $\alpha \in \mathbb{R}$:

$$e^{it} = \cos(\alpha t) + i \sin(\alpha t)$$

(4) In order now to find the parameter $\alpha \in \mathbb{R}$, let us look at what happens around $t = 0$. And here, we have the following basic estimate, obtained by truncating exp:

$$e^{it} \simeq 1 + it$$

On the other hand, according to the basic trigonometry estimates for sin and cos, from earlier in this chapter, we have as well the following estimate, again around $t = 0$:

$$\cos(\alpha t) + i \sin(\alpha t) \simeq 1 + i\alpha t$$

We conclude that we must have $\alpha = 1$, which gives the Euler formula.

(5) Alternatively, calculus shows that the derivative of $f(t) = e^{-it}(\cos t + i \sin t)$ vanishes, so $f(t) = f(0) = 1$. And, exercise for you, to learn more about all this, various proofs of the Euler formula, and to review as well the material from chapter 4. □

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

THEOREM 5.29. *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 5.23, by using the formula $e^{it} = \cos t + i \sin t$ from Theorem 5.28, 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 5.30. *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 5.24, reformulated by using the exponential notation from Theorem 5.29. □

We can investigate as well more complicated operations, as follows:

THEOREM 5.31. *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 5.30:

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

(2) Once again by using Theorem 5.30, 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 5.30, this operation $x \to x^a$ is indeed the correct one.

(4) This comes from the fact, that we know from Theorem 5.26, that the conjugation operation $x \to \bar{x}$ keeps the modulus, and switches the sign of the argument. □

5d. Equations, roots

Getting back to algebra, recall from Theorem 5.21 that any degree 2 real equation has 2 complex roots. We can in fact prove that any complex polynomial equation, of arbitrary degree $N \in \mathbb{N}$, has exactly N complex solutions, counted with multiplicities:

THEOREM 5.32. *Any polynomial $P \in \mathbb{C}[X]$ decomposes as*

$$P = c(X - a_1) \dots (X - a_N)$$

with $c \in \mathbb{C}$ and with $a_1, \dots, a_N \in \mathbb{C}$.

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. □

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 5.33. *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 5.30. 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 $w = e^{2\pi i/3}$.

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 5.31, the precise statement being as follows:

THEOREM 5.34. *Any nonzero $x = re^{it}$ has 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 reformulates as:

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

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

5e. Exercises

Welcome to geometry in 2 dimensions, and as exercises here, we have:

EXERCISE 5.35. *Solve the Pythagoras equation $a^2 + b^2 = c^2$ over the integers.*

EXERCISE 5.36. *Learn about the theorems of Menelaus and Ceva.*

EXERCISE 5.37. *Learn about the theorems of Desargues and Pappus.*

EXERCISE 5.38. *Not to forget the theorems of Pascal and Brianchon.*

EXERCISE 5.39. *Compute the sines and cosines of multiples of $15°$, $7.5°$, $3.75°$.*

EXERCISE 5.40. *Work out some approximations of π, using N-gons with $N = 2^s$.*

EXERCISE 5.41. *Solve some easy geometry problems, using affine coordinates.*

EXERCISE 5.42. *Solve as well a few geometry problems, using complex numbers.*

As bonus exercise, review what we previously said about $\cos, \sin$ in chapter 4.

CHAPTER 6

Waves and light

6a. Heat, revised

Back now to physics, we have many things to be done, as a continuation of what we learned in 1 dimension, in chapters 2 and 4. Which was actually quite a mess, with all sorts of physics involved, and thinking a bit, the 2D work would fall in two classes:

(1) First we have physics coming from motion, with or without collisions, including would-be thermodynamics, notably with the heat diffusion equation, then elasticity and the closely related wave equation, and finally various topics related to light, which is well-known to be a wave, including things like color, optics, and relativity theory.

(2) And then we have more die-hard, old physics, based on gravity only. With this including the movement of the pendulum, and the related topics of harmonic oscillators and resonance, then the solution of the planetary movement problem, following Kepler and Newton, and then other things, such as more on the equations, and on energy.

So, what to do? I don't know about you, but personally I feel a bit lost here. I mean, guided by the history of the subject, always a safe choice, we should go with (2) first, and do (1) after. However, still thinking history, the mathematics of Einstein for instance is in fact much simpler than that of Newton, so maybe we humans got it wrong somewhere, we should have produced Einstein before Newton, and so (1) comes before (2).

Well, as usual in such delicate situations, time to ask our 2D expert, the spider. But will the spider, who is rather into engineering, computations and hunting, all concrete things, be good at answering philosophical questions? And fortunately, so he is:

SPIDER 6.1. *Not many flies these days, I spend my time thinking at math and physics. You should go with waves, even without flies, small breezes, and 2D is curved.*

Okay, thanks spider, so my intuition was right, and we will go for (1) in this chapter. As for (2), that can stay for later, and by the way, we will probably need an extra math chapter, dealing with calculus in 2D, before attacking that Kepler-Newton theorem.

Getting started now, as a continuation of what we did in chapter 4, let us first discuss would-be thermodynamics, namely pressure and temperature basics, this time in 2D.

For our first result, dealing with pressure, and generalizing things from chapter 4, it is actually convenient to assume that we are in arbitrary $N = 1, 2, 3$ dimensions, with $N = 1$ being here for explaining the link with chapter 4, then $N = 2$ being here for our new result, and with $N = 3$, making the link with the real life, being included too.

So, let us call N-dimensional gas with $N = 1, 2, 3$ a usual 3-dimensional gas, with all molecules assumed to move in N directions only. That is, at $N = 1$ all molecules are assumed to move in the same direction, at $N = 2$ all molecules are assumed to move in the same plane, and at $N = 3$ the molecules are assumed to freely move in space. And with these various constraints referring to the speed vectors of the molecules.

With this convention, we have the following result, regarding pressure:

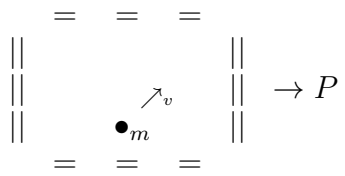
THEOREM 6.2. *The pressure P, volume V and total kinetic energy K of a gas, having point molecules, with no collisions between them, satisfy*

$$PV = \frac{2K}{N}$$

where $N = 1, 2, 3$ is the dimensionality of the gas.

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

(1) Let us first assume that the gas is enclosed in a cubic volume, $V = L^3$. 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. We must compute the pressure P exterted by this molecule on the right wall:



(2) We first look at a 1D gas. Our molecule hits the right wall at every $\Delta t = 2L/v$ interval, with its change of momentum being $\Delta p = 2mv$. We obtain, as desired:

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

(3) In the case of a N-dimensional gas, exactly the same computation takes place, but this time with v being replaced by its horizontal component v_1. Thus, we have:

$$P = \frac{mv_1^2}{V}$$

But, we have the following formula, with the equality on the right being understood in a statistical sense, our molecule being assumed to follow a random direction:

$$||v||^2 = v_1^2 + \ldots + v_N^2 = Nv_1^2$$

Thus, the pressure in this case is given by the following formula, as desired:

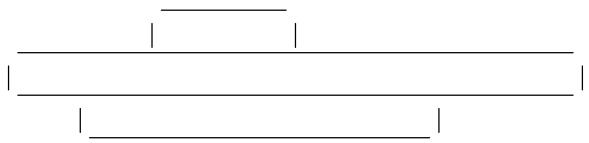
$$P = \frac{m||v||^2}{NV} = \frac{2K}{NV}$$

(4) It remains to extend our result to arbitrary volume shapes. For this purpose, let us first redo the above computations for a parallelepiped, $V = L_1L_2L_3$. Here the above 1D gas computation carries on, and gives the same result, as follows:

$$P = \frac{F}{L_2L_3} = \frac{\Delta p}{L_2L_3\Delta t} = \frac{2mv}{L_2L_3 \cdot 2L_1/v} = \frac{mv^2}{L_1L_2L_3} = \frac{2K}{V}$$

Thus the N-dimensional computation carries on too, and gives the result.

(5) In order now to reach to arbitrary shapes, the idea will be that of stacking thin parallelepipeds, best approximating the shape that we have in mind, as follows:



(6) But for this purpose it is better to drop our assumption that the gas has 1 molecule, and use M molecules instead. With $\rho = M/V$ being the molecular density, and K_0 being the kinetic energy of a single molecule, our computation in (4) for the parallelepiped, with now M molecules instead of 1, reformulates as follows:

$$P = \frac{2K}{NV} = \frac{2MK_0}{NV} = \frac{2\rho VK_0}{NV} = \frac{2\rho K_0}{N}$$

(7) But this latter formula shows that the pressure has nothing to do with the precise volume V, but just with the molecular density $\rho = M/V$. Thus, we can stack indeed parallelepipeds, with of course the assumption that ρ is constant over these parallelepipeds, and we obtain that the above formula holds for an arbitrary volume shape V:

$$P = \frac{2\rho K_0}{N}$$

Now by getting back to the volume V, we obtain the following formula:

$$P = \frac{2\rho K_0}{N} = \frac{2MK_0}{NV} = \frac{2K}{NV}$$

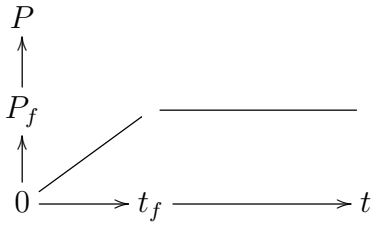
Thus, we are led to the conclusion in the statement. □

The above result is quite interesting, so let us keep building on it. Again by generalizing some 1D things that we know from chapter 4, we have the following result:

THEOREM 6.3 (addendum). *In the context of a gas consisting of point molecules, with no collisions between them, the correct time for reading the correct pressure is*

$$t_f = \frac{2\sqrt{N}\,V^{1/3}}{||v||} \quad : \quad P_f = \frac{2K}{NV}$$

with $||v||$ being the average molecular speed, with the precise pressure reading being



and with this being taken in an approximate, statistical sense.

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

(1) At $N = 1$, as explained in chapter 4, we can assume that we are in a cube, $V = L^3$, and we know that each molecule i hits the right wall, where P is measured, at $\Delta t_i = 2L/|v_i|$ intervals. But with this picture in hand, it is quite clear that, on average, the pressure reading process will be linear, starting from $P = 0$, up to time $t_f = 2L/|v|$, with $|v|$ being the average molecular speed, where the correct pressure $P_f = 2K/V$ will be read, and constant at P_f afterwards. Now since $L = V^{1/3}$, this gives, as desired:

$$t_f = \frac{2V^{1/3}}{|v|} \quad : \quad P_f = \frac{2K}{V}$$

(2) In the general case now, that of a N-dimensional gas, with $N = 1, 2, 3$, the same argument carries on, with the only change being that each molecular speed $v_i \in \mathbb{R}^N$ is now replaced by its horizontal component $v_{i1} \in \mathbb{R}$, which by statistical reasons has squared magnitude as follows, as explained in the proof of Theorem 6.2:

$$v_{i1}^2 = ||v_i||^2/N$$

Thus, with respect to (1), the correct final pressure must be adjusted by a N factor, and becomes $P_f = 2K/NV$, as in Theorem 6.2. As for the correct reading time, this must be adjusted by a $\sqrt{N}$ factor, and becomes $t_f = 2\sqrt{N}\,V^{1/3}/||v||$, as claimed. □

So long for pressure. Regarding now heat, bringing us into more advanced thermodynamics, this is notoriously a 3D phenomenon, but we can nevertheless have a look at formal heat diffusion, as we previously did in chapter 4, in 1 dimension.

But here, a bit as before in the context of our pressure considerations, it is most convenient to do our study in arbitrary $N = 1, 2, 3$ dimensions, with $N = 1$ being here

for explaining the link with the material from chapter 4, then $N = 2$ being here for our new result, and with $N = 3$, making the link with the real life, being included too.

So, here is the our result regarding heat, extending our 1D study from chapter 4:

THEOREM 6.4. *Heat diffusion in $N = 1, 2, 3$ dimensions is described by the equation*

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

where $\alpha > 0$ is the thermal diffusivity of the medium, 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. The study here is quite similar to the 1D study from chapter 4, as follows:

(1) To start with, in what regards the phenomenology of heat diffusion, and the statement itself, let us recall from chapter 4 that in 1 dimension the heat diffusion equation is as follows, with $\alpha > 0$ being the thermal diffusivity of the medium:

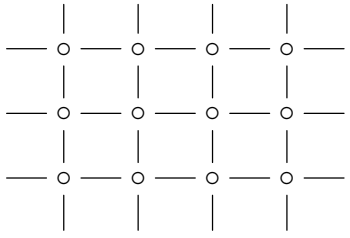
$$\dot{\varphi} = \alpha\varphi''$$

In $N \geq 2$ dimensions now, we can expect the same arguments to carry on, and more on this in a moment. However, what about the final result, that we are supposed to get? The point indeed is that we are dealing now with functions $\varphi : \mathbb{R}^N \to \mathbb{R}$, no longer having a well-defined second space derivative φ'', at $N \geq 2$. In answer, we can expect all N directional second derivatives to contribute, and so, our equation should be:

$$\dot{\varphi} = \alpha \sum_{i=1}^{N} \frac{d^2\varphi}{dx_i^2}$$

Thus, as a conclusion to this preliminary study, things fine with our statement.

(2) Getting to work now, let us first discuss 2 dimensions. Here we can use a lattice model as follows, with all lengths being $l > 0$, for simplifying:



(3) 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:

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

(4) 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:

$$\begin{aligned} \frac{\varphi(x,y,t+\delta) - \varphi(x,y,t)}{\delta} &= \alpha \cdot \frac{\varphi(x+l,y,t) - 2\varphi(x,y,t) + \varphi(x-l,y,t)}{l^2} \\ &+ \alpha \cdot \frac{\varphi(x,y+l,t) - 2\varphi(x,y,t) + \varphi(x,y-l,t)}{l^2} \end{aligned}$$

(5) But here, as previously in 1D, in chapter 4, 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:

$$\dot{\varphi}(x,y,t) = \alpha \cdot \Delta\varphi(x,y,t)$$

(6) Finally, in $N = 3$ dimensions, or even higher, the same argument carries over, namely a straightforward lattice model, and we will leave the details here, based on the above, as an exercise, and gives the heat equation, as formulated in the statement. □

Getting now to the solutions of the heat equation, again by following the 1D material from chapter 4, we have the following result, dealing with $N = 1, 2, 3$ dimensions:

THEOREM 6.5. *The heat diffusion equation, $\dot{\varphi} = \alpha\Delta\varphi$ with $\alpha > 0$, with initial condition $\varphi(x, 0) = f(x)$, has as solution the function*

$$\varphi(x,t) = \int_{\mathbb{R}^N} K_t(x-y)f(y)dy$$

where the function $K_t : \mathbb{R}^N \to \mathbb{R}$, called heat kernel, given by

$$K_t(x) = \frac{1}{\sqrt{(4\pi\alpha t)^N}}\, e^{-||x||^2/4\alpha t}$$

is the standard solution, coming from $f = \delta_0$, normalized as to have mass 1.

PROOF. The study here is similar to our 1D study from chapter 4, as follows:

(1) Let us first prove that the function K_t in the statement satisfies the heat equation. For this purpose, our equation $\dot{\varphi} = \alpha\Delta\varphi$ being linear, it is convenient to get rid of the $1/\sqrt{(4\pi\alpha)^N}$ factor, that we will only need later. Thus, we want to prove that the following function, called unnormalized heat kernel, satisfies the heat equation $\dot{\varphi} = \alpha\Delta\varphi$:

$$U = \frac{1}{\sqrt{t^N}}\, e^{-||x||^2/4\alpha t}$$

(2) Before starting the computations, yes I can hear you screaming that we have not learned yet about partial derivatives and multivariable calculus. In answer, partial derivatives are something trivial, and 1-dimensional, I mean just differentiate with respect to your variable, as you usually do in 1D calculus, by keeping the other variables fixed, and this is what we will do in what follows. As for multivariable calculus, this is something else, namely the art of doing advanced mathematics using partial derivatives, that we will not need for our computations here, and that we will learn later, in chapter 7.

(3) Getting to work now, the time derivative of our function U is given by:

$$\dot{U} = -\frac{N}{2t\sqrt{t^N}}\, e^{-||x||^2/4\alpha t} + \frac{||x||^2}{4\alpha t^2\sqrt{t^N}}\, e^{-||x||^2/4\alpha t}$$

Regarding the first space derivatives, these can be computed as indicated in (2), differentiate with respect to the chosen variable, and are given by the following formula:

$$\frac{dU}{dx_i} = -\frac{x_i}{2\alpha t\sqrt{t^N}}\, e^{-||x||^2/4\alpha t}$$

As for the second space derivatives, again these can be computed as indicated in (2), namely differentiate with respect to the chosen variable, and are given by:

$$\frac{d^2U}{dx_i^2} = -\frac{1}{2\alpha t\sqrt{t^N}}\, e^{-||x||^2/4\alpha t} + \frac{x_i^2}{4\alpha^2 t^2\sqrt{t^N}}\, e^{-||x||^2/4\alpha t}$$

And with this, good news, end of our derivative computations. Indeed, by summing over i the second space derivatives, we obtain the following formula:

$$\Delta U=-\frac{N}{2\alpha t\sqrt{t^N}}\,e^{-||x||^2/4\alpha t}+\frac{||x||^2}{4\alpha^2t^2\sqrt{t^N}}\,e^{-||x||^2/4\alpha t}$$

We conclude that the heat equation $\dot{U}=\alpha\Delta U$ is indeed satisfied, as claimed. Thus, by linearity, we have as well $\dot{K}=\alpha\Delta K$, for the heat kernel itself.

(4) Next, when perturbing the fundamental solution K that we found by an arbitrary function f, according to the convolution procedure in the statement, all the computations from (3) will perturb well, according to the following two formulae:

$$\dot{\varphi}(x,t)=\int_{\mathbb{R}^N}\dot{K}(x-y)f(y)dy$$

$$\Delta\varphi(x,t)=\int_{\mathbb{R}^N}\Delta K(x-y)f(y)dy$$

Thus, we can see that the heat equation $\dot{\varphi}=\alpha\Delta\varphi$ is indeed satisfied.

(5) Regarding now normalization, again by following the material from chapter 4, observe that the heat kernel has indeed mass 1, as shown by:

$$\begin{aligned}\int_{\mathbb{R}^N}K_t(x)dx&=\frac{1}{\sqrt{(4\pi\alpha t)^N}}\int_{\mathbb{R}^N}e^{-||x||^2/4\alpha t}dx\\&=\frac{1}{\sqrt{(4\pi\alpha t)^N}}\left(\int_{\mathbb{R}}e^{-x^2/4\alpha t}dx\right)^N\\&=\frac{1}{\sqrt{(4\pi\alpha t)^N}}\left(\int_{\mathbb{R}}e^{-y^2}\sqrt{4\alpha t}\,dy\right)^N\\&=\frac{1}{\sqrt{(4\pi\alpha t)^N}}\cdot(\sqrt{4\alpha t}\cdot\sqrt{\pi})^N\\&=1\end{aligned}$$

(6) Next, we can see that by plugging $f=\delta_0$ in the statement, we obtain precisely the standard solution K_t. And with this being something which is not surprising, K_t coming from the simplest situation, that of a radiating point body placed at 0.

(7) Finally, more generally, we can see that by using an arbitrary function f, we obtain the solution of the heat equation with initial condition $\varphi(x,0)=f(x)$. Moreover, it can be shown that this solution is unique, and more on such things, later in this book. □

And with this, end of our discussion on 2D thermodynamics. As a conclusion, we learned many interesting things, but looking back at what we did, nothing was in fact of genuine 2D nature. So, it is either that something is wrong with thermodynamics, or with $N=2$, or with the combination of the two, 2D thermodynamics. Mystery.

6b. Waves and light

Getting now to our main topic of discussion for this chapter, light, and everything coming from it, to start with, we can talk about waves in N dimensions, as follows:

THEOREM 6.6. *The wave equation in $N = 1, 2, 3$ dimensions is as follows,*

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

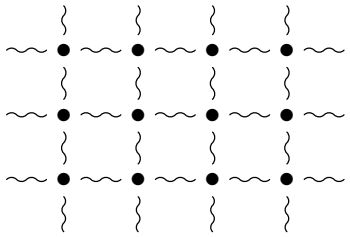
with $v > 0$ being the propagation speed of the wave, and with Δ given by

$$\Delta\varphi = \sum_{i=1}^{N} \frac{d^2\varphi}{dx_i^2}$$

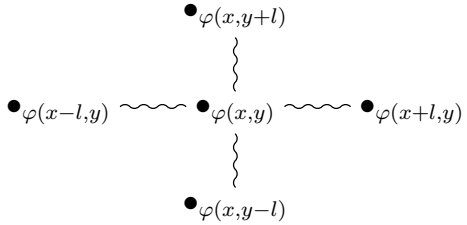
being as usual the Laplace operator, playing the role of a numeric second derivative.

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 in chapter 4 carries on. Indeed, we can use a lattice model as follows, with all 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:

$$\ddot{\varphi}(x,y)=\frac{KL^2}{M}\left(\frac{d^2\varphi}{dx^2}+\frac{d^2\varphi}{dy^2}\right)(x,y)$$

(3) 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.

(4) Finally, in $N = 3$ dimensions, or even higher, the same argument carries over, namely a straightforward lattice model, and we will leave the details here, based on the above, as an exercise, and gives the wave equation, as formulated in the statement. □

Getting now to light, things are quite complicated here. The idea is that light is a wave predicted by electrodynamics, that is, by the Maxwell equations that we will learn later in this book, 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. To be more precise, we first have:

FACT 6.7. *An accelerating or decelerating charge produces electromagnetic radiation, called light, whose frequency and wavelength can be explicitly computed.*

This phenomenon can be observed is a variety of situations, such as the usual light bulbs, where electrons get decelerated by the filament, acting as a resistor, or in usual fire, which is a chemical reaction, with the electrons moving around, as they do in any chemical reaction, or in more complicated machinery like nuclear plants, particle accelerators, and so on, leading there to all sorts of eerie glows, of various colors.

In order to discuss now frequency, wavelength and color, which are in fact the same thing, let us first recall from chapter 4 that in the 1D case, we have:

THEOREM 6.8. *The 1D wave equation $\ddot{\varphi} = v^2\varphi''$ 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. Any solution of the 1D wave equation appears as a linear combination of such basic solutions.

PROOF. This is something that we basically know from chapter 4, with a few things added, coming from our present familiarity with 2D, 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:

$$\ddot{\varphi} = -w^2\varphi \quad , \quad \varphi'' = -k^2\varphi$$

Thus, the wave equation is indeed satisfied, with speed $v = w/k$:

$$\ddot{\varphi} = \left(\frac{w}{k}\right)^2 \varphi'' = v^2 \varphi''$$

(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, with the linear combination mentioned there being of course infinite, is something mathematical and standard, coming from Fourier analysis, and with the d'Alembert theorem from chapter 4 helping. More on this later. □

As a first observation, the above result invites the use of complex numbers. Indeed, we can write the solutions that we found in a more convenient way, as follows:

$$\varphi(x) = Re\left[A\, e^{i(kx-wt+\delta)}\right]$$

And we can in fact do even better, by absorbing the quantity $e^{i\delta}$ into the amplitude A, which becomes now a complex number, and writing our formula as:

$$\varphi = Re(\widetilde{\varphi}) \quad , \quad \widetilde{\varphi} = \widetilde{A} e^{i(kx-wt)}$$

Moving ahead now towards what we wanted to do, color, let us formulate:

DEFINITION 6.9. *A monochromatic plane wave is a solution of the 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 6.8, 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 6.8. And second is the assumption, in connection with the Fourier decomposition result from the end of Theorem 6.8, 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.

So, this was for the story of the monochromatic plane waves, or light. Many more things can be said here, such as talking about polarization too, which is an important extra feature that light might have, in 3D. However, all this heavily relies on the mechanism from Fact 6.7, coming from electromagnetism and the Maxwell equations, which are of genuine 3D nature, scheduled for later in this book. So, in the hope that you got it, we sort of have light in 2D, perhaps by cheating a bit, and more on all this, later.

In practice now, we have various types of light, depending on frequency and wavelength. These are usually referred to as "electromagnetic waves", but for keeping things

simple, we will keep using the term "light". The classification 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 6.10. *The wave equation inside a linear, homogeneous medium is*

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

with v being the speed of light inside the medium, which is 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 coming, as usual with such things, from the Maxwell equations, that we have not learned yet. So, more on this, and also on the formula of n, with details regarding the various quantities involved there, later in this book. □

As an observation here, while the above is something quite simple, mathematically speaking, as we will discover later in this book, when talking Maxwell equations and applications, 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 6.10 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, in order to talk about optics, we have:

FACT 6.11. *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 now of the above, of great practical interest, we have:

THEOREM 6.12. *Light can be decomposed, by using a prism.*

PROOF. This follows from Fact 6.11. 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 6.13. *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.

And with this, end of our discussion regarding light. We learned many interesting things here, substantially improving our previous knowledge from chapter 4, but when looking more in detail at what we did, everything remains quite complicated, in need of electromagnetism, and the Maxwell equations. So, more on this, later in this book.

6c. Relativity, speeds

With light discussed, as a main application, time to go back to relativity theory, as developed in chapter 2, with a continuation of the story there, at $N=2$ and higher.

Let us first recall from chapter 2 that, based on various experiments, by Fizeau, and then Michelson-Morley and others, on some previous theory and findings by Lorentz and

others, in the related context of electromagnetism, and on a bit of abstract thinking too, Einstein came upon the following general principles, which agree of course with the light theory developed above, and which obviously contradict classical mechanics:

FACT 6.14 (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,* $v \not> c$.

In view of this, there are many things to be done, namely fix the speed addition formula, as for $c + v = c$ to hold, then, in view of $v = d/t$, see if something is wrong with distance d, or with time t, or with both, and finally, in view of $p = mv$ and $E = mv^2/2$, have a look as well at momentum and energy, I mean find the bugs, and fixes.

Let us begin with the speed addition problem, $c + v = c$, which in $c = 1$ units, that we will use in what follows, simply reads $1 + v = 1$. We already know the solution to this problem in 1D, from chapter 2, and in higher dimensions, we have:

QUESTION 6.15. *What is the correct analogue of the Einstein summation formula,*

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

in 2 *and* 3 *dimensions?*

In order to discuss this question, let us attempt to construct $u +_e v$ in arbitrary dimensions, just by using our common sense and intuition. When the vectors $u, v \in \mathbb{R}^N$ are proportional, we are basically in 1D, and so our addition formula must satisfy:

$$u \sim v \implies u +_e v = \frac{u + v}{1 + < u, v >}$$

To be more precise, we are making use here of the scalar product operation $< u, v >$ for the vectors $u, v \in \mathbb{R}^N$, which is given by the following formula:

$$< u, v >= \sum_{i=1}^{N} u_i v_i$$

Looking now at what we have, our $u \sim v$ formula 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, involving the scalar product $< u, v >$, is not bad, as a start, but we must add a correction term to it, for the above requirements to be satisfied, and with the correction term vanishing when $u \sim v$. So, we are led to:

PUZZLE 6.16. *How to correctly define the speed addition as*

$$u +_e v = \frac{u + v + \gamma_{uv}}{1 + < u, v >}$$

as for the following to happen, in analogy with what happens in 1D:

(1) γ_{uv} *vanishes when* $u \sim v$.
(2) $||u|| = 1 \implies u +_e v = u$.
(3) $||v|| = 1 \implies u +_e v = v$.

Getting to work now, let us first study the condition $||u|| = 1 \implies u +_e v = u$. We first have the following computation, valid for any two vectors $u, v \in \mathbb{R}^N$:

$$\begin{aligned} u +_e v = u &\iff \frac{u + v + \gamma_{uv}}{1 + < u, v >} = u \\ &\iff u + v + \gamma_{uv} = u + < u, v > u \\ &\iff \gamma_{uv} = < u, v > u - v \end{aligned}$$

Now if we want this to happen, under the assumption $||u|| = 1$, which in terms of scalar products reads $< u, u >= 1$, and with the extra condition that the correction term γ_{uv} must vanish when $u \sim v$, the solution is a no-brainer, as follows:

$$\gamma_{uv} = < u, v > u - < u, u > v$$

Next, let us study the other condition that we have, namely $||v|| = 1 \implies u +_e v = v$. As before, we have the following computation, valid for any two vectors $u, v \in \mathbb{R}^N$:

$$\begin{aligned} u +_e v = v &\iff \frac{u + v + \gamma_{uv}}{1 + < u, v >} = v \\ &\iff u + v + \gamma_{uv} = v + < u, v > v \\ &\iff \gamma_{uv} = < u, v > v - u \end{aligned}$$

Now if we want this to happen, under the assumption $||v|| = 1$, which in terms of scalar products reads $< v, v >= 1$, and with the extra condition that the correction term γ_{uv} must vanish when $u \sim v$, the solution is again a no-brainer, as follows:

$$\gamma_{uv} = < u, v > v - < v, v > u$$

Which is quite bad news, because this new correction term γ_{uv} is different from the previous correction term γ_{uv} that we computed, unless of course at $N = 1$, where both correction terms are $\gamma_{uv} = 0$. And, there does not seem to be any fix, for this.

Damn, I would say. So, we are in trouble, our conclusion being as follows:

CONCLUSION 6.17. *Our puzzle cannot really be solved at $N \geq 2$, and if we want to develop relativity theory there, we basically have to choose between:*

(1) $\gamma_{uv} =< u, v > u - < u, u > v$, *leading to* $||u|| = 1 \implies u +_e v = u$.
(2) $\gamma_{uv} =< u, v > v - < v, v > u$, *leading to* $||v|| = 1 \implies u +_e v = v$.

So, what to do? Relax a bit I guess, stop coffee and cigarettes, and take it easy. After all, what we did is not that bad, and going for the first choice, which is the most natural one, physically speaking, we have material for a modest Proposition, as follows:

PROPOSITION 6.18. *When defining the Einstein speed summation as*

$$u +_e v = \frac{u + v + < u, v > u - < u, u > v}{1 + < u, v >}$$

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. We already know this from the above discussion, with (1,2) coming from our computations before, and with (3,4) coming from the same computations, via:

$$< u, v > u - < u, u > v \neq < u, v > v - < v, v > u$$

Thus, proposition proved, and good work that we did. Very nice. □

Looking now at Proposition 6.18 from an abstract, mathematical perspective, there are still a few things missing from there, which can be summarized as follows:

QUESTION 6.19. *Can we fine-tune the Einstein speed summation into*

$$u +_e v = \frac{u + v + \lambda(< u, v > u - < u, u > v)}{1 + < u, v >}$$

with $\lambda \in \mathbb{R}$, chosen such that $||u|| = 1 \implies \lambda = 1$, as to have:

(1) $||u||, ||v|| < 1 \implies ||u +_e v|| < 1$.
(2) $||v|| = 1 \implies ||u +_e v|| = 1$.

Obviously, as simplest answer, λ must be some well-chosen function of $||u||$, or rather of $||u||^2$, because it is always better to use square norms, when possible. But then, with this idea in mind, after a few computations we are led to the following solution:

$$\lambda = \frac{1}{1 + \sqrt{1 - ||u||^2}}$$

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 6.20. *When defining the Einstein speed summation as*

$$u +_e v = \frac{1}{1+ < u, v >} \left(u + v + \frac{< u, v > u - < u, u > v}{1 + \sqrt{1 - ||u||^2}} \right)$$

in $c = 1$ units, 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 already, know from Proposition 6.18.

(2) In order to simplify notation, let us set $\delta = \sqrt{1 - ||u||^2}$, which is the inverse of the quantity $\gamma = 1/\sqrt{1 - ||u||^2}$. With this convention, we have:

$$\begin{aligned} u +_e v &= \frac{1}{1+ < u, v >} \left(u + v + \frac{< u, v > u - ||u||^2 v}{1 + \delta} \right) \\ &= \frac{(1 + \delta + < u, v >)u + (1 + \delta - ||u||^2)v}{(1+ < u, v >)(1 + \delta)} \end{aligned}$$

Taking the squared norm and computing gives the following formula:

$$||u +_e v||^2 = \frac{(1 + \delta)^2 ||u + v||^2 + (||u||^2 - 2(1 + \delta))(||u||^2 ||v||^2 - < u, v >^2)}{(1+ < u, v >)^2 (1 + \delta)^2}$$

Now this formula can be further processed by using $\delta = \sqrt{1 - ||u||^2}$, and we get:

$$||u +_e v||^2 = \frac{||u + v||^2 - ||u||^2 ||v||^2 + < u, v >^2}{(1+ < u, v >)^2}$$

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 already know, from Proposition 6.18.

(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. □

Still with me I hope, after all these computations. And the good news is that, we did indeed some first-class work, because the formula in Theorem 6.20 is the good one, confirmed by experimental physics. For more on all this, see Einstein [28].

6d. Mass, energy

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 6.21. *Relativistic time is subject to Lorentz dilation*

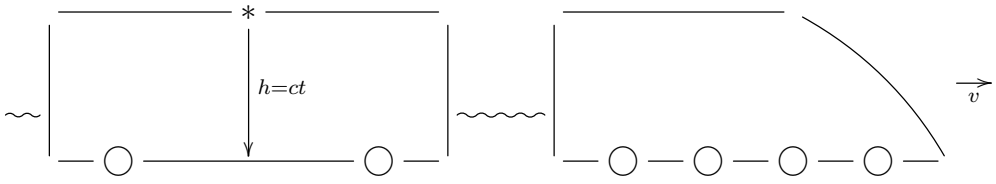
$$t \to \gamma t$$

where the number $\gamma \geq 1$, called Lorentz factor, is given by the formula

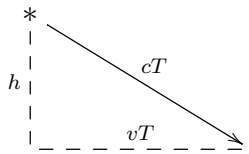
$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}$$

with v being the moving speed, at which time is measured.

PROOF. Assume indeed that we have a train, moving to the right with speed v, through vacuum. In order to compute the height h of the train, the passenger onboard switches on the ceiling light bulb, measures the time t that the light needs to hit the floor, by traveling at speed c, and concludes that the train height is $h = ct$:



On the other hand, an observer on the ground will see here something different, namely a right triangle, with on the vertical the height of the train h, on the horizontal the distance vT that the train has traveled, and on the hypotenuse the distance cT that light has traveled, with T being the duration of the event, according to his watch:



Now by Pythagoras applied to this triangle, we have the following formula:

$$h^2 + (vT)^2 = (cT)^2$$

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. We have here the following result:

THEOREM 6.22. *Relativistic length is subject to Lorentz contraction*

$$L \to L/\gamma$$

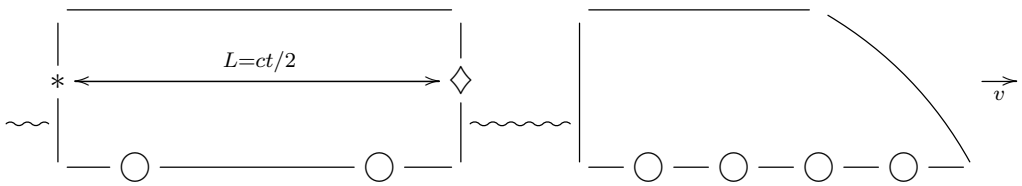
where the number $\gamma \geq 1$, called Lorentz factor, is given by the usual formula

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}$$

with v being the moving speed, at which length is measured.

PROOF. As before in the proof of Theorem 6.21, 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. □

Next, for our theory to be complete, we still have to discuss mass and energy. Things here are quite tricky, and as a first objective, we would like to fix the mass and momentum conservation equations for the plastic collisions, from chapter 2, namely:

$$m = m_1 + m_2 \quad , \quad mv = m_1 v_1 + m_2 v_2$$

Which 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 6.23. *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:

$$M = M_1 + M_2 \quad , \quad Mv = M_1v_1 + M_2v_2$$

Finally, it remains to discuss energy. And here, still following Einstein, we have:

THEOREM 6.24. *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 6.23, as follows:

(1) Given an object of rest mass $m > 0$, consider its relativistic mass $M = \gamma m$, as appearing in Fact 6.23, and then consider the following quantity:

$$\mathcal{E} = Mc^2$$

We know from Fact 6.23 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.

(2) Let us compute now $\mathcal{E}$. 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. □

What is next? Some more discussion I guess about $E = mc^2$, in order to better understand this formula. And here, we can report, with pleasure, the following:

THEOREM 6.25. *An atomic bomb based on a glass of water releasing all its $E = mc^2$ energy is equivalent to the Giza Pyramid hitting you at 3 km/s.*

PROOF. When converting $E = mc^2$ into kinetic energy of a body M, the formula is:

$$mc^2 = \frac{Mv^2}{2} \implies v = \sqrt{\frac{2m}{M}}\, c$$

In our case the glass of water m, glass included, is about 300 grams, and the Giza Pyramid M is about 6 million tons. Also, the speed of light in vacuum is $c \simeq 300,000\,\mathrm{km/s}$, and more on this later, when discussing electromagnetism. Thus the impact speed is:

$$v = \sqrt{\frac{2 \times 3 \times 10^{-1}}{6 \times 10^9}}\, c = \frac{c}{10^5} = 3\,\mathrm{km/s}$$

This being said, do not worry about this, because the glass of water, as long as it stays far away from a big source of energy, needed for triggering the $E = mc^2$ effect, will certainly not explode. However, we will see later, when doing quantum mechanics, that certain atomic bombs, based on other materials, can be constructed. Not to mention of course the stars in the sky, whose functioning is based on $E = mc^2$ too. □

6e. Exercises

This was an exciting physics chapter, and as exercises, we have:

EXERCISE 6.26. *Learn more about pressure and temperature, in 3D.*

EXERCISE 6.27. *Learn more about the heat kernel, and normal variables too.*

EXERCISE 6.28. *Learn about the 2D solutions of the wave equation.*

EXERCISE 6.29. *Learn a bit about the 3D mechanism producing light.*

EXERCISE 6.30. *Learn more about the Snell law, and optics in general.*

EXERCISE 6.31. *Scan for potential bugs our Lorentz factor proofs.*

EXERCISE 6.32. *Learn more about relativistic mass and momentum.*

EXERCISE 6.33. *Do some more numerics, of your own, for $E = mc^2$.*

As bonus exercise, learn a bit about the various types of stars.

CHAPTER 7

Vector calculus

7a. Linear algebra

Welcome back to mathematics, and many things for us to be done here, as a continuation of what we did in chapter 5. We have seen indeed in chapter 6, when doing physics, that of great use are the scalar products of vectors $<,>$, whose functioning we must therefore better understand. Also, we got in chapter 6 into partial derivatives and the Laplace operator Δ too, whose functioning we must better understand too.

In addition to this, remember gravity, and the Kepler-Newton theorem, stating that the planets move around the Sun on elliptic orbits, that we haven't talked about yet. And with this being certainly something quite complicated, most likely requiring a number of serious mathematical preliminaries, of both algebraic and analytic nature.

Getting started now, in what regards the scalar products, we have:

THEOREM 7.1. *We can talk about scalar products and lengths in* $\mathbb{R}^N$*,*

$$< x, y >= \sum_i x_i y_i \quad , \quad ||x|| = \sqrt{\sum_i x_i^2}$$

which are related by the following conversion formulae,

$$||x|| = \sqrt{< x, x >} \quad , \quad < x, y >= \frac{||x+y||^2 - ||x-y||^2}{4}$$

and the following happen:

(1) $< \lambda x, y >=< x, \lambda y >= \lambda < x, y >$.
(2) $< x + y, z >=< x, z > + < y, z >$.
(3) $< x, y + z >=< x, z > + < y, z >$.
(4) $||\lambda x|| = |\lambda| \cdot ||x||$.
(5) $| < x, y > | \leq ||x|| \cdot ||y||$.
(6) $||x + y|| \leq ||x|| + ||y||$.
(7) $x \perp y \iff < x, y >= 0$*, by definition.*
(8) $< x, y >= ||x|| \cdot ||y|| \cdot \cos t$*, with* t *being the angle between* x, y*, by definition.*
(9) $< x, y >=< x', y >=< x, y' >$*, prime being the projection on the other vector.*

PROOF. We can certainly talk about scalar products and lengths, as above, and with the second conversion formula, called polarization identity, coming from:

$$\begin{aligned}||x+y||^2-||x-y||^2 &= <x+y,x+y>-<x-y,x-y>\\ &= ||x||^2+||y||^2+2<x,y>-||x||^2-||y||^2+2<x,y>\\ &= 4<x,y>\end{aligned}$$

By the way, talking useful general identities, we have as well a parallelogram rule, that I forgot to mention in the above statement, which is as follows:

$$\begin{aligned}||x+y||^2+||x-y||^2 &= <x+y,x+y>+<x-y,x-y>\\ &= ||x||^2+||y||^2+2<x,y>+||x||^2+||y||^2-2<x,y>\\ &= 2(||x||^2+||y||^2)\end{aligned}$$

As for the various claims in the statement, their proof is as follows:

(1-4) All the verifications here are trivial, easy exercise for you.

(5) Given two vectors $x,y\in\mathbb{R}^N$, consider the following function $f:\mathbb{R}\to\mathbb{R}$:

$$\begin{aligned}f(t) &= ||x+ty||^2\\ &= <x+ty,x+ty>\\ &= ||x||^2+2t<x,y>+t^2||y||^2\end{aligned}$$

Thus f is a degree 2 polynomial, and since this polynomial is positive, its discriminant must be negative, $\Delta\leq 0$. But the discriminant is given by the following formula:

$$\Delta=4<x,y>^2-4||x||^2||y||^2$$

Thus we have the Cauchy-Schwarz inequality, $|<x,y>|\leq||x||\cdot||y||$, as claimed.

(6) This can be proved by raising to the square and simplifying, as follows:

$$\begin{aligned}||x+y||\leq||x||+||y|| &\iff ||x+y||^2\leq||x||^2+||y||^2+2||x||\cdot||y||\\ &\iff ||x||^2+||y||^2+2<x,y>\leq||x||^2+||y||^2+2||x||\cdot||y||\\ &\iff <x,y>\leq||x||\cdot||y||\end{aligned}$$

Indeed, the last inequality holds by (5), so we have our triangle inequality.

(7-9) These assertions are something more subtle, because in the lack of good intuition, do we really know what orthogonality and angles really are, in $\mathbb{R}^N$. So, the best is to proceed as indicated, with the notions of orthogonality and angles being defined as in the statement, and with this being compatible with what happens in 2D, say exercise for you, and with the last formula being something elementary, again exercise for you. □

Next, let us discuss linear algebra, that we will soon discover to be related to everything advanced in mathematics, of both algebraic and analytic nature. We will be interested in the transformations $f:\mathbb{R}^N\to\mathbb{R}^M$ which are linear, in the following sense:

THEOREM 7.2. *The maps $f : \mathbb{R}^N \to \mathbb{R}^M$ which are linear, in the sense that*

$$f(x+y) = f(x) + f(y) \quad , \quad f(\lambda x) = \lambda f(x)$$

are the maps of the following form, with $A \in M_{M\times N}(\mathbb{R})$ being a rectangular matrix:

$$f(x) = Ax$$

Moreover, the matrix A can be recaptured via the formula $A_{ij} =< f(e_j), e_i >$.

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

(1) A linear map $f : \mathbb{R}^N \to \mathbb{R}^M$ must send a vector $x \in \mathbb{R}^N$ to a certain vector $f(x) \in \mathbb{R}^M$, all whose components are linear combinations of the components of x, so:

$$f\begin{pmatrix} x_1 \\ \vdots \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} A_{11}x_1 + \ldots + A_{1N}x_N \\ \vdots \\ \vdots \\ A_{M1}x_1 + \ldots + A_{MN}x_N \end{pmatrix}$$

Now the parameters $A_{ij} \in \mathbb{R}$ can be regarded as being the entries of a rectangular matrix $A \in M_{M\times N}(\mathbb{R})$, and with the usual convention for matrix multiplication, namely "multiply the rows of the first matrix by the columns of the second matrix", we have:

$$f(x) = Ax$$

(2) The second assertion is clear too, because if we denote as usual by $e_1, e_2, e_3, \ldots$ the standard bases of our vector spaces, we have the following computation:

$$< f(e_j), e_i >=< Ae_j, e_i >= (Ae_j)_i = A_{ij}$$

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

At the level of examples, we have the following statement, which is a must-know:

THEOREM 7.3. *The following happen, in the plane:*

(1) *The rotation of angle $t \in \mathbb{R}$ is given by the following matrix:*

$$R_t = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}$$

(2) *The symmetry with respect to the Ox axis rotated by $t/2 \in \mathbb{R}$ is given by:*

$$S_t = \begin{pmatrix} \cos t & \sin t \\ \sin t & -\cos t \end{pmatrix}$$

(3) *The projection on the Ox axis rotated by $t/2 \in \mathbb{R}$ is given by:*

$$P_t = \frac{1}{2}\begin{pmatrix} 1+\cos t & \sin t \\ \sin t & 1-\cos t \end{pmatrix}$$

(4) *However, the non-trivial translations are not linear maps, in our sense.*

PROOF. All this is well-known and elementary, the idea being as follows:

(1) The rotation being linear, it must correspond to a certain matrix:

$$R_t = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

But we can guess this matrix, via its action on the standard coordinate vectors $\binom{1}{0}$ and $\binom{0}{1}$. Indeed, 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.

(2) This is again clear on a picture, drawn with $t/2$ instead of t, as indicated.

(3) Again, picture drawn with $t/2$, as indicated, plus some easy trigonometry.

(4) Indeed, the non-trivial translations satisfy $f(0) \neq 0$, so they are not linear. □

As a next piece of general theory, we have the following key definition:

DEFINITION 7.4. *A linear map $f : \mathbb{R}^N \to \mathbb{R}^N$ is called diagonalizable if there is a basis $v_1, \ldots, v_N \in \mathbb{R}^N$ such that f multiplies by λ_i in the direction v_i:*

$$f(v_i) = \lambda_i v_i$$

In terms of the writing $f(x) = Ax$, we say that the corresponding matrix $A \in M_N(\mathbb{R})$ is diagonalizable, with eingenvectors v_i and eigenvalues λ_i.

Here the system of vectors $v_1, \ldots, v_N \in \mathbb{R}^N$ being a basis means by definition that each vector $x \in \mathbb{R}^N$ must decompose uniquely as $x = \sum_i c_i v_i$, and we can see that the diagonalizability assumption uniquely determines f, which follows to be given by:

$$f\left(\sum_i c_i v_i\right) = \sum_i \lambda_i c_i v_i$$

Obviously, being diagonalizable means to be "good", and being not diagonalizable means to be "bad". In order to understand this, what good and bad mean in linear algebra, let us work out some examples. We have here the following result:

PROPOSITION 7.5. *The following happen:*

(1) *The rotation R_t is not diagonalizable, unless at $t = 0$ where it is the identity, $R_0 = 1$, and at $t = \pi$ where it is minus the identity, $R_\pi = -1$.*
(2) *The symmetry S_t is diagonalizable, with eigenvectors on the symmetry axis, and on its orthogonal, with respective eigenvalues $1, -1$.*
(3) *The projection P_t is diagonalizable, with the eigenvectors exactly as for the symmetry S_t, this time with respective eigenvalues $1, 0$.*
(4) *In fact, any projection is diagonalizable, with eigenvectors on it image, and on the orthogonal of its image, with respective eigenvalues $1, 0$.*

PROOF. All this is self-explanatory and, we insist, with no need for any computation, and exercise for you, to figure out how all this works, just by drawing pictures, and thinking. Of course, if eager for computations, do not worry, we will have some, later. □

Still in relation with diagonalization, at the general level, we have:

THEOREM 7.6. *Assuming that a matrix $A \in M_N(\mathbb{R})$ is diagonalizable, with eigenvectors $v_1, \ldots, v_N$ and corresponding eigenvalues $\lambda_1, \ldots, \lambda_N$, we have*

$$A = PDP^{-1}$$

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. We have $Pe_i = v_i$, where $\{e_i\}$ 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. □

As an illustration for this, you can have some fun with the various matrices from Proposition 7.3, with in each case the corresponding diagonalization formula $A = PDP^{-1}$ coming without much pain, because Proposition 7.5 tells us what both P, D are, in each case, and the only piece of work remaining is that of figuring out what P^{-1} is. Enjoy.

Summarizing, the diagonalizable matrices are the "good" ones, and their diagonalization is quite often a matter of doing some geometry. Regarding the non-diagonalizable matrices, these actually fall into two classes, "bad" and "evil". The bad ones are those which diagonalize over $\mathbb{C}$, with a main example here being the rotation R_t:

THEOREM 7.7. *The rotation of angle $t \in \mathbb{R}$ in the plane diagonalizes as:*

$$R_t = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}\begin{pmatrix} e^{-it} & 0 \\ 0 & e^{it} \end{pmatrix}\begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix}$$

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:

$$R_t\begin{pmatrix} 1 \\ i \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}\begin{pmatrix} 1 \\ i \end{pmatrix} = \begin{pmatrix} \cos t - i\sin t \\ i\cos t + \sin t \end{pmatrix} = e^{-it}\begin{pmatrix} 1 \\ i \end{pmatrix}$$

We have as well a second eigenvector, as follows:

$$R_t\begin{pmatrix} 1 \\ -i \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}\begin{pmatrix} 1 \\ -i \end{pmatrix} = \begin{pmatrix} \cos t + i\sin t \\ -i\cos t + \sin t \end{pmatrix} = e^{it}\begin{pmatrix} 1 \\ -i \end{pmatrix}$$

Thus our rotation matrix R_t is indeed diagonalizable over $\mathbb{C}$, with the passage matrix and diagonal form being, according to the above formulae, as follows:

$$P = \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} \quad , \quad D = \begin{pmatrix} e^{-it} & 0 \\ 0 & e^{it} \end{pmatrix}$$

Now by inverting P, an easy task, we are led to the conclusion in the statement. □

As for the evil matrices, these are truly evil, a basic example being as follows:

THEOREM 7.8. *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 $\binom{x}{0}$. □

As a next topic of discussion, given $A \in M_N(\mathbb{R})$, let us define its determinant as being the signed volume of the parallelepiped made by its column vectors $v_1, \ldots, v_N \in \mathbb{R}^N$:

$$\det A = \pm vol(< v_1, \ldots, v_N >)$$

To be more precise here, the above sign is by definition $+$ when one can continuously pass from the standard basis $e_1, \ldots, e_N$ to the system of vectors $v_1, \ldots, v_N$, and is $-$ otherwise. With this convention, we have the following key result, in the 2D case:

THEOREM 7.9. *In* 2 *dimensions we have the following formula,*

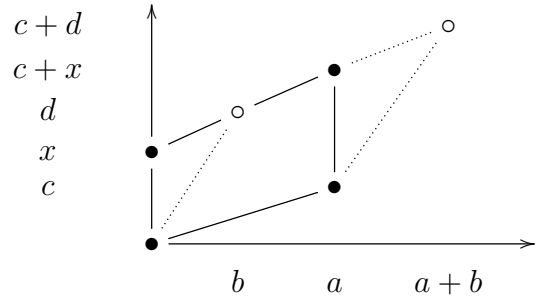
$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

with det *being as above, signed area of the parallelogram formed by* $\binom{a}{c}, \binom{b}{d}$.

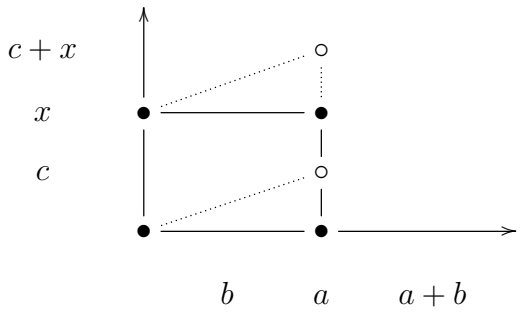
PROOF. Up to a straightforward sign discussion, we can assume $a, b, c, d > 0$. Moreover, by switching if needed the vectors $\binom{a}{c}, \binom{b}{d}$, we can assume that we have:

$$\frac{a}{c} > \frac{b}{d}$$

We want to compute the area of the parallelogram formed by $\binom{a}{c}, \binom{b}{d}$. In order to do so, let us slide the upper side of the parallelogram downwards left, until we reach Oy. Our parallelogram, which has not changed its area in this process, becomes:



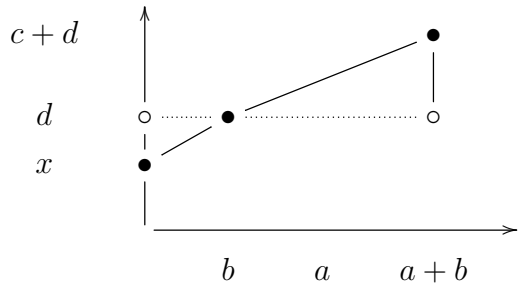
Moreover, we can further modify this parallelogram, once again by not altering its area, by sliding the right side downwards, until we reach the Ox axis:



Let us compute now the area. Since our two sliding operations have not changed the area of the original parallelogram, this area is given by the following formula:

$$A = ax$$

In order to compute the quantity x, observe that in the context of the first move, we have two similar triangles, according to the following picture:



Thus, we are led to the following equation for the number x:

$$\frac{d-x}{b} = \frac{c}{a}$$

But this gives $x = d - bc/a$, so the area of our parallelogram, or rather of the final rectangle obtained from it, which has the same area as the original parallelogram, is:

$$A = ax = ad - bc$$

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

Finally, let us discuss invertibility questions for the matrices. In 2D, we have:

THEOREM 7.10. *We have the following inversion formula, for the* 2×2 *matrices:*

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

When $ad - bc = 0$*, the matrix is not invertible.*

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

(1) As a first observation, when $ad - bc = 0$ we must have, for some $\lambda \in \mathbb{R}$:

$$b = \lambda a \quad , \quad d = \lambda c$$

Thus our matrix must be of the following special type:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & \lambda a \\ a & \lambda c \end{pmatrix}$$

But in this case the columns are proportional, so the linear map associated to the matrix is not invertible, and so the matrix itself is not invertible either. By the way, observe that all this follows as well as a consequence of Theorem 7.9.

(2) When $ad-bc\neq 0$, let us look for an inversion formula of the following type:

$$\begin{pmatrix}a&b\\c&d\end{pmatrix}^{-1}=\frac{1}{ad-bc}\begin{pmatrix}*&*\\*&*\end{pmatrix}$$

We must therefore solve the following equations:

$$\begin{pmatrix}a&b\\c&d\end{pmatrix}\begin{pmatrix}*&*\\*&*\end{pmatrix}=\begin{pmatrix}ad-bc&0\\0&ad-bc\end{pmatrix}$$

But the obvious solution here is as follows:

$$\begin{pmatrix}a&b\\c&d\end{pmatrix}\begin{pmatrix}d&-b\\-c&a\end{pmatrix}=\begin{pmatrix}ad-bc&0\\0&ad-bc\end{pmatrix}$$

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

So long for the foundations of linear algebra, and basic 2D illustrations and results. We will be back to all this on several occasions, in what follows, when really needing more specialized results. Among others, we will learn several tricks for diagonalizing the higher dimensional matrices, concrete or abstract, and we will see as well that Theorems 7.9 and 7.10 have some interesting higher dimensional analogues. More later.

7b. Ellipses, conics

With linear algebra understood, we are good for some astronomy. Looking up, to the sky, the first thing that you see is the Sun, seemingly moving around the Earth on a circle. However, this circle is rather a deformed circle, called ellipse. And good news, a full theory of ellipses is available, and this since the ancient Greeks, as follows:

THEOREM 7.11. *The ellipses, taken centered at the origin* 0*, and squarely oriented with respect to* Oxy*, can be defined in* 4 *possible ways, as follows:*

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

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

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

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

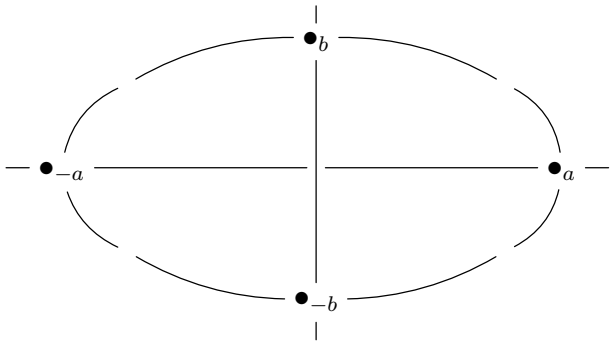
(3) *As the curves appearing when drawing a circle, from various perspectives:*

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

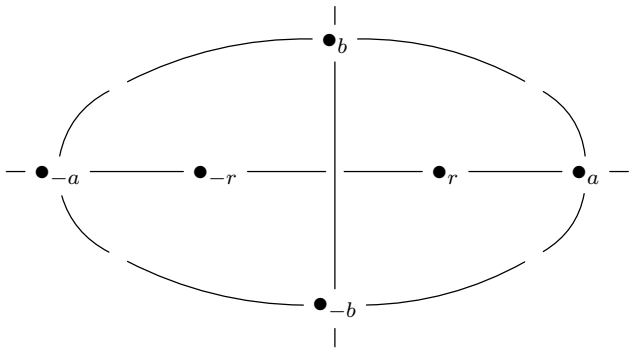
(4) *As the closed non-degenerate curves appearing by cutting a cone with a plane.*

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

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



(2) Let us prove now that such an ellipse has two focal points, as stated in (2). We must look for a number $r > 0$, and a number $l > 0$, such that our ellipse appears as $d(z, p) + d(z, q) = l$, with $p = (0, -r)$ and $q = (0, r)$, according to the following picture:



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

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

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

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

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

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

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

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

(6) Thus, our result holds indeed, and with the numbers $l, r > 0$ appearing, and no surprise here, via the formulae $l = 2a$ and $r = \sqrt{a^2 - b^2}$, found in (3) above.

(7) Getting back now to our theorem, we have two other assertions there at the end, labeled (3,4). But, thinking a bit, these assertions are in fact equivalent. And in what regards the assertion (4), this can be established indeed, by doing some 3D computations, that we will leave here as an instructive exercise, for you. And with the promise that we will come back to this in a moment, with a full proof, in a more general setting. □

Coming next, as a key analytic result regarding the ellipses, we have:

THEOREM 7.12. *The area of an ellipse, given by the equation*

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

with $a, b > 0$ being half the size of a box containing the ellipse, is $A = \pi ab$.

PROOF. The idea is that of cutting the ellipse into vertical slices. Indeed, the length of the vertical ellipse slice at x is given by the following formula:

$$l(x) = 2b\sqrt{1 - \frac{x^2}{a^2}}$$

We conclude that the area of the ellipse is given by the following formula:

$$\begin{aligned} A &= 2b \int_{-a}^{a} \sqrt{1 - \frac{x^2}{a^2}}\, dx \\ &= \frac{4b}{a} \int_{0}^{a} \sqrt{a^2 - x^2}\, dx \\ &= 4ab \int_{0}^{1} \sqrt{1 - y^2}\, dy \\ &= \pi ab \end{aligned}$$

Finally, as a verification, for $a = b = 1$ we get $A = \pi$, as we should. □

Still talking ellipses, in what regards the length things are quite tricky, as follows:

THEOREM 7.13. *The length of an ellipse, given by $(x/a)^2 + (y/b)^2 = 1$, is*

$$L = 4 \int_{0}^{\pi/2} \sqrt{a^2 \sin^2 t + b^2 \cos^2 t}\, dt$$

and with this integral being generically not computable.

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

(1) To start with, in the case where our ellipse is a circle, say of radius R, the area and length are related by the "pizza" formula $A = LR/2$, as we know well from chapter 5. The problem, however, is that such things will not work for general ellipses.

(2) So, what is the length of a curve $\gamma : [a, b] \to \mathbb{R}^2$? Good question, and in answer, a physicist would say that this is the quantity obtained by integrating the magnitude of the velocity vector over the curve, with respect to time, leading to:

$$L(\gamma) = \int_{a}^{b} ||\gamma'(t)||dt$$

(3) Regarding now mathematicians, these would say that the length of a curve is the following quantity, with $(t_0 = a, t_1, \ldots, t_{n-1}, t_n = b)$ being a uniform division of (a, b):

$$L(\gamma) = \lim_{n \to \infty} \sum_{i=1}^{n} ||\gamma(t_i) - \gamma(t_{i-1})||$$

But, by using the fundamental theorem of calculus, we can write this as follows:

$$L(\gamma) = \lim_{n\to\infty} \sum_{i=1}^{n} \left|\left| \int_{t_{i-1}}^{t_i} \gamma'(t)dt \right|\right|$$

And the point now is that, by doing some standard analysis, that we will leave here as an instructive exercise, we are led to the formula in (2).

(4) Getting back now to the ellipses, we can compute their length, as follows:

$$\begin{aligned} L &= 4\int_0^{\pi/2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt \\ &= 4\int_0^{\pi/2} \sqrt{\left(\frac{da\cos t}{dt}\right)^2 + \left(\frac{db\sin t}{dt}\right)^2}\, dt \\ &= 4\int_0^{\pi/2} \sqrt{a^2\sin^2 t + b^2\cos^2 t}\, dt \end{aligned}$$

(5) As for the last assertion, when $a = b = R$ we get of course $L = 2\pi R$, as we should, but in general, when $a \neq b$, there is no trick for computing the above integral. □

Getting back now to the astronomy considerations from the beginning of this section, let us settle as well the question of wandering asteroids. Observations show that these can travel on parabolas and hyperbolas, so what we need as mathematics is a unified theory of ellipses, parabolas and hyperbolas. And fortunately, this theory exists, based on:

DEFINITION 7.14. *A conic is a plane algebraic curve of the form*

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

with $P \in \mathbb{R}[x, y]$ being of degree ≤ 2.

As basic examples of conics, we have the ellipses, parabolas and hyperbolas. The simplest examples of these are as follows, with the ellipse actually being a circle:

$$x^2 + y^2 = 1 \quad , \quad x^2 = y \quad , \quad xy = 1$$

Observe that, due to our assumption $\deg P \leq 2$, we have as conics some degenerate curves as well, such as lines, $\emptyset$, and $\mathbb{R}^2$ itself, coming from $\deg P \leq 1$, as follows:

$$x = 0 \quad , \quad 1 = 0 \quad , \quad 0 = 0$$

This might suggest to replace our assumption $\deg P \leq 2$ by $\deg P = 2$, but we will not do so, because $\deg P = 2$ does not rule out degenerate situations, such as:

$$x^2 + y^2 = -1 \quad , \quad x^2 + y^2 = 0 \quad , \quad x^2 = 0 \quad , \quad xy = 0$$

In fact, what we get here are $\emptyset$, points, lines, and pairs of lines, so in the end, assuming $\deg P = 2$ instead of $\deg P \leq 2$ would only rule out $\mathbb{R}^2$ itself, which is not worth it.

Summarizing, our notion of conic from Definition 7.14 looks quite reasonable, so let us agree on this notion. Getting now to classification matters, we first have:

THEOREM 7.15. *Up to non-degenerate linear transformations of the plane,*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to A \begin{pmatrix} x \\ y \end{pmatrix}$$

with $\det A \neq 0$*, the conics fall into two classes, as follows:*

(1) *Non-degenerate: circles, parabolas, hyperbolas.*
(2) *Degenerate:* $\emptyset$*, points, lines, pairs of lines,* $\mathbb{R}^2$*.*

PROOF. As a first observation, looks like we forgot the ellipses, but via linear transformations these become circles, so things fine. As for the proof, this goes as follows:

(1) Consider an arbitrary conic, written as follows, with $a, b, c, d, e, f \in \mathbb{R}$:

$$ax^2 + by^2 + cxy + dx + ey + f = 0$$

(2) Assume first $a \neq 0$. By making a square out of ax^2, up to a linear transformation in (x, y), we can get rid of the term cxy, and we are left with:

$$ax^2 + by^2 + dx + ey + f = 0$$

In the case $b \neq 0$ we can make two obvious squares, and again up to a linear transformation in (x, y), we are left with an equation as follows:

$$x^2 \pm y^2 = k$$

In the case of positive sign, $x^2 + y^2 = k$, the solutions are the circle, when $k \geq 0$, the point, when $k = 0$, and $\emptyset$, when $k < 0$. As for the case of negative sign, $x^2 - y^2 = k$, which reads $(x-y)(x+y) = k$, here once again by linearity our equation becomes $xy = l$, which is a hyperbola when $l \neq 0$, and two lines when $l = 0$.

(3) In the case $b \neq 0$ the study is similar, with the same solutions, so we are left with the case $a = b = 0$. Here our conic is as follows, with $c, d, e, f \in \mathbb{R}$:

$$cxy + dx + ey + f = 0$$

If $c \neq 0$, by linearity our equation becomes $xy = l$, which produces a hyperbola or two lines, as explained before. As for the remaining case, $c = 0$, here our equation is:

$$dx + ey + f = 0$$

But this is generically the equation of a line, unless we are in the case $d = e = 0$, where our equation is $f = 0$, having as solutions $\emptyset$ when $f \neq 0$, and $\mathbb{R}^2$ when $f = 0$.

(4) So, this was the study of an arbitrary conic, and by putting now everything together, we are led to the conclusions in the statement. □

As a continuation of our classification work, and as final result here, we have:

THEOREM 7.16. *The conics fall into two classes, as follows:*

(1) *Non-degenerate: ellipses, parabolas, hyperbolas.*
(2) *Degenerate: $\emptyset$, points, lines, pairs of lines, $\mathbb{R}^2$.*

Also, the compact conics are $\emptyset$, the points, and the ellipses.

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

(1) To start with, in order to get to such a classification result, we just need to apply linear transformations to the curves that we found in Theorem 7.15. But this leaves the list there unchanged, up to the circles becoming ellipses, as stated above.

(2) In what regards the last assertion, this is clear from the first one, but since this assertion is quite interesting, let us give it a quick, independent proof as well. Consider an arbitary conic, written as follows, with $a, b, c, d, e, f \in \mathbb{R}$:

$$ax^2 + by^2 + cxy + dx + ey + f = 0$$

Compacity rules then out the case $c \neq 0$, and our conic must be in fact:

$$ax^2 + by^2 + dx + ey + f = 0$$

But then with $a, b \neq 0$ we must have by compacity $a, b > 0$ or $a, b < 0$, and we get an ellipsis, then with $a = 0, b \neq 0$ or $a \neq 0, b = 0$ we get by compacity either $\emptyset$ or a point, and finally with $a = b = 0$ the compacity rules out again everything, except for $\emptyset$. □

As a final general result about the conics, completing our study, we have:

THEOREM 7.17. *Up to some degenerate cases, the conics are exactly the curves which appear by cutting a 2-sided cone with a plane.*

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

(1) We can change the coordinates, as for our plane to be given by $z = 0$, then see how the cone equation $x^2 + y^2 = kz^2$ changes, under this change of coordinates, and then set $z = 0$, as to get the (x, y) equation of the intersection. But this leads, via some thinking or computations, to the conclusion that the cone equation $x^2 + y^2 = kz^2$ becomes in this way a degree 2 equation in (x, y), which can be arbitrary, as desired.

(2) Alternatively, and a bit more concretely, we can use the original coordinates, with the cone being $x^2+y^2 = kz^2$, and compute the intersection, with the conclusion that what we get, depending on the slope of the cone, and modulo degenerate cases, is an ellipse, hyperbola or parabola. So, by invoking Theorem 7.16, we get the result. □

And with this, end of our learning on plane curves. More soon, when discussing Kepler and Newton, and then later too, when doing more complicated math and physics.

7c. Derivatives, Taylor

Getting now to functions and analysis, in several variables, as a first result here, which is something fundamental, getting us into matrices and linear algebra, we have:

THEOREM 7.18. *A function $f : \mathbb{R}^N \to \mathbb{R}^M$ is continuously differentiable,*

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

with $f'(x)$ linear, and $x \to f'(x)$ continuous, precisely when it has partial derivatives,

$$\frac{df_i}{dx_j}(x) = \lim_{t\to 0} \frac{f_i(x+te_j) - f_i(x)}{t}$$

which depend continuously on x. In this case the derivative is

$$f'(x) = \left(\frac{df_i}{dx_j}(x)\right)_{ij} \in M_{M\times N}(\mathbb{R})$$

acting on the vectors $t \in \mathbb{R}^N$ by usual multiplication.

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

(1) As a first observation, the formula in the statement makes sense indeed, as an equality, or rather approximation, of vectors in $\mathbb{R}^M$, as follows:

$$f\begin{pmatrix} x_1+t_1 \\ \vdots \\ x_N+t_N \end{pmatrix} \simeq f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \frac{df_1}{dx_1}(x) & \ldots & \frac{df_1}{dx_N}(x) \\ \vdots & & \vdots \\ \frac{df_M}{dx_1}(x) & \ldots & \frac{df_M}{dx_N}(x) \end{pmatrix} \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix}$$

(2) Getting now to the proof, in the simplest possible case, $N = M = 1$, what we have is a usual 1-variable function $f : \mathbb{R} \to \mathbb{R}$, and the formula in the statement is something that we know well from 1-variable calculus, namely the definition of the derivative:

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

(3) Next, at $N = 2, M = 1$ the result holds as well, by using (2) twice, as follows:

$$\begin{aligned} f\begin{pmatrix} x_1+t_1 \\ x_2+t_2 \end{pmatrix} &\simeq f\begin{pmatrix} x_1+t_1 \\ x_2 \end{pmatrix} + \frac{df}{dx_2}(x)t_2 \\ &\simeq f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{df}{dx_1}(x)t_1 + \frac{df}{dx_2}(x)t_2 \\ &= f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} \frac{df}{dx_1}(x) & \frac{df}{dx_2}(x) \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \end{aligned}$$

(4) More generally, we can deal in this way with the general case $M = 1$, with the formula here, obtained via a straightforward recurrence, being as follows:

$$\begin{aligned} f\begin{pmatrix} x_1 + t_1 \\ \vdots \\ x_N + t_N \end{pmatrix} &\simeq f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \frac{df}{dx_1}(x)t_1 + \ldots + \frac{df}{dx_N}(x)t_N \\ &= f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \frac{df}{dx_1}(x) & \ldots & \frac{df}{dx_N}(x) \end{pmatrix}\begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \end{aligned}$$

(5) But this gives the result in the case where both $N, M \in \mathbb{N}$ are arbitrary too. Indeed, consider a function $f : \mathbb{R}^N \to \mathbb{R}^M$, and let us write it as follows:

$$f = \begin{pmatrix} f_1 \\ \vdots \\ f_M \end{pmatrix}$$

We can apply (4) to each of the components $f_i : \mathbb{R}^N \to \mathbb{R}$, and we get:

$$f_i\begin{pmatrix} x_1 + t_1 \\ \vdots \\ x_N + t_N \end{pmatrix} \simeq f_i\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \frac{df_i}{dx_1}(x) & \ldots & \frac{df_i}{dx_N}(x) \end{pmatrix}\begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix}$$

(6) But this collection of M formulae, taken altogether, tells us precisely that the formula in (1) happens indeed. Thus, we are led to the conclusion in the statement. □

Generally speaking, Theorem 7.18 is all you need to know, for extending to several variables the basic results from one-variable calculus. As a basic illustration, we have:

THEOREM 7.19. *We have the chain derivative formula*

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

as an equality of matrices.

PROOF. This is something standard in one variable, that we know from chapter 3, and in several variables the proof is similar, by using the notion of derivative coming from Theorem 7.18. To be more precise, consider a composition of functions, as follows:

$$f : \mathbb{R}^N \to \mathbb{R}^M \quad , \quad g : \mathbb{R}^K \to \mathbb{R}^N \quad , \quad f \circ g : \mathbb{R}^K \to \mathbb{R}^M$$

According to Theorem 7.18, the derivatives of these functions are certain linear maps, corresponding to certain rectangular matrices, as follows:

$$f'(g(x)) \in M_{M\times N}(\mathbb{R}) \quad , \quad g'(x) \in M_{N\times K}(\mathbb{R}) \qquad (f \circ g)'(x) \in M_{M\times K}(\mathbb{R})$$

Thus, our formula makes sense indeed. As for proof, this comes from:

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

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

Quite nice all this, and as a first application, we can finish some physics business started in chapter 4, namely uniqueness in the d'Alembert theorem. We have:

THEOREM 7.20. *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:*

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

Also, 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. We already talked about waves in this book, in chapters 4 and 6, but the above formula is still to be proved. Let us make the following change of variables:

$$y=x-vt\quad,\quad z=x+vt$$

We have the following computation, using the chain rule from Theorem 7.19:

$$\frac{d\varphi}{dt}=\frac{d\varphi}{dy}\cdot\frac{dy}{dt}+\frac{d\varphi}{dz}\cdot\frac{dz}{dt}=-v\,\frac{d\varphi}{dy}+v\,\frac{d\varphi}{dz}$$

By using the chain rule again, the second time derivative is given by:

$$\begin{aligned}\frac{d^2\varphi}{dt^2} &= -v\left(\frac{d^2\varphi}{dy^2}\cdot\frac{dy}{dt}+\frac{d^2\varphi}{dydz}\cdot\frac{dz}{dt}\right)+v\left(\frac{d^2\varphi}{dzdy}\cdot\frac{dy}{dt}+\frac{d^2\varphi}{dz^2}\cdot\frac{dz}{dt}\right)\\ &= -v\left(-v\,\frac{d^2\varphi}{dy^2}+v\,\frac{d^2\varphi}{dydz}\right)+v\left(-v\,\frac{d^2\varphi}{dzdy}+v\,\frac{d^2\varphi}{dz^2}\right)\\ &= v^2\left(\frac{d^2\varphi}{dy^2}+\frac{d^2\varphi}{dz^2}-2\frac{d^2\varphi}{dydz}\right)\end{aligned}$$

Regarding now the first space derivative, this is given by the following formula:

$$\frac{d\varphi}{dx}=\frac{d\varphi}{dy}\cdot\frac{dy}{dx}+\frac{d\varphi}{dz}\cdot\frac{dz}{dx}=\frac{d\varphi}{dy}+\frac{d\varphi}{dz}$$

By using the chain rule again, the second space derivative is given by:

$$\begin{aligned}\frac{d^2\varphi}{dt^2} &= \left(\frac{d^2\varphi}{dy^2}\cdot\frac{dy}{dx}+\frac{d^2\varphi}{dydz}\cdot\frac{dz}{dx}\right)+\left(\frac{d^2\varphi}{dzd\xi}\cdot\frac{dy}{dx}+\frac{d^2\varphi}{dz^2}\cdot\frac{dz}{dx}\right)\\ &= \left(\frac{d^2\varphi}{dy^2}+\frac{d^2\varphi}{dydz}\right)+\left(\frac{d^2\varphi}{dzdy}+\frac{d^2\varphi}{dz^2}\right)\\ &= \frac{d^2\varphi}{dy^2}+\frac{d^2\varphi}{dz^2}+2\frac{d^2\varphi}{dydz}\end{aligned}$$

Thus, our wave equation $\ddot{\varphi}=v^2\varphi''$ reformulates in a very simple way, as follows:

$$\frac{d^2\varphi}{dydz}=0$$

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

$$\varphi(x,t)=F(y)+G(z)=F(x-vt)+G(x+vt)$$

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. □

Back now to mathematics, time for some work at order 2. We have here:

THEOREM 7.21. *Given a twice differentiable function $f:\mathbb{R}^N\to\mathbb{R}$, we have*

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

with $f'(x)\in M_{1\times N}(\mathbb{R})$ being a row vector, and with $f''(x)\in M_N(\mathbb{R})$, given by

$$f''(x)=\left(\frac{d^2f}{dx_idx_j}\right)_{ij}(x)$$

being the Hessian matrix of f, at the point $x\in\mathbb{R}^N$.

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

(1) As a first remark, at $N=1$ the Hessian matrix is the 1×1 matrix having as entry the usual second derivative $f''(x)\in\mathbb{R}$, and the formula in the statement is something that we know well from 1-variable calculus, namely the Taylor formula at order 2:

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

(2) In general now, this is in fact something which does not need a whole new proof, because it follows from the one-variable formula above, applied to the restriction of f to the following segment in $\mathbb{R}^N$, which can be regarded as being a one-variable interval:

$$I=[x,x+t]$$

To be more precise, let $y \in \mathbb{R}^N$, and consider the following function, with $r \in \mathbb{R}$:

$$g(r) = f(x + ry)$$

We know from (1) that the Taylor formula for g, at the point $r = 0$, reads:

$$g(r) \simeq g(0) + g'(0)r + \frac{g''(0)r^2}{2}$$

And our claim is that, with $t = ry$, this is precisely the formula in the statement.

(3) So, let us see if our claim is correct. By using the chain rule, we have the following formula, with on the right, as usual, a row vector multiplied by a column vector:

$$g'(r) = f'(x + ry) \cdot y$$

By using again the chain rule, we can compute the second derivative as well:

$$\begin{aligned} g''(r) &= (f'(x + ry) \cdot y)' \\ &= \left(\sum_i \frac{df}{dx_i}(x + ry) \cdot y_i \right)' \\ &= \sum_i \sum_j \frac{d^2 f}{dx_i dx_j}(x + ry) \cdot \frac{d(x + ry)_j}{dr} \cdot y_i \\ &= \sum_i \sum_j \frac{d^2 f}{dx_i dx_j}(x + ry) \cdot y_i y_j \\ &= < f''(x + ry)y, y > \end{aligned}$$

(4) Time now to conclude. We know that we have $g(r) = f(x + ry)$, and according to our various computations above, we have the following formulae:

$$g(0) = f(x) \quad , \quad g'(0) = f'(x) \quad , \quad g''(0) =< f''(x)y, y >$$

Buit with this data in hand, the usual Taylor formula for our one variable function g, at order 2, at the point $r = 0$, takes the following form, with $t = ry$:

$$\begin{aligned} f(x + ry) &\simeq f(x) + f'(x)ry + \frac{< f''(x)y, y > r^2}{2} \\ &= f(x) + f'(x)t + \frac{< f''(x)t, t >}{2} \end{aligned}$$

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

Many other things can be said, as a continuation of the above, notably with a positivity discussion, regarding the minima and maxima of f. We will be back to this.

7d. Harmonic functions

With the above discussed, end of mathematics? You must be kidding. Indeed, at a more advanced level now, based on what we found in chapter 6, let us formulate:

DEFINITION 7.22. *The Laplace operator in 2 dimensions is:*

$$\Delta f = \frac{d^2 f}{dx^2} + \frac{d^2 f}{dy^2}$$

A function $f : \mathbb{R}^2 \to \mathbb{C}$ satisfying $\Delta f = 0$ will be called harmonic.

To be more precise here, we certainly know about the Laplace operator Δ from chapter 6, and in view of the material there, and of linear algebra too, telling us to look for the eigenvalues and eingenvectors of our linear maps, we are led to the Laplace equation above, $\Delta f = 0$, which mathematically is the equation of the 0-eigenvectors of Δ.

As a first good surprise, we have an interesting link with the complex functions. Indeed, with the standard convention that a function $f : \mathbb{C} \to \mathbb{C}$ is called holomorphic when it develops as a complex Taylor series around each point, which is the case for instance for the polynomials, the rational functions, or $\sin, \cos, \exp, \log$, we have:

THEOREM 7.23. *Any holomorphic function $f : \mathbb{C} \to \mathbb{C}$, when regarded as function*

$$f : \mathbb{R}^2 \to \mathbb{C}$$

is harmonic. Moreover, the conjugates $\bar{f}$ of holomorphic functions are harmonic too.

PROOF. The first assertion follows from the following computation, for the power functions $f(z) = z^n$, with the usual notation $z = x + iy$:

$$\begin{aligned}
\Delta z^n &= \frac{d^2 z^n}{dx^2} + \frac{d^2 z^n}{dy^2} \\
&= \frac{d(nz^{n-1})}{dx} + \frac{d(inz^{n-1})}{dy} \\
&= n(n-1)z^{n-2} - n(n-1)z^{n-2} \\
&= 0
\end{aligned}$$

As for the second assertion, this follows from $\Delta \bar{f} = \overline{\Delta f}$, which is clear from definitions, and which shows that if f is harmonic, then so is its conjugate $\bar{f}$. □

As a next goal, in order to understand the harmonic functions, we can try to find the homogeneous polynomials $P \in \mathbb{R}[x, y]$ which are harmonic. In order to do so, the most convenient is to use the variable $z = x + iy$, and think of these polynomials as being homogeneous polynomials $P \in \mathbb{R}[z, \bar{z}]$. With this convention, the result is as follows:

THEOREM 7.24. *The degree n homogeneous polynomials $P \in \mathbb{R}[x,y]$ which are harmonic are precisely the linear combinations of*

$$P = z^n \quad , \quad P = \bar{z}^n$$

with the usual identification $z = x + iy$.

PROOF. As explained above, any homogeneous polynomial $P \in \mathbb{R}[x,y]$ can be regarded as an homogeneous polynomial $P \in \mathbb{R}[z,\bar{z}]$, with the change of variables $z = x+iy$, and in this picture, the degree n homogeneous polynomials are as follows:

$$P(z) = \sum_{k+l=n} c_{kl} z^k \bar{z}^l$$

In oder to solve now the Laplace equation $\Delta P = 0$, we must compute the quantities $\Delta(z^k \bar{z}^l)$, for any k, l. And here, a routine computation gives the following formula:

$$f = z^k \bar{z}^l \implies \Delta f = \frac{4klf}{|z|^2}$$

Now let us get back to our homogeneous polynomial P, written as follows:

$$P(z) = \sum_{k+l=n} c_{kl} z^k \bar{z}^l$$

By using the above formula, it follows that the Laplacian of P is given by:

$$\Delta P(z) = \frac{4}{|z|^2} \sum_{k+l=n} klc_{kl} z^k \bar{z}^l$$

We conclude that the Laplace equation for P takes the following form:

$$\begin{aligned} \Delta P = 0 &\iff klc_{kl} = 0, \forall k, l \\ &\iff [k, l \neq 0 \implies c_{kl} = 0] \\ &\iff P = c_{n0} z^n + c_{0n} \bar{z}^n \end{aligned}$$

Thus, we are led to the conclusion in the statement. And with the observation that the real formulation of the final result is something quite complicated, and so, for one more time, the use of the complex variable $z = x + iy$ is something very useful. □

As a next objective, let us try now to find the harmonic functions which are radial, $f(z) = \varphi(|z|)$. However, things are quite tricky here, involving a blowup phenomenon at the dimension value $N = 2$, which is precisely the one that we are interested in.

In view of this phenomenon, it makes sense to move to arbitrary $N \in \mathbb{N}$ dimensions. As before in 2 dimensions, we can say that a function $f : \mathbb{R}^N \to \mathbb{C}$ is harmonic when $\Delta f = 0$. With this convention, the result about radial harmonics is as follows:

THEOREM 7.25. *The fundamental radial solutions of* $\Delta f = 0$ *are*

$$f(x) = \begin{cases} ||x||^{2-N} & (N \neq 2) \\ \log ||x|| & (N = 2) \end{cases}$$

with the log *at* $N = 2$ *basically coming from* $\log' = 1/x$.

PROOF. Consider indeed a radial function, defined outside the origin $x = 0$. This function can be written as follows, with $\varphi : (0, \infty) \to \mathbb{C}$ being a certain function:

$$f : \mathbb{R}^N - \{0\} \to \mathbb{C} \quad , \quad f(x) = \varphi(||x||)$$

Our first goal will be that of reformulating the Laplace equation $\Delta f = 0$ in terms of the one-variable function $\varphi : (0, \infty) \to \mathbb{C}$. For this purpose, observe that we have:

$$\frac{d||x||}{dx_i} = \frac{d\sqrt{\sum_{i=1}^N x_i^2}}{dx_i} = \frac{x_i}{||x||}$$

By using this formula, we have the following computation:

$$\frac{df}{dx_i} = \frac{d\varphi(||x||)}{dx_i} = \varphi'(||x||) \cdot \frac{x_i}{||x||}$$

By differentiating one more time, we obtain the following formula:

$$\begin{aligned} \frac{d^2 f}{dx_i^2} &= \left(\varphi''(||x||) \cdot \frac{x_i}{||x||}\right) \cdot \frac{x_i}{||x||} + \varphi'(||x||) \cdot \frac{||x|| - x_i \cdot x_i/||x||}{||x||^2} \\ &= \varphi''(||x||) \cdot \frac{x_i^2}{||x||^2} + \varphi'(||x||) \cdot \frac{||x||^2 - x_i^2}{||x||^3} \end{aligned}$$

Now by summing over $i \in \{1, \ldots, N\}$, this gives the following formula:

$$\begin{aligned} \Delta f &= \sum_{i=1}^N \varphi''(||x||) \cdot \frac{x_i^2}{||x||^2} + \sum_{i=1}^N \varphi'(||x||) \cdot \frac{||x||^2 - x_i^2}{||x||^3} \\ &= \varphi''(||x||) \cdot \frac{||x||^2}{||x||^2} + \varphi'(||x||) \cdot \frac{(N-1)||x||^2}{||x||^3} \\ &= \varphi''(||x||) + \varphi'(||x||) \cdot \frac{N-1}{||x||} \end{aligned}$$

Thus, with $r = ||x||$, the Laplace equation $\Delta f = 0$ can be reformulated as follows:

$$\varphi''(r) + \frac{(N-1)\varphi'(r)}{r} = 0$$

Equivalently, the equation that we want to solve is as follows:

$$r\varphi'' + (N-1)\varphi' = 0$$

Now observe that we have the following differentiation formula:

$$\begin{aligned}(r^{N-1}\varphi')' &= (N-1)r^{N-2}\varphi' + r^{N-1}\varphi'' \\ &= r^{N-2}((N-1)\varphi' + r\varphi'')\end{aligned}$$

Thus, the equation to be solved can be simply written as follows:

$$(r^{N-1}\varphi')' = 0$$

We conclude that $r^{N-1}\varphi'$ must be a constant K, and so, that we must have:

$$\varphi' = Kr^{1-N}$$

But the fundamental solutions of this latter equation are as follows:

$$\varphi(r) = \begin{cases} r^{2-N} & (N \neq 2) \\ \log r & (N = 2) \end{cases}$$

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

Quite interesting all this, and as a conclusion, something seems to be wrong with the dimension $N = 2$, that we are currently investigating, in this book. Nevermind. One more chapter about this, $N = 2$, and then we will get to $N = 3$, and higher.

7e. Exercises

Good mathematics chapter that we had here, and as exercises, we have:

EXERCISE 7.26. *Solve the questions that we left, regarding the scalar products.*

EXERCISE 7.27. *Learn a bit about orthonormal bases, and their properties.*

EXERCISE 7.28. *Compute the matrices of some 3D linear maps, of your choice.*

EXERCISE 7.29. *Try inverting, and computing the determinant, of the 3D matrices.*

EXERCISE 7.30. *Clarify the details, in the cutting of a 2-sided cone with a plane.*

EXERCISE 7.31. *Perform the final integrations, in the proof of d'Alembert.*

EXERCISE 7.32. *Learn more about Taylor at order 2, and its consequences.*

EXERCISE 7.33. *Try to work out the Taylor formula, at order 3, and higher.*

As bonus exercise, learn more on the Laplace equation, and harmonic functions.

CHAPTER 8

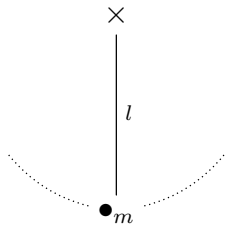
Standard mechanics

8a. The pendulum

According to the physics battle plan made in chapter 6, and approved by our regular 2D advisor, the spider, we are left in this chapter with discussing the fundamentals of 2D gravity. And good news here, with the mathematics that we learned in chapter 7, all quality tools, things will be quite easy, and the various problems that will appear on our way, we will eat them raw, and they will taste delicious, like big, meaty flies.

Getting to work now, we would first like to discuss a question which is rather something 1.5 dimensional, namely the movement of the pendulum. So, let us start with:

DEFINITION 8.1. *A simple pendulum is a device of type*



consisting of a bob of mass m, attached to a rigid rod of length l.

As a first observation, quite nice definition that we have here, and which is indeed something rather 1.5 dimensional, because the problem obviously happens in 2D, but from the perspective of the bob, which is what matters, and which will be ours in what follows, things seem to happen in 1D. I mean, imagine being that poor bob, you are there in your 1D that you are used to, but some strange force, coming from a 2D that you don't really know about, moves you to the left and to the right in your 1D. Annoying.

But are we here for talking philosophy. At the scientific level, we have:

ADVERTISEMENT 8.2. *The mechanics of the pendulum is the source of many good things, touching nearly all fundamental aspects of physics and engineering:*

(1) *Harmonic oscillators.*
(2) *Resonance and damping.*
(3) *Clocks and other machines.*

In order to study now the mechanics 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. So, let us recall from chapter 4 that we have:

PROPOSITION 8.3. *For a 1D particle of mass m, subject to a force F, if we define its kinetic and potential energy as follows, the prime being a space derivative,*

$$T = \frac{mv^2}{2} \quad , \quad V' = -F$$

and so with V being unique up to a constant, $E = T + V$ is constant over time.

PROOF. This is indeed something that we know from chapter 4, coming from:

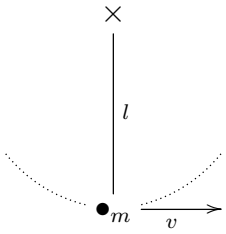
$$\begin{aligned} T = \frac{mv^2}{2} &\implies \dot{T} = mva = Fv \\ &\implies T = \int Fvdt = \int F\dot{x}dt = \int Fdx \\ &\implies T' = F \end{aligned}$$

Thus we have $E' = F - F = 0$, and so E is constant over time, as claimed. □

As mentioned above, what we have there holds in 1D. In higher dimensions things are more complicated, and the idea is that the equation $V' = -F$ must be replaced by $\nabla V = -F$, with $\nabla V(x) = V'(x)^t$ being the transpose of the first derivative, and with the energy conservation conclusion holding too, under suitable assumptions. More on this in chapter 10, and in the meantime, what we have in Proposition 8.3 will do.

Getting back now to the pendulum from Definition 8.1, by using this, we can have its mechanics understood with not too many computations involved, as follows:

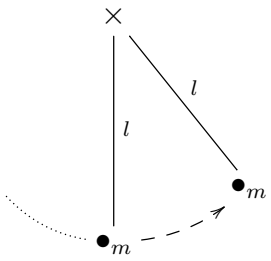
THEOREM 8.4. *For a pendulum starting with speed v from the equilibrium position,*



the motion will be confined if $v^2 < 4gl$, and circular if $v^2 > 4gl$.

PROOF. There are many ways of proving this result, along with working out several other useful related formulae, for which we will refer to the proof below, and with a quite elegant approach to this, using no computations or almost, being as follows:

(1) Let us first examine what happens when the bob has traveled an angular distance $\theta > 0$, with respect to the vertical. The picture here is as follows:



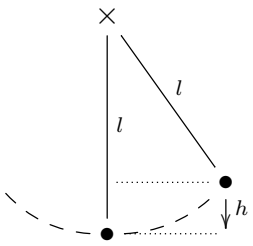
The distance traveled is then $x = l\theta$. As for the force acting, this is $F_{total} = mg$ oriented downwards, with the component alongside x being given by:

$$\begin{aligned} F &= -||F_{total}|| \sin\theta \\ &= -mg\sin\theta \\ &= -mg\sin\left(\frac{x}{l}\right) \end{aligned}$$

(2) But with this, we can compute the potential energy. With the convention that this vanishes at the equilibrium position, $V(0) = 0$, we obtain the following formula:

$$\begin{aligned} V' = -F \quad &\implies \quad V' = mg\sin\left(\frac{x}{l}\right) \\ &\implies \quad V = mgl\left(1 - \cos\left(\frac{x}{l}\right)\right) \\ &\implies \quad V = mgl(1 - \cos\theta) \end{aligned}$$

(3) Alternatively, in case this sounds too wizarding, we can compute the potential energy in a more standard way, by letting the bob fall, the picture being as follows:



The height of the fall is then $h = l - l\cos\theta$, and since for this fall the force is constant, $\mathcal{F} = -mg$, we obtain the following formula for the potential energy:

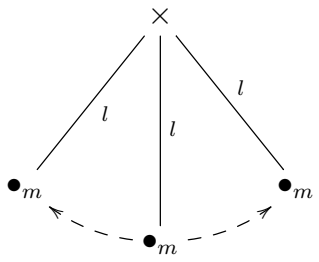
$$\begin{aligned} V' = -\mathcal{F} \quad &\implies \quad V' = mg \\ &\implies \quad V = mgh \\ &\implies \quad V = mgl(1 - \cos\theta) \end{aligned}$$

Summarizing, one way or another we have our formula for the potential energy V.

(4) Now comes the discussion. The motion will be confined when the initial kinetic energy, namely $E = mv^2/2$, satisfies the following condition:

$$\begin{aligned} E < \sup_\theta V = 2mgl &\iff \frac{mv^2}{2} < 2mgl \\ &\iff v^2 < 4gl \end{aligned}$$

In this case, the motion will be confined between two angles $-\theta, \theta$, as follows:



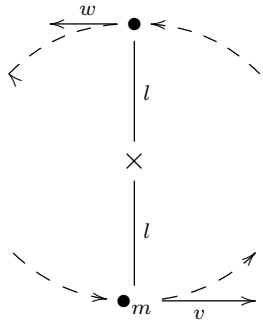
To be more precise here, the two extreme angles $-\theta, \theta \in (-\pi, \pi)$ can be explicitly computed, as being solutions of the following equation:

$$\begin{aligned} V = E &\iff mgl(1 - \cos\theta) = \frac{mv^2}{2} \\ &\iff 1 - \cos\theta = \frac{v^2}{2gl} \end{aligned}$$

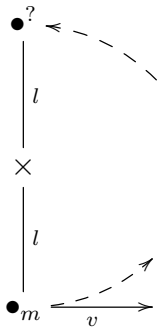
(5) Regarding now the case $v^2 > 4gl$, here the bob will certainly reach the upwards position, with the speed $w > 0$ there being given by the following formula:

$$\begin{aligned} \frac{mw^2}{2} = E - 2mgl &\implies \frac{mw^2}{2} = \frac{mv^2}{2} - 2mgl \\ &\implies w^2 = v^2 - 4gl \\ &\implies w = \sqrt{v^2 - 4gl} \end{aligned}$$

Thus, with the convention in the statement for v, that is, going to the right, the motion of the pendulum will be counterclockwise circular, and perpetual:



(6) Finally, in the case $v^2 = 4gl$, the bob will also reach the upwards position, but with speed $w = 0$ there, and then, at least theoretically, will remain there:



(7) Actually, it is quite interesting in this latter situation, $v^2 = 4gl$, to further speculate on what can happen, when making our problem more realistic. For instance, we can add to our setting the assumption that when the bob is stuck on top, with speed 0, there is a 33% chance for it to keep going, to the left, a 33% chance for it to come back, to the right, and a 33% chance for it to remain stuck. In this case there are infinitely many possible trajectories, which are best investigated by using probability. Welcome to chaos.

(8) As a final comment, yes I know that the figures in (7) don't add up to 100%. This is because there is as well a remaining 1% possibility, where a relativistic black cat appears, with a continuous effect on the bob, via a paw slap, when on top, with speed $w' \in (0.3c, 0.7c)$, with c being the speed of light. In this case, the set of possible trajectories becomes uncountable, and is again best investigated by using probability. □

8b. Harmonic oscillators

As a continuation of the above study of the pendulum, let us discuss now an important question in mechanics, namely the motion of a 1D 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 8.5. *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 8.6. *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 8.5, or from the explicit formulae below.

Very nice, so let us solve now the equation of motion. We have here:

THEOREM 8.7. *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 8.8. *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 8.7, as follows:

(1) In the case $k < 0$, with $x = ae^{pt} + be^{-pt}$ as in Theorem 8.7, 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 8.7, 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 8.9. *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 8.7 (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 8.7 (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 8.7 (2).

(2) Things are similar here. Indeed, as before in the proof of Theorem 8.7 (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 8.7 (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 8.7 (1). □

Let us study now the damped oscillators, obtained by adding to the picture friction, of some other extra force. We first have here the following result:

THEOREM 8.10. *For a damped oscillator, which is subject by definition to a force of type $F = -kx - \lambda\dot{x}$, the equation of motion is*

$$m\ddot{x} + \lambda\dot{x} + kx = 0$$

with m being as before the mass.

PROOF. This is clear indeed from $F = ma = m\ddot{x}$, which gives:

$$\begin{aligned}
F = -kx - \lambda\dot{x} &\iff m\ddot{x} = -kx - \lambda\dot{x}\\
&\iff m\ddot{x} + \lambda\dot{x} + kx = 0
\end{aligned}$$

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

Now let us try to solve the equation of motion. When looking for solutions of type $x = e^{pt}$, with $p \in \mathbb{C}$ constant, the equation of motion takes the following form:

$$mp^2 + \lambda p + k = 0$$

But this is a degree 2 equation, that we can solve right away, and we get:

$$p = \frac{-\lambda \pm \sqrt{\lambda^2 - 4mk}}{2m}$$

It is convenient to write these solutions that we found, and the overall final result about damping, in the following more convenient way:

THEOREM 8.11. *The generic solutions of $m\ddot{x} + \lambda\dot{x} + kx = 0$ are the real linear combinations of the functions e^{pt}, with the parameter $p \in \mathbb{C}$ being given by*

$$p = -\gamma \pm \sqrt{\gamma^2 - w^2} \quad , \quad \gamma = \frac{\lambda}{2m} \quad , \quad w = \sqrt{\frac{k}{m}}$$

with on the right w being the frequency of the usual, undamped oscillator.

PROOF. This follows indeed from the formula of p found above, by dividing everything by $2m$. Observe that w is indeed the frequency of the usual, undamped oscillator. □

Now assume that we are in the case $\lambda > 0$, which is the most usual one, in practice, meaning that our oscillator loses energy. We have then three cases, as follows:

PROPOSITION 8.12. *The oscillator damping with $\lambda > 0$, with this assumption meaning that our oscillator loses energy over the time, can be of three types:*

(1) *Large damping, with $\lambda > 0$ being such that $\gamma > w$. Here the roots found above $p = -\gamma \pm \sqrt{\gamma^2 - w^2}$ are both real, negative, and distinct.*
(2) *Small damping, with $\lambda > 0$ being such that $\gamma < w$. In this case the roots that we found $p = -\gamma \pm \sqrt{\gamma^2 - w^2}$ are complex and conjugate.*
(3) *Critical damping, with $\lambda > 0$ being such that $\gamma = w$. Here we have a double root, which is real and negative, namely $p = -\gamma$.*

PROOF. All this is clear indeed from the formula found in Theorem 8.11. □

Now let us study more in detail the above three types of damping. In what regards the large damping, things here are very simple and intuitive, as follows:

PROPOSITION 8.13. *For large oscillator damping, the trajectory is given by*

$$x = ae^{-\gamma_+ t} + be^{-\gamma_- t}$$

with the parameters $\gamma_+ > \gamma_- > 0$ being given by the formulae

$$\gamma_\pm = \gamma \pm \sqrt{\gamma^2 - w^2}$$

and with $a, b \in \mathbb{R}$. With $t >> 0$ we have $x \simeq be^{-\gamma_- t} \to 0$.

PROOF. All this is indeed self-explanatory, and clear from the formula that we found in Theorem 8.11, under the large damping assumption from Proposition 8.12 (1). □

In what regards the small damping, things here are again simple and intuitive, involving this time complex numbers and trigonometric functions, as follows:

PROPOSITION 8.14. *For small oscillator damping, the trajectory is given by*

$$x = 2re^{-\gamma t}\cos(\rho t - \theta)$$

with $\gamma = \lambda/2m$ as before, with the parameter $\rho > 0$ being given by the formula

$$\rho = \sqrt{w^2 - \gamma^2}$$

and with $r > 0$ and $\theta \in \mathbb{R}$. With $t >> 0$ we have $x \to 0$, exponentially.

PROOF. Assume indeed that we are in the small damping regime, where $\lambda > 0$ is such that $\gamma < w$. The roots that we found are then complex conjugate, as follows:

$$p = -\gamma \pm i\rho \quad , \quad \rho = \sqrt{w^2 - \gamma^2}$$

As for the solution itself, this is given by the following formula, with $c, d \in \mathbb{C}$:

$$\begin{aligned} x &= ce^{p_+ t} + de^{p_- t} \\ &= ce^{-\gamma t + i\rho t} + de^{-\gamma t - i\rho t} \\ &= e^{-\gamma t}(ce^{i\rho t} + de^{-i\rho t}) \end{aligned}$$

Now in order to have $x \in \mathbb{R}$, which is the same as saying $x = \bar{x}$, we must have $c = \bar{d}$. Thus we can write $c = re^{-i\theta}$ and $d = re^{i\theta}$ with $r > 0$ and $\theta \in \mathbb{R}$, and we obtain:

$$\begin{aligned} x &= e^{-\gamma t}(ce^{i\rho t} + \bar{c}e^{-i\rho t}) \\ &= 2e^{-\gamma t}Re(ce^{i\rho t}) \\ &= 2re^{-\gamma t}Re(e^{i(\rho t - \theta)}) \\ &= 2re^{-\gamma t}\cos(\rho t - \theta) \end{aligned}$$

Finally, the fact that we have indeed $x \to 0$, exponentially, is clear. □

As for the critical damping case, the result here is as follows:

THEOREM 8.15. *For critical oscillator damping, the trajectory is given by*

$$x = (a + bt)e^{-\gamma t}$$

with the parameter $\gamma > 0$ being given as usual by the formula

$$\gamma = \frac{\lambda}{2m}$$

and with $a, b \in \mathbb{R}$. With $t >> 0$ we have $x \simeq bte^{-\gamma t} \to 0$.

PROOF. Assume indeed that we are in the critical damping regime, where $\lambda > 0$ is such that $\gamma = w$. In this case the roots that we found in Theorem 8.11 are both given by $p = -\gamma$, and so that result provides us with solutions as follows, with $a \in \mathbb{R}$:

$$x = ae^{-\gamma t}$$

Thus, we must find in this case some further solutions of the equation $m\ddot{x}+\lambda\dot{x}+kx=0$, and our claim is that the following functions are solutions too:

$$x=bte^{-\gamma t}$$

Indeed, let us verify this, under the above critical damping assumption. By linearity it is enough to do this for $b=1$, and here the derivatives are computed as follows:

$$\begin{aligned}
& x=te^{-\gamma t}\\
\implies\quad & \dot{x}=e^{-\gamma t}-\gamma te^{-\gamma t}=(1-\gamma t)e^{-\gamma t}\\
\implies\quad & \ddot{x}=-\gamma e^{-\gamma t}-\gamma(1-\gamma t)e^{-\gamma t}=-\gamma(2-\gamma t)e^{-\gamma t}
\end{aligned}$$

We can verify now that the equation is indeed satisfied, as follows:

$$\begin{aligned}
& m\ddot{x}+\lambda\dot{x}+kx\\
= \quad & -m\gamma(2-\gamma t)e^{-\gamma t}+\lambda(1-\gamma t)e^{-\gamma t}+kte^{-\gamma t}\\
= \quad & (-2m\gamma+m\gamma^2t+\lambda-\lambda\gamma t+kt)e^{-\gamma t}\\
= \quad & (m\gamma^2t-\lambda\gamma t+kt)e^{-\gamma t}\\
= \quad & (m\gamma^2-\lambda\gamma+k)te^{-\gamma t}\\
= \quad & 0
\end{aligned}$$

Here we have used at the end the fact, that we know from Theorem 8.11 and its proof, that the solutions of the equation $mp^2+\lambda p+k=0$ are given by $p=-\gamma\pm\sqrt{\gamma^2-w^2}$. Indeed, in the present critical damping regime these solutions are both given by $p=-\gamma$, and so substituting this particular value in the equation gives zero, as needed:

$$m\gamma^2-\lambda\gamma+k=0$$

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

Many other things can be said, as a continuation of the above, and for more on all this, we refer to a standard mechanics book, such as Kibble [58] or Taylor [86].

8c. Kepler and Newton

Getting now to the real thing, astronomy, the result here, which is the pride of mathematics, physics, and human knowledge in general, is the following theorem:

THEOREM 8.16 (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 1D 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 the gravitational 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:

$$\begin{aligned}\ddot{r} &= \frac{\lambda(\varepsilon\cos\theta + \delta\sin\theta)\dot{\theta}}{c}\\ &= \frac{\lambda^2(\varepsilon\cos\theta + \delta\sin\theta)}{r^2 c}\\ &= \frac{\lambda^2}{r^2}\left(\frac{1}{r} - \frac{1}{c}\right)\end{aligned}$$

(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. □

Still with me, I hope. All the above was great, we deduced Kepler's first law from the Netwon principles of mechanics, by using nothing more than some standard calculus. It is possible of course to be more explicit in the above, by talking about explicit conics, and to deduce the second and third laws of Kepler as well, and more on this later.

In practice, Theorem 8.16 remains something quite theoretical, more adapted for doing mathematics than applied physics or engineering, and we will perform now a more detailed study of the orbits. Let us start with a brief account of what we have, not from Theorem 8.16 itself, but rather from its proof, sort of "best of" the formulae found there:

THEOREM 8.17 (Kepler, Newton). *In the context of a 2-body problem, with M fixed at 0, and m starting its movement from Ox, the equation of motion of m, namely*

$$\ddot{z} = -\frac{Kz}{||z||^3}$$

with $K = GM$, and $z = (x, y)$, becomes in polar coordinates, $x = r\cos\theta$, $y = r\sin\theta$,

$$\ddot{r} = \frac{\lambda^2}{r^2}\left(\frac{1}{r} - \frac{1}{c}\right) \quad , \quad \dot{\theta} = \frac{\lambda}{r^2}$$

for some $\lambda, c \in \mathbb{R}$, related by $\lambda^2 = Kc$. The value of r in terms of θ is given by

$$r = \frac{c}{1 + \varepsilon\cos\theta + \delta\sin\theta}$$

for some $\varepsilon, \delta \in \mathbb{R}$. At the level of the affine coordinates x, y, this means

$$x = \frac{c\cos\theta}{1 + \varepsilon\cos\theta + \delta\sin\theta} \quad , \quad y = \frac{c\sin\theta}{1 + \varepsilon\cos\theta + \delta\sin\theta}$$

with $\theta = \theta(t)$ being subject to $\dot{\theta} = \lambda^2/r$, as above. Finally, we have

$$x^2 + y^2 = (\varepsilon x + \delta y - c)^2$$

which is a degree 2 equation, and so the resulting trajectory is a conic.

PROOF. As already mentioned, this is a sort of "best of" the formulae found in the proof of Theorem 8.16. And in the hope of course that we have not forgotten anything. Finally, let us mention that the simplest illustration for this is the circular motion, and for details on this, not included in the above, we refer to the proof of Theorem 8.16. □

As a next question, we would like to understand how the various parameters appearing above, namely $\lambda, c, \varepsilon, \delta$, which via some basic math can only tell us more about the shape of the orbit, appear from the initial data. The formulae here are as follows:

THEOREM 8.18. *In the context of Theorem 8.17, and in polar coordinates, $x = r\cos\theta$, $y = r\sin\theta$, the initial data is as follows, with $R = r_0$:*

$$r_0 = \frac{c}{1 + \varepsilon} \quad , \quad \theta_0 = 0$$

$$\dot{r}_0 = -\frac{\delta\sqrt{K}}{\sqrt{c}} \quad , \quad \dot{\theta}_0 = \frac{\sqrt{Kc}}{R^2}$$

$$\ddot{r}_0 = \frac{\varepsilon K}{R^2} \quad , \quad \ddot{\theta}_0 = \frac{4\delta K}{R^2}$$

The corresponding formulae for the affine coordinates x, y can be deduced from this. Also, the various motion parameters c, ε, δ and $\lambda = \sqrt{Kc}$ can be recovered from this data.

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

(1) As mentioned in Theorem 8.17, the object m begins its movement on Ox. Thus we have $\theta_0 = 0$, and from this we get the formula of r_0 in the statement.

(2) Regarding the initial speed now, the formula of $\dot{\theta}_0$ follows from:

$$\dot{\theta} = \frac{\lambda}{r^2} = \frac{\sqrt{Kc}}{r^2}$$

Also, in what concerns the radial speed, the formula of $\dot{r}_0$ follows from:

$$\begin{aligned}
\dot{r} &= \frac{c(\varepsilon \sin\theta - \delta\cos\theta)\dot{\theta}}{(1 + \varepsilon\cos\theta + \delta\sin\theta)^2} \\
&= \frac{c(\varepsilon \sin\theta - \delta\cos\theta)}{c^2/r^2} \cdot \frac{\sqrt{Kc}}{r^2} \\
&= \frac{\sqrt{K}(\varepsilon \sin\theta - \delta\cos\theta)}{\sqrt{c}}
\end{aligned}$$

(3) Regarding now the initial acceleration, by using $\dot{\theta} = \sqrt{Kc}/r^2$ we find:

$$\ddot{\theta} = -2\sqrt{Kc} \cdot \frac{2r\dot{r}}{r^3} = -\frac{4\sqrt{Kc} \cdot \dot{r}}{r^2}$$

In particular at $t = 0$ we obtain the formula in the statement, namely:

$$\ddot{\theta}_0 = -\frac{4\sqrt{Kc} \cdot \dot{r}_0}{R^2} = \frac{4\sqrt{Kc}}{R^2} \cdot \frac{\delta\sqrt{K}}{\sqrt{c}} = \frac{4\delta K}{R^2}$$

(4) Also regarding acceleration, with $\lambda = \sqrt{Kc}$ our main motion formula reads:

$$\ddot{r} = \frac{Kc}{r^2}\left(\frac{1}{r} - \frac{1}{c}\right)$$

In particular at $t = 0$ we obtain the formula in the statement, namely:

$$\ddot{r}_0 = \frac{Kc}{R^2}\left(\frac{1}{R} - \frac{1}{c}\right) = \frac{Kc}{R^2} \cdot \frac{\varepsilon}{c} = \frac{\varepsilon K}{R^2}$$

(5) Finally, the last assertion is clear, and since the formulae look better anyway in polar coordinates than in affine coordinates, we will not get into details here. □

With the above formulae in hand, which are a precious complement to Theorem 8.17, 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 done here, and we will leave this as a long, instructive exercise.

8d. Lagrange points

With the above discussed, and getting now to the gravitational problem for a system of $k \geq 3$ bodies, many things can be said here. Let us start with:

DEFINITION 8.19. *Associated to a system of bodies $M_1, \ldots, M_k$, located at positions $c_1, \ldots, c_k \in \mathbb{R}^3$ 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 $M_1, \ldots, M_k$.

You are certainly a bit familiar with this concept, which is something very useful, from the real life, I mean try balancing a fork on your finger, the above is what you need. Getting to the mathematics now, things can be quite complicated, as follows:

PROPOSITION 8.20. *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:

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

The systems of $k \geq 3$ bodies might have several centers of gravity, usually uncomputable.

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 quite complicated. With c_i being the position of M_i, a center of gravity $x \in \mathbb{R}^3$ must satisfy:

$$\sum_i \frac{M_i(x - c_i)}{||x - c_i||^3} = 0$$

(3) So, let us first examine 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.

(4) In the 2D case now, and with $k \geq 3$, we are looking the the center of gravity of a triangle, with vertices weighted by masses M_1, M_2, M_3. The simplest possible situation is that of an equilateral triangle, with equal masses M, M, M at its vertices, and it is quite clear here that we will have 3 solutions, 1 lying on each of the 3 symmetry axes.

(5) Which is quite bad news, because we have now 3 solutions, instead of the 2 ones found in one dimension, in (3). And for the disaster to be complete, let us attempt now to compute these solutions. The best is to use here complex numbers, with our triangle being $1, w, w^2$ in the complex plane, with $w = e^{2\pi i/3}$. We will only look for the solution on the Ox axis, say $r \in \mathbb{R}$, the other solutions being $wr, w^2 r$. The equation is:

$$\frac{r-1}{|r-1|^3} + \frac{r-w}{|r-w|^3} + \frac{r-w^2}{|r-w^2|^3} = 0$$

By simplifying at left and using $1 + w + w^2 = 0$ for the right terms, this reads:

$$\frac{2r+1}{|r-w|^3} = \frac{1}{(1-r)^2}$$

(6) And that is pretty much it, for computing r we must raise this to the square, which leads us into a degree 6 equation, and we will not do this. As a conclusion, things are hard for the equilateral triangle equally weighted, and can only be harder in general. □

Now back to our mechanics questions, we would first like to understand the motion of an object of mass m around a dumbell of parameters M_1, M_2, d, as above, by using the notion of center of mass of that dumbell, constructed according to Definition 8.19:

$$\circ_m$$

$$\bullet_{M_1} \text{---} \star_c \text{---} \bullet_{M_2}$$

And the preliminary observations here are quite grim, as follows:

OBSERVATIONS 8.21. *In the context of a body of mass m moving around a dumbell of parameters M_1, M_2, d, the following happen:*

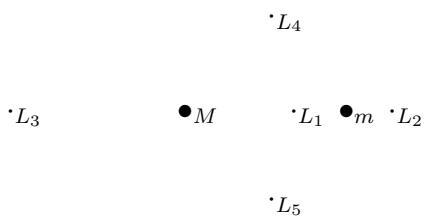
(1) *When m is at distance $x >> 0$ from the dumbell, the gravitation force acting on it is $F_1 + F_2 \simeq F_c$, so m should travel on some sort of approximate conic.*
(2) *However, when m is at distance $x \simeq 0$ from the dumbell, the physics and trajectory have nothing to do with the center of mass, $F_1 + F_2 \not\simeq F_c$.*

Here in what regards forces, (1) is something quite obvious and intuitive, and some math can be done here, of course. As for (2), this is obvious and intuitive too, because if you place m on the line joining M_1, M_2, bad things will happen, and the only possibility where c can be of help is when m was initially placed precisely on c, and with this, point on a line being at a prescribed location, being an event happening with probability 0.

As for the trajectory claim in (1), that is something sort of intuitive, but not really, because as you might know from observations with pendulums, balls rolling on various surfaces, and so on, involving equilibrium and non-equilibrium, there is no guarantee that the solution of a perturbed problem is a perturbation of the initial solution.

Nevermind. So, as a conclusion to all this, modesty, the k-body problem with $k \geq 3$ is obviously something very complicated. Talking positive results now, we have:

THEOREM 8.22. *The plane* 3*-body problem has* 5 *distinguished solutions,*



called Lagrange points L1-L5, whose positions with respect to M, m *are as above.*

PROOF. This is certainly something quite complicated, using all sorts of advanced mathematics, further building on what has been developed in the above, and we won't get into details here. Instead, let us describe at least how each L1-L5 functions:

(1) L1 is the most intuitive solution, placed in between M and m, and closer to m, at that precise point where the gravitation of M equals in magnitude the gravitation of m. The math here for finding the distances is very simple, but recall however that m is supposed to move around M, on an ellipse. Thus, the picture is that an object ε placed at L1 moves as well around M, by staying aligned with m, on a smaller ellipse.

(2) One problem with L1 comes from the fact that it is unstable, in the sense that an object ε placed around there, but not exactly there, will not stay there. Indeed, was ε to be placed a tiny little bit towards M, it will start slowly moving towards M, and gone it will be. And the same can happen in the other sense, with m being able to capture it too, since L1 is some sort of no man's land between the gravities of M, m, which look equal from there. Thus, L1 looks like some sort of fake solution to the problem.

(3) However, all this is useful in practice, because with just a little bit of homemade acceleration, from time to time, in order to correct the trajectory and keep it on L1, a satellite can be placed there at L1, and will stay there. In fact, most of the scientific

satellites are placed there at L1, first because its proximity to Earth, but also because there is no dirt like asteroids trapped there, due to the fact that L1 is unstable.

(4) Getting to L2 now, the functioning mechanism here is different, crucially relying on the fact that m, and so L2 too, does move around M, on an elliptic orbit. Indeed, what looks impossible in 1D, namely an object ε placed behind m not to be attracted by both M, m, and start going towards them, is now possible in the context of the 2D elliptical movement, with the normal movement of ε with respect to M being slightly altered by the presence of m, which in practice tends to pull ε away from M, m, and with the precise distance being that where equilibrium is achieved, in all this. As for L1, this point L2 is not far from Earth, and unstable, making it usable for satellites.

(5) In what regards now L3, yet another functioning mechanism going on here, which is this time something very simple, namely an object ε placed there at L3 will simply travel around M, in a standard elliptic way, basically on the same orbit as m, slightly adjusted as to take into account the gravity of m too. Again, this is an unstable point, suitable for satellites, but who would go up there to install one.

(6) Finally, regarding L4 and L5, these are two extra solutions, discovered later by Lagrange, located at the positions where they form, along with M, m, equilateral triangles. An object ε say placed at L4 will stay there, or rather travel on an elliptical orbit around M passing through L4, keeping its L4 relative position with respect to M, m, due to the fact that, due to the geometry of the equilateral triangle $\varepsilon - m - M$, the extra pull from m keeps it in tune with m, on that smaller orbit. And the same goes for L5.

(7) These two last points L4, L5 are stable, provided that the masses M, m of the two big objects satisfy $M/m > 24.96$, which is the case for instance for the Sun-Earth system, and for the Earth-Moon system too. However, the stability makes them unsuitable for satellite use, due to the tons of space garbage accumulated there, over the years. □

Moving forward, now that we talked about satellites, do we really have the means of sending them in space? At the basic level, the situation here is as follows:

THEOREM 8.23. *The escape velocity from a body having mass M and radius r is*

$$v = \sqrt{2gr}$$

where $g = GM/r^2$ is the surface gravitational acceleration.

PROOF. Many things can be said here, the idea being as follows:

(1) To start with, a rock launched towards the sky, at various angles and speeds, will normally come back to Earth, on an approximate parabolic trajectory. However, we know that this phenomenon has behind it the principle of energy conservation, and this raises the possibility of launching the rock with an enormous speed, as to beat the Earth gravity, and have no need to come back. This enormous speed is the escape velocity.

(2) In practice now, the escape velocity can be computed via energy, as follows:

$$\frac{mv^2}{2} = \frac{GMm}{r} \implies v = \sqrt{2gr}$$

(3) Thus, we have the formula in the statement. However, in relation with applications, this is just a beginning, on one hand because we have to beat as well drag, and on the other hand because the Earth spins, which modifies the escape velocity as well. □

In practice now, here are a few selected escape velocities, inside the Solar system:

Location	Beating	Escape velocity, km/s
Sun	Sun gravity	617.5
Mercury	Mercury gravity	4.25
Mercury	Sun gravity	67.7
Earth	Earth gravity	11.18
Earth	Sun gravity	42.1
Moon	Moon gravity	2.38
Moon	Earth gravity	1.4
Neptune	Neptune gravity	23.56
Neptune	Sun gravity	7.7

And we will stop our study here. More on such things, mechanics, in chapter 10.

8e. Exercises

We are now experts in classical mechanics, and as exercises, we have:

EXERCISE 8.24. *Learn more about potential energy, in several dimensions.*

EXERCISE 8.25. *Learn about the concept of resonance, and its applications.*

EXERCISE 8.26. *Build a clock, then a watch, using your pendulum knowledge.*

EXERCISE 8.27. *Learn more about the polar parametrization of the conics.*

EXERCISE 8.28. *Do more computations, as suggested, for the 2-body problem.*

EXERCISE 8.29. *Deduce the 1D free fall formula from what we have in 2D.*

EXERCISE 8.30. *Learn more about the center of mass, and its applications.*

EXERCISE 8.31. *Have a look at the equations of the plane 3-body problem.*

As bonus exercise, learn as well about fluid mechanics, drag and ballistics.

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. Lines, planes

Welcome to 3 dimensions, and in the hope that you already live there, enjoying of course mathematics and physics, but also things like sports, food, friends and parties, what a wonderful world we are blessed to live in. For us here, many things to be done, as a continuation of our previous learning, both in mathematics and physics.

Advising us will be the rat, who's very skilled at 3D. I mean, cats for instance tend to dissapear sometimes, that one who went outside in the garden reappearing a few minutes after in the room upstairs, without any explanation, and with such things taking their toll on your mental health, if not used to them. Rats, howewer, are much better at this business, quickly dissapearing on a regular basis, I mean that rat that you briefly saw near the stove, for 3–4 seconds, will be nowhere to be found, starting from $t = 5$.

This being said, I have my tricks, and I can sometimes have conversations with rat. And when asked about what to do, at this point in this book, here is his advice:

RAT 9.1. *For slow animals like you, lacking proper sight, smell, hearing, taste and touch, mathematics and vectors are the way, for some understanding of 3D.*

Thanks rat, this sounds reasonable, so we will proceed with some straightforward 3D mathematics, and in what regards physics, chemistry, catching small and big animals, and many other 3D things, we will see later. Getting started now, we first have:

DEFINITION 9.2. *The points $u \in \mathbb{R}^3$ can be represented as vectors*

$$u = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

and are subject to the addition and multiplication by scalars operations

$$u + u' = \begin{pmatrix} x + x' \\ y + y' \\ z + z' \end{pmatrix} \quad , \quad \lambda u = \begin{pmatrix} \lambda x \\ \lambda y \\ \lambda z \end{pmatrix}$$

geometrically corresponding to forming a parallelogram, and dilating by λ.

As a first observation, we already met these notions in $N = 2$ dimensions, in chapter 5 when doing basic plane geometry, and then in chapter 7, when doing more advanced geometry. And the point is that our computations from chapter 7, involving linear maps and matrices, show that the vectors must be written indeed as above, vertically.

We would like to first study the simplest objects in $\mathbb{R}^3$, namely the lines and planes. For this purpose, let us first record the following result in $\mathbb{R}^2$, regarding the simplest objects there, the lines, with the usual 2D convention that the vectors are denoted $u = \binom{x}{y}$:

PROPOSITION 9.3. *The equation of lines in $\mathbb{R}^2$ is as follows, with $(a, b) \neq (0, 0)$,*

$$L \quad : \quad ax + by = k$$

and with (a, b, k) unique, up to multiplication by a scalar. Given a second such line,

$$L' \quad : \quad a'x + b'x = k'$$

assumed to be different from the first one, $L \neq L'$, so $(a, b, k) \not\sim (a', b', k')$, we have:

$$L||L' \iff (a, b) \sim (a', b')$$

In the opposite case, namely $L \not|| L'$, the intersection $L \cap L'$ is a point.

PROOF. Many things can be said here, the idea being as follows:

(1) To start with, the result is something intuitive, and elementary, and we will leave its proof, done via some easy equation solving, as an exercise. And with the comment that we will prove in a moment, with details, a similar result, more difficult, in 3D.

(2) Next, let us point out the fact that the notion of scalar product that we learned in chapter 7 can be of help, in relation with this. Indeed, the equation of L reads:

$$\left\langle \begin{pmatrix} a \\ b \end{pmatrix}, \begin{pmatrix} x \\ y \end{pmatrix} \right\rangle = k$$

And with this being something quite intuitive, the point being that $v = \binom{a}{b}$ is a vector orthogonal to our line, $v \perp L$, and therefore fully determining it, as above.

(3) As a further piece of mathematics, still in relation with things that we learned in chapter 7, observe that the equation of L can be written as well as follows:

$$\det \begin{pmatrix} b & x \\ -a & y \end{pmatrix} = k$$

So, challenge for us, can we have some understanding of this? And in answer, sure yes, the starting observation here being that $w = \binom{b}{-a}$ is parallel to L.

(4) In short, many things to be done here, for a proper understanding of the lines in $\mathbb{R}^2$, and up to you to spend some time on all this, with some exercises. And with all this being worth the effort, remember the advice Rat 9.1, which applies to 2D as well. □

In 3 dimensions now, we have a similar result, regarding the planes, as follows:

THEOREM 9.4. *The equation of planes in $\mathbb{R}^3$ is as follows, with $(a,b,c) \neq (0,0,0)$,*

$$P \quad : \quad ax + by + cz = k$$

and with (a,b,c,k) unique, up to multiplication by a scalar. Given a second such plane,

$$P' \quad : \quad a'x + b'x + c'z = k'$$

assumed different from the first one, $P \neq P'$, so $(a,b,c,k) \not\sim (a',b',c',k')$, we have:

$$P||P' \iff (a,b,c) \sim (a',b',c')$$

In the opposite case, $P \not|| P'$, the intersection $P \cap P'$ is a line.

PROOF. Many computations needed here, the idea being as follows:

(1) To start with, it is clear that the equation of planes is $ax + by + cz = k$. The parallelism claim is clear too, so we are left with intersecting two non-parallel planes:

$$ax + by + cz = k$$

$$a'x + b'y + c'z = k'$$

(2) In order to deal with this system, observe that our non-parallelism assumption $(a,b,c) \not\sim (a',b',c')$ shows that one of the following 3 things must happen:

$$(a,b) \not\sim (a',b') \quad , \quad (a,c) \not\sim (a',c') \quad , \quad (b,c) \not\sim (b',c')$$

So assume, by permuting if necessary all our vectors, that we have $(a,b) \not\sim (a',b')$. In this case, we can write our system in the following way:

$$ax + by = k - cz$$

$$a'x + b'y = k' - c'z$$

(3) But now, by multiplying the first equation by b', the second equation by b, and then substracting these two equations, we obtain the following formula:

$$(ab' - a'b)x = b'(k - cz) - b(k' - c'z)$$

And the point is that, due to our assumption $(a,b) \not\sim (a',b')$, which tells us that we have $ab' \neq a'b$, we obtain in this way a formula for x in terms of z, as follows:

$$x = \frac{b'(k - cz) - b(k' - c'z)}{ab' - a'b}$$

(4) Similarly, by multiplying the first equation in (2) by a', the second equation by a, and then substracting these two equations, we obtain the following formula:

$$(a'b - ab')x = a'(k - cz) - a(k' - c'z)$$

Thus, as before, we have as well a formula for y in terms of z, as follows:

$$y = \frac{a'(k - cz) - a(k' - c'z)}{a'b' - ab'}$$

(5) Time now to conclude. The solutions of the system in (2) are as follows:

$$(x,y,z)=\left(\frac{b'(k-cz)-b(k'-c'z)}{ab'-a'b},\frac{a'(k-cz)-a(k'-c'z)}{a'b'-ab'},z\right)$$

With $z=t$, and by doing a number of simple manipulations, the solution is:

$$\begin{aligned}(x,y,z)&=\left(\frac{b'(k-ct)-b(k'-c't)}{ab'-a'b},\frac{a'(k-ct)-a(k'-c't)}{a'b'-ab'},t\right)\\&=\left(\frac{b'(k-ct)-b(k'-c't)}{ab'-a'b},\frac{a(k'-c't)-a'(k-ct)}{ab'-a'b},t\right)\\&=\left(\frac{(b'k-bk')+(bc'-b'c)t}{ab'-a'b},\frac{(ak'-a'k)+(a'c-ac')t}{ab'-a'b},t\right)\\&=\left(\frac{b'k-bk'}{ab'-a'b},\frac{ak'-a'k}{ab'-a'b},0\right)+\left(\frac{bc'-b'c}{ab'-a'b},\frac{a'c-ac'}{ab'-a'b},1\right)t\end{aligned}$$

We conclude that the intersection of our two planes is indeed a line, as stated.

(6) This being said, time perhaps for some doublechecks? In what regards the point appearing above, at $t=0$, this certainly belongs to our first plane, as shown by:

$$a\cdot\frac{b'k-bk'}{ab'-a'b}+b\cdot\frac{ak'-a'k}{ab'-a'b}+c\cdot 0=\frac{(ab'-a'b)k}{ab'-a'b}=k$$

Also, this $t=0$ point belongs as well to our second plane, as shown by:

$$a'\cdot\frac{b'k-bk'}{ab'-a'b}+b'\cdot\frac{ak'-a'k}{ab'-a'b}+c'\cdot 0=\frac{(ab'-a'b)k'}{ab'-a'b}=k'$$

(7) Next, let us see what happens to this, when adding that t part. With respect to the first plane, things fine, as shown by the following computation:

$$a\cdot\frac{bc'-b'c}{ab'-a'b}+b\cdot\frac{a'c-ac'}{ab'-a'b}+c\cdot 1=\frac{(a'b-ab')c}{ab'-a'b}+c=0$$

As for the situation with respect to the second plane, things fine too, according to:

$$a'\cdot\frac{bc'-b'c}{ab'-a'b}+b'\cdot\frac{a'c-ac'}{ab'-a'b}+c'\cdot 1=\frac{(a'b-ab')c'}{ab'-a'b}+c'=0$$

Summarizing, things fine, theorem proved, checked and doublechecked.

(8) Finally, in analogy with what we said in the proof of Proposition 9.3, observe that scalar products can help in relation with this, with the equation of P reading:

$$\left\langle\begin{pmatrix}a\\b\\c\end{pmatrix},\begin{pmatrix}x\\y\\z\end{pmatrix}\right\rangle=k$$

And with this being something quite intuitive, the point being that the vector on the left is orthogonal to our plane, and therefore fully determines it, as above. □

Getting now to higher dimensions, inspired by the above results, let us formulate:

THEOREM 9.5. *The equation of the hyperplanes in $\mathbb{R}^N$, which generalize the lines at $N = 2$, and the usual planes at $N = 3$, is as follows,*

$$P \quad : \quad a_1x_1 + \ldots + a_Nx_N = k$$

with $(a_1, \ldots, a_N) \neq (0, \ldots, 0)$, and with $(a_1, \ldots, a_N, k)$ being unique, up to the multiplication by a nonzero scalar. Given a second such hyperplane,

$$P' \quad : \quad a'_1x_1 + \ldots + a'_Nx_N = k'$$

assuming $P \neq P'$, we have $P||P'$ precisely when $(a_1, \ldots, a_N) \sim (a'_1, \ldots, a'_N)$. In the opposite case, $P \not||\, P'$, the intersection $P \cap P'$ is a $(N-2)$-dimensional space.

PROOF. This is actually a bit tricky, the idea being as follows:

(1) To start with, we don't have much intuition in higher dimensions, so we will take as definition for the hyperplanes $P \subset \mathbb{R}^N$ what is said in the statement.

(2) Next, the claims regarding the parametrization and parallelism are both clear. Thus, we are left with intersecting two non-parallel hyperplanes, and see what we get.

(3) But here, we are a bit in trouble, because even if willing to do computations, as those in the proof of Theorem 9.4, it is a bit unclear what exactly we want to prove.

(4) And so, we will have to stop here, say by declaring that a $(N-2)$-dimensional subspace of $\mathbb{R}^N$ appears by definition as an intersection of non-parallel hyperplanes. □

Not very bright, what we did. In order to further clarify all this, let us start with:

DEFINITION 9.6. *A vector space is a subspace $V \subset \mathbb{R}^N$ satisfying*

$$x, y \in V \implies x + y \in V \quad , \quad x \in V, \lambda \in \mathbb{R} \implies \lambda x \in V$$

in analogy with what happens for lines, planes and so on, passing through the origin.

To be more precise here, observe that we must have $0 \in V$, as we can see by taking $\lambda = 0$ in the second condition. In practice, the main examples are as follows:

– The total space, $\mathbb{R}^N$ itself.

– The hyperplanes from Theorem 9.5, assumed to pass through the origin, meaning given by equations of type $a_1x_1 + \ldots + a_Nx_N = 0$, with $(a_1, \ldots, a_N) \neq (0, \ldots, 0)$. Indeed, both the vector space axioms are trivially satisfied, for such a hyperplane.

– The intersections of hyperplanes discussed in Theorem 9.5, again assumed to pass through the origin. Indeed, this is clear, and more generally we can say that, given two vector spaces $V, W \subset \mathbb{R}^N$, their intersection $V \cap W \subset \mathbb{R}^N$ is a vector space too.

– And so on, more and more intersections, up to reaching the planes, then the lines, and finally the origin $\{0\}$ itself, which is indeed a vector space, satisfying our axioms.

With this discussed, what is next? Based on what we have so far, definitions and theorems alike, let us formulate the following claim, which fits with everything:

CLAIM 9.7. *The vector spaces $V \subset \mathbb{R}^N$ are precisely the intersections of hyperplanes*

$$V = P_1 \cap \ldots \cap P_k$$

with such an intersection having, in the generic case, dimension $N - k$.

In order to understand now what exactly is going on, with this claim, again inspired by the basic examples discussed before, let us formulate the following definition:

DEFINITION 9.8. *Let $V \subset \mathbb{R}^N$ be a vector space. We call a family $\{v_i\} \subset V$:*

(1) *Generating, if each $x \in V$ decomposes as $x = \sum_i \lambda_i v_i$.*
(2) *Linearly independent, when $\sum_i \lambda_i v_i = 0 \implies \lambda_i = 0$.*
(3) *A basis, if each $x \in V$ decomposes uniquely as $x = \sum_i \lambda_i v_i$.*

In practice, this is something quite intuitive, coming from what we know about $\mathbb{R}^N$ and its basic subspaces. At the level of things that can be said, these are as follows:

(1) Any generating set can be shrinked to a basis.

(2) Any linearly independent system can be extended to a basis.

(3) Being a basis means to be generating, and linearly independent.

And we will leave the formal proofs of these facts, which are all very intuitive, as an exercise. Next, we can talk about the dimension of vector spaces $V \subset \mathbb{R}^N$, as follows:

DEFINITION 9.9. *The dimension of a vector space $V \subset \mathbb{R}^N$ is the number of elements $\dim V = M$ of any of its bases $\{v_1, \ldots, v_M\}$, which is independent of the basis.*

As before with other such things, this definition is based on some mathematics. Indeed, we must first prove that V has a basis, and this comes either by starting with an arbitrary generating set, and shrinking it to a basis, or by starting with an arbitrary linearly independent system, and extending it to a basis. As for the fact that any two bases have the same cardinality, this is again something very intuitive, also exercise for you.

All this is very nice, and clarifies what we said before in Theorem 9.5, and in Claim 9.7 too. Now, time for a main theorem, in relation with this? Here that theorem is:

THEOREM 9.10. *Given a map $f : \mathbb{R}^N \to \mathbb{R}^N$ which is linear, in the sense that*

$$f(x + y) = f(x) + f(y) \quad , \quad f(\lambda x) = \lambda f(x)$$

its kernel $\ker f = \{v | f(v) = 0\}$ and image are both vector spaces, and we have:

$$\dim(\ker f) + \dim(Imf) = N$$

In particular, we have f injective $\iff$ f surjective $\iff$ f bijective.

PROOF. Many things going on here, the idea being as follows:

(1) To start with, the most accessible assertion in our theorem is the last one. Indeed, with $v_i = f(e_i)$, where $\{e_i\}$ is the standard basis of $\mathbb{R}^N$, our map f being injective means that $\{v_i\}$ is linearly independent, being surjective means that $\{v_i\}$ is generating, and being bijective means that $\{v_i\}$ is a basis. But, according to our general theory of bases developed before, for a subset $\{v_1, \ldots, v_N\} \subset \mathbb{R}^N$, being generating is the same as being linearly independent, or as being a basis, so we get the above equivalences regarding f.

(2) As for the formula $\dim(\ker f) + \dim(Imf) = N$, which is something more general, and which holds in fact for the arbitrary linear maps, of type $f : \mathbb{R}^N \to \mathbb{R}^M$, this can be proved too, either by fine-tuning the above arguments in (1), or by using some linear algebra, as developed below, in chapter 13. So, exercise here for you, to have a look at this, and if needed come back here later, after reading chapter 13. □

So long for basic vector calculus, and various things that must be known. All this might be a bit hard to digest, I agree, but the trick is to not insist, and come back here later, say after reading chapter 13, which will be dedicated to linear algebra.

9b. Bodies, shape

Let us look now at curves, surfaces, and other bodies, in 3D or more. We would like to understand if our object has holes, and how many holes, of which type, and so on. In order to discuss this, let us start with something a bit abstract, as follows:

DEFINITION 9.11. *An object X is called path connected when any two points $x, y \in X$ can be connected by a path. That is, given any two points $x, y \in X$, we can find a continuous function $f : [0, 1] \to X$ such that $f(0) = x$ and $f(1) = y$.*

The problem is now, given a connected object X, how to count its "holes". And this is quite subtle problem, because as examples of such spaces we have:

(1) The sphere, the donut, the double-holed donut, the triple-holed donut, and so on. These objects are quite simple, and intuition suggests to declare that the number of holes of the N-holed donut is, and you guessed right, N.

(2) However, we have as well as example the empty sphere, I mean just the crust of the sphere, and while this obviously falls into the class of "one-holed objects", this is not the same thing as a donut, its hole being of different nature.

(3) As another example, consider again the sphere, but this time with two tunnels drilled into it, in the shape of a cross. Whether that missing cross should account for 1 hole, or for 2 holes, or for something in between, I will leave it up to you.

Summarizing, things are quite tricky, suggesting that the "number of holes" of an object X is not an actual number, but rather something more complicated. Now with this in mind, let us formulate the following abstract definition:

DEFINITION 9.12. *A group is a set G endowed with a multiplication operation*

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

which must satisfy the following conditions:

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

As a first observation, this is an abstract definition, making use of some abstract notations, that we can fine-tune and adapt to our situation, if needed. For instance, in order to avoid certain confusions, we can use a dot for the multiplication operation:

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

This being said, we can use in fact any symbol for the multiplication operation, as for instance $(g,h)\to g\heartsuit h$. The same goes for the unit, which can be denoted as $\bullet\in G$ if we want to, and for the inverses as well, which can be denoted $\not{g}$, if that is convenient.

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

PROPOSITION 9.13. *We have the following groups, and non-groups:*

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

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

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

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

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

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

Observe now that $(\mathbb{Z},+)$ has the following special property:

$$g+h=h+g$$

However, we have not included this property in our axioms from Definition 9.12, which with the notations there would read $gh=hg$, because we would like our notion of group to cover as well examples which do not satisfy this, $gh\neq hg$. More on these later.

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

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

Observe that these latter groups all have the property $g+h=h+g$.

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

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

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

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

Finally, observe that $(\mathbb{Q}^*,\cdot\,)$ has the commutativity property $gh=hg$.

(5) Once again, the group axioms are satisfied for $(\mathbb{R}^*,\cdot\,),(\mathbb{C}^*,\cdot\,)$, for the same reasons as for $(\mathbb{Q}^*,\cdot\,)$, and with the remark that for $\mathbb{C}^*$ we are using here the fact that the nonzero complex numbers can be inverted, with this coming from the following formula:

$$\frac{1}{a+ib}=\frac{a-ib}{a^2+b^2}$$

Observe that these latter groups both have the property $gh=hg$.

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

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

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

Getting back now to our geometry questions, good news, we can formulate:

DEFINITION 9.14. *The homotopy group $\pi_1(X)$ of a connected object X is the group of loops based at a given point $*\in X$, with the following conventions,*

(1) *Two such loops are identified when one can pass continuously from one loop to the other, via a family of loops indexed by $t\in[0,1]$,*
(2) *The composition of two such loops is the obvious one, namely is the loop obtained by following the first loop, then the second loop,*
(3) *The unit loop is the null loop at $*$, which stays there, and the inverse of a given loop is the loop itself, followed backwards,*

with the remark that the group $\pi_1(X)$ defined in this way does not depend on the choice of the given point $*\in X$, where the loops are based.

This definition is obviously something non-trivial, based on some preliminary thinking on the subject, the technical details being as follows:

– The fact that the set $\pi_1(X)$ defined as above is indeed a group is obvious, with all the group axioms being clear from definitions, good exercise for you.

– Obvious as well is the fact that, since X is assumed to be connected, this group does not depend on the choice of the given point $* \in X$, where the loops are based.

– Finally, we will see in a moment that the loop composition does not necessarily satisfy $gh = hg$. Thus, Definition 9.12 as stated is the exact framework what we need.

As basic examples now, for objects having "no holes", such as $\mathbb{R}$ itself, or $\mathbb{R}^N$, and so on, we have $\pi_1 = \{1\}$. In fact, having no holes can only mean, by definition, $\pi_1 = \{1\}$:

DEFINITION 9.15. *An object X is called simply connected when:*

$$\pi_1(X) = \{1\}$$

That is, any loop inside our object X must be contractible.

As further illustrations for Definition 9.14, here are a few basic computations, for some objects that we are familiar with, with the convention that $F_N =< g_1, \ldots, g_N >$ is the free group, generated by variables $g_1, \ldots, g_N$, with no relations between them:

THEOREM 9.16. *We have the following computations of homotopy groups:*

(1) *For the circle, we have $\pi_1 = \mathbb{Z}$.*
(2) *For the torus, we have $\pi_1 = \mathbb{Z} \times \mathbb{Z}$.*
(3) *For the disk minus 2 points, we have $\pi_1 = F_2$.*
(4) *In fact, for the disk minus N points, we have $\pi_1 = F_N$.*

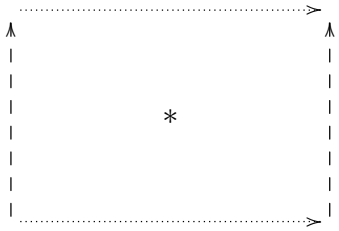
PROOF. These results are all standard, as follows:

(1) The first assertion is clear, because a loop on the circle must wind $n \in \mathbb{Z}$ times around the center, and this parameter $n \in \mathbb{Z}$ uniquely determines the loop, up to the identification in Definition 9.14. Thus, the homotopy group of the circle is the group of such parameters $n \in \mathbb{Z}$, which is of course the group $\mathbb{Z}$ itself.

(2) In what regards now the second assertion, the torus being a product of two circles, we are led to the conclusion that its homotopy group must be some kind of product of $\mathbb{Z}$ with itself. But pictures show that the two standard generators of $\mathbb{Z}$, and so the two copies of $\mathbb{Z}$ themselves, commute, $gh = hg$, so we obtain the product of $\mathbb{Z}$ with itself, subject to commutation, which is the usual product $\mathbb{Z} \times \mathbb{Z}$:

$$\left\langle g, h \middle| gh = hg \right\rangle = \mathbb{Z} \times \mathbb{Z}$$

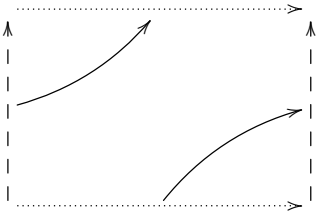
It is actually instructive here to work out explicitly the proof of the commutation relation. We can use the usual drawing convention for the torus, which is as follows, the idea being that the similar looking arrows must be identified, in the obvious way:



The standard generators g, h of the homotopy group are then as follows:

Regarding now the two compositions gh, hg, these are as follows:

But these two pictures coincide, up to homotopy, with the following picture:



Thus we have indeed $gh = hg$, as desired, which gives the formula in (2).

(3) Regarding now the disk minus 2 points, the result here is quite clear, because the homotopy group is generated by the 2 loops around the 2 missing points, and these 2 loops are obviously free, algebrically speaking. Thus, we obtain a free product of the group $\mathbb{Z}$ with itself, which is by definition the free group on 2 generators F_2.

(4) This is again clear, because the homotopy group is generated here by the N loops around the N missing points, which are free, algebrically speaking. Thus, we obtain a N-fold free product of $\mathbb{Z}$ with itself, which is the free group on N generators F_N. □

There are many other interesting things that can be said about homotopy groups. Also, another thing that can be done with the arbitrary objects X, again in relation with studying their "shape", is that of looking at the fiber bundles over them, again up to continuous deformation. We are led in this way into a group, called $K_0(X)$. Moreover, both $\pi_1(X)$ and $K_0(X)$ have higher analogues $\pi_n(X)$ and $K_n(X)$ as well, and a main goal of topology is that of understanding all these groups, along with some other groups, of similar flavor, which can be constructed as well. Many things to be learned here.

9c. Curves, surfaces

Getting back now to more standard mathematics, in relation with vectors and other things that we know well, as a next objective, we can do some algebraic geometry in $\mathbb{R}^3$, in continuation to what we did before in chapter 7 in $\mathbb{R}^2$, in relation with the conics.

But here, in 3 dimensions, 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 9.17. *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. I mean, just try, and you will understand.

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 9.18. *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 $C \subset \mathbb{R}^2$ before, in chapter 7, we have the following result:

THEOREM 9.19. *The degree 2 surfaces $S \subset \mathbb{R}^3$, called quadrics, are the ellipsoid*

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

which is the only compact one, plus 16 more, which can be explicitly listed.

PROOF. We will be quite brief here, because we intend to rediscuss all this in a moment, with 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 9.20. *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 9.21. *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. However, this is not an issue, the idea being that, with some algebra helping, allowing $r = \infty$ forces in fact $r < \infty$, and with this being called the Hilbert basis theorem, from algebraic geometry.

Finally, observe the relation with our previous vector space considerations, from the beginning of this chapter, and more specifically with Claim 9.7, stating that any vector space $V \subset \mathbb{R}^N$ must appear as an intersection of hyperplanes, as follows:

$$V = P_1 \cap \ldots \cap P_k$$

Thus, as a conclusion I guess, you are learning here mathematics from an algebraic geometer, viewing all things algebra in a geometric way. In case you like this type of approach, don't hesitate later to learn a bit of systematic algebraic geometry.

Back to work now, let us look more in detail at the hypersurfaces. We have here:

THEOREM 9.22. *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 empty set 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. So, theorem proved, modulo some linear algebra, that we will learn later, in chapter 13. □

As a conclusion to all this, foundations of algebraic geometry understood, and with this being potentially a very good thing, useful for both mathematics and physics. However, we should mention that the basics are not over with this, because Definition 9.21 has a well-known rival, namely the definition of a smooth manifold, and with the branch of mathematics studying such beasts, which are equally useful in mathematics and physics, being called differential geometry. And, more on this, later in this book.

9d. Regular polyhedra

Changing a bit topics, but still about the very nature of our 3 dimensions that we live in, let us discuss now the regular polyhedra. Following Plato, we have:

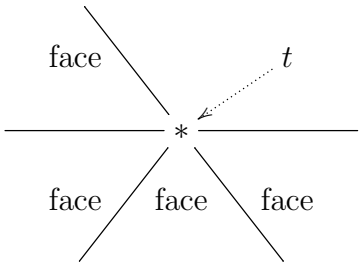
THEOREM 9.23. *There are* 5 *regular polyhedra, called Platonic solids, namely:*

(1) *Tetrahedron, having* 4 *vertices and* 4 *faces.*
(2) *Octahedron, having* 6 *vertices and* 8 *faces.*
(3) *Cube, having* 8 *vertices and* 6 *faces.*
(4) *Icosahedron, having* 12 *vertices and* 20 *faces.*
(5) *Dodecahedron, having* 20 *vertices and* 12 *faces.*

PROOF. Many things can be said here, the idea being as follows:

(1) Let us try to figure out how a regular polyhedron looks like. There are a number of faces meeting at each vertex, ≥ 3 faces to be more precise, and when flattening the

polyhedron there, we can see appear an angle t, called angle defect at that vertex:



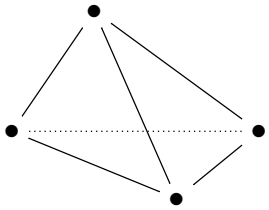
Now since hexagons and higher have angles $\geq 120°$, these cannot be used for constructing polyhedra, due to $t > 0$. In fact, still due to $t > 0$, we are left with 5 cases:

– Polyhedron made of triangles, with 3 or 4 or 5 faces meeting at each vertex.

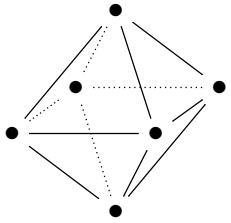
– Polyhedron made of squares, with 3 faces meeting at each vertex.

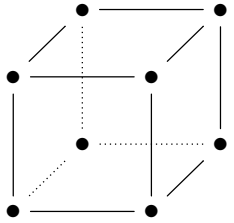
– Polyhedron made of pengatons, with 3 faces meeting at each vertex.

(2) Now let us try to construct the solutions. In the first case, polyhedron made of triangles, with 3 faces meeting at each vertex, we obtain the tetrahedron:



(3) Two other obvious solutions, corresponding to the second and fourth cases above, triangles meeting $\times 4$, and squares meeting $\times 3$, are the octahedron and the cube:



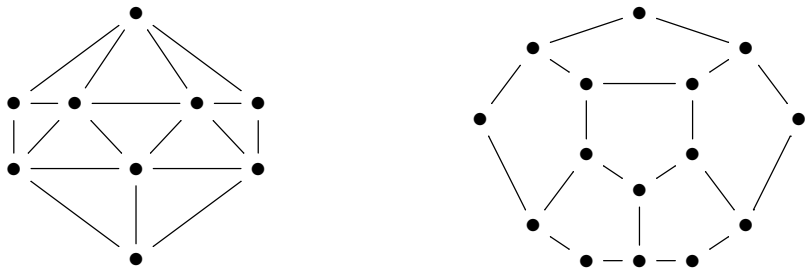


Before going further, observe that there is a relation between these two polyhedra, with the vertices of the octahedron appearing at the middle of the faces of the cube, and vice versa. Due to this, we say that the octahedron and cube are dual, and with this explaining why their number of vertices and faces are interchanged, as follows:

$$(6, 8) \leftrightarrow (8, 6)$$

By the way, observe that the tetrahedron is self-dual, $(4, 4) \leftrightarrow (4, 4)$.

(4) Back to constructing solutions, we are left with studying the third and fifth cases in (1), namely triangles meeting $\times 5$, and pentagons meeting $\times 3$. And here, by some kind of miracle, we have indeed solutions, namely the icosahedron and dodecahedron, which look as follows, with in each case half of the faces, those facing us, represented:



As before with the octahedron and cube, these two latter polyhedra are dual, with this intechanging their number of vertices and faces, $(12, 20) \leftrightarrow (20, 12)$. □

As a continuation of this, let us have a look at the various numbers appearing in Theorem 9.23, namely the number of vertices v, and the number of faces f. We can equally talk about the number of edges e, and the data is as follows:

	T	O	C	I	D
v	4	6	8	12	20
f	4	8	6	20	12
e	6	12	12	30	30

Now if you are a number theory nerd, and hope so are you, I am pretty much sure that, after contemplating a bit this table, you will notice that $v + f = e + 2$. Quite remarkably, this is something which is true, and under very general assumptions, for any connected planar graph, with the precise result here, due to Euler, being as follows:

THEOREM 9.24. *For a connected planar graph we have the Euler formula*

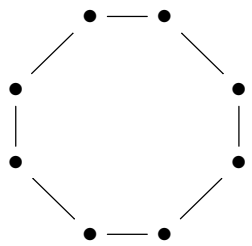
$$v - e + f = 2$$

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 the N-gon graph:



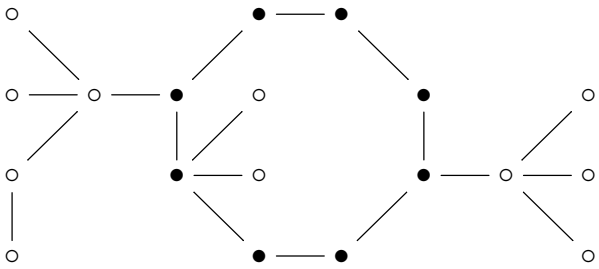
Then, for this graph, the Euler formula holds indeed, as follows:

$$N - N + 2 = 2$$

(3) With this example 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:

$$N - (N - 1) + 1 = 2$$

(4) 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:

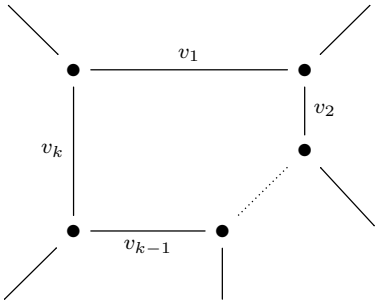


But here we can argue, again based on the fact that for a rooted tree, the non-root vertices are in obvious bijection with the edges, that removing all these trees won't change the problem. So, we are left with the problem for the N-gon, already solved in (2).

(5) And so on, the idea being that we can first remove all the trees, by using the argument in (4), 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 (2), or by recurrence.

(6) 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:



(7) 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:

THEOREM 9.25. *For a convex polyhedron we have the Euler formula*

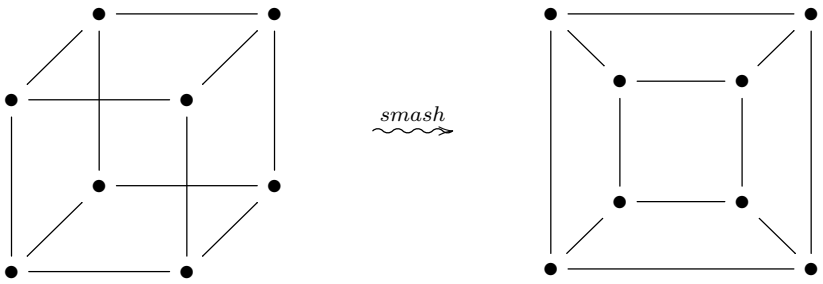
$$v - e + f = 2$$

with v, e, f being the number of vertices, edges and faces.

PROOF. This is more or less the same thing as Theorem 9.24, save for getting rid of the internal trees of the planar graph there, the idea being as follows:

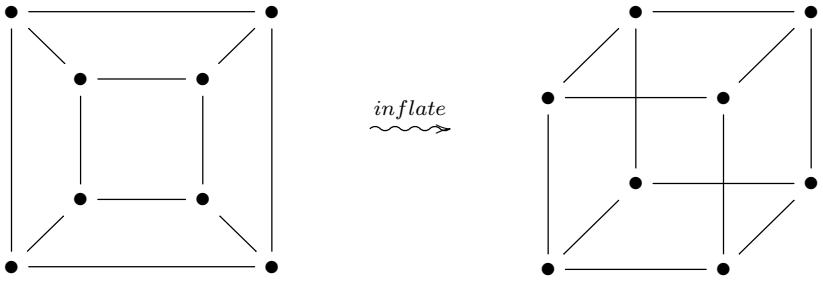
(1) In one sense, consider a convex polyhedron P. We can then enlarge one face, as much as needed, and then smash our polyhedron with a big hammer, as to get a planar

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 9.24.

(2) Conversely, consider a connected planar graph X. Then, save for getting rid of the internal trees, as explained in the proof of Theorem 9.24, we can assume that we are dealing with an agglomeration of N-gons, again as explained in the proof of Theorem 9.24. 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. Getting back now to the regular polyhedra, we have the following version of Theorem 9.23, obtained by using this technology:

THEOREM 9.26. *The Euler formula $v + f = e + 2$ allows for only 5 regular polyhedra, which are the 5 Platonic solids, as follows:*

(1) *Tetrahedron:* $4 + 4 = 6 + 2$.
(2) *Octahedron:* $6 + 8 = 12 + 2$.
(3) *Cube:* $8 + 6 = 12 + 2$.
(4) *Icosahedron:* $12 + 20 = 30 + 2$.
(5) *Dodecahedron:* $20 + 12 = 30 + 2$.

PROOF. As mentioned, this is Euler's remake of the Plato theorem, using his formula $v + f = e + 2$. Indeed, given a regular polyhedron, let us denote by m the number of

edges meeting at any vertex, and by n the size of the polygons used. We have then:

$$v = \frac{2e}{m} \quad , \quad f = \frac{2e}{n}$$

Thus, for a regular polyhedron, the Euler formula $v + f = e + 2$ becomes:

$$\frac{2e}{m} + \frac{2e}{n} = e + 2$$

Now observe that this latter formula gives the following estimate:

$$\frac{1}{m} + \frac{1}{n} = \frac{1}{2} + \frac{1}{e} > \frac{1}{2}$$

Thus, $m, n \geq 3$ must be small, and more specifically, the allowed values are:

$$(m, n) = (3, 3), (4, 3), (5, 3), (3, 4), (3, 5)$$

But these are exactly the 5 cases that we met in the proof of Plato, and the continuation is the same as there, further helped by the fact that, now that we have m, n, we can recapture e, and then v, f as well, with the corresponding 5 cases being as follows:

$$(v, f, e) = (4, 4, 6), (6, 8, 12), (12, 20, 30), (8, 6, 12), (20, 12, 30)$$

To be more precise, as before in the proof of Theorem 9.23, we obtain in this way the tetrahedron, the octahedron, the icosahedron, the cube and the dodecahedron. □

As a last topic of discussion for this chapter, let us look now at the symmetry groups of the Platonic solids, which is something that can be useful for many purposes, including mathematics, physics, chemistry and biology. The result here is as follows:

THEOREM 9.27. *The symmetry groups $G \subset O_3$ and the orientation-preserving symmetry groups $SG \subset SO_3$ of the Platonic solids are as follows:*

(1) *Tetrahedron: $G = S_4$, $SG = A_4$.*
(2) *Octahedron and cube: $G = S_4 \times \mathbb{Z}_2$, $SG = S_4$.*
(3) *Icosahedron and dodecahedron: $G = A_5 \times \mathbb{Z}_2$, $SG = A_5$.*

PROOF. This is something a bit more advanced, the idea with all this, explanations for the various groups appearing in the statement, and proof itself, being as follows:

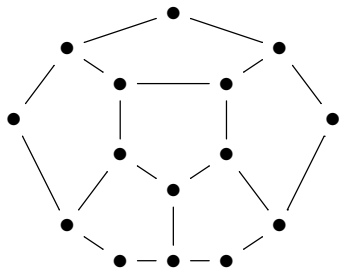
(1) To start with, we denote by O_3 the group of linear transformations of $\mathbb{R}^3$ which preserve the vector lengths $||x||$, or equivalently, the scalar products $< x, y >$. Also, we denote by $SO_3 \subset O_3$ the subgroup of transformations which are orientation-preserving.

(2) In what regards the tetrahedron, we certainly have $G = S_4$, permutation group of 4 points, and then $SG = A_4$, called alternating subgroup, or tetrahedral group.

(3) Regarding now the cube, here we have $SG = S_4$, with this S_4 being best understood as permuting the diagonals of the cube, and then $G = S_4 \times \mathbb{Z}_2$.

(4) As for the ocathedron, this being dual to the cube, the symmetry groups are the same. Let us mention also that $SG = S_4$ is also called octahedral group.

(5) In what regards now the icosahedron and dodecahedron, these are dual too, so they have the same symmetry groups. In order to compute these common symmetry groups, let us look at the dodecahedron, whose picture, facing us, was as follows:



But then, the idea is that we have exactly 5 cubes having vertices among the 20 vertices of the dodecahedron, and the symmetries $g \in SG$ come from permutations of these 5 cubes, which must be alternating. Thus $SG = A_5$, and $G = A_5 \times \mathbb{Z}_2$, as claimed. □

9e. Exercises

Welcome to algebraic geometry I guess, and as exercises on this, we have:

EXERCISE 9.28. *Learn more about vector space bases, and related topics.*

EXERCISE 9.29. *Do not forget to learn as well about orthonormal bases.*

EXERCISE 9.30. *Learn about the various groups formed by the square matrices.*

EXERCISE 9.31. *Compute some more homotopy groups, for bodies of your choice.*

EXERCISE 9.32. *Learn about the various plane curves of small degree $d > 2$.*

EXERCISE 9.33. *Learn more about the quadrics, and their classification.*

EXERCISE 9.34. *Learn about the genus of a surface, or of a graph.*

EXERCISE 9.35. *Work out the details, for the symmetry groups of Platonic solids.*

As bonus exercise, start reading an introductory algebraic geometry book.

CHAPTER 10

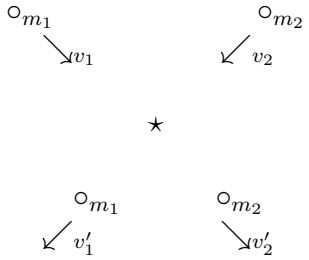
Advanced mechanics

10a. Collisions, scattering

Back now to physics, I have this feeling that, with all the geometry that we learned in chapter 9, we are good for some serious 3D mechanics. And, looking at what we did before in this book, in chapters 2-4-6-8, on various mechanics topics, the first question to be discussed here would be 3D motion, in the absence of forces.

Let us start with a study of the collisions, as a continuation of our 1D discussion from chapter 2. However, things in higher dimensions are quite tricky, due to:

FACT 10.1. *A collision between two particles is normally a 2D problem, happening in the plane $E = span(v_1, v_2)$. Schematically, we can represent the collision as follows:*



However, as bad news, we have to compute two vectors v_1', v_2', accounting for a total of 4 numbers. But what we have is one vector equation, coming from momentum, and one scalar equation, coming from energy, so a total of 3 equations. Thus, we won't be able to do the math, unless we know something more on the collision mechanism.

To be more precise here, as already mentioned on several occasions, we don't really know what happens, at the microscopic level, when collisions take place, be them plastic or elastic, or of any other kind. And in the case of elastic collisions in $N \geq 2$ dimensions, there are several possible mechanisms, which in practice each correspond to a different way in which the original angles of v_1, v_2 get transformed into the output angles of v_1', v_2'. More specifically, the extra parameter that we need is some sort of "scattering angle" θ, coming from the precise mechanism of collision that we are investigating.

Nevermind. By doing the math, partially, as indicated above, we are led to:

THEOREM 10.2. *In the context of an elastic collision between two bodies, assumed as above to happen in $\mathbb{R}^2$, the output speeds are*

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

with the vector parameter $q \in \mathbb{R}^2$ being subject to the following equation:

$$2 < v_2 - v_1, q >= \left(\frac{1}{m_1} + \frac{1}{m_2} \right) ||q||^2$$

Thus, the collision problem is solved, up to an angle $\theta \in \mathbb{R}$.

PROOF. We can do this by using momentum and energy conservation, as follows:

(1) According to our momentum and energy conservation principles, from chapter 2, the output speeds v_1', v_2' are subject to the following two equations:

$$m_1 v_1 + m_2 v_2 = m_1 v_1' + m_2 v_2'$$

$$m_1 ||v_1||^2 + m_2 ||v_2||^2 = m_1 ||v_1'||^2 + m_2 ||v_2'||^2$$

Let us first look at the first equation. This equation can be written as follows:

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

Now if we call q this quantity, which is the individual change of momentum, from the perspective of m_1, and from the opposite perspective of m_2, we have:

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

(2) Now let us plug these values into the second equation above. We obtain:

$$m_1 ||v_1||^2 + m_2 ||v_2||^2 = m_1 \left|\left| v_1 + \frac{q}{m_1} \right|\right|^2 + m_2 \left|\left| v_2 - \frac{q}{m_2} \right|\right|^2$$

By expanding the scalar products and then simplifying, we obtain:

$$\frac{||q||^2}{m_1} + \frac{||q||^2}{m_2} + 2 < v_1, q > -2 < v_2, q >= 0$$

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

(3) As for the last assertion, this is something rather philosophical. To be more precise, we know that our parameter $q \in \mathbb{R}^2$ is subject to an equation of the following type:

$$< v, q >= \lambda ||q||^2$$

Now if we denote by $\theta \in \mathbb{R}$ the oriented angle between v, q, this equation reads:

$$||v|| \cos\theta = \lambda ||q||$$

Thus we can recover $||q||$, and so also $q \in \mathbb{R}^2$ itself, out of this angle $\theta \in \mathbb{R}$, and this leads to the conclusion that our problem is indeed solved up to an angle $\theta \in \mathbb{R}$.

(4) As an illustration for this, let us work out what happens in the case of a 1D collison. Here we already know the 1D answer, from chapter 2, but wait for it. So, let us get back to our usual 1D scheme, with m_1, m_2 about to collide, on a line:

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

In the context of the present result, the fact that we are now in 1D simply tells us that the speed vectors $v_1, v_2 \in \mathbb{R}^2$ are proportional, $v_1 = \lambda v_2$. But this does not force in any way the vector $q \in \mathbb{R}^2$ there to be aligned with v_1, v_2, or if you prefer, the angle $\theta \in \mathbb{R}$ there to be 0. Thus, we are led to the conclusion that our general collision formalism, leading to the present result, theoretically allows escapes from 1D to 2D.

(5) And, is this reasonable, or not? As an example here, you can go to your chemistry lab, pick two big plastic H_2O molecules, and make them collide on a table. And, even if you always throw them with the same speeds v_1, v_2, what happens at the moment of the impact can vary, depending on which side of the first H_2O will hit which side of the second H_2O. To be more precise, you can observe in this way the above-mentioned angle θ, and if you throw hard enough, one of the two H_2O might even take off upon impact, making it clear that there might be as well a 2nd parameter involved there.

(6) Finally, for the mess to be complete, the above chemistry experiments, once the lab guy is not there, of course, and as usual with such interesting chemistry experiments, suggests that we might have as well escapes from 2D to 3D. Thus, our 2D assumption from Fact 10.1, and from the present theorem too, does not really correspond to what happens in the real life, and things are certainly more complicated than that. □

With the above discussed, what is next? You would probably say, time to upgrade our collision formalism, say by assuming that our particles are small elastic spheres, and with what comes out from elasticity theory providing us with the missing data.

Which is certainly interesting and doable, but thinking a bit, what collisions are good for are especially gases and thermodynamics, and here the above H_2O molecule phenomenology, with all sorts of scattering angles θ involved, and with the statistics of all this mattering, is what governs the behavior of vapor. So, quite complicated all this, and more about such things in chapter 12, when talking thermodynamics.

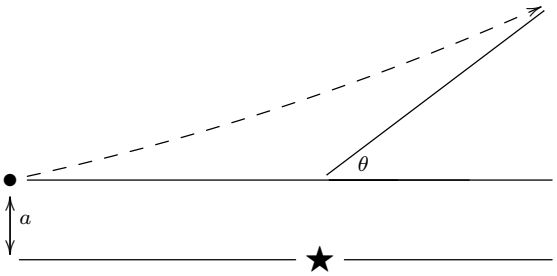
As an alternative to this, which in passing makes an interesting link with the notion of decay, studied in chapter 2, we can have a look at particle physics, according to:

FACT 10.3. *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,* $*_a + *_b \to *_1 + \ldots + *_n$.

So, let us get into this. We already know a bit about the mathematics of decay from chapter 2, and regarding now scattering, let us formulate things as follows:

DEFINITION 10.4. *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 $\theta = \theta(a)$.

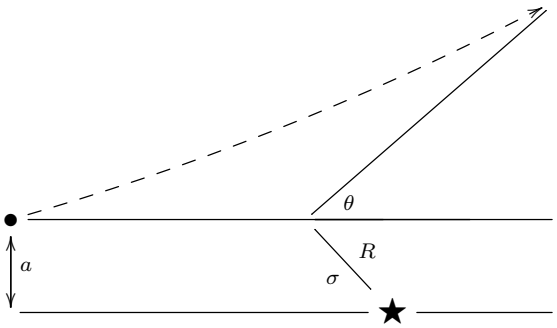
As a first computation now, coming from classical mechanics, we have:

PROPOSITION 10.5. *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 $\theta = 2 \cos^{-1}(a/R)$.

PROOF. In the context from the statement, which is all classical mechanics, with a pinch of basic elasticity added, 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 ★ being pictured, but you get the point. Now with σ being the angle appearing above, 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. □

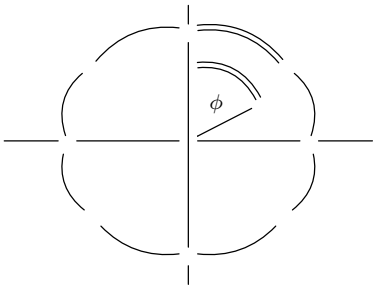
With this understood, let us try to make something more 3D, and statistical, out of this. We can indeed further build on Definition 10.4, as follows:

DEFINITION 10.6. *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, in chapter 2, that is, the main quantity to look at. In order to understand now how the cross section works, we have:

PROPOSITION 10.7. *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 10.6, 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:

$$D(\theta) = \frac{d\sigma}{d\Omega} = \left|\frac{a \cdot da \cdot d\phi}{\sin\theta \cdot d\theta \cdot d\phi}\right| = \left|\frac{a}{\sin\theta} \cdot \frac{da}{d\theta}\right|$$

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 10.8. *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 10.7 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:

$$D(\theta) = \left|\frac{a}{\sin\theta} \cdot \frac{da}{d\theta}\right| = \frac{R\cos(\theta/2)}{\sin\theta} \cdot \frac{R\sin(\theta/2)}{2} = \frac{R^2}{4}$$

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. □

And with this, end of our discussion about collisions. Quite interesting all this, we started as apprentice mechanicians, and ended as particle physics experts. Many more things can be said, as a continuation of the above, following Fermi and others, but it would be probably wiser to stop here, and focus on basic mechanics instead.

10b. Kepler, Coriolis

What is next? Still some 3D mechanics, I guess, by adding forces to the picture, and trying to use the 3D mathematics that we learned in chapter 9, in order to get some sharp results. Unfortunately, browsing through what we did in chapter 9, all interesting mathematics there, no question about this, but frankly, can all that stuff really be of help in relation with our basic mechanics problems. Not clear, hope you agree with me.

Nevermind, and this is actually an issue that we already met in this book, the mathematicians in charge with the math chapters seem to be more poets than scientists, and we will have to live with that. So, for our 3D physics, here is what we will need:

DEFINITION 10.9. *The vector product of two vectors in $\mathbb{R}^3$ is given by*

$$x \times y = ||x|| \cdot ||y|| \cdot \sin\theta \cdot n$$

where $n \in \mathbb{R}^3$ with $n \perp x, y$ and $||n|| = 1$ is constructed using the right-hand rule:

$$\begin{matrix} & \uparrow_{x\times y} \\ \leftarrow_{x} & \\ \swarrow_{y} & \end{matrix}$$

Alternatively, in usual vertical linear algebra notation for all vectors,

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} x_2y_3 - x_3y_2 \\ x_3y_1 - x_1y_3 \\ x_1y_2 - x_2y_1 \end{pmatrix}$$

the rule being that of computing 2×2 determinants, and adding a middle sign.

Obviously, this definition is something quite subtle, and also something very annoying, because you always need this, and always forget the formula. Here are my personal methods. With the first definition, what I always remember is that:

$$||x \times y|| \sim ||x||, ||y|| \quad , \quad x \times x = 0 \quad , \quad e_1 \times e_2 = e_3$$

So, here's how it works. We are looking for a vector $x \times y$ whose length is proportional to those of x, y. But the second formula tells us that the angle θ between x, y must be involved via $0 \to 0$, and so the factor can only be $\sin\theta$. And with this we are almost there, it's just a matter of choosing the orientation, and this comes from $e_1 \times e_2 = e_3$.

As with the second definition, that I like the most, what I remember here is simply:

$$\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} = ?$$

Indeed, when trying to compute this determinant, by developing over the first column, what you get as coefficients are the entries of $x \times y$. And with the good middle sign.

It is also good to know that $x \times y$ exists only in 3 dimensions, with our only tool in $N \neq 3$ dimensions being the scalar product $< x, y >$. And with this being something quite important in both mathematics and physics, philosophically speaking.

Getting to work now, on gravitational mechanics, we already know the main result, elliptic orbits, from chapter 8. However, that elliptic orbit result, proved by Newton, is just a part of the story, with the 3 motion laws discovered by Kepler being as follows:

FACT 10.10 (Kepler laws). *The following happen:*

(1) *The planetary orbits are elliptical, with the Sun at a focus.*
(2) *The radius vector from the Sun to a planet sweeps equal areas in equal times.*
(3) *The ratio of the square of the period of revolution and the cube of the ellipse semimajor axis is the same for all planets.*

So, can we prove such things? We already know well about Kepler 1, from chapter 8, and with the idea in mind of discussing Kepler 2, let us get now into the notion of angular momentum, which is something of fundamental importance, in physics in general.

Many things can be said here, and in relation with gravity, we have:

THEOREM 10.11. *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. □

While the above principle looks like something quite trivial, the mathematics behind it is quite interesting, and has several notable consequences, as follows:

THEOREM 10.12. *In the context of a 2-body problem, the following happen:*

(1) *The fact that the direction of J is fixed tells us that the trajectory of one body with respect to the other lies in a plane.*
(2) *The fact that the magnitude of J is fixed tells us that the Kepler 2 law holds, namely that we have same areas sweeped by Ox over the same times.*

PROOF. This follows indeed from Theorem 10.11, as follows:

(1) We have by definition $J = m(x \times v)$, and since a vector product is orthogonal on both the vectors it comes from, we deduce from this that we have:

$$J \perp x, v$$

But this can be written as follows, with $J^\perp$ standing for the plane orthogonal to J:

$$x, v \in J^\perp$$

Now since J is fixed by Theorem 10.11, we conclude that both x, v, and in particular the position x, and so the whole trajectory, lie in this fixed plane $J^\perp$, as claimed.

(2) Conversely now, forget about Theorem 10.11, and assume that the trajectory lies in a certain plane E. Thus $x \in E$, and by differentiating we have $v \in E$ too, and so $x, v \in E$. Thus $E = J^\perp$, and so $J = E^\perp$, so the direction of J is fixed, as claimed.

(3) Regarding now the last assertion, we already know from the various formulae from chapter 8 that the Kepler 2 law is more or less equivalent to the formula $\dot{\theta} = \lambda/r^2$ there. However, the derivation of $\dot{\theta} = \lambda/r^2$ was something tricky, and what we would like to prove now is that this appears as a simple consequence of $||J|| = \text{constant}$.

(4) In order to to so, let us compute J, according to its definition $J = x \times p$, but in polar coordinates, which will change everything. Since $p = m\dot{x}$, we have:

$$J = r \begin{pmatrix} \cos\theta \\ \sin\theta \\ 0 \end{pmatrix} \times m \begin{pmatrix} \dot{r}\cos\theta - r\sin\theta \cdot \dot{\theta} \\ \dot{r}\sin\theta + r\cos\theta \cdot \dot{\theta} \\ 0 \end{pmatrix}$$

Now recall from the definition of the vector product that we have:

$$\begin{pmatrix} a \\ b \\ 0 \end{pmatrix} \times \begin{pmatrix} c \\ d \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ ad - bc \end{pmatrix}$$

Thus J is a vector of the above form, with its last component being:

$$\begin{aligned} J_z &= rm \begin{vmatrix} \cos\theta & \dot{r}\cos\theta - r\sin\theta \cdot \dot{\theta} \\ \sin\theta & \dot{r}\sin\theta + r\cos\theta \cdot \dot{\theta} \end{vmatrix} \\ &= rm \cdot r(\cos^2\theta + \sin^2\theta)\dot{\theta} \\ &= r^2 m \cdot \dot{\theta} \end{aligned}$$

(5) Now with the above formula in hand, our claim is that the magnitude $||J||$ is constant precisely when $\dot{\theta} = \lambda/r^2$, for some $\lambda \in \mathbb{R}$. Indeed, up to the obvious fact that the orientation of J is a binary parameter, who cannot just switch like that, let us just agree on this, knowing J is the same as knowing J_z, and is also the same as knowing $||J||$. Thus, our claim is proved, and this leads to the conclusion in the statement. □

Moving on, as another basic application of the vector products to questions in mechanics, we have all sorts of useful formulae regarding rotating frames. We first have:

THEOREM 10.13. *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 10.14. *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 10.13, 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 10.15. *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 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 10.15. Good to know.

10c. Relativity, revised

Moving on, as a further application of our vector product technology, we can say something interesting about 3D relativity too, as follows:

THEOREM 10.16. *The Einstein speed summation in 3D is*

$$u +_e v = \frac{1}{1 + < u, v >} \left(u + v + \frac{u \times (u \times v)}{1 + \sqrt{1 - ||u||^2}} \right)$$

in $c = 1$ units, with the $\times$ being usual vector products.

PROOF. Many things can be said here, the idea being as follows:

(1) To start with, we know from chapter 6 that the summation formula is:

$$u +_e v = \frac{1}{1+ < u, v >}\left(u + v + \frac{< u, v > u - < u, u > v}{1 + \sqrt{1 - ||u||^2}}\right)$$

But this gives the formula in the statement, thanks to the following identity:

$$u \times (u \times v) =< u, v > u - < u, u > v$$

(2) This being said, the work in chapter 6, leading to the formula in (1), was in arbitrary N dimensions, and a bit mysterious. And the point is that, in the case $N = 3$ that we are mainly interested in, the vector products provide a better approach.

(3) So, let us discuss now this, which is something quite interesting. Getting back to discussions from chapter 6, the puzzle there was to define a 3D sum, as follows:

$$u +_e v = \frac{u + v + \gamma_{uv}}{1+ < u, v >}$$

But, we know from 1D that the correction term γ_{uv} must vanish when $u \sim v$, and this suggests to use the vector product $u \times v$, which notoriously satisfies:

$$u \sim v \implies u \times v = 0$$

(4) Thus, the correction term γ_{uv} must be something involving $w = u \times v$. Now by thinking a bit, in order to get γ_{uv}, what we have to do is to "transport" $w = u \times v$ to the plane spanned by u, v. But this is simplest done by taking the vector product with any vector in this plane, so as a reasonable candidate for our correction term, we have:

$$\gamma_{uv} = (\alpha u + \beta v) \times w$$

(5) And with this, almost done. Indeed, and making now the link with our previous computations from chapter 6, we can use next the following two formulae:

$$u \times (u \times v) =< u, v > u - < u, u > v$$

$$v \times (u \times v) =< v, v > u - < u, v > v$$

(6) To be more precise, by using these formulae, the continuation of the discussion is exactly as in chapter 6, leading via some phenomenology to the conclusion that we must choose between $\alpha = 0$ and $\beta = 0$. And with the $\beta = 0$ choice, which is the good one on physics grounds, leading via some normalization work to the formula in the statement.

(7) Summarizing, good work that we did, with the Einstein speed summation in 3D being better understood, with our previous N-dimensional considerations, which were a bit mysterious, being now replaced by the more solid arguments from (3,4). Nice. □

With this discussed, what is next? More relativity I guess, because thinking well, there are still some gaps in what we did so far in this book, about it, coming from our poor knowledge of the inertial frames, which play a key role in the Einstein principles.

So, let us discuss this now, inertial frames. Their definition is as follows:

DEFINITION 10.17. *An inertial frame is a frame where all the basic formulae,*

$$||F|| = \frac{GM_1M_2}{||x_1 - x_2||^2} \quad , \quad F = Ma \quad , \quad a = \dot{v} \quad , \quad v = \dot{x} \quad , \quad F_{12} = -F_{21}$$

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

To be more precise, the first 4 formulae are something very familiar, that we have already heavily used, 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 as follows:

$$F_{21} = -F_{12}$$

Regarding this third law of Newton, you probably heard of it since high school, but thinking a bit, this law does not really hold in the context of the computations that we did in chapter 8, for the 2-body problem. Indeed, assuming as there that M_1 is fixed at 0, its acceleration is $\ddot{0} = 0$, and so the force acting on it is 0 too, which is certainly not the inverse of the force acting on M_2, which does exist, and drags M_2 on a conic.

All this might seem a bit perplexing, so let us formulate, as conclusion:

CONCLUSION 10.18. *Newton's third law is something quite tricky, and it's not about whether this law holds or not, but rather about if our frame is inertial or not.*

In short, you got the point, we are now into advanced physics, with no less than 5 useful formulae in our pocket, those in Definition 10.17. If these formulae mix well, we call the frame inertial, and this is very good. And if they don't, that is because the Newton third law is not satisfied in our frame, and that is good too, and we call our frame non-inertial, and we can of course still use it for computations.

But all this is perhaps a bit too abstract and mathematical. As a complement to this, here is an alternative definition for the inertial frames, that I personally prefer:

DEFINITION 10.19. *Your frame is inertial if your coffee does not spill.*

As examples of non-inertial frames, we have for instance elevators, rollercoaster rides, or you having a cat on your lap. As for inertial frames, you would probably say the Earth, but according to our previous results on the rotating bodies, if you fill your mug really up to the very top, it will spill, and so no good. But more such things later.

Getting back to mathematics now, and to our inertial frames as axiomatized in Definition 10.17, here is a list of useful basic facts regarding them:

THEOREM 10.20. *The following hold, regarding the inertial frames:*

(1) *In the context of the 2-body problem, the standard frames used for computations, with M_1 or M_2 fixed, are not inertial.*
(2) *In fact, the linear combination frames or type $\lambda_1 M_1 + \lambda_2 M_2$, including the center of mass frame, are all non-inertial.*

PROOF. These facts are all well-known, the idea being as follows:

(1) This was something already discussed above, but let us present now the full mathematical proof. We want to check whether the forces between M_1, M_2 satisfy:

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

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 general case now, with parameters λ_1, λ_2 satisfying $\lambda_1 + \lambda_2 = 0$, as in the statement, the positions of our bodies M_1, M_2 are:

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

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 must satisfy:

$$M_1\lambda_1 = M_2\lambda_2$$

But these are exactly the parameters of the center of mass of the system formed by M_1, M_2. 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 relation now with the angular momentum, we have the following result:

THEOREM 10.21. *The following hold, regarding the inertial frames:*

(1) *The total angular momentum of a system of bodies $M_1, \ldots, M_k$ is conserved, when computed with respect to an inertial frame.*
(2) *However, this is not the end of the story with angular momentum, because at $k = 2$, this is conserved as well in the frames of type $\lambda_1 M_1 + \lambda_2 M_2$.*

PROOF. As before with Theorem 10.20, all this is elementary, as follows:

(1) Our inertial frame assumption tells us that we can use at will all the formulae of Newton, and by using them, and notably using $F_{12} = -F_{21}$, we obtain:

$$\begin{aligned}
\dot{J} &= \sum_i x_i \times \sum_{j \neq i} F_{ji} \\
&= \sum_{i<j} x_i \times F_{ji} + x_j \times F_{ij} \\
&= \sum_{i<j} x_i \times F_{ji} - x_j \times F_{ji} \\
&= \sum_{i<j} (x_i - x_j) \times F_{ji} \\
&= 0
\end{aligned}$$

Now since we have $\dot{J} = 0$, the angular momentum J is conserved, as claimed.

(2) This follows from a similar computation, and more specifically from the computation from the proof of Theorem 10.11, which gives $\dot{J} = 0$, and so conservation too. □

Now, let us start hunting for inertial frames. A simple idea here is to start with a familiar problem, such as the 2-body one, written in some standard non-inertial frame that is well-adapted for computations, and then try to transform that non-inertial frame into an inertial one. And regarding now the transformations, we can use here:

PROPOSITION 10.22. *Any frame can be obtained from another frame, by performing the following operations, which are independent of each other:*

(1) *Changing the origin,* $x \to x - d$.
(2) *Rescaling the coordinates,* $x \to r \cdot x$.
(3) *Rotating the coordinates,* $x \to Ux$ *with* $U \in O_3$.

PROOF. This is something obvious, written perhaps in a somewhat fancy way:

(1) This amounts in moving the origin 0 at an arbitrary location $d \in \mathbb{R}^3$.

(2) This amounts in rescaling the coordinates by $r_1, r_2, r_3 \in \mathbb{R}$, respectively.

(3) This amounts in rotating the frame, O_3 being the group of 3D rotations.

But these three operations clearly allow passing from one frame to another, as claimed, and we are led to the conclusion in the statement. □

Let us examine now the 2-body problem. Here we know that the center of mass frame, obtained via a transformation (1), is the best one that we have, so the problem is whether we can make it inertial using (2). And here, we have the following result:

THEOREM 10.23. *In the standard context of the 2-body problem, by moving the origin at the center of mass, and then rescaling all three coordinates by $r = r(t) \in \mathbb{R}$ given by*

$$(rx)'' = -\frac{GM_1(M_1 + M_2)(rx)}{||(rx)^3||}$$

with this meaning that $y = rx$ must be the trajectory of a body of mass M_1 traveling around a body of mass $M_1 + M_2$, the resulting frame is inertial.

PROOF. We will be actually looking at all the inertial frames that can be obtained via the operations (1) and (2) in Proposition 10.22, and only at the end we will particularize to the frame discussed in the statement. Our study goes as follows:

(1) Assume that M_1 is fixed at 0, and that M_2 moves around it, with position vector $x \in \mathbb{R}^3$. By changing the origin at $d \in \mathbb{R}^3$, then rescaling the coordinates by $r \in \mathbb{R}^3$, as in Proposition 10.22 (1,2), the new coordinates of M_1, M_2 are as follows:

$$z_1 = -r \cdot d \quad , \quad , z_2 = r \cdot (x - d)$$

We want to check whether the forces between M_1, M_2 satisfy the following formula:

$$M_1\ddot{z}_1 = -M_2\ddot{z}_2 = \frac{GM_1M_2(z_2 - z_1)}{||z_2 - z_1||^3}$$

By using the above formulae of z_1, z_2, this formula to be checked reads:

$$-M_1(r \cdot d)'' = -M_2(r \cdot x)'' + M_2(r \cdot d)'' = \frac{GM_1M_2(r \cdot x)}{||r \cdot x||^3}$$

(2) This looks complicated, so let us look for a uniform solution, $r = (r, r, r)$. In this case the componentwise dot product is a usual product, and our equation becomes:

$$-M_1(rd)'' = -M_2(rx)'' + M_2(rd)'' = \frac{GM_1M_2x}{r^2||x||^3}$$

By using now the formula of gravity in the initial frame, this equation becomes:

$$-M_1(rd)'' = -M_2(rx)'' + M_2(rd)'' = -M_1 \cdot \frac{x''}{r^2}$$

But this formula can be written, more conveniently, as follows:

$$(rd)'' = \frac{M_2}{M_1 + M_2}(rx)'' = \frac{x''}{r^2}$$

Thus, we have our equations, which look good, and we can now solve for r on the right, and then solve for d on the left, as to get our inertial frame.

(3) Let us first solve for r on the right. It is convenient here to replace x'' by the Newtonian gravitation formula it came from, and our equation becomes:

$$\frac{M_2}{M_1 + M_2}(rx)'' = -\frac{GM_1M_2x}{r^2||x||^3}$$

But this equation can be written in the following more convenient way:

$$(rx)'' = -\frac{GM_1(M_1+M_2)(rx)}{||(rx)^3||}$$

We conclude that $y = rx$ must be the trajectory of a body of mass M_1 travelling around a body of mass $M_1 + M_2$, with arbitrary initial data y_0, w_0.

(4) Let us solve now for d on the left, in the equations found in (2) above. Here the situation is very simple, the solutions rd being as follows, with $a, b \in \mathbb{R}^3$:

$$rd = \frac{M_2}{M_1+M_2}\, rx + at + b$$

Thus, with r being as in (3) above, the solutions are as follows, with $a, b \in \mathbb{R}^3$:

$$d = \frac{M_2}{M_1+M_2}\, x + \frac{at+b}{r}$$

Thus with c being the center of mass, the formula is as follows, with $a, b \in \mathbb{R}^3$:

$$d = c + \frac{at+b}{r}$$

Now by taking $a = b = 0$, we are led to the conclusion in the statement. □

Quite nice all this, but getting now to the general, k-body case, things there look infinitely more complicated, for a myriad reasons. So, as a conclusion, let us formulate:

QUESTION 10.24. *Is there an inertial frame, somewhere in the universe? Can we prove its existence, mathematically? And if not, what about the Einstein principles?*

And we will end I guess our discussion with this. Not your average easy questions, what we have here, and with this, we are certainly into advanced physics. Mystery.

10d. Lagrange, Hamilton

As a last topic of discussion for this chapter, let us get now into energy matters. We know since chapter 4 that in 1D the potential energy V, which is unique up to a scalar, can be defined according to the formula $V' = -F$. In higher dimensions, we have:

THEOREM 10.25. *Assuming that an object of mass m moves under the influence of a force F which is conservative, in the sense that*

$$F = -\nabla V$$

for a certain function V, if we call this function V potential energy of m, and set

$$E = T + V$$

with $T = m||v||^2/2$ being as usual the kinetic energy of m, then E is constant.

PROOF. We have indeed the following computation, based on Leibnitz and the chain rule, with ∇V being as usual the vector formed by the spatial derivatives of V:

$$\begin{aligned}
\dot{T} &= m < \dot{v}, v > \\
&= m < a, v > \\
&= < F, v > \\
&= - < \nabla V, v > \\
&= -\sum_i (\nabla V)_i v_i \\
&= -\sum_i \frac{dV}{dx_i} \cdot \frac{dx_i}{dt} \\
&= -\frac{dV}{dt} \\
&= -\dot{V}
\end{aligned}$$

Thus we have $\dot{E} = \dot{T} + \dot{V} = 0$, so the energy E is constant, as claimed. □

As a main illustration now for Theorem 10.25, in relation with gravity, we have:

THEOREM 10.26. *The gravitation force F is conservative, $F = -\nabla V$, coming from*

$$V = -\frac{Km}{||x||}$$

and therefore the total energy $E = T + V$, with $T = m||v||^2/2$ as usual, is constant.

PROOF. According to the Newton principles, and with $K = GM$ as usual, the force applied to m by the object M positioned at 0 is given by:

$$F = -||F|| \cdot \frac{x}{||x||} = -G \cdot \frac{mM}{||x||^2} \cdot \frac{x}{||x||} = -\frac{Kmx}{||x||^3}$$

Now let us compute ∇V, formed by the 3 spatial derivatives of V. We have:

$$\begin{aligned}
\frac{dV}{dx_i} &= -\frac{dKm/\sqrt{x_1^2 + x_2^2 + x_3^2}}{dx_i} \\
&= -Km \cdot \frac{d(x_1^2 + x_2^2 + x_3^2)^{-1/2}}{dx_i} \\
&= -Km \cdot \left(-\frac{1}{2}\right) \cdot \frac{2x_i}{(x_1^2 + x_2^2 + x_3^2)^{3/2}} \\
&= \frac{Kmx_i}{||x||^3}
\end{aligned}$$

Thus we have $\nabla V = Kmx/||x||^3$, which by the above equals $-F$, as desired. □

In order to further clarify our concept of conservative force, we will need:

DEFINITION 10.27. *The work done by a force $F = F(x)$ for moving a particle from point $p \in \mathbb{R}^3$ to point $q \in \mathbb{R}^3$ via a given path $\gamma : p \to q$ is the following quantity:*

$$W(\gamma) = \int_\gamma < F(x), dx >$$

We say that F is conservative if this work quantity $W(\gamma)$ does not depend on the chosen path $\gamma : p \to q$, and in this case we denote this quantity by $W(p, q)$.

We will see in a moment that this definition is compatible with our previous definition for the conservative forces. Getting now to the mathematics of such forces, we have:

THEOREM 10.28. *The work done by a conservative force F on a mass m object is*

$$W(p, q) = T(q) - T(p)$$

with $T = m||v||^2/2$ standing as usual for the kinetic energy of the object.

PROOF. Assuming that F is conservative, we have the following computation:

$$\begin{aligned}
W(p, q) &= \int_p^q < F(x), dx > \\
&= m \int_p^q < a(x), dx > \\
&= m \int_p^q \left\langle \frac{dv(x)}{dt}, v(x)dt \right\rangle \\
&= \frac{m}{2} \int_p^q \frac{d < v(x), v(x) >}{dt}\, dt \\
&= \frac{m}{2} \int_p^q \frac{d||v(x)||^2}{dt}\, dt \\
&= \frac{m}{2} \left(||v(q)||^2 - ||v(p)||^2 \right) \\
&= T(q) - T(p)
\end{aligned}$$

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

Next, we have the following result, which uses some more advanced mathematics:

THEOREM 10.29. *A force F is conservative precisely when it is of the form*

$$F = -\nabla V$$

for a certain function V, and in this case the work done by it is given by:

$$W(p, q) = V(p) - V(q)$$

Also, the gravitation force is conservative, coming from $V = -Km/||x||$.

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

(1) In one sense, assume that F is conservative. Since the work $W(p,q) = W(\gamma)$ is independent of the chosen path $\gamma : p \to q$, we can find a function V such that:

$$W(p,q) = V(p) - V(q)$$

Observe that this function V is well-defined up to an additive constant. Now with this formula in hand, we further obtain, as desired:

$$\begin{aligned} W(p,q) = V(p) - V(q) &\implies < F, dx >= -dV \\ &\implies < F, x_i >= -\frac{dV}{dx_i} \\ &\implies F = -\nabla V \end{aligned}$$

(2) In the other sense now, assuming $F = -\nabla V$, we have the following computation, valid for any loop $\circ : p \to p$, which shows that F is indeed conservative:

$$W(\circ) = -\int_{\circ} \nabla V = 0$$

More generally, regarding the work done by such a force $F = -\nabla V$, along a path $\gamma : p \to q$, which is independent on this path γ, this is given by:

$$W(p,q) = -\int_{p}^{q} \nabla V = V(p) - V(q)$$

(3) Finally, the last assertion is something that we know from Theorem 10.26. □

We can put now everything together, and we have the following result:

THEOREM 10.30. *Given a conservative force F, appearing as follows, with V being uniquely determined up to an additive constant,*

$$F = -\nabla V$$

the movements of a particle under F preserve the total energy, given by

$$E = T + V$$

with $T = m||v||^2/2$ being the kinetic energy, and with V being called potential energy.

PROOF. This is something that we already know from Theorem 10.25, but we can now provide a new proof for this fact, based on Theorems 10.28 and 10.29, which give:

$$W(p,q) = T(q) - T(p)$$

$$W(p,q) = V(p) - V(q)$$

Indeed, we obtain from these equalities the following formula:

$$T(p) + V(p) = T(q) + V(q)$$

We conclude that the total energy $E = T + V$ is conserved, as claimed. □

With this discussed, what is next? Some applications I guess, which do not look obvious, and with our 3D expert the rat being not available, now that it's daytime, here I am outside in the garden, looking for rodents, under the bemused eye of the cats. And fortunately, my search comes to an end when I hear, up from the old cherry tree:

SQUIRREL 10.31. *You won't believe me, but our rodent work in mechanics, and life in general, is based on a very simple principle, that of the least action.*

Humm, quite strange all this, but I should perhaps take this advice seriously, time and again these wild animals have proved to be more intelligent than me, poor domesticated being. So, with such ideas in mind, let us begin with some mathematics. We have:

THEOREM 10.32. *Given a function $f : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$, the integral*

$$I = \int_{x_0}^{x_1} f(u, \dot{u}) dx$$

is stationary, in the sense that it is left unchanged by small variations of $u = u(x)$, which vanish at the endpoints x_0, x_1, precisely when $u = u(x)$ satisfies the equations

$$\frac{df}{du_i} = \frac{d}{dx}\left(\frac{df}{d\dot{u}_i}\right)$$

called Euler-Lagrange equations.

PROOF. Let us just work out the case $N = 1$, the general case being similar. Consider a small variation $\Delta u(x)$, which vanishes at the endpoints x_0, x_1, as required above:

$$\Delta u(x_0) = \Delta u(x_1) = 0$$

The corresponding variation of $f(u, \dot{u})$, at first order, is then given by:

$$\Delta f = \frac{df}{du}\,\Delta u + \frac{df}{d\dot{u}}\,\Delta\dot{u}$$

Thus the corresponding variation of the integral in the statement is given by:

$$\begin{aligned}
\Delta I &= \int_{x_0}^{x_1} \frac{df}{du}\,\Delta u\,dx + \int_{x_0}^{x_1} \frac{df}{d\dot{u}}\,\Delta\dot{u}\,dx \\
&= \int_{x_0}^{x_1} \frac{df}{du}\,\Delta u\,dx + \int_{x_0}^{x_1} \frac{df}{d\dot{u}} \cdot \frac{d(\Delta u)}{dx}\,dx \\
&= \int_{x_0}^{x_1} \frac{df}{du}\,\Delta u\,dx + \left[\frac{df}{d\dot{u}}\,\Delta u\right]_{x_0}^{x_1} - \int_{x_0}^{x_1} \frac{d}{dx}\left(\frac{df}{d\dot{u}}\right)\Delta u\,dx \\
&= \int_{x_0}^{x_1} \frac{df}{du}\,\Delta u\,dx - \int_{x_0}^{x_1} \frac{d}{dx}\left(\frac{df}{d\dot{u}}\right)\Delta u\,dx \\
&= \int_{x_0}^{x_1} \left(\frac{df}{du} - \frac{d}{dx}\left(\frac{df}{d\dot{u}}\right)\right)\Delta u\,dx
\end{aligned}$$

We conclude that I is stationary precisely when the following equation is satisfied:

$$\frac{df}{du} = \frac{d}{dx}\left(\frac{df}{d\dot{u}}\right)$$

But this is the Euler-Lagrange equation in the statement, as desired. □

Now back to mechanics, we have the following result, called Hamilton principle:

THEOREM 10.33. *In the context of a conservative force F acting, let us call*

$$L = T - V$$

Lagrangian of the system. The following integral, called action integral, is then stationary,

$$I = \int_{t_0}^{t_1} L\, dt$$

and the corresponding Euler-Lagrange equations are precisely the equations of motion.

PROOF. According to Theorem 10.30, the Lagrangian is given by the following formula, with V being the potential associated to our conservative force, via $F = -\nabla V$:

$$L = \frac{m}{2}(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) - V(x, y, z)$$

Thus, we are in the framework of Theorem 10.32, with the function u there being played by the coordinates x, y, z. Now let us pick one of these coordinates, $s = x, y, z$, and compute the derivatives of L with respect to $s, \dot{s}$. By using $F = -\nabla V$ we have:

$$\frac{dL}{ds} = -\frac{dV}{ds} = F_s$$

Also since the potential V is time-independent, we have:

$$\frac{dL}{d\dot{s}} = m\dot{s} = ma_s$$

Now consider the equation of motion, under the influence of the force F:

$$F = ma$$

This is a vector equation, with 3 components, and according to the above formulae its 3 components can be written as follows, in terms of the Lagrangian L:

$$\frac{dL}{ds} = \frac{d}{dt}\left(\frac{dL}{d\dot{s}}\right)$$

But these are precisely the Euler-Lagrange equations for the stationarity of the action integral $I = \int L$, and we are therefore led to the conclusions in the statement. □

The point now with the above result is that it leads right away into another result, which this time is something fundamental and powerful, as follows:

THEOREM 10.34. *The Euler-Lagrange equations for the action integral*

$$I = \int_{t_0}^{t_1} L\,dt$$

hold in any system of coordinates (q_1, q_2, q_3), and are as follows:

$$\frac{dL}{dq} = \frac{d}{dt}\left(\frac{dL}{d\dot{q}}\right)$$

These latter equations are called the Lagrange equations of motion.

PROOF. We know from Theorem 10.33 that the action integral is stationary, with respect to the standard coordinates (x, y, z). But this shows that the action integral is stationary with respect to any system of coordinates (q_1, q_2, q_3), and so the corresponding Euler-Lagrange equations, which are the equations in the statement, hold indeed. □

All this might seem a bit theoretical, but in practice, the applications abound. As a basic illustration here, the Newton solution to the Kepler problem, discussed in chapter 8, was using polar coordinates, or rather cylindrical coordinates, and that solution can be heavily simplified by starting directly with the Lagrange equations in cylindrical coordinates. And, we will leave the computations here as an instructive exercise.

Moving ahead now, let us discuss some further interesting manipulations on the Lagrangian and the Lagrange equations, due to Hamilton. Let us start with:

DEFINITION 10.35. *Given a system of coordinates $q_1, \ldots, q_n$, the quantities*

$$p_i = \frac{dL}{d\dot{q}_i}$$

appearing above are called generalized momenta. Also, the quantity

$$H(q, p) = \sum_i p_i \dot{q}_i - L$$

is called Hamiltonian of the system.

In practice, for many simple systems, H is in fact the total energy. However, the main interest in H is that it can successfully replace L, as a quantity which encapsulates as well what is going on, namely the equations of motion. The result here is as follows:

THEOREM 10.36. *The Hamiltonian $H = H(q, p)$ is subject to the equations*

$$\frac{dH}{dp_i} = \dot{q}_i \quad , \quad \frac{dH}{dq_i} = -\dot{p}_i$$

called Hamilton equations, which are equivalent to the usual equations of motion.

PROOF. By using the definition of the variables p_i, we have indeed:

$$\begin{aligned}\frac{dH}{dp_i} &= \frac{d\left(\sum_j p_j\dot{q}_j - L\right)}{dp_i}\\ &= \dot{q}_i + \sum_j p_j\,\frac{d\dot{q}_j}{dp_i} - \sum_j \frac{dL}{d\dot{q}_j}\cdot\frac{d\dot{q}_j}{dp_i}\\ &= \dot{q}_i + \sum_j p_j\,\frac{d\dot{q}_j}{dp_i} - \sum_j p_j\,\frac{d\dot{q}_j}{dp_i}\\ &= \dot{q}_i\end{aligned}$$

Also, by using the Lagrange equations, we obtain the following formula:

$$\begin{aligned}\frac{dH}{dq_i} &= \frac{d\left(\sum_j p_j\dot{q}_j - L\right)}{dq_i}\\ &= -\frac{dL}{dq_i} + \sum_j p_j\,\frac{d\dot{q}_j}{dq_i} - \sum_j \frac{dL}{d\dot{q}_j}\cdot\frac{d\dot{q}_j}{dq_i}\\ &= -\dot{p}_i + \sum_j p_j\,\frac{d\dot{q}_j}{dq_i} - \sum_j p_j\,\frac{d\dot{q}_j}{dq_i}\\ &= -\dot{p}_i\end{aligned}$$

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

10e. Exercises

Tough physics chapter we had here, good to have it done. As exercises, we have:

EXERCISE 10.37. *Learn more about 3D collisions, decay and scattering.*

EXERCISE 10.38. *Do the missing integration, for the hard sphere scattering.*

EXERCISE 10.39. *Learn about the Kepler 3 law, its proof and consequences.*

EXERCISE 10.40. *Work out some numerics, for the Coriolis force on Earth.*

EXERCISE 10.41. *Fully rewrite the theory of Einstein speed addition, in 3D.*

EXERCISE 10.42. *Think at inertial frames, in the context of the 3-body problem.*

EXERCISE 10.43. *Learn more about Lagrangian mechanics, and its applications.*

EXERCISE 10.44. *Learn more about Hamiltonian mechanics, and its applications.*

As bonus exercise, you are good to go, for reading a mechanics book. Enjoy.

CHAPTER 11

Multiple integrals

11a. The determinant

What is next? Many things I guess, 3 dimensions being certainly a vast business. Mathematically, we can build upon the material from chapter 9, which was just a beginning, with more geometry and topology, of all types, and their relation with algebra, or try something new, like 3D analysis or probability. While physically, we can keep building on the various types of mechanics developed in chapter 10, which were just a beginning too, or again, try something new, like electromagnetism or thermodynamics.

Well, it looks like we are a bit lost, in this 3D jungle. Perhaps more modestly, and remembering that we are here, with this book, for learning the basics, can we have at least something minimally coherent, going on? I mean, the mathematics in chapter 9 and physics in chapter 10, while each perfectly making sense, as 3D starters, on their own, were a bit orthogonal to each other, instead of being in harmony, as desirable. So, can we avoid this kind of issue in what follows, and develop some coherent 3D mathematics and physics, in harmony, in the remainder of the present Part III, that is my question.

And good question this is. In answer, as usual in such difficult situations, I will have to ask our 3D expert, the rat. And fortunately rat is here, we are now in the middle of the night, I can hear him as I type. With a bit of cheese helping, here is what I get:

RAT 11.1. *Contrary to smaller dimensions, there are many entry points to 3D, all leading to knowledge and understanding. Follow your own way, and Hare Krishna.*

Okay, thanks rat, guess you have a point here, let's relax a bit, about all this. So, forgetting about physics and applications, and looking at what we did so far matematics in this book, I would say that more linear algebra and vector calculus, in 3D, I mean a 3D mixture of algebra and analysis, coming as something new, and complementary to the geometry and topology that we did in chapter 9, would be a good thing.

So, let us get into this, and for physics and applications, we can see later. Let us start our study with the following fundamental result, which has countless concrete applications, and which is actually at the origin of the whole linear algebra theory:

THEOREM 11.2. *Any linear system of equations*

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1N}x_N &= v_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2N}x_N &= v_2 \\ \vdots \\ a_{N1}x_1 + a_{N2}x_2 + \ldots + a_{NN}x_N &= v_N \end{cases}$$

can be compactly written in matrix form, as follows,

$$Ax = v$$

and when A is invertible, its solution is given by $x = A^{-1}v$.

PROOF. Indeed, with standard linear algebra conventions, our system reads:

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1N} \\ a_{21} & a_{22} & \ldots & a_{2N} \\ \vdots & & & \vdots \\ a_{N1} & a_{N2} & \ldots & a_{NN} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{pmatrix}$$

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

In practice now, we are led in this way to the question of inverting the matrices $A \in M_N(\mathbb{R})$. And here, remember from chapter 7 the following key definition:

DEFINITION 11.3. *The determinant of square matrix $A \in M_N(\mathbb{R})$ is the signed volume of the parallelepiped made by its column vectors,*

$$A = \begin{bmatrix} v_1 & \ldots & v_N \end{bmatrix} \quad \implies \quad \det A = \pm vol < v_1, \ldots, v_N >$$

with the sign being $+$ when we can continuously pass from the standard basis $e_1, \ldots, e_N$ to the system of vectors $v_1, \ldots, v_N$, and being $-$ otherwise.

As a first observation, this is certainly related to Theorem 11.2, because we have the following equivalence, coming from our general vector space basics from chapter 9:

$$\exists A^{-1} \iff \det A \neq 0$$

Next, let us recall as well, also from chapter 7, that the determinant of the 2×2 matrices can be explicitly computed, via some geometry, and is given by:

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

And finally, we also saw in chapter 7 that the solution to the 2×2 matrix inversion problem, coming as a useful complement to Theorem 11.2, at $N = 2$, is as follows:

$$ad - bc \neq 0 \implies \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

In order to discuss now what happens in arbitrary dimensions, we will need a number of theoretical results. Here is a first series of formulae, coming from definitions:

THEOREM 11.4. *The determinant has the following properties:*

(1) *When multiplying by scalars, the determinant gets multiplied as well:*

$$\det(\lambda_1 v_1, \ldots, \lambda_N v_N) = \lambda_1 \ldots \lambda_N \det(v_1, \ldots, v_N)$$

(2) *When permuting two columns, the determinant changes the sign:*

$$\det(\ldots, u, \ldots, v, \ldots) = -\det(\ldots, v, \ldots, u, \ldots)$$

(3) *The determinant* $\det(e_1, \ldots, e_N)$ *of the standard basis of* $\mathbb{R}^N$ *is* 1.

PROOF. All this is clear indeed from definitions, with perhaps a bit of geometric thinking needed for (2), that we will leave here as an instructive exercise. □

As a first application of the above result, and with the standard convention that the determinant of a numeric matrix A is denoted as well $|A|$, we have:

THEOREM 11.5. *The determinant of a diagonal matrix is given by:*

$$\begin{vmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{vmatrix} = \lambda_1 \ldots \lambda_N$$

That is, we obtain the product of diagonal entries.

PROOF. The above formula is clear by using Theorem 11.4, which gives:

$$\begin{vmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{vmatrix} = \lambda_1 \ldots \lambda_N \begin{vmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{vmatrix} = \lambda_1 \ldots \lambda_N$$

As for the last assertion, this is rather a remark, and more on this later. □

In order to reach now to a more advanced theory, let us adopt the linear map point of view. In this setting, the definition of the determinant reformulates as follows:

THEOREM 11.6. *Given a linear map, written as* $f(v) = Av$*, its "inflation coefficient", obtained as the signed volume of the image of the unit cube, is given by:*

$$I_f = \det A$$

More generally, I_f *is the inflation ratio of any parallelepiped in* $\mathbb{R}^N$*, via the transformation* f*. In particular* f *is invertible precisely when* $\det A \neq 0$.

PROOF. The only non-trivial thing in all this is the fact that the inflation coefficient I_f, as defined above, is independent of the choice of the parallelepiped. But this is a generalization of the Thales theorem, which follows from the Thales theorem itself. □

As a first application of the above linear map viewpoint, we have:

THEOREM 11.7. *We have the following formula, valid for any matrices A, B:*

$$\det(AB) = \det A \cdot \det B$$

In particular, we have $\det(AB) = \det(BA)$.

PROOF. The first formula follows from the formula $f_{AB} = f_A f_B$ for the associated linear maps. As for $\det(AB) = \det(BA)$, this is clear from the first formula. □

In general now, still at the theoretical level, we have the following key result:

THEOREM 11.8. *The determinant has the additivity property*

$$\det(\ldots, u+v, \ldots) = \det(\ldots, u, \ldots) + \det(\ldots, v, \ldots)$$

valid for any choice of the vectors involved.

PROOF. This follows indeed by doing some elementary geometry, in the spirit of the 2×2 determinant computation that we did in chapter 7, by using the Thales theorem as main tool. We will leave the verifications here as an instructive exercise. □

As a basic application of the above result, we have:

THEOREM 11.9. *We have the following results:*

(1) *The determinant of a diagonal matrix is the product of diagonal entries.*
(2) *The same is true for the upper triangular matrices.*
(3) *The same is true for the lower triangular matrices.*

PROOF. Here (1) is something that we already know, and (2) comes as follows:

$$\begin{vmatrix} \lambda_1 & & * \\ & \lambda_2 & \\ & \ddots & \\ 0 & & \lambda_N \end{vmatrix} = \begin{vmatrix} \lambda_1 & 0 & * \\ & \lambda_2 & \\ & \ddots & \\ 0 & & \lambda_N \end{vmatrix}$$

$$\vdots$$

$$\vdots$$

$$= \begin{vmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ & \ddots & \\ 0 & & \lambda_N \end{vmatrix}$$

$$= \lambda_1 \ldots \lambda_N$$

As for (3), this comes in a similar way, by proceeding this time from right to left. □

As an important theoretical result now, allowing us to really take off, we have:

THEOREM 11.10. *The determinant* $\det : M_N(\mathbb{R}) \to \mathbb{R}$ *is the unique map satisfying:*

(1) $\det(\ldots, u+v, \ldots) = \det(\ldots, u, \ldots) + \det(\ldots, v, \ldots)$.
(2) $\det(\lambda_1 v_1, \ldots, \lambda_N v_N) = \lambda_1 \ldots \lambda_N \det(v_1, \ldots, v_N)$.
(3) $\det(\ldots, u, \ldots, v, \ldots) = -\det(\ldots, v, \ldots, u, \ldots)$.
(4) *The determinant* $\det(e_1, \ldots, e_N)$ *of the standard basis of* $\mathbb{R}^N$ *is* 1.

PROOF. The conditions in the statement are those from Theorems 11.4 and 11.8. As for the converse, it is routine to check that any map $\det' : M_N(\mathbb{R}) \to \mathbb{R}$ satisfying (1-4) must coincide with det on the upper triangular matrices, and then on all matrices. □

Here is now another important theoretical result, which is very useful in practice:

THEOREM 11.11. *The determinant is subject to the row expansion formula*

$$\begin{vmatrix} a_{11} & \ldots & a_{1N} \\ \vdots & & \vdots \\ a_{N1} & \ldots & a_{NN} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & \ldots & a_{2N} \\ \vdots & & \vdots \\ a_{N2} & \ldots & a_{NN} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} & \ldots & a_{2N} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N3} & \ldots & a_{NN} \end{vmatrix}$$
$$+ \ldots\ldots + (-1)^{N+1} a_{1N} \begin{vmatrix} a_{21} & \ldots & a_{2,N-1} \\ \vdots & & \vdots \\ a_{N1} & \ldots & a_{N,N-1} \end{vmatrix}$$

and this method fully computes it, by recurrence.

PROOF. This follows indeed from the fact that the formula in the statement produces a certain function $\det : M_N(\mathbb{R}) \to \mathbb{R}$, which has the 4 properties in Theorem 11.10. □

We can expand as well over the columns, with the result here being as follows:

THEOREM 11.12. *The determinant is subject to the column expansion formula*

$$\begin{vmatrix} a_{11} & \ldots & a_{1N} \\ \vdots & & \vdots \\ a_{N1} & \ldots & a_{NN} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & \ldots & a_{2N} \\ \vdots & & \vdots \\ a_{N2} & \ldots & a_{NN} \end{vmatrix} - a_{21} \begin{vmatrix} a_{12} & \ldots & a_{1N} \\ a_{32} & \ldots & a_{3N} \\ \vdots & & \vdots \\ a_{N2} & \ldots & a_{NN} \end{vmatrix}$$
$$+ \ldots\ldots + (-1)^{N+1} a_{N1} \begin{vmatrix} a_{12} & \ldots & a_{1N} \\ \vdots & & \vdots \\ a_{N-1,2} & \ldots & a_{N-1,N} \end{vmatrix}$$

and this method fully computes it, by recurrence.

PROOF. This follows indeed by using the same argument as for the rows. □

As a first application of the above methods, we can now prove:

THEOREM 11.13. *The determinant of the* 3×3 *matrices is given by*

$$\begin{vmatrix} a & b & c\\ d & e & f\\ g & h & i\end{vmatrix} = aei+bfg+cdh-ceg-bdi-afh$$

which can be memorized by using Sarrus' triangle method, "triangles parallel to the diagonal, minus triangles parallel to the antidiagonal".

PROOF. Here is the computation, using the above results:

$$\begin{aligned}\begin{vmatrix} a & b & c\\ d & e & f\\ g & h & i\end{vmatrix} &= a\begin{vmatrix} e & f\\ h & i\end{vmatrix} - b\begin{vmatrix} d & f\\ g & i\end{vmatrix} + c\begin{vmatrix} d & e\\ g & h\end{vmatrix}\\ &= a(ei-fh)-b(di-fg)+c(dh-eg)\\ &= aei-afh-bdi+bfg+cdh-ceg\\ &= aei+bfg+cdh-ceg-bdi-afh\end{aligned}$$

Thus, we obtain the formula in the statement. □

As a first true application of all this, and making the link with some previous tough computations from chapter 9, we can now invert the 3×3 matrices, as follows:

THEOREM 11.14. *The inverses of the* 3×3 *matrices are given by*

$$\begin{pmatrix} a & b & c\\ d & e & f\\ g & h & i\end{pmatrix}^{-1} = \frac{1}{D}\begin{pmatrix} ei-fh & ch-bi & bf-ce\\ fg-di & ai-cg & cd-af\\ dh-eg & bg-ah & ae-bd\end{pmatrix}$$

with D *being the determimant. When* $D=0$*, the matrix is not invertible.*

PROOF. As before for the 2×2 matrices, in order for our matrix to be invertible, we must have $D\neq 0$. The trick now is to look for solutions of the following problem:

$$\begin{pmatrix} a & b & c\\ d & e & f\\ g & h & i\end{pmatrix}\begin{pmatrix} * & * & *\\ * & * & *\\ * & * & *\end{pmatrix} = \begin{pmatrix} D & 0 & 0\\ 0 & D & 0\\ 0 & 0 & D\end{pmatrix}$$

We know from Theorem 11.13 that the determinant is given by:

$$D = aei+bfg+cdh-ceg-bdi-afh$$

But this leads, via some obvious choices, to the following solution:

$$\begin{pmatrix} * & * & *\\ * & * & *\\ * & * & *\end{pmatrix} = \begin{pmatrix} ei-fh & ch-bi & bf-ce\\ fg-di & ai-cg & cd-af\\ dh-eg & bg-ah & ae-bd\end{pmatrix}$$

Thus, by rescaling, we obtain the formula in the statement. □

In fact, we can now fully solve the inversion problem, as follows:

THEOREM 11.15. *The inverse of a square matrix having nonzero determinant is*

$$A^{-1}=\frac{1}{\det A}\begin{pmatrix}\det A^{(11)} & -\det A^{(21)} & \det A^{(31)} & \ldots\\ -\det A^{(12)} & \det A^{(22)} & -\det A^{(32)} & \ldots\\ \det A^{(13)} & -\det A^{(23)} & \det A^{(33)} & \ldots\\ \vdots & \vdots & \vdots & \end{pmatrix}$$

where $A^{(ij)}$ is the matrix A, with the i-th row and j-th column removed.

PROOF. This follows indeed by using the row expansion formula from Theorem 11.11, which in terms of the matrix A^{-1} in the statement reads $AA^{-1}=1$. □

Let us discuss now the general formula of the determinant, at arbitrary values $N\in\mathbb{N}$ of the matrix size, generalizing the formulae that we have at $N=2,3$. For this purpose, we will use the symmetric group S_N, that we already met in chapter 9. We have:

THEOREM 11.16. *The permutations have a signature function*

$$\varepsilon:S_N\to\{\pm1\}$$

which can be defined in the following equivalent ways:

(1) *As $(-1)^c$, where c is the number of inversions.*
(2) *As $(-1)^t$, where t is the number of transpositions.*
(3) *As $(-1)^o$, where o is the number of odd cycles.*
(4) *As $(-1)^x$, where x is the number of crossings.*
(5) *As the orientation of the corresponding permuted basis of $\mathbb{R}^N$.*

PROOF. Let us begin with the precise definition of c,t,o,x, as numbers modulo 2:

(1) The idea here is that given any two numbers $i<j$ among $1,\ldots,N$, the permutation can either keep them in the same order, $\sigma(i)<\sigma(j)$, or invert them:

$$\sigma(j)>\sigma(i)$$

Now by making $i<j$ vary over all pairs of numbers in $1,\ldots,N$, we can count the number of inversions, and call it c. This is an integer, $c\in\mathbb{N}$, which is well-defined.

(2) Here the idea, which is something quite intuitive, is that any permutation appears as a product of switches, also called transpositions:

$$i\leftrightarrow j$$

The decomposition as a product of transpositions is not unique, but the number t of the needed transpositions is unique, when considered modulo 2. This follows for instance from the equivalence of (2) with (1,3,4,5), explained below.

(3) Here the point is that any permutation decomposes, in a unique way, as a product of cycles, which are by definition permutations of the following type:

$$i_1 \to i_2 \to i_3 \to \ldots\ldots \to i_k \to i_1$$

Some of these cycles have even length, and some others have odd length. By counting those having odd length, we obtain a well-defined number $o \in \mathbb{N}$.

(4) Here the method is that of drawing the permutation, as we usually do, and by avoiding triple crossings, and then counting the number of crossings. This number x depends on the way we draw the permutations, but modulo 2, we always get the same number. Indeed, this follows from the fact that we can continuously pass from a drawing to each other, and that when doing so, the number of crossings can only jump by ± 2.

Summarizing, we have 4 different definitions for the signature of the permutations, which all make sense, constructed according to (1-4) above. Regarding now the fact that we always obtain the same number, this can be established as follows:

(1)=(2) This is clear, because any transposition inverts once, modulo 2.

(1)=(3) This is clear as well, because the odd cycles invert once, modulo 2.

(1)=(4) This comes from the fact that the crossings correspond to inversions.

(2)=(3) This follows by decomposing the cycles into transpositions.

(2)=(4) This comes from the fact that the crossings correspond to transpositions.

(3)=(4) This follows by drawing a product of cycles, and counting the crossings.

Finally, in what regards the equivalence of all these constructions with (5), here simplest is to use (2). Indeed, we already know that the sign of a system of vectors switches when interchanging two vectors, and so the equivalence between (2,5) is clear. □

Now back to linear algebra, we can formulate a key result, as follows:

THEOREM 11.17. *We have the following formula for the determinant,*

$$\det A = \sum_{\sigma \in S_N} \varepsilon(\sigma) A_{1\sigma(1)} \ldots A_{N\sigma(N)}$$

with the signature function being the one introduced above.

PROOF. This follows by recurrence over $N \in \mathbb{N}$, as follows:

(1) When developing the determinant over the first column, we obtain a signed sum of N determinants of size $(N-1) \times (N-1)$. But each of these determinants can be computed by developing over the first column too, and so on, and we are led to the conclusion that we have a formula as in the statement, with $\varepsilon(\sigma) \in \{-1, 1\}$ being certain coefficients.

(2) But these latter coefficients $\varepsilon(\sigma) \in \{-1, 1\}$ can only be the signatures of the corresponding permutations $\sigma \in S_N$, with this being something that can be viewed again by recurrence, with either of the definitions (1-5) in Theorem 11.16 for the signature. □

The above result is something quite tricky, and as a first illustration, in 2 dimensions we recover the usual formula of the determinant, from chapter 7, namely:

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

In 3 dimensions now, we recover the Sarrus formula from Theorem 11.13:

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - ceg - bdi - afh$$

Let us record as well the following general formula, which was not obvious with our previous theory of the determinant, but which follows from Theorem 11.17:

$$\det(A) = \det(A^t)$$

Importantly, the formula that we found in Theorem 11.17 allows us to deal with the complex matrices too, by formulating here things as follows:

THEOREM 11.18. *If we define the determinant of complex matrices $A \in M_N(\mathbb{C})$ as*

$$\det A = \sum_{\sigma \in S_N} \varepsilon(\sigma) A_{1\sigma(1)} \ldots A_{N\sigma(N)}$$

then this determinant has the same properties as the determinant of the real matrices.

PROOF. This follows indeed by doing some sort of reverse engineering, with respect to what we did above, and we reach to the conclusion that det has indeed all the algebraic properties that we are familiar with. We will leave this as an instructive exercise. □

And with this, good news, end of the general theory that we wanted to develop. We have now in our bag all the needed techniques for computing the determinant.

11b. Change of variables

Getting now to vector calculus, we already have some solid knowledge in what regards the derivatives, viewed as matrices, from chapter 7. So, as a complement to that material, let us discuss now integration. The key result here is the change of variable formula, and in order to discuss this, let us start with something that we know well, in 1D:

PROPOSITION 11.19. *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 quantity $\det(\varphi'(t))$, called Jacobian, and with this in mind, we are led to:

THEOREM 11.20. *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 11.19.

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

(1) Observe first that the above formula generalizes indeed the change of variable formula in 1 dimension, the point here being that the absolute value on the derivative appears as to compensate for the lack of explicit bounds for the integral.

(2) As a second observation, we can assume if we want, by linearity, that we are dealing with the constant function $f = 1$. For this function, our formula reads:

$$vol(E) = \int_{\varphi^{-1}(E)} |J_\varphi(t)|dt$$

In terms of $D = \varphi^{-1}(E)$, this amounts in proving that we have:

$$vol(\varphi(D)) = \int_D |J_\varphi(t)|dt$$

And here, as a first remark, our formula is clear for the linear maps φ, by using the definition of the determinant of real matrices, as a signed volume.

(3) However, the extension of this to the case of non-linear maps φ is something non-trivial, so we will not follow this path. In order to prove now the result, as stated, our first claim is that the validity of the theorem is stable under taking compositions of transformations φ. In order to prove this claim, consider a composition, as follows:

$$\varphi : E \to F \quad , \quad \psi : D \to E \quad , \quad \varphi \circ \psi : D \to F$$

Assuming that the theorem holds for φ, ψ, we deduce that we have, as desired:

$$\begin{aligned}\int_F f(x)dx &= \int_E f(\varphi(s))|J_\varphi(s)|ds\\ &= \int_D f(\varphi\circ\psi(t))|J_\varphi(\psi(t))|\cdot|J_\psi(t)|dt\\ &= \int_D f(\varphi\circ\psi(t))|J_{\varphi\circ\psi}(t)|dt\end{aligned}$$

(4) Next, as a key ingredient, let us examine the case where we are in $N = 2$ dimensions, and our transformation φ has one of the following special forms:

$$\varphi(x, y) = (\psi(x, y), y) \quad , \quad \varphi(x, y) = (x, \psi(x, y))$$

By symmetry, it is enough to deal with the first case. Here the Jacobian is $d\psi/dx$, and by replacing if needed $\psi \to -\psi$, we can assume that this Jacobian is positive, $d\psi/dx > 0$. Now by assuming as before that $D = \varphi^{-1}(E)$ is a rectangle, $D = [a, b] \times [c, d]$, we can prove our formula by using the change of variables in 1 dimension, as follows:

$$\begin{aligned}\int_E f(s)ds &= \int_{\varphi(D)} f(x, y)dxdy\\ &= \int_c^d \int_{\psi(a,y)}^{\psi(b,y)} f(x, y)dxdy\\ &= \int_c^d \int_a^b f(\psi(x, y), y)\frac{d\psi}{dx}\,dxdy\\ &= \int_D f(\varphi(t))J_\varphi(t)dt\end{aligned}$$

(5) But with this, we can now prove the theorem, in $N = 2$ dimensions. Indeed, given a transformation $\varphi = (\varphi_1, \varphi_2)$, consider the following two transformations:

$$\phi(x, y) = (\varphi_1(x, y), y) \quad , \quad \psi(x, y) = (x, \varphi_2 \circ \phi^{-1}(x, y))$$

We have then $\varphi = \psi \circ \phi$, and by using (4) for ψ, ϕ, which are of the special form there, and then (3) for composing, we conclude that the theorem holds indeed for φ, as desired. Thus, theorem proved in $N = 2$ dimensions, and the extension of the above proof to arbitrary N dimensions is straightforward, that we will leave here as an exercise. □

In order to discuss now some basic applications, in 2D, we will need:

PROPOSITION 11.21. *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 something elementary, the Jacobian being computed as follows:

$$J=\begin{vmatrix}\frac{d(r\cos t)}{dr} & \frac{d(r\cos t)}{dt}\\ \frac{d(r\sin t)}{dr} & \frac{d(r\sin t)}{dt}\end{vmatrix}=\begin{vmatrix}\cos t & -r\sin t\\ \sin t & r\cos t\end{vmatrix}=r$$

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

We can now compute the Gauss integral, which is the best calculus formula ever:

THEOREM 11.22. *We have the following formula,*

$$\int_{\mathbb{R}}e^{-x^2}dx=\sqrt{\pi}$$

called Gauss integral formula.

PROOF. This is something truly magic, the idea being as follows:

(1) To start with, we can certainly integrate e^{-x^2} by using the formula of the exponential series, and the primitive which is worth 0 at $x=0$ is given by:

$$\int e^{-x^2}=\sum_{k=0}^{\infty}(-1)^k\frac{x^{2k+1}}{(2k+1)k!}$$

However, this series is not computable, in terms of the known, familiar series.

(2) Thus, no primitive, but we can still ask for the computation of $\int_{\mathbb{R}}e^{-x^2}dx$, who knows. And here, another surprise awaits us, this is simply undoable, with bare hands, I mean all the formulae and tricks that we learned in chapter 3 fail, for this integral.

(3) Which seems to send our problem to the trash can. However, and here comes the magic, the Gauss integral can be computed by using two dimensions, as follows:

$$\begin{aligned}\left(\int_{\mathbb{R}}e^{-x^2}dx\right)^2&=\int_{\mathbb{R}}\int_{\mathbb{R}}e^{-x^2-y^2}dxdy\\ &=\int_0^{2\pi}\int_0^{\infty}e^{-r^2}rdrdt\\ &=2\pi\int_0^{\infty}\left(-\frac{e^{-r^2}}{2}\right)'dr\\ &=2\pi\left[0-\left(-\frac{1}{2}\right)\right]\\ &=\pi\end{aligned}$$

(4) Amazing all this, isn't it. There are of course some other known proofs of the Gauss formula, as for instance the one that we met in chapter 4, when first discussing the Gauss integral, but all using two dimensions, appearing as variations of the above. □

Quite nice all this, looks like with Theorem 11.20 complemented by polar coordinates, we have our nuclear weapon in analysis. Moving now to 3 dimensions, we have here:

PROPOSITION 11.23. *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 as well the general spherical coordinate formula, in arbitrary N dimensions. The formula here, which generalizes those at $N = 2, 3$, is as follows:

THEOREM 11.24. *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.

11c. Spherical integrals

As an application, let us compute now 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 11.25. *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 11.26. *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:

$$\begin{aligned} V^+ &= \int_{B^+} 1 \\ &= \int_0^1 \int_0^{\pi/2} \ldots \int_0^{\pi/2} r^{N-1} \sin^{N-2} t_1 \ldots \sin t_{N-2}\, dr dt_1 \ldots dt_{N-1} \\ &= \int_0^1 r^{N-1}\, dr \int_0^{\pi/2} \sin^{N-2} t_1\, dt_1 \ldots \int_0^{\pi/2} \sin t_{N-2} dt_{N-2} \int_0^{\pi/2} 1 dt_{N-1} \\ &= \frac{1}{N} \times \left(\frac{\pi}{2}\right)^{[N/2]} \times \frac{(N-2)!!}{(N-1)!!}\cdot\frac{(N-3)!!}{(N-2)!!}\cdots\frac{2!!}{3!!}\cdot\frac{1!!}{2!!}\cdot 1 \\ &= \frac{1}{N} \times \left(\frac{\pi}{2}\right)^{[N/2]} \times \frac{1}{(N-1)!!} \\ &= \left(\frac{\pi}{2}\right)^{[N/2]} \frac{1}{(N+1)!!} \end{aligned}$$

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

The formula in Theorem 11.26 is certainly nice, but in practice, we would like to have estimates for that sphere volumes too. For this purpose, we will need:

THEOREM 11.27. *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:

$$\log(N!) = \sum_{k=1}^{N} \log k \approx \int_1^N \log x \, dx = N \log N - N + 1$$

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 by trapezoids. In practice, this gives the following estimate:

$$\log(N!) \approx \int_1^N \log x \, dx + \frac{\log 1 + \log N}{2} = N \log N - N + 1 + \frac{\log N}{2}$$

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 proved the Stirling formula, as formulated in the statement. □

We can now estimate the volumes of the spheres, as follows:

THEOREM 11.28. *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 11.26, 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 11.26, we can compute in the same way the area of the sphere, the result being as follows:

THEOREM 11.29. *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 $N!!=(N-1)(N-3)(N-5)\ldots$ as usual, and we have the estimate

$$A\simeq\left(\frac{2\pi e}{N}\right)^{N/2}\sqrt{\frac{N}{\pi}}$$

in the $N\to\infty$ limit.

PROOF. There is actually no need to compute again, because the "pizza" argument from chapter 5 shows that the area and volume of the sphere in $\mathbb{R}^N$ are related by:

$$A=N\cdot V$$

Thus, we obtain the results, coming from Theorems 11.26 and 11.28. Of course, it is possible to do this as well directly, by suitably modifying the computations in the proofs of Theorems 11.26 and 11.28, and we will leave this as an instructive exercise. □

With this discussed, done with spherical integrals? You must be kidding. As a continuation of Proposition 11.25, and in relation with the comments there, we have:

PROPOSITION 11.30. *We have the following formula,*

$$\int_0^{\pi/2}\cos^p t\sin^q t\,dt=\left(\frac{\pi}{2}\right)^{\varepsilon(p)\varepsilon(q)}\frac{p!!q!!}{(p+q+1)!!}$$

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. We use the same idea as in Proposition 11.25. Let I_{pq} be the integral in the statement. In order to do the partial integration, observe that we have:

$$\begin{aligned}(\cos^p t\sin^q t)'&=p\cos^{p-1}t(-\sin t)\sin^q t\\&+\cos^p t\cdot q\sin^{q-1}t\cos t\\&=-p\cos^{p-1}t\sin^{q+1}t+q\cos^{p+1}t\sin^{q-1}t\end{aligned}$$

By integrating between 0 and $\pi/2$, we obtain, for $p,q>0$:

$$pI_{p-1,q+1}=qI_{p+1,q-1}$$

Thus we can compute I_{pq} by recurrence, with input coming from Proposition 11.25, good exercise for you here, and we are led to the formula in the statement. □

Good news, we can now integrate in full generality over the spheres, as follows:

THEOREM 11.31. *The polynomial integrals over the unit sphere* $S_{\mathbb{R}}^{N-1}\subset\mathbb{R}^N$*, with respect to the normalized, mass* 1 *measure, are given by the following formula,*

$$\int_{S_{\mathbb{R}}^{N-1}}x_1^{k_1}\ldots x_N^{k_N}\,dx=\frac{(N-1)!!k_1!!\ldots k_N!!}{(N+\Sigma k_i-1)!!}$$

valid when all exponents k_i *are even. If an exponent* k_i *is odd, the integral vanishes.*

PROOF. Assume first that one of the exponents k_i is odd. We can make then the following change of variables, which shows that the integral in the statement vanishes:

$$x_i\to -x_i$$

Assume now that all exponents k_i are even. As a first observation, the result holds indeed at $N=2$, due to the formula from Proposition 11.30, which reads:

$$\int_0^{\pi/2}\cos^p t\sin^q t\,dt=\left(\frac{\pi}{2}\right)^{\varepsilon(p)\varepsilon(q)}\frac{p!!q!!}{(p+q+1)!!}=\frac{p!!q!!}{(p+q+1)!!}$$

In the general case now, where the dimension $N \in \mathbb{N}$ is arbitrary, the integral in the statement can be written in spherical coordinates, as follows:

$$I = \frac{2^N}{A} \int_0^{\pi/2} \ldots \int_0^{\pi/2} x_1^{k_1} \ldots x_N^{k_N} \, J \, dt_1 \ldots dt_{N-1}$$

To be more precise, here A is the area of the sphere, J is the Jacobian, and the 2^N factor comes from the restriction to the $1/2^N$ part of the sphere where all the coordinates are positive. According to Theorem 11.29, the normalization constant is given by:

$$\frac{2^N}{A} = \left(\frac{2}{\pi}\right)^{[N/2]} (N-1)!!$$

As for the unnormalized integral, this is given by the following formula:

$$\begin{aligned} I' = \int_0^{\pi/2} \ldots \int_0^{\pi/2} \quad & (\cos t_1)^{k_1} (\sin t_1 \cos t_2)^{k_2} \\ & \vdots \\ & (\sin t_1 \sin t_2 \ldots \sin t_{N-2} \cos t_{N-1})^{k_{N-1}} \\ & (\sin t_1 \sin t_2 \ldots \sin t_{N-2} \sin t_{N-1})^{k_N} \\ & \sin^{N-2} t_1 \sin^{N-3} t_2 \ldots \sin^2 t_{N-3} \sin t_{N-2} \\ & dt_1 \ldots dt_{N-1} \end{aligned}$$

Now by rearranging the terms, and integrating with respect to each variable by using Proposition 11.30, we are led to the formula in the statement. □

And with this, end of our study, good work that we did, and good final formula that we have, in Theorem 11.31, that can be used for all sorts of practical purposes.

11d. Hyperspherical laws

As an application of what we learned, we can do some interesting probability. Going a bit fast here, not much time left in this chapter, as a starting point, we have:

DEFINITION 11.32. *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 $M_k(f) = \int_{\mathbb{R}} x^k d\mu_f(x)$.*

To be more precise 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. Next, we have:

DEFINITION 11.33. *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 quick 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 11.34. *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 11.35. *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. □

Getting now to more concrete things, we can introduce the normal laws, as follows:

DEFINITION 11.36. *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 $x = \sqrt{2t}\, y$:

$$\int_{\mathbb{R}} e^{-x^2/2t} dx = \sqrt{2t} \int_{\mathbb{R}} e^{-y^2} dy = \sqrt{2\pi t}$$

Generally speaking, the normal laws appear as bit everywhere, in real life. The reasons behind this phenomenon come from the Central Limit Theorem (CLT), namely:

THEOREM 11.37 (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 normal law of parameter* t*.*

PROOF. In terms of moments, the Fourier transform $F_f(x) = E(e^{ixf})$ is given by:

$$F_f(x) = E\left(\sum_{k=0}^{\infty} \frac{(ixf)^k}{k!} \right) = \sum_{k=0}^{\infty} \frac{i^k M_k(f)}{k!}\, x^k$$

Thus, the Fourier transform of the variable in the statement is:

$$\begin{aligned} F(x) &= \left[F_f\left(\frac{x}{\sqrt{n}}\right)\right]^n \\ &= \left[1-\frac{tx^2}{2n}+O(n^{-2})\right]^n \\ &\simeq e^{-tx^2/2} \end{aligned}$$

On the other hand, the Fourier transform of g_t is given by:

$$\begin{aligned} F_{g_t}(x) &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} e^{-y^2/2t+ixy}dy \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} e^{-(y/\sqrt{2t}-\sqrt{t/2}ix)^2-tx^2/2}dy \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} e^{-z^2-tx^2/2}\sqrt{2t}dz \\ &= \frac{1}{\sqrt{\pi}}e^{-tx^2/2}\int_{\mathbb{R}} e^{-z^2}dz \\ &= e^{-tx^2/2} \end{aligned}$$

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

Finally, let us record the following useful property of the normal laws:

THEOREM 11.38. *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 $k\in\mathbb{N}$:

$$\begin{aligned} M_k &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} y^k e^{-y^2/2t}dy \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} (ty^{k-1})\left(-e^{-y^2/2t}\right)'dy \\ &= \frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} t(k-1)y^{k-2}e^{-y^2/2t}dy \\ &= t(k-1)\times\frac{1}{\sqrt{2\pi t}}\int_{\mathbb{R}} y^{k-2}e^{-y^2/2t}dy \\ &= t(k-1)M_{k-2} \end{aligned}$$

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

In the relation now with spheres and geometry, we have the following key result:

THEOREM 11.39. *The moments of the hyperspherical variables are*

$$\int_{S_{\mathbb{R}}^{N-1}} x_i^p dx = \frac{(N-1)!!p!!}{(N+p-1)!!}$$

and the rescaled variables $y_i = \sqrt{N}x_i$ become normal and independent with $N \to \infty$.

PROOF. According to the integration formula in Theorem 11.31, we have:

$$\int_{S_{\mathbb{R}}^{N-1}} x_i^p dx \simeq N^{-p/2} \times p!! = N^{-p/2} M_p(g_1)$$

Thus, the rescaled variables $\sqrt{N}x_i$ become normal with $N \to \infty$, as claimed. As for the proof of the asymptotic independence, this is standard too, coming from:

$$\begin{aligned} \int_{S_{\mathbb{R}}^{N-1}} x_1^{k_1} \ldots x_N^{k_N}\, dx &= \frac{(N-1)!!k_1!! \ldots k_N!!}{(N+\Sigma k_i-1)!!} \\ &\simeq N^{-\Sigma k_i} \times k_1!! \ldots k_N!! \end{aligned}$$

By rescaling, the joint moments of the variables $y_i = \sqrt{N}x_i$ are given by:

$$\int_{S_{\mathbb{R}}^{N-1}} y_1^{k_1} \ldots y_N^{k_N}\, dx \simeq k_1!! \ldots k_N!!$$

Thus, we have multiplicativity, and so independence with $N \to \infty$, as claimed. □

11e. Exercises

Good analytic learning this was, in this chapter, and as exercises, we have:

EXERCISE 11.40. *Clarify the geometric details, in our theory of the determinant.*

EXERCISE 11.41. *Clarify the details, in the proof of the change of variable formula.*

EXERCISE 11.42. *Try proving the Gauss formula without using 2 dimensions.*

EXERCISE 11.43. *Figure out the ranges of angles, in our spherical coordinate formulae.*

EXERCISE 11.44. *Learn more about the Stirling formula, and its various proofs.*

EXERCISE 11.45. *Learn about discrete probability, binomial and Poisson laws.*

EXERCISE 11.46. *Learn more about normal variables, real, complex and vectorial.*

EXERCISE 11.47. *Explicitly compute the hyperspherical laws in small dimensions.*

As bonus exercise, you are good to go, for reading a probability book. Enjoy.

CHAPTER 12

Charges, matter

12a. Charges, Gauss

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, $q < 0$.*

As our first result now on this, due to Coulomb, and that will come as a physics fact instead of a mathematical theorem, because, well, I must admit that what we have in Fact 12.1 is 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 soon discover, passed a few simple facts, things will be far more complicated here, the point being that moving charges are something totally crazy. More on this later.

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:

$$\begin{aligned} F &= sgn(q_1q_2)\cdot||F||\cdot\frac{x_2-x_1}{||x_2-x_1||}\\ &= sgn(q_1q_2)\cdot K\cdot\frac{|q_1q_2|}{||x_2-x_1||^2}\cdot\frac{x_2-x_1}{||x_2-x_1||}\\ &= K\cdot\frac{q_1q_2(x_2-x_1)}{||x_2-x_1||^3}\end{aligned}$$

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

In relation with the value of the constant K appearing above, we have:

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 clarified 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 indeed comes from the following alternative definition of the coulomb, in terms of the ampere, which is something more complicated, to be discussed later:

$$1C = 1A \cdot 1s$$

Quite fun, all this, as always with units and constants. Let us end this discussion 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. □

Let us develop now some basic mathematics for electrostatics. We first have:

DEFINITION 12.6. *Given charges $q_1, \ldots, q_k \in \mathbb{R}$ located at positions $x_1, \ldots, x_k \in \mathbb{R}^3$, we define their electric field to be the vector function*

$$E(x) = K \sum_i \frac{q_i(x - x_i)}{||x - x_i||^3}$$

so that their force applied to a charge $Q \in \mathbb{R}$ positioned at $x \in \mathbb{R}^3$ is given by $F = QE$.

More generally, we will be interested in the electric fields of various non-discrete configurations of charges, such as charged curves, surfaces and solid bodies:

DEFINITION 12.7. *The electric field of a charge configuration $L \subset \mathbb{R}^3$, with charge density function $\rho : L \to \mathbb{R}$, is the vector function*

$$E(x) = K \int_L \frac{\rho(z)(x-z)}{||x-z||^3}\, dz$$

so that the force of L applied to a charge Q positioned at x is given by $F = QE$.

It is convenient now to forget about 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:

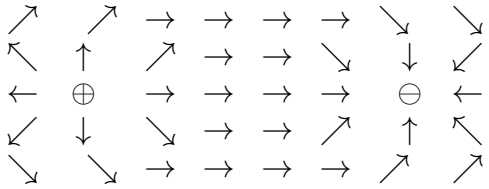
DEFINITION 12.8. *The field lines of $E : \mathbb{R}^3 \to \mathbb{R}^3$ are the oriented curves*

$$\gamma \subset \mathbb{R}^3$$

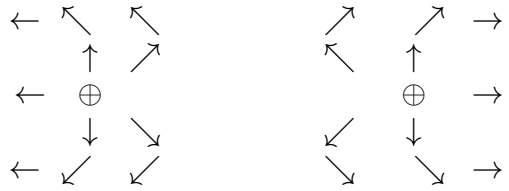
pointing at every point $x \in \mathbb{R}^3$ at the direction of the field, $E(x) \in \mathbb{R}^3$.

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 the geometry here to be far more complicated than that of gravity.

Let us introduce now a key definition, for the understanding of this, as follows:

DEFINITION 12.9. *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.

To be more precise, here by orientation of S we mean a choice of unit vectors $n(x)$ orthogonal to S, which must vary continuously with x. For instance a sphere has two possible orientations, one with all these vectors $n(x)$ pointing inside, and one with all these vectors $n(x)$ pointing outside. More generally, any surface has locally two possible orientations, so if it is connected, it has two possible orientations. In what follows the convention is that the closed surfaces are oriented, with each $n(x)$ pointing outside.

As a first illustration for this, let us do a basic computation, as follows:

PROPOSITION 12.10. *For a point charge $q \in \mathbb{R}$ at the center of a sphere S,*

$$\Phi_E(S) = \frac{q}{\varepsilon_0}$$

where the constant is $\varepsilon_0 = 1/(4\pi K)$, independently of the radius of S.

PROOF. Assuming that S has radius r, we have the following computation:

$$\begin{aligned}
\Phi_E(S) &= \int_S < E(x), n(x) > dx \\
&= \int_S \left\langle \frac{Kqx}{r^3}, \frac{x}{r} \right\rangle dx \\
&= \int_S \frac{Kq}{r^2}\, dx \\
&= \frac{Kq}{r^2} \times 4\pi r^2 \\
&= 4\pi Kq
\end{aligned}$$

Thus with $\varepsilon_0 = 1/(4\pi K)$ as above, we obtain the result. □

More generally now, we have the following result:

THEOREM 12.11. *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) By using now spherical coordinates, the integral from (3) becomes:

$$\begin{aligned}
\Phi_E(S) &= Kq \int_S \frac{1-rx}{(1-2xr+r^2)^{3/2}}\, du \\
&= Kq \int_0^{2\pi} \int_0^{\pi} \frac{1-r\cos s}{(1-2r\cos s+r^2)^{3/2}} \cdot \sin s\, ds\, dt \\
&= 2\pi Kq \int_0^{\pi} \frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}}\, ds \\
&= \frac{q}{2\varepsilon_0} \int_0^{\pi} \frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}}\, ds
\end{aligned}$$

(5) The point now is that the integral on the right can be computed with the change of variables $x = \cos s$. Indeed, we have $dx = -\sin s\, ds$, and we obtain:

$$\begin{aligned}\int_0^\pi \frac{(1-r\cos s)\sin s}{(1-2r\cos s+r^2)^{3/2}}\,ds &= \int_{-1}^1 \frac{1-rx}{(1-2rx+r^2)^{3/2}}\,dx\\ &= \left[\frac{x-r}{\sqrt{1-2rx+r^2}}\right]_{-1}^1\\ &= \frac{1-r}{\sqrt{1-2r+r^2}}-\frac{-1-r}{\sqrt{1+2r+r^2}}\\ &= \frac{1-r}{|1-r|}+1\\ &= 2\delta_{r<1}\end{aligned}$$

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

Even more generally now, we have the following result, due to Gauss:

THEOREM 12.12 (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 12.11, or even from Proposition 12.10, 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 12.11, by assuming this time that the integration takes place on an arbitrary surface as follows:

$$S_\lambda = \left\{\lambda(u)u\,\middle|\,u \in S\right\}$$

And I will leave the computations here to you, as an interesting exercise.

(4) But with this, we can finish by using a standard math trick. We can assume indeed 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}. Then, 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:

$$\begin{aligned}
\Phi_E(S) &= \Phi_E(S - \cup U_i) + \Phi_E(\cup U_i) \\
&= 0 + \Phi_E(\cup U_i) \\
&= \sum_i \Phi_E(U_i) \\
&= \sum_i \frac{q_i}{\varepsilon_0} \\
&= \frac{Q_{enc}}{\varepsilon_0}
\end{aligned}$$

Thus, we are led to the conclusion in the statement. We will be actually back to the present theorem in a moment, with a second proof as well, even more conceptual. □

12b. Green and Stokes

In order to reach to a better understanding of the Gauss law, we will need some more mathematics. Let us start with the following basic result, in 2 dimensions:

THEOREM 12.13 (Green). *Given a plane curve $C \subset \mathbb{R}^2$, we have the formula*

$$\int_C P dx + Q dy = \int_D \left(\frac{dQ}{dx} - \frac{dP}{dy} \right) dx dy$$

where $D \subset \mathbb{R}^2$ is the domain enclosed by C.

PROOF. We can parametrize the enclosed domain D as follows:

$$D = \left\{ (x, y) \middle| a \leq x \leq b, f(x) \leq y \leq g(x) \right\}$$

We have then the following computation, which gives the result for P:

$$\begin{aligned}
\int_D -\frac{dP}{dy} dx dy &= \int_a^b \left(\int_{f(x)}^{g(x)} -\frac{dP}{dy}(x, y) dy \right) dx \\
&= \int_a^b P(x, f(x)) - P(x, g(x)) dx \\
&= \int_a^b P(x, f(x)) dx + \int_b^a P(x, g(x)) dx \\
&= \int_C P dx
\end{aligned}$$

As for the result for the Q term, the computation here is similar. □

Getting now to 3D, let us begin with a standard definition, as follows:

DEFINITION 12.14. *Given a function $f : \mathbb{R}^3 \to \mathbb{R}$, its usual derivative $f'(u) \in \mathbb{R}^3$ can be written as $f'(u) = \nabla f(u)^t$, 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.

As a continuation now of our calculus work, we have the following result:

THEOREM 12.15 (Stokes). *Given a smooth oriented surface $S \subset \mathbb{R}^3$, with boundary $C \subset \mathbb{R}^3$, and a vector field F, we have the following formula:*

$$\int_S < (\nabla \times F)(x), n(x) > dx = \int_C < F(x), dx >$$

In other words, the line integral of a vector field F over a loop C equals the surface integral of the curl of the vector field φ, over the enclosed surface S.

PROOF. This basically follows from the Green theorem, the idea being as follows:

(1) Let us first parametrize our surface S, and its boundary C. We can assume that we are in the situation where we have a closed oriented curve $\gamma : [a, b] \to \mathbb{R}^2$, with interior $D \subset \mathbb{R}^2$, and where the surface appears as $S = \psi(D)$, with $\psi : D \to \mathbb{R}^3$. In this case, the function $\delta = \psi \circ \gamma$ parametrizes the boundary of our surface, $C = \delta[a, b]$.

(2) Let us first look at the integral on the right in the statement. We have the following formula, with $J_y(\psi)$ standing for the Jacobian of ψ at the point $y = \gamma(t)$:

$$\begin{aligned}\int_C < F(x), dx > &= \int_\gamma < F(\psi(\gamma)), d\psi(\gamma) > \\ &= \int_\gamma < F(\psi(y)), J_y(\psi) d\gamma >\end{aligned}$$

In order to further process this formula, let us introduce the following function:

$$P(u,v) = \left\langle F(\psi(u,v)), \frac{d\psi}{du}(u,v) \right\rangle e_u + \left\langle F(\psi(u,v)), \frac{d\psi}{dv}(u,v) \right\rangle e_v$$

In terms of this function, we have the following formula for our line integral:

$$\int_C < F(x), dx >= \int_\gamma < P(y), dy >$$

(3) In order to compute now the other integral in the statement, we first have:

$$\begin{aligned}&\frac{dP_v}{du} - \frac{dP_u}{dv} \\ = &\left\langle \frac{d(F\psi)}{du}, \frac{d\psi}{dv} \right\rangle + \left\langle F\psi, \frac{d^2\psi}{dudv} \right\rangle - \left\langle \frac{d(F\psi)}{dv}, \frac{d\psi}{du} \right\rangle - \left\langle F\psi, \frac{d^2\psi}{dvdu} \right\rangle \\ = &\left\langle \frac{d(F\psi)}{du}, \frac{d\psi}{dv} \right\rangle - \left\langle \frac{d(F\psi)}{dv}, \frac{d\psi}{du} \right\rangle \\ = &\left\langle \frac{d\psi}{dv}, \left(J_{\psi(u,v)}F - (J_{\psi(u,v)}F)^t\right) \frac{d\psi}{du} \right\rangle \\ = &\left\langle \frac{d\psi}{dv}, (\nabla \times F) \times \frac{d\psi}{du} \right\rangle \\ = &\left\langle \nabla \times F, \frac{d\psi}{du} \times \frac{d\psi}{dv} \right\rangle\end{aligned}$$

We conclude that the integral on the left in the statement is given by:

$$\begin{aligned}&\int_S < (\nabla \times F)(x), n(x) > dx \\ = &\int_D \left\langle (\nabla \times F)(\psi(u,v)), \frac{d\psi}{du}(u,v) \times \frac{d\psi}{dv}(u,v) \right\rangle dudv \\ = &\int_D \left(\frac{dP_v}{du} - \frac{dP_u}{dv} \right) dudv\end{aligned}$$

(4) But with this, we are done, because the integrals computed in (2) and (3) are indeed equal, due to the Green theorem. Thus, the Stokes formula holds indeed. □

As a conclusion to what we did so far, we have the following statement:

THEOREM 12.16. *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, with (1) being something that we know well in 1D, namely fundamental theorem of calculus, and the general, N-dimensional formula following from that, and with (2) and (3) being established above. □

Getting back now to electrostatics, as a main application of the above, we have the following new point of view on the Gauss formula, which is more conceptual:

THEOREM 12.17 (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. □

12c. Maxwell equations

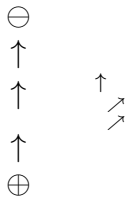
Time to make our charges move. Just by feeding a light bulb with a battery, and playing a bit with the cables, we are led to the following interesting conclusion:

FACT 12.18. *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 12.19 (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 12.18. 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 12.20. *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 result:

FACT 12.21 (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 12.19, and in the definition of $\times$. In fact, the Lorentz force law is just a fancy reformulation of Fact 12.19, 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 12.19.

As an interesting application of the Lorentz force law, we have:

THEOREM 12.22. *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 12.23. *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 12.21 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 12.24. *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 12.25. *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 12.26 (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 12.27 (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 12.28. *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. Indeed, 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, coming from the Gauss formula.

Together with Fact 12.27, 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 12.29. *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, we have:

$$\varepsilon_0\mu_0c^2 \simeq 8.854 \times 1.257 \times 3.000^2 \times 10^{16-12-6} = 1.001$$

Getting back now to theory and math, the Biot-Savart law has as consequence:

THEOREM 12.30. *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 12.31 (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 12.32. *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 12.30 and Theorem 12.31, which both follow from the Biot-Savart law. □

With this discussed, what is next? Not very clear, and in the lack of our usual 3D advisor, the rat, I mean I even tried with cheese, and he's not there, here I am knocking on the door of my neighbor, asking for a business meeting with her hamster. And with the meeting approved, and hamster being in a very good mood, here is what he says:

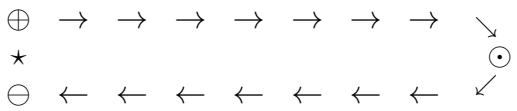
HAMSTER 12.33. *Sir, try making your electrons run around a loop, that can be quite fun, and good for their health, too.*

Thanks hamster, this sounds quite interesting. So, getting now to my garage, and doing some experiments, here is what I found, and with later my physics colleagues from Cergy telling me that this was actually known, long before, since Faraday:

FACT 12.34 (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 better explain 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 12.35. *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 12.34, these can be fine-tuned, and reformulated in terms of the emf, in the following way:

FACT 12.36 (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 12.37 (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 12.38 (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 12.39 (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. □

And with this, end of our discussion regarding electromagnetism. Good to have these Maxwell equations on the table, and more on them later, in Part IV, when discussing relativity and quantum mechanics, which are in fact behind all this. More later.

12d. Thermodynamics

Not much time left for thermodynamics, in this chapter, and present Part III, but we should definitely talk about this too. To start with, we already met gases, pressure and heat on several occasions in this book, and what we know on the subject makes it quite clear that any attempt of axiomatization would be something terribly complicated. So, modesty, and instead of doing mathematical physics, as usual in this book, we will be doing here plain physics, I mean phenomenology discussed, along with some formulae.

Let us start with some philosophy, which is actually very interesting in relation with the general topic of this book, "forces and vectors". We have been dealing so far with "clean" physics, clear forces acting on clear objets, given by exact formulae, or at least by exact equations. Of course, we have met a few dirty phenomena too, such as friction and drag in the context of gravity, but we have usually waived such phenomena with equations of the following type, with $\lambda \in (0,1)$ being a constant:

$$F_{real} = \lambda F_{abstract}$$

Getting now into dirty, real-life physics, in connection with nature surrounding us, obviously what happens around us is not the effect of a single force, or of a handful of such forces. At work are millions and millions of tiny little forces, acting between the various small particles that we are made of, and what makes things rolling is the resulting "force", which appears as an average of these tiny little foces, over time $t > 0$:

$$F_{real} = \int_{time} \int_{space} F_{tiny}$$

As a first comment, such kind of average is obviously not a force in the usual sense. It is rather "something", corresponding to a transformation of a complex system, over a perceptible period of time $t > 0$. So, as a conclusion, we will have to abandon our beloved concept of force, and look for more realistic physical quantities to deal with.

In thermodynamics, these quantities are pressure P and temperature T, which, as mentioned above, are quite difficult to axiomatize. So, take it easy, and since we already know a bit about pressure P, here is something about temperature T as well:

METHOD 12.40. *Temperature measures the heat present in the environement, and can be computed by exposing various materials, which can be solids, liquids or gases, to the environement, and measuring their volume, based on the following facts:*

(1) *Heat dilates the material, and cold contracts the material.*
(2) *Heat flows from warm to cold, aiming at equalizing temperature.*
(3) *That heat flow needs some perceptible time $t > 0$, to fully happen.*
(4) *Temperature can be regulated, as to be independent of the material used.*

All this looks quite reasonable, save perhaps for (1) which certainly does not apply to things like freezing water, then for the flow direction in (2), which is opposite to the direction where the wind goes, but this will be the least of our concerns, for the moment, and then for the mysterious time $t > 0$ needed in (3), and finally for some engineering concerns in relation with (4) which again, will be the least of our concerns.

So long for the axiomatics, based on Method 12.40 we have now thermometers, so we can talk about the beast T that these devices measure, even without precisely knowing what T is. Going now directly for the kill, at the start of everything modern, we have:

FACT 12.41. *An ideal gas is subject to the equation of state*

$$PV = kT \quad : \quad k = Nb$$

where N is the number of molecules present, and where

$$b = 1.380\ 649 \times 10^{-23}$$

is an exact constant, called Boltzmann constant.

As a first comment, this fact is something old as ever, going back to the 18th century greats, such as Boyle and Charles, and then with further important contributions in the 19th century, by Gay-Lussac and Clausius, and then Maxwell and Boltzmann.

As a second comment, the above fact lies somewhere between fact and theorem. Indeed, we know from chapter 6 that we have $PV = 2K/3$, and somehow the argument would be that the total kinetic energy K corresponds to temperature T, via:

$$\frac{2K}{3} = kT$$

Equivalently, by dividing everything by the number of molecules N, the individual molecular kinetic energy K_0 should correspond to temperature T, via:

$$\frac{2K_0}{3} = bT$$

But is this obvious? We know that temperature T measures heat Q, which is a form of energy, and with this in mind, having $K_0 \sim T$ perfectly makes sense. But finding the precise proportionality factor amounts in doing the math for the approximately 10^{10} collisions per second that must be added, and with this leading us into countless challenges. And in addition to this, above everything, even if you manage to do the math, you still need afterwards experimental physicists to confirm the fact that your model is indeed the correct one, physically speaking. So, things complicated, with Fact 12.41.

This being said, as a matter of having some mathematics going, here is a key result on the subject, due to Maxwell and Boltzmann, intimately related to $PV = kT$:

THEOREM 12.42. *The molecular speeds $v \in \mathbb{R}^3$ of a gas in thermal equilibrium are subject to the Maxwell-Boltzmann distribution formula*

$$P(v) = \left(\frac{m}{2\pi bT}\right)^{3/2} \exp\left(-\frac{m||v||^2}{2bT}\right)$$

with m being the mass of the molecules, and b being the Boltzmann constant.

PROOF. As before with other such things, this is something in between fact and theorem. Here is Maxwell's original argument, which is something quite elementary:

(1) We are looking for the precise probability distribution P of the molecular speeds $v = (v_1, v_2, v_3)$ which makes the mechanics of gases work. Intuition tells us that P has no corellations between the x, y, z directions of space, and so we must have:

$$P(v) = f(v_1)g(v_2)h(v_3)$$

Moreover, by rotational symmetry the functions f, g, h must coincide, and so:

$$P(v) = f(v_1)f(v_2)f(v_3)$$

(2) Further thinking, again invoking rotational symmetry, leads to the conclusion that $P(v)$ must depend only on the magnitude $||v||$ of the velocity $v \in \mathbb{R}^3$, and not on the direction. Thus, we must have as well a formula of the following type:

$$P(v) = \varphi(||v||^2)$$

(3) Now by comparing the requirements in (1) and (2), we are led via some math to the conclusion that φ must be an exponential, which amounts in saying that:

$$P(v) = \lambda \exp\left(-C||v||^2\right)$$

(4) Obviously we must have $C > 0$, for things to be bounded, and then by integrating we can obtain λ as function of C, and our formula becomes:

$$P(v) = \left(\frac{C}{\pi}\right)^{3/2} \exp\left(-C||v||^2\right)$$

(5) It remains to find the value of $C > 0$. But for this purpose, observe that, now that we have our distribution, be that still depending on $C > 0$, we can compute everything that we want to, just by integrating. In particular, we find that on average:

$$v_1^2 = v_2^2 = v_3^2 = \frac{1}{2C}$$

Thus the average magnitude of the molecular speed is given by:

$$||v|| = \frac{3}{2C}$$

It follows that the average kinetic energy of the molecules is:

$$K_0 = \frac{m||v||^2}{2} = \frac{3m}{4C}$$

(6) On the other hand, recall from our discussion before that one of the many equivalent formulations of $PV = kT$, using $PV = 2K/3$, was as follows:

$$\frac{2K_0}{3} = bT$$

(7) Thus we obtain $m/(2C) = bT$, and so $C = m/(2bT)$, as desired. □

Still regarding the gases, at a more advanced level now, we have:

THEOREM 12.43. *Beyond the ideal gas setting, stating that we should have*

$$PV = kT$$

the gases are subject to the Van der Waals equation

$$\left(P + \frac{\alpha}{V^2}\right)(V - \beta) = kT$$

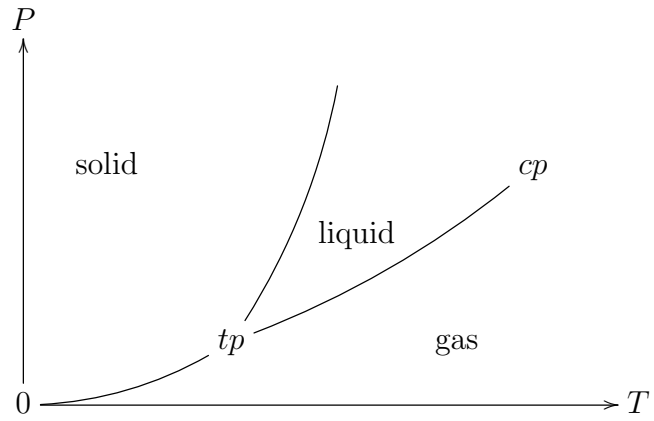
depending on two parameters $\alpha, \beta > 0$.

PROOF. This is something quite tricky, with the correction parameters $\alpha, \beta > 0$ appearing from a detailed study of the gas, from a kinetic viewpoint. For an introduction to all this, we refer for instance to Huang [50], or Schroeder [76]. □

The above result is of key importance, and takes us into rethinking everything that what we know about the ideal gases, which must be replaced with Van der Waals gases, at the advanced level. Among the main consequences of this replacement, the isobars, isochores, isothermals and adiabatics of the ideal gases, given by simple formulae, must be replaced by isobars, isochores, isothermals and adiabatics for the Van der Waals gases, which are no longer something trivial, with some interesting math being now involved.

Among others, this Van der Waals gas study makes appear some interesting points on the isothermals, called triple and critical points of the gas. Which makes the connection with the general theory of matter, which can be summarized as follows:

FACT 12.44. *Ordinary matter appears in* 3 *forms, namely solid, liquid and gaseous, roughly appearing according to the following generic diagram*



with tp, cp *standing for the triple and critical points. Also, at low or high temperatures we have interesting phenonema like Bose-Einstein condensation, and plasma.*

Needless to say, the quantity of things that can be said about all this is enormous. Have a look at some standard books here, such as [9], [18], [41], [48], [59], [97].

12e. Exercises

Quite compact chapter that we had here, doing just 1/2 of the things that I was planning to talk about, and as exercises on this, all about learning, we have:

EXERCISE 12.45. *Learn more about the Coulomb law and constant.*

EXERCISE 12.46. *Learn more about the Biot-Savart law, and constants involved.*

EXERCISE 12.47. *Learn more about the Ampère law, and its Maxwell correction.*

EXERCISE 12.48. *Learn about the light basics, coming from the Maxwell equations.*

EXERCISE 12.49. *Learn about basic thermodynamics, and the work of Joule.*

EXERCISE 12.50. *Learn about Carnot and engines, and absolute temperature.*

EXERCISE 12.51. *Learn from Ohm about the relation between electricity and heat.*

EXERCISE 12.52. *Then learn from Planck more about this, radiation and heat.*

As bonus exercise, which is part of the standard initiation to physics, choose between electrodynamics and thermodynamics, and start reading some more.

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

Welcome to mathematics and physics in higher dimensions, typically $N = 4$, following Einstein's relativity theory, where space and time are related, or $N = \infty$, following the theory of quantum mechanics, as developed by Heisenberg, Schrödinger and the others, and with anything in between, $4 < N < \infty$, being of course not ruled out either.

Things will be quite difficult, with relativity being related to motion and light, and so to classical mechanics and electromagnetism, and with the relation with gravity being present too, but in a more complicated way. As for quantum mechanics, that originally appeared as a continuation of electromagnetism, meant to be exact, but non-relativistic, which is a bit contradictory, electromagnetism being relativistic. But with the relativistic correction being possible, leading to something quite complicated, namely quantum electrodynamics. Which itself has troubles in being unified with the relativistic gravitational theory, the finer particle theory and thermodynamics being not known yet to us.

In short, expect something quite complicated, using everything that we know, and featuring some substantial logical loops, as we like them in physics, that would make any pure mathematician jump up to the ceiling. As for the pure physicists, these will be not spared either, with the main principle of theoretical physics, namely that God made this world by using simple principles and equations, being quite often replaced in our context by some fairly complicated equations, as we sometimes enjoy them in mathematics.

So, God bless, and let us have our exploration started. Advising us will be the cats, and with this being a good choice I guess, cats being the most philosopher creations on this planet, except perhaps for big trees, oceans and mountains. So, going now for my good old lady, always there grumpy on the porch, meditating, here is what she says:

CAT 13.1. *Does not look easy, what you want to understand. Start with some mathematics, things that you know, and improve them a bit.*

Which looks like a good advice, thanks cat. So, getting now into feline mode, I would say that what we have as most valuable technology so far is linear algebra, I mean thinking a bit, algebra, geometry, analysis and probability all come from linear algebra, and this would be a good starting point, more linear algebra, developed without blinking.

In practice, we already have some good knowledge of linear algebra from chapters 7 and 11, notably with the notion of determinant, developed there. However, as explained in chapter 7, the main topic in linear algebra is diagonalization, which still remains to be discussed. The basic diagonalization theory, that we already know, is as follows:

THEOREM 13.2. *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, that we know from chapter 7, 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 too, because we have $Pe_i = v_i$, and so:

$$PDP^{-1}v_i = PDe_i = P\lambda_i e_i = \lambda_i Pe_i = \lambda_i v_i$$

Thus, the matrices A and PDP^{-1} coincide, as stated. □

Let us discuss now some basic examples, namely the rotations, symmetries and projections in 2 dimensions. Regarding the symmetries and projections, we have:

PROPOSITION 13.3. *The symmetry with respect to the Ox axis rotated by an angle $t/2 \in \mathbb{R}$, and the projection on this same rotated Ox axis, given by*

$$S_t = \begin{pmatrix} \cos t & \sin t \\ \sin t & -\cos t \end{pmatrix} \quad , \quad P_t = \frac{1}{2}\begin{pmatrix} 1+\cos t & \sin t \\ \sin t & 1-\cos t \end{pmatrix}$$

are both diagonalizable, their diagonal forms being as follows,

$$S_t \sim \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad , \qquad P_t \sim \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

and this regardless of the value of the angle $t/2$.

PROOF. Both assertions are clear, and this even without computations, as follows:

(1) 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.

(2) With L being as before, now 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. □

Regarding now the rotations, here the situation is more complicated, as follows:

PROPOSITION 13.4. *The rotation of angle $t \in \mathbb{R}$ in the plane, given by*

$$R_t = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}$$

diagonalizes 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 the diagonalization is impossible, unless $t = 0, \pi$.

PROOF. We know this from chapter 7, with the eigenvectors coming from:

$$R_t \begin{pmatrix} 1 \\ i \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} 1 \\ i \end{pmatrix} = \begin{pmatrix} \cos t - i \sin t \\ i \cos t + \sin t \end{pmatrix} = e^{-it} \begin{pmatrix} 1 \\ i \end{pmatrix}$$

$$R_t \begin{pmatrix} 1 \\ -i \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} 1 \\ -i \end{pmatrix} = \begin{pmatrix} \cos t + i \sin t \\ -i \cos t + \sin t \end{pmatrix} = e^{it} \begin{pmatrix} 1 \\ -i \end{pmatrix}$$

Indeed, we are led in this way to the conclusions in the statement. □

As a non-example now, still in 2D, nightmare of all mathematicians, we have:

PROPOSITION 13.5. *The following matrix is not diagonalizable,*

$$J = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

because it has only 1 eigenvector.

PROOF. This is something that we know from chapter 7 too, the point being that for the matrix J in the statement, the eigenvector/eigenvalue equation $Jv = \lambda v$ reads:

$$\begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{pmatrix} \lambda x \\ \lambda y \end{pmatrix}$$

Since at $\lambda \neq 0$ we have no non-trivial eigenvectors, we are left with the case $\lambda = 0$, where the eigenvectors are the multiples of $\binom{1}{0}$, leading to the above conclusions. □

Finally, as a matter of having something in $N \geq 3$ dimensions too, let us record:

THEOREM 13.6. *The all-one, or flat matrix, diagonalizes according to*

$$\frac{1}{N}\begin{pmatrix}1&\ldots&1\\ \vdots&&\vdots\\ 1&\ldots&1\end{pmatrix}=\frac{1}{N}F_N\begin{pmatrix}1&&&\\ &0&&\\ &&\ddots&\\ &&&0\end{pmatrix}F_N^*$$

where $F_N=(w^{ij})_{ij}$ with $w=e^{2\pi i/N}$ is the Fourier matrix, and $F_N^*=(w^{-ij})_{ij}$.

PROOF. This is something quite tricky, worth a detailed discussion, as follows:

(1) Let us denote by $\mathbb{I}_N$ the all-1 matrix, set $P_N=\mathbb{I}_N/N$, and consider as well the all-1 vector $\xi=(1)_i\in\mathbb{R}^N$. The matrices $\mathbb{I}_N,P_N$ act on the vectors $v\in\mathbb{R}^N$ as follows:

$$\mathbb{I}_Nv=\sum_i v_i\cdot\xi\quad,\quad P_Nv=\frac{\sum_i v_i}{N}\cdot\xi$$

Now the point is that the formula on the right shows that P_N is the orthogonal projection on ξ, and so, as for any rank 1 projection, its diagonal form must be:

$$P_N\sim\begin{pmatrix}1&&&\\ &0&&\\ &&\ddots&\\ &&&0\end{pmatrix}$$

(2) Next, how to diagonalize P_N? And the problem here is that, when looking for 0-eigenvectors, we must explicitly solve the following equation:

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

Which is not an easy task, if we want a nice basis for the space of solutions. I mean, try this over the reals, at $N=3,4$, in order to understand what I am talking about.

(3) Fortunately, the complex numbers, and more specifically the N-th roots of unity, come to the rescue. Indeed, with $w=e^{2\pi i/N}$ as in the statement, we have the following magic formula, coming from an obvious barycenter computation, in the plane:

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

But this shows that the columns of the Fourier matrix $F_N=(w^{ij})_{ij}$ are the desired eigenvectors for P_N, and with the standard discrete Fourier analysis convention that the matrix indices are $i,j=0,1,\ldots,N-1$, we are led to the formula in the statement.

(4) To be more precise here, we have the diagonal form from (1), and the passage matrix from (3), and the point is that the root of unity formula in (3) shows that the rescaling $U=F_N/\sqrt{N}$ is unitary, in the sense that $U^{-1}=U^*$, with $U^*=(\bar{U}_{ji})_{ij}$. Thus, by putting everything together, we are led to the formula in the statement. □

Time now for some theory? In order to study the diagonalization problem, the idea is that the eigenvectors can be grouped into linear spaces, called eigenspaces:

THEOREM 13.7. *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 13.8. *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 13.9. *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. Indeed, by summing the inequalities $\dim(E_\lambda) \leq m_\lambda$ from Theorem 13.8, 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 13.7 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 13.9. As a first consequence of Theorem 13.9, which is very useful in practice, we have:

THEOREM 13.10. *If a matrix $A \in M_N(\mathbb{C})$ has distinct eigenvalues, then it is diagonalizable. Moreover, this is indeed the case, for the generic matrices.*

PROOF. The first assertion is clear from Theorem 13.9, because the criterion there for diagonalization is trivially satisfied when the eigenvalues are different, as follows:

$$\sum_\lambda \dim(E_\lambda) = \sum_\lambda 1 = N$$

As for the second assertion, this is something quite intuitive, coming from the fact that N numbers $\lambda_1, \ldots, \lambda_N \in \mathbb{C}$ picked at random must be distinct. Of course, this does not stand as a formal proof, but we will come back to this in a moment, with a proof. □

Getting back now to Theorem 13.9, the main problem raised by the diagonalization procedure is the computation of the eigenvalues. As here, as a useful trick, we have:

THEOREM 13.11. *The complex eigenvalues of a matrix $A \in M_N(\mathbb{C})$, counted with multiplicities, have the following properties:*

(1) *Their sum is the trace.*
(2) *Their product is the determinant.*

PROOF. Consider indeed the characteristic polynomial P of the matrix:

$$\begin{aligned} P(X) &= \det(A - X1_N) \\ &= (-1)^N X^N + (-1)^{N-1} Tr(A) X^{N-1} + \ldots + \det(A) \end{aligned}$$

We can factorize this polynomial, by using its N complex roots, and we obtain:

$$\begin{aligned} P(X) &= (-1)^N (X - \lambda_1) \ldots (X - \lambda_N) \\ &= (-1)^N X^N + (-1)^{N-1} \left(\sum_i \lambda_i \right) X^{N-1} + \ldots + \prod_i \lambda_i \end{aligned}$$

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

Regarding now the intermediate terms, we have here the following result:

THEOREM 13.12. *Assume that $A \in M_N(\mathbb{C})$ has eigenvalues $\lambda_1, \ldots, \lambda_N \in \mathbb{C}$, counted with multiplicities. The basic symmetric functions of these eigenvalues, namely*

$$c_k = \sum_{i_1 < \ldots < i_k} \lambda_{i_1} \ldots \lambda_{i_k}$$

are then given by the fact that the characteristic polynomial of the matrix is:

$$P(X) = (-1)^N \sum_{k=0}^{N} (-1)^k c_k X^k$$

Moreover, all symmetric functions of the eigenvalues, such as the sums of powers

$$d_s = \lambda_1^s + \ldots + \lambda_N^s$$

appear as polynomials in these characteristic polynomial coefficients c_k.

PROOF. As explained in the previous proof, the characteristic polynomial is:

$$\begin{aligned} P(X) &= (-1)^N (X - \lambda_1) \ldots (X - \lambda_N) \\ &= (-1)^N X^N + (-1)^{N-1} \left(\sum_i \lambda_i \right) X^{N-1} + \ldots + \prod_i \lambda_i \\ &= (-1)^N \left(X^N - c_1 X^{N-1} + \ldots + (-1)^N c_N \right) \end{aligned}$$

Thus, with the convention $c_0 = 1$, we are led to the first assertion. As for the second assertion, this comes by doing some abstract algebra, say good exercise for you. □

Let us go back now to Theorem 13.10, and more specifically, to the last assertion there, which was a quite strong statement. In order to discuss this, we first have:

THEOREM 13.13. *For a matrix $A \in M_N(\mathbb{C})$ the following conditions are equivalent,*

(1) *The eigenvalues are different, $\lambda_i \neq \lambda_j$,*
(2) *The characteristic polynomial P has simple roots,*
(3) *The characteristic polynomial satisfies $(P, P') = 1$,*
(4) *The discriminant of P is nonzero, $\Delta(P) \neq 0$,*

and in this case, the matrix is diagonalizable.

PROOF. The equivalences in the statement are all standard, the idea being as follows:

(1) $\iff$ (2) This follows indeed from Theorem 13.9.

(2) $\iff$ (3) This is standard, the double roots of P being roots of P'.

(3) $\iff$ (4) This is something standard too, coming from the well-known fact that the discriminant $\Delta(P)$ appears as a normalized version of the resultant $R(P, P')$. And for more on this, you can check any standard calculus book, such as mine [11].

As for the last assertion, this is something that we know, from Theorem 13.10. □

We can now put everything together, along with some more, and we obtain the following useful collection of density results, which are all nuclear-grade weapons:

THEOREM 13.14. *The following happen, inside $M_N(\mathbb{C})$:*

(1) *The invertible matrices are dense.*
(2) *The matrices having distinct eigenvalues are dense.*
(3) *The diagonalizable matrices are dense.*

PROOF. These are quite advanced results, which can be proved as follows:

(1) This is clear, intuitively speaking, because the invertible matrices are given by the condition $\det A \neq 0$. Thus, the set formed by these matrices appears as the complement of the hypersurface $\det A = 0$, and so must be dense inside $M_N(\mathbb{C})$, as claimed.

(2) Here we can use a similar argument, this time by saying that the set formed by the matrices having distinct eigenvalues appears as the complement of the hypersurface given by $\Delta(P_A) = 0$, and so must be dense inside $M_N(\mathbb{C})$, as claimed.

(3) This follows from (2), via the fact that the matrices having distinct eigenvalues are diagonalizable, that we know from Theorem 13.10. There are of course some other proofs as well, for instance by putting the matrix in Jordan form. □

As an interesting list of applications of the above density results, we have:

THEOREM 13.15. *The following happen:*

(1) *We have $P_{AB} = P_{BA}$, for any two matrices $A, B \in M_N(\mathbb{C})$.*
(2) *AB, BA have the same eigenvalues, with the same multiplicities.*
(3) *If A has eigenvalues $\lambda_1, \ldots, \lambda_N$, then $f(A)$ has eigenvalues $f(\lambda_1), \ldots, f(\lambda_N)$.*

PROOF. These results can be deduced by using Theorem 13.14, as follows:

(1) It follows from definitions that the characteristic polynomial of a matrix is invariant under conjugation, in the sense that we have the following formula:

$$P_C = P_{ACA^{-1}}$$

Now observe that, when assuming that A is invertible, we have:

$$AB = A(BA)A^{-1}$$

Thus, we have the result when A is invertible. By using now Theorem 13.14 (1), we conclude that this formula holds for any matrix A, by continuity.

(2) This is a reformulation of (1) above, via the fact that P encodes the eigenvalues, with multiplicities, and which is hard to prove with bare hands.

(3) This is something more informal, clear for the diagonal matrices D, then for the diagonalizable matrices PDP^{-1}, and finally for all matrices, by using Theorem 13.14 (3), provided that f has suitable regularity properties. We will be back to this. □

13b. Spectral theorems

In order to reach to a more advanced diagonalization theory, we need to have a more geometric look at the matrices $A \in M_N(\mathbb{C})$, and at linear algebra in general. We first have the following result, regarding the space $\mathbb{C}^N$ itself, which is straightforward:

THEOREM 13.16. *We can talk about scalar products and lengths in $\mathbb{C}^N$,*

$$< x, y >= \sum_i x_i \bar{y}_i \qquad , \qquad ||x|| = \sqrt{\sum_i |x_i|^2}$$

which are related by the following conversion formulae,

$$||x|| = \sqrt{< x, x >} \quad , \quad < x, y >= \frac{||x+y||^2 - ||x-y||^2 + i||x+iy||^2 - i||x-iy||^2}{4}$$

and the following happen:

(1) $< \lambda x, y >=< x, \bar{\lambda} y >= \lambda < x, y >$.
(2) $< x + y, z >=< x, z > + < y, z >$.
(3) $< x, y + z >=< x, z > + < y, z >$.
(4) $< x, y >= \overline{< y, x >}$.
(5) $||\lambda x|| = |\lambda| \cdot ||x||$.
(6) $| < x, y > | \leq ||x|| \cdot ||y||$.
(7) $||x + y|| \leq ||x|| + ||y||$.
(8) $x \perp y \iff < x, y >= 0$, *by definition.*

PROOF. This is a straightforward complex remake of our previous result regarding the real scalar products, from chapter 7, the details being as follows:

(1) To start with, we can certainly talk about scalar products and lengths in $\mathbb{C}^N$, as above, and with the scalar product convention, which is the so-called mathematical one, coming from the cats, I mean we had a cat seminar the other day, carefully examined what has been done in quantum physics, from Heisenberg to Witten, and concluded that the complex scalar products are best denoted $< x, y >$, and taken linear at left.

(2) Next, the proof of the complex polarization identity is as follows:

$$\begin{aligned} & ||x+y||^2 - ||x-y||^2 + i||x+iy||^2 - i||x-iy||^2 \\ = \; & ||x||^2 + ||y||^2 - ||x||^2 - ||y||^2 + i||x||^2 + i||y||^2 - i||x||^2 - i||y||^2 \\ & +2Re(< x, y >) + 2Re(< x, y >) + 2iIm(< x, y >) + 2iIm(< x, y >) \\ = \; & 4 < x, y > \end{aligned}$$

(3) We have as well a parallelogram rule, whose proof is similar, by developing the scalar products, that I forgot to mention in the above statement, as follows:

$$||x + y||^2 + ||x - y||^2 = 2(||x||^2 + ||y||^2)$$

(4) As for the various claims in the statement, their proof is similar to the one from the real case, from chapter 7, and with Cauchy-Schwarz, which is the key statement there, using the function $f(t) = ||wx + ty||^2$, with $t \in \mathbb{R}$ and $|w| = 1$, good exercise for you. □

Back now to the complex matrices, the point is that these do not come alone, but rather in pairs (A, A^*), with the basics of the $A \to A^*$ operation being as follows:

THEOREM 13.17. *Given $A \in M_N(\mathbb{C})$, define its adjoint matrix by $(A^*)_{ij} = \bar{A}_{ji}$.*

(1) $< Ax, y >=< x, A^*y >$, *for any two vectors $x, y \in \mathbb{C}^N$.*
(2) $x \to Ux$ *with $U \in M_N(\mathbb{C})$ is an isometry precisely when $U^* = U^{-1}$.*
(3) $x \to Px$ *with $P \in M_N(\mathbb{C})$ is a projection precisely when $P^2 = P^* = P$.*

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

(1) The formula $< Ax, y >=< x, A^*y >$ is indeed clear from definitions on the basis vectors, $x, y \in \{e_1, \ldots, e_N\}$, and then by linearity, it must hold in general, as stated.

(2) Given a matrix $U \in M_N(\mathbb{C})$, we have indeed the following equivalences, with the first one coming from the polarization identity, and the other ones being clear:

$$\begin{aligned} ||Ux|| = ||x|| &\iff < Ux, Uy >=< x, y > \\ &\iff < x, U^*Uy >=< x, y > \\ &\iff U^*Uy = y \\ &\iff U^*U = 1 \\ &\iff U^* = U^{-1} \end{aligned}$$

(3) Given a matrix $P \in M_N(\mathbb{C})$, in order for $x \to Px$ to be an oblique projection, we must have $P^2 = P$. Now observe that this projection is orthogonal when:

$$\begin{aligned} < Px - x, Py >= 0 &\iff < P^*Px - P^*x, y >= 0 \\ &\iff P^*Px - P^*x = 0 \\ &\iff P^*P - P^* = 0 \\ &\iff P^*P = P^* \end{aligned}$$

The point now is that by conjugating the last formula, we get $P^*P = P$. Now by comparing with $P^*P = P^*$, we obtain $P = P^*$, and this gives the result. □

Back now to our usual diagonalization business, we first have:

THEOREM 13.18. *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:

$$\begin{aligned} \lambda < v, v > &= < Av, v > \\ &= < v, Av > \\ &= \bar{\lambda} < v, v > \end{aligned}$$

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, by using the fact that we have $\mu \in \mathbb{R}$:

$$\begin{aligned} \lambda < v, w > &= < Av, w > \\ &= < v, Aw > \\ &= \mu < v, w > \end{aligned}$$

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. □

As an interesting consequence of the above result, in the real case, we have:

THEOREM 13.19. *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 13.18, and its proof. □

An important class of self-adjoint matrices, which includes for instance all the orthogonal projections, are the positive matrices. The theory here is as follows:

THEOREM 13.20. *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. This is something very standard, the idea being as follows:

(1) $\implies$ (2) This is clear, because we can take $C = B$.

(2) $\implies$ (3) This follows from the following computation:

$$< Ax, x >=< CC^*x, x >=< C^*x, C^*x > \geq 0$$

(3) $\implies$ (4) By using the fact that $< Ax, x >$ is real, we have:

$$< Ax, x >=< x, A^*x >=< A^*x, x >$$

Thus we have $A = A^*$, and the remaining assertion, regarding the eigenvalues, follows from the following computation, assuming $Ax = \lambda x$:

$$< Ax, x >=< \lambda x, x >= \lambda < x, x > \geq 0$$

(4) $\implies$ (5) This follows indeed by using Theorem 13.18.

(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 we have $B^2 = A$, as desired. □

Finally, let us record the following technical version of the above result:

THEOREM 13.21. *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, and write $A > 0$:*

(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 13.20, by adding the extra assumptions in the statement, or from the proof of Theorem 13.20, by modifying where needed. □

13c. Normal matrices

With the above discussed, done with spectral theorems, and time to get a beer? You must be kidding. Indeed, we know from Proposition 13.4 that the basic rotations are diagonalizable too, so let us have a look now into this. Here is our general result:

THEOREM 13.22. *Any matrix $U \in M_N(\mathbb{C})$ which is unitary, $U^* = U^{-1}$, is diagonalizable, with the eigenvalues on the unit circle $\mathbb{T}$. 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:

$$U^* = VD^*V^* = (V^*)^{-1}D^{-1}V^{-1} = U^{-1}$$

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. □

Well done all this, and I have this feeling that you are thirsty again. Well, not yet, because we still have to unify Theorem 13.18, dealing with the self-adjoints, with Theorem 13.22, dealing with the unitaries. And fortunately, this can be done, as follows:

THEOREM 13.23. *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:

$$AA^* = UDD^*U^* = UD^*DU^* = A^*A$$

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 $A - \lambda 1_N$:

$$||(A^* - \bar{\lambda}1_N)v|| = ||(A - \lambda 1_N)^*v|| = ||(A - \lambda 1_N)v|| = 0$$

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 13.22. □

As a first application of our spectral theorems, we have the following result:

THEOREM 13.24. *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 self-adjoint. Thus, we can diagonalize this matrix as follows, with $U \in U_N$, and with D diagonal:

$$A^*A = UDU^*$$

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:

$$\sqrt{A^*A} = U\sqrt{D}U^*$$

And with this, done, because 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 useful polar decomposition result, as follows:

THEOREM 13.25. *Given a matrix $A \in M_N(\mathbb{C})$, the following happen:*

(1) *When A is invertible, we have $A = U|A|$, with U being a unitary.*
(2) *In general, we still have $A = U|A|$, with U being a partial isometry.*

PROOF. According to our definition of the modulus, $|A| = \sqrt{A^*A}$, we have:

$$\begin{aligned} < |A|x, |A|y > &= < x, |A|^2 y > \\ &= < x, A^*Ay > \\ &= < Ax, Ay > \end{aligned}$$

We conclude that the following linear application is well-defined, and isometric:

$$U : Im|A| \to Im(A) \quad , \quad |A|x \to Ax$$

Next, we can further extend this linear isometric map U that we constructed into a partial isometry $U : \mathbb{C}^N \to \mathbb{C}^N$, in a straightforward way, by setting:

$$Ux = 0 \quad , \quad \forall x \in Im|A|^\perp$$

But with this convention, we have indeed the formula $A = U|A|$, as desired. □

And with this, curtain, good work that we did, and time now to celebrate.

13d. Spectral measures

Welcome back, and we would like to discuss now some interesting applications of our spectral theorems, to advanced probability theory. Let us formulate indeed:

DEFINITION 13.26. *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 $M_k(A) = \int_{\mathbb{R}} x^k d\mu_A(x)$.*

And isn't this clever, and potentially interesting. To be more precise, what we have here looks quite reasonable, and of distinct Heisenberg flavor, but as before with the usual random variables $f \in L^\infty(X)$, that we discussed in chapter 11, some work 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 13.27. *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 13.26.

PROOF. The first formula in the statement comes indeed from definitions. 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:

$$\int_{\mathbb{R}} x^k d\mu_A(x) = \frac{1}{N}\sum_i \int_{\mathbb{R}} x^k d\delta_{\lambda_i}(x) = \frac{1}{N}\sum_i \lambda_i^k$$

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 13.27, dealing with the general case:

THEOREM 13.28. *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 13.25.

PROOF. According to Theorem 13.18 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 13.29. *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 13.29 as stated.

Regarding now the notion of law, this extends too, the result being as follows:

THEOREM 13.30. *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 13.29 has the following straightforward analogue:

DEFINITION 13.31. *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 13.32. *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 indeed the following two computations:

$$tr(JJ^* \cdot JJ^*) = tr\left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\right) = tr\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \frac{1}{2}$$

$$tr(JJ \cdot J^*J^*) = tr\left(\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\right) = tr\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = 0$$

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

Summarizing, things are quite complicated, when attempting to talk about the law of an arbitrary matrix $A \in M_N(\mathbb{C})$. However, as a solution to this, we have:

DEFINITION 13.33. *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.

So, let us try to compute some matrix laws, in this sense, and see what we get. Based on what we have, we can formulate the following result, with mixed conclusions:

THEOREM 13.34. *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 are led to the conclusion in (4), 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 13.35. *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 13.23, 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. □

Quite interesting, all this. Shall you ever get into things like quantum groups or random matrices, Theorem 13.35 and its generalizations are what you will need.

13e. Exercises

Good linear algebra learning this was, and as exercises on this, we have:

EXERCISE 13.36. *Find the passage matrix, for the plane symmetries and rotations.*

EXERCISE 13.37. *Learn more about the Fourier matrix, and its interpretations.*

EXERCISE 13.38. *Diagonalize 100 matrices of your choice,* 3×3 *of course.*

EXERCISE 13.39. *Learn more tricks, for computing the roots of polynomials.*

EXERCISE 13.40. *Learn about the resultant and discriminant of polynomials.*

EXERCISE 13.41. *Work out the missing details, for the complex scalar products.*

EXERCISE 13.42. *Learn more about polar decomposition, and singular values.*

EXERCISE 13.43. *Find some further interesting examples of normal matrices.*

As bonus exercise, review the calculus material from chapter 7. You will enjoy.

CHAPTER 14

Relativity theory

14a. Light, revised

Getting now to physics, we would like to review in this chapter what we know about relativity, and go further. And in what regards the revision, tough task ahead, because we have bits of useful knowledge scattered all over our physics chapters so far:

– In chapter 2 we talked about the Einstein principles, and speed addition in 1D.

– In chapter 4 we discussed waves in 1D, notably with the d'Alembert theorem.

– In chapter 6 we heavily talked about relativity and related topics, mainly in 2D.

– In chapter 8 we discussed 2D mechanics, useful in relation with inertial frames.

– In chapter 10 we talked about inertial frames, and about speed addition in 3D.

– In chapter 12 we discussed the Maxwell equations, needed for producing light.

So, now at this point of this book, with us being quite advanced in our learning, both in mathematics and physics, can we make something coherent, out of this? In answer, sure yes, we can review the above material in a pleasant way, as a matter of better understanding all this, and then go a bit further, our plan being as follows:

PLAN 14.1. *In order to advance in our study of relativity, we will:*

(1) *Review the phenomenology, notably light and inertial frames, in 3D.*
(2) *Then talk speed addition in 1D, followed by Lorentz transform in 3D.*
(3) *Use the 3D Lorentz transform in order to recapture the 3D speed sum.*
(4) *Discuss mass, energy, curved spacetime, and with a look at gravity too.*

Which sounds like a nice plan, and no wonder here, this closely follows the vision of Einstein himself on relativity, as explained in his book [28], which is a must-read. Needless to say, we could not have done this before due to pedagogical reasons, I mean the mess in the previous chapters was there for a good reason, namely learning the basics.

By the way, talking people who discovered things, like Einstein, and the books that they wrote, like [28], let me warmly recommend such books, written by inventors. These are solid gold, the thing being that in theoretical science, discoveries often take a long time to be accepted and digested by the community, typically 10-20 years, and with this

being not the end of the story, because once your discovery is accepted and digested, the community can decide to go either forward or backwards, in developing science, with respect to your discovery, which in practice can add an extra 10-20 years, or even much more, to the true acceptation of your invention. And with examples of this abounding, especially in mathematics, where interesting things were often forgotten, and rediscovered a few centuries after. So, in a word, trust the inventors themselves, even a long time after their invention, and always be skeptical about what the community says, about it.

Be said in passing, the modern way of doing theoretical science, which is basically that of constantly hunting for the tons of grants, prizes and other honors available, small or big, has not fixed the problem, quite the opposite. The point indeed is that this system encourages all players to have their work close to the average of the surrounding players, as a matter of getting good evaluations when asking for this and that, and with this having bad effects on both originality, and acceptation of originality. Mathematically, this comes from the following simple fact, related to our theory from chapter 7:

FACT 14.2. *Given a connected graph X, and a function $f : X \to \mathbb{R}$ which is harmonic, that is, subject to the following averaging formula,*

$$f(x) = \frac{1}{|x \sim y|} \sum_{x \sim y} f(y)$$

this function must be constant, $f(x) = c$, for all $x \in X$.

And we will end with this our philosophical discussion about science, stay away from connectivity and harmony. Like in fact Einstein himself did, most of his early work being done away from the academia, while working at the Swiss Patent Office in Bern.

Time now to get to work? According to Plan 14.1, we must first have a discussion about light, including understanding how this appears, from the Maxwell equations from chapter 12, and then commenting on its speed in vacuum c. Let us start with:

FACT 14.3. *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, inside a certain elastic 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, and of foremost importance for physics, the behavior of all the above waves is described by the same equation. In order to understand how this equation appears, let us first have a look at the mechanical waves, which are the simplest ones to study. Regarding them, we have the following result, that we know from chapter 6:

THEOREM 14.4. *The equation of mechanical waves in N dimensions is*

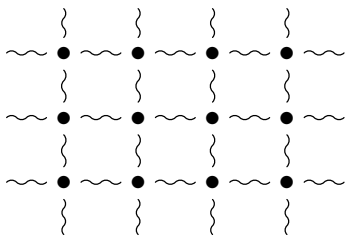
$$\ddot{\varphi} = v^2 \Delta\varphi$$

with $v > 0$ being the propagation speed of the wave, and with Δ given by

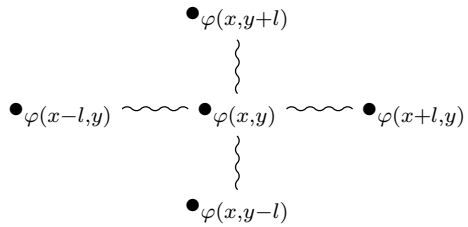
$$\Delta\varphi = \sum_{i=1}^{N} \frac{d^2\varphi}{dx_i^2}$$

being the Laplace operator, playing the role of a numeric second derivative.

PROOF. The standard proof here, that we met in 1D in chapter 4, and then in 2D and higher in chapter 6, is by using a lattice model. Indeed, taking $N = 2$ for simplifying, let us represent the medium by a network of balls and springs, as follows:



Now let us send an impulse, and zoom on one ball. The situation here is as follows, with φ being the position function of the balls, and l being the spring length:



We have 5 forces acting at (x, y), namely the Newton force, mass m times acceleration, and the 4 Hooke forces, displacement of the spring, times spring constant k. Since all these forces must cancel each other, we conclude that the equation of motion is:

$$\begin{aligned} m \cdot \ddot{\varphi}(x,y) &= k(\varphi(x+l,y) - 2\varphi(x,y) + \varphi(x-l,y)) \\ &+ k(\varphi(x,y+l) - 2\varphi(x,y) + \varphi(x,y-l)) \end{aligned}$$

But with this, done, because as explained in chapter 6, by taking the continuum limit of our model, $l \to 0$, we are led via some standard calculus to the wave equationn:

$$\ddot{\varphi}(x,y) = v^2 \Delta\varphi(x,y)$$

Thus, theorem proved in 2D, and the proof in general is similar, via a lattice model. □

In order to talk now about electromagnetic waves, let us recall that we have:

THEOREM 14.5. *Electrodynamics is governed by the formulae*

$$< \nabla, E >= \frac{\rho}{\varepsilon_0} \quad , \quad < \nabla, B >= 0$$

$$\nabla \times E = -\dot{B} \quad , \quad \nabla \times B = \mu_0 J + \mu_0 \varepsilon_0 \dot{E}$$

called Maxwell equations.

PROOF. This is something discussed in chapter 12, the idea being as follows:

(1) To start with, electrodynamics is the science of moving electrical charges. And this is something quite complicated, because unlike in classical mechanics, where the Newton law is good for both the static and the dynamic setting, the Coulomb law is only static, and no longer describes well the situation when the charges are moving.

(2) The problem comes from the fact that moving charges produce magnetism, and with this being visible when putting together two electric wires, which will attract or repel, depending on orientation. Thus, in contrast with classical mechanics, where static or dynamic problems are described by a unique field, the gravitational one, in electrodynamics we have two fields, namely the electric field E, and the magnetic field B.

(3) Fortunately, there is a full set of equations relating E and B, those above. To be more precise, the first formula is the Gauss law, ρ being the charge, and ε_0 being a constant. The second formula is something basic, and anonymous. The third formula is the Faraday law. As for the fourth formula, this is the Ampère law, as modified by Maxwell, with J being the volume current density, and μ_0 being a constant.

(4) Finally, let us mention that the Maxwell equations are something of statistical nature, not applying well to the microscopic setting, say an electron spinning around a proton. But more on such things later, when discussing quantum mechanics. □

As a useful complement now to Theorem 14.5, dealing with the various constants involved, and making a crucial link with the speed of light in vacuum c, we have:

THEOREM 14.6. *The constants involved in the Maxwell equations are the electric permittivity of free space ε_0 and the magnetic permeability of free space μ_0, given by:*

$$\varepsilon_0 = 8.854\ 187\ 8128(13) \times 10^{-12}$$

$$\mu_0 = 1.256\ 637\ 062\ 12(19) \times 10^{-6}$$

These two constants are subject to the Biot-Savart formula, namely

$$\varepsilon_0 \mu_0 = \frac{1}{c^2}$$

where c is the observed speed of light in vacuum, numerically given by

$$c = 299\,792\,458$$

and with this being an exact formula, as per the latest SI regulations.

PROOF. Certainly not a mathematical theorem, what we have here, but with the importance of this being on par with that of the Maxwell equations themselves, we chose to call it so, for symmetry reasons with Theorem 14.5. As for the story with this:

(1) In what regards the electric permittivity of free space ε_0, things quite simple with it, because this comes from the Gauss law from electrostatics, namely:

$$< \nabla, E >= \frac{\rho}{\varepsilon_0}$$

To be more precise, ε_0 appears as a natural rescaling of the Coulomb constant K, in the context of the mathematics of the Gauss law, according to the following formula:

$$K = \frac{1}{4\pi\varepsilon_0}$$

(2) In what regards now the magnetic permeability of free space μ_0, this is trickier, coming from the Biot-Savart formula for the magnetic field of a steady line current:

$$B = \frac{\mu_0}{4\pi} \int \frac{I \times x}{||x||^3}$$

Alternatively, this new constant μ_0 appears in the Ampère law for the curl of the magnetic field, in its original formulation by Ampère himself, namely:

$$\nabla \times B = \mu_0 J$$

And with the story being not over here, because the true occurrence of μ_0 is that in the modification by Maxwell of the Ampère law, which is as follows:

$$\nabla \times B = \mu_0 J + \mu_0 \varepsilon_0 \dot{E}$$

(3) With this discussed, it remains to explain the Biot-Savart formula in the statement, connecting the constants ε_0, μ_0 to the observed speed of light in vacuum c, namely:

$$\varepsilon_0 \mu_0 = \frac{1}{c^2}$$

And here, unfortunately, I'm afraid that I have no simple explanation to offer. This is magic, and the only way of understanding this magic is by rewriting all physics from basic principles, with this being something quite complicated, called quantum field theory, and then recovering the Maxwell equations and the constants involved as something statistical, coming at the macroscopic level from quantum electrodynamics (QED). Tough.

(4) Finally, a word on figures and units. As mentioned in the statement, the formula there for the speed of light in vacuum, which is as follows, is exact:

$$c = 299\,792\,458 \text{ m/s}$$

To be more precise here, the figure $c \simeq 3 \times 10^8$ m/s, coming from various experiments, is something quite old. As an interesting feature, however, as mentioned, the above formula is exact, due to the well-known mess with regulating meters, seconds and other

units. Indeed, there is a rock-solid modern definition for the second, based on atomic clocks, but not for the meter, and by a recent SI decision the above formula was declared exact, and any further changes in the observed value of c will be blamed on the meter.

(5) Be said in passing, would such a wise decision have been taken much earlier, the speed of light would be $c = 3 \times 10^8$ m/s, exactly. But now it is too late. Finally, all these issues do not bother much theoretical physicists, who use tailored units depending on the precise problems they are working on, with around 50 orders of magnitude on the menu, and who usually, when dealing often with c, arrange things as to take $c = 1$.

(6) As a conclusion now to all this, the Maxwell equations come in the form of two statements, Theorem 14.5 dealing with the mathematics, and the present Theorem 14.6 dealing with the physics. And with the physics being, as usual, 100 times more complicated than the mathematics, I mean just have a look at (3) above, that obviously introduces a logical loop in our learning, and we will have to live with that. □

Coming next, we have the following key consequence of the Maxwell equations:

THEOREM 14.7. *In regions of space where there is no charge or current present the Maxwell equations for electrodynamics read*

$$< \nabla, E >=< \nabla, B >= 0$$

$$\nabla \times E = -\dot{B} \quad , \quad \nabla \times B = \dot{E}/c^2$$

and both the electric field E and magnetic field B are subject to the wave equation

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

where $\Delta = \sum_i d^2/dx_i^2$ is the Laplace operator, and c is the speed of light.

PROOF. Under the circumstances in the statement, namely no charge or current present, with this meaning $\rho = 0$ and $J = 0$, we obtain indeed the equations in the statement. Next, by applying the curl operator to the last two equations, we obtain:

$$\nabla \times (\nabla \times E) = -\nabla \times \dot{B} = -(\nabla \times B)' = -\ddot{E}/c^2$$

$$\nabla \times (\nabla \times B) = \nabla \times \dot{E}/c^2 = (\nabla \times E)'/c^2 = -\ddot{B}/c^2$$

Now observe that the double curl operator is subject to the following formula:

$$\nabla \times (\nabla \times \varphi) = \nabla < \nabla, \varphi > -\Delta \varphi$$

But in our case, $\varphi = E, B$, by using the first two Maxwell equations, this reads:

$$\nabla \times (\nabla \times \varphi) = -\Delta \varphi$$

Thus, both E and B are subject to the wave equation of speed c, as stated. □

So, what is light? Light is the wave predicted by Theorem 14.7, traveling in vacuum at the maximum possible speed, c, its creation mechanism being as follows:

FACT 14.8. *An accelerating or decelerating charge produces electromagnetic radiation, subject in vacuum to the wave equation* $\ddot{\varphi} = c^2\Delta\varphi$*, called light.*

As explained in chapter 6, this phenomenon can be observed is a variety of situations, such as in light bulbs, where the electrons get decelerated by the filament, acting as a resistor, or in fire, which is a chemical reaction, with the electrons moving around. In general, Fact 14.8 is something that can be deduced from the Maxwell equations. However, the computations, that we will not really need in what follows, are a bit complicated, and we refer here to any standard book on electrodynamics, such as Griffiths [44].

Moving on, let us record as well the following version of Theorem 14.7:

THEOREM 14.9. *Inside a linear, homogeneous medium, where there is no free charge or free current present, both fields* E *and* B *are subject to the wave equation*

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

with $v < c$ *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 Theorem 14.7, and the proof in general is identical, with suitable definitions for ε, μ, the speed being:

$$v = \frac{1}{\sqrt{\varepsilon\mu}}$$

But this formula can be written is a more familiar from, as above. □

And with this, end of our discussion about light. Good learning this was, although still containing a logical loop, in relation with Theorem 14.6, and a technical gap, in relation with Fact 14.8. Well, as said time and again, such things are complicated, modesty.

Finally, for other aspects of light, such as color, we refer to the material in chapter 6. In what follows next, in relation with relativity, we will not need color.

14b. Questions, answers

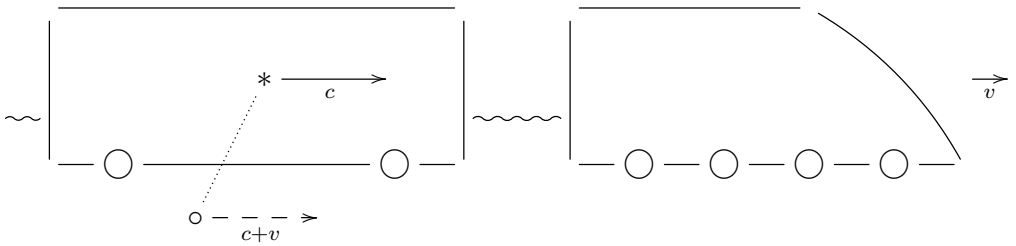
Getting now to the point where we wanted to get, speed of light in vacuum c and its philosophy, many things can be said. To start with, we know from the above that we have $c < \infty$, and more specifically $c \simeq 3 \times 10^8$, and with this drastically improving the previous figure, $c = \infty$, from the ancient times. Not bad, as a start.

Next comes the observation, or rather feeling, that nothing can travel faster than light, $v \not> c$. Which is something intuitive, in tune with what our senses feel, although you can still argue that God's thought most likely travels at $v = \infty$, contradicting this.

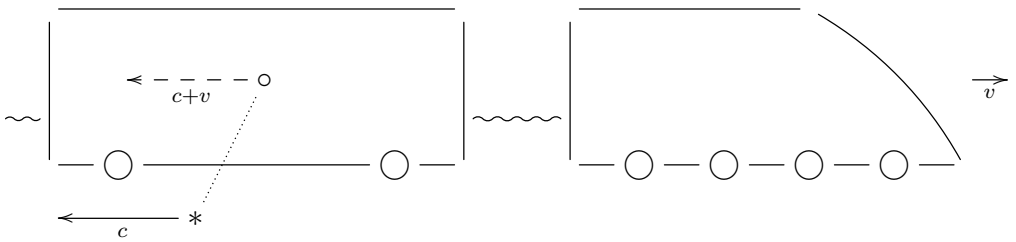
Well, in relation with this, think at time t. This is the slow and down-to-earth beast needed for a room to warm in the Summer, when opening a window, or for the same room to cool in the Winter, again when opening a window. So, nothing much divine about time t, this is a vulgar quantity, not bothering God in his thoughts. So, no contradiction.

With this said, what is next? Well, next comes the relativity puzzle:

PUZZLE 14.10. *Assuming 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 traveling at speed $c + v > c$, which is a contradiction:*



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



And the problem is, how to fix the laws of motion, as to avoid these contradictions?

So, let us get now into this. As a first observation, in order to get to a finer understanding of the problem, we must talk about inertial frames. With the convention that such a frame is one which is not subject to acceleration, like for instance are both the train and ground frames in the above considerations, our puzzle becomes:

PUZZLE 14.11 (update). *The speed of light in vacuum c must be the same for all inertial observers. So, we should invent a theory where $c + v = c$, for any $v < c$.*

Which sounds quite good, our knowledge has evolved, instead of having a contradiction to be understood, as in Puzzle 14.10, we have now a problem to be solved. Nice.

This being said, before going further with our study, let us meditate a bit on the inertial frames, which are quite tricky objects. Indeed, in the absence of anything, in this

universe, any frame is of course inertial. However, add one particle M to this universe, and with you being a second particle m, is your frame inertial? Certainly not, because you can feel the attraction of M, I mean even if not knowing about M, if you have a cup of coffee, that will spill, which is a clear sign that your frame is not inertial.

As explained in chapter 10, all this leads to some interesting mathematics, and equations to be solved, with the conclusion that, in the above context of the gravitational 2-body problem, an inertial frame exists indeed. However, as also explained in chapter 10, in the context of the gravitational k-body problem with $k \geq 3$, and with $k >> 0$ corresponding to our usual universe, by leaving aside other forces than gravity, the mathematics becomes substantially more complicated, and the answer is unclear.

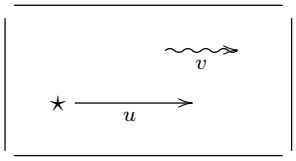
So, are there any inertial frames at all, in our universe? Not clear, and I will have to ask the cats, I guess. And here, going as usual for my good old lady, she's really old and small, like these 5-feet grandmas from the times, here is what her grumpiness says:

CAT 14.12. *We cats are sensitive and fast, but not that sensitive and fast, especially at my age. No idea about this inertial frame business, young man.*

Well, thanks cat, but this does not fit me at all, what is my dudeness supposed to do, with this. Guess we will add this to the list of bugs present in this chapter, coming as item #3, after the troubles with Biot-Savart, #1, and with the light creation, #2.

Moving on, with Puzzle 14.11 being now a bit dented by Cat 14.12, what about having something new as a starting point, for relativity theory? And here, following Einstein's book [28], we have indeed the following remarkable finding, by Fizeau:

FACT 14.13 (Fizeau, 1851). *Assume that light moves through a liquid at speed $u < c$. Then, when this liquid moves through a tube at speed $v > 0$,*



the observed speed of light is not the Galilean $u +_g v = u + v$, but rather

$$u +_f v = u + v\left(1 - \frac{1}{n^2}\right)$$

where $n = c/u$ is the index of refraction of the liquid.

And good input this is, more or less leading to the same question as before, namely inventing some new mathematics for $c + v = c$, as asked in Puzzle 14.11, but this time avoiding the trains in Puzzle 14.10, and the subtleties of inertial frames, in general.

So, let us reformulate our puzzle as follows, and in $c = 1$ units, for simplifying:

PUZZLE 14.14 (update). *How to define an addition operation $u +_e v$ on the space of 1D speeds, which is the interval $[-1, 1]$, with the $c = 1$ convention, as to have:*

(1) $u +_e v \simeq u + v$ *at low speeds, as required by Galileo.*
(2) $u +_e v \simeq u + v(1 - u^2)$ *for $v << u < 1$, as required by Fizeau.*
(3) $1 +_e v = 1$ *for any $v < 1$, as suggested by bulbs in moving trains.*

And with this, good news, end of physics, we are into mathematics. In answer now, fortunately, our puzzle has a simple solution coming from algebra. Obviously, what we have to do with $u + v$ is to divide it by a factor ≥ 1, that we can call $1 + r$:

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

But then our Galileo and train requirements in Puzzle 14.14 read, in terms of r:

$$\begin{cases} u, v \simeq 0 & \implies \quad r \simeq 0 \\ u = 1 & \implies \quad r = v \end{cases}$$

And at this point, it is a no-brainer to take $r = uv$, and see if this works, meaning compatibility with Fizeau. And, by some kind of miracle, this works indeed:

THEOREM 14.15. *If we sum the speeds according to the Einstein formula*

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

in $c = 1$ units, then the following happen:

(1) $u +_e v \simeq u + v$ *at low speeds, as required by Galileo.*
(2) $u +_e v \simeq u + v(1 - u^2)$ *for $v << u < 1$, as required by Fizeau.*
(3) $1 +_e v = 1$ *for any $v < 1$, as suggested by bulbs in moving trains.*

PROOF. This is clear from the above discussion, with the only needed computations being for (2). But here, by using our assumption $v << u < 1$, we have indeed:

$$\begin{aligned} u +_e v &= \frac{u + v}{1 + uv} \\ &\simeq (u + v)(1 - uv) \\ &= u + v - u^2 v - uv^2 \\ &\simeq u + v - u^2 v \\ &= u + v(1 - u^2) \end{aligned}$$

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

Next, in relation with geometry, we have the following interesting observation:

THEOREM 14.16. *The Einstein speed summation in 1D is given by*

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

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

PROOF. 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, as you might have heard 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 14.15. So, we have an idea here, and in practice, we have:

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

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

As a conclusion to all this, speed addition formula well understood in 1D. And with this formula being the correct one, compatible with other physics experiments as well. For details here, full story with all this, have a look at the book of Einstein [28].

14c. Lorentz transform

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. And the result here, which might seem quite surprising, is as follows:

THEOREM 14.17. *Time and length are subject to Lorentz dilation/contraction*

$$t \to \gamma t \quad , \quad L \to L/\gamma$$

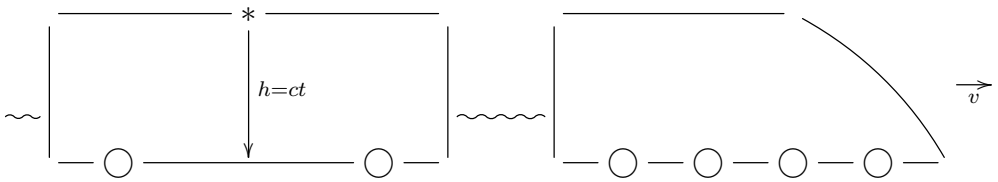
where the number $\gamma \geq 1$*, called Lorentz factor, is given by the formula*

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}$$

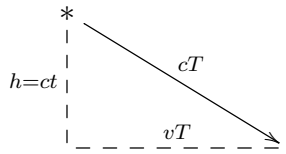
with v *being the moving speed, at which time and length are measured.*

PROOF. This is something that we know from chapter 6, and which is not exactly 1D, because based on the Pythagoras theorem, which is 2D, the idea being as follows:

(1) Regarding time, assume that we have a train, moving to the right with speed v, through vacuum. In order to compute the height h of the train, the passenger onboard switches on the ceiling light bulb, measures the time t that the light needs to hit the floor, by traveling at speed c, and concludes that the train height is $h = ct$:



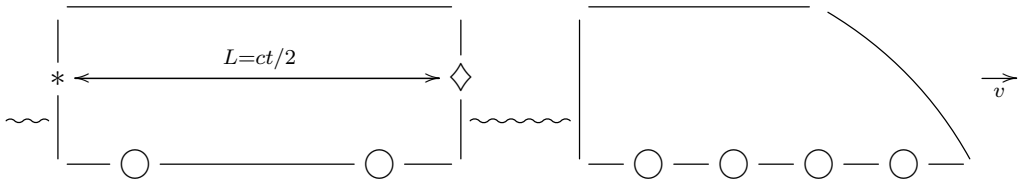
But, an observer on the ground will see here something different, namely a right triangle, as follows, with T being the duration of the event, according to his watch:



Now by Pythagoras applied to this latter triangle, we obtain, as claimed:

$$(ct)^2 + (vT)^2 = (cT)^2 \implies T = \frac{t}{\sqrt{1 - v^2/c^2}}$$

(2) In the same train as before, 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:



Now viewed from the ground, the duration of the event is $T = T_1 + T_2$, where $T_1 > T_2$ are the time needed for the light to travel forward, and travel back. 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:

$$l = \frac{\gamma L(c^2 - v^2)}{c^2} = \gamma L\left(1 - \frac{v^2}{c^2}\right) = \frac{\gamma L}{\gamma^2} = \frac{L}{\gamma}$$

We are therefore led to the length contraction conclusion in the statement. □

As a comment here, there is a bug in the above proof of (2), coming from the use of speeds $c \pm v$, which are forbidden, I mean both being equal to c. No idea who invented this damn thing, blame Feynman I guess for such things, although not sure for this one, but since I find this lovely, I used this here in this book, like everyone else does.

As a fix, the standard way, which is however less elegant, is to use time dilation, that was rigorously established in (1), in order to talk about the Lorentz transform, as we will actually do next, and deduce the length contraction by using this Lorentz transform.

Which brings us into the philosophy of science, and the meaning and value of elegance, I mean to us humans, because God certainly knows what that means. And here, always good to remember the following famous quote, by Hermann Weyl:

Weyl 14.18. *Among the correct and the beautiful, I always chose the beautiful.*

And we will leave things like this, theorem officialy proved, but making appear a bug #4 in this chapter, after #1 Biot-Savart, #2 light creation, and #3 inertial frames. And with this latest bug being actually relatively easy to fix, as indicated above.

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 usual $t' = t$. In the relativistic setting now, things are more tricky, as follows:

THEOREM 14.19. *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 should use here the reverse Lorentz transformation, given by the following formulae:

$$x = \gamma(x' + vt')$$

$$y = y'$$

$$z = z'$$

Indeed, by using our previous formula for x' we can compute t', and we obtain:

$$t' = \frac{x - \gamma x'}{\gamma v} = \frac{x - \gamma^2(x - vt)}{\gamma v} = \frac{\gamma^2 vt + (1 - \gamma^2)x}{\gamma v}$$

On the other hand, we have the following computation, regarding γ:

$$\gamma^2 = \frac{c^2}{c^2 - v^2} \implies \gamma^2(c^2 - v^2) = c^2 \implies (\gamma^2 - 1)c^2 = \gamma^2 v^2$$

Thus we can finish the computation of t' as follows:

$$t' = \frac{\gamma^2 vt + (1 - \gamma^2)x}{\gamma v} = \frac{\gamma^2 vt - \gamma^2 v^2 x/c^2}{\gamma v} = \gamma\left(t - \frac{vx}{c^2}\right)$$

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

As a continuation now, since y, z are irrelevant, it is better to put them at the end, and also put time t first, as to be close to x. In linear algebra terms, we are led to:

THEOREM 14.20. *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. With the above manipulations done, our system can be written as:

$$ct' = \gamma(ct - vx/c)$$
$$x' = \gamma(x - vt)$$
$$y' = y$$
$$z' = z$$

Now with $\beta = v/c$, replacing v, we are led to the formula in the statement. □

As an illustration, let us verify that the inverse Lorentz transformation appears indeed by reversing the speed, $v \to -v$. With notations as in Theorem 14.19, the result is:

THEOREM 14.21. *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 14.20, reversing the speed $v \to -v$ amounts in reversing $\beta \to -\beta$. 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. □

Moving on, let us discuss now the general frame change formula, meaning with the speed v no longer assumed to be along the x axis. In the classical case, we have:

THEOREM 14.22. *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 14.23. *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 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, that 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. □

What we found in Theorem 14.23 can be reformulated as follows:

THEOREM 14.24. *In the relativistic case, the general frame change formula is*

$$x_1' = x_1 + (\gamma - 1)\frac{< v, x > v_1}{||v||^2} - \gamma t v_1$$

$$\vdots$$

$$x_N' = x_N + (\gamma - 1)\frac{< v, x > v_N}{||v||^2} - \gamma t v_N$$

$$ct' = \gamma\left(ct - \frac{< v, x >}{c}\right)$$

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 14.23, 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 14.25. *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 14.24, at $N = 3$, and with the rescaled time variable being put in the first position. □

14d. Curved spacetime

Time to get back to physics, with the Lorentz transformation being our main weapon. There is a lot of physics happening with $\mathbb{R}^4$, and we will start of course with plain motion, which is the simplest to examine. As a continuation of our work above, we have:

THEOREM 14.26. *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. We know the above formula from chapter 6, deduced there via some speculations, and time now to have a more conceptual look at this, as follows:

(1) 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. □

As a comment here, as explained in chapter 10, in the case $N = 3$, where the vector product operation $\times$ is available, an equivalent formula is as follows:

$$u +_e v = \frac{1}{1+ < u, v >}\left(u + v + \frac{u \times (u \times v)}{1 + \sqrt{1 - ||u||^2}}\right)$$

Moving on with more physics in $\mathbb{R}^4$, we have the following key result regarding electromagnetism, which was found by Lorentz himself, using his transform, originally designed precisely for this purpose, some time before Einstein's relativity theory:

THEOREM 14.27. *The Maxwell equations are invariant under Lorentz transformations*

$$x' = \gamma(x - vt)$$
$$y' = y$$
$$z' = z$$
$$t' = \gamma(t - vx/c^2)$$

with $\gamma = 1/\sqrt{1 - v^2/c^2}$ being as usual the Lorentz factor.

PROOF. Consider indeed an electromagnetic field (E, B). This is altered by a Lorentz transformation into a field (E', B'), the equations for E' being as follows:

$$E'_x = E_x$$
$$E'_y = \gamma(E_y - vB_z)$$
$$E'_z = \gamma(E_z + vB_y)$$

As for the equations of B', these are quite similar, as follows:

$$B'_x = B_x$$
$$B'_y = \gamma\left(B_y + \frac{v}{c^2}E_z\right)$$
$$B'_z = \gamma\left(B_z - \frac{v}{c^2}E_y\right)$$

In order to do the math, consider the following matrices, with $\beta = v/c$ as usual:

$$D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \gamma & 0 \\ 0 & 0 & \gamma \end{pmatrix} \quad , \quad M = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\beta\gamma \\ 0 & \beta\gamma & 0 \end{pmatrix}$$

In terms of these matrices, the formulae for the new field (E', B') read:

$$E' = DE + cMB$$
$$B' = DB - \frac{M}{c}E$$

Which is not that bad, and starting from these formulae, one can prove that (E', B') satisfies as well the Maxwell equations, as desired, say good exercise for you. □

Moving on, as another piece of physics happening in $\mathbb{R}^4$, we have everything fundamental in relation with mass, momentum and energy. And the result here, which is famous, due to Einstein, that we already know since chapter 6, is as follows:

THEOREM 14.28. *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. The relativistic energy of an object of rest mass $m > 0$,

$$\mathcal{E} = Mc^2$$

which is conserved too, 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. We know this chapter 6, but as always with $E = mc^2$, good to talk about it again. In the classical case, the conservation equations for plastic collisions are:

$$m = m_1 + m_2 \quad , \quad mv = m_1v_1 + m_2v_2$$

The point now is that, as experimental physics shows, these equations are in fact the low-speed approximations of the correct, relativistic equations, which are as follows:

$$M = M_1 + M_2 \quad , \quad Mv = M_1v_1 + M_2v_2$$

Regarding now energy, it makes sense to look at $\mathcal{E} = Mc^2$, which is given by:

$$\mathcal{E} = \frac{mc^2}{\sqrt{1 - v^2/c^2}} \simeq mc^2\left(1 + \frac{v^2}{2c^2}\right) = mc^2 + \frac{mv^2}{2}$$

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

Getting now to some philosophy, in non-relativistic physics two events are separated by space Δx and 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 14.29. *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, describing K, 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. The result here, which is quite subtle, is as follows:

THEOREM 14.30 (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, as discussed in Theorem 14.29, and by having the $F = \dot{P}$ idea from (1) in mind too.

(6) So, this was the main idea behind general relativity, and in practice, all this needs a bit of technical geometry and formulae. For more on all this, we refer to [28]. □

14e. Exercises

Nice chapter that we had here, and as exercises on all this, we have:

EXERCISE 14.31. *Further meditate on constants, units, and* $\varepsilon_0\mu_0 = 1/c^2$.

EXERCISE 14.32. *Learn more about accelerating charges producing light.*

EXERCISE 14.33. *Learn more about light, including color, polarization, optics.*

EXERCISE 14.34. *Have some work done, on the inertial frame question.*

EXERCISE 14.35. *Learn more about the hyperbolic trigonometric functions.*

EXERCISE 14.36. *Fix the bug, in the proof of the Lorentz contraction formula.*

EXERCISE 14.37. *Finish the work, for the invariance of the Maxwell equations.*

EXERCISE 14.38. *Learn more about geometry, and general relativity.*

As bonus exercise, and no surprise here, read Einstein's book [28].

CHAPTER 15

Infinite matrices

15a. Infinite matrices

With this book approaching its end, we still have a difficult task ahead, namely talking about quantum mechanics. And here, based on the notorious fact that electrons, and other particles appearing there, are of "slippery" nature, having no exact positions and speeds, it is pretty much clear that, as mathematics that we need to develop, we need infinite dimensional tools. More specifically, we would like for instance to have an infinite dimensional generalization of the linear algebra theory developed in chapter 13.

In order to get started, let us formulate the following key definition:

DEFINITION 15.1. *A Hilbert space is a complex vector space H with a scalar product $< x, y >$, taken linear at left and antilinear at right,*

$$< \lambda x, y >= \lambda < x, y > \quad , \quad < x, \lambda y >= \bar{\lambda} < x, y >$$

which is complete with respect to corresponding norm

$$||x|| = \sqrt{< x, x >}$$

in the sense that any sequence $\{x_n\}$ which is a Cauchy sequence, having the property $||x_n - x_m|| \to 0$ with $n, m \to \infty$, has a limit, $x_n \to x$.

Here the fact that $||x|| = \sqrt{< x, x >}$ is indeed a norm, satisfying $||x+y|| \leq ||x|| + ||y||$, follows from Cauchy-Schwarz, $| < x, y > | \leq ||x|| \cdot ||y||$, which itself follows as in the $H = \mathbb{C}^N$ case, by looking at $f(t) = ||wx + ty||^2$, as explained before. At the level of examples of Hilbert spaces, generalizing the basic $H = \mathbb{C}^N$ example, we have:

THEOREM 15.2. *Given an index set I, which can be finite or not, the space of square-summable vectors having indices in I, namely*

$$l^2(I) = \left\{ (x_i)_{i \in I} \Big| \sum_i |x_i|^2 < \infty \right\}$$

is a Hilbert space, with scalar product as follows:

$$< x, y >= \sum_i x_i \bar{y}_i$$

When I is finite, $I = \{1, \ldots, N\}$, we obtain in this way the usual space $H = \mathbb{C}^N$.

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

(1) We know that $l^2(I) \subset \mathbb{C}^I$ is the space of vectors satisfying $||x|| < \infty$. We want to prove that $l^2(I)$ is a vector space, that $< x, y >$ is a scalar product on it, that $l^2(I)$ is complete with respect to $||.||$, and finally that for $|I| < \infty$ we have $l^2(I) = \mathbb{C}^{|I|}$.

(2) The last assertion, $l^2(I) = \mathbb{C}^{|I|}$ for $|I| < \infty$, is clear, because in this case the sums are finite, so the condition $||x|| < \infty$ is automatic. So, we know at least one thing.

(3) Regarding the rest, we have Cauchy-Schwarz, $| < x, y > | \leq ||x|| \cdot ||y||$, proved as usual, by looking at $f(t) = ||wx + ty||^2$, and we deduce that we have $||x + y|| \leq ||x|| + ||y||$. Thus $l^2(I)$ is indeed a vector space, the other vector space conditions being trivial.

(4) Also, the quantity $< x, y >$ is surely a scalar product on this vector space, because all the conditions for a scalar product are trivially satisfied.

(5) Finally, the fact that our space $l^2(I)$ is indeed complete with respect to its norm $||.||$ follows in the obvious way, the limit of a Cauchy sequence $\{x_n\}$ being the vector $y = (y_i)$ given by $y_i = \lim_{n \to \infty} x_{ni}$, with all the verifications here being trivial. □

Going now a bit more abstract, we have, more generally, the following result:

THEOREM 15.3. *Given an arbitrary space X with a positive measure μ on it, the space of square-summable complex functions on it, namely*

$$L^2(X) = \left\{ f : X \to \mathbb{C} \Big| \int_X |f(x)|^2 \, d\mu(x) < \infty \right\}$$

is a Hilbert space, with scalar product as follows:

$$< f, g > = \int_X f(x) \overline{g(x)} \, d\mu(x)$$

When $X = I$ is discrete, meaning that the measure μ on it is the counting measure, $\mu(\{x\}) = 1$ for any $x \in X$, we obtain in this way the previous spaces $l^2(I)$.

PROOF. This is something routine, remake of Theorem 15.2, as follows:

(1) The proof of the first, and main assertion is something perfectly similar to the proof of Theorem 15.2, by replacing everywhere the sums by integrals.

(2) With the remark that I forgot to say in the statement that the L^2 functions are by definition taken up to equality almost everywhere, $f = g$ when $||f - g|| = 0$.

(3) As for the last assertion, when μ is the counting measure all our integrals here become usual sums, and so we recover in this way Theorem 15.2. □

As a third and last theorem about Hilbert spaces, that we will need, we have:

THEOREM 15.4. *Any Hilbert space H has an orthonormal basis $\{e_i\}_{i\in I}$, which is by definition a set of vectors whose span is dense in H, and which satisfy*

$$< e_i, e_j >= \delta_{ij}$$

with δ being a Kronecker symbol. The cardinality $|I|$ of the index set, which can be finite, countable, or worse, depends only on H, and is called dimension of H. We have

$$H \simeq l^2(I)$$

in the obvious way, mapping $\sum \lambda_i e_i \to (\lambda_i)$. The Hilbert spaces with $\dim H = |I|$ being countable, including $l^2(\mathbb{N})$ and $L^2(\mathbb{R})$, are all isomorphic, and are called separable.

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

(1) In finite dimensions an orthonormal basis $\{e_i\}_{i\in I}$ can be constructed by starting with any vector space basis $\{x_i\}_{i\in I}$, and using the well-known Gram-Schmidt procedure. As for the other assertions, these are all clear, from basic linear algebra.

(2) In general, the same method works, namely Gram-Schmidt, with a subtlety coming from the fact that the basis $\{e_i\}_{i\in I}$ will not span in general the whole H, but just a dense subspace of it, as it is in fact obvious by looking at the standard basis of $l^2(\mathbb{N})$.

(3) And there is a second subtlety as well, coming from the fact that the recurrence procedure needed for Gram-Schmidt must be replaced by some sort of "transfinite recurrence", using scary tools from logic, and more specifically the Zorn lemma.

(4) Finally, everything at the end is clear from definitions, except perhaps for the fact that $L^2(\mathbb{R})$ is separable. But here we can argue that, since functions can be approximated by polynomials, we have a countable algebraic basis, namely $\{x^n\}_{n\in\mathbb{N}}$, called the Weierstrass basis, that we can orthogonalize afterwards by using Gram-Schmidt. □

Moving ahead, now that we know what our vector spaces are, we can talk about infinite matrices with respect to them. And the situation here is as follows:

THEOREM 15.5. *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

$$B(H) \subset M_I(\mathbb{C})$$

with the correspondence $T \to M$ coming via the usual linear algebra formulae, namely:

$$T(x) = Mx \quad , \quad M_{ij} =< Te_j, e_i >$$

In infinite dimensions, the inclusion $B(H) \subset M_I(\mathbb{C})$ is not 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 $x \to Tx$ is linear.

(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 embedding, 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:

$$B(H) \subset M_I(\mathbb{C})$$

In finite dimensions we have an isomorphism, because any matrix $M \in M_N(\mathbb{C})$ determines a linear operator $T : \mathbb{C}^N \to \mathbb{C}^N$, given by the following formula:

$$< Te_j, e_i >= M_{ij}$$

However, in infinite dimensions we have matrices not producing operators, as for instance the all-one matrix, so the embedding $B(H) \subset M_I(\mathbb{C})$ is not an isomorphism. □

Finally, as a second and last basic result regarding the operators, we will need:

THEOREM 15.6. *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. □

15b. Spectral radius

Let us discuss now the diagonalization problem for the operators $T \in B(H)$, in analogy with the diagonalization problem for the usual matrices $A \in M_N(\mathbb{C})$. We first have:

DEFINITION 15.7. *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 15.8. *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 $0 \in \sigma(T) - \varepsilon(T)$. □

The above result might look quite surprising, and philosophically, the best way of thinking at all 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, where some things are upside-down.

As a second basic result about the spectrum, which is of great use, we have:

PROPOSITION 15.9. *The spectrum of an operator $T \in B(H)$ satisfies:*

$$\sigma(T) \subset D_0(||T||))$$

In other words, the spectral radius $\rho(T) = \sup_{\lambda \in \sigma(T)} |\lambda|$ satisfies $\rho(T) \leq ||T||$.

PROOF. We use the standard fact that for $||S|| < 1$, we have the following formula:

$$(1 - S)^{-1} = 1 + S + S^2 + \ldots$$

Indeed, by using this inversion formula, we have the following computation:

$$\begin{aligned}\lambda > ||T|| &\implies \left|\left|\frac{T}{\lambda}\right|\right| < 1\\ &\implies 1-\frac{T}{\lambda} \in B(H)^{-1}\\ &\implies \lambda - T \in B(H)^{-1}\\ &\implies \lambda \notin \sigma(T)\end{aligned}$$

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

Let us develop now some more general theory. Here is a key result about products:

THEOREM 15.10. *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) Let us first prove the main assertion, stating that the sets $\sigma(ST), \sigma(TS)$ 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 $RST = R - 1$, and by using this, we have the following computation:

$$\begin{aligned}(1 + TRS)(1 - TS) &= 1 + TRS - TS - TRSTS\\ &= 1 + TRS - TS - TRS + TS\\ &= 1\end{aligned}$$

Similarly, $(1 - TS)(1 + TRS) = 1$. Thus $1 - TS$ is invertible, proving our claim. Now by multiplying by scalars, we deduce that for any $\lambda \in \mathbb{C} - \{0\}$ we have, as desired:

$$\lambda \notin \sigma(ST) \implies \lambda \notin \sigma(TS)$$

(2) 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, which is the following operator:

$$S^*(e_i) = \begin{cases} e_{i-1} & \text{if } i > 0\\ 0 & \text{if } i = 0\end{cases}$$

Let us compose now these two operators. In one sense, we have:

$$S^*S = 1 \implies 0 \notin \sigma(SS^*)$$

In the other sense, however, the situation is different, as follows:

$$SS^* = Proj(e_0^\perp) \implies 0 \in \sigma(SS^*)$$

Thus, the spectra do not match on 0, and we have our counterexample, as desired. □

As our next result, inspired from the theory from before, we have:

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. This can be proved in two steps, as follows:

(1) Assume first that our rational function $f \in \mathbb{C}(X)$ is a usual polynomial $P \in \mathbb{C}[X]$. We pick a scalar $\lambda \in \mathbb{C}$, and we decompose the polynomial $P - \lambda$, as follows:

$$P(X) - \lambda = c(X - r_1) \ldots (X - r_n)$$

We have then the following equivalences, which give the result:

$$\begin{aligned}
\lambda \notin \sigma(P(T)) &\iff P(T) - \lambda \in B(H)^{-1} \\
&\iff c(T - r_1) \ldots (T - r_n) \in B(H)^{-1} \\
&\iff T - r_1, \ldots, T - r_n \in B(H)^{-1} \\
&\iff r_1, \ldots, r_n \notin \sigma(T) \\
&\iff \lambda \notin P(\sigma(T))
\end{aligned}$$

(2) In general now, we pick a scalar $\lambda \in \mathbb{C}$, we write $f = P/Q$, and we set $F = P - \lambda Q$. By using what we found in (1), for this polynomial $F \in \mathbb{C}[X]$, 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 a first application of the above methods, we have the following key result:

THEOREM 15.12. *The following happen:*

(1) *For a unitary operator, $U^* = U^{-1}$, we have $\sigma(U) \subset \mathbb{T}$.*
(2) *For a self-adjoint operator, $T = T^*$, we have $\sigma(T) \subset \mathbb{R}$.*

PROOF. This is something quite tricky, based on Theorem 15.11, as follows:

(1) 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, using Proposition 15.9:

$$\sigma(U) \subset D$$

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.

(2) Consider the following rational function, depending on a parameter $r \in \mathbb{R}$:

$$f(z) = \frac{z + ir}{z - ir}$$

Then 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 (1) we have $\sigma(T) \subset f^{-1}(\mathbb{T}) = \mathbb{R}$, as desired. □

Next, we have the following key result, which is something quite far-reaching:

THEOREM 15.13. *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:

$$u_n \leq \frac{1}{n}(lmu_{lm} + ru_r) \leq u_m + \frac{r}{n}\, u_1$$

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 the following estimate:

$$||T^{n-1}|| \leq \frac{1}{2\pi}\int_{|z|=\rho} \left|\left|\frac{z^n}{z-T}\right|\right| dz \leq \rho^n \cdot \sup_{|z|=\rho} \left|\left|\frac{1}{z-T}\right|\right|$$

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 15.14. *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 $n = 2^k$, 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 15.14 we can say a number of non-trivial things about the normal operators, commonly known as "spectral theorem for normal operators". We first have:

THEOREM 15.15. *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 15.11 for polynomials, 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. □

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

THEOREM 15.16. *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 15.15. Indeed, by Stone-Weierstrass, that morphism has 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. □

Even more generally now, we have the following result:

THEOREM 15.17. *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:

(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:

$$< S(x+y), x+y >=< Sx, x > + < Sy, y > + < Sx, y > + < Sy, x >$$

By replacing $y \to iy$, we have as well the following formula:

$$< S(x+iy), x+iy >=< Sx, x > + < Sy, y > -i < Sx, y > +i < Sy, x >$$

By multiplying this formula by i, and summing with the first one, we obtain:

$$\begin{aligned} < S(x+y), x+y > +i < S(x+iy), x+iy > &= (1+i)[< Sx, x > + < Sy, y >] \\ &+ 2 < Sx, y > \end{aligned}$$

(3) But with this, we can 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 we obtain in this way a certain operator $f(T) \in B(H)$, having all the desired properties. □

Good news, we can now diagonalize the normal operators. Let us start with:

THEOREM 15.18. *Any self-adjoint 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 $f : X \to \mathbb{R}$ 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:

$$\overline{span\left(T^k x \middle| n \in \mathbb{N}\right)} = H$$

For this purpose, let us go back to the proof of Theorem 15.17. 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. □

We have the following technical generalization of the above result:

THEOREM 15.19. *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 15.18, by suitably modifying the measurable calculus formula, and the measure μ itself, as to have this formula working for all the operators T_i. With this modification done, everything extends. □

We can now discuss the case of the arbitrary normal operators, as follows:

THEOREM 15.20. *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. Consider the decomposition of T into its real and imaginary parts:

$$T = \frac{T + T^*}{2} + i \cdot \frac{T - T^*}{2i}$$

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 15.19 applies to these real and imaginary parts, and gives the result. □

Getting now to applications, the above results are quite powerful, and many things can be said, in analogy with what we know about usual matrices. Let us record here:

THEOREM 15.21. *Any bounded operator $T \in B(H)$ can be decomposed as*

$$T = U|T|$$

with U being a partial isometry, and with $|T| = \sqrt{T^*T}$.

PROOF. The operator T^*T being self-adjoint, and even positive, in the sense that we have $< T^*Tx, x > \geq 0$ for any $x \in H$, we can extract its square root $|T| = \sqrt{T^*T}$, by using the spectral theorem. Now observe that we have the following formula:

$$\begin{aligned} < |T|x, |T|y > &= < x, |T|^2 y > \\ &= < x, T^*Ty > \\ &= < Tx, Ty > \end{aligned}$$

We conclude that the following linear application is well-defined, and isometric:

$$U : Im|T| \to Im(T) \quad , \quad |T|x \to Tx$$

Now by continuity we can extend this isometry U into an isometry between certain Hilbert subspaces of H, as follows:

$$U : \overline{Im|T|} \to \overline{Im(T)} \quad , \quad |T|x \to Tx$$

Moreover, we can further extend U into a partial isometry $U : H \to H$, by setting $Ux = 0$, for any $x \in \overline{Im|T|}^{\perp}$, and with this convention, the result follows. □

15d. Compact operators

We will restrict now the attention to the compact operators, which share many properties with the usual matrices. Let us start with a basic definition, as follows:

DEFINITION 15.22. *An operator $T \in B(H)$ is said to be of finite rank if its image*

$$Im(T) \subset H$$

is finite dimensional. The set of such operators is denoted $F(H)$.

There are many interesting examples of finite rank operators, the most basic ones being the finite rank projections, on the finite dimensional subspaces $K \subset H$. We have:

PROPOSITION 15.23. *The set of finite rank operators*

$$F(H) \subset B(H)$$

is a two-sided $*$-ideal.

PROOF. It is clear that $F(H)$ is a vector space. Let us prove now that $F(H)$ is stable under $*$. Given $T \in F(H)$, we can regard it as an invertible operator, as follows:

$$T : (\ker T)^\perp \to Im(T)$$

We conclude that we have the following dimension equality:

$$\dim((\ker T)^\perp) = \dim(Im(T))$$

On the other hand, we have equalities as follows, which give the result:

$$\begin{aligned} \dim(Im(T^*)) &= \dim(\overline{Im(T^*)}) \\ &= \dim((\ker T)^\perp) \\ &= \dim(Im(T)) \end{aligned}$$

Regarding now the ideal property, this follows from the following two formulae, valid for any $S, T \in B(H)$, which are once again clear from definitions:

$$\dim(Im(ST)) \leq \dim(Im(T))$$

$$\dim(Im(TS)) \leq \dim(Im(T))$$

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

Let us discuss now the compact operators. These are introduced as follows:

DEFINITION 15.24. *An operator $T \in B(H)$ is said to be compact if the closed set*

$$\overline{T(B_1)} \subset H$$

is compact, where $B_1 \subset H$ is the unit ball. The set of such operators is denoted $K(H)$.

In finite dimensions any operator is compact. In general, as a first observation, any finite rank operator is compact. We have in fact the following result:

PROPOSITION 15.25. *Any finite rank operator is compact,*

$$F(H) \subset K(H)$$

and the finite rank operators are dense inside the compact operators.

PROOF. The first assertion is clear, because if $Im(T)$ is finite dimensional, then the following subset is closed and bounded, and so it is compact:

$$\overline{T(B_1)} \subset Im(T)$$

Regarding the second assertion, let us pick a compact operator $T \in K(H)$, and a number $\varepsilon > 0$. By compactness of T we can find a finite set $S \subset B_1$ such that:

$$T(B_1) \subset \bigcup_{x \in S} B_\varepsilon(Tx)$$

Consider now the orthogonal projection P onto the following finite dimensional space:

$$E = span\left(Tx \middle| x \in S\right)$$

Since the set S is finite, this space E is finite dimensional, and so P is of finite rank, $P \in F(H)$. Now observe that for any norm one $y \in H$ and any $x \in S$ we have:

$$\begin{aligned} ||Ty - Tx||^2 &= ||Ty - PTx||^2 \\ &= ||Ty - PTy + PTy - PTx||^2 \\ &= ||Ty - PTy||^2 + ||PTx - PTy||^2 \end{aligned}$$

Now by picking $x \in S$ such that the ball $B_\varepsilon(Tx)$ covers the point Ty, we conclude from this that we have the following estimate:

$$||Ty - PTy|| \leq ||Ty - Tx|| \leq \varepsilon$$

Thus we have $||T - PT|| \leq \varepsilon$, which gives the density result. □

Quite remarkably, the set of compact operators is closed, and we have:

THEOREM 15.26. *The set of compact operators*

$$K(H) \subset B(H)$$

is a closed two-sided $*$-ideal.

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

(1) It is clear that $K(H)$ is a vector space. In order to prove now that $K(H)$ is closed, assume that $T_n \in K(H)$ converges to $T \in B(H)$. Given $\varepsilon > 0$, pick $N \in \mathbb{N}$ such that:

$$||T - T_N|| \leq \varepsilon$$

By compactness of T_N we can find a finite set $S \subset B_1$ such that:

$$T_N(B_1) \subset \bigcup_{x \in S} B_\varepsilon(T_N x)$$

We conclude that for any $y \in B_1$ there exists $x \in S$ such that:

$$\begin{aligned} ||Ty - Tx|| &\leq ||Ty - T_N y|| + ||T_N y - T_N x|| + ||T_N x - Tx|| \\ &\leq \varepsilon + \varepsilon + \varepsilon \\ &= 3\varepsilon \end{aligned}$$

Thus, we have an inclusion as follows, showing that T is indeed compact:

$$T(B_1) \subset \bigcup_{x \in S} B_{3\varepsilon}(Tx)$$

(2) Regarding now the fact that $K(H)$ is stable under involution, this follows from Proposition 15.23, Proposition 15.25 and (1). Indeed, by using Proposition 15.25, given $T \in K(H)$ we can write it as a limit of finite rank operators, as follows:

$$T = \lim_{n \to \infty} T_n$$

Now by applying the adjoint, we have as well $T^* = \lim_{n \to \infty} T_n^*$. We know from Proposition 15.23 that the operators T_n^* are of finite rank, and so compact by Proposition 15.25, and by using (1) we obtain that T^* is compact too, as desired.

(3) Finally, regarding the ideal property of $K(H)$, this is clear from definitions. □

Here is now a second key result regarding the compact operators:

THEOREM 15.27. *A bounded operator $T \in B(H)$ is compact precisely when*

$$Te_n \to 0$$

for any orthonormal system $\{e_n\} \subset H$.

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

" $\implies$ " Assume that T is compact. By contradiction, assume $Te_n \not\to 0$. This means that there exists $\varepsilon > 0$ and a subsequence satisfying $||Te_{n_k}|| > \varepsilon$, and by replacing $\{e_n\}$ with this subsequence, we can assume that the following holds, with $\varepsilon > 0$:

$$||Te_n|| > \varepsilon$$

Since T is compact, a certain subsequence $\{Te_{n_k}\}$ must converge. Thus, by replacing once again $\{e_n\}$ with a subsequence, we can assume that we have, with $x \neq 0$:

$$Te_n \to x$$

But this is a contradiction, as desired, because we obtain in this way:

$$\begin{aligned} < x, x > &= \lim_{n \to \infty} < Te_n, x > \\ &= \lim_{n \to \infty} < e_n, T^* x > \\ &= 0 \end{aligned}$$

"$\Longleftarrow$" Assume $Te_n \to 0$, for any orthonormal system $\{e_n\} \subset H$. In order to prove $T \in K(H)$, we will prove that T is in the closure of the space of finite rank operators:

$$T \in \overline{F(H)}$$

We do this by contradiction. So, assume that there exists $\varepsilon > 0$ such that:

$$S \in F(H) \implies ||T - S|| > \varepsilon$$

As a first observation, by using $S = 0$ we obtain $||T|| > \varepsilon$. Thus, we can find a norm one vector $e_1 \in H$ such that the following holds:

$$||Te_1|| > \varepsilon$$

Our claim, which will bring the desired contradiction, is that we can construct by recurrence vectors $e_1, \ldots, e_n$ such that the following holds, for any i:

$$||Te_i|| > \varepsilon$$

Indeed, assume that we have constructed $e_1, \ldots, e_n$. Let $E \subset H$ be the linear space spanned by these vectors, and set $P = Proj(E)$. Since TP has finite rank, our assumption above shows that we have $||T - TP|| > \varepsilon$. Thus, we can find $x \in H$ such that:

$$||(T - TP)x|| > \varepsilon$$

We have then $x \notin E$, and so we can consider the following nonzero vector:

$$y = (1 - P)x$$

With this nonzero vector y constructed, now let us set:

$$e_{n+1} = \frac{y}{||y||}$$

This vector e_{n+1} is then orthogonal to E, has norm one, and satisfies:

$$||Te_{n+1}|| \geq ||y||^{-1}\varepsilon \geq \varepsilon$$

Thus we are done with our construction by recurrence, and this contradicts our assumption that $Te_n \to 0$, for any orthonormal system $\{e_n\} \subset H$, as desired. □

Let us discuss now the spectral theory of the compact operators. We first have:

PROPOSITION 15.28. *Assuming that $T \in B(H)$, with $\dim H = \infty$, is compact and self-adjoint, the following happen:*

(1) *The eigenvalues of T form a sequence $\lambda_n \to 0$.*
(2) *All eigenvalues $\lambda_n \neq 0$ have finite multiplicity.*

PROOF. We prove both the assertions at the same time. For this purpose, we fix a number $\varepsilon > 0$, we consider all the eigenvalues satisfying $|\lambda| \geq \varepsilon$, and for each such eigenvalue we consider the corresponding eigenspace $E_\lambda \subset H$. Let us set:

$$E = span\left(E_\lambda \,\middle|\, |\lambda| \geq \varepsilon\right)$$

Our claim, which will prove both (1) and (2), is that this space E is finite dimensional. In now to prove now this claim, we can proceed as follows:

(1) We know that we have $E \subset Im(T)$. Our claim is that we have:

$$\bar{E} \subset Im(T)$$

Indeed, assume that we have a sequence $g_n \in E$ which converges, $g_n \to g \in \bar{E}$. Let us write $g_n = Tf_n$, with $f_n \in H$. By definition of E, the following condition is satisfied:

$$h \in E \implies ||Th|| \geq \varepsilon ||h||$$

Now since the sequence $\{g_n\}$ is Cauchy we obtain from this that the sequence $\{f_n\}$ is Cauchy as well, and with $f_n \to f$ we have $Tf_n \to Tf$, as desired.

(2) Consider now the projection $P \in B(H)$ onto the above space $\bar{E}$. The composition PT is then as follows, surjective on its target:

$$PT : H \to \bar{E}$$

On the other hand since T is compact so must be PT, and if follows from this that the space $\bar{E}$ is finite dimensional. Thus E itself must be finite dimensional too, and as explained in the beginning of the proof, this gives (1) and (2), as desired. □

In order to construct now eigenvalues, we will need:

PROPOSITION 15.29. *If T is compact and self-adjoint, one of the numbers*

$$||T|| \ , \ -||T||$$

must be an eigenvalue of T.

PROOF. We know from the spectral theory of the self-adjoint operators that the spectral radius $||T||$ of our operator T is attained, and so one of the numbers $||T||, -||T||$ must be in the spectrum. In order to prove now that one of these numbers must actually appear as an eigenvalue, we must use the compactness of T, as follows:

(1) First, we can assume $||T|| = 1$. By functional calculus this implies $||T^3|| = 1$ too, and so we can find a sequence of norm one vectors $x_n \in H$ such that:

$$| < T^3 x_n, x_n > | \to 1$$

By using our assumption $T = T^*$, we can rewrite this formula as follows:

$$| < T^2 x_n, Tx_n > | \to 1$$

Now since T is compact, and $\{x_n\}$ is bounded, we can assume, up to changing the sequence $\{x_n\}$ to one of its subsequences, that the sequence Tx_n converges:

$$Tx_n \to y$$

Thus, the convergence formula found above reformulates as follows, with $y \neq 0$:

$$| < Ty, y > | = 1$$

(2) Our claim now, which will finish the proof, is that this latter formula implies $Ty = \pm y$. Indeed, by using Cauchy-Schwarz and $||T|| = 1$, we have:

$$| < Ty, y > | \leq ||Ty|| \cdot ||y|| \leq 1$$

We know that this must be an equality, so Ty, y must be proportional. But since T is self-adjoint the proportionality factor must be ± 1, and so we obtain, as claimed:

$$Ty = \pm y$$

Thus, we have constructed an eigenvector for $\lambda = \pm 1$, as desired. □

We can further build on the above results in the following way:

PROPOSITION 15.30. *If T is compact and self-adjoint, there is an orthogonal basis of H made of eigenvectors of T.*

PROOF. We use Proposition 15.28. According to the results there, we can arrange the nonzero eigenvalues of T, taken with multiplicities, into a sequence $\lambda_n \to 0$. Let $y_n \in H$ be the corresponding eigenvectors, and consider the following space:

$$E = \overline{span(y_n)}$$

The result follows then from the following observations:

(1) Since we have $T = T^*$, both E and its orthogonal $E^\perp$ are invariant under T.

(2) On the space E, our operator T is by definition diagonal.

(3) On the space $E^\perp$, our claim is that we have $T = 0$. Indeed, assuming that the restriction $S = T_{E^\perp}$ is nonzero, we can apply Proposition 15.29 to this restriction, and we obtain an eigenvalue for S, and so for T, contradicting the maximality of E. □

With the above results in hand, we can now formulate a first theorem, as follows:

THEOREM 15.31. *Assuming that $T \in B(H)$, with $\dim H = \infty$, is compact and self-adjoint, the following happen:*

(1) *The spectrum $\sigma(T) \subset \mathbb{R}$ consists of a sequence $\lambda_n \to 0$.*
(2) *All spectral values $\lambda \in \sigma(T) - \{0\}$ are eigenvalues.*
(3) *All eigenvalues $\lambda \in \sigma(T) - \{0\}$ have finite multiplicity.*
(4) *There is an orthogonal basis of H made of eigenvectors of T.*

PROOF. In view of our various results above, it is enough to establish (2). So, assume that $\lambda \neq 0$ belongs to the spectrum $\sigma(T)$, but is not an eigenvalue. By using Proposition 15.30, pick an orthonormal basis $\{e_n\}$ of H consisting of eigenvectors of T, and set:

$$Sx = \sum_n \frac{< x, e_n >}{\lambda_n - \lambda} e_n$$

Then S is an inverse for $T - \lambda$, and so we have $\lambda \notin \sigma(T)$, as desired. □

Finally, we have the following result, regarding the general case:

THEOREM 15.32. *The compact operators $T \in B(H)$, with $\dim H = \infty$, are the operators of the following form, with $\{e_n\}$, $\{f_n\}$ being orthonormal families, and with $\lambda_n \searrow 0$:*

$$T(x) = \sum_n \lambda_n < x, e_n > f_n$$

The numbers λ_n, called singular values of T, are the eigenvalues of $|T|$. In fact, the polar decomposition of T is given by $T = U|T|$, with

$$|T|(x) = \sum_n \lambda_n < x, e_n > e_n$$

and with U being given by $Ue_n = f_n$, and $U = 0$ on the complement of $span(e_i)$.

PROOF. This basically follows from Theorem 15.31, as follows:

(1) Given two orthonormal families $\{e_n\}$, $\{f_n\}$, and a sequence of real numbers $\lambda_n \searrow 0$, consider the linear operator given by the formula in the statement, namely:

$$T(x) = \sum_n \lambda_n < x, e_n > f_n$$

Our first claim is that T is bounded. Indeed, when assuming $|\lambda_n| \leq \varepsilon$ for any n, which is something that we can do if we want to prove that T is bounded, we have:

$$\begin{aligned} ||T(x)||^2 &= \left| \sum_n \lambda_n < x, e_n > f_n \right|^2 \\ &= \sum_n |\lambda_n|^2 | < x, e_n > |^2 \\ &\leq \varepsilon^2 \sum_n | < x, e_n > |^2 \\ &\leq \varepsilon^2 ||x||^2 \end{aligned}$$

(2) The next observation is that this operator is indeed compact, because it appears as the norm limit, $T_N \to T$, of the following sequence of finite rank operators:

$$T_N = \sum_{n \leq N} \lambda_n < x, e_n > f_n$$

(3) Regarding now the polar decomposition assertion, for the above operator, this follows once again from definitions. Indeed, the adjoint is given by:

$$T^*(x) = \sum_n \lambda_n < x, f_n > e_n$$

Thus, when composing T^* with T, we obtain the following operator:

$$T^*T(x) = \sum_n \lambda_n^2 < x, e_n > e_n$$

Now by extracting the square root, we obtain the formula in the statement, namely:

$$|T|(x) = \sum_n \lambda_n < x, e_n > e_n$$

(4) Conversely now, assume that $T \in B(H)$ is compact. Then T^*T, which is self-adjoint, must be compact as well, and so by Theorem 15.31 we have a formula as follows, with $\{e_n\}$ being a certain orthonormal family, and with $\lambda_n \searrow 0$:

$$T^*T(x) = \sum_n \lambda_n^2 < x, e_n > e_n$$

By extracting the square root we obtain the formula of $|T|$ in the statement, and then by setting $U(e_n) = f_n$ we obtain a second orthonormal family, $\{f_n\}$, such that:

$$T(x) = U|T| = \sum_n \lambda_n < x, e_n > f_n$$

Thus, our compact operator $T \in B(H)$ appears indeed as in the statement. □

15e. Exercises

This was a quite straightforward chapter, and as exercises, we have:

EXERCISE 15.33. *In relation with separability, learn about orthogonal polynomials.*

EXERCISE 15.34. *Clarify the details in the proof of the measurable calculus theorem.*

EXERCISE 15.35. *Jointly diagonalize the families of commuting normal operators.*

EXERCISE 15.36. *Check the positivity details, in the proof of the polar decomposition.*

EXERCISE 15.37. *Learn as well about the strictly positive operators, $T > 0$.*

EXERCISE 15.38. *Work out other decomposition results, for the linear operators.*

EXERCISE 15.39. *Learn about the trace class operators, and their properties.*

EXERCISE 15.40. *Learn also about Hilbert-Schmidt operators, and their properties.*

As bonus exercise, in addition to this, learn some operator algebras as well.

CHAPTER 16

Quantum mechanics

16a. Atomic theory

Welcome to quantum mechanics, and in the hope that we will survive. Going back in time, about 100 years ago, the main question was that of understanding the intimate structure of matter, at the atomic level. There is of course a long story here, regarding the intimate structure of matter, going back centuries and even millennia ago, and our presentation here will be quite simplified. As a starting point, because we do need a starting point, let us agree on the following fact, or rather claim, from around 1900:

CLAIM 16.1. *Ordinary matter is made of small particles called atoms, with each atom appearing as a mix of even smaller particles, namely protons* $+$*, neutrons* 0 *and electrons* $-$*, with the same number of protons* $+$ *and electrons* $-$*.*

As a first observation, this is something which does not look obvious at all, with probably lots of work, by many people, being involved, as to lead to this claim. And so it is. The story goes back to the discovery of charges and electricity, which were attributed to a small particle, the electron $-$. Now since matter is by default neutral, this naturally leads to the consideration to the proton $+$, having the same charge as the electron.

But, as a natural question, why should be these electrons $-$ and protons $+$ that small? And also, what about the neutron 0? These are not easy questions, and the fact that it is so came from several clever experiments. Let us first recall that careful experiments with tiny particles are practically impossible. However, all sorts of brutal experiments, such as bombarding matter with other pieces of matter, accelerated to the extremes, or submitting it to huge electric and magnetic fields, do work. And it is such kind of experiments, due to Thomson, Rutherford and others, "peeling off" protons $+$, neutrons 0 and electrons $-$ from matter, and observing them, that led to the conclusion that these small beasts $+, 0, -$ exist indeed, in agreement with Claim 16.1.

Of particular importance here was as well the radioactivity theory of Becquerel and Pierre and Marie Curie, involving this time such small beasts, or perhaps some related radiation, peeling off by themselves, in heavy elements such as uranium ${}_{92}\mathrm{U}$, polonium ${}_{84}\mathrm{Po}$ and radium ${}_{88}\mathrm{Ra}$. And there was also Einstein's work on the photoelectric effect, light interacting with matter, suggesting that even light itself might have associated to it some kind of particle, called photon. All this goes of course beyond Claim 16.1, with

further particles involved, and more on this later, but as a general idea, all this deluge of small particle findings, all coming around 1900-1910, further solidified Claim 16.1.

So, taking now Claim 16.1 for granted, how are then the atoms organized, as mixtures of protons $+$, neutrons 0 and electrons $-$? The answer here lies again in the above-mentioned "brutal" experiments of Thomson, Rutherford and others, which not only proved Claim 16.1, but led to an improved version of it, as follows:

CLAIM 16.2. *The atoms are formed by a core of protons $+$ and neutrons 0, surrounded by a cloud of electrons $-$, gravitating around the core.*

And with this, good news, we are into some kind of mechanics, that we can start investigating. However, before doing that, we need more data regarding this mechanics, and the question is, how to manage, via clever experiments, to obtain this data?

In answer, let there be light. Recall indeed from chapter 6 the following key fact:

FACT 16.3. *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.*

In view of this, which is something truly smart, our privileged method for studying the atoms, which is something very simple, that you can do in your own garage, armed with only a prism, will be that of heating them, and recording the spectral signature.

In practice now, there is a long story with this, involving many discoveries of many people, around 1890-1900, focusing on hydrogen H. We will present here things a bit retrospectively. First on our list is the following discovery, by Lyman in 1906:

FACT 16.4 (Lyman). *The hydrogen atom has spectral lines given by the formula*

$$\frac{1}{\lambda} = R\left(1 - \frac{1}{n^2}\right)$$

where $R \simeq 1.097 \times 10^7$ and $n \geq 2$, which are as follows,

n	Name	Wavelength	Color
	$-$	$-$	
2	α	121.567	UV
3	β	102.572	UV
4	γ	97.254	UV
$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	limit	91.175	UV

called Lyman series of the hydrogen atom.

Observe that all the Lyman series lies in UV, which is invisible to the naked eye. Due to this fact, this series, while theoretically being the most important, was discovered only second. The first discovery, which was the big one, and the breakthrough, was by Balmer, the founding father of all this, back in 1885, in the visible range, as follows:

FACT 16.5 (Balmer). *The hydrogen atom has spectral lines given by the formula*

$$\frac{1}{\lambda} = R\left(\frac{1}{4} - \frac{1}{n^2}\right)$$

where $R \simeq 1.097 \times 10^7$ *and* $n \geq 3$*, which are as follows,*

n	Name	Wavelength	Color
	–	–	
3	α	656.279	red
4	β	486.135	aqua
5	γ	434.047	blue
6	δ	410.173	violet
7	ε	397.007	UV
$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	limit	346.600	UV

called Balmer series of the hydrogen atom.

So, this was Balmer's original result, which started everything. As a third main result now, this time in IR, due to Paschen in 1908, we have:

FACT 16.6 (Paschen). *The hydrogen atom has spectral lines given by the formula*

$$\frac{1}{\lambda} = R\left(\frac{1}{9} - \frac{1}{n^2}\right)$$

where $R \simeq 1.097 \times 10^7$ *and* $n \geq 4$*, which are as follows,*

n	Name	Wavelength	Color
	–	–	
4	α	1875	IR
5	β	1282	IR
6	γ	1094	IR
$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	limit	820.4	IR

called Paschen series of the hydrogen atom.

Observe the striking similarity between the above three results. In fact, we have here the following fundamental, grand result, due to Rydberg in 1888, based on the Balmer series, and with later contributions by Ritz in 1908, using the Lyman series as well:

CONCLUSION 16.7 (Rydberg, Ritz). *The spectral lines of the hydrogen atom are given by the Rydberg formula, depending on integer parameters $n_1 < n_2$,*

$$\frac{1}{\lambda_{n_1n_2}} = R\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

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:

$$\frac{1}{\lambda_{n_1n_2}} + \frac{1}{\lambda_{n_2n_3}} = \frac{1}{\lambda_{n_1n_3}}$$

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 16.8. *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.

Getting now to quantum mechanics, we have some serious data here. The spectral lines are something basic and beautiful, and in order to get started with our theory, we first need to solve the puzzle of the Ritz-Rydberg combination principle.

But, how to do this? Fortunately, matrix theory comes to the rescue, as follows:

THOUGHT 16.9. *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, which are the elementary matrices given by

$$e_{ij} : e_j \to e_i$$

perhaps taken in infinite dimensions, as to allow infinite-ranging indices.

Which sounds quite good and conceptual, and definitely mathematical. In a word, saved by linear algebra, as always or almost, and we can now start dreaming of:

THOUGHT 16.10. *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.*

Which is a considerable advance, because we are now into familiar territory, namely some kind of mathematical physics. And with this in mind, all the pieces of our puzzle start fitting together, and we are led to the following grand conclusion:

CLAIM 16.11 (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 isn't this beautiful. Moreover, some further claims, also by Bohr and others, are that the theory can be further extended and fine-tuned as to explain many other phenomena, such as the above-mentioned findings of Einstein, and of Becquerel and Pierre and Marie Curie, and generally speaking, all the physics and chemistry known.

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 Thought 16.10.

16b. Schrödinger equation

Before explaining what Heisenberg was saying, 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 16.12. *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:

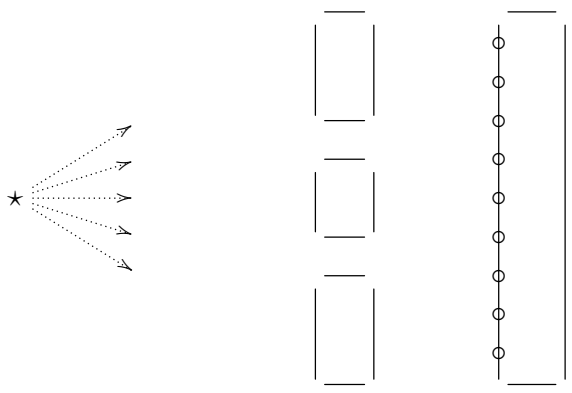
$$\varphi_t(x) = \varphi_0(x) + \gamma(t)$$

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 16.13. *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.*

To be more precise, the above experiments, which are nicely described, with extensive comments, in Feynman's book [35], are performed with a machinery as follows:



Here on the left we have a multi-purpose gun $\star$, which can shoot bullets, water waves, or electrons. On the middle we have a wall with two holes in it. And on the right we have a solid wall, with sensors $\circ$, adapted to the matter that we are shooting.

The first experiment, performed with 22 cal ammo, assumed to be idealized, as to be indestructible, and we refer again to Feynman [35] for full details, goes as follows:

EXPERIMENT 16.14. *When shooting bullets, the density function on the right wall, stopping them, as recorded by our sensors there, is given by*

$$\varphi = \varphi_1 + \varphi_2$$

where φ_1 is the same density, but measured with the lower hole closed, and φ_2 is also the same density, but measured with the upper hole closed.

Nothing surprising so far. Now let us shoot water waves, or rather assume that our gun $\star$ is a wave source. For this experiment, the sensors at right are set to measure the energy of the incoming wave, which is proportional to the square of the height.

The experiment, best performed with our favorite drinkable, tap water, gives:

EXPERIMENT 16.15. *When shooting waves, the energy density functions at right, measured with one or the other hole open, or both holes open, are related by*

$$\varphi_1 = |\psi_1|^2 \quad , \quad \varphi_2 = |\psi_1|^2 \quad , \quad \varphi = |\psi_1 + \psi_2|^2$$

where $\psi_1, \psi_2 \in \mathbb{C}$ are the amplitudes of the waves passing through one hole, with the other hole closed. This phenomenon and formula of φ are due to interference.

This is again, something not surprising, that we know, coming from the fact that two colliding waves can add up in various ways, depending on their phases.

Now let us shoot electrons. At first sight, these behave like particles, because our sensors beep for them one at the time. However, when examining the results regarding probability distributions, these don't add up as for bullets, the conclusions being:

EXPERIMENT 16.16. *When shooting electrons, these come up one at the time, exactly as bullets. However, in what regards the density functions, these don't add up:*

$$\varphi \neq \varphi_1 + \varphi_2$$

Thus, we have some interference, and most likely the correct formula is

$$\varphi_1 = |\psi_1|^2 \quad , \quad \varphi_2 = |\psi_1|^2 \quad , \quad \varphi = |\psi_1 + \psi_2|^2$$

with $\psi_1, \psi_2 \in \mathbb{C}$ being certain amplitudes, exactly as for the waves.

This is a bit surprising, showing that the electrons have a mix of particle and wave behavior, at least with respect to this experiment. Let us also mention too that, contrary to the previous two experiments which are simple and real, this is rather a Gedankenexperiment, and so the wave formulae are to be taken with care. See Feynman [35].

Finally, as a last experiment in our series, again with electrons, we have:

EXPERIMENT 16.17. *When shooting electrons as before, but by putting a light bulb behind one hole, whose light is scattered by electrons passing through that hole:*

$$\varphi = \varphi_1 + \varphi_2$$

That is, observing the electrons passing through one hole, via them scattering light, has killed the interference process, and we have now usual particles, like bullets.

And this is probably the most surprising experiment of them all. Indeed, the fact that in our previous Experiment 16.16 we had particles when counting and waves when looking at densities might seem odd, but after all, why not. So these are our beasts, electrons, and this is how their properties are, a bit odd, but at least we know one thing. However, what happens now seems to defy any logic. Observing the electrons has changed their properties, and that's how things are. Welcome to quantum mechanics.

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 $\varphi_t(x) = |\psi_t(x)|^2$. So, let us reformulate Question 16.12, in the following way:

QUESTION 16.18. *In the context of the hydrogen atom, assuming that the proton is fixed, what is the amplitude function $\psi_t(x)$ of the position of the electron e, at time t,*

$$P_t(e \in V) = \int_V |\psi_t(x)|^2 dx$$

as function of an intial amplitude function $\psi_0(x)$? Moreover, can the corresponding equation be solved, and will this prove the Bohr claims for hydrogen, statistically?

Mathematically, what we did here is to replace the density $\varphi_t(x) > 0$ by the amplitude function $\psi_t(x) \in \mathbb{C}$. 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, experience with mathematics shows that such manipulations can be crucial, raising for instance the possibility that the amplitude function satisfies some simple equation, while the density itself, maybe not. So, let us hope for this to happen.

And this is what happens indeed. Schrödinger was led in this way to:

CLAIM 16.19 (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 interesting things can be said here, about the origins and various interpretations of the Schrödinger equation, and more on this later.

In practice now, let us start with some computations. As a first question, we would like to see how the probability density $\varphi = |\psi|^2$ evolves in time, and we have here:

PROPOSITION 16.20. *In the context of the Schrödinger equation we have*

$$\dot{\varphi} = \frac{ih}{2m}\left(\Delta\psi\cdot\bar{\psi} - \Delta\bar{\psi}\cdot\psi\right)$$

for the time derivative of the probability density function $\varphi = |\psi|^2$.

PROOF. According to the Leibnitz product rule, we have the following formula:

$$\dot{\varphi} = \frac{d}{dt}|\psi|^2 = \frac{d}{dt}(\psi\bar{\psi}) = \dot{\psi}\bar{\psi} + \psi\dot{\bar{\psi}}$$

On the other hand, the Schrödinger equation and its conjugate read:

$$\dot{\psi} = \frac{ih}{2m}\left(\Delta\psi - \frac{2m}{h^2}V\psi\right) \quad , \quad \dot{\bar{\psi}} = -\frac{ih}{2m}\left(\Delta\bar{\psi} - \frac{2m}{h^2}V\bar{\psi}\right)$$

By plugging this data, we obtain the following formula:

$$\dot{\varphi} = \frac{ih}{2m}\left[\left(\Delta\psi - \frac{2m}{h^2}V\psi\right)\bar{\psi} - \left(\Delta\bar{\psi} - \frac{2m}{h^2}V\bar{\psi}\right)\psi\right]$$

But this gives, after simplifying, the formula in the statement. □

As an important application of Proposition 16.20, we have:

THEOREM 16.21. *The Schrödinger equation conserves probability amplitudes,*

$$\int_{\mathbb{R}^3}|\psi_0|^2 = 1 \implies \int_{\mathbb{R}^3}|\psi_t|^2 = 1$$

in agreement with the basic probabilistic requirement, $P = 1$ overall.

PROOF. According to the formula in Proposition 16.20, we have:

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^3}|\psi|^2\,dx &= \int_{\mathbb{R}^3}\frac{d}{dt}|\psi|^2\,dx\\ &= \int_{\mathbb{R}^3}\dot{\varphi}\,dx\\ &= \frac{ih}{2m}\int_{\mathbb{R}^3}\left(\Delta\psi\cdot\bar{\psi} - \Delta\bar{\psi}\cdot\psi\right)dx\end{aligned}$$

Now by remembering the definition of the Laplace operator, we have:

$$\begin{aligned}
\frac{d}{dt}\int_{\mathbb{R}^3}|\psi|^2\,dx &= \frac{ih}{2m}\int_{\mathbb{R}^3}\sum_i\left(\frac{d^2\psi}{dx_i^2}\cdot\bar{\psi}-\frac{d^2\bar{\psi}}{dx_i^2}\cdot\psi\right)dx\\
&= \frac{ih}{2m}\sum_i\int_{\mathbb{R}^3}\frac{d}{dx_i}\left(\frac{d\psi}{dx_i}\cdot\bar{\psi}-\frac{d\bar{\psi}}{dx_i}\cdot\psi\right)dx\\
&= \frac{ih}{2m}\sum_i\int_{\mathbb{R}^2}\left[\frac{d\psi}{dx}\cdot\bar{\psi}-\frac{d\bar{\psi}}{dx}\cdot\psi\right]_{-\infty}^{\infty}\frac{dx}{dx_i}\\
&= \frac{ih}{2m}\sum_i\int_{\mathbb{R}^2}0\,\frac{dx}{dx_i}\\
&= 0
\end{aligned}$$

Here we have used at the end the assumption, which is physically speaking, something reasonable, that the wave function and its derivatives vanish at ∞. Now with this in hand, since the quantity under consideration is constant, we obtain the result. □

Let us do now some more computations, in order to get some insight into the quantum mechanics of the particle, as dictated by the Schrödinger equation. We first have:

THEOREM 16.22. *The average position and momentum of the particle are*

$$< x >= \int_{\mathbb{R}^3} x|\psi|^2\,dx$$

$$< p >= -ih\int_{\mathbb{R}^3}\nabla\psi\cdot\bar{\psi}\,dx$$

with the convention that the average speed is the derivative of the average position.

PROOF. This follows again by doing some mathematics, as follows:

(1) The formula for the average position $< x >$ is clear from definitions. Regarding now the average speed $< v >$, we have here the following computation:

$$\begin{aligned}
< v > &= \frac{d < x >}{dt}\\
&= \int_{\mathbb{R}^3} x\cdot\frac{d}{dt}|\psi|^2\,dx\\
&= \int_{\mathbb{R}^3} x\,\dot{\varphi}\,dx\\
&= \frac{ih}{2m}\int_{\mathbb{R}^3} x\left(\Delta\psi\cdot\bar{\psi}-\Delta\bar{\psi}\cdot\psi\right)dx
\end{aligned}$$

(2) But each of the components can be computed as follows, by taking into account the vanishing formula found in the proof of Theorem 16.21:

$$\begin{aligned}
< v >_i &= \frac{ih}{2m}\int_{\mathbb{R}^3} x_i\left(\Delta\psi\cdot\bar{\psi}-\Delta\bar{\psi}\cdot\psi\right)dx\\
&= \frac{ih}{2m}\int_{\mathbb{R}^3} x_i\sum_j\left(\frac{d^2\psi}{dx_j^2}\cdot\bar{\psi}-\frac{d^2\bar{\psi}}{dx_j^2}\cdot\psi\right)dx\\
&= \frac{ih}{2m}\sum_j\int_{\mathbb{R}^3} x_i\left(\frac{d^2\psi}{dx_j^2}\cdot\bar{\psi}-\frac{d^2\bar{\psi}}{dx_j^2}\cdot\psi\right)dx\\
&= \frac{ih}{2m}\int_{\mathbb{R}^3} x_i\left(\frac{d^2\psi}{dx_i^2}\cdot\bar{\psi}-\frac{d^2\bar{\psi}}{dx_i^2}\cdot\psi\right)dx
\end{aligned}$$

(3) We can now finish the computation by doing two partial integrations, as follows:

$$\begin{aligned}
< v >_i &= \frac{ih}{2m}\int_{\mathbb{R}^3} x_i\cdot\frac{d}{dx_i}\left(\frac{d\psi}{dx_i}\cdot\bar{\psi}-\frac{d\bar{\psi}}{dx_i}\cdot\psi\right)dx\\
&= -\frac{ih}{2m}\int_{\mathbb{R}^3}\left(\frac{d\psi}{dx_i}\cdot\bar{\psi}-\frac{d\bar{\psi}}{dx_i}\cdot\psi\right)dx\\
&= -\frac{ih}{m}\int_{\mathbb{R}^3}\frac{d\psi}{dx_i}\cdot\bar{\psi}\,dx
\end{aligned}$$

(4) We conclude that the average speed is given by the following formula:

$$< v >= -\frac{ih}{m}\int_{\mathbb{R}^3}\nabla\psi\cdot\bar{\psi}\,dx$$

By multiplying by the mass, we obtain the formula for $< p >$ in the statement. □

As an interesting speculation now, based on the above two formulae that we found, and inspired from Heisenberg's idea of matrix mechanics, we have:

SPECULATION 16.23. *The average position and momentum formulae, written as*

$$< x >= \int_{\mathbb{R}^3}\bar{\psi}\cdot x\cdot\psi\,dx$$

$$< p >= \int_{\mathbb{R}^3}\bar{\psi}\cdot(-ih\nabla)\cdot\psi\,dx$$

suggest that x represents position, and $-ih\nabla$ represents momentum.

To be more precise, the quantities x and $-ih\nabla$ appearing above are, mathematically speaking, linear operators, which are actually unbounded, in the sense of chapter 15. Now with this convention, the above tells us that for computing the average value of x, p, we must "sandwich" the corresponding operator between $\bar{\psi}, \psi$, and then integrate.

Which is something quite remarkable, and we are now very tempted to formulate something extremely general, and of course still a bit vague, as follows:

SPECULATION 16.24. *The average value of an observable O should appear as*

$$< O >= \int_{\mathbb{R}^3} \bar{\psi} \cdot \widehat{O} \cdot \psi \, dx$$

"sandwich between $\bar{\psi}, \psi$ and integrate", where $\widehat{O}$ is the operator associated to O.

As an illustration, let us see if this sandwiching method works for the kinetic energy of the particle. The kinetic energy is given by the following formula:

$$T = \frac{m||v||^2}{2} = \frac{< p, p >}{2m}$$

Thus, the operator associated to the energy should be given by:

$$\widehat{T} = \frac{< -ih\nabla, -ih\nabla >}{2m} = -\frac{h^2\Delta}{2m}$$

We obtain in this way something which looks quite reasonable, as follows:

$$< T >= -\frac{h^2}{2m} \int_{\mathbb{R}^3} \Delta\psi \cdot \bar{\psi} \, dx$$

More generally now, we can incorporate into this the potential energy too, and we are led in this way to the following interesting, conceptual conclusion:

CONCLUSION 16.25. *According to the above speculations, the operator associated to the total energy, or Hamiltonian, $H = T + V$ is given by*

$$\widehat{H} = -\frac{h^2\Delta}{2m} + V$$

and so the Schrödinger equation itself appears as

$$ih\dot{\psi} = \widehat{H}\psi$$

which is something quite conceptual, in terms of this operator.

To be more precise, according to the above, $\widehat{H}$ appears indeed via the formula in the statement. But now, let us look back at the Schrödinger equation, namely:

$$ih\dot{\psi} = -\frac{h^2}{2m}\Delta\psi + V\psi$$

We recognize on the right the operator $\widehat{H}$ acting on ψ, which leads to the above conclusion. Which is, most likely, the good viewpoint on the Schrödinger equation.

Summarizing, theory doing well, and time now for some axioms? Following Heisenberg and Schrödinger, and then especially Dirac, who did the axiomatization, we have:

DEFINITION 16.26. *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 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 operators on a certain Hilbert space H.

So long for the axiomatization of quantum mechanics, quickly explained as above, by taking some shortcuts, and needless to say, all the above will take us some time, to understand. As a first result now of our new theory, which is famous, we have:

THEOREM 16.27 (Heisenberg). *We have the following uncertainty principle,*

$$\sigma_S \cdot \sigma_T \geq \left| \frac{< [S, T] >}{2} \right|$$

regarding the variances of any two observables S, T. In particular, we have

$$\sigma_x \cdot \sigma_p \geq \frac{h}{2}$$

implying that you cannot measure position and momentum at the same time.

PROOF. This follows indeed by doing some computations with operators and their commutators, and for details here, we refer to any standard introduction to quantum mechanics, such as Griffiths [45]. In what follows we will not really need this, and focus instead on a more basic question, the functioning of the hydrogen atom. □

16c. Hydrogen atom

Moving now towards hydrogen, which remains our main objective, we first have:

THEOREM 16.28. *In the case of time-independent potentials V, including the Coulomb potential of the proton, the separated solutions of the Schrödinger equation*

$$\psi_t(x) = w_t\phi(x)$$

are given by the following formulae, with E being a certain constant,

$$w = e^{-iEt/h}w_0 \quad , \quad E\phi = -\frac{h^2}{2m}\Delta\phi + V\phi$$

with the equation for ϕ being called time-independent Schrödinger equation.

PROOF. By dividing by ψ, the Schrödinger equation becomes:

$$ih \cdot \frac{\dot{w}}{w} = -\frac{h^2}{2m} \cdot \frac{\Delta\phi}{\phi} + V$$

Now since the left-hand side depends only on time, and the right-hand side depends only on space, both quantities must equal a constant E, and this gives the result. □

Next, we can reformulate the Schrödinger equation in spherical coordinates:

THEOREM 16.29. *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. We use the following formula for the Laplace operator in spherical coordinates, whose proof can be found in any advanced calculus book, such as mine [11]:

$$\Delta = \frac{1}{r^2} \cdot \frac{d}{dr}\left(r^2 \cdot \frac{d}{dr}\right) + \frac{1}{r^2\sin s} \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d}{ds}\right) + \frac{1}{r^2\sin^2 s} \cdot \frac{d^2}{dt^2}$$

By using this formula, the time-independent Schrödinger equation reformulates as:

$$(V-E)\phi = \frac{h^2}{2m}\left[\frac{1}{r^2} \cdot \frac{d}{dr}\left(r^2 \cdot \frac{d\phi}{dr}\right) + \frac{1}{r^2\sin s} \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\phi}{ds}\right) + \frac{1}{r^2\sin^2 s} \cdot \frac{d^2\phi}{dt^2}\right]$$

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:

$$(V-E)\rho\alpha=\frac{h^2}{2m}\left[\frac{\alpha}{r^2}\cdot\frac{d}{dr}\left(r^2\cdot\frac{d\rho}{dr}\right)+\frac{\rho}{r^2\sin s}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{\rho}{r^2\sin^2 s}\cdot\frac{d^2\alpha}{dt^2}\right]$$

By multiplying everything by $2mr^2/(h^2\rho\alpha)$, and then 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 $-K$, and this gives the result. □

Let us first study the angular equation. The result here is as follows:

THEOREM 16.30. *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 a routine study, and with everything above being taken up to linear combinations. For details here, see for instance Griffiths [45]. □

In order to finish our study, it remains to solve the radial equation, for the Coulomb potential V of the proton. As a first manipulation on the radial equation, we have:

PROPOSITION 16.31. *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,

$$Eu=-\frac{h^2}{2m}\cdot u''+\left(V+\frac{h^2l(l+1)}{2mr^2}\right)u$$

which is a time-independent 1D Schrödinger equation.

PROOF. With $\rho = u/r$ as in the statement, we have the following formulae:

$$\rho = \frac{u}{r} \quad , \quad \rho' = \frac{u'r - u}{r^2} \quad , \quad (r^2\rho')' = u''r$$

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 desired. □

It remains to solve the above equation, for the Coulomb potential of the proton. And this leads to the following result, which proves the original atomic claims by Bohr:

THEOREM 16.32 (Schrödinger). *In the case of the hydrogen atom, with V being the Coulomb potential of the proton, we have 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 with all this 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 $\gamma = \sqrt{-2mE}/h$, this equation takes the following form:

$$\frac{u''}{\gamma^2} = \left(1 + \frac{Ke^2}{Er} + \frac{l(l+1)}{(\gamma r)^2}\right) u$$

In terms of the new variable $p = \gamma r$, this latter equation reads:

$$u'' = \left(1 + \frac{\gamma Ke^2}{Ep} + \frac{l(l+1)}{p^2}\right) u$$

Now let us introduce a new constant S for our problem, as follows:

$$S = -\frac{\gamma Ke^2}{E}$$

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:

$$S=-\frac{\gamma Ke^2}{E}\quad:\quad\gamma=\frac{\sqrt{-2mE}}{h}$$

Thus, we have the following formula for the square of S:

$$S^2=\frac{\gamma^2K^2e^4}{E^2}=-\frac{2mE}{h^2}\cdot\frac{K^2e^4}{E^2}=-\frac{2mK^2e^4}{h^2E}$$

Now by using the formula $S=2n$ from (6), the energy E must be of the form:

$$E=-\frac{2mK^2e^4}{h^2S^2}=-\frac{mK^2e^4}{2h^2n^2}$$

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:

$$K=8.988\times10^9\quad,\quad e=1.602\times10^{-19}$$

$$h=1.055\times10^{-34}\quad,\quad m=9.109\times10^{-31}$$

But this gives the formula of E_1 in the statement. □

As a first remark, the above result agrees with the Rydberg formula, due to:

THEOREM 16.33. *The Rydberg constant for hydrogen is given by*

$$R=-\frac{E_1}{h_0c}$$

where E_1 is the Bohr binding energy, and the Rydberg formula itself, namely

$$\frac{1}{\lambda_{n_1n_2}}=R\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right)$$

simply reads, via the energy formula in Theorem 16.32,

$$\frac{1}{\lambda_{n_1n_2}}=\frac{E_{n_2}-E_{n_1}}{h_0c}$$

which is in agreement with the Planck formula $E=h_0c/\lambda$.

PROOF. Here the first assertion is something numeric, coming from:

$$R=\frac{-E_1}{h_0c}=\frac{2.177\times10^{-18}}{6.626\times10^{-34}\times2.998\times10^8}=1.096\times10^7$$

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 16.32:

$$\frac{R}{n^2}=-\frac{E_n}{h_0c}$$

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

Back now to general theory, let us formulate our conclusions so far as follows:

THEOREM 16.34. *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 that we have 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 16.35. *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 $a=1/\gamma$ is the inverse of the parameter appearing in our computations above,

$$\gamma=\frac{\sqrt{-2mE}}{h}$$

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 $a\simeq5.291\times10^{-11}$.

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

(1) According to the various formulae in the proof of Theorem 16.32, taken at $n = 1$, the parameter γ appearing in the computations there is given by:

$$\gamma = \frac{\sqrt{-2mE}}{h} = \frac{1}{h} \cdot m \cdot \frac{Ke^2}{h} = \frac{mKe^2}{h^2}$$

Thus, the inverse $a = 1/\gamma$ is indeed given by the formula in the statement.

(2) Regarding the wave function, according to Theorem 16.34 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. □

In order to improve our results, we will need the following standard fact:

PROPOSITION 16.36. *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,

$$L_q(x) = \frac{e^x}{q!} \left(\frac{d}{dx}\right)^q (e^{-x} x^q)$$

called Rodrigues formula for the Laguerre polynomials.

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:

$$< f, g >= \int_0^\infty f(x)g(x)e^{-x}\, dx$$

The orthogonal basis obtained by applying Gram-Schmidt to the Weierstrass basis $\{x^q\}$ is then formed by the Laguerre polynomials $\{L_q\}$, and this gives the results. □

With the above result in hand, we can now improve our main result, as follows:

THEOREM 16.37. *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, and then doing some remaining work, concerning the normalization of the wave function. □

Quite nice all this, who could have guessed, at the beginning of this book, that we will get sometimes soon into a formula as deep and as complicated, as the above one.

16d. Fine structure

What is next? All sorts of corrections to the solution that we found, due to various phenomena that we neglected in our computations, or rather in our modeling of the problem, which can be both of electric and relativistic nature. But before getting into that, let us first enjoy what we found, say by taking it as a final, exact result regarding the hydrogen atom. As a first conclusion, of quite philosophical nature, we have:

CONCLUSION 16.38. *The phenomenon of quantization appears, mathematically speaking, from certain equations which generically blow up, and force the various separation constants $C\in\mathbb{R}$ which appear to be integers, $C\in\mathbb{N}$.*

To be more precise, the phenomenon of quantization that we are talking about is of course the Bohr energy one, allowing discrete energies only, E_n with $n\in\mathbb{N}$, which is the mother of everything, in quantum mechanics. Looking back at the proof of this fact, separation constants $C\in\mathbb{R}$ which mysteriously became integers, $C\in\mathbb{N}$, was indeed the mathematical phenomenon behind this. Which appeared no less than 3 times:

(1) First when the azimuthal/polar separation parameter, denoted m^2, turned to be the square of an integer, $m\in\mathbb{Z}$.

(2) Then when the radial/angular separation constant K turned to be of a similar form, $K=l(l+1)$ with $l\in\mathbb{N}$.

(3) And finally in the context of the radial equation, where the parameter S there turned to be of the form $S=2n$, with $n\in\mathbb{N}$.

As another comment now, in our study we dismissed several times all sorts of solutions, on various physical grounds, usually unacceptable blow up. But, at a more advanced level, some of these solutions make sense of course, due to the following fact:

FACT 16.39. *The hydrogen atom is not the general 2-body problem in quantum mechanics, but rather the case of confined, stable orbits. Some of the solutions which blow up correspond to scattering, in the context of an electron/proton meeting.*

Again, this is something a bit philosophical. In analogy with classical mechanics, what we did is to solve the planetary motion problem. But things like comets and asteroids still need to be investigated, for having a full theory, and many things can be said here.

Back to work now, corrections, we will focus on energy only. Let us start with:

THEOREM 16.40 (Schrödinger). *The energy of the ϕ_{nlm} state of the hydrogen atom is independent on the quantum numbers l, m, given by the Bohr formula*

$$E_n = -\frac{\alpha^2}{n^2} \cdot \frac{mc^2}{2}$$

where α is a dimensionless constant, called fine structure constant, given by

$$\alpha = \frac{Ke^2}{hc}$$

which in practice means $\alpha \simeq 1/137$.

PROOF. This is the Bohr energy formula that we know, proved by Schrödinger, and reformulated in terms of Sommerfeld's fine structure constant:

(1) We know from Theorem 16.32 that we have the following formula, which can be written as in the statement, by using the fine structure constant α:

$$E_n = -\frac{m}{2}\left(\frac{Ke^2}{h}\right)^2 \cdot \frac{1}{n^2}$$

(2) Observe now that our modified Bohr formula can be further reformulated as follows, with T_c being the kinetic energy of the electron traveling at speed c:

$$E_n = -\frac{\alpha^2}{n^2} \cdot T_c$$

Thus α^2, and so α too, is dimensionless, as being a quotient of energies.

(3) Let us doublecheck however this latter fact, the check being instructive. With respect to the SI system that we use, the units for K, e, h, c are:

$$U_K = \frac{m^3 \cdot kg}{s^2 \cdot C^2} \quad , \quad U_e = C \quad , \quad U_h = \frac{m^2 \cdot kg}{s} \quad , \quad U_c = \frac{m}{s}$$

Thus the units for the fine structure constant α are, as claimed:

$$U_\alpha = U_C \cdot U_e^2 \cdot U_h^{-1} \cdot U_c^{-1} = \frac{m^3 \cdot kg}{s^2 \cdot C^2} \cdot C^2 \cdot \frac{s}{m^2 \cdot kg} \cdot \frac{s}{m} = 1$$

(4) In what regards now the numerics, these are as follows:

$$\alpha = \frac{Ke^2/h}{c} \simeq \frac{2.186 \times 10^6}{2.998 \times 10^8} = 7.291 \times 10^{-3} \simeq \frac{1}{137}$$

Here we have used an estimate for Ke^2/h, from the proof of Theorem 16.32. □

Moving ahead now with our correction work, we will be quite brief, and for further details, we refer as usual to our favorite books, namely Feynman [35], Griffiths [45] and Weinberg [91]. We first have the following result, which is something non-trivial:

THEOREM 16.41. *There is a relativistic correction to be made to the Bohr energy E_n of the state ϕ_{nlm}, depending on the quantum number l, given by*

$$\mathcal{E}_{nl} = \frac{\alpha^2 E_n}{n^2}\left(\frac{n}{l+1/2} - \frac{3}{4}\right)$$

coming by replacing the kinetic energy by the relativistic kinetic energy.

PROOF. According to Einstein, the relativistic kinetic energy is given by:

$$T = \frac{p^2}{2m} - \frac{p^4}{8m^3c^2} + \ldots$$

The Schrödinger equation, based on $T = p^2/2m$, must be therefore corrected with a $\mathcal{T} = -p^4/(8m^3c^2)$ term, and this leads to the above correction term $\mathcal{E}_{nl}$. □

Equally non-trivial is the following correction, independent from the above one:

THEOREM 16.42. *There is a spin-related correction to be made to the Bohr energy E_n of the state ϕ_{nlm}, depending on the number $j = l \pm 1/2$, given by*

$$\mathcal{E}_{nj} = -\frac{\alpha^2 E_n}{n^2} \cdot \frac{n(j-l)}{(l+1/2)(j+1/2)}$$

coming from the torque of the proton on the magnetic moment of the electron.

PROOF. The idea here is that the electron has a spin $\pm 1/2$, which is naturally associated to the quantum number l, leading to the parameter $j = l \pm 1/2$. But, knowing now that the electron has a spin, the proton which moves around it certainly acts on its magnetic moment, and this leads to the above correction term $\mathcal{E}_{nj}$. □

So, these are the first two corrections to be made, and again, we refer to Feynman [35], Griffiths [45], Weinberg [91] for details. Obviously we don't quite know what we're doing here, but let us add now the above corrections to E_n, and see what we get. We obtain in this way one of the most famous formulae in quantum mechanics, namely:

THEOREM 16.43. *The energy levels of the hydrogen atom, taking into account the fine structure coming from the relativistic and spin-related correction, are given by*

$$E_{nj} = E_n\left[1 + \frac{\alpha^2}{n^2}\left(\frac{n}{j+1/2} - \frac{3}{4}\right)\right]$$

with $j = l \pm 1/2$ being as above, and with α being the fine structure constant.

PROOF. We have the following computation, based on the above formulae:

$$\begin{aligned}\mathcal{E}_{nl}+\mathcal{E}_{nj} &= \frac{\alpha^2E_n}{n^2}\left(\frac{n}{l+1/2}-\frac{3}{4}-\frac{n(j-l)}{(l+1/2)(j+1/2)}\right)\\ &= \frac{\alpha^2E_n}{n^2}\left(\frac{n}{l+1/2}\left(1-\frac{j-l}{j+1/2}\right)-\frac{3}{4}\right)\\ &= \frac{\alpha^2E_n}{n^2}\left(\frac{n}{j+1/2}-\frac{3}{4}\right)\end{aligned}$$

Thus the corrected formula of the energy is as follows:

$$E_{nj}=E_n+\mathcal{E}_{nl}+\mathcal{E}_{nj}=E_n+\frac{\alpha^2E_n}{n^2}\left(\frac{n}{j+1/2}-\frac{3}{4}\right)$$

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

Summarizing, quantum mechanics is more complicated than what originally appears from Schrödinger's solution of the hydrogen atom. And the story is not over here, because on top of the above fine structure correction, which is of order α^2, we have afterwards the Lamb shift, which is an order α^3 correction, then the hyperfine splitting, and more.

And we will end this book with this. It's been a pleasure to have you here, and in the hope that.. but wait, cat is here, meowing something, let's see what she has to say:

CAT 16.44. *What is the formula of α?*

Well, no idea, katie, but wasn't it supposed to be the other way around, with me asking difficult physics questions, and you answering? In any case, good question, I guess.

16e. Exercises

Congratulations for having read this book, and no exercises for this final chapter. However, if interested in learning more, in mathematics and physics, you can have a look at the various books referenced below. As for some good questions to work on, in mathematics you can try to solve $\sum_n n^z=0$ over the complex numbers, a good question due to Riemann, and in physics you can try to compute the fine structure constant α, as suggested above by cat. Be said in passing, no one really knows which of these two questions is simpler, this is I guess something that you will have to figure out.

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 (2026).

[12] T. Banica, Introduction to modern physics (2025).

[13] T. Banica, The joys of relativity theory (2026).

[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