Stone Age mathematics

Teo Banica

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

2010 *Mathematics Subject Classification.* 01A67

Key words and phrases. Algebra, Geometry

Abstract. This is an introduction to mathematics, with emphasis on algebra and geometry. We first discuss numbers, fractions and percentages, and their basic applications, followed by real numbers, and with a look into philosophy and logic too. Then we get into plane geometry and trigonometry, and coordinates and some space geometry. We then go back to numbers, with more advanced theory, in relation with divisibility, prime numbers and related topics, and polynomials and their roots. Finally, we provide an introduction to functions and analysis, with the basics of the theory, followed by exponentials, logarithms and more trigonometry, and with the derivatives explained too.

Preface

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

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

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

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

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

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

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

Part III goes back to numbers, which remain something extremely useful, and discusses more advanced aspects of them, sometimes in relation with geometry.

Part IV is an introduction to more advanced mathematics, namely functions and analysis, again with motivations coming from craftsmanship, and geometry.

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

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

Cergy, January 2025
Teo Banica

Contents

Part I

Numbers

Oh, Shenandoah
I long to hear you
Look away, we're bound away
Across the wide Missouri

CHAPTER 1

Numbers

1a. Numbers

You certainly know a bit about numbers $1,2,3,4,\ldots$, and we will be here, with this book, for learning more about them. Many things can be said here, but instead of starting right away with some complicated mathematics, it is wiser to relax, and go back to these small numbers $1,2,3,4,\ldots$ that you know well, and have some more thinking at them. After all, these small numbers are something quite magic, worth some more thinking. And with the thinking work that we will be doing here being something useful.

So, reviewing the material from elementary school. Shall we start with 7×8, or perhaps with 6×7? I don't know about you, but personally I found these two computations both quite difficult, as a kid, these multiples of 7 are no joke, when learning arithmetic.

In answer, these are indeed tough computations, forget about them, and let us start with the very basics. Here will be our method, which is quite philosophical:

Method 1.1. *In order to better understand the small numbers* $1,2,3,4,\ldots$ *and their arithmetic, the best is to forget about these numbers, and reinvent them. With this being guaranteed to work, an inventor being not supposed to ever forget his invention.*

Ready for this? Hang on, and getting started now, here we are, in the dark. It is actually most convenient here to do assume that we are in the dark, say in a Stone Age cavern, lit only by a small fire, and with a pile of bloody ribs waiting to be counted, cooked, and eaten by our community. So, how to count these bloody ribs?

As a simple solution, we can invent some words for counting, ribs or any other type of objects. And going here with English, here is a proposal, for our first numbers:

$$\text{one, two, three, four, } \ldots$$

However, this method obviously has some limitations, because the more objects we want to count, the more words we will have to invent for them, and this is not very funny. In fact, we even risk, as leaders, to be killed and eaten by the tribe, on the grounds that our mathematics is too complicated and annoying. Well, this is how things were going during the Stone Age, people being honest and direct, nothing to do with the students nowadays, politely listening to whatever their math professor teaches them.

In short, we are in trouble here, and as problem to be solved, we have:

Problem 1.2. *Words are not very good for counting, we must invent something else, say some sort of bizarre signs.*

So, let us attempt to invent some suitable signs, doing the counting. The first thought here goes to the ribs themselves, that we want to count, which can be designated, pictorially, by vertical bars $|$. And with this, we certainly have our improved numeration system, which starts as follows, and can be continued indefinitely:

$$|\,,\ ||\,,\ |||\,,\ ||||\,,\ \ldots$$

However, there are still some bugs, with this new system, which remains not very practical for big numbers, say when counting small fruits. In addition, it is a bit of a pity to completely give up language, and to have no words for our signs, after all our one, two, three, four were not that bad, for the small numbers, and we are missing them.

A good solution to this, again by thinking at ribs, comes by thinking as well at the animals these ribs come from. Indeed, and by going now a bit abstract, we can group ribs into animals, and we reach to an even better numeration system. Moreover, animals can be grouped into herds, and so on, and with this, we reach to something even better.

Here comes now a big discovery, in relation with all this:

Discovery 1.3. *For best results with our system, it is ideal to assume that the number of ribs of an animal equals the number of animals in a herd.*

Which is, obviously, something quite far-reaching, answering most of the questions formulated above. But the next question is now, what should be this common number that we are using, of ribs of an animal, or of animals in a herd, and so on?

This is a quite subtle question, whose answer is not obvious, and this even if you know well math, as many of our ancestors did, over the centuries, so let us work out some examples. As a first example here, which is something a bit formal, we have:

Example 1.4. *Numeration basis two.*

Many things can be said here, and we can even start, with this, to do some serious mathematics, with tables and rules for addition and multiplication, and for substraction and division too, and with many other interesting things that can be said, about this.

As a comment here, this system is not that unuseful or obsolote, because this is more or less what computer scientists are using, nowadays. But more on this later.

Next on our list, coming natually after numeration basis two, is of course:

Example 1.5. *Numeration basis three.*

As before, many things can be said here. Pros and cons. Note in passing that we are learning good mathematics here, with our numeration systems.

Coming next, we have:

Example 1.6. *Numeration basis four.*

This is somehow better than numeration basis two. Good to know.

Coming next, we have:

Example 1.7. *Numeration basis five.*

Quite interesting, and nice pictorially, still used on prison walls.

Coming next, we have:

Example 1.8. *Numeration basis six.*

Again, this is something quite interesting, nicely mixing two and three.

And we will stop here with our list of examples, with seven being reputed to be something quite complicated, as a number, better not mess with it. But the question comes now, which system to use? And we have here several schools of thought:

(1) Numeration basis two, or better, four, or even better, eight, or perhaps even sixteen, or why not sixty-four, are something very natural and useful. In practice, and in view of what we can do, and what we can't, the choice is between eight and sixteen.

(2) Numeration basis three, or much better, because even, six, or why now twelve, or twenty-four are something natural and useful too. In practice now, again in view of what we can do, and what we can't, the choice here is between six and twelve.

(3) Finally, we have numeration basis five, or much better, because even, ten. Not very clear what the advantage of using ten would be, but at least, as an interesting observation, at least there is no dillema here, with fifty being barred, as being too big.

So, this was for the story of the bases of numeration, and in what follows we will use, as everyone or almost nowadays, basis ten, somehow for the reasons discussed above. As for the ten digits needed, my proposal would be to use the following signs:

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 0$$

And with this, we are ready to go, into some serious arithmetic.

1b.

1c.

1d.

1e. Exercises

Exercises:

Exercise 1.9.

Exercise 1.10.

Exercise 1.11.

Exercise 1.12.

Exercise 1.13.

Exercise 1.14.

Exercise 1.15.

Exercise 1.16.

Bonus exercise.

CHAPTER 2

Fractions

2a. Fractions

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

Definition 2.1. *The rational numbers are the quotients of type*

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

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

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

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

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

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

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

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

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

Theorem 2.2. *The number of possibilities of choosing k objects among n objects is*

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

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

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

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

However, when thinking well, the number N that we computed is in fact the number of possibilities of choosing k ordered objects among n objects. Thus, we must divide everything by the number M of orderings of the k objects that we chose:

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

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

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

And this is the correct answer, because, well, that is how things are. □

As an important adding to Theorem 2.2, we should mention that, by definition, we must declare that $0!=1$, as for the following computation to work:

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

Going ahead now with more mathematics and less philosophy, with Theorem 2.2 complemented by this convention being in final form, we have:

Theorem 2.3. *We have the binomial formula*

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

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^kb^{n-k}. Thus, we have a formula as follows:

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

In order to finish, it remains to compute the coefficients C_k. But, according to our product formula, C_k is the number of choices for the k needed a variables among the n available a variables. Thus, according to Theorem 2.2, we have:

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

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

Theorem 2.3 is something quite interesting, so let us doublecheck it with some numerics. At small values of n we obtain the following formulae, which are all correct:

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

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

Theorem 2.4. *The Pascal triangle, formed by the binomial coefficients* $\binom{n}{k}$,

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

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

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

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

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

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

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

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

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

By using the binomial formula, this formula becomes:

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

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

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

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

As an application to this, let us do some probability. We first have:

Theorem 2.5. *The probabilities at poker are as follows:*

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

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

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

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

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

Thus, the probability of having a pair is:

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

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

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

Thus, the probability of having two pairs is:

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

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

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

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

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

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

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

Thus, the probability of having full house is:

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

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

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

Thus, the probability of having a straight is:

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

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

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

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

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

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

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

Thus, the probability of having a flush is:

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

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

$$N=4\cdot 4$$

Thus, the probability of having a straight flush is:

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

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

Here is now a theorem about flipping coins:

Theorem 2.6. *When flipping a coin k times what you can win are quantities of type $\$k-2s$, with $s=0,1,\ldots,k$, with the probability for this to happen being:*

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

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

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

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

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

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

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

$$\begin{aligned}\sum_{s=0}^kP(k-2s)&=\sum_{s=0}^k\frac{1}{2^k}\binom{k}{s}\\&=\frac{1}{2^k}\sum_{s=0}^k\binom{k}{s}\\&=\frac{1}{2^k}\sum_{s=0}^k\binom{k}{s}1^s1^{k-s}\\&=\frac{1}{2^k}(1+1)^k\\&=\frac{1}{2^k}\times2^k\\&=1\end{aligned}$$

But shall we really trust this. Imagine for instance that you play your game for \$1000 instead of \$1 as basic gain, your life is obviously at stake, so all this is worth a second doublecheck, before being used in practice. So, as second doublecheck, let us verify that, on average, what you win is exactly \$0, which is something very intuitive, the game itself obviously not favoring you, nor your partner. But this can be checked as follows:

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

Here we have used a change of indices, namely $s=k-t$, along with the following formula, which is clear from the definition of binomial coefficients:

$$\binom{k}{t}=\binom{k}{k-t}$$

Summarizing, we have a good and valid theorem here, ready to be used in practice. □

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

2b.

2c.

2d.

2e. Exercises

Exercises:

Exercise 2.7.

Exercise 2.8.

Exercise 2.9.

Exercise 2.10.

Exercise 2.11.

Exercise 2.12.

Exercise 2.13.

Exercise 2.14.

Bonus exercise.

CHAPTER 3

Real numbers

3a. Real numbers

In more advanced mathematical terms, the operations on the rationals, namely sum, product and inversion, tell us that $\mathbb{Q}$ is a field, in the following sense:

Definition 3.1. *A field is a set F with a sum operation $+$ and a product operation $\times$, subject to the following conditions:*

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

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

$$1+1=0$$

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

Theorem 3.2. *The following happen:*

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

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

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

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

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

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

Getting back now to our philosophical discussion regarding numbers, what we have in Theorem 3.2 is not exactly good news, suggesting that, on purely mathematical grounds, there is a certain rivalry between $\mathbb{Q}$ and $\mathbb{F}_p$, as being the simplest field. So, which of them shall we study, as being created first? Not an easy question, and as answer, we have:

Answer 3.3. *Ignoring what pure mathematics might say, and trusting instead physics and chemistry, we will choose to trust in $\mathbb{Q}$, as being the simplest field.*

In short, welcome to science, and with this being something quite natural for us, science being the topic of the present book. Moving ahead now, many things can be done with $\mathbb{Q}$, but getting straight to the point, one thing that fails is solving $x^2=2$:

Theorem 3.4. *The field $\mathbb{Q}$ does not contain a square root of 2:*

$$\sqrt{2}\notin\mathbb{Q}$$

In fact, among integers, only the squares, $n=m^2$ with $m\in\mathbb{N}$, have square roots in $\mathbb{Q}$.

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

(1) In what regards $\sqrt{2}$, assuming that $r=a/b$ with $a,b\in\mathbb{N}$ prime to each other satisfies $r^2=2$, we have $a^2=2b^2$, and so $a\in 2\mathbb{N}$. But then by using again $a^2=2b^2$ we obtain $b\in 2\mathbb{N}$ as well, which contradicts our assumption $(a,b)=1$.

(2) Along the same lines, any prime number $p\in\mathbb{N}$ has the property $\sqrt{p}\notin\mathbb{Q}$, with the proof here being as the above one for $p=2$, by congruence and contradiction.

(3) More generally, our claim is that any $n\in\mathbb{N}$ which is not a square has the property $\sqrt{n}\notin\mathbb{Q}$. Indeed, we can argue here that our number decomposes as $n=p_1^{a_1}\ldots p_k^{a_k}$, with $p_1,\ldots,p_k$ distinct primes, and our assumption that n is not a square tells us that one of the exponents $a_1,\ldots,a_k\in\mathbb{N}$ must be odd. Moreover, by extracting all the obvious squares from n, we can in fact assume $a_1=\ldots=a_k=1$. But with this done, we can set $p=p_1$, and the congruence argument from (2) applies, and gives $\sqrt{n}\notin\mathbb{Q}$, as desired. □

In short, in order to advance with our mathematics, we are in need to introduce the field of real numbers $\mathbb{R}$. You would probably say that this is very easy, via decimal writing, like everyone does, but before doing that, let me ask you a few questions:

(1) Honestly, do you really like the addition of real numbers, using the decimal form? Let us take, as example, the following computation:

$$12.456\,783\,872$$

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

This computation can surely be done, but, annoyingly, it must be done from right to left, instead of left to right, as we would prefer. I mean, personally I would be most interested in knowing first what happens at left, if the integer part is 39 or 40, but go do all the computation, starting from the right, in order to figure out that. In short, my feeling is that this addition algorithm, while certainly good, is a bit deceiving.

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

$$12.456\,783\,872$$

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

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

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

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

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

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

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

$$\ldots$$

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

FACT 3.5. *The real numbers $x \in \mathbb{R}$ can be certainly introduced via their decimal form, but with this, the field structure of $\mathbb{R}$ remains something quite unclear.*

Well, it looks like we are a bit stuck. Fortunately, there is a clever solution to this, due to Dedekind. His definition for the real numbers is as follows:

Definition 3.6. *The real numbers $x \in \mathbb{R}$ are formal cuts in the set of rationals,*

$$\mathbb{Q} = A_x \sqcup B_x$$

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

$$p \in A_x \ , \ q \in B_x \implies p < q \qquad , \qquad \inf B_x \notin B_x$$

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

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

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

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

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

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

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

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

Theorem 3.7. *The equation $x^2 = 2$ has two solutions over the real numbers, namely the positive solution, denoted $\sqrt{2}$, and its negative counterpart, which is $-\sqrt{2}$.*

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

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

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

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

Theorem 3.8. *The solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{R}$ are*

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

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

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

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

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

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

THEOREM 3.9. *The real numbers $x\in\mathbb{R}$ can be written in decimal form,*

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

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

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

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

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

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

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

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

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

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

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

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

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

$$A_x=\bigcup_{n\geq1}\left\{p\in\mathbb{Q}\Big|p\leq 0.b_1\ldots b_n\right\}$$

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

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

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

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

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

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

The above theorem was of course quite difficult, but this is how things are. Let us record as well the following result, coming as a useful complement to the above:

Theorem 3.10. *A real number $r\in\mathbb{R}$ is rational precisely when*

$$r=\pm a_1\ldots a_m.b_1\ldots b_n(c_1\ldots c_p)$$

that is, when its decimal writing is periodic.

Proof. In one sense, this follows from the following computation, which shows that a number as in the statement is indeed rational:

$$\begin{aligned}
r &= \pm\frac{1}{10^n}\,a_1\ldots a_mb_1\ldots b_n.c_1\ldots c_pc_1\ldots c_p\ldots\\
&= \pm\frac{1}{10^n}\left(a_1\ldots a_mb_1\ldots b_n+c_1\ldots c_p\left(\frac{1}{10^p}+\frac{1}{10^{2p}}+\ldots\right)\right)\\
&= \pm\frac{1}{10^n}\left(a_1\ldots a_mb_1\ldots b_n+\frac{c_1\ldots c_p}{10^p-1}\right)
\end{aligned}$$

As for the converse, given a rational number $r=k/l$, we can find its decimal writing by performing the usual division algorithm, k divided by l. But this algorithm will be surely periodic, after some time, so the decimal writing of r is indeed periodic, as claimed. □

At a more advanced level, passed the rationals, our problem remains the same, namely how to recognize the arithmetic properties of the real numbers $r\in\mathbb{R}$, as for instance being square roots of rationals, and so on, when written in decimal form. Many things can be said here, and for more on all this, we refer to Part III of the present book.

Getting back now to Theorem 3.9, that was definitely something quite difficult. Alternatively, we have the following definition for the real numbers:

Theorem 3.11. *The field of real numbers $\mathbb{R}$ can be defined as well as the completion of $\mathbb{Q}$ with respect to the usual distance on the rationals, namely*

$$d\left(\frac{a}{b},\frac{c}{d}\right)=\left|\frac{a}{b}-\frac{c}{d}\right|$$

and with the operations on $\mathbb{R}$ coming from those on $\mathbb{Q}$, via Cauchy sequences.

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

(1) Getting back to chapter 2, we know from there what the rational numbers are. But, as a continuation of the material there, we can talk about the distance between such rational numbers, as being given by the formula in the statement, namely:

$$d\left(\frac{a}{b},\frac{c}{d}\right)=\left|\frac{a}{b}-\frac{c}{d}\right|=\frac{|ad-bc|}{|bd|}$$

(2) Very good, so let us get now into Cauchy sequences. We say that a sequence of rational numbers $\{r_n\}\subset\mathbb{Q}$ is Cauchy when the following condition is satisfied:

$$\forall\varepsilon>0,\exists N\in\mathbb{N},m,n\geq N\implies d(r_m,r_n)<\varepsilon$$

Here of course $\varepsilon\in\mathbb{Q}$, because we do not know yet what the real numbers are.

(3) With this notion in hand, the idea will be to define the reals $x\in\mathbb{R}$ as being the limits of the Cauchy sequences $\{r_n\}\subset\mathbb{Q}$. But since these limits are not known yet to

exist to us, precisely because they are real, we must employ a trick. So, let us define instead the reals $x \in \mathbb{R}$ as being the Cauchy sequences $\{r_n\} \subset \mathbb{Q}$ themselves.

(4) The question is now, will this work. As a first observation, we have an inclusion $\mathbb{Q} \subset \mathbb{R}$, obtained by identifying each rational $r \in \mathbb{Q}$ with the constant sequence $r_n = r$. Also, we can sum and multiply our real numbers in the obvious way, namely:

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

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

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

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

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

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

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

Very nice all this, so have have two equivalent definitions for the real numbers. Finally, getting back to the decimal writing approach, that can be recycled too, with some analysis know-how, and we have a third possible definition for the real numbers, as follows:

Theorem 3.12. *The real numbers $\mathbb{R}$ can be defined as well via the decimal form*

$$x = \pm a_1 \ldots a_n . a_{n+1} a_{n+2} a_{n+3} \ldots\ldots$$

with $a_i \in \{0, 1, \ldots, 9\}$, with the usual convention for such numbers, namely

$$\ldots a999 \ldots = \ldots (a+1)000 \ldots$$

and with the sum and multiplication coming by writing such numbers as

$$x = \pm \sum_{k \in \mathbb{Z}} a_k 10^{-k}$$

and then summing and multiplying, in the obvious way.

Proof. This is something which looks quite intuitive, but which in practice, and we insist here, is not exactly beginner level, the idea with this being as follows:

(1) Let us first forget about the precise decimal writing in the statement, and define the real numbers $x \in \mathbb{R}$ as being formal sums as follows, with the sum being over integers $k \in \mathbb{Z}$ assumed to be greater than a certain integer, $k \geq k_0$:

$$x = \pm \sum_{k \in \mathbb{Z}} a_k 10^{-k}$$

(2) Now by truncating, we can see that what we have here are certain Cauchy sequences of rationals, and with a bit more work, we conclude that the $\mathbb{R}$ that we constructed is precisely the $\mathbb{R}$ that we constructed in Theorem 3.11. Thus, we get the result.

(3) Alternatively, by getting back to Theorem 3.9 and its proof, we can argue, based on that, that the $\mathbb{R}$ that we constructed coincides with the old $\mathbb{R}$ from Definition 3.6, the one constructed via Dedekind cuts, and this gives again all the assertions. □

Many things can be said about rationals and irrationals, and we have:

Theorem 3.13. *The number e from analysis, given by*

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

which numerically means $e = 2.7182818284\ldots$, is irrational.

Proof. Following Fourier, we will do this by contradiction. So, assume $e = m/n$, with $m, n \in \mathbb{N}$, and let us look at the following number:

$$x = n!\left(e - \sum_{k=0}^{n} \frac{1}{k!}\right)$$

As a first observation, x is an integer, as shown by the following computation:

$$\begin{aligned} x &= n!\left(\frac{m}{n} - \sum_{k=0}^{n} \frac{1}{k!}\right) \\ &= m(n-1)! - \sum_{k=0}^{n} n(n-1)\ldots(n-k+1) \\ &\in \mathbb{Z} \end{aligned}$$

On the other hand $x > 0$, and we have as well the following estimate:

$$\begin{aligned}
x &= n!\sum_{k=n+1}^\infty\frac{1}{k!}\\
&= \frac{1}{n+1}+\frac{1}{(n+1)(n+2)}+\ldots\\
&< \frac{1}{n+1}+\frac{1}{(n+1)^2}+\ldots\\
&= \frac{1}{n}
\end{aligned}$$

Thus $x \in (0,1)$, which contradicts our previous finding $x \in \mathbb{Z}$, as desired. □

We will be back to this in Part IV of the present book, when doing analysis.

3b.

3c.

3d.

3e. Exercises

Exercises:

Exercise 3.14.

Exercise 3.15.

Exercise 3.16.

Exercise 3.17.

Exercise 3.18.

Exercise 3.19.

Exercise 3.20.

Exercise 3.21.

Bonus exercise.

CHAPTER 4

About infinity

4a. About infinity

Let us start with some set theory. Many things can be said here, and as a very useful result, which leads to non-trivial consequences, we have the inclusion-exclusion principle. And, as a beautiful application of this inclusion-exclusion principle, we have:

Theorem 4.1. *The probability for a random $\sigma \in S_N$ to have no fixed points is*

$$P \simeq \frac{1}{e}$$

in the $N \to \infty$ limit, where $e = 2.7182\ldots$ is the usual constant from analysis. More generally, the main character of S_N, which counts the fixed points, and is given by

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

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

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

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

Proof. Many things going on here, the idea being as follows:

(1) In order to prove the first assertion, which is the key, and probably the most puzzling one, we will use the inclusion-exclusion principle. Let us set:

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

The set of permutations having no fixed points, called derangements, is then:

$$X_N = \left(\bigcup_k S_N^k \right)^c$$

Now the inclusion-exclusion principle tells us that we have:

$$\begin{aligned}
|X_N| &= \left|\left(\bigcup_k S_N^k\right)^c\right|\\
&= |S_N| - \sum_k |S_N^k| + \sum_{k<l}|S_N^k\cap S_N^l| - \ldots + (-1)^N \sum_{k_1<\ldots<k_N}|S_N^{k_1}\cup\ldots\cup S_N^{k_N}|\\
&= N! - N(N-1)! + \binom{N}{2}(N-2)! - \ldots + (-1)^N\binom{N}{N}(N-N)!\\
&= \sum_{r=0}^N (-1)^r\binom{N}{r}(N-r)!
\end{aligned}$$

Thus, the probability that we are interested in, for a random permutation $\sigma\in S_N$ to have no fixed points, is given by the following formula:

$$P = \frac{|X_N|}{N!} = \sum_{r=0}^N \frac{(-1)^r}{r!}$$

Since on the right we have the expansion of $1/e$, this gives the result.

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

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

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

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

We already know, from (1), that this formula holds at $r=0$. In the general case now, we have to count the permutations $\sigma\in S_N$ having exactly r points. Now since having such a permutation amounts in choosing r points among $1,\ldots,N$, and then permuting the $N-r$ points left, without fixed points allowed, we have:

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

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

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

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

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

Thus, we obtain as limiting measure the Poisson law of parameter 1, as stated.

(3) Finally, let us construct the truncated characters of S_N, as in the statement. As before in the case $t=1$, we have the following computation, coming from definitions:

$$\chi_t(\sigma)=\sum_{i=1}^{[tN]}\delta_{\sigma(i)i}=\#\left\{i\in\{1,\ldots,[tN]\}\Big|\sigma(i)=i\right\}$$

Also before in the proofs of (1) and (2), we obtain by inclusion-exclusion that:

$$\begin{aligned}P(\chi_t=0)&=\frac{1}{N!}\sum_{r=0}^{[tN]}(-1)^r\sum_{k_1<\ldots<k_r<[tN]}|S_N^{k_1}\cap\ldots\cap S_N^{k_r}|\\&=\frac{1}{N!}\sum_{r=0}^{[tN]}(-1)^r\binom{[tN]}{r}(N-r)!\\&=\sum_{r=0}^{[tN]}\frac{(-1)^r}{r!}\cdot\frac{[tN]!(N-r)!}{N!([tN]-r)!}\end{aligned}$$

Now with $N\to\infty$, we obtain from this the following estimate:

$$P(\chi_t=0)\simeq\sum_{r=0}^{[tN]}\frac{(-1)^r}{r!}\cdot t^r\simeq e^{-t}$$

More generally, by counting the permutations $\sigma\in S_N$ having exactly r fixed points among $1,\ldots,[tN]$, as in the proof of (2), we obtain:

$$P(\chi_t=r)\simeq\frac{t^r}{r!e^t}$$

Thus, we obtain in the limit a Poisson law of parameter t, as stated. □

Now back to our number theory business, with our accumulated set theory knowledge, we can talk about cardinalities of various sets, ordinal numbers, and generally speaking, talk about ∞ in all its flavors. There are many interesting questions here.

4b.

4c.

4d.

4e. Exercises

Exercises:

Exercise 4.2.

Exercise 4.3.

Exercise 4.4.

Exercise 4.5.

Exercise 4.6.

Exercise 4.7.

Exercise 4.8.

Exercise 4.9.

Bonus exercise.

Part II

Geometry

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

CHAPTER 5

Triangles

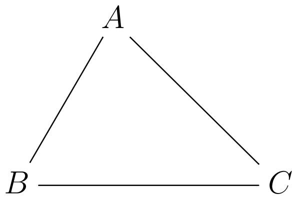
5a. Triangles

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

Theorem 5.1. *Given a triangle ABC, the following happen:*

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

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



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

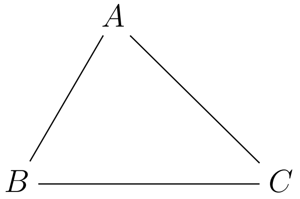
(1) Come with a small circle, inside ABC, and then inflate it, as to touch all 3 edges. The center of the circle will be then at equal distance from all 3 edges, so it will lie on all 3 angle bisectors. Thus, we have constructed the incenter, as required.

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

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

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

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



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

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

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

Fact 5.2. *Besides the above* 4 *centers, many more remarkable points can be associated to a triangle ABC, and most of these lie on a line, called Euler line of ABC.*

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

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

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

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

Definition 5.3. *A projective space is a space consisting of points and lines, subject to the following conditions:*

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

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

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

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

Thus, what we have is a projective space in the sense of Definition 5.3. More generally now, we have the following construction, in arbitrary dimensions:

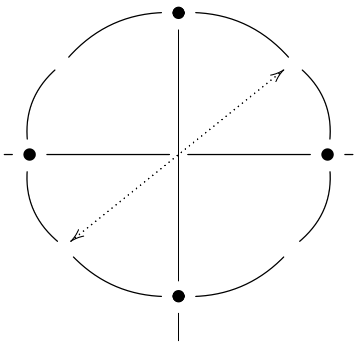
Theorem 5.4. *We can define the projective space $P^{N-1}_\mathbb{R}$ as being the space of lines in $\mathbb{R}^N$ passing through the origin, and in small dimensions:*

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

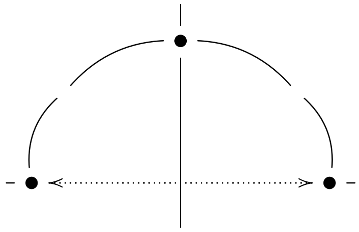
Proof. We have several assertions here, with all this being of course a bit informal, and self-explanatory, the idea and some further details being as follows:

(1) To start with, the fact that the space $P^{N-1}_\mathbb{R}$ constructed in the statement is indeed a projective space in the sense of Definition 5.3 follows from definitions, exactly as in the discussion preceding the statement, regarding the case $N = 3$.

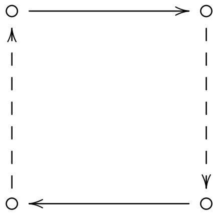
(2) At $N = 2$ now, a line in $\mathbb{R}^2$ passing through the origin corresponds to 2 opposite points on the unit circle $\mathbb{T} \subset \mathbb{R}^2$, according to the following scheme:



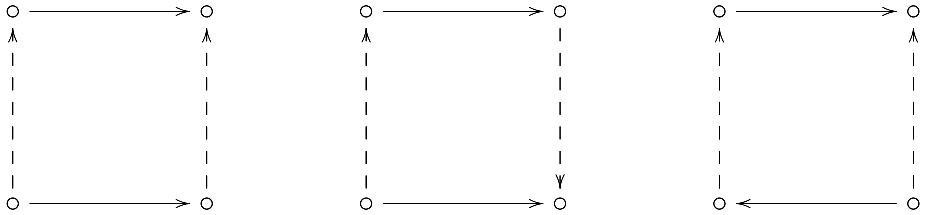
Thus, $P^1_\mathbb{R}$ corresponds to the upper semicircle of $\mathbb{T}$, with the endpoints identified, and so we obtain a circle, $P^1_\mathbb{R} = \mathbb{T}$, according to the following scheme:



(3) At $N = 3$, the space $P^2_\mathbb{R}$ corresponds to the upper hemisphere of the sphere $S^2_\mathbb{R} \subset \mathbb{R}^3$, with the points on the equator identified via $x = -x$. Topologically speaking, we can deform if we want the hemisphere into a square, with the equator becoming the boundary of this square, and in this picture, the $x = -x$ identification corresponds to a "identify opposite edges, with opposite orientations" folding method for the square:



(4) Thus, we have our space. In order to understand now what this beast is, let us look first at the other 3 possible methods of folding the square, which are as follows:



Regarding the first space, the one on the left, things here are quite simple. Indeed, when identifying the solid edges we get a cylinder, and then when further identifying the dotted edges, what we get is some sort of closed cylinder, which is a torus.

(5) Regarding the second space, the one in the middle, things here are more tricky. Indeed, when identifying the solid edges we get again a cylinder, but then when further identifying the dotted edges, we obtain some sort of "impossible" closed cylinder, called Klein bottle. This Klein bottle obviously cannot be drawn in 3 dimensions, but with a bit of imagination, you can see it, in its full splendor, in 4 dimensions.

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

(7) With these preliminaries made, and getting back now to the projective space $P^2_{\mathbb{R}}$, we can see that this is something more complicated, of the same type, reminding the torus and the Klein bottle. So, we will call it "sort of twisted sphere", as in the statement, and exercise for you to figure out how this beast looks like, in 4 dimensions. □

All this is quite exciting, and reminds childhood and primary school, but is however a bit tiring for our neurons, guess that is pure mathematics. It is possible to come up with some explicit formulae for the embedding $P^2_{\mathbb{R}} \subset \mathbb{R}^4$, which are useful in practice, allowing us to do some analysis over $P^2_{\mathbb{R}}$, and we will leave this as an instructive exercise.

All this is very interesting, but we will pause our study here, because we still have many other things to say. Getting now to finite fields, we have:

Theorem 5.5. *Given a field F, we can talk about the projective space P^{N-1}_F, as being the space of lines in F^N passing through the origin. At $N = 3$ we have*

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

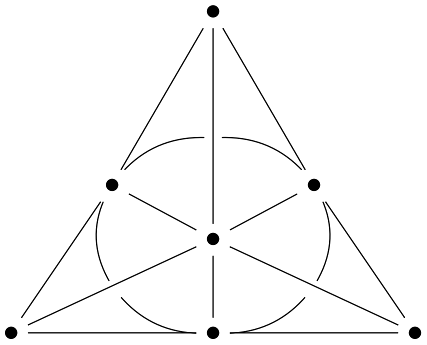
where $q = |F|$, in the case where our field F is finite.

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

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

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

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



Here the circle in the middle is by definition a line, and with this convention, the basic axioms in Definition 5.3 are satisfied, in the sense that any two points determine a line, and any two lines determine a point. And isn't this beautiful.

5b.

5c.

5d.

5e. Exercises

Exercises:

Exercise 5.6.

Exercise 5.7.

Exercise 5.8.

Exercise 5.9.

Exercise 5.10.

Exercise 5.11.

Exercise 5.12.

Exercise 5.13.

Bonus exercise.

CHAPTER 6

Trigonometry

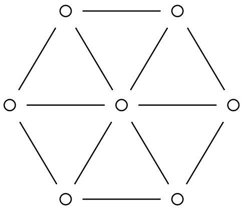
6a. Trigonometry

Let us get now into angles and trigonometry. For this purpose, the best is to talk first about circles and π. And here, to start with, we have the following result:

Theorem 6.1. *The following two definitions of π are equivalent:*

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

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



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

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

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

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

Getting now to trigonometry, the basics here are as follows:

Theorem 6.2. *The following happen:*

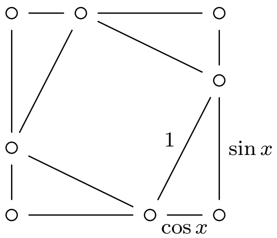
(1) *We can talk about angles $x \in \mathbb{R}$, by using the unit circle, in the usual way, and in this correspondence, the right angle has a value of $\pi/2$.*

(2) *Associated to any $x \in \mathbb{R}$ are numbers $\sin x, \cos x \in \mathbb{R}$, constructed in the usual way, by using a triangle. These numbers satisfy $\sin^2 x + \cos^2 x = 1$.*

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

(1) The formula $L = 2\pi$ from Theorem 6.1 shows that the length of a quarter of the unit circle is $l = \pi/2$, and so the right angle has indeed this value, $\pi/2$.

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



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

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

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

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

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

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

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

Theorem 6.3. *The sines and cosines of sums are given by*

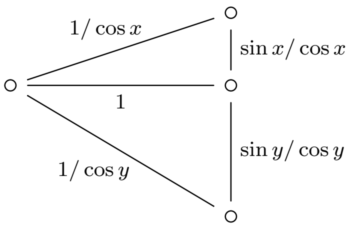
$$\sin(x + y) = \sin x \cos y + \cos x \sin y$$

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

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

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

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



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

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

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

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

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

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

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

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

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

6b.

6c.

6d.

6e. Exercises

Exercises:

Exercise 6.4.

Exercise 6.5.

Exercise 6.6.

Exercise 6.7.

Exercise 6.8.

Exercise 6.9.

Exercise 6.10.

Exercise 6.11.

Bonus exercise.

CHAPTER 7

Coordinates

7a. Coordinates

Looking up, to the sky, the first thing that you see is the Sun, seemingly moving around the Earth on a circle, but a more careful study reveals that this circle is rather a deformed circle, called ellipsis. As for the other stars and planets, these have all sort of weird trajectories, but a more careful study reveals that, with due attention to what the best "center" is, replacing our Earth, the trajectories are often ellipses:

(1) Indeed, this applies to all the planets in our Solar System, which move around the biggest object in the system, which is by far the Sun, on ellipses.

(2) The same trick applies to the trajectories of various distant stars, the rule being always the same, "small moves around big, on an ellipsis".

(3) However, there are counterexamples too, such as asteroids reaching our Solar system, but then travelling outwards, never to be seen again.

Summarizing, modulo some annoying asteroids that we will leave for later, we are led in this way to ellipses, and their mathematics. And good news, a full theory of ellipses is available, and this since the ancient Greeks, whose main findings were as follows:

Theorem 7.1. *The ellipses, taken centered at the origin* 0*, and squarely oriented with respect to* Oxy*, can be defined in* 4 *possible ways, as follows:*

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

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

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

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

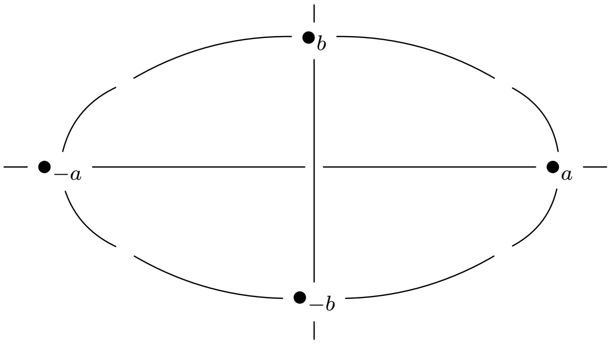
(3) *As the curves appearing when drawing a circle, from various perspectives:*

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

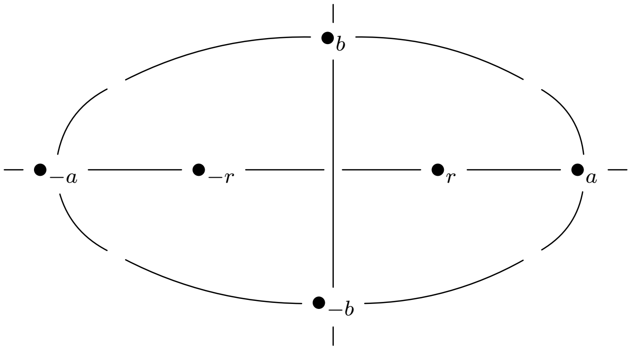
(4) *As the closed non-degenerate curves appearing by cutting a cone with a plane.*

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

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



(2) Let us prove now that such an ellipsis has two focal points, as stated in (2). We must look for a number $r > 0$, and a number $l > 0$, such that our ellipsis appears as $d(z,p) + d(z,q) = l$, with $p = (0, -r)$ and $q = (0, r)$, according to the following picture:



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

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

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

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

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

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

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

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

(6) Thus, our result holds indeed, and with the numbers $l, r > 0$ appearing, and no surprise here, via the formulae $l = 2a$ and $r = \sqrt{a^2 - b^2}$, found in (3) above.

(7) Getting back now to our theorem, we have two other assertions there at the end, labelled (3,4). But, thinking a bit, these assertions are in fact equivalent, and in what concerns us, we will rather focus on (4), which looks more mathematical. And in what regards this assertion (4), this can be established indeed, by doing some 3D computations, that we will leave here as an instructive exercise, for you. And with the promise that we will come back to this in a moment, with a full proof, in a more general setting. □

All this is very nice, but before getting into physics, with some explanations for the fact that planets travel indeed on ellipses, which is something that we must surely understand, before going with some further math, 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, also since the ancient Greeks, summarized as follows:

Theorem 7.2. *The conics, which are the algebraic curves of degree* 2 *in the plane,*

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

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

Proof. This follows by further building on Theorem 7.1, as follows:

(1) Let us first classify the conics up to non-degenerate linear transformations of the plane, which are by definition transformations as follows, with $\det A\neq 0$:

$$\begin{pmatrix}x\\y\end{pmatrix}\to A\begin{pmatrix}x\\y\end{pmatrix}$$

Our claim is that as solutions we have the circles, parabolas, hyperbolas, along with some degenerate solutions, namely $\emptyset$, points, lines, pairs of lines, $\mathbb R^2$.

(2) As a first remark, it looks like we forgot precisely the ellipses, but via linear transformations these become circles, so things fine. As a second remark, all our claimed solutions can appear. Indeed, the circles, parabolas, hyperbolas can appear as follows:

$$x^2+y^2=1\quad,\quad x^2=y\quad,\quad xy=1$$

As for $\emptyset$, points, lines, pairs of lines, $\mathbb R^2$, these can appear too, as follows, and with our polynomial P chosen, whenever possible, to be of degree exactly 2:

$$x^2=-1\quad,\quad x^2+y^2=0\quad,\quad x^2=0\quad,\quad xy=0\quad,\quad 0=0$$

Observe here that, when dealing with these degenerate cases, assuming $\deg P=2$ instead of $\deg P\leq 2$ would only rule out $\mathbb R^2$ itself, which is not worth it.

(3) Getting now to the proof of our claim in (1), classification up to linear transformations, consider an arbitrary conic, written as follows, with $a,b,c,d,e,f\in\mathbb R$:

$$ax^2+by^2+cxy+dx+ey+f=0$$

Assume first $a\neq 0$. By making a square out of ax^2, up to a linear transformation in (x,y), we can get rid of the term cxy, and we are left with:

$$ax^2+by^2+dx+ey+f=0$$

In the case $b\neq 0$ we can make two obvious squares, and again up to a linear transformation in (x,y), we are left with an equation as follows:

$$x^2\pm y^2=k$$

In the case of positive sign, $x^2+y^2=k$, the solutions are the circle, when $k\geq 0$, the point, when $k=0$, and $\emptyset$, when $k<0$. As for the case of negative sign, $x^2-y^2=k$, which reads $(x-y)(x+y)=k$, here once again by linearity our equation becomes $xy=l$, which is a hyperbola when $l\neq 0$, and two lines when $l=0$.

(4) In the case $b \neq 0$ the study is similar, with the same solutions, so we are left with the case $a = b = 0$. Here our conic is as follows, with $c, d, e, f \in \mathbb{R}$:

$$cxy + dx + ey + f = 0$$

If $c \neq 0$, by linearity our equation becomes $xy = l$, which produces a hyperbola or two lines, as explained before. As for the remaining case, $c = 0$, here our equation is:

$$dx + ey + f = 0$$

But this is generically the equation of a line, unless we are in the case $d = e = 0$, where our equation is $f = 0$, having as solutions $\emptyset$ when $f \neq 0$, and $\mathbb{R}^2$ when $f = 0$.

(5) Thus, done with the classification, up to linear transformations as in (1). But this classification leads to the classification in general too, by applying now linear transformations to the solutions that we found. So, done with this, and very good.

(6) It remains to discuss the cone cutting. By suitably choosing our coordinate axes (x, y, z), we can assume that our cone is given by an equation as follows, with $k > 0$:

$$x^2 + y^2 = kz^2$$

In order to prove the result, we must in principle intersect this cone with an arbitrary plane, which has an equation as follows, with $(a, b, c) \neq (0, 0, 0)$:

$$ax + by + cz = d$$

(7) However, before getting into computations, observe that what we want to find is a certain degree 2 equation in the above plane, for the intersection. Thus, it is convenient to change the coordinates, as for our plane to be given by the following equation:

$$z = 0$$

(8) But with this done, what we have to do is to see how the cone equation $x^2+y^2 = kz^2$ changes, under this change of coordinates, and then set $z = 0$, as to get the (x, y) equation of the intersection. But this leads, via some thinking or computations, to the conclusion that the cone equation $x^2 + y^2 = kz^2$ becomes in this way a degree 2 equation in (x, y), which can be arbitrary, and so to the final conclusion in the statement. □

Ready for some physics? We have the following result:

Theorem 7.3. *Planets and other celestial bodies move around the Sun on conics,*

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

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

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

7b.

7c.

7d.

7e. Exercises

Exercises:

Exercise 7.4.

Exercise 7.5.

Exercise 7.6.

Exercise 7.7.

Exercise 7.8.

Exercise 7.9.

Exercise 7.10.

Exercise 7.11.

Bonus exercise.

CHAPTER 8

Space geometry

8a. Space geometry

Getting started with some applications, here is the notion what we will need:

Definition 8.1. *The vector product of two vectors in $\mathbb{R}^3$ is given by*

$$x\times y=||x||\cdot||y||\cdot\sin\theta\cdot n$$

where $n\in\mathbb{R}^3$ with $n\perp x,y$ and $||n||=1$ is constructed using the right-hand rule:

$$\begin{matrix} & \uparrow_{x\times y}\\ \leftarrow_x & \\ \swarrow_y & \end{matrix}$$

Alternatively, in usual vertical linear algebra notation for all vectors,

$$\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\times\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}x_2y_3-x_3y_2\\x_3y_1-x_1y_3\\x_1y_2-x_2y_1\end{pmatrix}$$

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 $0\to0$, 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.

In practice now, in order to get familiar with the vector products, nothing better than doing some classical mechanics. We have here the following key result:

Theorem 8.2. *In the gravitational 2-body problem, the angular momentum*

$$J = x \times p$$

with $p = mv$ being the usual momentum, is conserved.

Proof. There are several things to be said here, the idea being as follows:

(1) First of all the usual momentum, $p = mv$, is not conserved, because the simplest solution is the circular motion, where the moment gets turned around. But this suggests precisely that, in order to fix the lack of conservation of the momentum p, what we have to do is to make a vector product with the position x. Leading to J, as above.

(2) Regarding now the proof, consider indeed a particle m moving under the gravitational force of a particle M, assumed, as usual, to be fixed at 0. By using the fact that for two proportional vectors, $p \sim q$, we have $p \times q = 0$, we obtain:

$$\begin{aligned}\dot{J} &= \dot{x} \times p + x \times \dot{p}\\ &= v \times mv + x \times ma\\ &= m(v \times v + x \times a)\\ &= m(0 + 0)\\ &= 0\end{aligned}$$

Now since the derivative of J vanishes, this quantity is constant, as stated. □

As another basic application of the vector products, still staying with classical mechanics, we have all sorts of useful formulae regarding rotating frames. We first have:

Theorem 8.3. *Assume that a 3D body rotates along an axis, with angular speed w. For a fixed point of the body, with position vector x, the usual 3D speed is*

$$v = \omega \times x$$

where $\omega = wn$, with n unit vector pointing North. When the point moves on the body

$$V = \dot{x} + \omega \times x$$

is its speed computed by an inertial observer O on the rotation axis.

Proof. We have two assertions here, both requiring some 3D thinking, as follows:

(1) Assuming that the point is fixed, the magnitude of $\omega \times x$ is the good one, due to the following computation, with r being the distance from the point to the axis:

$$||\omega \times x|| = w||x|| \sin t = wr = ||v||$$

As for the orientation of $\omega \times x$, this is the good one as well, because the North pole rule used above amounts in applying the right-hand rule for finding n, and so ω, and this right-hand rule was precisely the one used in defining the vector products $\times$.

(2) Next, when the point moves on the body, the inertial observer O can compute its speed by using a frame (u_1, u_2, u_3) which rotates with the body, as follows:

$$\begin{aligned} V &= \dot{x}_1 u_1 + \dot{x}_2 u_2 + \dot{x}_3 u_3 + x_1 \dot{u}_1 + x_2 \dot{u}_2 + x_3 \dot{u}_3 \\ &= \dot{x} + (x_1 \cdot \omega \times u_1 + x_2 \cdot \omega \times u_2 + x_3 \cdot \omega \times u_3) \\ &= \dot{x} + w \times (x_1 u_1 + x_2 u_2 + x_3 u_3) \\ &= \dot{x} + \omega \times x \end{aligned}$$

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

In what regards now the acceleration, the result, which is famous, is as follows:

Theorem 8.4. *Assuming as before that a $3D$ body rotates along an axis, the acceleration of a moving point on the body, computed by O as before, is given by*

$$A = a + 2\omega \times v + \omega \times (\omega \times x)$$

with $\omega = wn$ being as before. In this formula the second term is called Coriolis acceleration, and the third term is called centripetal acceleration.

Proof. This comes by using twice the formulae in Theorem 8.3, as follows:

$$\begin{aligned} A &= \dot{V} + \omega \times V \\ &= (\ddot{x} + \dot{\omega} \times x + \omega \times \dot{x}) + (\omega \times \dot{x} + \omega \times (\omega \times x)) \\ &= \ddot{x} + \omega \times \dot{x} + \omega \times \dot{x} + \omega \times (\omega \times x) \\ &= a + 2\omega \times v + \omega \times (\omega \times x) \end{aligned}$$

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

The truly famous result is actually the one regarding forces, obtained by multiplying everything by a mass m, and writing things the other way around, as follows:

$$ma = mA - 2m\omega \times v - m\omega \times (\omega \times x)$$

Here the second term is called Coriolis force, and the third term is called centrifugal force. These forces are both called apparent, or fictious, because they do not exist in the inertial frame, but they exist however in the non-inertial frame of reference, as explained above. And with of course the terms centrifugal and centripetal not to be messed up.

In fact, even more famous is the terrestrial application of all this, as follows:

THEOREM 8.5. *The acceleration of an object m subject to a force F is given by*

$$ma = F - mg - 2m\omega \times v - m\omega \times (\omega \times x)$$

with g pointing upwards, and with the last terms being the Coriolis and centrifugal forces.

PROOF. This follows indeed from the above discussion, by assuming that the acceleration A there comes from the combined effect of a force F, and of the usual g. □

We refer to any standard undergraduate mechanics book, such as Feynman [33], Kibble [57] or Taylor [91] for more on the above, including various numerics on what happens here on Earth, the Foucault pendulum, history of all this, and many other things. Let us just mention here, as a basic illustration for all this, that a rock dropped from 100m deviates about 1cm from its intended target, due to the formula in Theorem 8.4.

8b.

8c.

8d.

8e. Exercises

Exercises:

EXERCISE 8.6.

EXERCISE 8.7.

EXERCISE 8.8.

EXERCISE 8.9.

EXERCISE 8.10.

EXERCISE 8.11.

EXERCISE 8.12.

EXERCISE 8.13.

Bonus exercise.

Part III

Arithmetic

When it’s summer in Siam
And the moon is full of rainbows
When it’s summer in Siam
And we go through many changes

CHAPTER 9

Divisibility

9a. Divisibility

Time now to get into prime numbers, which will be a main theme of discussion, for this Part II. How many primes do you know? The more the better, and those under 100 are mandatory, at the beginner level, here they are, in all their beauty:

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 97$$

We have already met prime numbers in the above, and even used some of their basic properties, that you were certainly very familiar with, but time now to review all this, on a more systematic basis. First, as definition for them, we have:

Definition 9.1. *The prime numbers are the integers $p > 1$ satisfying*

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

with each of these properties uniquely determining them.

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

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

Theorem 9.2. *Any integer $n > 1$ decomposes uniquely as*

$$n = p_1^{a_1} \ldots p_k^{a_k}$$

with $p_1 < \ldots < p_k$ primes, and with exponents $a_1, \ldots, a_k \geq 1$.

Proof. This is something that you certainly know, related to the equivalent conditions (1,2,3) in Definition 9.1, and exercise for you, to remember how all this works. Exercise as well, work out this for all integers $n \leq 100$, with no calculators allowed. □

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

Theorem 9.3. *There is an infinity of prime numbers.*

Proof. Indeed, assuming that we have finitely many prime numbers are $p_1,\ldots,p_k$, we can set $n=p_1\ldots p_k+1$, and this number n cannot factorize, contradiction. □

In practice, we can obtain the prime numbers as follows:

Theorem 9.4. *The set of prime numbers P can be obtained as follows:*

(1) *Start with* $2,3,4,5,6,7,8,9,10,11,12,\ldots$
(2) *Mark the first number,* 2*, as prime, and remove its multiples.*
(3) *Mark the new first number,* 3*, as prime, and remove its multiples.*
(4) *Mark the new first number,* 5*, as prime, and remove its multiples.*
(5) *And so on, with at each step a new prime number found.*

Proof. This algorithm for finding the primes, which is very old, and called "sieve method", is something obvious, with the first steps being as follows:

$$\begin{array}{ccccccccccccccccccc}
\underline{2} & 3 & \not4 & 5 & \not6 & 7 & \not8 & 9 & \not{1}0 & 11 & \not{1}2 & 13 & \not{1}4 & 15 & \not{1}6 & 17 & \not{1}8 & 19 & \not{2}0\\
 & \underline{3} & & 5 & & 7 & & \not9 & & 11 & & 13 & & \not{1}5 & & 17 & & 19 & \\
 & & & \underline{5} & & 7 & & & & 11 & & 13 & & & & 17 & & 19 & \\
 & & & & & \underline{7} & & & & 11 & & 13 & & & & 17 & & 19 & \\
 & & & & & & & & & \underline{11} & & 13 & & & & 17 & & 19 & \\
 & & & & & & & & & & & \underline{13} & & & & 17 & & 19 & \\
 & & & & & & & & & & & \vdots & & & & & & &
\end{array}$$

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

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

Definition 9.5. *We say that $a,b\in\mathbb Z$ are congruent modulo $c\in\mathbb Z$, and write $a=b(c)$, when c divides $b-a$.*

A first interesting question concerns solving $a=0(n)$, with n fixed and small. By writing $n=p_1^{s_1}\ldots p_k^{s_k}$, the problem reduces to solving $a=0(q)$, with $q=p^s$ small prime power. And as you surely know, there are many tricks here, summarized as follows:

Proposition 9.6. *Given a positive integer $a=a_1\ldots a_r$, we have:*

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

Proof. Here the $q=2^k,5$ assertions follow from $10=2\times5$, the $q=3,9$ assertions follow from $10=9+1$, and the $q=11$ assertion follows from $10=11-1$. □

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

Proposition 9.7. *Assuming that we have convinced mankind to change the numeration basis from* 10 *to* 14*, given a positive integer $a=a_1\ldots a_r$, we have:*

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

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

In short, we have solved the $q=7,13$ problems, but as a caveat, we have now $q=11$ not working. And is this worth it or not, up to you to decide, and launch an online petition if enthusiastic about it. Be said in passing, our Proposition 9.7 is a bit ill-formulated, mixing things written in basis 10 and basis 14, and we will leave fixing all this, with a fully correct mathematical statement, as another instructive exercise for you.

Moving ahead, congruences in general, but at a more advanced level, the mother of all results here is the following key theorem of Fermat:

Theorem 9.8. *We have the following congruence, for any prime p,*

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

called Fermat's little theorem.

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

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

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

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

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

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

9b.

9c.

9d.

9e. Exercises

Exercises:

Exercise 9.9.

Exercise 9.10.

Exercise 9.11.

Exercise 9.12.

Exercise 9.13.

Exercise 9.14.

Exercise 9.15.

Exercise 9.16.

Bonus exercise.

CHAPTER 10

Prime numbers

10a. Prime numbers

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

Theorem 10.1. *We have the following formula, implying $|P|=\infty$:*

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

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

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

valid for any integer $N\geq 2$.

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

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

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

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

In what concerns the sum on the left, this is well-known to be ∞. In what concerns now the product on the right, this can be estimated by using log, as follows:

$$\begin{aligned}
\log\left[\prod_{p\in P}\left(1-\frac{1}{p}\right)^{-1}\right] &= -\sum_{p\in P}\log\left(1-\frac{1}{p}\right)\\
&= \sum_{p\in P}\frac{1}{p}+\frac{1}{2p^2}+\frac{1}{3p^3}+\frac{1}{4p^4}+\ldots\\
&< \sum_{p\in P}\frac{1}{p}+\frac{1}{2p^2}+\frac{1}{2p^3}+\frac{1}{2p^4}+\ldots\\
&= \sum_{p\in P}\frac{1}{p}+\frac{1}{2}\sum_{p\in P}\frac{1}{p^2}\cdot\frac{1}{1-1/p}\\
&= \sum_{p\in P}\frac{1}{p}+\frac{1}{2}\sum_{p\in P}\frac{1}{p(p-1)}\\
&< \sum_{p\in P}\frac{1}{p}+\frac{1}{2}\sum_{n=2}^{\infty}\frac{1}{n(n-1)}\\
&= \sum_{p\in P}\frac{1}{p}+\frac{1}{2}
\end{aligned}$$

We therefore obtain the following estimate, which gives the first assertion:

$$\sum_{p\in P}\frac{1}{p}+\frac{1}{2}>\log\left(\sum_{n=1}^{\infty}\frac{1}{n}\right)=\infty$$

Regarding now the second assertion, the idea is to replace in the above computations the set P of all primes by the set of all primes $p<N$. We obtain in this way the following estimate, and with exercise for you, to work out the details:

$$\begin{aligned}
\sum_{p<N}\frac{1}{p}+\frac{1}{2} &> \log\left(\sum_{n=1}^{N}\frac{1}{n}\right)\\
&> \log\left(\int_1^N\frac{1}{x}\,dx\right)\\
&= \log\log N
\end{aligned}$$

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

The Euler formula and its proof are something of utter beauty, suggesting doing an enormous amount of things, and yes indeed, doing such things has been one of the favorite pastimes of mathematicians, since. Here is a brief account, of all this:

(1) The Euler formula $\sum_{p\in P}1/p=\infty$ basically tells us that there are "many primes", but what about the opposite, trying now to prove that there are "few primes"? Well, this comes too from the Euler formula, but in its refined version, with $\log\log N$:

$$\sum_{p<N}\frac{1}{p}\simeq\log\log N$$

Many things can be done here, one of the conclusions being that the N-th prime $\pi(N)$ satisfies $\pi(N)\sim N/\log N$. We will be back to this later in this book.

(2) Still talking analysis, an interesting observation, by Erdős, coming from his own proof of the Euler formula, regards the sets $S\subset\mathbb{N}$ satisfying the following condition:

$$\sum_{s\in S}\frac{1}{s}=\infty$$

Based on this, Erdős conjectured that such sets S contain arbitrarily long arithmetic progessions. And the point is that this is a very difficult and fascinating problem, with the case $S=P$ being settled only recently, by Green and Tao.

(3) Leaving aside now estimates and analysis, and going back to the beginning of Euler's proof, let us look more in detail at the formula there, namely:

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

This formula is something really beautiful, and the more you look at it, thinking at versions and so on, the more you are lost into the mysteries of number theory.

(4) To be more precise, the above formula suggests introducing the following function, depending on a parameter s, which can be integer, real, or even complex:

$$\zeta(s)=\sum_{n=1}^\infty\frac{1}{n^s}$$

And this is the famous Riemann zeta function, which obsesses all number theorists, be them algebraists, analysts, geometers, physicists, or amateurs. We will be talking about this magical function later in this book, in Part IV, after learning some analysis.

10b.

10c.

10d.

10e. Exercises

Exercises:

Exercise 10.2.

Exercise 10.3.

Exercise 10.4.

Exercise 10.5.

Exercise 10.6.

Exercise 10.7.

Exercise 10.8.

Exercise 10.9.

Bonus exercise.

CHAPTER 11

Squares, residues

11a. Squares, residues

Let us go back to what we did in chapter 9 with congruences. Our aim here will be that of further building on some of the theorems there. To be more precise, we will be interested in solving the following ubiquitous equation, over the integers:

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

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

Definition 11.1. *The Legendre symbol is defined as follows,*

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

with $p\geq 3$ prime.

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

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

Theorem 11.2. *The Legendre symbol is given by the formula*

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

called Euler formula for the Legendre symbol.

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

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

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

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

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

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

$$a^{\frac{p-1}{2}} = \begin{cases} 1 & \text{if } \exists\, b, a = b^2 \\ -1 & \text{if } \nexists\, b, a = b^2 \end{cases}$$

By comparing now with Definition 3.1, we obtain the formula in the statement. □

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

Proposition 11.3. *We have the following formula, valid for any $a, b \in \mathbb{Z}$:*

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

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

Proof. This is clear indeed from the Euler formula, because $a^{\frac{p-1}{2}}(p)$ is obviously multiplicative in $a \in \mathbb{Z}$. Alternatively, this can be proved as well directly, with no need for the Fermat formula used in the proof of Euler, just by thinking at what is quadratic residue and what is not in $\mathbb{F}_p$, along the lines of (2) in the proof of Theorem 11.2. □

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

Proposition 11.4. *We have the following formula, telling us that modulo any prime number p, a product of non-squares is a square:*

$$\left(\frac{a}{p}\right) = -1\ ,\ \left(\frac{b}{p}\right) = -1 \implies \left(\frac{ab}{p}\right) = 1$$

Also, the Legendre symbol, regarded as a function

$$\chi : \mathbb{F}_p - \{0\} \to \{-1, 1\} \quad , \quad \chi(a) = \left(\frac{a}{p}\right)$$

is a character, in the sense that it is multiplicative.

Proof. The first asssertion is a consequence of Proposition 11.3, more or less equivalent to it, and with the remark that this formally holds at $p = 2$ too, as $\emptyset \implies \emptyset$. As for the second assertion, this is just a fancy reformulation of Proposition 11.3. □

It is possible to say some further conceptual things, some sounding very fancy, in relation with Proposition 11.3 and Proposition 11.4. But remember that, according to the plan made in the beginning of this chapter, we are here for the kill, namely computing the Legendre symbol, no matter what, and with no prisoners taken.

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

THEOREM 11.5. *We have the quadratic reciprocity formula*

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

valid for any primes $p,q\geq 3$.

PROOF. This is obviously something tough, because how on Earth, you would say, the above two Legendre symbols can be related to each other. Good point, and in answer, I do not have any simple explanation to offer for this, at this point of writing. We will see however a proof for this, later in this chapter, by using some calculus with the roots of unity, and more specifically, with certain beasts called quadratic Gauss sums. □

As a comment now, the above result is extremely powerful, here being an illustration, computing the seemingly uncomputable number on the left in a matter of seconds:

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

In fact, when combining Theorem 11.5 with Proposition 11.3, it is quite clear that, no matter how big p is, if a has only small prime factors, we are saved.

Besides Proposition 11.3, the quadratic reciprocity formula comes accompanied by two other statements, which are very useful in practice. First, at $a=-1$, we have:

PROPOSITION 11.6. *We have the following formula,*

$$\left(\frac{-1}{p}\right)=\begin{cases}1&\text{if } p=1(4)\\-1&\text{if } p=3(4)\end{cases}$$

solving in practice the equation $b^2=-1(p)$.

PROOF. This follows from the Euler formula, which at $a=-1$ reads:

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

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

As a second useful result, this time at $a = 2$, we have:

Theorem 11.7. *We have the following formula,*

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

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

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

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

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

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

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

As a continuation of this, speaking Legendre symbol for small values of the upper variable, we can try to compute these for $a = \pm\, 3, 4, 5, 6, 7, 8, \ldots$ But by multiplicativity plus Proposition 11.6 plus Theorem 11.7 we are left with the case where $a = q$ is an odd prime, and we can solve the problem with quadratic reciprocity, so done.

Let us record however a few statements here, which can be useful in practice, and with this being mostly for illustration purposes, for Theorem 11.5. We first have:

Proposition 11.8. *We have the following formula,*

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

valid for any prime $p \geq 5$.

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

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

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

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

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

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

PROPOSITION 11.9. *We have the following formula,*

$$\left(\frac{5}{p}\right)=\begin{cases}1 & \text{if } p=1,4(5)\\ -1 & \text{if } p=2,3(5)\end{cases}$$

valid for any odd prime $p\neq 5$.

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

$$\left(\frac{5}{p}\right)=(-1)^{\frac{5-1}{2}\cdot\frac{p-1}{2}}\left(\frac{p}{5}\right)=\left(\frac{p}{5}\right)$$

Thus, we have the result. Alternatively, we have the following formula:

$$\left(\frac{5}{p}\right)=(-1)^{\left[\frac{2p+2}{5}\right]}$$

Indeed, this is the simplest implementation of the formula in the statement. □

Moving ahead now, we have the following interesting generalization of the Legendre symbol, to the case of denominators not necessarily prime, due to Jacobi:

THEOREM 11.10. *The theory of Legendre symbols can be extended by multiplicativity into a theory of Jacobi symbols, according to the formula*

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

with the denominator being not necessarily prime, but just an arbitrary odd number, and this theory has as results those imported from the Legendre theory.

PROOF. This is something self-explanatory, and we will leave listing the basic properties of the Jacobi symbols, based on the theory of Legendre symbols, as an exercise. □

The story is not over with Jacobi, because the denominator there is still odd, and positive. So, we have a problem to be solved, the solution to it being as follows:

THEOREM 11.11. *The theory of Jacobi symbols can be further extended into a theory of Kronecker symbols, according to the formula*

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

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

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

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

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

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

The answer for the second symbol is as follows:

$$\left(\frac{a}{-1}\right)=\begin{cases}1&\text{if } a\geq 0\\-1&\text{if } a<0\end{cases}$$

As for the answer for the third symbol, this is as follows:

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

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

11b.

11c.

11d.

11e. Exercises

Exercises:

Exercise 11.12.

Exercise 11.13.

Exercise 11.14.

Exercise 11.15.

Exercise 11.16.

Exercise 11.17.

Exercise 11.18.

Exercise 11.19.

Bonus exercise.

CHAPTER 12

Polynomials, roots

12a. Polynomials, roots

We have seen that many number theory questions lead us into computing roots of polynomials $P \in \mathbb{Q}[X]$. We will investigate here such questions, with a detailed study of the arbitrary polynomials $P \in \mathbb{C}[X]$, and their roots, often by using analytic methods.

Let us start with something that we know well, but is always good to remember:

THEOREM 12.1. *The solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{C}$ are*

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

with the square root of complex numbers being defined as $\sqrt{re^{it}} = \sqrt{r}e^{it/2}$.

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

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

Here we have used the fact, mentioned in the statement, that any complex number $z = re^{it}$ has indeed a square root, given by $\sqrt{z} = \sqrt{r}e^{it/2}$, plus in fact a second square root as well, namely $-\sqrt{z}$. Thus, we are led to the conclusion in the statement. □

Very nice all this, and you would probably say that the story is over here, with degree 2. However, not really. Here are a few tricks, in order to deal with degree 2 questions:

TRICKS 12.2. *The following happen:*

1. *The roots of $x^2 - ax + b$ can be computed by using $r + s = a$, $rs = b$.*
2. *The eigenvalues of $A \in M_2(\mathbb{C})$ are given by $r + s = Tr(A)$, $rs = \det A$.*

To be more precise, (1) is clear, and the equations there are usually the fastest way for computing, via instant thinking, the roots r,s, provided of course that these roots are simple numbers, say integers. As for (2), consider indeed a 2×2 matrix:

$$A=\begin{pmatrix}m&n\\ p&q\end{pmatrix}$$

In order to find the eigenvalues r,s, you are certainly very used to compute the characteristic polynomial, then apply Theorem 12.1. But my point is that this characteristic polynomial is of the form x^2-ax+b, with $a=Tr(a)$ and $b=\det A$, so we can normally apply the trick in (1), provided of course that r,s are simple numbers, say integers.

Finally, for this discussion to be complete, let us mention too:

Warning 12.3. *The above tricks work in pure mathematics, where the numbers r,s that we can meet are usually integers, or rationals. In applied mathematics, however, the numbers that we meet are integers or rationals with probability $P=0$, so no tricks.*

I am saying this of course in view of the fact that in applied mathematics the numbers that can appear, say via reading certain scientific instruments, are quite "random", and to be more precise, oscillating in a random way around an average value. Thus, we are dealing here with the continuum, and the probability of being rational is $P=0$.

Moving now to degree 3 and higher, things here are far more complicated, and as a first objective, we would like to understand what the analogue of the discriminant $\Delta=b^2-4ac$ is. But even this is something quite tricky, because we would like to have $\Delta=0$ precisely when $(P,P')\neq 1$, which leads us into the question of deciding, given two polynomials $P,Q\in\mathbb C[X]$, if these polynomials have a common root, $(P,Q)\neq 1$, or not.

Fortunately this latter question has a nice answer. We will need:

Theorem 12.4. *Given a monic polynomial $P\in\mathbb C[X]$, factorized as*

$$P=(X-a_1)\ldots(X-a_k)$$

the following happen:

1. *The coefficients of P are symmetric functions in $a_1,\ldots,a_k$.*
2. *The symmetric functions in $a_1,\ldots,a_k$ are polynomials in the coefficients of P.*

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

(1) By expanding our polynomial, we have the following formula:

$$P=\sum_{r=0}^k(-1)^r\sum_{i_1<\ldots<i_r}a_{i_1}\ldots a_{i_r}\cdot X^{k-r}$$

Thus the coefficients of P are, up to some signs, the following functions:

$$f_r = \sum_{i_1<\ldots<i_r} a_{i_1}\ldots a_{i_r}$$

But these are indeed symmetric functions in $a_1,\ldots,a_k$, as claimed.

(2) Conversely now, let us look at the symmetric functions in the roots $a_1,\ldots,a_k$. These appear as linear combinations of the basic symmetric functions, given by:

$$S_r = \sum_i a_i^r$$

Moreover, when allowing polynomials instead of linear combinations, we need in fact only the first k such sums, namely $S_1,\ldots,S_k$. That is, the symmetric functions $\mathcal{F}$ in our variables $a_1,\ldots,a_k$, with integer coefficients, appear as follows:

$$\mathcal{F} = \mathbb{Z}[S_1,\ldots,S_k]$$

(3) The point now is that, alternatively, the symmetric functions in our variables $a_1,\ldots,a_k$ appear as well as linear combinations of the functions f_r that we found in (1), and that when allowing polynomials instead of linear combinations, we need in fact only the first k functions, namely $f_1,\ldots,f_k$. That is, we have as well:

$$\mathcal{F} = \mathbb{Z}[f_1,\ldots,f_k]$$

But this gives the result, because we can pass from $\{S_r\}$ to $\{f_r\}$, and vice versa.

(4) This was for the idea, and in practice now up to you to clarify all the details. In fact, we will also need in what follows the extension of all this to the case where P is no longer assumed to be monic, and with this being, again, exercise for you. □

Getting back now to our original question, namely that of deciding whether two polynomials $P,Q\in\mathbb{C}[X]$ have a common root or not, this has the following nice answer:

Theorem 12.5. *Given two polynomials $P,Q\in\mathbb{C}[X]$, written as*

$$P = c(X-a_1)\ldots(X-a_k) \quad , \quad Q = d(X-b_1)\ldots(X-b_l)$$

the following quantity, which is called resultant of P,Q,

$$R(P,Q) = c^l d^k \prod_{ij}(a_i - b_j)$$

is a certain polynomial in the coefficients of P,Q, with integer coefficients, and we have $R(P,Q)=0$ precisely when P,Q have a common root.

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

(1) Given two polynomials $P,Q\in\mathbb{C}[X]$, we can certainly construct the quantity $R(P,Q)$ in the statement, with the role of the normalization factor c^ld^k to become clear later on, and then we have $R(P,Q)=0$ precisely when P,Q have a common root:

$$R(P,Q)=0\iff\exists i,j,a_i=b_j$$

(2) As bad news, however, this quantity $R(P,Q)$, defined in this way, is a priori not very useful in practice, because it depends on the roots a_i,b_j of our polynomials P,Q, that we cannot compute in general. However, and here comes our point, as we will prove below, it turns out that $R(P,Q)$ is in fact a polynomial in the coefficients of P,Q, with integer coefficients, and this is where the power of $R(P,Q)$ comes from.

(3) You might perhaps say, nice, but why not doing things the other way around, that is, formulating our theorem with the explicit formula of $R(P,Q)$, in terms of the coefficients of P,Q, and then proving that we have $R(P,Q)=0$, via roots and everything. Good point, but this is not exactly obvious, the formula of $R(P,Q)$ in terms of the coefficients of P,Q being something terribly complicated. In short, trust me, let us prove our theorem as stated, and for alternative formulae of $R(P,Q)$, we will see later.

(4) Getting started now, let us expand the formula of $R(P,Q)$, by making all the multiplications there, abstractly, in our head. Everything being symmetric in $a_1,\ldots,a_k$, we obtain in this way certain symmetric functions in these variables, which will be therefore certain polynomials in the coefficients of P. Moreover, due to our normalization factor c^l, these polynomials in the coefficients of P will have integer coefficients.

(5) With this done, let us look now what happens with respect to the remaining variables $b_1,\ldots,b_l$, which are the roots of Q. Once again what we have here are certain symmetric functions in these variables $b_1,\ldots,b_l$, and these symmetric functions must be certain polynomials in the coefficients of Q. Moreover, due to our normalization factor d^k, these polynomials in the coefficients of Q will have integer coefficients.

(6) Thus, we are led to the conclusion in the statement, that $R(P,Q)$ is a polynomial in the coefficients of P,Q, with integer coefficients, and with the remark that the c^ld^k factor is there for these latter coefficients to be indeed integers, instead of rationals. □

All the above might seem a bit complicated, so as an illustration, let us work out an example. Consider the case of a polynomial of degree 2, and a polynomial of degree 1:

$$P=ax^2+bx+c\quad,\quad Q=dx+e$$

In order to compute the resultant, let us factorize our polynomials:

$$P=a(x-p)(x-q)\quad,\quad Q=d(x-r)$$

The resultant can be then computed as follows, by using the method above:

$$\begin{aligned}
R(P,Q) &= ad^2(p-r)(q-r)\\
&= ad^2(pq-(p+q)r+r^2)\\
&= cd^2+bd^2r+ad^2r^2\\
&= cd^2-bde+ae^2
\end{aligned}$$

Finally, observe that $R(P,Q)=0$ corresponds indeed to the fact that P,Q have a common root. Indeed, the root of Q is $r=-e/d$, and we have:

$$P(r)=\frac{ae^2}{d^2}-\frac{be}{d}+c=\frac{R(P,Q)}{d^2}$$

Regarding now the explicit formula of the resultant $R(P,Q)$, this is something quite complicated, and there are several methods for dealing with this problem. We have:

Theorem 12.6. *The resultant of two polynomials, written as*

$$P=p_kX^k+\ldots+p_1X+p_0\quad,\quad Q=q_lX^l+\ldots+q_1X+q_0$$

appears as the determinant of an associated matrix, as follows,

$$R(P,Q)=\begin{vmatrix}
p_k & & & q_l & & \\
\vdots & \ddots & & \vdots & \ddots & \\
p_0 & & p_k & q_0 & & q_l\\
 & \ddots & \vdots & & \ddots & \vdots\\
 & & p_0 & & & q_0
\end{vmatrix}$$

with the matrix having size $k+l$, and having 0 coefficients at the blank spaces.

Proof. This is something clever, due to Sylvester, as follows:

(1) Consider the vector space $\mathbb{C}_k[X]$ formed by the polynomials of degree $<k$:

$$\mathbb{C}_k[X]=\left\{P\in\mathbb{C}[X]\Big|\deg P<k\right\}$$

This is a vector space of dimension k, having as basis the monomials $1,X,\ldots,X^{k-1}$. Now given polynomials P,Q as in the statement, consider the following linear map:

$$\Phi:\mathbb{C}_l[X]\times\mathbb{C}_k[X]\to\mathbb{C}_{k+l}[X]\quad,\quad(A,B)\to AP+BQ$$

(2) Our first claim is that with respect to the standard bases for all the vector spaces involved, namely those consisting of the monomials $1,X,X^2,\ldots$, the matrix of Φ is the matrix in the statement. But this is something which is clear from definitions.

(3) Our second claim is that $\det\Phi=0$ happens precisely when P,Q have a common root. Indeed, our polynomials P,Q having a common root means that we can find A,B such that $AP+BQ=0$, and so that $(A,B)\in\ker\Phi$, which reads $\det\Phi=0$.

(4) Finally, our claim is that we have $\det\Phi = R(P,Q)$. But this follows from the uniqueness of the resultant, up to a scalar, and with this uniqueness property being elementary to establish, along the lines of the proofs of Theorems 12.4 and 12.5. $\square$

In what follows we will not really need the above formula, so let us just check now that this formula works indeed. Consider our favorite polynomials, as before:

$$P = ax^2 + bx + c \quad , \quad Q = dx + e$$

According to the above result, the resultant should be then, as it should:

$$R(P,Q) = \begin{vmatrix} a & d & 0 \\ b & e & d \\ c & 0 & e \end{vmatrix} = ae^2 - bde + cd^2$$

We can go back now to our original question, and we have:

Theorem 12.7. *Given a polynomial $P \in \mathbb{C}[X]$, written as*

$$P(X) = aX^N + bX^{N-1} + cX^{N-2} + \ldots$$

its discriminant, defined as being the following quantity,

$$\Delta(P) = \frac{(-1)^{\binom{N}{2}}}{a} R(P,P')$$

is a polynomial in the coefficients of P, with integer coefficients, and $\Delta(P) = 0$ happens precisely when P has a double root.

Proof. The fact that the discriminant $\Delta(P)$ is a polynomial in the coefficients of P, with integer coefficients, comes from Theorem 12.5, coupled with the fact that the division by the leading coefficient a is indeed possible, under $\mathbb{Z}$, as being shown by the following formula, which is of course a bit informal, coming from Theorem 12.6:

$$R(P,P') = \begin{vmatrix} a & & & Na & & \\ \vdots & \ddots & & \vdots & \ddots & \\ z & & a & y & & Na \\ & \ddots & \vdots & & \ddots & \vdots \\ & & z & & & y \end{vmatrix}$$

Also, the fact that we have $\Delta(P) = 0$ precisely when P has a double root is clear from Theorem 12.5. Finally, let us mention that the sign $(-1)^{\binom{N}{2}}$ is there for various reasons, including the compatibility with some well-known formulae, at small values of $N \in \mathbb{N}$, such as $\Delta(P) = b^2 - 4ac$ in degree 2, that we will discuss in a moment. $\square$

As already mentioned, by using Theorem 12.6, we have an explicit formula for the discriminant, as the determinant of a certain matrix. There is a lot of theory here, and in order to get into this, let us first see what happens in degree 2. Here we have:

$$P=aX^2+bX+c\quad,\quad P'=2aX+b$$

Thus, the resultant is given by the following formula:

$$\begin{aligned}R(P,P')&=ab^2-b(2a)b+c(2a)^2\\&=4a^2c-ab^2\\&=-a(b^2-4ac)\end{aligned}$$

It follows that the discriminant of our polynomial is, as it should:

$$\Delta(P)=b^2-4ac$$

Alternatively, we can use the formula in Theorem 12.6, and we obtain:

$$\begin{aligned}\Delta(P)=&=-\frac{1}{a}\begin{vmatrix}a&2a&\\b&b&2a\\c&&b\end{vmatrix}\\&=-\begin{vmatrix}1&2&\\b&b&2a\\c&&b\end{vmatrix}\\&=-b^2+2(b^2-2ac)\\&=b^2-4ac\end{aligned}$$

We will be back later to such formulae, in degree 3, and in degree 4 as well, with the comment however, coming in advance, that these formulae are not very beautiful.

At the theoretical level now, we have the following result, which is not trivial:

THEOREM 12.8. *The discriminant of a polynomial P is given by the formula*

$$\Delta(P)=a^{2N-2}\prod_{i<j}(r_i-r_j)^2$$

where a is the leading coefficient, and $r_1,\ldots,r_N$ are the roots.

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

(1) The first thought goes to the formula in Theorem 12.5, so let us see what that formula teaches us, in the case $Q=P'$. Let us write P,P' as follows:

$$P=a(x-r_1)\ldots(x-r_N)$$

$$P'=Na(x-p_1)\ldots(x-p_{N-1})$$

According to Theorem 12.5, the resultant of P, P' is then given by:

$$R(P, P') = a^{N-1}(Na)^N \prod_{ij}(r_i - p_j)$$

And bad news, this is not exactly what we wished for, namely the formula in the statement. That is, we are on the good way, but certainly have to work some more.

(2) Obviously, we must get rid of the roots $p_1, \ldots, p_{N-1}$ of the polynomial P'. In order to do this, let us rewrite the formula that we found in (1) in the following way:

$$\begin{aligned} R(P, P') &= N^N a^{2N-1} \prod_i \left(\prod_j (r_i - p_j) \right) \\ &= N^N a^{2N-1} \prod_i \frac{P'(r_i)}{Na} \\ &= a^{N-1} \prod_i P'(r_i) \end{aligned}$$

(3) In order to compute now P', and more specifically the values $P'(r_i)$ that we are interested in, we can use the Leibnitz rule. So, consider our polynomial:

$$P(x) = a(x - r_1) \ldots (x - r_N)$$

The Leibnitz rule for derivatives tells us that $(fg)' = f'g + fg'$, but then also that $(fgh)' = f'gh + fg'h + fgh'$, and so on. Thus, for our polynomial, we obtain:

$$P'(x) = a \sum_i (x - r_1) \ldots \underbrace{(x - r_i)}_{missing} \ldots (x - r_N)$$

Now when applying this formula to one of the roots r_i, we obtain:

$$P'(r_i) = a(r_i - r_1) \ldots \underbrace{(r_i - r_i)}_{missing} \ldots (r_i - r_N)$$

By making now the product over all indices i, this gives the following formula:

$$\prod_i P'(r_i) = a^N \prod_{i \neq j}(r_i - r_j)$$

(4) Time now to put everything together. By taking the formula in (2), making the normalizations in Theorem 12.7, and then using the formula found in (3), we obtain:

$$\begin{aligned} \Delta(P) &= (-1)^{\binom{N}{2}} a^{N-2} \prod_i P'(r_i) \\ &= (-1)^{\binom{N}{2}} a^{2N-2} \prod_{i \neq j}(r_i - r_j) \end{aligned}$$

(5) This is already a nice formula, which is very useful in practice, and that we can safely keep as a conclusion, to our computations. However, we can do slightly better, by grouping opposite terms. Indeed, this gives the following formula:

$$\begin{aligned}
\Delta(P) &= (-1)^{\binom{N}{2}}a^{2N-2}\prod_{i\neq j}(r_i-r_j)\\
&= (-1)^{\binom{N}{2}}a^{2N-2}\prod_{i<j}(r_i-r_j)\cdot\prod_{i>j}(r_i-r_j)\\
&= (-1)^{\binom{N}{2}}a^{2N-2}\prod_{i<j}(r_i-r_j)\cdot(-1)^{\binom{N}{2}}\prod_{i<j}(r_i-r_j)\\
&= a^{2N-2}\prod_{i<j}(r_i-r_j)^2
\end{aligned}$$

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

As applications now, the formula in Theorem 12.8 is quite useful for the real polynomials $P\in\mathbb{R}[X]$ in small degree, because it allows to say when the roots are real, or complex, or at least have some partial information about this. For instance, we have:

PROPOSITION 12.9. *Consider a polynomial with real coefficients, $P\in\mathbb{R}[X]$, assumed for simplicity to have nonzero discriminant, $\Delta\neq 0$.*

(1) *In degree 2, the roots are real when $\Delta>0$, and complex when $\Delta<0$.*
(2) *In degree 3, all roots are real precisely when $\Delta>0$.*

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

(1) The first assertion is something that you certainly know, coming from Theorem 12.1, but let us see how this comes via the formula in Theorem 12.8, namely:

$$\Delta(P)=a^{2N-2}\prod_{i<j}(r_i-r_j)^2$$

In degree $N=2$, this formula looks as follows, with r_1,r_2 being the roots:

$$\Delta(P)=a^2(r_1-r_2)^2$$

Thus $\Delta>0$ amounts in saying that we have $(r_1-r_2)^2>0$. Now since r_1,r_2 are conjugate, and with this being something trivial, meaning no need here for the computations in Theorem 12.1, we conclude that $\Delta>0$ means that r_1,r_2 are real, as stated.

(2) In degree $N=3$ now, we know from analysis that P has at least one real root, and the problem is whether the remaining 2 roots are real, or complex conjugate. For this purpose, we can use the formula in Theorem 12.8, which in degree 3 reads:

$$\Delta(P)=a^4(r_1-r_2)^2(r_1-r_3)^2(r_2-r_3)^2$$

We can see that in the case $r_1, r_2, r_3 \in \mathbb{R}$, we have $\Delta(P) > 0$. Conversely now, assume that $r_1 = r$ is the real root, coming from analysis, and that the other roots are $r_2 = z$ and $r_3 = \bar{z}$, with z being a complex number, which is not real. We have then:

$$\begin{aligned}
\Delta(P) &= a^4(r-z)^2(r-\bar{z})^2(z-\bar{z})^2\\
&= a^4|r-z|^4(2iIm(z))^2\\
&= -4a^4|r-z|^4Im(z)^2\\
&< 0
\end{aligned}$$

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

In relation with the above, for our result to be truly useful, we must of course compute the discriminant in degree 3. We will do this, along with applications, right next.

12b.

12c.

12d.

12e. Exercises

Exercises:

Exercise 12.10.

Exercise 12.11.

Exercise 12.12.

Exercise 12.13.

Exercise 12.14.

Exercise 12.15.

Exercise 12.16.

Exercise 12.17.

Bonus exercise.

Part IV

Functions

Dancing like there's no one there
Before she ever seemed to care
Now she wouldn't dare
It's so rock and roll to be alone

CHAPTER 13

Functions

13a. Functions

13b.

13c.

13d.

13e. Exercises

Exercises:

EXERCISE 13.1.

EXERCISE 13.2.

EXERCISE 13.3.

EXERCISE 13.4.

EXERCISE 13.5.

EXERCISE 13.6.

EXERCISE 13.7.

EXERCISE 13.8.

Bonus exercise.

CHAPTER 14

Powers, logarithms

14a. Powers, logarithms

14b.

14c.

14d.

14e. Exercises

Exercises:

Exercise 14.1.

Exercise 14.2.

Exercise 14.3.

Exercise 14.4.

Exercise 14.5.

Exercise 14.6.

Exercise 14.7.

Exercise 14.8.

Bonus exercise.

CHAPTER 15

More trigonometry

15a. More trigonometry

15b.

15c.

15d.

15e. Exercises

Exercises:

EXERCISE 15.1.

EXERCISE 15.2.

EXERCISE 15.3.

EXERCISE 15.4.

EXERCISE 15.5.

EXERCISE 15.6.

EXERCISE 15.7.

EXERCISE 15.8.

Bonus exercise.

CHAPTER 16

Derivatives

16a. Derivatives

16b.

16c.

16d.

16e. Exercises

Congratulations for having read this book, and no exercises for this final chapter.

Bibliography

[1] A.A. Abrikosov, Fundamentals of the theory of metals, Dover (1988).

[2] V.I. Arnold, Ordinary differential equations, Springer (1973).

[3] V.I. Arnold, Lectures on partial differential equations, Springer (1997).

[4] V.I. Arnold, Catastrophe theory, Springer (1984).

[5] N.W. Ashcroft and N.D. Mermin, Solid state physics, Saunders College Publ. (1976).

[6] T. Banica, Calculus and applications (2024).

[7] T. Banica, Linear algebra and group theory (2024).

[8] T. Banica, Introduction to modern physics (2024).

[9] G.K. Batchelor, An introduction to fluid dynamics, Cambridge Univ. Press (1967).

[10] M.J. Benton, Vertebrate paleontology, Wiley (1990).

[11] M.J. Benton and D.A.T. Harper, Introduction to paleobiology and the fossil record, Wiley (2009).

[12] S.J. Blundell and K.M. Blundell, Concepts in thermal physics, Oxford Univ. Press (2006).

[13] B. Bollobás, Modern graph theory, Springer (1998).

[14] S.M. Carroll, Spacetime and geometry, Cambridge Univ. Press (2004).

[15] P.M. Chaikin and T.C. Lubensky, Principles of condensed matter physics, Cambridge Univ. Press (1995).

[16] A.R. Choudhuri, Astrophysics for physicists, Cambridge Univ. Press (2012).

[17] J. Clayden, S. Warren and N. Greeves, Organic chemistry, Oxford Univ. Press (2012).

[18] D.D. Clayton, Principles of stellar evolution and nucleosynthesis, Univ. of Chicago Press (1968).

[19] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).

[20] A. Cottrell, An introduction to metallurgy, CRC Press (1997).

[21] C. Darwin, On the origin of species (1859).

[22] P.A. Davidson, Introduction to magnetohydrodynamics, Cambridge Univ. Press (2001).

[23] P.A.M. Dirac, Principles of quantum mechanics, Oxford Univ. Press (1930).

[24] S. Dodelson, Modern cosmology, Academic Press (2003).

[25] S.T. Dougherty, Combinatorics and finite geometry, Springer (2020).

[26] M. Dresher, The mathematics of games of strategy, Dover (1981).

[27] R. Durrett, Probability: theory and examples, Cambridge Univ. Press (1990).

[28] F. Dyson, Origins of life, Cambridge Univ. Press (1984).

[29] A. Einstein, Relativity: the special and the general theory, Dover (1916).

[30] L.C. Evans, Partial differential equations, AMS (1998).

[31] W. Feller, An introduction to probability theory and its applications, Wiley (1950).

[32] E. Fermi, Thermodynamics, Dover (1937).

[33] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics I: mainly mechanics, radiation and heat, Caltech (1963).

[34] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics II: mainly electromagnetism and matter, Caltech (1964).

[35] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics III: quantum mechanics, Caltech (1966).

[36] R.P. Feynman and A.R. Hibbs, Quantum mechanics and path integrals, Dover (1965).

[37] P. Flajolet and R. Sedgewick, Analytic combinatorics, Cambridge Univ. Press (2009).

[38] A.P. French, Special relativity, Taylor and Francis (1968).

[39] J.H. Gillespie, Population genetics, Johns Hopkins Univ. Press (1998).

[40] C. Godsil and G. Royle, Algebraic graph theory, Springer (2001).

[41] H. Goldstein, C. Safko and J. Poole, Classical mechanics, Addison-Wesley (1980).

[42] D.L. Goodstein, States of matter, Dover (1975).

[43] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).

[44] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).

[45] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).

[46] D.J. Griffiths, Revolutions in twentieth-century physics, Cambridge Univ. Press (2012).

[47] V.P. Gupta, Principles and applications of quantum chemistry, Elsevier (2016).

[48] W.A. Harrison, Solid state theory, Dover (1970).

[49] W.A. Harrison, Electronic structure and the properties of solids, Dover (1980).

[50] R.A. Horn and C.R. Johnson, Matrix analysis, Cambridge Univ. Press (1985).

[51] C.E. Housecroft and A.G. Sharpe, Inorganic chemistry, Pearson (2018).

[52] K. Huang, Introduction to statistical physics, CRC Press (2001).

[53] K. Huang, Fundamental forces of nature, World Scientific (2007).

[54] S. Huskey, The skeleton revealed, Johns Hopkins Univ. Press (2017).

[55] L. Hyman, Comparative vertebrate anatomy, Univ. of Chicago Press (1942).

[56] L.P. Kadanoff, Statistical physics: statics, dynamics and renormalization, World Scientific (2000).

[57] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).

[58] C. Kittel, Introduction to solid state physics, Wiley (1953).

[59] D.E. Knuth, The art of computer programming, Addison-Wesley (1968).

[60] M. Kumar, Quantum: Einstein, Bohr, and the great debate about the nature of reality, Norton (2009).

[61] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).

[62] L.D. Landau and E.M. Lifshitz, Mechanics, Pergamon Press (1960).

[63] L.D. Landau and E.M. Lifshitz, The classical theory of fields, Addison-Wesley (1951).

[64] L.D. Landau and E.M. Lifshitz, Quantum mechanics: non-relativistic theory, Pergamon Press (1959).

[65] S. Lang, Algebra, Addison-Wesley (1993).

[66] P. Lax, Linear algebra and its applications, Wiley (2007).

[67] P. Lax, Functional analysis, Wiley (2002).

[68] P. Lax and M.S. Terrell, Calculus with applications, Springer (2013).

[69] P. Lax and M.S. Terrell, Multivariable calculus with applications, Springer (2018).

[70] S. Ling and C. Xing, Coding theory: a first course, Cambridge Univ. Press (2004).

[71] J.P. Lowe and K. Peterson, Quantum chemistry, Elsevier (2005).

[72] S.J. Marshall, The story of the computer: a technical and business history, Create Space Publ. (2022).

[73] M.L. Mehta, Random matrices, Elsevier (2004).

[74] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge Univ. Press (2000).

[75] R.K. Pathria and and P.D. Beale, Statistical mechanics, Elsevier (1972).

[76] T.D. Pollard, W.C. Earnshaw, J. Lippincott-Schwartz and G. Johnson, Cell biology, Elsevier (2022).

[77] J. Preskill, Quantum information and computation, Caltech (1998).

[78] R. Rojas and U. Hashagen, The first computers: history and architectures, MIT Press (2000).

[79] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[80] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[81] W. Rudin, Functional analysis, McGraw-Hill (1973).

[82] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[83] B. Ryden and B.M. Peterson, Foundations of astrophysics, Cambridge Univ. Press (2010).

[84] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).

[85] R. Shankar, Fundamentals of physics I: mechanics, relativity, and thermodynamics, Yale Univ. Press (2014).

[86] R. Shankar, Fundamentals of physics II: electromagnetism, optics, and quantum mechanics, Yale Univ. Press (2016).

[87] N.J.A. Sloane and S. Plouffe, Encyclopedia of integer sequences, Academic Press (1995).

[88] A.M. Steane, Thermodynamics, Oxford Univ. Press (2016).

[89] S. Sternberg, Dynamical systems, Dover (2010).

[90] D.R. Stinson, Combinatorial designs: constructions and analysis, Springer (2006).

[91] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[92] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[93] J. von Neumann and O. Morgenstern, Theory of games and economic behavior, Princeton Univ. Press (1944).

[94] J. Watrous, The theory of quantum information, Cambridge Univ. Press (2018).

[95] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[96] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[97] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).

[98] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).

[99] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).

[100] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).

Index