

Introduction to arithmetic

Teo Banica

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CERGY-PONTOISE, F-95000
CERGY-PONTOISE, FRANCE. teo.banica@gmail.com

2010 *Mathematics Subject Classification.* 97F60

Key words and phrases. Numbers, Primes

ABSTRACT. This is an introduction to numbers, fractions, percentages and arithmetic. We first discuss what can be done with integers and their quotients, namely basic arithmetic, all sorts of counting results, and with a look into abstract algebra and quadratic residues too. We then upgrade our knowledge by introducing the real numbers, and exploring what can be done with them, in relation with number theory questions. Then we further upgrade our methods, by introducing and using the complex numbers. Finally, we provide an introduction to the zeta function, and the Riemann hypothesis.

Preface

Number theory is the Queen of Mathematics, who has not wished to deal with numbers in computations, instead of that complicated trigonometry things. This book is an introduction to numbers, and their theory. You will learn from here all you need to know about numbers, fractions and percentages, followed by some basic number theory, also known as basic arithmetic, and then followed by more advanced aspects.

The story of numbers, or at least numbers employed by us humans, is long. Things go back to the Stone Age, where the sighting of a bison was reported with a “Ha” shout, the sighting of two bisons was reported with a “Ya”, and of three, with a “Rg”. And one day, an interesting thing happened. Gronk came back to camp, from his morning walk, shouting “Rg”, and pointing towards the plains. While Kelc and Tay, one coming back from the lake, and the other, from the hill nearby, both started yelling “Ya”.

So, which way to go? Times were hard, it was Winter, not much food left, and the more bisons hunted, the better. Big chief started thinking, then drinking, singing and dancing, and in the end, he cut his finger, and wrote on the wall of the cavern:

$$\text{Ya} + \text{Ya} > \text{Rg}$$

And with this, arithmetic was born. They went towards the lake, hunted the Ya + Ya bisons there, and had enough food for the rest of the Winter. Also, during the long Winter nights, they thought some more, and convened for “Uy” to designate the sighting of Ya + Ya bisons. And a few years after, after countless other hunts, they came upon the following fomula, that they wrote on the cavern wall too, and called Theorem:

$$\text{Ha} + \text{Rg} = \text{Uy}$$

So, this was for the beginnings, and many things have happened since, with countless improvings to this bison counting system. Romans in particular came with a system that no one really understands nowadays, I, II, III, IV, V, VI, . . . , apart from certain fine intellectuals, and sports fans, but as a matter of telling the whole story, we will employ here that system too, for labeling the parts of the present book.

As already mentioned, this book is an introduction to arithmetic, with the aim of telling you the basic theory of numbers, counting, fractions and percentages, followed by

some basic number theory, also known as basic arithmetic, and then followed by more advanced aspects. The book is organized in four parts, as follows:

Part I deals with numbers, counting, fractions and percentages, all you need to know, with the basics explained, and all sorts of useful tricks and formulae, in the spirit of the above Stone Age Theorem. We will also discuss some more advanced algebraic aspects, such as the finite groups, basic field theory, and quadratic residues.

Part II deals with the real numbers, and what can be done with them, in relation with arithmetic. In particular, we will get to know more about the primes, via the Euler formula, and other analysis tricks. Among others, we will have a look at the Mertens theorems, and Chebycheff estimates, and discuss the p -adic numbers as well.

Part III deals with the complex numbers, which are substantially more far-reaching, and what can be done with them, again in relation with arithmetic. As main topics here, we will talk about the Cardano formula in degree 3 and 4, higher degree curves, Gauss sums with their sign computed, and the transcendence of e and π .

Part IV goes back to questions from basic arithmetic, notably regarding the prime numbers, with more on the subject, by benefiting from the knowledge of real and complex numbers. We will talk here about the Riemann zeta function, first real and then complex, the Prime Number theorem, via various proofs, and the Riemann hypothesis.

In the hope that you will like this book, and get to love numbers and their theory. We will be quite quick on certain aspects, especially in what regards analysis, and for full preliminaries here, you have my introductory book to functions [8]. For more linear algebra and geometry, you have my advanced linear algebra book [9]. As for more abstract algebra in general, you have here my book on finite groups [10].

Many thanks to everyone, having helped me to learn about numbers, since childhood and up to nowadays, and still counting. Thanks as well to my cats, it's a bit hard to talk to them because they use quaternions, but I learned from them many things too.

Cergy, January 2026

Teo Banica

Contents

Preface	3
Part I. Number theory	9
Chapter 1. Numbers, counting	11
1a. Numbers	11
1b. Basic arithmetic	16
1c. Congruences, primes	21
1d. Binomial formula	25
1e. Exercises	32
Chapter 2. Fractions, percentages	33
2a. Fractions, rules	33
2b. Decimals, percentages	40
2c. Games, probability	44
2d. Binomial laws	50
2e. Exercises	56
Chapter 3. Groups, algebra	57
3a. Groups, examples	57
3b. General results	62
3c. Abelian groups	67
3d. Rings and fields	71
3e. Exercises	80
Chapter 4. Squares, residues	81
4a. Squares, residues	81
4b. Legendre symbol	89
4c. Further results	94
4d. Some applications	99
4e. Exercises	104

Part II. Real numbers	105
Chapter 5. Real numbers	107
5a. Real numbers	107
5b. Limits, series	111
5c. The number e	119
5d. The number π	124
5e. Exercises	128
Chapter 6. Functions, calculus	129
6a. Functions, continuity	129
6b. Derivatives, rules	133
6c. Higher derivatives	140
6d. Integrals, primitives	149
6e. Exercises	152
Chapter 7. Primes, revised	153
7a. Euler formula	153
7b. Stirling formula	160
7c. Mertens theorems	167
7d. Chebycheff estimates	172
7e. Exercises	176
Chapter 8. Algebra, fields	177
8a. Abstract algebra	177
8b. Manifolds, ideals	183
8c. Fields, extensions	189
8d. p -adic numbers	195
8e. Exercises	200
Part III. Complex numbers	201
Chapter 9. Complex numbers	203
9a. Complex numbers	203
9b. Polar coordinates	208
9c. Exponential writing	213
9d. Equations, roots	219
9e. Exercises	224

Chapter 10. Polynomials	225
10a. Resultant, discriminant	225
10b. Cardano formulae	232
10c. Spirals, lemniscates	239
10d. Nullstellensatz	245
10e. Exercises	248
Chapter 11. Gauss sums	249
11a. Gauss sums	249
11b. Further summing	254
11c. The Gauss sign	258
11d. Some applications	267
11e. Exercises	272
Chapter 12. Transcendence	273
12a. More about e	273
12b. Transcendence of e	282
12c. Approximating π	287
12d. Transcendence of π	293
12e. Exercises	296
Part IV. Analytic aspects	297
Chapter 13. Complex analysis	299
13a. Complex functions	299
13b. Holomorphic functions	303
13c. Cauchy formula	307
13d. Harmonic functions	313
13e. Exercises	320
Chapter 14. Zeta function	321
14a. Zeta function	321
14b. Basel formula	327
14c. Bernoulli numbers	335
14d. Special values	339
14e. Exercises	344
Chapter 15. Complex zeta	345

15a. Gamma function	345
15b. Complex zeta	353
15c. Riemann formula	359
15d. Further results	363
15e. Exercises	368
Chapter 16. Prime distribution	369
16a. Back to primes	369
16b. Prime distribution	372
16c. Selberg formula	378
16d. Riemann hypothesis	390
16e. Exercises	392
Bibliography	393
Index	397

Part I

Number theory

*I'm only happy when it rains
I'm only happy when it's complicated
And though I know you can't appreciate it
I'm only happy when it rains*

CHAPTER 1

Numbers, counting

1a. Numbers

You certainly know a bit about numbers $1, 2, 3, 4, \dots$, and we will be here, with this book, for learning more about them. Many things can be said, but instead of starting right away with some complicated mathematics, it is wiser to relax, and go back to these small numbers $1, 2, 3, 4, \dots$ 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 short, thinking well, it is probably wise to leave the multiplications for later, and start more modestly, with the sums. But before doing sums, we must first remember what the numbers themselves are. Which is a very good question, and in answer:

$$|\circ| = \text{one}$$

$$|\circ\circ| = \text{two}$$

$$|\circ\circ\circ| = \text{three}$$

$$|\circ\circ\circ\circ| = \text{four}$$

$$|\circ\circ\circ\circ\circ| = \text{five}$$

$$|\circ\circ\circ\circ\circ\circ| = \text{six}$$

$$|\circ\circ\circ\circ\circ\circ\circ| = \text{seven}$$

$$|\circ\circ\circ\circ\circ\circ\circ\circ| = \text{eight}$$

$$|\circ\circ\circ\circ\circ\circ\circ\circ\circ| = \text{nine}$$

$$|\circ\circ\circ\circ\circ\circ\circ\circ\circ\circ| = \text{ten}$$

$$\vdots$$

Which sounds a bit boring, so what about stopping there at ten, and also replacing, for purely computational purposes, the various English words above by some mathematical symbols which are easy to draw, say by symbols $1, \dots, \diamond$, as follows:

$$\begin{aligned}
 | \circ | &= 1 \\
 | \circ \circ | &= 2 \\
 | \circ \circ \circ | &= 3 \\
 | \circ \circ \circ \circ | &= 4 \\
 | \circ \circ \circ \circ \circ | &= 5 \\
 | \circ \circ \circ \circ \circ \circ | &= 6 \\
 | \circ \circ \circ \circ \circ \circ \circ | &= 7 \\
 | \circ \circ \circ \circ \circ \circ \circ \circ | &= 8 \\
 | \circ \circ \circ \circ \circ \circ \circ \circ \circ | &= 9 \\
 | \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ | &= \diamond
 \end{aligned}$$

In order to count past ten, without the need of more symbols, we can trick. Let us introduce indeed a new symbol 0, called zero, in the following way:

$$\diamond = 10$$

That is, we are making here the tricky convention that the 1 in 10 stands for ten, and the 0 in 10 stands for nothing. And with this convention, is it quite clear that we can count now beyond 10, a bit in the same way, as follows:

$$\begin{aligned}
 | \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ | &= 10 \\
 | \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ | &= 11 \\
 | \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ | &= 12 \\
 | \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ | &= 13 \\
 | \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ | &= 14 \\
 | \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ | &= 15 \\
 | \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ | &= 16 \\
 | \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ | &= 17 \\
 | \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ | &= 18 \\
 | \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ \circ | &= 19 \\
 | \circ | &= 20 \\
 | \circ | &= 21 \\
 | \circ | &= 22 \\
 | \circ | &= 23 \\
 &\vdots
 \end{aligned}$$

Very nice all this, so let us formulate our findings in the following way:

DEFINITION 1.1. *The numbers are sequences of type*

$$n = a_1 a_2 \dots a_k$$

with $a_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, and $a_1 \neq 0$. These numbers count, as follows:

- (1) If a set has $a_1 \in \{1, \dots, 9\}$ objects, the set count is $n = a_1$,
- (2) If a set has $a_1 \in \{1, \dots, 9\}$ chunks of 10 objects, followed by $a_2 \in \{0, \dots, 9\}$ objects, the set count is $n = a_1 a_2$,
- (3) If a set has $a_1 \in \{1, \dots, 9\}$ chunks of 100 objects, followed by $a_2 \in \{0, \dots, 9\}$ chunks of 10 objects, and $a_3 \in \{0, \dots, 9\}$ objects, the count is $n = a_1 a_2 a_3$,

.. and so on, the idea being that we can count any set, no matter how big, in this way.

In mathematical notation, the counting rules above can be summarized as follows, with obvious meanings for the sum and product operations $+$ and \times :

$$a_1 = a_1$$

$$a_1 a_2 = 10 \times a_1 + a_2$$

$$a_1 a_2 a_3 = 100 \times a_1 + 10 \times a_2 + a_3$$

$$a_1 a_2 a_3 a_4 = 1000 \times a_1 + 100 \times a_2 + 10 \times a_3 + a_4$$

$$\vdots$$

We conclude that, again in standard mathematical notation, we have the following formula, for an arbitrary number $n = a_1 a_2 \dots a_k$, as in Definition 1.1:

$$a_1 a_2 \dots a_k = 10^{k-1} \times a_1 + 10^{k-2} \times a_2 + \dots + 10 \times a_{k-1} + a_k$$

With this understood, the problem is now, how these numbers $n = a_1 a_2 \dots a_k$ add? And here, no matter how we approach this question, via counting as in Definition 1.1, or via a mathematical formula with powers of 10 as above, we are led to the conclusion that things are quite clear, up to adding the digits $a \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ themselves.

But, in what regards this latter question, we know the answer to it, learned the hard way in school. So, problem solved, and as our first theorem in this book, we have:

THEOREM 1.2. *The digits $a \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ add according to the table*

+	1	2	3	4	5	6	7	8	9
1	2	3	4	5	6	7	8	9	10
2	3	4	5	6	7	8	9	10	11
3	4	5	6	7	8	9	10	11	12
4	5	6	7	8	9	10	11	12	13
5	6	7	8	9	10	11	12	13	14
6	7	8	9	10	11	12	13	14	15
7	8	9	10	11	12	13	14	15	16
8	9	10	11	12	13	14	15	16	17
9	10	11	12	13	14	15	16	17	18

and then the numbers $n = a_1a_2 \dots a_k$ add in the obvious way, using this.

PROOF. As already mentioned, this is something that we learned the hard way in school, and which is undoubtedly difficult, so we will not attempt to prove this here, with full details. Here are, however, a few comments about all this:

(1) In what regards the table, that comes by counting sets, by hand, and recording the results. Not to forget, of course, to duly memorize all this afterwards.

(2) As a further remark about the table, observe that this consists of the numbers $1, 2, 3, 4, \dots$ themselves, filling the / diagonals, in the obvious way. You can call this observation theorem if you want, and in any case, this is certainly something that does happen, and of course I used this myself, when typing in the table. Easy work.

(3) Once the addition table for digits digested, in order to add two arbitrary numbers, $n = a_1a_2 \dots a_k$ and $m = b_1b_2 \dots b_s$, we can do this in the following way:

$$\begin{aligned}
 & a_1a_2 \dots a_k + b_1b_2 \dots b_s \\
 = & (10^{k-1} \times a_1 + 10^{k-2} \times a_2 + \dots + 10 \times a_{k-1} + a_k) \\
 & + (10^{s-1} \times b_1 + 10^{s-2} \times b_2 + \dots + 10 \times b_{s-1} + b_s) \\
 = & 10^{k-1} \times a_1 + 10^{s-1} \times b_1 + \dots + 10 \times a_{k-1} + 10 \times b_{s-1} + a_k + b_s \\
 = & 10(10^{k-2} \times a_1 + 10^{s-2} \times b_1 + \dots + a_{k-1} + b_{s-1}) + a_k + b_s
 \end{aligned}$$

Thus, proceeding from right to left, the last digit will obviously be $a_k + b_s$, or rather the last digit of $a_k + b_s$, in case $a_k + b_s \geq 10$, and so on, up to the first digit.

(4) Equivalently, we have here the basic algorithm for addition, obtained by putting $n = a_1a_2 \dots a_k$ on top of $m = b_1b_2 \dots b_s$, and summing as in (3), that you know well. \square

Getting now to multiplication, things are considerably tougher here, again learned the hard way in school, with the relevant theorem here being as follows:

THEOREM 1.3. *The digits $a \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ multiply according to the table*

\times	1	2	3	4	5	6	7	8	9
1	1	2	3	4	5	6	7	8	9
2	2	4	6	8	10	12	14	16	18
3	3	6	9	12	15	18	21	24	27
4	4	8	12	16	20	24	28	32	36
5	5	10	15	20	25	30	35	40	45
6	6	12	18	24	30	36	42	48	54
7	7	14	21	28	35	42	49	56	63
8	8	16	24	32	40	48	56	64	72
9	9	18	27	36	45	54	63	72	81

and then the numbers $n = a_1 a_2 \dots a_k$ multiply in the obvious way, using this.

PROOF. As before with Theorem 1.2, this is something that we learned the hard way in school, and which is undoubtedly difficult, so we will not attempt to prove this here, with full details. Here are, however, a few comments about all this:

(1) In what regards the table, again that comes by counting sets, by hand, and recording the results. Not to forget, of course, to duly memorize all this afterwards.

(2) As a further remark about the table, this was harder for me to type in, algorithmically, than the addition one, and what I did is to type in each line, starting from the left, with on line i by adding every time the number i to the previous one. Mathematically, this corresponds to the fact that each line of the table is an arithmetic progression.

(3) Once the multiplication table for digits digested, in order to multiply two arbitrary numbers, $n = a_1 a_2 \dots a_k$ and $m = b_1 b_2 \dots b_s$, we can do this in the following way:

$$\begin{aligned}
 & a_1 a_2 \dots a_k \times b_1 b_2 \dots b_s \\
 = & (10^{k-1} \times a_1 + 10^{k-2} \times a_2 + \dots + 10 \times a_{k-1} + a_k) \\
 & \times (10^{s-1} \times b_1 + 10^{s-2} \times b_2 + \dots + 10 \times b_{s-1} + b_s) \\
 = & 10^{k+s-2} \times a_1 b_1 + \dots + 10 \times a_{k-1} b_s + 10 \times a_k b_{s-1} + a_k b_s \\
 = & 10(10^{k+s-3} \times a_1 b_1 + \dots + a_{k-1} b_s + a_k b_{s-1}) + a_k b_s
 \end{aligned}$$

Thus, when proceeding from right to left, the last digit will obviously be $a_k b_s$, or rather the last digit of $a_k b_s$, in case $a_k b_s \geq 10$, and so on, up to the first digit.

(4) Equivalently, we have the algorithm for multiplication, obtained by putting $n = a_1 a_2 \dots a_k$ on top of $m = b_1 b_2 \dots b_s$, and multiplying as in (3), that you know well. \square

And with this, good news, done with the hard mathematics, with the rest of the present book being more or less trivialities, coming from the above.

1b. Basic arithmetic

With counting and numbers understood, let us develop now some basic arithmetic. We have here the following key notion, which often appears in the real life:

DEFINITION 1.4. *We say that b divides a , and we write $b|a$, when*

$$a = bc$$

for some number c . In this case we also use the following notation,

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

with this being called fraction, for designating this quotient number c .

All this is quite intuitive, and the fractions are subject to a number of simple formulae, which are all useful, in the real life, which can be summarized as follows:

THEOREM 1.5. *The fractions add according to the following formula,*

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

they subtract according to the following formula,

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

they multiply according to the following formula,

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

and they divide according to the following formula,

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

provided that the quotient a/b is a multiple of the quotient c/d .

PROOF. You certainly know this, but for making sure that this is good science, let us see how the mathematical proof, based on Definition 1.4, goes. According to Definition 1.4 we have the following equality, that we will use many times, in what follows:

$$\frac{a}{b} = \frac{ad}{bd} \quad , \quad \forall d$$

(1) In what regards the addition formula, this can be established as follows:

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

(2) The proof of the subtraction formula is similar, as follows:

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

(3) In what regards now the multiplication formula, this comes from:

$$\left(\frac{a}{b} \cdot \frac{c}{d}\right) bd = \left(\frac{a}{b} \cdot b\right) \left(\frac{c}{d} \cdot d\right) = ac$$

(4) As for the division formula, this can be proved as follows:

$$\left(\frac{a}{b} : \frac{c}{d}\right) bc = \frac{abc}{b} : \frac{c}{d} = ac : \frac{c}{d} = ad$$

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

And more on this, divisibility of numbers, and on fractions too, in the above sense, and in some generalized sense too, when $b \nmid a$, later in this book.

Moving ahead with more arithmetic, we have the following result:

THEOREM 1.6. *Given two numbers a, b , we can talk about their greatest common divisor (a, b) , and their least common multiple $[a, b]$. We can write*

$$a = da' \quad , \quad b = db'$$

with a', b' being numbers having no common divisor, and we have:

$$(a, b) = d \quad , \quad [a, b] = da'b'$$

Also, (a, b) and $[a, b]$ are subject to the formula $(a, b)[a, b] = ab$.

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

(1) To start with, we can talk indeed about the greatest common divisor $d = (a, b)$ of any two numbers $a, b \in \mathbb{N}$, in several equivalent ways, as follows:

– The simplest way is to argue that since all common divisors $e|a, b$ satisfy $e \leq a, b$, we can pick the greatest such common divisor, $d = (a, b)$.

– Alternatively, we can say that $d = (a, b)$ is the smallest number having the property that any common divisor $e|a, b$ must divide it, $e|d$.

– Yet another approach is by recurrence on $a + b$. Indeed, assuming $a > b$, we can perform the division $a = bn + c$, and then set $(a, b) = (b, c)$, by recurrence.

(2) In practice now, there is some discussion needed here, in order to prove that the above 3 methods yield indeed the same number $d = (a, b)$. But this is best seen via the third method, which produces the same numbers as the first and second methods.

(3) So, this was for the story of the greatest common divisor (a, b) , and in what regards the least common multiple $[a, b]$, the story here is similar, and we will leave the details as an exercise. Alternatively, if looking for a quick formal proof here, we can define $[a, b]$ by starting with (a, b) , and using the formula $(a, b)[a, b] = ab$, discussed below.

(4) Next, with $d = (a, b)$, we can decompose our two numbers as follows:

$$a = da' \quad , \quad b = db'$$

We have then the following implications, coming from definitions:

$$(a, b) = d \implies (da', db') = d \implies (a', b') = 1$$

Thus, we managed to write a, b as in the statement, and with this done, it is clear that we must have $[a, b] = da'b'$, so we have both formulae in the statement, namely:

$$(a, b) = d \quad , \quad [a, b] = da'b'$$

(5) Finally, observe that we have the following formula:

$$ab = d^2 a' b' = d \times da' b' = (a, b)[a, b]$$

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

We can basically do the same with three numbers, as follows:

THEOREM 1.7. *Given three numbers a, b, c , we can talk about their greatest common divisor (a, b, c) , and their least common multiple $[a, b, c]$, and if we write*

$$a = da' \quad , \quad b = db' \quad , \quad c = dc'$$

with $(a', b', c') = 1$, and then further decompose each pair (a', b') , (a', c') , (b', c') , by using their respective greatest common divisors, we are led to a decomposition as follows,

$$a = dpqx \quad , \quad b = dpry \quad , \quad c = dqrz$$

and in terms of this decomposition, we have the following formulae:

$$(a, b, c) = d \quad , \quad [a, b, c] = dpqrxzy$$

Also, $(a, b, c)^2[a, b, c]$ divides abc , but with these numbers being different, in general.

PROOF. As before, the fact that we can talk indeed about greatest common divisors (a, b, c) , and about least common multiples $[a, b, c]$ too, is something which is clear. Now if we set $d = (a, b, c)$, we can decompose our three numbers as follows:

$$a = da' \quad , \quad b = db' \quad , \quad c = dc'$$

We have then the following implications, coming from definitions:

$$(a, b, c) = d \implies (da', db', dc') = d \implies (a', b', c') = 1$$

Thus, we have managed to write a, b, c as in the statement, and we have:

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

In order to compute now $[a', b', c']$, we can look at the pairs (a', b') , (a', c') , (b', c') , and apply to them the theory that we learned in Theorem 1.6. Indeed, let us set:

$$p = (a', b') \quad , \quad q = (a', c') \quad , \quad r = (b', c')$$

We are led in this way to decompositions as follows, for the numbers a', b', c' :

$$a' = pqx \quad , \quad b' = pry \quad , \quad c' = qrz$$

As a conclusion, our original numbers a, b, c decompose as follows:

$$a = dpqx \quad , \quad b = dpry \quad , \quad c = dqrz$$

But with these formulae in hand, the numbers that we were looking for are:

$$(a, b, c) = d \quad , \quad [a, b, c] = dpqrxyz$$

Now when multiplying our numbers a, b, c , we have the following formula:

$$\begin{aligned} abc &= dpqx \cdot dpry \cdot dqrz \\ &= d^3(pqr)^2xyz \\ &= d^2 \cdot dpqrxyz \cdot pqr \\ &= (a, b, c)^2 \cdot [a, b, c] \cdot pqr \end{aligned}$$

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

In the general case now, that of k numbers, we have the following result:

THEOREM 1.8. *Given numbers a_1, \dots, a_k , we can talk about their greatest common divisor (a_1, \dots, a_k) , and their least common multiple $[a_1, \dots, a_k]$, and we can write*

$$a_1 = da'_1 \quad , \quad \dots \quad , \quad a_k = da'_k$$

with $d = (a_1, \dots, a_k)$, and with $(a'_1, \dots, a'_k) = 1$. Also, the product

$$(a_1, \dots, a_k)^{k-1} [a_1, \dots, a_k]$$

divides the product $a_1 \dots a_k$, but these numbers are different, in general.

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

(1) As before, the fact that we can talk indeed about greatest common divisors, and about least common multiples, is something which is clear from definitions.

(2) Also as before, we can divide our numbers a_1, \dots, a_k by their common divisor $d = (a_1, \dots, a_k)$, and we reach to a decomposition as follows, with $(a'_1, \dots, a'_k) = 1$:

$$a_1 = da'_1 \quad , \quad \dots \quad , \quad a_k = da'_k$$

(3) However, and here comes the point, when it comes to suitably decomposing our numbers a_1, \dots, a_k , or rather their reduced versions a'_1, \dots, a'_k , by using their various common divisors, as we did in Theorem 1.6 at $k = 2$, and in Theorem 1.7 at $k = 3$, things become considerably more complicated at $k = 4$ and higher. We can only recommend here doing some computations at $k = 4$, in order to understand what the difficulty is.

(4) Summarizing, we cannot say much, as a continuation of this, along the lines of what we did before at $k = 2, 3$, and the only obvious thing that can be said, completing our theorem, is the last assertion, which is something that we know, from Theorem 1.7. \square

Moving on, still talking divisibility, here is a key result:

THEOREM 1.9. *Given two numbers satisfying $(a, b) = 1$, we can write*

$$ae + bf = 1$$

for certain integers $e, f \in \mathbb{Z}$.

PROOF. This might sound a bit surprising, but give me any two numbers which are prime to each other, say 7 and 10, and after thinking a bit, here is what I find:

$$7 \times 3 - 10 \times 2 = 1$$

So, let us try to prove now the result, in general. For this purpose, let us look at:

$$b, 2b, 3b, \dots, ab$$

This is a certain collection of a numbers, and our claim is that the remainders of these numbers, modulo a , are different. Indeed, this is something coming from:

$$\begin{aligned} cb = db(a) &\iff a|(c-d)b \\ &\iff a|c-d \\ &\iff c = d \end{aligned}$$

Thus, our a numbers above are distinct modulo a , and so we have, still modulo a :

$$\{b, 2b, 3b, \dots, ab\} = \{1, 2, 3, \dots, a\}$$

But this does the job, because we get a certain $f \in \{1, \dots, a\}$ such that:

$$bf = 1(a)$$

Thus we must have $ae + bf = 1$, for a certain $e \in \mathbb{Z}$, as desired. \square

Here is a useful generalization of the above result:

THEOREM 1.10. *Given numbers satisfying $(a_1, \dots, a_k) = 1$, we can write*

$$a_1e_1 + a_2e_2 + \dots + a_ke_k = 1$$

for certain integers $e_1, \dots, e_k \in \mathbb{Z}$.

PROOF. We already know from Theorem 1.9 that this holds at $k = 2$, and in general, the result will follow from this, the idea being as follows:

(1) Let us first see how the proof goes at $k = 3$. Given three numbers a, b, c having no common divisor, $(a, b, c) = 1$, let us write them as in Theorem 1.7, as follows:

$$a = pqx \quad , \quad b = pry \quad , \quad c = qrz$$

Now let us look at the sums of the following type, with $e, f, g \in \mathbb{Z}$:

$$ae + bf + cg = pqxe + pryf + qrzg$$

– As a first observation, since we have $(qx, ry) = 1$, we know from Theorem 1.9 that the quantities of type $qxe + ryf$ will range over the whole \mathbb{Z} .

– Next, and getting now towards what we want to prove, we conclude from this that the quantities of type $pqxe + pryf$ range over the whole $p\mathbb{Z}$.

– But then, since we have $(p, qrz) = 1$, we conclude, again by using Theorem 1.9, that the above sums $pqxe + pryf + qrzg$ range over the whole \mathbb{Z} .

(2) Thus, theorem proved at $k = 3$, and the proof in general is similar, by recurrence on k . To be more precise, it is technically convenient to look at a slightly more general statement, saying that given a_1, \dots, a_k , we can always find $e_1, \dots, e_k \in \mathbb{Z}$ such that:

$$a_1e_1 + a_2e_2 + \dots + a_ke_k = (a_1, \dots, a_k)$$

But with this picture in hand, it is quite clear that the above arguments used at $k = 3$ will apply in general, and will give the result, by recurrence on $k \in \mathbb{N}$. \square

Summarizing, we have up and working a useful theory of greatest common divisors, and least common multiples, but obviously this is just the tip of the iceberg, and many interesting questions remain open. We will be back to them, later in this book.

1c. Congruences, primes

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

DEFINITION 1.11. *We say that $a, b \in \mathbb{Z}$ are congruent modulo $c \in \mathbb{Z}$, and write*

$$a \equiv b \pmod{c}$$

when c divides $b - a$.

A first interesting question concerns solving $a \equiv 0 \pmod{n}$, with n fixed and small. There is a bit of recursivity that can be used, in relation with this, as shown by:

$$\begin{aligned} 6|a &\iff 2|a \text{ and } 3|a \\ 10|a &\iff 2|a \text{ and } 5|a \\ 12|a &\iff 3|a \text{ and } 4|a \\ 14|a &\iff 2|a \text{ and } 7|a \\ 15|a &\iff 3|a \text{ and } 5|a \\ 18|a &\iff 2|a \text{ and } 9|a \\ 20|a &\iff 4|a \text{ and } 5|a \\ 21|a &\iff 3|a \text{ and } 7|a \\ 22|a &\iff 2|a \text{ and } 11|a \\ 24|a &\iff 3|a \text{ and } 8|a \end{aligned}$$

In general, based on these observations, the idea is that by writing $n = n_1 \dots n_k$ with the factors n_i having no common divisor, we just have to solve this question for certain special values of n , excluding $n = 6, 10, 12, 14, 15, 18, 20, 21, 22, 24, \dots$

These special values of n are called “powers of primes”, and many things can be said about them. More on them later in this chapter, and then later in this book.

In practice, the first such numbers, powers of primes, are as follows:

$$q = 2, 3, 4, 5, 7, 8, 9, 11, 13, 16, 17, 19, 23, \dots$$

And in what regards solving $a = 0(q)$, with respect to these powers of primes, there are many useful tricks here, which can be summarized as follows:

THEOREM 1.12. *Given a positive integer $a = a_1 \dots a_k$, we have:*

- (1) $2|a$ when $2|a_k$.
- (2) $3|a$ when $3|\sum a_i$.
- (3) $4|a$ when $4|a_{k-1}a_k$.
- (4) $5|a$ when $5|a_k$.
- (5) $8|a$ when $8|a_{k-2}a_{k-1}a_k$.
- (6) $9|a$ when $9|\sum a_i$.
- (7) $11|a$ when $11|\sum (-1)^i a_i$.
- (8) $16|a$ when $16|a_{k-3}a_{k-2}a_{k-1}a_k$.

PROOF. Here the $q = 2^r$, 5 assertions follow from $10 = 2 \times 5$, the $q = 3, 9$ assertions follow from $10 = 9 + 1$, and the $q = 11$ assertion follows from $10 = 11 - 1$. \square

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 1.13. *Assuming that we have convinced mankind to change the numeration basis from 10 to 14, given a positive integer $a = a_1 \dots a_k$, we have:*

- (1) $2|a$ when $2|a_k$.
- (2) $3|a$ when $3|\sum (-1)^i a_i$.
- (3) $4|a$ when $4|a_{k-1}a_k$.
- (4) $5|a$ when $5|\sum (-1)^i a_i$.
- (5) $7|a$ when $7|a_k$.
- (6) $8|a$ when $8|a_{k-2}a_{k-1}a_k$.
- (7) $9|a$ when $9|\sum (-1)^i a_i$.
- (8) $13|a$ when $13|\sum a_i$.
- (9) $16|a$ when $16|a_{k-3}a_{k-2}a_{k-1}a_k$.

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

In short, good news, we have solved indeed 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 here, and launch an online petition if enthusiastic about it.

Time now to get into prime numbers, which will be a main theme of discussion, in this book. 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

Actually those between 101 and 200 are mandatory too, here they are:

101, 103, 107, 109, 113, 127, 131, 137, 139, 149

151, 157, 163, 167, 173, 179, 181, 191, 193, 197, 199

But then, can we ignore those between 201 and 300. These are as follows:

211, 223, 227, 229, 233, 239, 241

251, 257, 263, 269, 271, 277, 281, 283, 293

Not to forget the primes between 301 and 400, which are as follows:

307, 311, 313, 317, 331, 337, 347, 349

353, 359, 367, 373, 379, 383, 389, 397

And we have kept the best for the end, primes between 401 and 500, which are:

401, 409, 419, 421, 431, 433, 439, 443, 449

457, 461, 463, 467, 479, 487, 491, 499

So, these are our beasts, in arithmetic, and try to get familiar with them, and learn some of their powers too, because these prime powers are very useful too.

We have already met prime numbers in the above, when talking divisibility, 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, with proofs and everything.

First, as a formal definition for the prime numbers, we have:

DEFINITION 1.14. *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 the conditions (1,2,3) is something intuitive and standard, which can be deduced by using our common divisor technology developed above. Observe also 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 we know, already used in the above, we have:

THEOREM 1.15. *Any integer $n > 1$ decomposes uniquely as*

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

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

PROOF. This is something very standard, related to the equivalent conditions (1,2,3) in Definition 1.14, which formally comes by recurrence on n . As an interesting exercise here, work out this for all the integers $n \leq 100$, with no calculators allowed. \square

As a first result now about the prime numbers themselves, we have:

THEOREM 1.16. *There is an infinity of prime numbers.*

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

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

THEOREM 1.17. *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, ...*
- (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:

<u>2</u>	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	<u>3</u>		5		7		9		11		13		15		17		19	
			<u>5</u>		7				11		13				17		19	
					<u>7</u>				11		13				17		19	
									<u>11</u>		13				17		19	
											<u>13</u>				17		19	

\vdots

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

The above algorithm, while mathematically rather trivial, is something quite fascinating, because it suggests all sorts of mechanical ways of dealing with the primes, via analysis and physics and engineering. Let us record this as a conjecture:

CONJECTURE 1.18. *A good analyst, physicist and engineer would probably have no troubles in elucidating everything about primes, using the sieve method.*

And we will end the present section with this. Mystery.

1d. Binomial formula

Time now for some more advanced mathematics, in relation with counting. We have the following key result, solving a problem which often appears in real life:

THEOREM 1.19. *The number of possibilities of choosing k objects among n objects is*

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

called binomial number, where $n! = 1 \cdot 2 \cdot 3 \dots (n-2)(n-1)n$, 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:

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

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 2 choices for the object to be designated # $(k-1)$, and 1 choice for the object to be designated # k . We conclude that we have:

$$M = k(k-1) \dots 2 \cdot 1 = k!$$

By putting now everything together, the number that we were computing is:

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

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

The binomial numbers, as constructed above, are quite fascinating objects, and the more you know about them, the better your mathematics will be. Trust me here.

To start with, here are some basic formulae for binomial coefficients that you should definitely memorize, and pull right away, when needed in your computations:

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

$$\binom{n}{3} = \frac{n(n-1)(n-2)}{6}$$

$$\binom{n}{4} = \frac{n(n-1)(n-2)(n-3)}{24}$$

Here are as well some numerics, with $n = k, k+1, k+2, \dots$ in each case, that you should know well too, and pull out instantly, when needed in your computations:

$$\binom{n}{2} = 1, 3, 6, 10, 15, 21, 28, \dots$$

$$\binom{n}{3} = 1, 4, 10, 20, 35, 56, \dots$$

$$\binom{n}{4} = 1, 5, 15, 35, 70, \dots$$

Finally, talking numerics, as an important adding to Theorem 1.19, we have:

CONVENTION 1.20. *By definition we have the formula*

$$0! = 1$$

as for the following binomial coefficient computation to work,

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

in agreement with what Theorem 1.19 says, requiring $\binom{n}{n} = 1$.

Going ahead now with more mathematics and less philosophy, with Theorem 1.19 complemented by this convention being in final form, we have:

THEOREM 1.21. *We have the binomial formula*

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

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

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

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

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

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

In order to finish, it remains to compute the coefficients C_k . But, according to our product formula, C_k is the number of choices for the k needed a variables among the n available a variables. Thus, according to Theorem 1.19, we have the following formula:

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

We are therefore led to the formula in the statement. \square

What we have in Theorem 1.21 is quite interesting. At small values of n we obtain:

$$\begin{aligned} (a + b)^0 &= 1 \\ (a + b)^1 &= a + b \\ (a + b)^2 &= a^2 + 2ab + b^2 \\ (a + b)^3 &= a^3 + 3a^2b + 3ab^2 + b^3 \\ (a + b)^4 &= a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4 \\ (a + b)^5 &= a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5 \\ (a + b)^6 &= a^6 + 6a^5b + 15a^4b^2 + 20a^3b^3 + 15a^2b^4 + 6ab^5 + b^6 \\ &\vdots \end{aligned}$$

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

$$\begin{array}{ccccccc} & & & & 1 & & \\ & & & & & & \\ & & & 1 & , & 1 & \\ & & 1 & , & 2 & , & 1 \\ & 1 & , & 3 & , & 3 & , & 1 \\ & & 1 & , & 4 & , & 6 & , & 4 & , & 1 \\ & 1 & , & 5 & , & 10 & , & 10 & , & 5 & , & 1 \\ 1 & , & 6 & , & 15 & , & 20 & , & 15 & , & 6 & , & 1 \\ & & & & & & & & & & & \\ & & & & & & & & & & & \vdots \end{array}$$

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

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

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

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

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

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

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

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

By using the binomial formula, this formula becomes:

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

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

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

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 later, on a regular basis. However, the idea will be always the same, namely that in order to find such formulae you have a choice between algebra and combinatorics, a bit as in the above, and that when it comes to formal proofs, the brute-force computation method is something 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.

Getting now to more advanced aspects of the binomial coefficients, let us formulate, as a complement to the various particular cases discussed before:

DEFINITION 1.23. *The central and middle binomial coefficients are the numbers*

$$D_k = \binom{2k}{k} \quad , \quad E_n = \binom{n}{\lfloor n/2 \rfloor}$$

with $\lfloor \cdot \rfloor$ standing as usual for the integer part.

Observe that we can recover the central binomial coefficients as particular cases of the middle binomial coefficients, due to the following trivial formula:

$$D_k = E_{2k}$$

However, in practice, the central binomial coefficients D_k are the truly interesting quantities, and the middle binomial coefficients E_n remain something secondary. Regarding now the numerics for the central binomial coefficients, these are as follows:

$$D_k = 1, 2, 6, 20, 70, 252, 924, 3432, 12870, 48620, \dots$$

This sequence is actually something quite fascinating, and if you are a number theory nerd, and hope so are you, one of the first things that you will discover, by playing with it, is that these central binomial coefficients factorize as follows:

$$\begin{aligned} D_k = & 1 \times 1, 2 \times 1, 3 \times 2, 4 \times 5, 5 \times 14, 6 \times 42, \\ & 7 \times 132, 8 \times 429, 9 \times 1430, 10 \times 4862, \dots \end{aligned}$$

Thus, we are led in this way to the following conjecture:

CONJECTURE 1.24. *The central binomial coefficients factorize as*

$$D_k = (k+1)C_k$$

with $C_k = 1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, \dots$ being certain integers.

So, how to prove this? In answer, this is undoable with bare hands, and we will have to trick. Obviously, we need to know what the quotients $C_k = D_k/(k+1)$ count, and for this purpose, we can use the following fact, which fits with the above numerics:

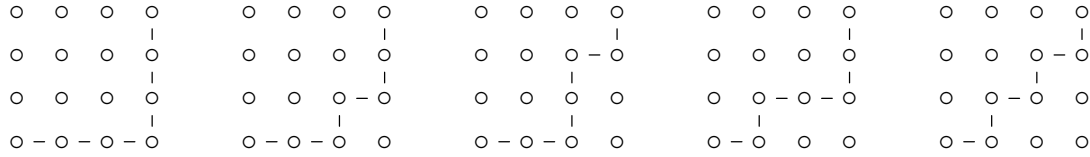
THEOREM 1.25. *The following objects are counted by the same numbers C_k , called Catalan numbers, given by $C_0 = C_1 = 1$ and $C_{k+1} = \sum_{a+b=k} C_a C_b$:*

- (1) *The length $2k$ loops on \mathbb{N} , based at 0.*
- (2) *The noncrossing pairings of $1, \dots, 2k$.*
- (3) *The noncrossing partitions of $1, \dots, k$.*
- (4) *The length $2k$ Dyck paths in the plane.*

PROOF. All this is standard combinatorics, the idea being as follows:

(1) To start with, in what regards the various objects involved, the length $2k$ loops on \mathbb{N} are the length $2k$ loops on \mathbb{N} that we know, and the same goes for the noncrossing pairings of $1, \dots, 2k$, and for the noncrossing partitions of $1, \dots, k$, the idea here being that you must be able to draw the pairing or partition in a noncrossing way.

(2) Regarding now the length $2k$ Dyck paths in the plane, these are by definition the paths from $(0, 0)$ to (k, k) , marching North-East over the integer lattice $\mathbb{Z}^2 \subset \mathbb{R}^2$, by staying inside the square $[0, k] \times [0, k]$, and staying as well under the diagonal of this square. As an example, here are the 5 possible Dyck paths at $n = 3$:



(3) Thus, we have definitions for all objects involved, and in each case, if you start counting them, you end up with the same sequence, namely that in Conjecture 1.24:

$$1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, \dots$$

(4) In order to prove now that (1-4) produce indeed the same numbers, many things can be said. The idea is that, leaving aside mathematical brevity, and more specifically abstract reasonings of type $a = b, b = c \implies a = c$, what we have to do, in order to fully understand what is going on, is to establish $\binom{4}{2} = 6$ equalities, via bijective proofs.

(5) But this can be done, indeed. As an example here, the noncrossing pairings of $1, \dots, 2k$ from (2) are in bijection with the noncrossing partitions of $1, \dots, k$ from (3), via fattening the pairings and shrinking the partitions. We will leave the details here as an instructive exercise, and exercise as well, to add (1) and (4) to the picture.

(6) However, matter of having our theorem formally proved, I mean by me professor and not by you student, here is a less elegant argument, which is however very quick, and does the job. The point is that, in each of the cases (1-4) under consideration, the numbers C_k that we get are easily seen to be subject to the following recurrence:

$$C_{k+1} = \sum_{a+b=k} C_a C_b$$

Now the initial data being the same, namely $C_0 = C_1 = 1$, in each of the cases (1-4) under consideration, we get indeed the same numbers, as desired. \square

Now we can pass to the second step, namely selecting in the above list the objects that we find the most convenient to count, and count them. This leads to:

THEOREM 1.26. *The Catalan numbers are given by the formula*

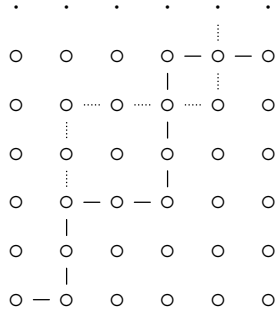
$$C_k = \frac{1}{k+1} \binom{2k}{k}$$

with this being best seen by counting the length $2k$ Dyck paths in the plane.

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

(1) Let us count indeed the Dyck paths in the plane. For this purpose, we use a trick. Indeed, if we ignore the assumption that our path must stay under the diagonal of the square, we have $\binom{2k}{k}$ such paths. And among these, we have the “good” ones, those that we want to count, and then the “bad” ones, those that we want to ignore.

(2) So, let us count the bad paths, those crossing the diagonal of the square, and reaching the higher diagonal next to it, the one joining $(0, 1)$ and $(k, k+1)$. In order to count these, the trick is to “flip” their bad part over that higher diagonal, as follows:



(3) Now observe that, as it is obvious on the above picture, due to the flipping, the flipped bad path will no longer end in (k, k) , but rather in $(k-1, k+1)$. Moreover, more is true, in the sense that, by thinking a bit, we see that the flipped bad paths are precisely those ending in $(k-1, k+1)$. Thus, we can count these flipped bad paths, and so the bad paths, and so the good paths too, and so good news, we are done.

(4) To finish now, by putting everything together, we have:

$$\begin{aligned} C_k &= \binom{2k}{k} - \binom{2k}{k-1} \\ &= \binom{2k}{k} - \frac{k}{k+1} \binom{2k}{k} \\ &= \frac{1}{k+1} \binom{2k}{k} \end{aligned}$$

Thus, we are led to the formula in the statement. And good work that we did, among others with Conjecture 1.24 being now proved. Nothing can resist us. \square

As a basic application now of our binomial formula technology, we have:

THEOREM 1.27. *We have the following congruence, for any prime p ,*

$$a^p \equiv a \pmod{p}$$

called Fermat's little theorem.

PROOF. We can prove this by recurrence on $a \in \mathbb{N}$, as follows:

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

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 coefficients, given by:

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

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 $a \rightarrow -a$ we get it for any $a \in \mathbb{Z}$, as desired. \square

1e. Exercises

This was a very basic chapter, save for some more advanced material at the end, in relation with counting and binomial coefficients. As exercises on this, we have:

EXERCISE 1.28. *Review if needed the usual rules for addition and multiplication.*

EXERCISE 1.29. *Clarify all the details, in relation with $abc = (a, b, c)^2[a, b, c]pqr$.*

EXERCISE 1.30. *Then study what happens, via suitable examples, for a, b, c, d .*

EXERCISE 1.31. *Find some clever criteria for divisibility by 7, and by 13.*

EXERCISE 1.32. *Count, before sleeping, the prime numbers listed in the above.*

EXERCISE 1.33. *Compute and memorize the coefficients $\binom{n}{k}$, for all $n \leq 10$.*

EXERCISE 1.34. *Study a bit the arithmetic properties of the numbers $\binom{n}{k}$.*

EXERCISE 1.35. *Learn more about the Catalan numbers C_k , and their occurrences.*

As bonus exercise, read a bit about atoms, molecules, cells, animals, the origin of mankind, then civilizations, and numbers. It is important to be aware of all this.

CHAPTER 2

Fractions, percentages

2a. Fractions, rules

Time now for some more complicated mathematics, going beyond what we know about the integers. We will be talking here about the mathematics of fractions:

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”.

So, this is the definition of the rational numbers, which will take us some time, in order to properly understand. Generally speaking, the idea will be as follows:

(1) We will usually treat, based on a number of abstract results that we will prove, the rational numbers as usual numbers, that is, as integers.

(2) With the remark of course that the rational numbers are not necessarily integers. It is only that their arithmetic is quite similar to that of the integers.

As a first observation, we have inclusions as follows, coming from definitions:

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q}$$

Getting now to operations and algebra, we must first talk about addition and subtraction, and then about multiplication and division. As a first result, we have:

THEOREM 2.2. *We can add the rational numbers $r = a/b$ according to the rule*

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

and with this convention, we have the following formulae,

$$(r + s) + t = r + (s + t) \quad , \quad r + s = s + r$$

called associativity and commutativity of the addition operation.

PROOF. We can certainly define an operation as in the statement, and with this done, our operation is indeed associative, as shown by the following computation:

$$\begin{aligned}\left(\frac{a}{b} + \frac{c}{d}\right) + \frac{e}{f} &= \frac{(ad + bc)f + bde}{bdf} \\ &= \frac{adf + b(cf + de)}{bdf} \\ &= \frac{a}{b} + \left(\frac{c}{d} + \frac{e}{f}\right)\end{aligned}$$

As for the commutativity of the sum, this is clear from definitions, as shown by:

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

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

Next, we must talk about subtraction. The result here is similar, as follows:

PROPOSITION 2.3. *We can invert the rational numbers $r = a/b$ according to the rule*

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

and with this convention, we have the following formula:

$$r + (-r) = (-r) + r = 0$$

More generally, we can subtract the rationals according to the rule

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

and with this convention, all the basic formulae that we know for fractions still hold.

PROOF. We can certainly define an inversion operation as in the statement, and with this done, the first formula in the statement holds indeed, as shown by:

$$\frac{a}{b} + \left(-\frac{a}{b}\right) = \frac{a}{b} + \frac{-a}{b} = \frac{ab - ab}{b} = 0$$

As for the second formula in the statement, its proof is similar, as follows:

$$\left(-\frac{a}{b}\right) + \frac{a}{b} = \frac{-a}{b} + \frac{a}{b} = \frac{-ab + ab}{b} = 0$$

Regarding now the subtraction operation in the statement, that comes, somewhat by definition, from the following computation:

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

Finally, in what regards the last assertion, this is something rather informal, and we will leave doing some computations here as an exercise. \square

Summarizing, everything fine with the addition of rationals. Next, we must talk about multiplication. The result here is similar to Theorem 2.2, as follows:

THEOREM 2.4. *We can multiply the rational numbers $r = a/b$ according to the rule*

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

and with this convention, we have the following formulae,

$$(rs)t = r(st) \quad , \quad rs = sr$$

called associativity and commutativity of the multiplication operation.

PROOF. We can certainly define an operation as in the statement, and with this done, our operation is indeed associative, as shown by the following computation:

$$\left(\frac{a}{b} \cdot \frac{c}{d}\right) \frac{e}{f} = \frac{ace}{bdf} = \frac{a}{b} \left(\frac{c}{d} \cdot \frac{e}{f}\right)$$

As for the commutativity property, this is clear from definitions, as shown by:

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

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

Next, we must talk about division. The result here is again routine, as follows:

PROPOSITION 2.5. *We can invert the nonzero numbers $r = a/b$ according to*

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

and with this convention, we have the following formula:

$$r \cdot r^{-1} = r^{-1} \cdot r = 1$$

More generally, we can divide the rationals according to the usual rule

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

and with this convention, all the basic formulae that we know for fractions still hold.

PROOF. We can certainly define an inversion operation as in the statement, and with this done, the first formula in the statement holds indeed, as shown by:

$$\frac{a}{b} \left(\frac{a}{b}\right)^{-1} = \frac{a}{b} \cdot \frac{b}{a} = \frac{ab}{ba} = 1$$

As for the second formula in the statement, its proof is similar, as follows:

$$\left(\frac{a}{b}\right)^{-1} \frac{a}{b} = \frac{b}{a} \cdot \frac{a}{b} = \frac{ba}{ab} = 1$$

Regarding now the division operation in the statement, that comes, somewhat by definition, from the following computation:

$$\frac{a}{b} : \frac{c}{d} = \frac{a}{b} \left(\frac{c}{d} \right)^{-1} = \frac{a}{b} \cdot \frac{d}{c} = \frac{ad}{bc}$$

Finally, in what regards the last assertion, this is something rather informal, and we will leave doing some computations here as an exercise. \square

As a conclusion now, things fine with both the addition and the multiplication, and you might probably think that we are done with all this algebra. Well, you must be kidding. Many other basic algebraic things remain to be done, such as:

THEOREM 2.6. *The addition and multiplication of rationals are subject to*

$$r(s + t) = rs + rt \quad , \quad (r + s)t = rt + st$$

called distributivity formulae.

PROOF. The verification of the first formula goes as follows:

$$\begin{aligned} \frac{a}{b} \left(\frac{c}{d} + \frac{e}{f} \right) &= \frac{acf + ade}{bdf} \\ &= \frac{ac}{bd} + \frac{ae}{bf} \\ &= \frac{a}{b} \cdot \frac{c}{d} + \frac{a}{b} \cdot \frac{e}{f} \end{aligned}$$

As for the second formula, this follows from the first one, and commutativity:

$$(r + s)t = t(r + s) = tr + ts = rt + st$$

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

Summarizing, many operations that we have here, and job for us to get familiar with all this, via practice, tricks and so on. Here is my favorite trick for fractions, which is something quite trivial, but I will call this Theorem, because this is perhaps the mathematical formula that I use the most, in my complicated, daily quantum physics work:

THEOREM 2.7. *We have the following subtraction formula,*

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

valid for any $n \in \mathbb{N}$. As illustrations for this, we have

$$1 - \frac{1}{2} = \frac{1}{2} \quad , \quad \frac{1}{2} - \frac{1}{3} = \frac{1}{6} \quad , \quad \frac{1}{3} - \frac{1}{4} = \frac{1}{12} \quad , \quad \frac{1}{4} - \frac{1}{5} = \frac{1}{20} \quad \dots$$

and with the knowledge of these latter formulae being mandatory too.

PROOF. This is something trivial, but since we called our result Theorem, as mathematicians do, let us pull out now a complete proof, also as mathematicians do:

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

Thus, theorem proved, and for the particular cases at the end, I will leave it to you. The more such particular cases you know well, the better your mathematics will be. \square

At a more abstract level now, we have the following result, regarding the sums:

THEOREM 2.8. *The sum of two fractions is always of the following form,*

$$\frac{a}{b} + \frac{c}{d} = \frac{e}{[b, d]}$$

with $e \in \mathbb{Z}$ being a certain number. More generally, the sum of n fractions is of the form

$$\frac{a_1}{b_1} + \dots + \frac{a_n}{b_n} = \frac{e}{[b_1, \dots, b_n]}$$

with $e \in \mathbb{Z}$ being a certain number.

PROOF. In what regards the first assertion, we know from chapter 1 that the least common multiple $[b, d]$ appears as follows, for certain integers p, q :

$$[b, d] = bp = dq$$

But with this, we have the following computation, proving the first assertion:

$$\frac{a}{b} + \frac{c}{d} = \frac{ap}{bp} + \frac{cq}{dq} = \frac{ap}{[b, d]} + \frac{cq}{[b, d]} = \frac{ap + cq}{[b, d]}$$

As for the second assertion, its proof is similar. We know that the least common multiple $[b_1, \dots, b_n]$ appears as follows, for certain integers p_1, \dots, p_n :

$$[b_1, \dots, b_n] = b_1 p_1 = \dots = b_n p_n$$

But with these formulae in hand, we have the following computation:

$$\begin{aligned} \frac{a_1}{b_1} + \dots + \frac{a_n}{b_n} &= \frac{a_1 p_1}{b_1 p_1} + \dots + \frac{a_n p_n}{b_n p_n} \\ &= \frac{a_1 p_1}{[b_1, \dots, b_n]} + \dots + \frac{a_n p_n}{[b_1, \dots, b_n]} \\ &= \frac{a_1 p_1 + \dots + a_n p_n}{[b_1, \dots, b_n]} \end{aligned}$$

Thus, theorem proved, but as before with other such things, a lot of practice is needed, meaning working out a lot of exercises, in order to master well this method. \square

Next, going beyond what was said in the above, we must talk about the ordering of the fractions. The result here, formulated a bit informally, is as follows:

THEOREM 2.9. *We can order the positive fractions as follows,*

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

and with this, all the basic results about the ordering of numbers extend.

PROOF. This is obviously something a bit informal, but there is a reason for this, there are in fact countless things to be checked here, and my main concern as math professor is you getting a bit bored by all this, and starting to attend a chemistry class instead. This being said, as a sample verification, let us attempt to prove the following fact:

$$\frac{a}{b} < \frac{c}{d}, \frac{e}{f} < \frac{g}{h} \implies \frac{a}{b} + \frac{e}{f} < \frac{c}{d} + \frac{g}{h}$$

But this can be proved indeed, with some patience, as follows:

$$\begin{aligned} \frac{a}{b} + \frac{e}{f} < \frac{c}{d} + \frac{g}{h} &\iff \frac{af + be}{bf} < \frac{ch + dg}{dh} \\ &\iff (af + be)dh < bf(ch + dg) \\ &\iff adfh + bdeh < bcfh + bdgf \\ &\iff adfh - bcfh < bdgf - bdeh \\ &\iff (ad - bc)fh < bd(gf - eh) \end{aligned}$$

Indeed, what we have at the end does hold, because our assumptions give:

$$(ad - bc)fh < 0 < bd(gf - eh)$$

So, verification done, and for the rest, exercise of course for you to formulate a more precise version of the statement, and perform the other verifications that are needed. \square

Finally, as a last useful piece of abstract mathematics for the fractions, we have:

THEOREM 2.10. *We can talk about the integer part of fractions,*

$$\left[\frac{a}{b} \right] = c \in \mathbb{Z}$$

with $c \in \mathbb{Z}$ being by definition the unique number such that $c \leq \frac{a}{b} < c + 1$.

PROOF. This is again something a bit informal and compact, due to the above-mentioned chemistry class fears, and I will leave it to you to learn more about this, the integer part, and to learn a bit about the fractional part too, given by:

$$\left\{ \frac{a}{b} \right\} = \frac{a}{b} - \left[\frac{a}{b} \right]$$

So, exercise for you, to explore a bit all this, as to get familiar with $[.]$ and $\{.\}$. \square

Getting now a bit abstract, the basic operations on the rational numbers, namely sum, product and inversion, tell us that \mathbb{Q} is a field, in the following sense:

DEFINITION 2.11. 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 $a(b + c) = ab + ac$.

So, this is the much feared definition of the fields, and more on this later in this book. In the meantime, let us record the following result, coming from the above:

THEOREM 2.12. The rational numbers \mathbb{Q} form a field, with operations given by:

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

In fact, \mathbb{Q} is the smallest field containing \mathbb{Z} , and even the smallest field containing \mathbb{N} .

PROOF. This is more or less clear from the above, as follows:

(1) We have indeed a field, with the operations in the statement, with the verification of the various field axioms coming from our various theorems above.

(2) Also, when searching for a field containing \mathbb{N} , by looking at the equation $c = a - b$ we are led into \mathbb{Z} . But then, by looking at the equation $c = a/b$, we are led into \mathbb{Q} . \square

Still staying a bit abstract, as a first nice result about \mathbb{Q} , in relation with what we like to do the most, since the beginning of this book, namely counting, we have:

THEOREM 2.13. The field of rational numbers \mathbb{Q} is countable.

PROOF. We can count indeed the positive rationals, with some redundancies, by arranging them in a table, and snaking our way inside this table, as follows:

$$\begin{array}{cccccc}
 1/1 & \rightarrow & 1/2 & 1/3 & \rightarrow & 1/4 & 1/5 & \dots \\
 & \swarrow & & \nearrow & \swarrow & & \nearrow & \\
 2/1 & & 2/2 & & 2/3 & & 2/4 & 2/5 & \dots \\
 & \downarrow & \nearrow & & \swarrow & & \nearrow & & \\
 3/1 & & 3/2 & & 3/3 & & 3/4 & 3/5 & \dots \\
 & \swarrow & & \nearrow & & & & & \\
 4/1 & & 4/2 & & 4/3 & & 4/4 & 4/5 & \dots \\
 & \downarrow & \nearrow & & & & & & \\
 5/1 & & 5/2 & & 5/3 & & 5/4 & 5/5 & \dots \\
 & & & & & & & & \\
 \vdots & & \vdots & & \vdots & & \vdots & \vdots & \ddots
 \end{array}$$

Thus, after eliminating the redundancies, and then adding the negatives, which must be countable too, say via an alternating $+/-$ scheme, theorem proved. \square

2b. Decimals, percentages

As explained earlier, in chapter 1, and which is something key for mathematics, the best way to deal with the integers is by using their decimal form:

$$n = a_1 \dots a_k$$

The point now is that, in what regards the fractions, we can write them in decimal form too. Indeed, in the case where our fraction is of the special form $n/10$, with n as above, we can use the following convention, with a dot appearing on the right:

$$\frac{n}{10} = a_1 \dots a_{k-1}.a_k$$

Next, in the case where our fraction is of the special form $n/100$, with n as above, we can use the following convention, also with a dot appearing on the right:

$$\frac{n}{100} = a_1 \dots a_{k-2}.a_{k-1}a_k$$

And so on, and we are led in this way to the following notion:

DEFINITION 2.14. *We can write any fraction of the form $n/10^l$ in decimal form, according to the following convention:*

$$\frac{a_1 \dots a_k b_1 \dots b_l}{10^l} = a_1 \dots a_k.b_1 \dots b_l$$

Moreover, this writing is bijective, in the sense that any expression $a_1 \dots a_k.b_1 \dots b_l$ comes from such a fraction $n/10^l$, as above.

For many practical purposes, the case $l = 2$, so $10^l = 100$, and $n \in \{0, 1, 2, \dots, 100\}$, is of particular interest. In this case the corresponding fraction is called percentage, and is written in the following simplified way, with the symbol % standing for $/100$:

$$\frac{n}{100} = n\%$$

With this understood, let us discuss now what can be done with the arbitrary fractions. These can be written as well in decimal form, with the result here being as follows:

THEOREM 2.15. *We can write any rational number $r = p/q$ in decimal form, with this decimal form being finite or not,*

$$r = \pm a_1 a_2 \dots a_k . b_1 b_2 b_3 \dots$$

with the rule $999 \dots = 1$, at the end, according to the following convention:

$$r = \pm \left(a_1 a_2 \dots a_k + \frac{b_1}{10} + \frac{b_2}{100} + \frac{b_3}{1000} + \dots \right)$$

Moreover, the decimal expressions that we can obtain in this way are periodic, and with this taken into account, our writing is bijective.

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

(1) Let us first prove that any rational number $r = p/q$ can be written in decimal form, as in the statement. We can assume $r > 0$, and we want to write r as:

$$r = a_1a_2 \dots a_k.b_1b_2b_3 \dots$$

But this can be done by using the integer part. Indeed, we can first set:

$$a_1 \dots a_k = [r]$$

Next, let us look for b_1 . This must be a digit $b_1 \in \{0, \dots, 9\}$, subject to:

$$a_1 \dots a_k.b_1 \leq r < a_1 \dots a_k.(b_1 + 1)$$

According to our convention from Definition 2.14, these inequalities read:

$$\frac{a_1 \dots a_k b_1}{10} \leq r < \frac{a_1 \dots a_k b_1 + 1}{10}$$

Thus, the digit b_1 that we are looking for must appear according to:

$$a_1 \dots a_k b_1 = [10r]$$

Similarly, the next digit b_2 must appear according to the following formula:

$$a_1 \dots a_k b_1 b_2 = [100r]$$

And so on, with at step l , the needed digit b_l appearing according to:

$$a_1 \dots a_k b_1 b_2 \dots b_l = [10^l r]$$

(2) Summarizing, we have our decimal writing for r , and the question is now, what this decimal writing exactly means. But this meaning comes from the above formula that we used, in order to define b_l , which corresponds to inequalities as follows:

$$a_1 \dots a_k b_1 b_2 \dots b_l \leq 10^l r < a_1 \dots a_k b_1 b_2 \dots b_l + 1$$

Indeed, by dividing all terms by 10^l , these inequalities become as follows:

$$\frac{a_1 \dots a_k b_1 \dots b_l}{10^l} \leq r < \frac{a_1 \dots a_k b_1 \dots b_l + 1}{10^l}$$

Equivalently, we can write the above inequalities in the following way:

$$a_1 a_2 \dots a_k + \frac{b_1}{10} + \frac{b_2}{100} + \dots + \frac{b_l}{10^l} \leq r < a_1 a_2 \dots a_k + \frac{b_1}{10} + \frac{b_2}{100} + \dots + \frac{b_l}{10^l} + \frac{1}{10^l}$$

We conclude that, alternatively, and without reference to the above algorithm, the decimal writing corresponds to the infinite sum formula in the statement, namely:

$$r = a_1 a_2 \dots a_k + \frac{b_1}{10} + \frac{b_2}{100} + \frac{b_3}{1000} + \dots$$

Summarizing, main assertion of the theorem proved, and with this coming either from the above explicit algorithm involving the integer part, or just from $1/10^l \rightarrow 0$, and some

elementary analysis. All this is quite intuitive, and we will come back to this, with further details, and a generalization, in chapter 5 below, when discussing the real numbers.

(3) Getting now to bijectivity issues, one bug comes from the following equality:

$$0.999\dots = 1$$

Indeed, in case you don't believe me, let us compute the number on the left. According to our conventions for the decimal writing, this number is given by:

$$\begin{aligned} 0.999\dots &= \frac{9}{10} + \frac{9}{100} + \frac{9}{1000} + \dots \\ &= \frac{9}{10} \left(1 + \frac{1}{10} + \frac{1}{100} + \dots \right) \\ &= \frac{9}{10} \cdot \frac{1}{1 - \frac{1}{10}} \\ &= \frac{9}{10} \cdot \frac{10}{9} \\ &= 1 \end{aligned}$$

More generally now, the same computation shows that we have:

$$\pm a_1 a_2 \dots a_k . b_1 b_2 \dots b_l 999\dots = \pm a_1 a_2 \dots a_k . b_1 b_2 \dots b_l 1$$

Thus, we must make the convention in the statement, $999\dots = 1$, at the end.

(4) Still talking bijectivity, our claim now is that the decimal expressions that we can obtain in this way, from rational numbers, are precisely those which are periodic:

$$r = \pm a_1 \dots a_k . b_1 \dots b_l (c_1 \dots c_p)$$

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

$$\begin{aligned} r &= \pm \frac{1}{10^l} a_1 \dots a_k b_1 \dots b_l . c_1 \dots c_p c_1 \dots c_p \dots \\ &= \pm \frac{1}{10^l} \left(a_1 \dots a_k b_1 \dots b_l + c_1 \dots c_p \left(\frac{1}{10^p} + \frac{1}{10^{2p}} + \dots \right) \right) \\ &= \pm \frac{1}{10^l} \left(a_1 \dots a_k b_1 \dots b_l + \frac{c_1 \dots c_p}{10^p - 1} \right) \end{aligned}$$

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. \square

Getting back now to percentages, and their mathematics, which are something very useful, we can use them for approximating rational numbers $r \in [0, 1]$, as follows:

PROPOSITION 2.16. *Any rational number $r \in [0, 1]$, written in decimal form as*

$$r = 0.a_1a_2a_3 \dots$$

can be approximated by a percentage, as follows,

$$r \simeq a_1a_2\%$$

with the error term being smaller than 1%.

PROOF. This is more of an empty statement, coming by comparing our conventions for percentages and decimals, from Definition 2.14 and afterwards, and Theorem 2.15. \square

The above result is something quite intuitive, and very useful in practice. Of course, there are certain situations in life where our 1% error margin is too big, and in this case we can add decimals to our percentages, as shown by the following result:

THEOREM 2.17. *Any rational number $r \in [0, 1]$, written in decimal form as*

$$r = 0.a_1a_2a_3 \dots$$

can be approximated by a percentage, as follows,

$$r \simeq a_1a_2\%$$

$$r \simeq a_1a_2.a_3\%$$

$$r \simeq a_1a_2.a_3a_4\%$$

$$\vdots$$

with the error term being smaller than 1%, 0.1%, 0.01% and so on.

PROOF. Again, this is more of an empty statement, coming by comparing our various conventions from Definition 2.14 and afterwards, and from Theorem 2.15. \square

Finally, in relation with all this, as a question that you might have, shall the assumption $r \in [0, 1]$ bother us, or no. In answer, not at all, because we have:

FACT 2.18. *The probability for anything in this world to happen is a number*

$$r \in [0, 1]$$

and so, this probability can be approximated as a percentage, as above.

As a comment here, any probability $r \in [0, 1]$ is in fact a percentage, even without approximating it, because in agreement with Theorem 2.17, we can say that we have:

$$r = 0.a_1a_2a_3 \dots \iff r = a_1a_2.a_3 \dots \%$$

Quite nice all this, philosophically speaking, hope you agree with me. In short, when you hear economists saying that everything in life is about percentages, they are in fact right, and this even in relation with complicated questions from theoretical physics.

2c. Games, probability

As an application to what we learned so far, let us do some probability. We first have the following result, solving a well-known problem, of key importance in real life:

THEOREM 2.19. *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:

$$\{7, 8, 9, 10, J, Q, K, A\} \times \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$$

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. □

At a more advanced level now, many things can be learned by flipping coins, and recording your findings. Let us start with something very basic, as follows:

FACT 2.20. *The probability of winning when flipping a coin is $1/2$.*

Obvious you would say, but there are some subtleties here, even in this simplest possible probability question. The first thing is that I said “winning”, like everyone says when it comes to flipping coins, but winning against whom?

So, this is a first subtlety. Flipping a coin is best regarded as being a game, with you choosing between heads and tails, let us say heads, then flipping the coin, and winning if heads. But now, that we talked about a game, you need a partner for your game. That is, you are not playing a game alone, but with someone else, who wins when it's tails.

Which brings us into a second question, winning what? Many options here, like winning apples, or oranges, or luxury cars, assuming that both you and your partner have a considerable stock of those. Or why not, for making the game even more exciting, the right to slap your partner, or why not pulling a knife, and killing your partner.

So, what to choose? The answer here is money, this is what money is made for, for simplifying such things, transactions between humans. In the hope that we agree on this, and now with this discussion made, let us record our findings, as follows:

CONCLUSION 2.21. *Flipping a coin is best regarded as being a game, between you and a partner, the rules being:*

- (1) *Every time it is heads, you win \$1 from your partner.*
- (2) *Every time it is tails, your partner wins \$1 from you.*

With this conclusion recorded, we can see now more clearly what is behind coin flipping. Obviously, all sorts of interesting things that we can explore, and we will do that, and with the main question, which is surely on everyone's mind, being:

QUESTION 2.22. *Who wins?*

So, let us study now this question. Thus is a matter of understanding how the game axiomatized in Conclusion 2.21 evolves over the time, taking into account the $1/2$ mathematics from Fact 2.20, and here are a few preliminary observations, about this:

PROPOSITION 2.23. *When flipping a coin k times, the following happen,*

- (1) *The probability of you winning \$ k is $1/2^k$.*
- (2) *The probability of you winning \$ $k - 1$ is 0.*
- (3) *The probability of you winning \$ $k - 2$ is $k/2^k$.*
- (4) *The probability of you winning \$ $k - 3$ is again 0.*
- (5) *The probability of you winning \$ $k - 4$ is $k(k - 1)/2^{k+1}$.*

and so on, with the probability increasing, up to the tie situation, and then decreasing.

PROOF. This follows indeed from some simple mathematics, as follows:

(1) You winning $\$k$ means you winning every time, over k attempts, so your probability here is $P = (1/2) \times \dots \times (1/2)$, with k terms in the product, which reads $P = 1/2^k$.

(2) The point here is that you cannot win $\$k - 1$, exactly. Indeed, you must lose at least once, and so you profit will be $\leq (k - 1) - 1 = k - 2$.

(3) Here we have a similar computation as in (1). For winning $\$k - 2$ you need to lose exactly once, and since there are k possibilities of losing exactly once, $P = k/2^k$.

(4) Here the situation is similar to that in (2). Indeed, for winning exactly $\$k - 3$ you would certainly need to lose twice, so you profit will be $\leq (k - 2) - 2 = k - 4$.

(5) With the same reasoning as in (3), here you need to lose exactly twice, and since there are $k(k - 1)/2$ possibilities of losing exactly twice, $P = k(k - 1)/2^{k+1}$.

(6) Finally, regarding the last assertion, which is a bit informal, we will leave the clarification here, both statement and proof, to you, as an instructive exercise. \square

Obviously, some interesting mathematics is going on here, that needs to be better understood. We have the following result, generalizing Proposition 2.23:

THEOREM 2.24. *When flipping a coin k times what you can win are quantities of type $\$k - 2s$, with $s = 0, 1, \dots, k$, 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, by fine-tuning our various observations from Proposition 2.23 and its proof, the point here 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:

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

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

With Theorem 2.24 in hand, we are somehow done with math, and time now to turn to Question 2.22. Let us first examine a more concrete question, namely:

QUESTION 2.25. *What and how do you win, depending on your strategy?*

However, this appears to be a bit of a bad question, at least in the context of our very simple flipping game, because you have not so many options for developing a strategy. In view of this, let us reformulate our strategy question as follows:

QUESTION 2.26. *Assuming that the rules are biased, favoring you, by allowing you to pull off at any moment of your choice, what is your best strategy?*

In relation now with this, the legend has it that what you have to do is to play and play, until you reached a sum of money that you fixed as objective in advance, say \$100. Then you pull off, with the money in your pocket. Simple like that.

So, let us see how this works. To start with, this can only work, I mean just play and play, as indicated above, and you will certainly end up with \$100 in your pocket, no question about this. However, this might take some precious time t , and the mathematics, based on our formula in Theorem 2.24, shows that this time t is as follows:

Time spent playing	Probability to win
$t = 100$	$1/2^{100} = 0.000$
$t = 102$	$102/2^{102} = 0.000$
$t = 104$	$\binom{104}{2}/2^{104} = 0.000$
$t = 106$	$\binom{106}{3}/2^{106} = 0.000$
$t = 108$	$\binom{108}{4}/2^{108} = 0.000$
$t = 110$	$\binom{110}{5}/2^{110} = 0.000$
\vdots	\vdots

Which does not look very good, hope you agree with me. Obviously, we are here into some sort of very abstract math, not corresponding to anything in the real life. So, in order to reach to something more reasonable, good moment to remember that:

FACT 2.27. *Time is money.*

In view of this, let us downgrade our ambitions, and only wish to win a modest \$10. Here we reach to a more reasonable winning scheme, as follows:

Time spent playing	Probability to win
$t = 10$	$1/2^{10} = 0.001$
$t = 12$	$12/2^{12} = 0.003$
$t = 14$	$\binom{14}{2}/2^{14} = 0.006$
$t = 16$	$\binom{16}{3}/2^{16} = 0.009$
$t = 18$	$\binom{18}{4}/2^{18} = 0.012$
$t = 20$	$\binom{20}{5}/2^{20} = 0.015$
\vdots	\vdots

However, this is still not interesting, financially speaking. Not good, all this.

Moving ahead now, we talked in the above about “time is money”, which is something that must be taken into account, but thinking well, what really matters in all this is the maximum amount of money that you can afford to lose. So, let us further reformulate our strategy question, by making it more realistic, as follows:

QUESTION 2.28. *What is your best strategy, assuming that the game is asymmetric:*

- (1) *With the rules being biased, favoring you, by allowing you to pull off from the game, at any moment of your choice.*
- (2) *With the capital being unequal, favoring your partner, who has N money that he can afford to lose, compared to your $n < N$ money.*
- (3) *And perhaps with a fee for playing the game too, again favoring your partner, to be paid by you, and this because N, n are normally secret.*

And good news, this is the good, final question, which perfectly makes sense, and is fully realistic. There is some math to be done here, and getting started with this, we can solve a simple case right away, namely that when your partner has endless money:

$$N = \infty$$

A player having this feature is called “the bank”, and with this convention made, the answer to our various questions, and notably to Question 2.22 that we started with, is:

ANSWER 2.29. *The bank wins.*

To be more precise here, as already mentioned, we can certainly do some math here, and we will do this later. But, for our purposes now, the simplest is to argue that, in your situation, when you have $\$n$ and you lose $\$n$, say with $n = 1,000,000$ for having a precise figure, you are dead, say with this coming from drug overdose, after reaching the street, after your bankruptcy. So, your strategy of pulling off once you won a precise sum of money, say $\$100$, is certainly flawed, because you can meet death on the way:

Time spent playing	Probability to win	Other outcomes
$t = 100$	small	losing
$t = 102$	small	losing
$t = 104$	small	losing
\vdots	\vdots	\vdots
$t = 1,000,000$	attractive	death
$t = 1,000,002$	attractive	death
$t = 1,000,004$	attractive	death
\vdots	\vdots	\vdots

As for the variations of this strategy, these can be certainly investigated too, but it is quite clear that all this will not lead to anything good. So, we are led to Answer 2.29.

2d. Binomial laws

At a more advanced level, we can roll dice. The difference with coins comes from the fact that the basic $1/2 - 1/2$ probabilities get now replaced by a better readable $1/6 - 5/6$. To be more precise, let us first convene for the following rules for the game:

RULES 2.30. *Rolling the die is played with the following rules:*

- (1) *Every time it is 1, 2, 3, 4, 5, your partner wins \$1 from you.*
- (2) *And every time it is 6, you win \$5 from your partner.*

Of course, you might say that this is not very standard, but hey, we are just doing some math here, and we will complicate the rules later on, no worries for that. Now with these rules agreed on, we have the following analogue of Theorem 2.24:

THEOREM 2.31. *When rolling a die k times what you can win are quantities of type $\$6w - k$, with $w = 0, 1, \dots, k$, with the probability for this to happen being:*

$$P(6w - k) = \frac{5^{k-w}}{6^k} \binom{k}{w}$$

Geometrically, your winning curve starts with probability $(5/6)^k$ of losing $\$k$, then increases, up to some point, and then decreases, up to probability $1/6^k$ of winning $\$5k$.

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

(1) When rolling the die k times, what will happen is that you will win w times and lose l times, with $k = w + l$. And in this situation, your profit will be, as stated:

$$\begin{aligned} \$ &= 5w - l \\ &= 5w - (k - w) \\ &= 6w - k \end{aligned}$$

(2) As for the probability for this to happen, this is the total number of possibilities for you to win w times, which is $5^{k-w} \binom{k}{w}$, because this amounts in choosing the w times when you will win, among k , then multiplying by 5^{k-w} possibilities, at places where your partner wins, and finally dividing by the total number of possibilities, which is 6^k :

$$P(6w - k) = \frac{5^{k-w}}{6^k} \binom{k}{w}$$

We are therefore led to the various conclusions in the statement. □

Now with Theorem 2.31 in hand, it is quite clear that the basic $1/6 - 5/6$ probabilities at dice can be repaced with something of type $p - (1 - p)$, with $p \in [0, 1]$ being arbitrary. We are led in this way to the following notions, which are quite general:

DEFINITION 2.32. *Given $p \in [0, 1]$, the Bernoulli law of parameter p is given by:*

$$P(\text{win}) = p \quad , \quad P(\text{lose}) = 1 - p$$

More generally, the k -th binomial law of parameter p , with $k \in \mathbb{N}$, is given by

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

with the Bernoulli law appearing at $k = 1$, with $s = 1, 0$ here standing for win and lose.

To be more precise, what we call here “law” is something intuitive, based on what we did before with coins and dice, basically standing for “outcome of a game”. As a first observation, the Bernoulli law generalizes indeed what we did before with coins and dice, which come respectively from the following choices of the parameter $p \in [0, 1]$:

$$p_{\text{coin}} = 1/2 \quad , \quad p_{\text{die}} = 1/6$$

Observe also that the last assertion holds indeed, because at $k = 1$ the binomial law is as follows, coinciding indeed with the Bernoulli law of parameter p :

$$P(1) = p \quad , \quad P(0) = 1 - p$$

Finally, regarding the binomial law, observe that is indeed a “law”, or what we can expect from a game, because the various probabilities sum up to 1, as they should:

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

With this discussed, let us try now to understand the relation between Bernoulli and binomial laws. Indeed, we know from both coins and dice that the Bernoulli laws produce the binomial laws, simply by iterating the game, from 1 throw to $k \in \mathbb{N}$ throws.

The reasons behind this obviously come from the “independence” of our coin or dice throwings, when iterating. So, let us record this finding, as follows:

CONCLUSION 2.33. *The Bernoulli laws produce the binomial laws, by iterating the game, via the independence of the throws.*

Of course, this finding is something quite intuitive, and temporary, and it remains to work out the precise mathematics of independence, producing the explicit formula of the binomial laws, out of the explicit formula of the Bernoulli laws. In the meantime:

(1) The idea is to encapsulate the data from Definition 2.32 into so-called “probability measures” associated to the Bernoulli and binomial laws. For the Bernoulli law, the

corresponding measure is as follows, with the δ symbols standing for Dirac masses:

$$\mu_{ber} = (1 - p)\delta_0 + p\delta_1$$

As for the binomial law, here the measure is as follows, constructed in a similar way, you get the point I hope, again with the δ symbols standing for Dirac masses:

$$\mu_{bin} = \sum_{s=0}^k p^s (1 - p)^{k-s} \binom{k}{s} \delta_s$$

(2) Getting now to independence, and to our finding from Conclusion 2.33, the mathematics there is that of the following formula, with $*$ standing for the convolution operation for probability measures, which on Dirac masses is given by $\delta_x * \delta_y = \delta_{x+y}$:

$$\mu_{bin} = \underbrace{\mu_{ber} * \dots * \mu_{ber}}_{k \text{ terms}}$$

(3) To be more precise, this latter formula does hold indeed, as a straightforward application of the binomial formula, the formal proof being as follows:

$$\begin{aligned} \mu_{ber}^{*k} &= ((1 - p)\delta_0 + p\delta_1)^{*k} \\ &= \sum_{s=0}^k p^s (1 - p)^{k-s} \binom{k}{s} \delta_0^{*(k-s)} * \delta_1^{*s} \\ &= \sum_{s=0}^k p^s (1 - p)^{k-s} \binom{k}{s} \delta_s \\ &= \mu_{bin} \end{aligned}$$

(4) Summarizing, question solved, save for a bit of formal mathematics in relation with independence, Dirac masses and convolution, that we will do in a moment.

All this is very nice, and is perhaps worth a reformulation of Conclusion 2.33. We reach in this way to a quite drastic statement, as follows:

CONCLUSION 2.34. *Most of what we did with coins and dice reduces to the formula*

$$\mu_{ber}^{*k} = \mu_{bin}$$

relating the Bernoulli and binomial laws, via the convolution operation $$.*

And isn't this magic. We have proof here for the abstract power of mathematics. Or perhaps of physics, because the Dirac masses, involved in all this, come from Dirac.

Getting to formal mathematical work now, in relation with this, let us start with:

DEFINITION 2.35. *Given a set X , which can be finite, countable, or even uncountable, a discrete probability measure on it is a linear combination as follows,*

$$\mu = \sum_{x \in X} \lambda_x \delta_x$$

with $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. Given a function $f : X \rightarrow \mathbb{Q}$ we set

$$\int_X f(x) d\mu(x) = \sum_{x \in X} \lambda_x f(x)$$

with the convention that each Dirac mass integrates up to 1.

Observe that, with this, we are now into pure mathematics. However, and we insist on this, it is basic probability which is behind all this. Next, we have:

DEFINITION 2.36. *A random variable on a probability space X is a function*

$$f : X \rightarrow \mathbb{Q}$$

and the expectation of such a random variable is the quantity

$$E(f) = \sum_{x \in X} f(x) P(x)$$

which is best thought as being the average gain, when the game is played.

As a basic illustration for this, in case we are dealing with a usual die, what we win is what the die says, and on average, what we win is the following quantity:

$$E = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

Along the same lines, at a more advanced level, we can talk as well about variance. Imagine for instance that you want to set up some sort of business, with your random variable $f : X \rightarrow \mathbb{Q}$. You are of course mostly interested in the expectation $E(f) \in \mathbb{Q}$, but passed that, the way this expectation comes in matters too. For instance:

(1) When your variable is constant, $f = c$, you certainly have $E(f) = c$, and your business will run smoothly, with not so many surprises on the way.

(2) On the opposite, for a complicated variable satisfying $E(f) = c$, your business will be more bumpy, with wins or loses on the way, depending on your skills.

In short, and extrapolating now from business to mathematics, physics, chemistry and everything else, we must complement Definition 2.36 with something finer, regarding the “quality” of the expectation $E(f) \in \mathbb{Q}$ appearing there. And the first thought here, which is the correct one, goes to the following number, called variance of our variable:

$$V(f) = E((f - E(f))^2) = E(f^2) - E(f)^2$$

However, let us not stop here. For a total control of your business, be that of financial, mathematical, physical or chemical type, you will certainly want to know more about your variable $f : X \rightarrow \mathbb{Q}$. Which leads us into general moments, constructed as follows:

DEFINITION 2.37. *The moments of a variable $f : X \rightarrow \mathbb{Q}$ are the numbers*

$$M_k = E(f^k)$$

which satisfy $M_0 = 1$, then $M_1 = E(f)$, and then $V(f) = M_2 - M_1^2$.

And with this, good news, we have now all the needed tools in our bag for doing some good business. To put things in a very compacted way, M_0 is about foundations, M_1 is about running some business, M_2 is about running that business well, and M_3 and higher are advanced level, about ruining all the competing businesses.

As a further piece of basic probability, coming this time as a theorem, we have:

THEOREM 2.38. *Given a random variable $f : X \rightarrow \mathbb{Q}$, if we define its law as being*

$$\mu = \sum_{x \in X} P(x) \delta_{f(x)}$$

regarded as probability measure on \mathbb{Q} , then the moments are given by the formula

$$E(f^k) = \int_{\mathbb{Q}} y^k d\mu(y)$$

with the usual convention that each Dirac mass integrates up to 1.

PROOF. To start with, given a random variable $f : X \rightarrow \mathbb{Q}$, we can certainly talk about its law μ , as being the formal linear combination of Dirac masses in the statement. Also talking basics, let us record as well the following alternative formula for the law, which is clear from definitions, and that we will often use, in what follows:

$$\mu = \sum_{y \in \mathbb{Q}} P(f = y) \delta_y$$

Now let us compute the moments of f . With the usual convention that each Dirac mass integrates up to 1, as mentioned in the statement, we have:

$$\begin{aligned} E(f^k) &= \sum_{x \in X} P(x) f(x)^k \\ &= \sum_{y \in \mathbb{Q}} y^k \sum_{f(x)=y} P(x) \\ &= \int_{\mathbb{Q}} y^k d\mu(y) \end{aligned}$$

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

Next, we have the following definition, inspired by what happens for coins, dice and cards, as explored above, and which is the foundation for everything advanced:

DEFINITION 2.39. *We say that two variables $f, g : X \rightarrow \mathbb{Q}$ are independent when*

$$P(f = x, g = y) = P(f = x)P(g = y)$$

happens, for any $x, y \in \mathbb{Q}$.

As already mentioned, this is something very intuitive, inspired by what happens for coins, dice and cards. As a first result now regarding independence, we have:

THEOREM 2.40. *Assuming that $f, g : X \rightarrow \mathbb{Q}$ are independent, we have:*

$$E(fg) = E(f)E(g)$$

More generally, we have the following formula, for the mixed moments,

$$E(f^k g^l) = E(f^k)E(g^l)$$

and the converse holds, in the sense that this formula implies the independence of f, g .

PROOF. We have indeed the following computation, using the independence of f, g :

$$\begin{aligned} E(f^k g^l) &= \sum_{xy} x^k y^l P(f = x, g = y) \\ &= \sum_{xy} x^k y^l P(f = x)P(g = y) \\ &= \sum_x x^k P(f = x) \sum_y y^l P(g = y) \\ &= E(f^k)E(g^l) \end{aligned}$$

As for the last assertion, this is clear too, because having the above computation work, for any $k, l \in \mathbb{N}$, amounts in saying that the independence formula for f, g holds. \square

Regarding now the convolution operation, motivated and inspired by what we found before, in Conclusion 2.34, we have the following general result:

THEOREM 2.41. *Assuming that $f, g : X \rightarrow \mathbb{Q}$ are independent, we have*

$$\mu_{f+g} = \mu_f * \mu_g$$

with the $$ operation being given by $\delta_x * \delta_y = \delta_{x+y}$. The converse of this holds too.*

PROOF. We have indeed the following straightforward computation:

$$\begin{aligned}
 \mu_{f+g} &= \sum_{x \in \mathbb{Q}} P(f+g=x) \delta_x \\
 &= \sum_{y,z \in \mathbb{Q}} P(f=y, g=z) \delta_{y+z} \\
 &= \sum_{y,z \in \mathbb{Q}} P(f=y) P(g=z) \delta_y * \delta_z \\
 &= \left(\sum_{y \in \mathbb{Q}} P(f=y) \delta_y \right) * \left(\sum_{z \in \mathbb{Q}} P(g=z) \delta_z \right) \\
 &= \mu_f * \mu_g
 \end{aligned}$$

As for the converse, this comes too from the above computation. \square

Good news, with all this understood, we can now get back to what we found in Conclusion 2.34, and formulate a precise theorem about that, as follows:

THEOREM 2.42. *The following happen, in the context of the biased coin game:*

- (1) *The Bernoulli laws μ_{ber} produce the binomial laws μ_{bin} , by iterating the game $k \in \mathbb{N}$ times, via the independence of the throws.*
- (2) *We have in fact $\mu_{bin} = \mu_{ber}^{*k}$, with $*$ being the convolution operation for real probability measures, given by $\delta_x * \delta_y = \delta_{x+y}$, and linearity.*

PROOF. This is something that we knew from before, from Conclusion 2.34, in a vague form, and which is now clear indeed, by using Theorem 2.41. \square

2e. Exercises

This was yet another very standard chapter, and as exercises on this, we have:

EXERCISE 2.43. *How to avoid redundancies, in our method for counting \mathbb{Q}_+ ?*

EXERCISE 2.44. *Review if needed the classical algorithm for dividing integers.*

EXERCISE 2.45. *Learn how computers deal, or not, with the 999... issue.*

EXERCISE 2.46. *Does mathematics simplify with \mathbb{Q} introduced via decimal form?*

EXERCISE 2.47. *Play some poker, matter of verifying the figures computed above.*

EXERCISE 2.48. *Learn more about the binomial laws, and their various properties.*

EXERCISE 2.49. *Mediate about the variance V , and its exact probabilistic meaning.*

EXERCISE 2.50. *Learn more about Dirac masses, from mathematicians and physicists.*

As bonus exercise, find and start reading a nice book on probability.

CHAPTER 3

Groups, algebra

3a. Groups, examples

Getting now to the algebraic aspects of numbers, we would like to talk about groups, fields and other algebraic beasts, which are all intimately related to numbers. Let us first talk about groups. Their definition is something very simple, as follows:

DEFINITION 3.1. *A group is a set G endowed with a multiplication operation*

$$(g, h) \rightarrow gh$$

which must satisfy the following conditions:

- (1) *Associativity: we have, $(gh)k = g(hk)$, for any $g, h, k \in G$.*
- (2) *Unit: there is an element $1 \in G$ such that $g1 = 1g = g$, for any $g \in G$.*
- (3) *Inverses: for any $g \in G$ there is $g^{-1} \in G$ such that $gg^{-1} = g^{-1}g = 1$.*

The multiplication law is not necessarily commutative. In the case where it is, in the sense that $gh = hg$, for any $g, h \in G$, we call G abelian, en hommage to Abel, and we usually denote its multiplication, unit and inverse operation as follows:

$$(g, h) \rightarrow g + h \quad , \quad 0 \in G \quad , \quad g \rightarrow -g$$

However, this is not a general rule, and rather the converse is true, in the sense that if a group is denoted as above, this means that the group must be abelian.

There are many examples of groups, with typically all the basic systems of numbers that we know being groups. Here are some standard illustrations, for this fact:

PROPOSITION 3.2. *We have the following groups, and non-groups:*

- (1) $(\mathbb{Z}, +)$ *is a group.*
- (2) $(\mathbb{Q}, +)$ *is a group too.*
- (3) $(\mathbb{N}, +)$ *is not a group.*
- (4) (\mathbb{Q}^*, \cdot) *is a group.*
- (5) (\mathbb{N}^*, \cdot) , (\mathbb{Z}^*, \cdot) *are not groups.*

PROOF. All this is clear from the definition of the groups, as follows:

- (1) The group axioms are indeed satisfied for $(\mathbb{Z}, +)$, with the group operation being the sum $(g, h) \rightarrow g + h$, the unit element being 0, and the inverse map being $g \rightarrow -g$.

(2) Once again, the group axioms are satisfied for $(\mathbb{Q}, +)$, for the same reasons as for $(\mathbb{Z}, +)$, and with the remark that we are using here the following formula:

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

(3) In $(\mathbb{N}, +)$ the first two group axioms are satisfied, for the same reasons as for $(\mathbb{Z}, +)$. However, we do not have inverses, so we do not have a group:

$$-1 \notin \mathbb{N}$$

As a comment here, such beasts, which satisfy the first two group axioms, but not the third one, are called semigroups. Thus, $(\mathbb{N}, +)$ is a semigroup, which is not a group.

(4) The group axioms are indeed satisfied for (\mathbb{Q}^*, \cdot) , with the product gh being the usual product, 1 being the usual 1, and g^{-1} being the usual g^{-1} . Observe that we must remove indeed the element $0 \in \mathbb{Q}$, because in a group, any element must be invertible:

$$0 \notin \mathbb{Q}^*$$

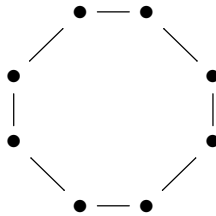
(5) In what regards (\mathbb{N}^*, \cdot) , (\mathbb{Z}^*, \cdot) , here the first two group axioms are satisfied, but not the third one, for instance due to the fact that the element 2 has no inverse:

$$\frac{1}{2} \notin \mathbb{Z}^*$$

Thus, both (\mathbb{N}^*, \cdot) , (\mathbb{Z}^*, \cdot) are semigroups, which are not groups. □

Getting now to the finite group case, which is of particular interest in relation with numbers, as a basic example here we have the cyclic group \mathbb{Z}_N , constructed as follows:

DEFINITION 3.3. *The cyclic group \mathbb{Z}_N is the group formed by the N rotations of the regular N -gon, with the group operation being the composition of these rotations:*



Alternatively, $\mathbb{Z}_N = \{0, 1, 2, \dots, N-1\}$ is the group of remainders modulo N , with the usual addition operation for such remainders.

Here the fact that the above two definitions of \mathbb{Z}_N are indeed equivalent comes from the fact that, with the first approach, if we set $\mathbb{Z}_N = \{R_0, R_1, R_2, \dots, R_{N-1}\}$, with $R_0 = id$, and with R_1, R_2, \dots being the other rotations, listed in increasing counterclockwise order, the group law is given by $R_k R_l = R_{k+l}$, with $k+l$ taken modulo N . Thus, we have $\mathbb{Z}_N = \{0, 1, 2, \dots, N-1\}$, with the group operation being $(k, l) \rightarrow k+l$, modulo N .

As a first observation, the above cyclic groups \mathbb{Z}_N are all abelian. We can construct further abelian groups by taking products of such cyclic groups, as follows:

PROPOSITION 3.4. *The following groups are all finite, and abelian,*

$$G = \mathbb{Z}_{N_1} \times \dots \times \mathbb{Z}_{N_k}$$

for any choice of the numbers $N_1, \dots, N_k \in \mathbb{N}$.

PROOF. This is something trivial, with the obvious definition for \times , coming from the fact that a product of abelian groups must be abelian too. We will see later in this chapter that any finite abelian group must appear as above, as a product of cyclic groups. \square

Let us examine a bit the groups appearing in Proposition 3.4. In the simplest case, that of a product of two cyclic groups, we have the following useful result:

THEOREM 3.5. *Given two integers satisfying $(M, N) = 1$, we have:*

$$\mathbb{Z}_M \times \mathbb{Z}_N = \mathbb{Z}_{MN}$$

In the case $(M, N) > 1$ this fails, and the group on the left is not cyclic.

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

(1) In order to establish the identification in the statement, consider the following map, which is obviously well-defined, and is a group morphism, for any $M, N \in \mathbb{N}$:

$$f : \mathbb{Z}_M \times \mathbb{Z}_N \rightarrow \mathbb{Z}_{MN} \quad , \quad f(a, b) = Na + Mb$$

In order to prove that f is bijective, when $(M, N) = 1$, we can invoke a standard result from chapter 1, stating that for $(M, N) = 1$, we can always find $p, q \in \mathbb{Z}$ such that:

$$Mp + Nq = 1$$

Indeed, this shows that we have $f(q, p) = 1$, and by further multiplying everything by a given $c \in \mathbb{Z}_{MN}$, taken arbitrary, we can have $c \in \text{Im}(f)$, as follows:

$$f(cq, cp) = Ncq + Mcp = (Nq + Mp)c = c$$

Thus f is surjective, and with the cardinalities matching, f is bijective, as desired.

(2) For the second assertion, assume that we have an isomorphism as follows:

$$g : \mathbb{Z}_{MN} \rightarrow \mathbb{Z}_M \times \mathbb{Z}_N \quad , \quad g(1) = (a, b)$$

Now if we set $K = [M, N]$, then we have, with K terms in the sum:

$$g(K) = (a, b) + \dots + (a, b) = (Ka, Kb) = (0, 0)$$

Thus by injectivity we must have $K = MN$, and so $(M, N) = 1$, as claimed. \square

As a second basic example of a finite group, we have the symmetric group S_N :

THEOREM 3.6. *The permutations of $\{1, \dots, N\}$, which are the bijections*

$$\sigma : \{1, \dots, N\} \rightarrow \{1, \dots, N\}$$

form a group S_N , called symmetric group, having $N!$ elements.

PROOF. It is indeed clear from definitions that S_N is a group. Regarding now the element count, in order to construct a permutation $\sigma \in S_N$, we have:

- N choices for the value of $\sigma(N)$.
- $(N - 1)$ choices for the value of $\sigma(N - 1)$.
- $(N - 2)$ choices for the value of $\sigma(N - 2)$.
- \vdots
- and so on, up to 1 choice for the value of $\sigma(1)$.

Thus, we have indeed $N!$ choices for a permutation $\sigma \in S_N$, as claimed. \square

At the level of the general theory now, by using the symmetric groups, we have the following fundamental result regarding the finite groups, due to Cayley:

THEOREM 3.7. *Given a finite group G , we have an embedding as follows,*

$$G \subset S_N \quad , \quad g \rightarrow (h \rightarrow gh)$$

with $N = |G|$. Thus, any finite group is a permutation group.

PROOF. Given a group element $g \in G$, we can associate to it the following map:

$$\sigma_g : G \rightarrow G \quad , \quad h \rightarrow gh$$

Since $gh = gh'$ implies $h = h'$, this map is bijective, and so is a permutation of G , viewed as a set. Thus, with $N = |G|$, we can view this map as a usual permutation, $\sigma_g \in S_N$. Summarizing, we have constructed so far a map as follows:

$$G \rightarrow S_N \quad , \quad g \rightarrow \sigma_g$$

Our first claim is that this is a group morphism. Indeed, this follows from:

$$\sigma_g \sigma_h(k) = \sigma_g(hk) = ghk = \sigma_{gh}(k)$$

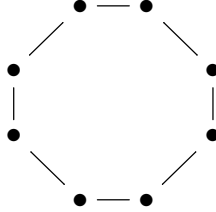
It remains to prove that this group morphism is injective. But this follows from:

$$\begin{aligned} g \neq h &\implies \sigma_g(1) \neq \sigma_h(1) \\ &\implies \sigma_g \neq \sigma_h \end{aligned}$$

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

Moving on, as a third basic example of finite group, lying in complexity somewhere between \mathbb{Z}_N and S_N , we have the dihedral group D_N , which appears as follows:

DEFINITION 3.8. *The dihedral group D_N is the symmetry group of*



that is, of the regular polygon having N vertices.

Many interesting things can be said about D_N . To start with, we have $D_2 = \mathbb{Z}_2$, and $D_3 = S_3$. In general, the dihedral group D_N has $2N$ elements, as follows:

– First we have the N rotations $R_1, \dots, R_N \in \mathbb{Z}_N$, with R_k being the rotation of angle $360^\circ k/N$. When labeling the vertices of the N -gon $1, \dots, N$ we have $R_k : i \rightarrow k + i$.

– Then we have N symmetries S_1, \dots, S_N , with S_k being the symmetry with respect to the Ox axis rotated by $180^\circ k/N$. The symmetry formula is $S_k : i \rightarrow k - i$.

Now let us see how these rotations and symmetries multiply. We have:

$$R_k R_l : i \rightarrow l + i \rightarrow k + l + i$$

$$R_k S_l : i \rightarrow l - i \rightarrow k + l - i$$

$$S_k R_l : i \rightarrow l + i \rightarrow k - l - i$$

$$S_k S_l : i \rightarrow l - i \rightarrow k - l + i$$

We conclude that we can talk about D_N abstractly, if we want to, as follows:

PROPOSITION 3.9. *D_N is the group formed by R_1, \dots, R_N and S_1, \dots, S_N , with*

$$R_k R_l = R_{k+l} \quad , \quad R_k S_l = S_{k+l}$$

$$S_k R_l = S_{k-l} \quad , \quad S_k S_l = R_{k-l}$$

being the multiplication formulae for these group elements.

PROOF. This follows indeed from the above discussion. □

As a continuation of this, observe that $\mathbb{Z}_N \times \mathbb{Z}_2$ is the group having $2N$ elements, r_1, \dots, r_N and s_1, \dots, s_N , which multiply according to the following rules:

$$r_k r_l = r_{k+l} \quad , \quad r_k s_l = s_{k+l}$$

$$s_k r_l = s_{k+l} \quad , \quad s_k s_l = r_{k+l}$$

We conclude that D_N must appear as some sort of “twisted version” of $\mathbb{Z}_N \times \mathbb{Z}_2$. In order to have a theorem, stating this, let us formulate the following definition:

DEFINITION 3.10. *Given groups H, K , with an action $K \curvearrowright H$, the crossed product*

$$G = H \rtimes K$$

is the set $H \times K$, with multiplication $(g, s)(h, t) = (gh^s, st)$.

It is routine to check that G is indeed a group. Observe that when the action is trivial, $h^s = h$ for any $h \in H$ and $s \in K$, we obtain the usual product $H \times K$. Now with this technology in hand, we can go back to the dihedral group D_N , and we have:

THEOREM 3.11. *We have a crossed product decomposition as follows,*

$$D_N = \mathbb{Z}_N \rtimes \mathbb{Z}_2$$

with $\mathbb{Z}_2 = \{1, \tau\}$ acting on \mathbb{Z}_N via switching signs, $k^\tau = -k$.

PROOF. We have an action $\mathbb{Z}_2 \curvearrowright \mathbb{Z}_N$ given by the formula in the statement, namely $k^\tau = -k$, so we can consider the corresponding crossed product group:

$$E_N = \mathbb{Z}_N \rtimes \mathbb{Z}_2$$

In order to understand the structure of E_N , observe that, according to Definition 3.10, the elements of E_N can be labeled ρ_1, \dots, ρ_N and $\sigma_1, \dots, \sigma_N$, as follows:

$$\rho_k = (k, 1) \quad , \quad \sigma_k = (k, \tau)$$

Now when computing the products of such elements, we obtain:

$$\rho_k \rho_l = \rho_{k+l} \quad , \quad \rho_k \sigma_l = \sigma_{k+l}$$

$$\sigma_k \rho_l = \sigma_{k+l} \quad , \quad \sigma_k \sigma_l = \rho_{k+l}$$

But these are exactly the multiplication formulae for D_N , from Proposition 3.9. Thus, we have an isomorphism $D_N \simeq E_N$ given by $R_k \rightarrow \rho_k$ and $S_k \rightarrow \sigma_k$, as desired. \square

3b. General results

Getting now to some general theory, for the finite groups, many interesting things can be said. We first have the following key result, due to Lagrange:

THEOREM 3.12 (Lagrange). *Given a group G and a subgroup $H \subset G$, the sets*

$$G/H = \{gH \mid g \in G\} \quad , \quad H \backslash G = \{Hg \mid g \in G\}$$

consist of partitions of G into subsets of size $|H|$, and we have the following formula:

$$|G/H| = |H \backslash G| = \frac{|G|}{|H|}$$

In particular, the order of the subgroup divides the order of the group, $|H| \mid |G|$.

PROOF. The partition claim for the set G/H constructed in the statement can be deduced as follows, and the proof for $H\backslash G$ is similar:

$$gH \cap kH \neq \emptyset \iff g^{-1}k \in H \iff gH = kH$$

But with this in hand, the cardinality formulae are all clear. \square

As a continuation of the above, which is something fundamental, we have:

THEOREM 3.13. *Given a subgroup $H \subset G$ which is normal, in the sense that*

$$gH = Hg \quad , \quad \forall g \in G$$

the space $G/H = H\backslash G$ is a group, with multiplication $(gH)(kH) = gkH$.

PROOF. We must first prove that the group operation in the statement is indeed well-defined. So, assume that we have group elements g, k, s, t such that:

$$gH = sH \quad , \quad kH = tH$$

We have then the following computation, as desired, using the normality condition:

$$gkH = gtH = gHt = sHt = stH$$

Regarding now the proof of associativity of our operation, this goes as follows:

$$\begin{aligned} [(gH)(hH)](kH) &= (ghH)(kH) \\ &= ghkH \\ &= (gH)(hkH) \\ &= (gH)[(hH)(kH)] \end{aligned}$$

Regarding now the unit and inverses, our claim is that these are as follows:

$$1 = H \quad , \quad (gH)^{-1} = g^{-1}H$$

Indeed, in what regards the unit, this comes from the following computation, which uses the normality condition $gH = Hg$ in order to compute $H(gH)$:

$$(gH)H = H(gH) = gH$$

As for the formula for inverses, the verification here goes as follows, once again by using the normality condition, which produces $gg^{-1} = g^{-1}g = 1$ quantities:

$$(gH)(g^{-1}H) = (g^{-1}H)(gH) = H$$

Thus, we have our abstract quotient group G/H , as desired. \square

Next in line, we have the following result, which is something very useful:

THEOREM 3.14. *Given an arbitrary morphism of groups $f : G \rightarrow K$, its kernel*

$$\ker f = \left\{ g \in G \mid f(g) = 1 \right\}$$

is a normal subgroup of G , and we have an isomorphism $G/\ker f = \text{Im}(f)$.

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

(1) The fact that $\ker f$ is stable under multiplication comes from:

$$\begin{aligned} g, h \in \ker f &\implies f(g) = 1, f(h) = 1 \\ &\implies f(gh) = 1 \\ &\implies gh \in \ker f \end{aligned}$$

Next, the fact that $\ker f$ contains the unit is trivial, coming from:

$$f(1) = 1 \implies 1 \in \ker f$$

Also, the fact that $\ker f$ is stable under taking inverses comes from:

$$g \in \ker f \implies f(g) = 1 \implies f(g^{-1}) = 1 \implies g^{-1} \in \ker f$$

And finally, the kernel of f is indeed a normal subgroup of G , as shown by:

$$h \in \ker f \implies f(h) = 1 \implies f(ghg^{-1}) = 1 \implies ghg^{-1} \in \ker f$$

(2) Getting now to the last assertion, by restricting the image we can assume that we are dealing with a surjective morphism $f : G \rightarrow K$, and we want to prove that:

$$G/\ker f = K$$

But this is something straightforward, with the isomorphism being as follows:

$$\varphi : g \cdot \ker f \rightarrow f(g)$$

(3) To be more precise, this latter map is well-defined, with this coming from:

$$g \cdot \ker f = h \cdot \ker f \implies g^{-1}h \in \ker f \implies f(g) = f(h)$$

Next, this map is a group morphism. Indeed, this can be checked as follows:

$$\begin{aligned} \varphi((g \cdot \ker f)(h \cdot \ker f)) &= \varphi((gh) \ker f) \\ &= f(gh) \\ &= f(g)f(h) \\ &= \varphi(g \cdot \ker f)\varphi(h \cdot \ker f) \end{aligned}$$

This map is injective. Indeed, its kernel can be computed as follows:

$$\begin{aligned} \ker(\varphi) &= \left\{ g \cdot \ker f \mid f(g) = 0 \right\} \\ &= \left\{ g \cdot \ker f \mid g \in \ker f \right\} \\ &= \ker f \ker f \\ &= \ker f \\ &= 0 \end{aligned}$$

Finally, our map is obviously surjective, and this finishes the proof. \square

With this discussed, time now for some examples. In relation with the dihedral groups, which are basically the only non-abelian groups that we know well, we have:

PROPOSITION 3.15. *The following happen, in relation with $D_N = \mathbb{Z}_N \rtimes \mathbb{Z}_2$:*

- (1) $\mathbb{Z}_N \subset D_N$ is normal, and $D_N/\mathbb{Z}_N = \mathbb{Z}_2$.
- (2) The various copies $\mathbb{Z}_2 \subset D_N$ are in general not normal,
- (3) Save for $\mathbb{Z}_2 = \{1, R_{N/2}\}$ when N is even, for which $D_N/\mathbb{Z}_2 = D_{N/2}$.

PROOF. Let us recall that D_N is the group having $2N$ elements, R_1, \dots, R_N and S_1, \dots, S_N , which multiply as follows, with all the indices being modulo N :

$$\begin{aligned} R_k R_l &= R_{k+l} \quad , \quad R_k S_l = S_{k+l} \\ S_k R_l &= S_{k-l} \quad , \quad S_k S_l = R_{k-l} \end{aligned}$$

(1) The subgroup $\mathbb{Z}_N = \{R_1, \dots, R_N\}$ is indeed normal, because the conjugates of rotations by either symmetries or rotations are again rotations, as shown by:

$$\begin{aligned} S_k R_l S_k^{-1} &= S_{k-l} S_k = R_{-l} \\ R_k R_l R_k^{-1} &= R_{k+l} R_{-l} = R_l \end{aligned}$$

As for the quotient D_N/\mathbb{Z}_N , this having 2 elements, it must be \mathbb{Z}_2 .

(2) We have N copies $\mathbb{Z}_2 \subset D_N$, namely $\mathbb{Z}_2 = \{1, S_k\}$, with S_k being the N symmetries, and none of these is normal, at $N > 2$, due to the following conjugation formulae:

$$\begin{aligned} S_l S_k S_l^{-1} &= R_{l-k} S_l = S_{2l-k} \\ R_l S_k R_l^{-1} &= S_{l+k} R_{-l} = S_{2l+k} \end{aligned}$$

(3) Thus, we are left with examining $\mathbb{Z}_2 = \{1, R_{N/2}\}$ when N is even, and with this covering the case $N = 2$ too. And the subgroup $\mathbb{Z}_2 = \{1, R_{N/2}\}$ is indeed normal, since $R_{N/2}$ commutes with everything, and the quotient is easily seen to be $D_N/\mathbb{Z}_2 = D_{N/2}$. \square

The above result is quite interesting, and we have the following extension of it:

PROPOSITION 3.16. *The following happen, regarding an arbitrary group G :*

- (1) Any index 2 subgroup $H \subset G$ is normal, and $G/H = \mathbb{Z}_2$.
- (2) A subgroup $\mathbb{Z}_2 \subset G$ is normal when it is central, $\mathbb{Z}_2 \subset Z(G)$.

PROOF. This is something a bit more abstract, the idea being as follows:

(1) Assuming that $H \subset G$ has index $[G : H] = 2$, meaning $|G| = 2|H|$, let us try to prove $gH = Hg$. But for $g \in H$ this is clear, and for $g \notin H$ this is clear too, due to:

$$gH = Hg = G - H$$

As for the quotient G/H , this having 2 elements, it must be \mathbb{Z}_2 .

(2) With $\mathbb{Z}_2 = \{1, h\}$, the normality condition on $\mathbb{Z}_2 \subset G$ reads $ghg^{-1} = h$, so $gh = hg$, for any $g \in G$. Thus, h must belong to the center $Z(G) = \{h \in G | gh = hg\}$. \square

Moving on, as another continuation of Theorem 3.12, which is something fundamental too, we can talk about the order of group elements, inside any finite group, as follows:

THEOREM 3.17. *Given a finite group G , any $g \in G$ generates a cyclic subgroup*

$$\langle g \rangle = \{1, g, g^2, \dots, g^{k-1}\}$$

with $k = \text{ord}(g)$ being the smallest number $k \in \mathbb{N}$ satisfying $g^k = 1$. Also, we have

$$\text{ord}(g) \mid |G|$$

that is, the order of any group element divides the order of the group.

PROOF. In order to prove the first assertion, let $g \in G$, and consider the semigroup $\langle g \rangle \subset G$ formed by the sequence of powers of g :

$$\langle g \rangle = \{1, g, g^2, g^3, \dots\} \subset G$$

Since G was assumed to be finite, the sequence of powers must cycle, $g^n = g^m$ for some $n < m$. But this shows that $g^k = 1$, with $k = m - n$, which gives:

$$\langle g \rangle = \{1, g, g^2, \dots, g^{k-1}\}$$

Moreover, we can choose the number $k \in \mathbb{N}$ to be minimal with this property, and with this choice, we have a set without repetitions. Thus $\langle g \rangle \subset G$ is indeed a group, and more specifically a cyclic group, whose order is as follows:

$$|\langle g \rangle| = k = \text{ord}(g)$$

Thus, we proved the first assertion, and with this in hand, the second assertion, namely $\text{ord}(g) \mid |G|$, follows from Theorem 3.12, applied to the subgroup $\langle g \rangle \subset G$. \square

As an interesting application of Theorem 3.17, in relation with numbers, we have:

THEOREM 3.18. *We have the following congruence, for any prime p ,*

$$a^p \equiv a \pmod{p}$$

called Fermat's little theorem.

PROOF. There is a simple and a complicated proof for this, as follows:

(1) The complicated proof, that we know since chapter 1, is as follows:

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

(2) As for the simple proof, we can argue here that (\mathbb{Z}_p^*, \cdot) being obviously a group, having $p-1$ elements, its elements must satisfy $a^{p-1} = 1$. Just like that. \square

3c. Abelian groups

Time now for a true, tough theorem? We know since Proposition 3.4 that the products of cyclic groups are abelian, and since it looks impossible to come up with finite abelian groups G which are not of this form, as a challenge, we would like to prove that:

$$G = \mathbb{Z}_{N_1} \times \dots \times \mathbb{Z}_{N_k}$$

However, this does not look obvious at all. So, in the lack of a bright idea, we will have to ask the cat, who's quite good at both algebra and physics, not to talk of course about analysis, differential equations, and catching mice. And cat says:

CAT 3.19. *Difficult things in arithmetic can be proved by localizing them at primes $p \in \mathbb{N}$. As for the needed lemmas, make groups act, this is what they are made for.*

Humm, this sounds terribly advanced, guess one of these days I should start trading my usual tomatoes and potatoes for raw mice and birds. This being said, there is a suggestion at the end to make the groups act, which is something understandable, and which sounds quite exciting. So, getting now into this, group actions, we have here:

PROPOSITION 3.20. *Given an action $G \curvearrowright X$ and a point $x \in X$, we have*

$$|G(x)| = |G|/|G_x|$$

where $G_x = \{g \in G | g(x) = x\}$. In particular, the cardinality of orbits divides $|G|$.

PROOF. In order to prove this, we will construct a bijection, as follows:

$$\varphi : G/G_x \rightarrow G(x)$$

But the formula of φ can only be something straightforward, as follows:

$$\varphi(gG_x) = g(x)$$

So, let us see if this works. To start with, φ is well-defined and injective, due to:

$$\begin{aligned} gG_x = hG_x &\iff g^{-1}h \in G_x \\ &\iff g^{-1}h(x) = x \\ &\iff g(x) = h(x) \end{aligned}$$

But φ is clearly surjective too, and we therefore obtain the result. □

As an application of the above technology, we have the following key result:

THEOREM 3.21 (Cauchy). *Given a finite group G , and a prime number satisfying*

$$p \mid |G|$$

G has an element of order p . Equivalently, G has a subgroup of order p .

PROOF. We must find $g \neq 1$ with $g^p = 1$. In order to do so, let us set:

$$X = \left\{ (g_1, \dots, g_p) \in G^p \mid g_1 \dots g_p = 1 \right\}$$

We have then an obvious action $\mathbb{Z}_p \curvearrowright X$, by rotation, as follows:

$$k(g_1, \dots, g_p) = (g_{k+1}, \dots, g_{k+p})$$

Now let us decompose X into orbits. This gives the following formula, with $F \subset X$ being the fixed points, and with the sum being over the non-trivial orbits O :

$$|X| = |F| + \sum_{|O| \geq 2} |O|$$

Next, let us look at this equality modulo p . To start with, we have:

$$|X| = |G|^{p-1} = 0(p)$$

Also, in what regards the fixed points, we can say here that we have:

$$(1, \dots, 1) \in F \implies |F| \geq 1$$

Finally, by Proposition 3.20 the size of any orbit must divide $|\mathbb{Z}_p| = p$, and so:

$$|O| \geq 2 \implies |O| = p$$

Now by putting everything together, modulo our $p \geq 2$, we conclude that:

$$|F| \geq 2$$

But this is exactly what we need, because the fixed points are precisely the elements $(g, \dots, g) \in G^p$ with $g^p = 1$. Thus, we have found $g \neq 1$ with $g^p = 1$, as desired. \square

Let us go back now to the finite abelian groups, with the aim of proving that these are exactly the products of cyclic groups. Following Cat 3.19, let us start with:

PROPOSITION 3.22. *Given a finite abelian group G , and $p \mid |G|$, the set*

$$G_p = \left\{ g \in G \mid \exists k \in \mathbb{N}, g^{p^k} = 1 \right\}$$

is a subgroup, having as order the biggest power of p dividing $|G|$.

PROOF. This follows from the Cauchy theorem, the idea being as follows:

(1) To start with, the fact that the set in the statement $G_p \subset G$ is a subgroup is clear, coming from the following computation, valid inside any abelian group:

$$g^a = 1, h^b = 1 \implies (gh)^{ab} = g^a h^b = 1$$

Indeed, given two elements $g, h \in G$, having as orders powers of p , this computation shows that $gh \in G$ has as order a certain power of p too, as desired.

(2) Next, assuming $|G| = p^k n$ with $(n, p) = 1$, we must show that we have $|G_p| = p^k$. But this is best seen by contradiction. Indeed, assuming $p \mid |G/G_p|$, by Cauchy we would

have a certain non-trivial element $hG_p \in G/G_p$ of order p . But this means $h \notin G_p$, $h^p \in G_p$, which in turn reads $h \notin G_p$, $h \in G_p$, which is contradictory. \square

As a continuation of this, we have the following key result:

THEOREM 3.23. *Given a finite abelian group G , we have*

$$G = \prod_p G_p$$

with $G_p \subset G$ with p prime being the subgroups constructed above.

PROOF. By using the fact that our group G is abelian, we have a group morphism as follows, with the order of the factors when computing $\prod_p g_p$ being irrelevant:

$$\prod_p G_p \rightarrow G \quad , \quad (g_p) \rightarrow \prod_p g_p$$

(1) Our first claim is that this morphism is injective. Indeed, let us consider an element in its kernel, which amounts in having an equation of the following type:

$$g_1 \dots g_k = 1$$

Now since the elements g_1 and $g_2 \dots g_k$, which are inverse to each other, must have the same order, and the order of g_1 is a certain prime power, and that of $g_2 \dots g_k$ is not divisible by that prime, we conclude that the kernel is trivial, as claimed.

(2) It remains to prove that our morphism is surjective. But this can be done in the pedestrian way, by picking $g \in G$, writing its order as $\text{ord}(g) = p_1^{a_1} \dots p_k^{a_k}$, and doing some arithmetic in order to reach to a writing of type $g = g_1 \dots g_k$, with $g_i \in G_{p_i}$. \square

Getting now to what we wanted to do, structure theorem for the abelian groups, Theorem 3.23 does half of the job. For the other half, we must decompose the components G_p . With the convention that p -group means $|G| = p^k$, for some $k \in \mathbb{N}$, we have:

THEOREM 3.24. *The abelian p -groups decompose as follows:*

$$G = \mathbb{Z}_{p^{r_1}} \times \dots \times \mathbb{Z}_{p^{r_s}}$$

That is, the abelian p -groups are the products of cyclic p -groups.

PROOF. We can do this by recurrence on $|G|$, as follows:

(1) Let us pick $g \in G$ of maximal order, say $\text{ord}(g) = p^k$, and consider the subgroup $H = \langle g \rangle$ that it generates, inside G . By recurrence, the quotient group G/H must decompose as follows, with the components C_i being cyclic groups:

$$G/H = C_1 \times \dots \times C_n$$

Our goal will be that of producing, out of this, an isomorphism as follows:

$$G = H \times C_1 \times \dots \times C_n$$

(2) Let us start by fixing some notation. The subgroups $C_i \subset G/H$ appearing above being cyclic, we can denote them as $C_i = \{z_i^a H\}$, with $z_i H \in C_i$ being some chosen generators for them. And with this, the isomorphism that we have is:

$$\varphi : C_1 \times \dots \times C_n \rightarrow G/H \quad , \quad (z_1^{a_1} H, \dots, z_n^{a_n} H) \rightarrow z_1^{a_1} \dots z_n^{a_n} H$$

Our more precise claim now, which will prove the result, is that, with a suitable choice of the generators $z_i H \in C_i$, we can lift this into an isomorphism as follows:

$$\psi : H \times C_1 \times \dots \times C_n \rightarrow G \quad , \quad (g^a, z_1^{a_1} H, \dots, z_n^{a_n} H) \rightarrow g^a z_1^{a_1} \dots z_n^{a_n}$$

(3) In order to do this, let us look at one of the components, $C = C_i$. If we pick an arbitrary generator $zH \in C$, with $z \in G$, the following happens, trivially:

$$\text{ord}(zH) | \text{ord}(z)$$

And our claim now, which will provide us with what is needed in (2), is that we can always arrange for our generator $zH \in C$, with $z \in G$, as to have equality:

$$\text{ord}(zH) = \text{ord}(z)$$

(4) Summarizing, we have eventually found something concrete to prove, in relation with what we want to do, so let us prove this. Let us start with an arbitrary generator $xH \in C$, with $x \in G$. Consider the two orders mentioned in (3), namely:

$$p^r = \text{ord}(xH) \quad , \quad p^s = \text{ord}(x) \quad , \quad r \leq s$$

Our goal will be that of suitably modifying our generator xH , as to have $r = s$.

(5) In order to do so, let us look at the following group element $y \in G$:

$$y = x^{p^r} \quad , \quad \text{ord}(y) = p^{s-r}$$

Since $\text{ord}(xH) = p^r$ we have $\text{ord}(yH) = 1$, which means $y \in H$. Now since $H = \langle g \rangle$ was the group generated by g , we can write y as follows, with $(n, p) = 1$:

$$y = g^{np^t}$$

Now recall that $g \in G$ was chosen of maximal order p^k . Thus, we have:

$$\text{ord}(y) = p^{k-t}$$

We conclude that we have $s - r = k - t$. Now consider the following element:

$$z = xg^{-np^{t-r}}$$

Our claim is that this is the element $z \in G$ that we were looking for, in (3).

(6) Indeed, we first have the following computation, which gives $\text{ord}(z) \leq p^r$:

$$z^{p^r} = x^{p^r} g^{-np^t} = y \cdot y^{-1} = 1$$

Also, $zH = xH = C$, and so $\text{ord}(zH) = |C| = p^r$. Thus we have, as desired:

$$\text{ord}(zH) = \text{ord}(z) = p^r$$

(7) Time for the endgame. Let us go back to the isomorphism in (2), which was as follows, and with the generators $z_i H \in C_i$ with $z_i \in G$ being chosen as above:

$$\varphi : C_1 \times \dots \times C_n \rightarrow G/H \quad , \quad (z_1^{a_1} H, \dots, z_n^{a_n} H) \rightarrow z_1^{a_1} \dots z_n^{a_n} H$$

Our claim is that this lifts into an isomorphism as follows:

$$\psi : H \times C_1 \times \dots \times C_n \rightarrow G \quad , \quad (g^a, z_1^{a_1} H, \dots, z_n^{a_n} H) \rightarrow g^a z_1^{a_1} \dots z_n^{a_n}$$

(8) Indeed, this latter map is well-defined, due to $\text{ord}(z_i H) = \text{ord}(z_i)$. It is also clear that ψ is a group morphism. Also, since φ is surjective, so must be ψ . Finally, since the cardinalities of the domain and range match, ψ must be an isomorphism, as desired. \square

Time now to put everything together. We obtain the following remarkable result:

THEOREM 3.25. *The finite abelian groups are the products of cyclic groups:*

$$G = \mathbb{Z}_{N_1} \times \dots \times \mathbb{Z}_{N_k}$$

Moreover, we can choose the numbers N_i to be prime powers.

PROOF. This follows indeed by putting together all the above, and more specifically, by combining Theorem 3.23 and Theorem 3.24. As further remarks on this:

(1) In fact, what you need to know is just the first part of the present theorem, because the second part is easy to recover, thanks to the following elementary isomorphisms:

$$\mathbb{Z}_{p_1^{a_1} \dots p_k^{a_k}} = \mathbb{Z}_{p_1^{a_1}} \times \dots \times \mathbb{Z}_{p_k^{a_k}}$$

(2) There is a uniqueness assertion too, which is elementary, stating that with G fully split, with N_i prime powers, the components will be unique, up to permutation. \square

As a conclusion to all this, we managed to find our way inside group theory, helped at all key moments by our good old friends, the prime numbers p . And with this being certainly a good thing, the symmetry groups being present a bit everywhere, in physics, chemistry, biology and so on. Not to mention mathematics, of course.

3d. Rings and fields

With the groups understood, time now for some more algebra. Remember from our discussion regarding fields and \mathbb{Q} , from chapter 2, that we have:

PRINCIPLE 3.26. *\mathbb{Q} is the simplest field.*

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

$$1 + 1 = 0$$

To be more precise, according to our field axioms, we certainly must have:

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

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

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

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

And, what we get in this way is a field, denoted as follows:

$$\mathbb{F}_2 = \{0, 1\}$$

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

THEOREM 3.27. *The following happen:*

- (1) \mathbb{Q} is the simplest field having the property $1 + \dots + 1 \neq 0$, in the sense that any field F having this property must contain it, $\mathbb{Q} \subset F$.
- (2) The property $1 + \dots + 1 \neq 0$ can hold or not, and if not, the smallest number of terms needed for having $1 + \dots + 1 = 0$ is a certain prime number p .
- (3) $\mathbb{F}_p = \{0, 1, \dots, p-1\}$, with p prime, is the simplest field having the property $1 + \dots + 1 = 0$, 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 $1 + \dots + 1 \neq 0$ 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 $1 + \dots + 1 = 0$, with $p = ab$ terms, chosen minimal, we would have a formula as follows, which is a contradiction:

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

(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 $1 + \dots + 1$, which must cycle modulo p , as shown by (2). \square

Getting back now to our philosophical discussion regarding numbers, what we have in Theorem 3.27 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 these two fields shall we study here, say as having been created first? Not an easy question, but as an answer to this, let us update Principle 3.26 as follows:

PRINCIPLE 3.28 (update). *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, mathematics and science being the topic of the present book.

This being said, in view of the above, let us study a bit more the fields \mathbb{F}_p , with this being good learning, trust me. At $p = 2$, the situation is as follows:

PROPOSITION 3.29. *The integers modulo 2 add and multiply according to the tables*

$$\begin{array}{cc|cc} + & 1 & & \times & 1 \\ & 1 & 0 & & 1 & 1 \end{array}$$

and they form a finite field, $\mathbb{F}_2 = \{0, 1\}$.

PROOF. This is indeed something self-explanatory, that we know from the above. Observe the similarity with the tables for numeration basis 2, which are:

$$\begin{array}{cc|cc} + & 1 & & \times & 1 \\ & 1 & 10 & & 1 & 1 \end{array}$$

And we will leave some further thinking here, as an instructive exercise. \square

Coming next, at $p = 3$ the situation is quite similar, as follows:

PROPOSITION 3.30. *The integers modulo 3 add and multiply according to the tables*

$$\begin{array}{ccc|ccc} + & 1 & 2 & & \times & 1 & 2 \\ & 1 & 2 & 0 & & 1 & 1 & 2 \\ & 2 & 0 & 1 & & 2 & 2 & 1 \end{array}$$

and they form a finite field, $\mathbb{F}_3 = \{0, 1, 2\}$.

PROOF. This is again something self-explanatory, that we know from the above. Observe the similarity with the tables for numeration basis 3, namely:

$$\begin{array}{ccc|ccc} + & 1 & 2 & & \times & 1 & 2 \\ & 1 & 2 & 10 & & 1 & 1 & 2 \\ & 2 & 10 & 11 & & 2 & 2 & 11 \end{array}$$

As before, we we will leave some further thinking here, as an exercise. \square

Let us study as well the case $p = 4$. This is not a prime, and the result here is:

PROPOSITION 3.31. *The integers modulo 4 add and multiply according to the tables*

$$\begin{array}{cccc|cccc} + & 1 & 2 & 3 & & \times & 1 & 2 & 3 \\ & 1 & 2 & 3 & 0 & & 1 & 1 & 2 & 3 \\ & 2 & 3 & 0 & 1 & & 2 & 2 & 0 & 2 \\ & 3 & 0 & 1 & 2 & & 3 & 3 & 2 & 1 \end{array}$$

and they form a beast $\mathbb{Z}_4 = \{0, 1, 2, 3\}$ which is not a field.

PROOF. This is again something self-explanatory, that we know from the above. Observe the similarity with the tables for numeration basis 4, namely:

+	1	2	3		×	1	2	3
1	2	3	10		1	1	2	3
2	3	10	11	,	2	2	10	12
3	10	11	12		3	3	12	21

As before, we we will leave some further thinking here, as an exercise. \square

The above result is quite concerning, and more on this later. In the meantime, let us record, for future reference, what is wrong with $\mathbb{Z}_4 = \{0, 1, 2, 3\}$, namely:

$$2 \times 2 = 0$$

Coming next, at $p = 5$ the situation is as follows:

PROPOSITION 3.32. *The integers modulo 5 add and multiply according to the tables*

+	1	2	3	4		×	1	2	3	4
1	2	3	4	0		1	1	2	3	4
2	3	4	0	1	,	2	2	4	1	3
3	4	0	1	2		3	3	1	4	2
4	0	1	2	3		4	4	3	2	1

and they form a finite field, $\mathbb{F}_5 = \{0, 1, 2, 3, 4\}$.

PROOF. This is again something self-explanatory, that we know from the above. Observe the similarity with the tables for numeration basis 5, namely:

+	1	2	3	4		×	1	2	3	4
1	2	3	4	10		1	1	2	3	4
2	3	4	10	11	,	2	2	4	11	13
3	4	10	11	12		3	3	11	14	22
4	10	11	12	13		4	4	13	22	31

As before, we we will leave some further thinking here, as an exercise. \square

At $p = 6$ now, which is not a prime, the result is as follows:

PROPOSITION 3.33. *The integers modulo 6 add and multiply according to the tables*

+	1	2	3	4	5		×	1	2	3	4	5
1	2	3	4	5	0		1	1	2	3	4	5
2	3	4	5	0	1		2	2	4	0	2	4
3	4	5	0	1	2	,	3	3	0	3	0	3
4	5	0	1	2	3		4	4	2	0	4	2
5	0	1	2	3	4		5	5	4	3	2	1

and they form a beast $\mathbb{Z}_6 = \{0, 1, 2, 3, 4, 5\}$ which is not a field.

PROOF. This is again something self-explanatory, that we know from the above. Observe the similarity with the tables for numeration basis 6, namely:

+	1	2	3	4	5		×	1	2	3	4	5
1	2	3	4	5	10		1	1	2	3	4	5
2	3	4	5	10	11		2	2	4	10	12	14
3	4	5	10	11	12	,	3	3	10	13	20	23
4	5	10	11	12	13		4	4	12	20	24	32
5	10	11	12	13	14		5	5	14	23	32	41

As before, we we will leave some further thinking here, as an exercise. \square

As before with $p = 4$, this is not good, and more on this later. In the meantime, let us record, for future reference, what is wrong with $\mathbb{Z}_6 = \{0, 1, 2, 3, 4, 5\}$, namely:

$$2 \times 3 = 0$$

Finally, at $p = 7$, which is a prime again, the result is as follows:

PROPOSITION 3.34. *The integers modulo 7 add and multiply according to the tables*

+	1	2	3	4	5	6		×	1	2	3	4	5	6
1	2	3	4	5	6	0		1	1	2	3	4	5	6
2	3	4	5	6	0	1		2	2	4	6	1	3	5
3	4	5	6	0	1	2	,	3	3	6	2	5	1	4
4	5	6	0	1	2	3		4	4	1	5	2	6	3
5	6	0	1	2	3	4		5	5	3	1	6	4	2
6	0	1	2	3	4	5		6	6	5	4	3	2	1

and they form a finite field, $\mathbb{F}_7 = \{0, 1, 2, 3, 4, 5, 6\}$.

PROOF. This is again something self-explanatory, that we know from the above. Observe the similarity with the tables for numeration basis 7, namely:

+	1	2	3	4	5	6		×	1	2	3	4	5	6
1	2	3	4	5	6	10		1	1	2	3	4	5	6
2	3	4	5	6	10	11		2	2	4	6	11	13	15
3	4	5	6	10	11	12	,	3	3	6	12	15	21	24
4	5	6	10	11	12	13		4	4	11	15	22	26	33
5	6	10	11	12	13	14		5	5	13	21	26	34	42
6	10	11	12	13	14	15		6	6	15	24	33	42	51

As before, we we will leave some further thinking here, as an exercise. \square

Summarizing, quite fun all this, obviously related to numeration bases, and to many other interesting things regarding the essence of numbers, that we met in chapter 1.

But the question that you surely have in mind is, what to do with $p = 4, 6$ and other composite numbers? And as a lazy answer here, let us formulate:

DEFINITION 3.35. A commutative ring is a set R 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 $0 \in R$ such that $a + 0 = a$.
- (2) Any $a \in R$ has an inverse $-a \in R$, satisfying $a + (-a) = 0$.
- (3) $ab = ba$, $a(bc) = (ab)c$, there exists $1 \in R$ such that $a1 = a$.
- (4) The sum and product are compatible via $a(b + c) = ab + ac$.

That is, this is something as a field, save for the existence of multiplicative inverses.

Which is a lazy definition indeed, as mathematicians usually pull out when stuck with something complicated. Anyway, with this we can formulate right away, as a theorem:

THEOREM 3.36. The following happen:

- (1) The integers modulo $p \in \mathbb{N}$ form a commutative ring \mathbb{Z}_p .
- (2) When p is prime, this ring is a field, denoted \mathbb{F}_p .

PROOF. This is clear indeed from definitions, because the integers modulo $p \in \mathbb{N}$ satisfy all the properties from Definition 3.35, trivially. \square

And more on commutative rings later in this book, the idea being that, while these are certainly useful for doing trivial things as above, these are key notions for understanding something else, namely algebraic geometry, and the geometry of numbers and the equations relating them, in particular. But, all this is quite advanced. More later.

Getting back now to fields, here is something non-trivial about them, which is a few levels above Theorem 3.36, and that we would like to discuss, in what follows:

FACT 3.37. Given a prime power $q = p^k$, it is still possible to manufacture a certain finite field \mathbb{F}_q , having q elements, which in the case $q = p$ is the usual \mathbb{F}_p .

So, our next objects of study will be the finite fields. Let us start with a basic result, including what we already know from Theorem 3.27, along with a little bit more:

THEOREM 3.38. Given a field F , define its characteristic $p = \text{char}(F)$ as being the smallest $p \in \mathbb{N}$ such that the following happens, and as $p = 0$, if this never happens:

$$\underbrace{1 + \dots + 1}_{p \text{ times}} = 0$$

Then, assuming $p > 0$, this characteristic p must be a prime number, we have a field embedding $\mathbb{F}_p \subset F$, and $q = |F|$ must be of the form $q = p^k$, with $k \in \mathbb{N}$.

PROOF. Crowded statement that we have here, the idea being as follows:

- (1) The fact that $p > 0$ must be prime comes by contradiction, by using:

$$\underbrace{(1 + \dots + 1)}_{a \text{ times}} \times \underbrace{(1 + \dots + 1)}_{b \text{ times}} = \underbrace{1 + \dots + 1}_{ab \text{ times}}$$

Indeed, assuming that we have $p = ab$ with $a, b > 1$, the above formula corresponds to an equality of type $AB = 0$ with $A, B \neq 0$ inside F , which is impossible.

(2) Back to the general case, F has a smallest subfield $E \subset F$, called prime field, consisting of the various sums $1 + \dots + 1$, and their quotients. In the case $p = 0$ we obviously have $E = \mathbb{Q}$. In the case $p > 0$ now, the multiplication formula in (1) shows that the set $S = \{1 + \dots + 1\}$ is stable under taking quotients, and so $E = S$.

(3) Now with $E = S$ in hand, we obviously have $(E, +) = \mathbb{Z}_p$, and since the multiplication is given by the formula in (1), we conclude that we have $E = \mathbb{F}_p$, as a field. Thus, in the case $p > 0$, we have constructed an embedding $\mathbb{F}_p \subset F$, as claimed.

(4) In the context of the above embedding $\mathbb{F}_p \subset F$, we can say that F is a vector space over \mathbb{F}_p , and so we have $|F| = p^k$, with $k \in \mathbb{N}$ being the dimension of this space. \square

Next, we have a weak analogue of the Fermat theorem, as follows:

PROPOSITION 3.39. *In a field F of characteristic $p > 0$ we have*

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

for any two elements $a, b \in F$.

PROOF. We have indeed the computation, exactly as in the proof of Fermat, by using the fact that the non-trivial binomial coefficients are all multiples of p :

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

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

Observe that we can iterate the Fermat formula, and we obtain $(a + b)^r = a^r + b^r$ for any power $r = p^s$. In particular we have, with $q = |F|$, the following formula:

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

But this is something quite interesting, showing that the following subset of F , which is closed under multiplication, is closed under addition too, and so is a subfield:

$$E = \left\{ a \in F \mid a^q = a \right\}$$

So, what is this subfield $E \subset F$? In the lack of examples, or general theory for subfields $E \subset F$, we are a bit in the dark here, but it seems quite reasonable to conjecture that we have $E = F$. Thus, our conjecture would be that we have the following formula, for any $a \in F$, and with this being the field extension of the Fermat theorem itself:

$$a^q = a$$

Now that we have our conjecture, let us think at a potential proof. And here, in the lack of anything obvious, we have the following theorem, which comes to the rescue:

THEOREM 3.40. *Given a field F , any finite subgroup of its multiplicative group*

$$G \subset F - \{0\}$$

must be cyclic.

PROOF. This can be done via some standard arithmetic, as follows:

(1) Let us pick an element $g \in G$ of highest order, $n = \text{ord}(g)$. Our claim, which will easily prove the result, is that the order $m = \text{ord}(h)$ of any $h \in G$ satisfies $m|n$.

(2) In order to prove this claim, let $d = (m, n)$, write $d = am + bn$ with $a, b \in \mathbb{Z}$, and set $k = g^a h^b$. We have then the following computations:

$$k^m = g^{am} h^{bm} = g^{am} = g^{d-bn} = g^d$$

$$k^n = g^{an} h^{bn} = h^{bn} = h^{d-am} = h^d$$

By using either of these formulae, say the first one, we obtain:

$$k^{[m,n]} = k^{mn/d} = (k^m)^{n/d} = (g^d)^{n/d} = g^n = 1$$

Thus $\text{ord}(k) | [m, n]$, and our claim is that we have in fact $\text{ord}(k) = [m, n]$.

(3) In order to prove this latter claim, assume first that we are in the case $d = 1$. But here the result is clear, because the formulae in (2) read $g = k^m, h = g^n$, and since $n = \text{ord}(g), m = \text{ord}(g)$ are prime to each other, we conclude that we have $\text{ord}(k) = mn$, as desired. As for the general case, where d is arbitrary, this follows from this.

(4) Summarizing, we have proved our claim in (2). Now since the order $n = \text{ord}(g)$ was assumed to be maximal, we must have $[m, n] | n$, and so $m | n$. Thus, we have proved our claim in (1), namely that the order $m = \text{ord}(h)$ of any $h \in G$ satisfies $m | n$.

(5) But with this claim in hand, the result follows. Indeed, since the polynomial $x^n - 1$ has all the elements $h \in G$ as roots, its degree must satisfy $n \geq |G|$. On the other hand, from $n = \text{ord}(g)$ with $g \in G$, we have $n | |G|$. We therefore conclude that we have $n = |G|$, which shows that G is indeed cyclic, generated by the element $g \in G$. \square

We can now extend the Fermat theorem to the finite fields, as follows:

THEOREM 3.41. *Given a finite field F , with $q = |F|$ we have*

$$a^q = a$$

for any $a \in F$.

PROOF. According to Theorem 3.40 the multiplicative group $F - \{0\}$ is cyclic, of order $q - 1$. Thus, the following formula is satisfied, for any $a \in F - \{0\}$:

$$a^{q-1} = 1$$

Now by multiplying by a , we are led to the conclusion in the statement, with of course the remark that the formula there trivially holds for $a = 0$. \square

The Fermat polynomial $X^p - X$ is something very useful, and its field generalization $X^q - X$, with $q = p^k$ prime power, can be used in order to elucidate the structure of finite fields. In order to discuss this question, let us start with a basic fact, as follows:

PROPOSITION 3.42. *Given a finite field F , we have*

$$X^q - X = \prod_{a \in F} (X - a)$$

with $q = |F|$.

PROOF. We know from the Fermat theorem established above that we have $a^q = a$, for any $a \in F$. We conclude from this that all the elements $a \in F$ are roots of the polynomial $X^q - X$, and so this polynomial must factorize as in the statement. \square

The continuation of the story is more complicated, as follows:

THEOREM 3.43. *For any prime power $q = p^k$ there is a unique field \mathbb{F}_q having q elements. At $k = 1$ this is the usual \mathbb{F}_p . In general, this is the splitting field of:*

$$P = X^q - X$$

Moreover, we can construct an explicit model for \mathbb{F}_q , at $q = p^2$ or higher, as

$$\mathbb{F}_q = \mathbb{F}_p[X]/(Q)$$

with $Q \in \mathbb{F}_p[X]$ being a suitable irreducible polynomial, of degree k .

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

(1) The first assertion, regarding the existence and uniqueness of \mathbb{F}_q , follows from some abstract algebra. Indeed, we know from Proposition 3.42 that given a finite field, $|F| = q$ with $k \in \mathbb{N}$, the Fermat polynomial $P = X^q - X$ factorizes as:

$$X^q - X = \prod_{a \in F} (X - a)$$

But this shows, via some abstract algebra, that we will learn later in this book, that our field F must be the splitting field of P , and so is unique. As for the existence, this follows again from abstract algebra, telling us that the splitting field always exists.

(2) In what regards now the modeling of \mathbb{F}_q , at $q = p$ there is nothing to do, because we have our usual \mathbb{F}_p here. At $q = p^2$ and higher, we know from commutative algebra that we have an isomorphism as follows, whenever $Q \in \mathbb{F}_p[X]$ is taken irreducible:

$$\mathbb{F}_q = \mathbb{F}_p[X]/(Q)$$

(3) Regarding now the best choice of the irreducible polynomial $Q \in \mathbb{F}_p[X]$, providing us with a good model for the finite field \mathbb{F}_q , that we can use in practice, this question depends on the value of $q = p^k$, and many things can be said here. All in all, our models are quite similar to $\mathbb{C} = \mathbb{R}[i]$, with i being a formal number satisfying $i^2 = -1$.

(4) To be more precise, at the simplest exponent, $q = 4$, to start with, we can use $Q = X^2 + X + 1$, with this being actually the unique possible choice of a degree 2 irreducible polynomial $Q \in \mathbb{F}_2[X]$, and this leads to a model as follows:

$$\mathbb{F}_4 = \left\{ 0, 1, a, a+1 \mid a^2 = a+1 \right\}$$

To be more precise here, we assume of course that the characteristic of our model is $p = 2$, which reads $x + x = 0$ for any x , and so determines the addition table. As for the multiplication table, this is uniquely determined by $a^2 = -a - 1 = a + 1$.

(5) Next, at exponents of type $q = p^2$ with $p \geq 3$ prime, we can use $Q = X^2 - r$, with r being a non-square modulo p , and with $(p-1)/2$ choices here. We are led to:

$$\mathbb{F}_{p^2} = \left\{ a + b\gamma \mid \gamma^2 = r \right\}$$

Here, as before with \mathbb{F}_4 , our formula is something self-explanatory. Observe the analogy with $\mathbb{C} = \mathbb{R}[i]$, with i being a formal number satisfying $i^2 = -1$.

(6) Finally, at $q = p^k$ with $k \geq 3$ things become more complicated, but the main idea remains the same. We have for instance models for \mathbb{F}_8 , \mathbb{F}_{27} using $Q = X^3 - X - 1$, and a model for \mathbb{F}_{16} using $Q = X^4 + X + 1$. Many other things can be said here.

(7) Summarizing, we are led to the conclusions in the statement, modulo some abstract algebra, that we still have to learn. We will be back to this, later in this book. \square

3e. Exercises

This was a quite standard chapter, and as exercises on this, we have:

EXERCISE 3.44. *Learn more about finite groups, as much as you can.*

EXERCISE 3.45. *Learn a bit about rings, and their use in algebraic geometry.*

EXERCISE 3.46. *Learn also about modules and ideals, and their use in geometry.*

EXERCISE 3.47. *Read about rational functions, and what can be done with them.*

EXERCISE 3.48. *Have a look as well at the quadratic fields, such as $\mathbb{Q}[\sqrt{2}]$.*

EXERCISE 3.49. *Review as well the other basic examples of characteristic 0 fields.*

EXERCISE 3.50. *Learn full Galois theory, including all the needed preliminaries.*

EXERCISE 3.51. *Experiment a bit, with the modeling of various finite fields.*

As bonus exercise, read about the Nullstellensatz. This is what algebra is about.

CHAPTER 4

Squares, residues

4a. Squares, residues

Let us go back now to what we did in chapter 1 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:

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

Many things can be said here, of various levels of difficulty. At the simplest level, we have the following result, which is something very useful, in practice:

THEOREM 4.1. *The squares modulo 4 can be of the form*

$$a^2 = 0, 1(4)$$

that is, with the cases $a^2 = 2, 3(4)$ being excluded.

PROOF. When our number to be squared is even, $a = 2k$, we have:

$$a^2 = (2k)^2 = 4k^2 = 0(4)$$

As for the case where our number is odd, $a = 2k + 1$, here we have:

$$a^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 1(4)$$

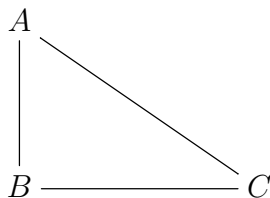
Thus, we are led to the conclusion in the statement. □

As a comment here, in the odd case, $a = 2k + 1$, since $k(k + 1)$ is even, we obtain in fact $a^2 = 1(8)$. Thus, we have in fact $a^2 = 0, 1, 4(8)$, improving what Theorem 4.1 says. However, for most applications, Theorem 4.1 as stated will be just fine.

The above statement is quite interesting, and it is quite clear that this is just the tip of the iceberg, with the equation $a^2 = b(c)$, with $c = p^k$ being a prime power, being something of interest, waiting to be systematically investigated. We will be back to this question, with several answers to it, later in this chapter.

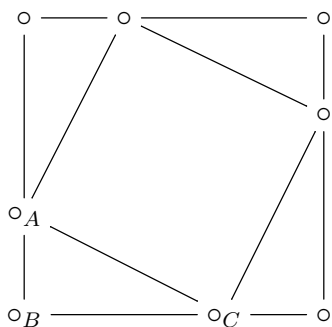
Before that, however, let us discuss some older mathematics in relation with the squares. We have indeed the following key result, due to Pythagoras:

THEOREM 4.2 (Pythagoras). *In a right triangle ABC ,*



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

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



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

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

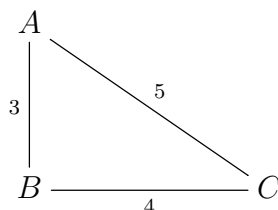
On the other hand, the outer square is made of the smaller square, having side AC , and of four identical right triangles, having sizes AB, BC . Thus:

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

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

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

THEOREM 4.3. *A triangle having sides 3, 4, 5 is a right triangle:*



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

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

$$16 + 9 = 25$$

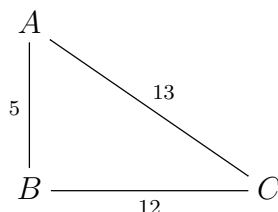
As for the second assertion, what does that exactly mean, and how can that be used in practice, we will leave this as an interesting engineering exercise. \square

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

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

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

THEOREM 4.4. *A triangle having sides 5, 12, 13 is a right triangle:*



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

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

$$144 + 25 = 169$$

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

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

THEOREM 4.5. *The Pythagoras equation, namely*

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

can be fully solved over the integers, the solutions being

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} \frac{a}{b} &= \frac{1}{2} \left(\frac{m}{n} - \frac{n}{m} \right) = \frac{m^2 - n^2}{2mn} \\ \frac{c}{b} &= \frac{1}{2} \left(\frac{m}{n} + \frac{n}{m} \right) = \frac{m^2 + n^2}{2mn} \end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Observe that $40 + 9 + 41 = 90$, so as good news here, we have solved as well a quite difficult exercise left, namely the one at the end of the proof of Theorem 4.4. \square

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

$$a^n + b^n = c^n$$

And we will leave some reading about this as an exercise, but with the comment however that this is quite difficult, and in relation for instance with the more advanced algebraic number theory, that we will discuss later in this book.

Getting back now to congruences, we have seen that the proof of Theorem 4.5 uses Theorem 4.1, and with this providing us with some further motivation for congruences. So, time to study more in detail the equations of type $a^2 = b(c)$, with $c = p^k$ prime power. At $p = 2$ we have the following result, coming as a continuation of Theorem 4.1:

THEOREM 4.6. *The quadratic residues modulo 2^k are as follows:*

- (1) *At $k = 1$ there is no restriction, these are $0, 1(2)$.*
- (2) *At $k = 2$ these are $0, 1(4)$.*
- (3) *At $k = 3$ these are $0, 1, 4(8)$.*
- (4) *At $k = 4$ these are $0, 1, 4, 9(16)$.*
- (5) *At $k = 5$ these are $0, 1, 4, 9, 16, 17, 25(32)$.*

PROOF. This is very elementary mathematics, as follows:

- (1) Nothing to be done at $k = 1$, with both cases $0, 1(2)$ being obviously possible.
- (2) At $k = 2$, this is something that we already know, from Theorem 4.1.
- (3) At $k = 3$ now, the possible numbers modulo 8 are as follows:

$$a = 0, \pm 1, \pm 2, \pm 3, \pm 4$$

The squares of these numbers, taken modulo 8, are then as follows:

$$a^2 = 0, 1, 4, 1, 0$$

Thus, we are led to the conclusion in the statement, $a^2 = 0, 1, 4(8)$.

- (4) At $k = 4$ now, the possible numbers modulo 16 are as follows:

$$a = 0, \pm 1, \pm 2, \pm 3, \pm 4, \pm 5, \pm 6, \pm 7, 8$$

The squares of these numbers, taken modulo 16, are then as follows:

$$a^2 = 0, 1, 4, 9, 0, 9, 4, 1, 0$$

Thus, we are led to the conclusion in the statement, $a^2 = 0, 1, 4, 9(16)$.

- (5) At $k = 5$ now, the possible numbers modulo 32 are as follows:

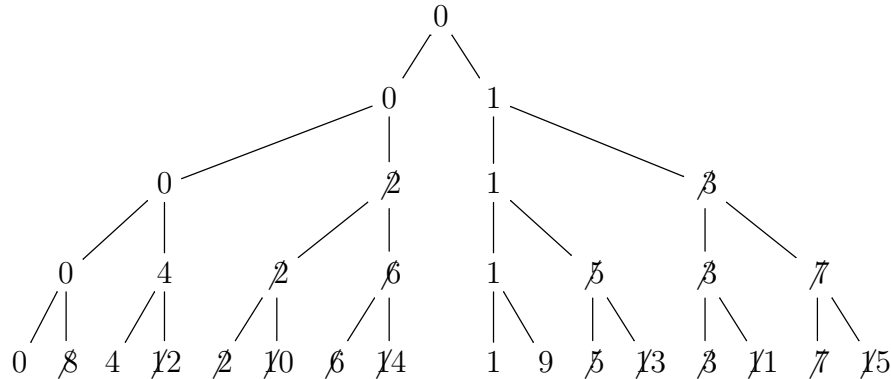
$$a = 0, \pm 1, \pm 2, \pm 3, \pm 4, \pm 5, \pm 6, \pm 7, \pm 8, \pm 9, \pm 10, \pm 11, \pm 12, \pm 13, \pm 14, \pm 15, 16$$

The squares of these numbers, taken modulo 32, are then as follows:

$$a^2 = 0, 1, 4, 9, 16, 25, 4, 17, 0, 17, 4, 25, 16, 9, 4, 1, 0$$

Thus, we are led to the conclusion in the statement, $a^2 = 0, 1, 4, 9, 16, 17, 25(32)$. \square

The above result is certainly good to know as such, but at the level of the mathematics, it is quite unclear what happens there. One way of interpreting the results is by drawing a tree, as follows, with the quadratic residues which are forbidden being barred:



Let us see as well what happens at $p = 3$. Here the result is as follows:

THEOREM 4.7. *The quadratic residues modulo 3^k are as follows:*

- (1) *At $k = 1$ these are $0, 1(3)$.*
- (2) *At $k = 2$ these are $0, 1, 4, 7(9)$.*
- (3) *At $k = 3$ these are $0, 1, 4, 7, 9, 10, 13, 16, 19, 22, 25(27)$.*

PROOF. This is again very elementary mathematics, as follows:

- (1) At $k = 1$ this comes from the following computation:

$$(3n \pm 1)^2 = 1(3)$$

- (2) At $k = 2$ now, the possible numbers modulo 9 are as follows:

$$a = 0, \pm 1, \pm 2, \pm 3, \pm 4$$

The squares of these numbers, taken modulo 9, are then as follows:

$$a^2 = 0, 1, 4, 0, 7$$

Thus, we are led to the conclusion in the statement, $a^2 = 0, 1, 4, 7(9)$.

- (3) At $k = 3$ now, the possible numbers modulo 27 are as follows:

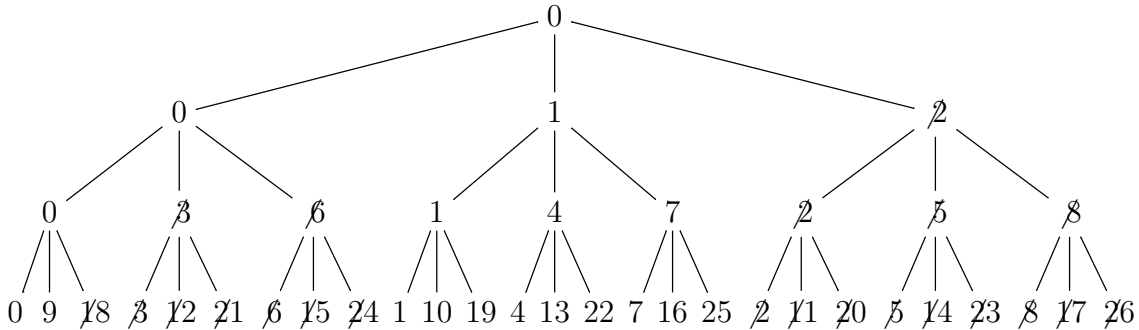
$$a = 0, \pm 1, \pm 2, \pm 3, \pm 4, \pm 5, \pm 6, \pm 7, \pm 8, \pm 9, \pm 10, \pm 11, \pm 12, \pm 13$$

The squares of these numbers, taken modulo 27, are then as follows:

$$a^2 = 0, 1, 4, 9, 16, 25, 9, 22, 10, 0, 19, 13, 9, 7$$

Thus, we are led to $a^2 = 0, 1, 4, 7, 9, 10, 13, 16, 19, 22, 25(27)$, as stated. \square

As before with $p = 2$, the above result is certainly good to know, but at the level of the mathematics, it is quite unclear what happens there. We can interpret the results by drawing a tree, as follows, with the forbidden quadratic residues being barred:



Let us see as well what happens at $p = 5$. Here the result is as follows:

THEOREM 4.8. *The quadratic residues modulo 5^k are as follows:*

- (1) *At $k = 1$ these are $0, 1, 4(5)$.*
- (2) *At $k = 2$ these are $0, 1, 4, 9, 11, 14, 16, 19, 21, 24(25)$.*

PROOF. This is again very elementary mathematics, as follows:

(1) At $k = 1$, the possible numbers modulo 5 are as follows:

$$a = 0, \pm 1, \pm 2$$

The squares of these numbers, taken modulo 5, are then as follows:

$$a^2 = 0, 1, 4$$

Thus, we are led to the conclusion in the statement, $a^2 = 0, 1, 4(5)$.

(2) At $k = 2$ now, the possible numbers modulo 25 are as follows:

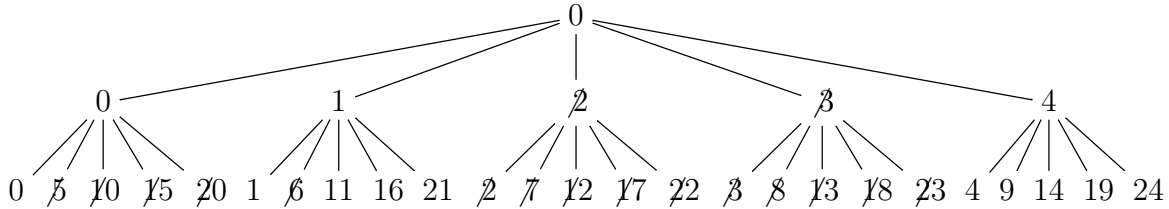
$$a = 0, \pm 1, \pm 2, \pm 3, \pm 4, \pm 5, \pm 6, \pm 7, \pm 8, \pm 9, \pm 10, \pm 11, \pm 12$$

The squares of these numbers, taken modulo 25, are then as follows:

$$a^2 = 0, 1, 4, 9, 16, 0, 11, 24, 14, 6, 0, 21, 19$$

Thus, we are led to $a^2 = 0, 1, 4, 9, 11, 14, 16, 19, 21, 24(25)$, as stated. \square

As before with $p = 2, 3$, at the level of the mathematics, it is quite unclear what is going on. We can interpret the above results by drawing a tree, as follows:



And we will stop our elementary study here, all this being obviously a bit too amateurish. We will be back in a moment to such things, which more advanced tools.

4b. Legendre symbol

Getting back now to what we wanted to do in this chapter, understand $a = b^2(c)$, we have the following definition, putting everything on a solid basis:

DEFINITION 4.9. *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 4.10. *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 the Fermat theorem 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 we have:

$$a^{p-1} = 1(p)$$

Now since our prime number p was assumed to be odd, $p \geq 3$, we can write this formula by factorizing, 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 $b \rightarrow -b$, 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:

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

(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 4.9, we obtain the formula in the statement. \square

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

PROPOSITION 4.11. *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 follows from what we have, the idea being as follows:

(1) To start with, this is clear indeed indeed from the Euler formula from Theorem 4.10, because the quantity $a^{\frac{p-1}{2}}(p)$ is obviously multiplicative in $a \in \mathbb{Z}$.

(2) 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 4.10. \square

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

PROPOSITION 4.12. *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

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

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

PROOF. The first assertion is a consequence of Proposition 4.11, 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 4.11. \square

Question now, how to compute 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 in this chapter, of course with full proofs, but with some of these proofs being a bit formal and mysterious, in need of more study, that we will leave for later. As a first and main result, which is something heavy, we have:

THEOREM 4.13. *We have the quadratic reciprocity formula*

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

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

PROOF. This is something quite tricky, one proof being as follows:

(1) First we have a combinatorial formula for the Legendre symbol, called Gauss lemma. Given a prime number $q \geq 3$, and $a \neq 0(q)$, consider the following sequence:

$$a, 2a, 3a, \dots, \frac{q-1}{2}a$$

The Gauss lemma tells us that if we look at these numbers modulo q , and denote by n the number of residues modulo q which are greater than $q/2$, then:

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

(2) In order to prove this lemma, the idea is to look at the following product:

$$Z = a \times 2a \times 3a \times \dots \times \frac{q-1}{2} a$$

Indeed, on one hand we have the following formula, with Euler used at the end:

$$Z = a^{\frac{q-1}{2}} \left(\frac{q-1}{2}\right)! = \left(\frac{a}{q}\right) \left(\frac{q-1}{2}\right)!$$

(3) On the other hand, we can compute Z in more complicated way, but leading to a simpler answer. Indeed, let us define the following function:

$$|x| = \begin{cases} x & \text{if } 0 < x < q/2 \\ q - x & \text{if } q/2 < x < q \end{cases}$$

With this convention, our product Z is given by the following formula, with n being as in (1), namely the number of residues modulo q which are greater than $q/2$:

$$Z = (-1)^n \times |a| \times |2a| \times |3a| \times \dots \times \left|\frac{q-1}{2} a\right|$$

(4) But, the numbers $|ra|$ appearing in the above formula are all distinct, so up to a permutation, these must be exactly the numbers $1, 2, \dots, \frac{q-1}{2}$. That is, we have:

$$\left\{|a|, |2a|, |3a|, \dots, \left|\frac{q-1}{2} a\right|\right\} = \left\{1, 2, 3, \dots, \frac{q-1}{2}\right\}$$

Now by multiplying all these numbers, we obtain, via the formula in (3):

$$Z = (-1)^n \left(\frac{q-1}{2}\right)!$$

(5) But this is what we need, because when comparing with what we have in (2), we obtain the following formula, which is exactly the one claimed by the Gauss lemma:

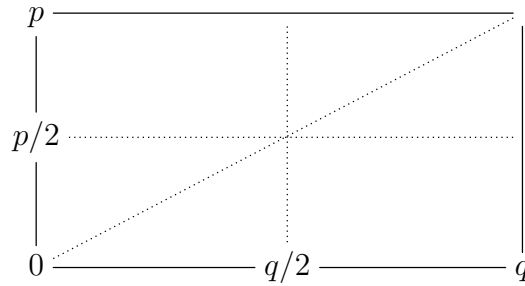
$$\left(\frac{a}{q}\right) = (-1)^n$$

(6) Next, we have a variation of this formula, due to Eisenstein. His formula for the Legendre symbol, this time involving a prime number numerator $p \geq 3$ in the symbol, is

as follows, with the quantities on the right being integer parts, and with the proof being very similar to the proof of the Gauss lemma, that we will leave here as an exercise:

$$\left(\frac{p}{q}\right) = (-1)^n \quad , \quad n = \sum_{k=0}^{(q-1)/2} \left\lfloor \frac{2kp}{q} \right\rfloor$$

(7) The key point now is that, in this latter formula of Eisenstein, the number n itself counts the points of the lattice \mathbb{Z}^2 lying in the triangle $(0,0), (q,0), (q,p)$. So, based on this observation, let us draw a picture, as follows:



(8) We must count the points of \mathbb{Z}^2 lying in the triangle $(0,0), (q,0), (q,p)$, modulo 2. This triangle has 3 components, when split by the dotted lines above. Since the points at right, in the small rectangle, and in the small triangle above it, will cancel modulo 2, we are left with the points at left, in the small triangle there, and the conclusion is that, if we denote by m the number of integer points there, we have the following formula:

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

(9) Now by flipping the diagram, we have as well the following formula, with r being the number of integer points in the small triangle above the small triangle in (8):

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

(10) But, since our two small triangles add up to a small rectangle, we have:

$$m + r = \frac{p-1}{2} \cdot \frac{q-1}{2}$$

Thus, by multiplying the formulae in (8) and (9), we are led to the result. \square

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:

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

In fact, when combining Theorem 4.13 with Proposition 4.11, it is quite clear that, no matter how big p is, if a has only small prime factors, we are saved.

4c. Further results

Besides Proposition 4.11, the quadratic reciprocity formula comes accompanied by two other statements, which are very useful in practice. First, at $a = -1$, we have:

PROPOSITION 4.14. *We have the following formula,*

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

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 4.15. *We have the following formula,*

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

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

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

(1) As a first observation, the Euler formula at $a = 2$ is as follows, obviously well below the quality of the very precise formula in the statement:

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

As a second observation, the quadratic reciprocity formula, assuming that known, cannot help either, because in that formula $p, q \geq 3$ are odd primes.

(2) Thus, we must improvise, and prove the result. The proof will come via the following formula, which is equivalent to the formula in the statement:

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

Let us also mention that, despite 2 being an even prime, the problematics here is a bit similar to the one of quadratic reciprocity, and the proof below will contain many good ideas, that we will use later, in an alternative proof of quadratic reciprocity.

(3) Getting started now, let us introduce a formal number i satisfying $i^2 = -1$, then a formal number w satisfying $w^2 = i$, and then let us set $t = w + w^{-1}$:

$$i^2 = -1 \quad , \quad w^2 = i \quad , \quad t = w + w^{-1}$$

Quite crazy all this, you would say, and in answer, no worries, we will get to know more, about where all this comes from, as this book further develops.

(4) As a second remark that you might have, in case you know a bit about complex numbers, you might say that the above is not crazy, but rather stupid, because $t = \sqrt{2}$. In answer, yes I know, but it is better to forget this, and do formal arithmetic instead, with integers as scalars, based on our rules above, and the following computation:

$$\begin{aligned} t^2 &= 2 + w^2 + w^{-2} \\ &= 2 + i - i \\ &= 2 \end{aligned}$$

(5) Now by using the Euler formula for the Legendre symbol, we have:

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

(6) By multiplying now by t we obtain from this, in a formal sense, and I will leave it you to clarify all the details here, namely what this formal sense exactly means:

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

(7) On the other hand, by using the binomial formula, and the standard fact that all non-trivial binomial coefficients are multiples of p , we obtain, again formally:

$$\begin{aligned} t^p &= (w + w^{-1})^p \\ &= \sum_{k=0}^p \binom{k}{p} w^k w^{k-p} \\ &= w^p + w^{-p} \pmod{p} \end{aligned}$$

(8) Now let us look at the quantity $w^p + w^{-p}$. Since we have $w^4 = -1$, this quantity will depend only on p modulo 8, and more precisely, we have:

$$w^p + w^{-p} = \begin{cases} w + w^{-1} & \text{if } p \equiv \pm 1 \pmod{8} \\ -w - w^{-1} & \text{if } p \equiv \pm 3 \pmod{8} \end{cases}$$

Thus $w^p + w^{-p} = \pm t$, with the sign depending on p modulo 8, and more specifically:

$$w^p + w^{-p} = (-1)^{\frac{p^2-1}{8}} t$$

(9) Time now to put everything together. By combining (6,7,8) we obtain:

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

By dividing by t , this gives the following formula:

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

But the mod p symbol can now be dropped, because our equality is between two ± 1 quantities, and we obtain the formula in the statement. \square

As already mentioned, the above result was something quite tricky, and we will be back to this later in this book, with some further explanations on all this.

As a continuation of this, speaking Legendre symbol for small values of the upper variable, we can try to compute these for $a = \pm 3, 4, 5, 6, 7, 8, \dots$. But by multiplicativity plus Proposition 4.14 plus Theorem 4.15 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 4.13. We first have:

PROPOSITION 4.16. *We have the following formula,*

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

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:

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

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. \square

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

PROPOSITION 4.17. *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:

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

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

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

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

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. \square

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

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

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:

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

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. \square

As a further plot to the story, the theory of Kronecker symbols can be further generalized into a theory of Hilbert symbols. But then, you guessed it right, Hilbert was not the last one in the series, which contains some other illustrious mathematicians.

So bad these times are over. More recently, Prince made an attempt to join the series, with a very interesting symbol, which was however not accepted by the community.

4d. Some applications

Time now for some applications. Some very interesting objects in linear algebra are the Hadamard matrices, which are defined in a very simple way, as follows:

DEFINITION 4.20. *An Hadamard matrix is a square binary matrix*

$$H \in M_N(\pm 1)$$

whose rows are pairwise orthogonal.

Observe that the orthogonality condition states that, when comparing any two rows, the number of matches must equal the numbers of mismatches. As a basic example now, at $N = 2$ we have the following matrix, called first Walsh matrix:

$$W_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

This matrix is quite trivial, of size 2×2 , but by taking tensor powers of it, we have as examples the higher Walsh matrices as well, having size $2^k \times 2^k$, given by:

$$W_{2^k} = W_2^{\otimes k}$$

As an example here, the second Walsh matrix, obtained by tensoring W_2 with itself, is as follows, with respect to the lexicographic order on the double indices:

$$W_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

Which raises the question, what happens then in arbitrary size $N \times N$? Do we still have there Hadamard matrices, that we can use for various applied linear algebra purposes?

In order to discuss this question, as a first remark, it is clear that we must have $2|N$. Along the same lines, it is easy to see, by playing around with the first rows, that once your matrix has $N \geq 3$ rows, we must have $4|N$, the precise result being as follows:

PROPOSITION 4.21. *The size of an Hadamard matrix $H \in M_N(\pm 1)$ must satisfy*

$$N \in \{2\} \cup 4\mathbb{N}$$

with this coming from the orthogonality condition between the first 3 rows.

PROOF. By permuting the rows and columns or by multiplying them by -1 , as to rearrange the first 3 rows, we can always assume that our matrix looks as follows:

$$H = \begin{pmatrix} \underbrace{1 \dots 1}_x & \underbrace{1 \dots 1}_y & \underbrace{1 \dots 1}_z & \underbrace{1 \dots 1}_t \\ 1 \dots 1 & 1 \dots 1 & -1 \dots -1 & -1 \dots -1 \\ 1 \dots 1 & -1 \dots -1 & 1 \dots 1 & -1 \dots -1 \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

Now if we denote by x, y, z, t the sizes of the block columns, as indicated, the orthogonality conditions between the first 3 rows give the following system of equations:

$$(1 \perp 2) \quad : \quad x + y = z + t$$

$$(1 \perp 3) \quad : \quad x + z = y + t$$

$$(2 \perp 3) \quad : \quad x + t = y + z$$

The numbers x, y, z, t being such that the average of any two equals the average of the other two, and so equals the global average, the solution of our system is:

$$x = y = z = t$$

Thus the matrix size $N = x + y + z + t$ must be a multiple of 4, as claimed. \square

The above result is something quite interesting, and a similar analysis with 4 rows or more does not give any further restriction on the possible values of the size $N \in \mathbb{N}$.

In fact, we are led in this way to the following famous conjecture:

CONJECTURE 4.22 (Hadamard). *There is an Hadamard matrix of order N ,*

$$H \in M_N(\pm 1)$$

for any $N \in 4\mathbb{N}$.

Normally this is an analytic question, because in practice the number of Hadamard matrices grows exponentially with N , and so in order to prove the conjecture, you just need a modest lower estimate on this number. But, no one knows how to do this, and this despite the Hadamard conjecture being open for more than 100 years.

This being said, what we can do with our number theory methods is to verify at least the Hadamard conjecture at small values of $N \in 4\mathbb{N}$. And here, with $N = 4, 8$ being solved by the Walsh matrices, we are faced with constructing a matrix at $N = 12$.

In order to solve this question, let $q = p^k$ be an odd prime power, and consider the quadratic character of \mathbb{F}_q , given by the following formula:

$$\chi(a) = \begin{cases} 0 & \text{if } a = 0 \\ 1 & \text{if } a = b^2, b \neq 0 \\ -1 & \text{otherwise} \end{cases}$$

Next, consider as well the following matrix, with indices in \mathbb{F}_q :

$$Q_{ab} = \chi(b - a)$$

With these conventions, the Paley construction of Hadamard matrices, which works well at $N = 12$, and at many other values of N , is as follows:

THEOREM 4.23. *Given an odd prime power $q = p^k$, construct $Q_{ab} = \chi(b - a)$ as above. We have then constructions of Hadamard matrices, as follows:*

(1) *Paley 1: if $q = 3(4)$ we have a matrix of size $N = q + 1$, as follows:*

$$P_N^1 = 1 + \begin{pmatrix} 0 & 1 & \dots & 1 \\ -1 & & & \\ \vdots & & Q & \\ -1 & & & \end{pmatrix}$$

(2) *Paley 2: if $q = 1(4)$ we have a matrix of size $N = 2q + 2$, as follows:*

$$P_N^2 = \begin{pmatrix} 0 & 1 & \dots & 1 \\ 1 & & & \\ \vdots & & Q & \\ 1 & & & \end{pmatrix} : \quad 0 \rightarrow \begin{pmatrix} 1 & -1 \\ -1 & -1 \end{pmatrix} \quad , \quad \pm 1 \rightarrow \pm \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

These matrices are skew-symmetric ($H + H^t = 2$), respectively symmetric ($H = H^t$).

PROOF. In order to simplify the presentation, we denote by 1 all the identity matrices, of any size, and by \mathbb{I} all the rectangular all-one matrices, of any size as well. It is elementary to check that the matrix $Q_{ab} = \chi(a - b)$ has the following properties:

$$QQ^t = q1 - \mathbb{I} \quad , \quad Q\mathbb{I} = \mathbb{I}Q = 0$$

In addition, we have the following formulae, which are elementary as well, coming from the fact that -1 is a square in \mathbb{F}_q precisely when $q = 1(4)$:

$$q = 1(4) \implies Q = Q^t \quad , \quad q = 3(4) \implies Q = -Q^t$$

With these observations in hand, the proof goes as follows:

(1) With our above conventions for 1 and \mathbb{I} , the matrix in the statement is:

$$P_N^1 = \begin{pmatrix} 1 & \mathbb{I} \\ -\mathbb{I} & 1 + Q \end{pmatrix}$$

With this formula in hand, the Hadamard matrix condition follows from:

$$\begin{aligned} P_N^1 (P_N^1)^t &= \begin{pmatrix} 1 & \mathbb{I} \\ -\mathbb{I} & 1 + Q \end{pmatrix} \begin{pmatrix} 1 & -\mathbb{I} \\ \mathbb{I} & 1 - Q \end{pmatrix} \\ &= \begin{pmatrix} N & 0 \\ 0 & \mathbb{I} + 1 - Q^2 \end{pmatrix} \\ &= \begin{pmatrix} N & 0 \\ 0 & N \end{pmatrix} \end{aligned}$$

(2) If we denote by G, F the 2×2 matrices in the statement, which replace respectively the 0, 1 entries, then we have the following formula for our matrix:

$$P_N^2 = \begin{pmatrix} 0 & \mathbb{I} \\ \mathbb{I} & Q \end{pmatrix} \otimes F + 1 \otimes G$$

With this formula in hand, the Hadamard matrix condition follows from:

$$\begin{aligned} (P_N^2)^2 &= \begin{pmatrix} 0 & \mathbb{I} \\ \mathbb{I} & Q \end{pmatrix}^2 \otimes F^2 + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes G^2 + \begin{pmatrix} 0 & \mathbb{I} \\ \mathbb{I} & Q \end{pmatrix} \otimes (FG + GF) \\ &= \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix} \otimes 2 + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes 2 + \begin{pmatrix} 0 & \mathbb{I} \\ \mathbb{I} & Q \end{pmatrix} \otimes 0 \\ &= \begin{pmatrix} N & 0 \\ 0 & N \end{pmatrix} \end{aligned}$$

Finally, the last assertion is clear, from the above formulae relating Q, Q^t . \square

Getting back now to the Hadamard conjecture, we have:

THEOREM 4.24. *The Hadamard conjecture is verified up to $N = 88$, as follows:*

- (1) *At $N = 4, 8, 16, 32, 64$ we have Walsh matrices.*
- (2) *At $N = 12, 20, 24, 28, 44, 48, 60, 68, 72, 80, 84, 88$ we have Paley 1 matrices.*
- (3) *At $N = 36, 52, 76$ we have Paley 2 matrices.*
- (4) *At $N = 40, 56$ we have Paley 1 matrices tensored with W_2 .*

However, at $N = 92$ these constructions (Walsh, Paley, tensoring) don't work.

PROOF. First of all, the numbers in (1-4) are indeed all the multiples of 4, up to 88. As for the various assertions, the proof here goes as follows:

- (1) This is clear indeed from the definition of Walsh matrices.
- (2) Here the number $N - 1$ takes the following values:

$$q = 11, 19, 23, 27, 43, 47, 59, 67, 71, 79, 83, 87$$

But these are all prime powers, so we can apply the Paley 1 construction, as stated.

(3) Since $N = 4(8)$ here, and $N/2 - 1$ takes the values $q = 17, 25, 37$, all prime powers, we can indeed apply the Paley 2 construction, in these cases.

(4) At $N = 40$ we have indeed $P_{20}^1 \otimes W_2$, and at $N = 56$ we have $P_{28}^1 \otimes W_2$.

(5) Finally, at $N = 92$ we have $92 - 1 = 7 \times 13$, so Paley 1 does not work, and $92/2 = 46$, so Paley 2, or tensoring with W_2 , does not work either. \square

At $N = 92$ now, the situation is considerably more complicated, and we have:

THEOREM 4.25. *Assuming that $A, B, C, D \in M_K(\pm 1)$ are circulant, symmetric, pairwise commute and satisfy the condition*

$$A^2 + B^2 + C^2 + D^2 = 4K$$

the following $4K \times 4K$ matrix is Hadamard, called of Williamson type:

$$H = \begin{pmatrix} A & B & C & D \\ -B & A & -D & C \\ -C & D & A & -B \\ -D & -C & B & A \end{pmatrix}$$

Moreover, matrices A, B, C, D as above exist at $K = 23$, where $4K = 92$.

PROOF. Consider the standard quaternion units $1, i, j, k \in M_4(0, 1)$, which describe the positions of the A, B, C, D entries in the matrix H from the statement:

$$\begin{aligned} 1 &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, & i &= \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \\ j &= \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, & k &= \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{aligned}$$

Thus, the above matrix H can be written as follows, in terms of $1, i, j, k$:

$$H = A \otimes 1 + B \otimes i + C \otimes j + D \otimes k$$

Assuming now that A, B, C, D are symmetric, we have:

$$\begin{aligned} HH^t &= (A \otimes 1 + B \otimes i + C \otimes j + D \otimes k) \\ &\quad (A \otimes 1 - B \otimes i - C \otimes j - D \otimes k) \\ &= (A^2 + B^2 + C^2 + D^2) \otimes 1 - ([A, B] - [C, D]) \otimes i \\ &\quad -([A, C] - [B, D]) \otimes j - ([A, D] - [B, C]) \otimes k \end{aligned}$$

Now assume that our matrices A, B, C, D pairwise commute, and satisfy the condition in the statement. In this case, it follows from the above formula that we have:

$$HH^t = 4K$$

Thus, we obtain indeed an Hadamard matrix, as claimed. However, finding such matrices is in general a difficult task, and this is where Williamson's extra assumption in the statement, that A, B, C, D should be taken circulant, comes from. Finally, regarding the $K = 23$ and $N = 92$ example, this comes via a computer search. \square

At higher N things become more technical, and more complicated constructions, along the lines of those of Paley and Williamson, are needed. Quite curiously, as of now, early 21th century, the human knowledge stops at the number of the beast, namely:

$$\mathfrak{N} = 666$$

That is, explicit examples of Hadamard matrices have been constructed for all multiples of four $N \leq 664$, but no such matrix is known so far at $N = 668$.

4e. Exercises

This was another quite advanced arithmetic chapter, and as exercises here, we have:

EXERCISE 4.26. *Work out the numerics for the solutions of the Pythagoras equation.*

EXERCISE 4.27. *Compute more quadratic residues with respect to $q = p^k$ numbers.*

EXERCISE 4.28. *Learn some other proofs for the quadratic reciprocity formula.*

EXERCISE 4.29. *Fill in all the details, for the theory of the Jacobi symbols.*

EXERCISE 4.30. *Fill in all the details for the theory of Kronecker symbols too.*

EXERCISE 4.31. *And do not forget to learn a bit about more general symbols too.*

EXERCISE 4.32. *Learn more about the Hadamard conjecture, and its story.*

EXERCISE 4.33. *Learn also about the circulant Hadamard matrices.*

As bonus exercise, find a nice book on algebraic number theory, and start reading.

Part II

Real numbers

*Do you remember
Before the rain came down
You were so full of life
So bring that right back around*

CHAPTER 5

Real numbers

5a. Real numbers

We have certainly used a bit real numbers in the above, as everyone does, but time now to get more in detail into their definition, and philosophy. Among others, we will see how the knowledge of the real numbers tells us more about \mathbb{Q} , and even about \mathbb{Z} or \mathbb{N} .

But, what are the real numbers? Intuition tells us that these are the numbers missing from \mathbb{Q} , and based on this, let us formulate the following definition, due to Dedekind:

DEFINITION 5.1. *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:

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

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:

$$A_r = \{p \in \mathbb{Q} \mid p \leq r\} \quad , \quad B_r = \{q \in \mathbb{Q} \mid q > r\}$$

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 5.2. *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 $x \rightarrow -x$, 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:

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

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

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 5.3. *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 indeed our equation in the following way:

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

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

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

$$x = \pm a_1 \dots a_n . b_1 b_2 b_3 \dots$$

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

PROOF. This is something which requires a number of verifications, 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 $\pm a_1 \dots a_n$ and then each of the digits b_1, b_2, b_3, \dots 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 $x = \pm a_1 \dots a_n . b_1 b_2 b_3 \dots$ corresponds to a real number $x \in \mathbb{R}$, and that with the convention $\dots b999 \dots = \dots (b+1)000 \dots$, 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 $[x]$:

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

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 $b_1 \in \{0, \dots, 9\}$, and it is clear that such an approximation exists, simply by comparing x with the numbers $0.0, 0.1, \dots, 0.9$. 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 $0.b_10, 0.b_11, \dots, 0.b_19$. 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 $x = 0.b_1b_2b_3\dots$ 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:

$$A_x = \bigcup_{n \geq 1} \left\{ p \in \mathbb{Q} \mid p \leq 0.b_1 \dots b_n \right\}$$

(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 $x = \dots b999\dots$, 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 $\dots b999\dots$. Going back to the construction in (6), we are led to the conclusion that we have the following equality:

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

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

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

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

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 5.5. *The following happen, in relation with $\mathbb{Q} \rightarrow \mathbb{R}$:*

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

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

(1) We already know, from chapter 2, that the rationals \mathbb{Q} are countable. Regarding now the reals \mathbb{R} , assume by contradiction that $[0, 1]$ is countable, listed as follows, and with the convention that the writings of type $\dots 999\dots$ are avoided:

$$\begin{aligned}x_1 &= 0.a_1a_2a_3\dots \\x_2 &= 0.b_1b_2b_3\dots \\x_3 &= 0.c_1c_2c_3\dots \\&\dots\end{aligned}$$

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

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

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

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

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

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

(3) Finally, regarding the probability assertion, in order to avoid some troubles, we will prove that the probability for a real number $x \in [0, 1]$ to be rational is 0. So, let us write the rational numbers $r \in [0, 1]$ in the form of a sequence $r_1, r_2, r_3\dots$ as follows:

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

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

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

Here we have used the formula $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = 1$, which comes by dividing $[0, 1]$ into half, then one of the halves into half again, and so on, and then saying in the end that the pieces that we have must sum up to 1. Thus $P \leq c$, and so $P = 0$, as desired. \square

5b. Limits, series

Time now to do some analysis, among others in order to better understand what was said in the above, regarding the reals. At the start of everything, we have:

DEFINITION 5.6. *We say that a sequence $\{x_n\}_{n \in \mathbb{N}} \subset \mathbb{R}$ converges to $x \in \mathbb{R}$ when:*

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

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

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

As a basic example for all this, we have:

PROPOSITION 5.7. *We have $1/n \rightarrow 0$.*

PROOF. This is obvious, but let us prove it by using Definition 5.6. We have:

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

Thus we can take $N = [1/\varepsilon] + 1$ in Definition 5.6, and we are done. \square

There are many other examples, and more on this in a moment. Going ahead with more theory, let us complement Definition 5.6 with:

DEFINITION 5.8. *We write $x_n \rightarrow \infty$ when the following condition is satisfied:*

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

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

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

PROPOSITION 5.9. *We have $n^2 \rightarrow \infty$.*

PROOF. As before, this is obvious, but let us prove it using Definition 5.8. We have:

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

Thus we can take $N = [\sqrt{K}] + 1$ in Definition 5.8, and we are done. \square

We can unify and generalize Proposition 5.7 and Proposition 5.9, as follows:

PROPOSITION 5.10. *We have the following convergence,*

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

with $n \rightarrow \infty$.

PROOF. This follows indeed by using the same method as in the proof of Proposition 5.7 and Proposition 5.9, first for a rational, and then for a real as well. \square

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

THEOREM 5.11. *The following happen:*

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

PROOF. Again, this is something elementary, coming from definitions:

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

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

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

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

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

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

- (4) This is clear indeed from definitions. \square

Here are as well some general rules for computing limits:

THEOREM 5.12. *The following happen, with the conventions $\infty + \infty = \infty$, $\infty \cdot \infty = \infty$, $1/\infty = 0$, and with the conventions that $\infty - \infty$ and $\infty \cdot 0$ are undefined:*

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

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

- (1) This is something which is obvious from definitions.
- (2) This follows indeed from the following estimate:

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

- (3) This follows indeed from the following estimate:

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

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

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

PROPOSITION 5.13. *The $n \rightarrow \infty$ limits of quotients of polynomials are given by*

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

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

PROOF. The first assertion comes from the following computation:

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

As for the second assertion, this comes from Proposition 5.10. □

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

DEFINITION 5.14. *Given a sequence $\{x_n\}_{n \in \mathbb{N}} \subset \mathbb{R}$, we let*

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

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

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

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

THEOREM 5.15. *A sequence x_n converges, with finite limit $x \in \mathbb{R}$, precisely when*

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

called Cauchy condition.

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

The above result is quite interesting, and as an application, we can further clarify what was said before in this chapter, in relation with the reals. Indeed, we have:

THEOREM 5.16. *\mathbb{R} is the completion of \mathbb{Q} , in the sense that it is the space of Cauchy sequences over \mathbb{Q} , identified when the virtual limit is the same, in the sense that:*

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

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

PROOF. Let us denote indeed the completion operation by $X \rightarrow \bar{X} = C_X / \sim$, where C_X is the space of Cauchy sequences over X , and \sim is the above equivalence relation. Since by Theorem 5.15 any Cauchy sequence $(x_n) \in C_{\mathbb{Q}}$ has a limit $x \in \mathbb{R}$, we obtain $\bar{\mathbb{Q}} = \mathbb{R}$. As for the equality $\bar{\mathbb{R}} = \mathbb{R}$, this is clear again by using Theorem 5.15. \square

With this discussed, let us get now into series. Let us start with:

DEFINITION 5.17. *Given numbers $x_0, x_1, x_2, \dots \in \mathbb{R}$, we write*

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

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

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

THEOREM 5.18. *We have the “geometric series” formula*

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

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

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

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

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

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

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

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

THEOREM 5.19. *We have the following formula:*

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

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

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

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

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

(2) Regarding now the second assertion, we have that at $a = 1$, and so at any $a \leq 1$. Thus, it remains to prove that at $a > 1$ the series converges. Let us first discuss the case $a = 2$, which will prove the convergence at any $a \geq 2$. The trick here is as follows:

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

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

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

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

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

We therefore obtain the following estimate for the Riemann sum:

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

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

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

THEOREM 5.20. *We have the following convergence result:*

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

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

PROOF. Both the assertions follow from Theorem 5.19, as follows:

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

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

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

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

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

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

THEOREM 5.21. *The following hold, with the converses of (1) and (2) being wrong, and with (3) not holding when the assumption $x_n \geq 0$ is removed:*

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

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

(1) We know that $\sum_n x_n$ converges when $S_k = \sum_{n=0}^k x_n$ converges. Thus by Cauchy we have $x_k = S_k - S_{k-1} \rightarrow 0$, and this gives the result. As for the simplest counterexample for the converse, this is $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots = \infty$, coming from Theorem 5.19.

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

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

As for the simplest counterexample for the converse, this is $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots < \infty$, coming from Theorem 5.20, coupled with $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots = \infty$ from (1).

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

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

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

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

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

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

THEOREM 5.22. *The following happen, and in all cases, the situation where $c = 1$ is indeterminate, in the sense that the series can converge or diverge:*

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

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

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

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

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

Finally, generalizing the first assertion in Theorem 5.20, we have:

THEOREM 5.23. *If $x_n \searrow 0$ then $\sum_n (-1)^n x_n$ converges.*

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

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

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

Very nice all this, good math that we learned. Let us mention, however, that the above is not the end of the story, because most of the above results extend, and even look better in extended form, to the complex numbers. More on this later in this book.

5c. The number e

Time now to get into some truly interesting mathematics, featuring the famous number $e = 2.71828\dots$. In order to establish our results, we will need the following fact:

THEOREM 5.24. *We have the following inequality, for any $a_1, \dots, a_n \geq 0$,*

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

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

PROOF. This can be done in several steps, as follows:

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

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

(2) But with this, we can prove our inequality at $n = 4$ too, as follows:

$$\frac{a+b+c+d}{4} \geq \frac{\sqrt{ab} + \sqrt{cd}}{2} \geq \sqrt[4]{abcd}$$

(3) Next, we can prove our inequality, in the same way, at $n = 8$, then at $n = 16$, and so on. Thus, as a conclusion, we know how to prove the result at any $n = 2^s$.

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

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

Now pick $s \in \mathbb{N}$ such that $n \leq 2^s$, and let us complete our series a_1, \dots, a_n with $2^s - n$ copies of m . The arithmetic mean stays the same, and by using (3) we obtain:

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

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

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

THEOREM 5.25. *We have the following convergence*

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

where $e = 2.71828\dots$ is a certain number.

PROOF. This is something quite tricky, as follows:

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

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

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

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

In practice, this gives the following inequality:

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

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

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

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

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

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

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

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

In practice, this gives the following inequality:

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

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

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

And, by inverting this inequality that we found, we obtain, as desired:

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

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

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

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

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

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

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

More generally now, we have the following result:

THEOREM 5.26. *We have the following formula,*

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

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

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

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

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

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

THEOREM 5.27. *We have the following formula,*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

THEOREM 5.28. *We have the following formula,*

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

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

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

(1) In order to prove the result, consider the following function, whose convergence is clear by using our various criteria for series from Theorem 5.22:

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

Our claim, which is the key one, is that we have the following formula:

$$f(x + y) = f(x)f(y)$$

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

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

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

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

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

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

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

Observe that we used in the above proof a few intuitive things about functions, and their continuity properties. We will be back to this, with details, later on.

Finally, still talking about e and e^x , here is some more, regarding them:

FACT 5.29. *The following happen:*

- (1) e is the unique number satisfying $e^x \simeq 1 + x$, for $x \simeq 0$.
- (2) e^x is the unique function mapping $0 \rightarrow 1$, and equal to its own slope.

Obviously, some non-trivial things are going on here, and we will not attempt to explain all this, right now. More on this later, when talking calculus, but in the meantime, knowing all this is something useful, philosophically speaking. Indeed, we can see from the above that if there is an interesting number in advanced mathematics, that is not 1 or 2, but rather $e = 2.71828 \dots$. And more on this later, when doing analysis.

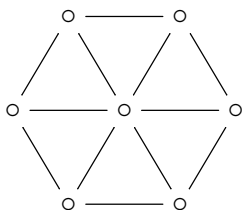
5d. The number pi

As the saying goes, there is no e without π , so let us get now into π and trigonometry. Many things can be said here, and to start with, sort of axiomatically, we have:

THEOREM 5.30. *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:

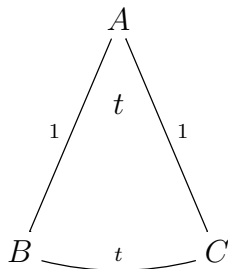
$$A = N \times \frac{HS}{2} = \frac{HP}{2} \simeq \frac{1 \times L}{2}$$

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

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 $\pi = 3.14159\dots$, but we will come back to this later, once we will have appropriate tools for dealing with such questions.

Now that we know about π , we can talk about measuring angles, as follows:

THEOREM 5.31. *We can measure angles by putting them in the middle of a circle of radius 1, and assigning to them the corresponding arc lengths:*



Equivalently, we can use twice the area of the disk slice, which equals the arc length. In this way, the multiples of 90° get converted into corresponding multiples of $\pi/2$.

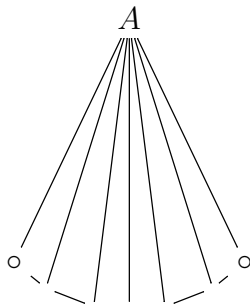
PROOF. We have two things to be proved here, as follows:

(1) First is the fact that our measuring method is indeed good, in the sense that doubling the angles will double their values, tripling the angles will triple their values, and so on. But this is something which is plainly obvious, so done with this.

(2) And then, there is the claim that we have the following formula, with on the left the area of the disk slice ABC , and on the right the arc length BC :

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

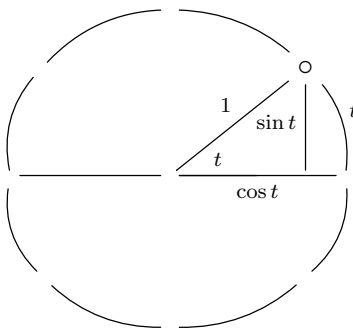
But this is something which is clear for isosceles triangles having altitude 1, and then our disk slice can be approximated by unions of such isosceles triangles, as follows:



Thus, we conclude that our area formula holds indeed, as desired. \square

As a question now that you might have, is doing the above, namely replacing our beloved 90° coming from astronomy by that crazy $\pi/2$ number, a good thing? In answer, our new convention shines when it comes to trigonometry. We first have:

THEOREM 5.32. *The sine and cosine of any $t \in \mathbb{R}$ can be computed according to*



with the convention that inverted segments count as negatives.

PROOF. This is something self-explanatory, coming from the usual definition of \sin and \cos , and with the comment that all this certainly brings some new light on \sin , \cos . For instance, we can see that the correspondence $t \rightarrow \sin t$, which is now something geometric, is such that $\sin t \simeq t$ for t small. And, more on this in a moment. \square

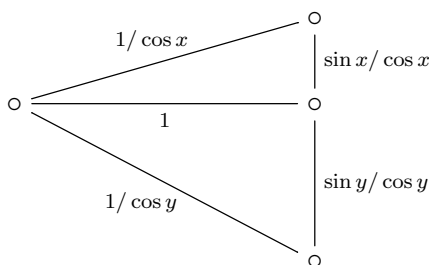
In order to further comment on this, we will need the following standard fact:

THEOREM 5.33. *The sine and cosine are subject to the following formulae:*

- (1) $\sin^2 x + \cos^2 x = 1$.
- (2) $\sin(x + y) = \sin x \cos y + \cos x \sin y$.
- (3) $\cos(x + y) = \cos x \cos y - \sin x \sin y$.

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

- (1) This is Pythagoras' theorem, that we know well from chapter 4.
- (2) Consider the following picture, consisting of a length 1 line segment, with angles x, y drawn on each side, and with the lengths being 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.

- (3) By using $\sin(x + y)$ we can deduce a formula for $\cos(x + y)$, as follows:

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

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

Let us get now into an interesting question, namely estimating the trigonometric functions. For this purpose, we can use the above formulae for sums, which allow us to transport our approximation questions around $t = 0$. Let us record this, as follows:

FACT 5.34. *We can use the standard formulae for the sums of angles in order to transport our various approximation questions around $t = 0$.*

So, let us get now to what happens with trigonometric functions around 0. And here, coming as a main result regarding the trigonometric functions, we have:

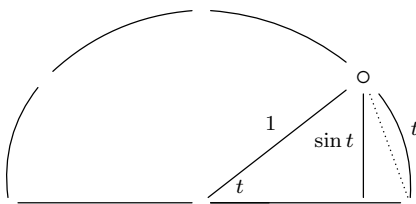
THEOREM 5.35. *We have the following estimates, for small angles*

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

coming from our new convention for numeric angles.

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

(1) The general idea is that the estimates are both clear from our circle picture for the angles, and trigonometric functions. Indeed, the picture for the sine is:



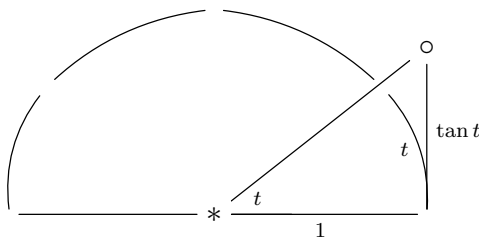
But with this picture in hand, we can draw the dotted segment above, having length $2 \sin(t/2)$, and with Pythagoras on the left, followed by shortest distance between two points being achieved by that dotted segment on the right, we obtain:

$$\sin t \leq 2 \sin(t/2) \leq t$$

(2) As an alternative proof, we can compare the area of the above isosceles triangle with the area of the disk slice, which gives right away the following estimate:

$$\frac{\sin t}{2} \leq \frac{t}{2}$$

(3) Regarding now the tangent, again for $t \in [0, \pi/2]$, the picture is as follows:



But with this picture in hand, by comparing areas we get, as desired:

$$\frac{t}{2} \leq \frac{\tan t}{2}$$

Summarizing, done, one way or another, with both the inequalities proved. \square

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

THEOREM 5.36. *The following happen, for small angles, again coming from our new convention for numeric angles, and best justifying this convention:*

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

PROOF. This can be indeed established as follows:

(1) This is clear indeed on the circle, by arguing like in the previous proof, and we will leave the various details here as an instructive exercise. Equivalently, this follows from $\sin t \leq t \leq \tan t$, by using $\tan t = \sin t / \cos t \simeq \sin t$, coming from $\cos t \simeq 1$.

(2) This comes from (1), and from Pythagoras. Indeed, knowing $\sin t \simeq t$, when looking for a quantity $\cos t$ making the Pythagoras formula $\sin^2 t + \cos^2 t = 1$ hold, we are led, via some quick thinking, to the formula $\cos t \simeq 1 - t^2/2$, according to:

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

(3) This is again clear on the circle, or simply follows from (1,2), by dividing. □

And with this, good news, we have all tools in our bag for doing some truly tough calculus, with the real numbers. Hang on, difficult material to come, right next.

5e. Exercises

This was a very basic analysis chapter, and as exercises here, we have:

EXERCISE 5.37. *Quickly rewrite the theory of \mathbb{R} , introduced as completion of \mathbb{Q} .*

EXERCISE 5.38. *Then, rewrite the theory of \mathbb{R} , introduced via the decimal form.*

EXERCISE 5.39. *Then, rewrite again, with \mathbb{R} introduced via binary numbers.*

EXERCISE 5.40. *Clarify what we said about rearranging terms of $\sum_k (-1)^k/k$.*

EXERCISE 5.41. *Fill in all the details, regarding the properties of $f(x) = \sum_k x^k/k!$.*

EXERCISE 5.42. *Further meditate on the two definitions of π , and their equivalence.*

EXERCISE 5.43. *Review, if needed, the definition and basic properties of \sin, \cos, \tan .*

EXERCISE 5.44. *Try to improve our estimates for \sin, \cos, \tan , at higher order.*

As bonus exercise, find 100 exercises on sequences and series, and solve them.

CHAPTER 6

Functions, calculus

6a. Functions, continuity

Before getting back to arithmetic, armed with our new knowledge of the reals \mathbb{R} , it is wiser to do a bit more analysis, this time in relation with the functions $f : \mathbb{R} \rightarrow \mathbb{R}$, which can be something very useful. Let us start with the following key notion:

DEFINITION 6.1. A function $f : \mathbb{R} \rightarrow \mathbb{R}$, or more generally $f : X \rightarrow \mathbb{R}$, with $X \subset \mathbb{R}$ being a subset, is called continuous when, for any $x_n, x \in X$:

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

Also, we say that $f : X \rightarrow \mathbb{R}$ is continuous at a given point $x \in X$ when the above condition is satisfied, for that point x .

Regarding the basic examples of continuous functions, there are many of them, and we will discuss them in a moment, once we will have some basic tools, in order to prove that this or that function is continuous or not, without much pain. As a matter, however, of having a first illustration for Definition 6.1, let us record the following fact:

PROPOSITION 6.2. The basic power functions, namely

$$f(x) = x^k$$

with $k \in \mathbb{N}$, are all continuous.

PROOF. According to Definition 6.1, we want to prove that we have:

$$x_n \rightarrow x \implies x_n^k \rightarrow x^k$$

Which looks quite clear, but you might want a rigorous proof for this. In answer:

(1) A first method is by using the results from chapter 5 regarding the sequences. To be more precise, we know from there that the following formula holds:

$$\lim_{n \rightarrow \infty} x_n y_n = \lim_{n \rightarrow \infty} x_n \lim_{n \rightarrow \infty} y_n$$

But with $x_n = y_n$, this leads to the following formula:

$$\lim_{n \rightarrow \infty} x_n^2 = \left(\lim_{n \rightarrow \infty} x_n \right)^2$$

Obviously, we can iterate this method, and so for any $k \in \mathbb{N}$, we have:

$$\lim_{n \rightarrow \infty} x_n^k = \left(\lim_{n \rightarrow \infty} x_n \right)^k$$

But now, assuming $x_n \rightarrow x$ as above, this formula gives, as desired:

$$\lim_{n \rightarrow \infty} x_n^k = x^k$$

(2) Thus, result proved, but let us try as well a second method, which is less conceptual, but is instructive too. Our idea here will be to use no idea at all. Obviously, in order to solve our question, we must estimate quantities of type $(x+t)^k - x^k$, with t small. But we can do this with the binomial formula, which gives, for $|t| \leq 1$:

$$\begin{aligned} |(x+t)^k - x^k| &= \left| \sum_{s=1}^k \binom{k}{s} x^{k-s} t^s \right| \\ &\leq \sum_{s=1}^k \binom{k}{s} |x|^{k-s} |t|^s \\ &\leq |t| \sum_{s=1}^k \binom{k}{s} |x|^{k-s} \\ &\leq |t| (1 + |x|)^k \end{aligned}$$

Now assume $x_n \rightarrow x$. We can then write $x_n = x + t_n$, and by choosing our $n \gg 0$ as to have $|t_n| \leq 1$, we can use the above estimate, which gives:

$$|x_n^k - x^k| \leq |t_n| (1 + |x|)^k$$

Now since we have $t_n \rightarrow 0$, we obtain from this $x_n^k \rightarrow x^k$, as desired. \square

Getting back now to the general theory, and to Definition 6.1 as stated, many things can be said. To start with, there are many other equivalent formulations of the notion of continuity, with a well-known, useful, and much feared one, being as follows:

THEOREM 6.3. *A function $f : X \rightarrow \mathbb{R}$ is continuous when*

$$\forall x \in X, \forall \varepsilon > 0, \exists \delta > 0, |x - y| < \delta \implies |f(x) - f(y)| < \varepsilon$$

holds.

PROOF. Let us prove this, with no fear. According to Definition 6.1, in order for our function f to be continuous, the following must happen, for any $x \in X$:

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

Now when reminding what convergence of a sequence exactly means, for both the convergences $x_n \rightarrow x$ and $f(x_n) \rightarrow f(x)$, we are led to the conclusion in the statement. \square

Next, we have the following useful theoretical result regarding continuity:

THEOREM 6.4. *If f, g are continuous, then so are:*

- (1) $f + g$.
- (2) fg .
- (3) f/g .
- (4) $f \circ g$.

PROOF. Before anything, we should mention that the claim is that (1-4) hold indeed, provided that at the level of domains and ranges, the statement makes sense. For instance in (1,2,3) we are talking about functions having the same domain, and with $g(x) \neq 0$ for the needs of (3), and there is a similar discussion regarding (4).

(1) The claim here is that if both f, g are continuous at a point x , then so is the sum $f + g$. But this is clear from the similar result for sequences, namely:

$$\lim_{n \rightarrow \infty} (x_n + y_n) = \lim_{n \rightarrow \infty} x_n + \lim_{n \rightarrow \infty} y_n$$

(2) Again, the statement here is similar, and the result follows from:

$$\lim_{n \rightarrow \infty} x_n y_n = \lim_{n \rightarrow \infty} x_n \lim_{n \rightarrow \infty} y_n$$

(3) Here the claim is that if both f, g are continuous at x , with $g(x) \neq 0$, then f/g is continuous at x . In order to prove this, observe that by continuity, $g(x) \neq 0$ shows that $g(y) \neq 0$ for $|x - y|$ small enough. Thus we can assume $g \neq 0$, and with this assumption made, the result follows from the similar result for sequences, namely:

$$\lim_{n \rightarrow \infty} \frac{x_n}{y_n} = \frac{\lim_{n \rightarrow \infty} x_n}{\lim_{n \rightarrow \infty} y_n}$$

(4) Here the claim is that if g is continuous at x , and f is continuous at $g(x)$, then $f \circ g$ is continuous at x . But this is clear, coming from:

$$\begin{aligned} x_n \rightarrow x &\implies g(x_n) \rightarrow g(x) \\ &\implies f(g(x_n)) \rightarrow f(g(x)) \end{aligned}$$

Alternatively, let us prove this as well by using that scary ε, δ condition from Theorem 6.3. So, let us pick an arbitrary $\varepsilon > 0$. We want in the end to have something of type $|f(g(x)) - f(g(y))| < \varepsilon$, so we must first use that ε, δ condition for the function f . So, let us start in this way. Since f is continuous at $g(x)$, we can find $\delta > 0$ such that:

$$|g(x) - z| < \delta \implies |f(g(x)) - f(z)| < \varepsilon$$

On the other hand, since g is continuous at x , we can find $\gamma > 0$ such that:

$$|x - y| < \gamma \implies |g(x) - g(y)| < \delta$$

Now by combining the above two inequalities, with $z = g(y)$, we obtain:

$$|x - y| < \gamma \implies |f(g(x)) - f(g(y))| < \varepsilon$$

Thus, the composition $f \circ g$ is continuous at x , as desired. □

As a first consequence of the above result, of great interest, we have:

THEOREM 6.5. *Any polynomial $P \in \mathbb{R}[X]$, regarded as function*

$$P : \mathbb{R} \rightarrow \mathbb{R} \quad , \quad x \rightarrow P(x)$$

is continuous, over its whole domain.

PROOF. This follows from Theorem 6.4, and from the extra fact, which is something trivial, and that I forgot to mention there, that if a function f is continuous, and $\lambda \in \mathbb{R}$ is a scalar, then the function λf is continuous too. Indeed, since any polynomial $P \in \mathbb{R}[X]$ can be obtained by starting with the function $x \rightarrow x$, which is continuous, and making sums, products, and multiplication by scalars, P must be continuous, as stated. \square

Back to theory, as a main result now regarding continuity, we have:

THEOREM 6.6 (Intermediate value property). *Given a continuous function*

$$f : [a, b] \rightarrow \mathbb{R}$$

its image is a closed bounded interval, $Im(f) = [c, d]$.

PROOF. It is convenient to make the convention that all intervals are by definition closed and bounded, and with $[a, b]$ denoting the numbers comprised between a, b , regardless on whether $a \leq b$, or $a > b$. With this convention, the proof goes as follows:

(1) Let us first prove that f takes its intermediate values, in the sense that any $u \in [f(a), f(b)]$ belongs to $Im(f)$. In order to do so, observe that we have:

$$[f(a), f(b)] \subset \left[f(a), f\left(\frac{a+b}{2}\right) \right] \cup \left[f\left(\frac{a+b}{2}\right), f(b) \right]$$

Thus $u \in [f(a), f(b)]$ must belong to one of the intervals on the right. Now by repeating this procedure, indefinitely, we are led to a certain decreasing sequence of closed intervals I_k , with the size of these intervals halving at each step, such that:

$$u \in f(I_k) \quad , \quad \forall k$$

Now consider the limiting point x of the intervals that we found, given by:

$$\bigcap_k I_k = \{x\}$$

And with this, we are done. Indeed, by continuity of our function f at this point x , from the above condition, $u \in f(I_k)$ for any k , we obtain that we have, as desired:

$$f(x) = u$$

(2) Next, we can apply what found to the restriction $f' : [a', b'] \rightarrow \mathbb{R}$ of our function to any interval $[a', b'] \subset [a, b]$, and with a bit of thinking here, that we will leave as an exercise, this shows that the image of our function is indeed an interval, as stated. \square

In practice now, Theorem 6.6 as stated is something quite compact, and in view of applications, the various findings there are best recalled as follows:

THEOREM 6.7. *The following happen for a continuous function $f : [a, b] \rightarrow \mathbb{R}$:*

- (1) *f takes all intermediate values between $f(a), f(b)$.*
- (2) *f has a minimum and maximum on $[a, b]$.*
- (3) *If $f(a), f(b)$ have different signs, $f(x) = 0$ has a solution.*

PROOF. All these statements follow indeed from Theorem 6.6. □

As a standard application now of our continuity technology, we have:

THEOREM 6.8. *Any polynomial $P \in \mathbb{R}[X]$ of odd degree has a root. In particular, an arbitrary degree 3 polynomial must decompose as*

$$P(x) = (x - r)Q(x)$$

with $r \in \mathbb{R}$ and with Q being of degree 2, having 0 or 2 roots.

PROOF. Here the first assertion comes from Theorem 6.7 (3), because if P has odd degree, taken with leading coefficient $c > 0$, it must travel as follows:

$$P(-\infty) = -\infty \quad \rightsquigarrow \quad P(\infty) = \infty$$

As for the second assertion, this is something self-explanatory, coming from this. □

6b. Derivatives, rules

Time now for calculus, as developed by Newton, Leibnitz and others. The basic idea is very simple. We are interested in the functions $f : \mathbb{R} \rightarrow \mathbb{R}$, and when f is continuous at a point x , this means that we can write an approximation formula as follows:

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

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

DEFINITION 6.9. *A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is called differentiable at x when*

$$f'(x) = \lim_{t \rightarrow 0} \frac{f(x + t) - f(x)}{t}$$

called derivative of f at that point x , exists.

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

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

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

PROPOSITION 6.10. *If f is differentiable at x , then f must be continuous at x . However, the converse is not true, a basic counterexample being $f(x) = |x|$, at $x = 0$.*

PROOF. The first assertion is something that we already know, from the above. As for the second assertion, regarding $f(x) = |x|$, this is something quite clear on the picture of f , but let us prove this mathematically, based on Definition 6.9. We have:

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

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

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

Thus, the limit in Definition 6.9 does not converge, as desired. \square

At the level of the general theory, we can solve our original question, as follows:

THEOREM 6.11. *Assuming that f is differentiable at x , we have:*

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

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

PROOF. Assuming that f is differentiable at x , we have, in the $t \rightarrow 0$ limit:

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

Thus, by multiplying by t , we are led to the formula in the statement. \square

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

PROPOSITION 6.12. *We have the differentiation formula*

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

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

PROOF. In the case $p \in \mathbb{N}$ we can use the binomial formula, which gives:

$$(x+t)^p = x^p + px^{p-1}t + \dots + t^p \simeq x^p + px^{p-1}t$$

Next, for $p \in \mathbb{Q}$, we can write $p = m/n$, with $m \in \mathbb{N}$ and $n \in \mathbb{Z}$, and we have:

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

Finally, the general case $p \in \mathbb{R}$ follows too, via a standard continuity argument. \square

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

PROPOSITION 6.13. *We have the following results:*

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

PROOF. Regarding the sine, the computation here goes as follows:

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

The computation for the cosine is quite similar, as follows:

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

For the exponential, the derivative can be computed as follows:

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

As for the logarithm, the computation here is standard too, as follows:

$$(\log x)' = \lim_{t \rightarrow 0} \frac{\log(x+t) - \log x}{t} = \lim_{t \rightarrow 0} \frac{\log(1+t/x)}{t} = \frac{1}{x}$$

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

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

THEOREM 6.14. *We have the following formulae:*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

PROPOSITION 6.15. *The derivatives of fractions are given by:*

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

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

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

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

PROOF. This statement is written a bit upside down, and for the proof it is better to proceed backwards. To be more precise, by using $(x^p)' = px^{p-1}$ and Theorem 6.14 (3), we obtain the third formula. Then, with $p = -1$, we obtain from this the second formula. And finally, by using this second formula and Theorem 6.14 (2), we obtain:

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

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

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

PROPOSITION 6.16. *We have the following formulae,*

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

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

PROOF. For \tan , we have the following computation:

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

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

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

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

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

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

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

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

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

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

At the theoretical level now, further building on Theorem 6.11, we have:

THEOREM 6.17. *The local minima and maxima of a differentiable function $f : \mathbb{R} \rightarrow \mathbb{R}$ appear at the points $x \in \mathbb{R}$ where:*

$$f'(x) = 0$$

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

PROOF. The first assertion follows from the formula in Theorem 6.11, written as:

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

Indeed, saying that our function f has a local maximum at $x \in \mathbb{R}$ means that there exists a number $\varepsilon > 0$ such that the following happens:

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

We conclude that we must have $f'(x)t \geq 0$ for sufficiently small t , and since this small t can be both positive or negative, this shows that we have, as claimed:

$$f'(x) = 0$$

The discussion for local maxima is similar. Finally, in what regards the last assertion, the standard counterexample here is $f(x) = x^3$, taken at $x = 0$. \square

In practice, the method in Theorem 6.17 is best recalled as follows:

ALGORITHM 6.18. *In order to find the minimum and maximum of $f : [a, b] \rightarrow \mathbb{R}$:*

- (1) *Compute the derivative f' .*
- (2) *Solve the equation $f'(x) = 0$.*
- (3) *Add a, b to your set of solutions.*
- (4) *Compute $f(x)$, for all your solutions.*
- (5) *Compute the min/max of all these $f(x)$ values.*
- (6) *Then this is the min/max of your function.*

Which is quite interesting, the general problem in science being that of finding the extrema of functions. As another important consequence of Theorem 6.17, we have:

THEOREM 6.19. *Assuming that $f : [a, b] \rightarrow \mathbb{R}$ is differentiable, we have*

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

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

PROOF. In the case $f(a) = f(b)$, the result, called Rolle theorem, states that we have $f'(c) = 0$ for some $c \in (a, b)$, and follows from Theorem 6.17. Now in what regards our statement, due to Lagrange, this follows from Rolle, applied to the following function:

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

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

As a key consequence of Theorem 6.19, of great practical interest, we have:

THEOREM 6.20. *For a differentiable function we have*

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

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

PROOF. This is indeed something self-explanatory, coming from Theorem 6.19. \square

The above result has countless applications, and here is a particularly useful one:

THEOREM 6.21. *We have the generalized binomial formula*

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

with the generalized binomial coefficients being given by

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

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

PROOF. The series in the statement f converges indeed at $|x| < 1$, thanks to our convergence results from chapter 5, which apply. Also, we have the following formula:

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

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

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

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

As another application of Theorem 6.20, more philosophical this time, we have:

THEOREM 6.22. *The exponential function is the unique solution of*

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

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

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

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

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

6c. Higher derivatives

The derivative theory that we have is already quite powerful, and can be used in order to solve all sorts of interesting questions, but with a bit more effort, we can do better. Indeed, at a more advanced level, we can come up with the following notion:

DEFINITION 6.23. *We say that $f : \mathbb{R} \rightarrow \mathbb{R}$ is twice differentiable if it is differentiable, and its derivative $f' : \mathbb{R} \rightarrow \mathbb{R}$ is differentiable too. The derivative of f' is denoted*

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

and is called second derivative of f .

So, what is the second derivative good for? Good question, and answer to this coming in a moment. But before that, let us get a bit familiar with f'' . We first have:

INTERPRETATION 6.24. *The second derivative $f''(x) \in \mathbb{R}$ is the number which:*

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

So, this is the truth about the second derivative, making it clear that what we have here is a very interesting notion. In practice now, the situation is as follows:

(1) This is something very intuitive, which follows from the usual interpretation of the derivative, both as a growth rate, and a slope.

(2) This is a reformulation of (1), by using the intuitive meaning of the word “acceleration”, with the relevant physics equation, due to Newton, being $a = x''$.

(3) This is something more subtle, and very useful too, which is of statistical nature, and that we will clarify with some mathematics, in a moment.

(4) This is something quite subtle too, and again very useful in practice, that we will again clarify with some mathematics, later in this section.

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

PROPOSITION 6.25. *The second derivatives of the basic functions are as follows:*

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

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

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

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

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

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

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

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

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

(2) Regarding now the counterexample, this is straightforward, as follows:

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

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

Getting back now to theory, we will need the following standard result:

PROPOSITION 6.26. *The 0/0 type limits can be computed according to the formula*

$$\frac{f(x)}{g(x)} \simeq \frac{f'(x)}{g'(x)}$$

called L'Hôpital's rule.

PROOF. The above formula holds indeed, as an application of the general first derivative theory from before, which gives, in the situation from the statement:

$$\frac{f(x+t)}{g(x+t)} \simeq \frac{f(x) + f'(x)t}{g(x) + g'(x)t} = \frac{f'(x)t}{g'(x)t} = \frac{f'(x)}{g'(x)}$$

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

We can now formulate the following key result, improving Theorem 6.11:

THEOREM 6.27. *Given a twice differentiable function $f : \mathbb{R} \rightarrow \mathbb{R}$, we have:*

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

That is, f is approximately locally quadratic.

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

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

We must have $a = f(x)$, and we also know from before that $b = f'(x)$ is the correct choice for the coefficient of t . Thus, our approximation must be as follows:

$$f(x+t) \simeq f(x) + f'(x)t + ct^2$$

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

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

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

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

In order to prove now that this formula holds indeed, we can use L'Hôpital's rule. Indeed, by using this, if we denote by $\varphi(t) \simeq P(t)$ the formula to be proved, we have:

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

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

The above result substantially improves Theorem 6.11, and there are many applications of it. As a first such application, justifying Interpretation 6.24 (3), we have:

PROPOSITION 6.28. *Intuitively speaking, the second derivative $f''(x) \in \mathbb{R}$ computes how much different is $f(x)$, compared to the average of $f(z)$, with $z \simeq x$.*

PROOF. This is something a bit heuristic, but which is good to know. Let us write the formula in Theorem 6.27, as such, and with $t \rightarrow -t$ too:

$$\begin{aligned} f(x+t) &\simeq f(x) + f'(x)t + \frac{f''(x)}{2} t^2 \\ f(x-t) &\simeq f(x) - f'(x)t + \frac{f''(x)}{2} t^2 \end{aligned}$$

By making the average, we obtain the following formula:

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

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

Back to rigorous mathematics, we can improve as well Theorem 6.17, as follows:

THEOREM 6.29. *The local minima and local maxima of a twice differentiable function $f : \mathbb{R} \rightarrow \mathbb{R}$ appear at the points $x \in \mathbb{R}$ where*

$$f'(x) = 0$$

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

PROOF. The first assertion is something that we already know. As for the second assertion, we can use the formula in Theorem 6.27, which in the case $f'(x) = 0$ reads:

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

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

As before with Theorem 6.17, the above result is not the end of the story with the mathematics of the local minima and maxima, because things are undetermined when:

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

And more on this later. Moving on, as a main concrete application now of the second derivative, which is related to Interpretation 6.24 (4), we have the following result:

THEOREM 6.30. *Given a convex function $f : \mathbb{R} \rightarrow \mathbb{R}$, we have the following Jensen inequality, for any $x_1, \dots, x_N \in \mathbb{R}$, and any $\lambda_1, \dots, \lambda_N > 0$ summing up to 1,*

$$f(\lambda_1 x_1 + \dots + \lambda_N x_N) \leq \lambda_1 f(x_1) + \dots + \lambda_N f(x_N)$$

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

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

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

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

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

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

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

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

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

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

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

$$f(\lambda_1 x_1 + \dots + \lambda_N x_N) \leq \lambda_1 f(x_1) + \dots + \lambda_N f(x_N)$$

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

$$f''(x) \geq 0$$

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

Back now to the general theory of the derivatives, and their theoretical applications, we can further develop our basic approximation method, at order 3, at order 4, and so on. Let us start with something nice and intuitive, as follows:

FACT 6.31. *The third derivative f''' can be thought of as being the jerk of f .*

Here the terminology comes from real life and classical mechanics, where the jerk is by definition the derivative of the acceleration, and so is the second derivative of the speed, and so is the third derivative of the position, according to the following formulae:

$$j = a' = v'' = x'''$$

As before with second derivatives, many other things can be said. Let us also record the formulae of the third derivatives of the basic functions, which are as follows:

PROPOSITION 6.32. *The third derivatives of the basic functions are as follows:*

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

PROOF. The various formulae in the statement all follow indeed from the various formulae for the second derivatives established before. \square

Getting now to the fourth derivatives, things are less intuitive here, in what regards the interpretation, but we can nevertheless do some computations, as follows:

THEOREM 6.33. *The fourth derivatives of the basic functions are as follows:*

- (1) $(x^p)'''' = p(p-1)(p-2)(p-3)x^{p-4}$.
- (2) $\sin'''' = \sin$.
- (3) $\cos'''' = \cos$.
- (4) $\exp'''' = \exp$.
- (5) $\log''''(x) = -6/x^4$.

PROOF. The various formulae in the statement follow indeed from those before. \square

With this discussed, and getting back now to our usual approximation business, the ultimate result on the subject, called Taylor formula, is as follows:

THEOREM 6.34. *Assuming that $f : \mathbb{R} \rightarrow \mathbb{R}$ is n times differentiable, we have*

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

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

PROOF. As before at order 2, in the proof of Theorem 6.27, when trying to best approximate the function $\varphi(t) = f(x+t)$ by a degree n polynomial $P(t)$, the natural conditions to be imposed are those stating that P and φ should match at $t = 0$, at the level of the actual value, of the derivative, second derivative, and so on up the n -th derivative. Thus, we are led to the approximation in the statement, namely:

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

In order to prove now that this approximation holds indeed, we can use L'Hôpital's rule, applied several times, as in the proof of Theorem 6.27. To be more precise, if we denote by $\varphi(t) \simeq P(t)$ the approximation to be proved, we have:

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

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

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

PROPOSITION 6.35. *For a polynomial of degree n , the Taylor approximation*

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

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

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

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

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

In relation now with the local extrema, we have the following result:

THEOREM 6.36. *Assuming that $f : \mathbb{R} \rightarrow \mathbb{R}$ is n times differentiable, and*

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

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

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

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

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

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

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

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

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

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

$$f(x+t) \simeq f(x) + \frac{f'''(x)}{6} t^3$$

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

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

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

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

As a first illustration for the Taylor formula, we have:

THEOREM 6.37. *We have the generalized binomial formula*

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

for any $p \in \mathbb{R}$, with the generalized binomial coefficients being given by

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

as a Taylor series approximation, and also as an exact formula, at any $|t| < 1$.

PROOF. Consider indeed the following function $f(x) = x^p$. We have:

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

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

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

But this is exactly our binomial formula. As for the fact that this formula is indeed exact, at any $|t| < 1$, this is something that we already know, from Theorem 6.21. \square

As another basic illustration for the Taylor series, we have:

THEOREM 6.38. *We have the following formulae,*

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

as well as the following formulae,

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

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

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

(1) Regarding the Taylor series statements, we can use here:

$$(\sin x)' = \cos x \quad , \quad (\cos x)' = -\sin x \quad , \quad (e^x)' = e^x \quad , \quad (\log x)' = x^{-1}$$

Thus, by computing, we are led to the various Taylor series in the statement.

(2) Next, regarding the fact that our Taylor approximation formulae are in fact exact, we already know that this happens for \exp . In order to deal now with \sin and \cos , we have the following trick, using a formal number satisfying $i^2 = -1$:

$$(\cos t + i \sin t)' = -\sin t + i \cos t = i(\cos t + i \sin t)$$

Thus $f(t) = \cos t + i \sin t$ satisfies $f' = if$, and by arguing like in the proof of Theorem 6.22, we conclude that we have $f(t) = e^{it}$. But this does the job, because:

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

Of course, I can hear you screaming that this is not rigorous. In answer, this is in fact rigorous, when assuming a better knowledge of $i^2 = -1$. More on this in chapter 9.

(3) Finally, regarding \log , if we set $f(1+x) = \sum_k (-1)^{k+1} x^k / k$, we have the following computation, using the generalized binomial formula, with exponent $-r$:

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

Thus f satisfies the functional equation of \log , and the slope at 1 being the correct one, we can proceed as in chapter 5 for the exponential, and we get $f = \log$, as stated. \square

6d. Integrals, primitives

Done with calculus? Not yet, because we still have to discuss one interesting topic, related in a subtle way to the derivatives, namely integration. Let us start with:

DEFINITION 6.39. *The integral of a continuous function $f : [a, b] \rightarrow \mathbb{R}$, denoted*

$$\int_a^b f(x)dx$$

is the area below the graph of f , signed $+$ where $f \geq 0$, and signed $-$ where $f \leq 0$.

Here it is of course understood that the above area can be computed, but let us not bother with such obvious things. Instead, following Riemann, let us formulate:

THEOREM 6.40. *We have the Riemann integration formula,*

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

which can serve as a definition for the integral.

PROOF. This is standard, by drawing rectangles. We have indeed the following formula, which can stand as a definition for the signed area below the graph of f :

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

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

As yet another interpretation of the integral, let us formulate as well:

THEOREM 6.41. *The integral of a function $f : [a, b] \rightarrow \mathbb{R}$ is given by*

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

where $A(f)$ is the average of f over the interval $[a, b]$.

PROOF. This is something coming from Definition 6.39, area being side times average height. Alternatively, this is something which certainly comes from Theorem 6.40. □

Time perhaps for some computations? As a first result, we have:

THEOREM 6.42. *We have the integration formula*

$$\int_a^b x^p dx = \frac{b^{p+1} - a^{p+1}}{p+1}$$

valid for any $p \in \mathbb{N}$.

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

(1) By linearity we can assume that our interval $[a, b]$ is of the form $[0, c]$, and the formula that we want to establish is as follows:

$$\int_0^c x^p dx = \frac{c^{p+1}}{p+1}$$

Moreover, we can further assume $c = 1$, and by expressing the left term as a Riemann sum, we are in need of the following estimate, in the $N \rightarrow \infty$ limit:

$$1^p + 2^p + \dots + N^p \simeq \frac{N^{p+1}}{p+1}$$

(2) So, let us try to prove this. At $p = 1, 2, 3$ we have the following well-known formulae, coming via combinatorics, geometry, or pedestrian recurrence, doing the job:

$$1 + 2 + \dots + N = \frac{N(N+1)}{2} \simeq \frac{N^2}{2}$$

$$1^2 + 2^2 + \dots + N^2 = \frac{N(N+1)(2N+1)}{6} \simeq \frac{N^3}{3}$$

$$1^3 + 2^3 + \dots + N^3 = \frac{N^2(N+1)^2}{4} \simeq \frac{N^4}{4}$$

(3) Next, with some patience, we can deal in the same way with $p = 4, 5, 6$. What we get, after lots of numerics, then guessing the answer, based on these numerics, and finally proving the formulae by recurrence, once we have them, is as follows:

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

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

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

(4) The problem is now, how to deal with the general case, $p \in \mathbb{N}$. In answer, we have the following general formula, which is provable by recurrence:

$$1^p + 2^p + \dots + N^p = \frac{1}{p+1} \sum_{k=0}^p (-1)^k \binom{p+1}{k} B_k N^{p+1-k}$$

To be more precise here, the coefficients B_k in this formula are the Bernoulli numbers, which are certain rationals, defined by the following recurrence relation:

$$\sum_{k=0}^m \binom{m+1}{k} B_k = \delta_{m0}$$

(5) In practice now, let us see how these Bernoulli numbers look like. By using the above recurrence formula for them, at $m = 0, 1, 2, 3, 4, \dots$, we obtain:

$$\begin{aligned} B_0 &= 1 \implies B_0 = 1 \\ B_0 + 2B_1 &= 0 \implies B_1 = -1/2 \\ B_0 + 3B_1 + 3B_2 &= 0 \implies B_2 = 1/6 \\ B_0 + 4B_1 + 6B_2 + 4B_3 &= 0 \implies B_3 = 0 \\ B_0 + 5B_1 + 10B_2 + 10B_3 + 5B_4 &= 0 \implies B_4 = -1/30 \\ &\vdots \end{aligned}$$

(6) Getting back now to our business, the formula in (4) is compatible with those before at $p = 1, \dots, 6$, with the data for the first few Bernoulli numbers being:

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

(7) Now the point is that the formula in (4) does the job, because we get:

$$1^p + 2^p + \dots + N^p \simeq \frac{N^{p+1}}{p+1}$$

And we will leave more learning about this, Bernoulli numbers, as an exercise. \square

Quite nice all this, but such things remain a bit amateurish. At a more advanced level, the point is that the integrals are related to derivatives, as follows:

THEOREM 6.43. *We have the following formula, for any differentiable function F ,*

$$\int_a^b F'(x)dx = F(b) - F(a)$$

called fundamental theorem of calculus.

PROOF. To start with, given a continuous function $f : [a, b] \rightarrow \mathbb{R}$, by integrating $\min f \leq f \leq \max f$ we obtain the following formula, called mean value property:

$$\exists c \in [a, b] \quad , \quad \int_a^b f(x)dx = (b-a)f(c)$$

Next, this mean value property shows that we have the following implication:

$$I(x) = \int_a^x f(s)ds \implies I' = f$$

Now given $F : \mathbb{R} \rightarrow \mathbb{R}$ as in the statement, by using this with $f = F'$, we obtain $I' = F'$. Since $I(a) = 0$, this reads $F(x) = I(x) + F(a)$, and with $x = b$ we get:

$$F(b) = \int_a^b F'(x)dx + F(a)$$

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

The above result, which is something very powerful, can be used as follows:

THEOREM 6.44. *Given a function $f : \mathbb{R} \rightarrow \mathbb{R}$, let us call primitive of it any function $F : \mathbb{R} \rightarrow \mathbb{R}$ satisfying $F' = f$. Then, the following happen:*

- (1) *The primitive, if it exists, is uniquely determined up to a scalar.*
- (2) *We have $\int_a^b f(x)dx = \left[F\right]_a^b$, with the convention $[F]_a^b = F(b) - F(a)$.*

PROOF. Here the first assertion is something trivial, coming from Theorem 6.20, and the second assertion is a reformulation of our main result, Theorem 6.43. \square

Here are as well some supplementary formulae, which must be known as well.

THEOREM 6.45. *We have the integration by parts, and change of variable formula,*

$$\int_a^b (f'g + fg')(x)dx = [fg]_a^b$$

$$\int_a^b f(x)dx = \int_{\varphi^{-1}(a)}^{\varphi^{-1}(b)} f(\varphi(t))\varphi'(t)dt$$

with our usual convention $[F]_a^b = F(b) - F(a)$, for the first formula.

PROOF. The integration by parts formula comes indeed by integrating the Leibnitz rule, and the change of variable formula comes by integrating the chain rule. \square

And with this, good news, we have all the needed calculus tools in our pocket.

6e. Exercises

Welcome to calculus, such a joy to have you here, and as exercises, we have:

EXERCISE 6.46. *Clarify all the details, in the proof of $(x^p)' = px^{p-1}$.*

EXERCISE 6.47. *Compute the derivatives of all trigonometric functions.*

EXERCISE 6.48. *Learn more about Rolle, Lagrange and the mean value property.*

EXERCISE 6.49. *Compute the second derivatives of all trigonometric functions.*

EXERCISE 6.50. *Further work on the various interpretations of the second derivative.*

EXERCISE 6.51. *Fill in all the details, for what we said on the convex functions.*

EXERCISE 6.52. *Learn more about higher derivatives, and the Taylor formula.*

EXERCISE 6.53. *Compute various areas and volumes, as many as you can.*

As bonus exercise, compute the Taylor series of all trigonometric functions. Enjoy.

CHAPTER 7

Primes, revised

7a. Euler formula

Good news, with our accumulated analysis knowledge, we can go back now to arithmetic, and more specifically to the prime numbers. Following Euler, we have:

THEOREM 7.1. *We have the following formula, with P being the set of primes,*

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

implying $|P| = \infty$.

PROOF. By using the factorization theorem for integers, we have:

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

Next, the logarithm of the product on the right can be estimated as follows:

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

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

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

As for the second assertion, namely $|P| = \infty$, which is of course something that we know well, since chapter 1, via a standard proof by contradiction, this is clear from what we have, because the sum $\sum_{p \in P} 1/p$ being infinite, so must be the indexing set P . \square

The Euler formula and its proof are something of utter beauty, suggesting doing an enormous amount of things. Here is a brief account, of what can be done:

(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, which is as follows:

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

Many things can be done here, one of the conclusions being that the N -th prime $\pi(N)$ satisfies $\pi(N) \sim N/\log N$. And more on this, on several occasions, in what follows.

(2) Still talking analysis, an interesting observation, by Erdős, coming from his own work on 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 progressions. And the point is that this is a very difficult and fascinating problem, with the case $S = P$ being settled only recently, in the mid 2000s, 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. In fact, this is a “local-global” formula, in the local sense suggested by cat, in chapter 3.

(4) Among others, 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, after learning some more analysis.

Obviously, many things going on here. Let us summarize this discussion as follows:

CONCLUSION 7.2. *Modern number theory, be that algebraic or analytic, comes from the Euler formula, and its ramifications.*

Which sounds quite good, not that we understand yet all this, still a long way to go, but we seem to have now a decent philosophical level, in number theory.

Getting to work now, what to start with? I would say, with (1) above, everything else looking quite complicated. So, getting to this, improvements of Theorem 7.1, we have:

THEOREM 7.3. *We have the following formula, with sum over primes,*

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

for the partial Euler sums.

PROOF. This is something quite straightforward, as follows:

(1) As before in the proof of Theorem 7.1, we can use the factorization theorem for integers, which gives by summing the following estimate:

$$\prod_{p < N} \left(1 - \frac{1}{p}\right)^{-1} = \prod_{p < N} \left(1 + \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^3} + \dots\right) > \sum_{n=1}^{N-1} \frac{1}{n}$$

(2) Also as before, the log of the product on the left can be estimated as follows:

$$\begin{aligned} - \sum_{p < N} \log \left(1 - \frac{1}{p}\right) &= \sum_{p < N} \frac{1}{p} + \frac{1}{2p^2} + \frac{1}{3p^3} + \frac{1}{4p^4} + \dots \\ &< \sum_{p < N} \frac{1}{p} + \frac{1}{2p^2} + \frac{1}{2p^3} + \frac{1}{2p^4} + \dots \\ &= \sum_{p < N} \frac{1}{p} + \frac{1}{2} \sum_{p < N} \frac{1}{p(p-1)} \\ &< \sum_{p < N} \frac{1}{p} + \frac{1}{2} \sum_{n=2}^{\infty} \frac{1}{n(n-1)} \\ &= \sum_{p < N} \frac{1}{p} + \frac{1}{2} \end{aligned}$$

(3) Now comes the new ingredient, a standard calculus computation, as follows:

$$\sum_{n=1}^{N-1} \frac{1}{n} > \int_1^N \frac{1}{x} dx = \int_1^N (\log x)' dx = \log N$$

(4) By putting now everything together, we obtain the following estimate:

$$\sum_{p < N} \frac{1}{p} + \frac{1}{2} > \log \left[\prod_{p < N} \left(1 - \frac{1}{p} \right)^{-1} \right] > \log \log N$$

We are therefore led to the conclusion in the statement. \square

Ready for some more analysis? As explained after Theorem 7.1, further improving the Euler formula is a key problem, which is definitely worth more study. And with a bit of work, we can further improve what we have, in the following way:

THEOREM 7.4 (update). *We have the following formula, with sum over primes,*

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

and the $1/2$ constant on the right can be improved to $\log(\pi^2/6) = 0.49770\dots$

PROOF. This is something more tricky, and with improving our constant 0.5 into $0.49770\dots$ being a good thing, worth doing, believe me here, the idea being as follows:

(1) The first assertion is something that we know, from Theorem 7.3, and a quick look at what we did in the proof there reveals four $>$ signs, which appear as follows:

$$\begin{aligned} \sum_{p < N} \frac{1}{p} + \frac{1}{2} &> \sum_{p < N} \frac{1}{p} + \frac{1}{2p^2} + \frac{1}{2p^3} + \frac{1}{2p^4} + \dots \\ &> \log \left[\prod_{p < N} \left(1 - \frac{1}{p} \right)^{-1} \right] \\ &> \log \left(\sum_{n=1}^{N-1} \frac{1}{n} \right) \\ &> \log \log N \end{aligned}$$

In principle, we can improve all these four estimates, if we want to. However, we would like to use instead a different technique, which is quite instructive.

(2) The point indeed is that we have a rival method, based by using the factorization $n = p_1 \dots p_k m^2$, with p_i distinct primes. This factorization gives:

$$\begin{aligned}
 \sum_{n=1}^{N-1} \frac{1}{n} &< \prod_{p < N} \left(1 + \frac{1}{p}\right) \sum_{m=1}^N \frac{1}{m^2} \\
 &< \prod_{p < N} \exp\left(\frac{1}{p}\right) \sum_{m=1}^{\infty} \frac{1}{(m-1/2)(m+1/2)} \\
 &= \exp\left(\sum_{p < N} \frac{1}{p}\right) \sum_{m=1}^{\infty} \frac{1}{m-1/2} - \frac{1}{m+1/2} \\
 &= 2 \exp\left(\sum_{p < N} \frac{1}{p}\right)
 \end{aligned}$$

Now by taking the logarithm, this gives the following formula:

$$\sum_{p < N} \frac{1}{p} + \log 2 > \log \left(\sum_{n=1}^{N-1} \frac{1}{n} \right)$$

We therefore obtain the following estimate, for our sum:

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

(3) However, $\log 2 = 0.69314..$ does not improve our $1/2$ constant, and we have to be more careful with our telescoping in (2). By separating the first term, we get closer:

$$\sum_{m=1}^{\infty} \frac{1}{m^2} < 1 + \frac{2}{3} = \frac{5}{3} \quad , \quad \log \left(\frac{5}{3} \right) = 0.51082..$$

By separating the first two terms, we get even closer, but still not there:

$$\sum_{m=1}^{\infty} \frac{1}{m^2} < 1 + \frac{1}{4} + \frac{2}{5} = \frac{33}{20} \quad , \quad \log \left(\frac{33}{20} \right) = 0.50077..$$

However, with the first three terms separated, what we get is a win:

$$\sum_{m=1}^{\infty} \frac{1}{m^2} < 1 + \frac{1}{4} + \frac{1}{9} + \frac{2}{7} = \frac{415}{252} \quad , \quad \log \left(\frac{415}{252} \right) = 0.49884..$$

(4) In practice now, in order to finish this discussion, in a neat way, we can invoke the Basel formula, due to Euler, which is however something quite complicated:

$$\sum_{m=1}^{\infty} \frac{1}{m^2} = \frac{\pi^2}{6}$$

To be more precise, the original proof by Euler of this formula was as follows, making some clever manipulations on the Taylor series expansion of $\sin x/x$, based on the fact that the zeroes of this function appear precisely at the points $x = k\pi$, with $k \in \mathbb{Z}$:

$$\begin{aligned}
 \frac{\sin x}{x} &= 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \dots \\
 &= \left(1 - \frac{x}{\pi}\right) \left(1 + \frac{x}{\pi}\right) \left(1 - \frac{x}{2\pi}\right) \left(1 + \frac{x}{2\pi}\right) \dots \\
 &= \left(1 - \frac{x^2}{\pi^2}\right) \left(1 - \frac{x^2}{4\pi^2}\right) \left(1 - \frac{x^2}{9\pi^2}\right) \dots \\
 &= 1 - \frac{1}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} x^2 + \dots
 \end{aligned}$$

This needs of course more justification, and later, Weierstrass came with the fix, by showing that the above factorization of $\sin x/x$ holds indeed, as a general property of the analytic functions. In what concerns us, we will be back to this later in this book, with a full proof the Basel formula, using Fourier analysis, and the Parseval formula.

(5) Anyway, with the Basel formula in hand, and by putting everything together, we conclude that the factorization method $n = p_1 \dots p_k m^2$ gives:

$$\begin{aligned}
 \exp\left(\sum_{p < N} \frac{1}{p}\right) \times \frac{\pi^2}{6} &> \prod_{p < N} \left(1 + \frac{1}{p}\right) \times \frac{\pi^2}{6} \\
 &= \prod_{p < N} \left(1 + \frac{1}{p}\right) \sum_{m=1}^N \frac{1}{m^2} \\
 &> \sum_{n=1}^{N-1} \frac{1}{n} \\
 &> \int_1^N \frac{1}{x} dx \\
 &= \log N
 \end{aligned}$$

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

As an interesting feature of the above proof, this brings us into the zeta function:

DEFINITION 7.5. *We can talk about the Riemann zeta function,*

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

with $s > 1$, and with this, the Basel formula above reads $\zeta(2) = \pi^2/6$.

To be more precise, the fact that the above Riemann series converges indeed at $s > 1$ is something that we know well, since chapter 5. As for the Basel formula $\zeta(2) = \pi^2/6$ and its various generalizations, more on this later in this book.

Motivated by this, we would like to discuss now what happens to the Euler estimate, when adding an exponent $s \in \mathbb{R}$ there. Let us start with a basic result, as follows:

PROPOSITION 7.6. *The Euler estimate can be generalized into*

$$\sum_{p < N} \frac{1}{p^s} > \log \left(\int_1^N \frac{1}{x^s} dx \right) - \frac{1}{2} \sum_{n=2}^{N-1} \frac{1}{n^s(n^s - 1)}$$

with the above integral given by the formula

$$\int_1^N \frac{1}{x^s} dx = \begin{cases} \frac{N^{1-s}-1}{1-s} & \text{if } s \neq 1 \\ \log N & \text{if } s = 1 \end{cases}$$

involving now a real parameter $s \in \mathbb{R}$, with exactly the same proof.

PROOF. By using the unique factorization $n = p_1^{a_1} \dots p_k^{a_k}$, as before, we have:

$$\begin{aligned} \prod_{p < N} \left(1 - \frac{1}{p^s} \right)^{-1} &= \prod_{p < N} \left(1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \frac{1}{p^{3s}} + \dots \right) \\ &> \sum_{n=1}^{N-1} \frac{1}{n^s} \\ &> \int_1^N \frac{1}{x^s} dx \end{aligned}$$

But the logarithm of the product on the left can be estimated as follows:

$$\begin{aligned} - \sum_{p < N} \log \left(1 - \frac{1}{p^s} \right) &= \sum_{p < N} \frac{1}{p^s} + \frac{1}{2p^{2s}} + \frac{1}{3p^{3s}} + \frac{1}{4p^{4s}} + \dots \\ &< \sum_{p < N} \frac{1}{p^s} + \frac{1}{2p^{2s}} + \frac{1}{2p^{3s}} + \frac{1}{2p^{4s}} + \dots \\ &= \sum_{p < N} \frac{1}{p^s} + \frac{1}{2} \sum_{p < N} \frac{1}{p^s(p^s - 1)} \\ &< \sum_{p < N} \frac{1}{p^s} + \frac{1}{2} \sum_{n=2}^{N-1} \frac{1}{n^s(n^s - 1)} \end{aligned}$$

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

In the case $s > 1$, which is the one where $\zeta(s) < \infty$, we obtain in this way:

THEOREM 7.7. *We have the following Euler type estimate*

$$\sum_{p < N} \frac{1}{p^s} > \log \left(\frac{1 - N^{1-s}}{s-1} \right) - \frac{\zeta(2s)}{2}$$

valid for any value of the parameter $s > 1$.

PROOF. In the case $s > 1$ the estimate that we found in Proposition 7.6 gives:

$$\begin{aligned} \sum_{p < N} \frac{1}{p^s} &> \log \left(\frac{1 - N^{1-s}}{s-1} \right) - \frac{1}{2} \sum_{n=2}^{N-1} \frac{1}{n^s(n^s-1)} \\ &> \log \left(\frac{1 - N^{1-s}}{s-1} \right) - \frac{1}{2} \sum_{n=2}^{\infty} \frac{1}{n^s(n^s-1)} \\ &> \log \left(\frac{1 - N^{1-s}}{s-1} \right) - \frac{1}{2} \sum_{n=2}^{\infty} \frac{1}{(n-1)^{2s}} \\ &> \log \left(\frac{1 - N^{1-s}}{s-1} \right) - \frac{\zeta(2s)}{2} \end{aligned}$$

Here we have used the following inequality, with $\varepsilon = 1/n < 1$, which is true:

$$\begin{aligned} \frac{1}{n^s(n^s-1)} < \frac{1}{(n-1)^{2s}} &\iff (n-1)^{2s} < n^s(n^s-1) \\ &\iff \left(1 - \frac{1}{n}\right)^{2s} < 1 - \frac{1}{n^s} \\ &\iff (1 - \varepsilon)^{2s} < 1 - \varepsilon^s \\ &\iff (1 - \varepsilon)^{2s-1} < \frac{1 - \varepsilon^s}{1 - \varepsilon} \end{aligned}$$

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

It is possible to further build along the above lines, but we will leave this discussion for later, in Part IV, when talking more in detail about the Riemann zeta function.

7b. Stirling formula

In order to advance with our estimates, we are in need of more calculus input. There are several possible paths that can be taken here, and perhaps the most straightforward one is via the Stirling formula. In order to discuss this, we will need:

THEOREM 7.8. *We have the following formula,*

$$\int_{\mathbb{R}} e^{-x^2} dx = \sqrt{\pi}$$

called Gauss integral formula.

PROOF. This is something truly magic, the idea being as follows:

(1) To start with, we can certainly integrate e^{-x^2} by using the formula of the exponential series, and the primitive which is worth 0 at $x = 0$ is given by:

$$\int e^{-x^2} = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)k!}$$

However, this series is not computable, in terms of the known, familiar series.

(2) Thus, no primitive, but we can still ask for the computation of $\int_{\mathbb{R}} e^{-x^2} dx$, who knows. And here, another surprise awaits us, this is simply undoable, with bare hands, I mean all the formulae and tricks that we learned in chapter 6 fail, for this integral.

(3) Which seems to send our problem to the trash can. However, and here comes the magic, the Gauss integral can be computed by using two dimensions, as follows:

$$\begin{aligned} \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-x^2-y^2} dx dy &= 4 \int_0^{\infty} \int_0^{\infty} e^{-x^2-y^2} dx dy \\ &= 4 \int_0^{\infty} \int_0^{\infty} e^{-t^2 y^2 - y^2} y dt dy \\ &= 4 \int_0^{\infty} \int_0^{\infty} y e^{-y^2(1+t^2)} dy dt \\ &= 2 \int_0^{\infty} \int_0^{\infty} \left(-\frac{e^{-y^2(1+t^2)}}{1+t^2} \right)' dy dt \\ &= 2 \int_0^{\infty} \frac{dt}{1+t^2} \\ &= 2 \int_0^{\infty} (\arctan t)' dt \\ &= \pi \end{aligned}$$

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

Getting now to the Stirling formula, which is obviously related to arithmetic, and is related to many other things too, as we will soon discover, this is as follows:

THEOREM 7.9. *We have the Stirling formula*

$$N! \simeq \left(\frac{N}{e} \right)^N \sqrt{2\pi N}$$

valid in the $N \rightarrow \infty$ limit.

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

(1) Let us first see what we can get with Riemann sums. We have:

$$\log(N!) = \sum_{k=1}^N \log k \approx \int_1^N \log x \, dx = N \log N - N + 1$$

By exponentiating, this gives the following estimate, which is not bad:

$$N! \approx \left(\frac{N}{e}\right)^N \cdot e$$

(2) We can improve our estimate by replacing the rectangles from the Riemann sum approach to the integrals by trapezoids. In practice, this gives the following estimate:

$$\begin{aligned} \log(N!) &= \sum_{k=1}^N \log k \\ &\approx \int_1^N \log x \, dx + \frac{\log 1 + \log N}{2} \\ &= N \log N - N + 1 + \frac{\log N}{2} \end{aligned}$$

By exponentiating, this gives the following estimate, which gets us closer:

$$N! \approx \left(\frac{N}{e}\right)^N \cdot e \cdot \sqrt{N}$$

(3) In order to conclude, we must take some kind of mathematical magnifier, and carefully estimate the error made in (2). Fortunately, this mathematical magnifier exists, called Euler-Maclaurin formula, and after some tough computations, we get to:

$$N! \simeq \left(\frac{N}{e}\right)^N \sqrt{2\pi N}$$

(4) However, all this remains a bit complicated, so we would like to present now an alternative approach to (3), which also misses some details, but better does the job, explaining where the $\sqrt{2\pi}$ factor comes from. First, by partial integration we have:

$$\begin{aligned} \int_0^\infty x^N e^{-x} dx &= N \int_0^\infty x^{N-1} e^{-x} dx \\ &= N(N-1) \int_0^\infty x^{N-2} e^{-x} dx \\ &\vdots \\ &= N! \int_0^\infty e^{-x} dx \\ &= N! \end{aligned}$$

(5) Since the integrand is sharply peaked at $x = N$, as you can see by computing the derivative of $\log(x^N e^{-x})$, this suggests writing $x = N + y$, and we obtain:

$$\begin{aligned}
 \log(x^N e^{-x}) &= N \log x - x \\
 &= N \log(N + y) - (N + y) \\
 &= N \log N + N \log \left(1 + \frac{y}{N}\right) - (N + y) \\
 &\simeq N \log N + N \left(\frac{y}{N} - \frac{y^2}{2N^2}\right) - (N + y) \\
 &= N \log N - N - \frac{y^2}{2N}
 \end{aligned}$$

By exponentiating, we obtain from this the following estimate:

$$x^N e^{-x} \simeq \left(\frac{N}{e}\right)^N e^{-y^2/2N}$$

(6) Now by integrating, we obtain from this the following estimate:

$$\begin{aligned}
 N! &= \int_0^\infty x^N e^{-x} dx \\
 &\simeq \int_{-N}^N \left(\frac{N}{e}\right)^N e^{-y^2/2N} dy \\
 &\simeq \left(\frac{N}{e}\right)^N \int_{\mathbb{R}} e^{-y^2/2N} dy \\
 &= \left(\frac{N}{e}\right)^N \sqrt{2N} \int_{\mathbb{R}} e^{-z^2} dz \\
 &= \left(\frac{N}{e}\right)^N \sqrt{2\pi N}
 \end{aligned}$$

Thus, we have proved the Stirling formula, as formulated in the statement. \square

As a first basic application of the Stirling formula, let us record:

THEOREM 7.10. *We have the following estimate for binomial coefficients,*

$$\binom{N}{K} \simeq \left(\frac{1}{t^t(1-t)^{1-t}}\right)^N \frac{1}{\sqrt{2\pi t(1-t)N}}$$

in the $K \simeq tN \rightarrow \infty$ limit, with $t \in (0, 1]$. In particular we have

$$\binom{2N}{N} \simeq \frac{4^N}{\sqrt{\pi N}}$$

in the $N \rightarrow \infty$ limit, for the central binomial coefficients.

PROOF. The first formula follows indeed from Stirling, as follows:

$$\begin{aligned}
\binom{N}{K} &= \frac{N!}{K!(N-K)!} \\
&\simeq \left(\frac{N}{e}\right)^N \sqrt{2\pi N} \left(\frac{e}{K}\right)^K \frac{1}{\sqrt{2\pi K}} \left(\frac{e}{N-K}\right)^{N-K} \frac{1}{\sqrt{2\pi(N-K)}} \\
&= \frac{N^N}{K^K (N-K)^{N-K}} \sqrt{\frac{N}{2\pi K(N-K)}} \\
&\simeq \frac{N^N}{(tN)^{tN} ((1-t)N)^{(1-t)N}} \sqrt{\frac{N}{2\pi tN(1-t)N}} \\
&= \left(\frac{1}{t^t(1-t)^{1-t}}\right)^N \frac{1}{\sqrt{2\pi t(1-t)N}}
\end{aligned}$$

As for the second formula, this follows from this, at $t = 1/2$. □

As a second application now of the Stirling formula, which reveals its true geometric nature, we can estimate the volumes of the spheres. Let us start with:

THEOREM 7.11. *The volume of the unit sphere in \mathbb{R}^N is given by*

$$V = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!}$$

with the convention $N!! = (N-1)(N-3)(N-5)\dots$

PROOF. If we denote by V_N the volume of the unit sphere in \mathbb{R}^N , we have:

$$\begin{aligned}
V_N &= \int_{-1}^1 (1-x^2)^{(N-1)/2} dx \cdot V_{N-1} \\
&= 2V_{N-1} \int_0^1 (1-x^2)^{(N-1)/2} dx \\
&= 2V_{N-1} \int_0^{\pi/2} (1-\sin^2 t)^{(N-1)/2} \cos t dt \\
&= 2V_{N-1} \int_0^{\pi/2} \cos^N t dt
\end{aligned}$$

Now let us compute the integral I_N on the right. For this purpose, we can use:

$$\begin{aligned}
(\cos^N t \sin t)' &= N \cos^{N-1} t (-\sin t) \sin t + \cos^N t \cos t \\
&= N \cos^{N+1} t - N \cos^{N-1} t + \cos^{N+1} t \\
&= (N+1) \cos^{N+1} t - N \cos^{N-1} t
\end{aligned}$$

By integrating this between 0 and $\pi/2$, we obtain the following formula:

$$(N+1)I_{N+1} = NI_{N-1}$$

Thus we can compute I_N by recurrence, and we obtain in this way the following formula, where $\varepsilon(N) = 1$ if N is even, and $\varepsilon(N) = 0$ if N is odd:

$$\begin{aligned} I_N &= \frac{N-1}{N} I_{N-2} \\ &= \frac{N-1}{N} \cdot \frac{N-3}{N-2} I_{N-4} \\ &= \frac{N-1}{N} \cdot \frac{N-3}{N-2} \cdot \frac{N-5}{N-4} I_{N-6} \\ &\vdots \\ &= \frac{N!!}{(N+1)!!} I_{1-\varepsilon(N)} \end{aligned}$$

The initial data being $I_0 = \frac{\pi}{2}$ and $I_1 = 1$, we end up with the following formula:

$$I_N = \left(\frac{\pi}{2}\right)^{\varepsilon(N)} \frac{N!!}{(N+1)!!}$$

Getting back now to our sphere computation, by recurrence, using this, we have:

$$\begin{aligned} V_N &= 2^N \int_0^{\pi/2} \cos^N t dt \int_0^{\pi/2} \cos^{N-1} t dt \dots \int_0^{\pi/2} \cos t dt \\ &= 2^N \left(\frac{\pi}{2}\right)^{\varepsilon(N)+\varepsilon(N-1)+\dots+\varepsilon(1)} \frac{N!!}{(N+1)!!} \cdot \frac{(N-1)!!}{N!!} \dots \frac{1!!}{2!!} \\ &= \left(\frac{\pi}{2}\right)^{\varepsilon(N)+\varepsilon(N-1)+\dots+\varepsilon(1)} \frac{2^N}{(N+1)!!} \\ &= \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!} \end{aligned}$$

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

We can now estimate the volumes of the spheres, using Stirling, as follows:

THEOREM 7.12. *The volume of the unit sphere in \mathbb{R}^N is given by*

$$V \simeq \left(\frac{2\pi e}{N}\right)^{N/2} \frac{1}{\sqrt{\pi N}}$$

in the $N \rightarrow \infty$ limit.

PROOF. This is very standard, based on the formula in Theorem 7.11, namely:

$$V = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!}$$

(1) In the case where $N = 2K$ is even, we have the following estimate:

$$\begin{aligned} (N+1)!! &= 2^K K! \\ &\simeq \left(\frac{2K}{e}\right)^K \sqrt{2\pi K} \\ &= \left(\frac{N}{e}\right)^{N/2} \sqrt{\pi N} \end{aligned}$$

(2) As for the case where $N = 2K - 1$ is odd, here the estimate goes as follows:

$$\begin{aligned} (N+1)!! &= \frac{(2K)!}{2^K K!} \\ &\simeq \frac{1}{2^K} \left(\frac{2K}{e}\right)^{2K} \sqrt{4\pi K} \left(\frac{e}{K}\right)^K \frac{1}{\sqrt{2\pi K}} \\ &= \left(\frac{2K}{e}\right)^K \sqrt{2} \\ &= \left(\frac{N+1}{e}\right)^{(N+1)/2} \sqrt{2} \\ &= \left(\frac{N}{e}\right)^{N/2} \left(\frac{N+1}{N}\right)^{N/2} \sqrt{\frac{N+1}{e}} \cdot \sqrt{2} \\ &\simeq \left(\frac{N}{e}\right)^{N/2} \sqrt{e} \cdot \sqrt{\frac{N}{e}} \cdot \sqrt{2} \\ &= \left(\frac{N}{e}\right)^{N/2} \sqrt{2N} \end{aligned}$$

(3) Now back to the spheres, when N is even, the estimate goes as follows:

$$\begin{aligned} V &= \left(\frac{\pi}{2}\right)^{N/2} \frac{2^N}{(N+1)!!} \\ &\simeq \left(\frac{\pi}{2}\right)^{N/2} 2^N \left(\frac{e}{N}\right)^{N/2} \frac{1}{\sqrt{\pi N}} \\ &= \left(\frac{2\pi e}{N}\right)^{N/2} \frac{1}{\sqrt{\pi N}} \end{aligned}$$

(4) As for the case where N is odd, here the estimate goes as follows:

$$\begin{aligned}
 V &= \left(\frac{\pi}{2}\right)^{(N-1)/2} \frac{2^N}{(N+1)!!} \\
 &\simeq \left(\frac{\pi}{2}\right)^{(N-1)/2} 2^N \left(\frac{e}{N}\right)^{N/2} \frac{1}{\sqrt{2N}} \\
 &= \sqrt{\frac{2}{\pi}} \left(\frac{2\pi e}{N}\right)^{N/2} \frac{1}{\sqrt{2N}} \\
 &= \left(\frac{2\pi e}{N}\right)^{N/2} \frac{1}{\sqrt{\pi N}}
 \end{aligned}$$

Thus, we are led to the uniform formula in the statement. \square

7c. Mertens theorems

Back to the Euler formula and primes, we already have a number of interesting estimates. The continuation of the story involves the work of Mertens, that we would like to discuss now. Let us start with some standard analysis conventions, as follows:

DEFINITION 7.13. *We use the following notations:*

- (1) *We write $f \simeq g$ when $f - g \rightarrow 0$.*
- (2) *We write $f \cong g$ when $f - g$ is bounded.*
- (3) *We write $f \sim g$ when $f/g \rightarrow 1$.*
- (4) *We write $f \approx g$ when f/g is bounded.*

Occasionally, we will use as well the Landau $O(f)$, $o(f)$ symbols, making it for just 2 notations, instead of 4. With these conventions, the formula of Mertens is as follows:

THEOREM 7.14. *We have the following formula, with sum over primes,*

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

and with $M = 0.26149\dots$ being a constant, called Mertens constant.

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

(1) The idea is to use the following formula, which comes as usual from the unique factorization of integers, namely $n = p_1^{a_1} \dots p_k^{a_k}$, with the product being over the various prime powers $p^k \leq N$, and with the exponent $[N/p^k]$ being an integer part:

$$N! = \prod_{p^k \leq N} p^{[N/p^k]}$$

(2) By taking the logarithm, we obtain from this the following estimate:

$$\begin{aligned}
 \log N! &= \sum_{p^k \leq N} \left[\frac{N}{p^k} \right] \log p \\
 &= \sum_{p^k \leq N} \left(\frac{N}{p^k} + o(1) \right) \log p \\
 &= N \sum_{p^k \leq N} \frac{\log p}{p^k} + o(1) \sum_{p^k \leq N} \log p
 \end{aligned}$$

(3) By using $\log N! = N \log N + O(N)$, coming from Stirling, this gives:

$$\begin{aligned}
 \sum_{p^k \leq N} \frac{\log p}{p^k} &= \frac{\log N!}{N} + o\left(\frac{1}{N}\right) \sum_{p^k \leq N} \log p \\
 &= \log N + o(1) + o\left(\frac{1}{N}\right) \sum_{p^k \leq N} \log p
 \end{aligned}$$

(4) Now let us analyze the sum on the right. We have:

$$\begin{aligned}
 \sum_{p^k \leq N} \log p &\leq \sum_{p \in (N, 2N]} \log p \\
 &\leq \log \binom{2N}{N} \\
 &= O(N)
 \end{aligned}$$

(5) We conclude that the estimate in (3) can be written as follows:

$$\sum_{p^k \leq N} \frac{\log p}{p^k} = \log N + o(1)$$

(6) Now since the sum of reciprocals of squares is finite, $\sum_{k \geq 1} 1/k^2 < \infty$, we can remove all the squares from the sum on the left, and we are left with:

$$\sum_{p \leq N} \frac{\log p}{p} = \log N + o(1)$$

(7) But now by doing a partial summation, in the obvious way, this gives a formula as follows, with $M \in \mathbb{R}$ being a certain constant:

$$\sum_{p \leq N} \frac{1}{p} \simeq \log \log N + M + O\left(\frac{1}{\log N}\right)$$

(8) Now since we have $\log \log N \simeq \log \log(N+1)$, if we replace $N \rightarrow N+1$ on the left, the right term is unchanged. Thus, our formula can be written as follows:

$$\sum_{p < N} \frac{1}{p} \simeq \log \log N + M + O\left(\frac{1}{\log N}\right)$$

We are therefore led to the conclusion in the statement, and of course with the precise numerics for the Mertens constant M remaining to be justified. \square

Observe that Theorem 7.14 improves our previous Euler estimates, and no wonder here, because while the arithmetic ingredient was the same, namely the unique factorization of the integers, in what regards the analysis part, the Mertens theorem is based on the Stirling formula, which is far more advanced than $(\log x)' = 1/x$, used for Euler.

Moving on, the continuation of the story, involving Mertens, and then Meissel and others, is quite long. The above proof can be of course improved, with some technical bounds for M , and for the rate of convergence too. Also, and getting now to the point, the Mertens constant M itself, there are several interesting formulae for it.

In order to discuss this, we will need the following standard fact:

THEOREM 7.15. *The following limit converges,*

$$\gamma = \lim_{N \rightarrow \infty} \sum_{n=1}^N \frac{1}{n} - \log N$$

the result being the Euler-Mascheroni constant $\gamma = 0.57721..$

PROOF. This is indeed something very standard, which is best viewed via basic calculus. Consider indeed the following integral, with $[.]$ being the integer part:

$$I = \int_1^{\infty} \left(\frac{1}{[x]} - \frac{1}{x} \right) dx$$

As a first observation, this integral is clearly positive, $I > 0$, and it converges too, $I < \infty$, with an easy upper bound for it being obtained as follows:

$$\begin{aligned} I &= \int_1^2 \left(\frac{1}{[x]} - \frac{1}{x} \right) dx + \int_2^{\infty} \left(\frac{1}{[x]} - \frac{1}{x} \right) dx \\ &= 1 - \log 2 + \int_2^{\infty} \left(\frac{1}{[x]} - \frac{1}{x} \right) dx \\ &< 1 - \log 2 + \int_2^{\infty} \left(\frac{1}{x-1} - \frac{1}{x} \right) dx \\ &= 1 - \log 2 + \log 2 \\ &= 1 \end{aligned}$$

On the other hand, the truncations of the above integral I are given by:

$$\begin{aligned}
 I_N &= \int_1^N \left(\frac{1}{[x]} - \frac{1}{x} \right) dx \\
 &= \sum_{n=1}^{N-1} \int_n^{n+1} \left(\frac{1}{[x]} - \frac{1}{x} \right) dx \\
 &= \sum_{n=1}^{N-1} \frac{1}{n} + \log n - \log(n+1) \\
 &= \sum_{n=1}^{N-1} \frac{1}{n} - \log N
 \end{aligned}$$

Thus we have in fact $I = \gamma$, leading to the conclusion in the statement. \square

Getting back to Mertens, he proved in fact three theorems regarding the prime numbers, with Theorem 7.14, the most famous one, being his second theorem. So, time now to discuss this. Here are the three formulae established by Mertens:

THEOREM 7.16. *We have the following Mertens estimates, in the $N \rightarrow \infty$ limit,*

$$\begin{aligned}
 \sum_{p < N} \frac{\log p}{p} &\cong \log N \\
 \sum_{p < N} \frac{1}{p} &\simeq \log \log N + M \\
 \sum_{p < N} \log \left(1 - \frac{1}{p} \right) &\simeq -\log \log N - \gamma
 \end{aligned}$$

$M = 0.26149\dots$ and $\gamma = 0.57721\dots$ being the Mertens and Euler-Mascheroni constants.

PROOF. This is something more advanced, that we will not prove here, among others requiring a better knowledge of Euler-Mascheroni constant, the idea being as follows:

(1) The first Mertens formula is something quite standard, in the spirit of Theorem 7.14, and with the precise upper bound obtained by Mertens being as follows:

$$\sum_{p < N} \frac{\log p}{p} < \log N + 2$$

As usual, exercise for you, to read more about this. Alternatively, you can come back here after reading Part IV, with the Prime Number theorem there implying this.

(2) The second Mertens formula, which is the main one, is something that we know, from Theorem 7.14. However, as already mentioned, the story here is not over yet, because we still have to comment on the numerics of M . More on this in a moment.

(3) The third Mertens formula is equivalent, by exponentiating, to:

$$\prod_{p < N} \left(1 - \frac{1}{p}\right) \approx \frac{e^{-\gamma}}{\log N}$$

In relation with this, observe that we have the following formula:

$$\begin{aligned} \prod_{p < N} \left(1 - \frac{1}{p}\right) \prod_{p < N} \left(1 + \frac{1}{p}\right) &= \prod_{p < N} \left(1 - \frac{1}{p^2}\right) \\ &= \left[\prod_{p < N} \left(1 + \frac{1}{p^2} + \frac{1}{p^4} + \dots\right) \right]^{-1} \\ &\simeq \left[\sum_{n < N} \frac{1}{n^2} \right]^{-1} \\ &\approx 1 \end{aligned}$$

Thus, another equivalent formulation of the third Mertens formula is as follows:

$$\prod_{p < N} \left(1 + \frac{1}{p}\right) \approx e^{-\gamma} \log N$$

As for the proof of this, as before, exercise for you to try, or read more about this, from any classical number theory book, such as Hardy and Wright [45]. \square

Now back to the Mertens constant, we have the following formula for it:

THEOREM 7.17. *The Mertens constant is given by the formula*

$$M = \gamma + \sum_p \left(\log \left(1 - \frac{1}{p}\right) + \frac{1}{p} \right)$$

with $\gamma = 0.57721\dots$ being the Euler-Mascheroni constant.

PROOF. We know that the Mertens constant appears by definition as follows:

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

But, according to Theorem 7.16, we have as well the following formula:

$$\sum_{p < N} \log \left(1 - \frac{1}{p}\right) \simeq -\log \log N - \gamma$$

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

The above considerations eventually lead to the numeric figure from Theorem 7.14, namely $M = 0.26149\dots$ We will leave some further learning here as an exercise.

7d. Chebycheff estimates

Let us investigate now some related questions, again regarding the primes and their distribution, which look more intuitive and appealing, but which require more complicated techniques. We would like to estimate the number $\pi(N)$, constructed as follows:

DEFINITION 7.18. *We define the function $\pi : \mathbb{N} \rightarrow \mathbb{N}$ by*

$$\pi(N) = \#\{p \leq N \text{ prime}\}$$

the first few values being 0, 0, 1, 2, 2, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, ...

Many things can be said about these numbers $\pi(N)$, for instance in relation with our previous Euler estimates, and the theorems of Mertens, which can be converted into results about them. However, according to our general policy for this opening chapter on analytic number theory, let us do things slowly. To start with, we have:

PROPOSITION 7.19. *We have the following estimate,*

$$\pi(N) \geq \log \log N$$

coming from the unique factorization of integers, $n = p_1^{a_1} \dots p_k^{a_k}$.

PROOF. This is something that I learned from my pure algebra colleagues. If we denote by p_n the n -th prime number, according to the unique factorization of integers, and more specifically to the related proof of the infinity of primes, we have:

$$p_{n+1} \leq p_1 \dots p_n + 1$$

But this gives, by recurrence on n , the following estimate:

$$p_n \leq 2^{2^n}$$

In terms of the function π from Definition 7.18, this estimate reads:

$$\pi(2^{2^n}) \geq n$$

Thus, we obtain an estimate as in the statement, but shifted by 1, and with \log_2 instead of \log . However, \log_2 being for computer scientists, \log_{10} for social science, and $\log = \log_e$ for mathematics, let us stick with \log . By using $e^{n-1} > 2^n$ for $n > 3$ we can pass from \log_2 to \log , and we obtain the formula in the statement. \square

Next in line, we have the following estimate, heavily improving Proposition 7.19:

PROPOSITION 7.20. *We have the following estimate,*

$$\pi(N) \geq \frac{\log N}{\log 4}$$

coming from the unique factorization $n = p_1 \dots p_k m^2$, with p_i distinct.

PROOF. This is again something that I learned from my algebra colleagues. Consider the first n primes, denoted p_1, \dots, p_n , and let us try to compute the number $f(N)$ of integers $K \leq N$ all whose prime factors are among $\{p_1, \dots, p_n\}$. By using the factorization in the statement, that we can write as $K = SM^2$ with S square-free, we get:

$$f(N) \leq 2^n \sqrt{N}$$

On the other hand we obviously have $f(N) \geq N$, and we obtain from this:

$$N \leq 4^n \leq 4^{\pi(N)}$$

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

Getting now to a more systematic study of the problem, by using more advanced techniques, following Chebycheff, let us introduce the following related function:

DEFINITION 7.21. *The Chebycheff theta function is given by*

$$\theta(N) = \sum_{p \leq N} \log p$$

with the sum being over primes.

In what follows, the idea will be that of estimating θ , and then converting our results in terms of π . Indeed, in what regards θ , we have a nice estimate for it, as follows:

THEOREM 7.22. *We have the following estimate,*

$$\theta(N) \leq \log 16 \cdot N$$

for the Chebycheff theta function introduced above.

PROOF. This is something quite tricky, using the central binomial coefficients, that we already met in the proof of the Mertens theorem. These coefficients are as follows:

$$\binom{2n}{n} = \frac{(2n)(2n-1) \dots (n+1)}{n!}$$

Since this coefficient is obviously divisible by all primes $n < p \leq 2n$, we have:

$$\prod_{n < p \leq 2n} p < \binom{2n}{n} < (1+1)^{2n} = 4^n$$

Now in terms of the Chebycheff theta function from Definition 7.21, this gives:

$$\theta(2n) - \theta(n) < \log 4 \cdot n$$

Now by summing, we are led to the formula in the statement. \square

We can now formulate a first key theorem of Chebycheff, as follows:

THEOREM 7.23. *We have an estimate as follows,*

$$\pi(N) < C \cdot \frac{N}{\log N}$$

with C being a certain constant, $C < \log 32 + 2$.

PROOF. We have the following estimate, relating the functions θ and π :

$$\begin{aligned} \theta(n) &= \sum_{p \leq n} \log p \\ &\geq \sum_{\sqrt{n} < p \leq n} \log p \\ &\geq \log \sqrt{n} (\pi(n) - \pi(\sqrt{n})) \end{aligned}$$

Now by taking into account the estimate found in Theorem 7.22, we obtain:

$$\begin{aligned} \pi(n) &\leq \frac{2\theta(n)}{\log n} + \sqrt{n} \\ &\leq \log 32 \cdot \frac{n}{\log n} + 2 \cdot \frac{n}{\log n} \\ &= (\log 32 + 2) \frac{n}{\log n} \end{aligned}$$

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

As a second theorem of Chebycheff, going now in the other sense, we have:

THEOREM 7.24. *We have an estimate as follows,*

$$\pi(N) > c \cdot \frac{N}{\log N}$$

with c being a certain constant.

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

(1) As before in the previous proof, we use the central binomial coefficients, but written this time, and estimated, in a different way, as follows:

$$\binom{2n}{n} = \frac{n+1}{1} \cdot \frac{n+2}{2} \cdots \frac{n+1}{n} \geq 2^n$$

If we denote by v_p the exponent of each p inside this coefficient, we obtain:

$$\prod_p p^{v_p} \geq 2^n$$

Equivalently, by taking the logarithm, this gives the following formula:

$$\sum_p v_p \log p \geq n \log 2$$

(2) On the other hand, the above exponents v_p are given by the following formula, with m_p standing for the highest number such that $p^{m_p} \leq 2n$:

$$\begin{aligned} v_p &= \sum_{k=1}^{m_p} \left[\frac{2n}{p^k} \right] - \left[\frac{n}{p^k} \right] \\ &\leq m_p \\ &= \left[\frac{\log 2n}{\log p} \right] \end{aligned}$$

(3) Now by putting the estimates in (1) and (2) together, we obtain:

$$\sum_{p < 2n} \left[\frac{\log 2n}{\log p} \right] \cdot \log p \geq n \log 2$$

(4) It is convenient now to split the sum into two parts, as follows:

$$\begin{aligned} n \log 2 &\leq \sum_{p < 2n} \left[\frac{\log 2n}{\log p} \right] \cdot \log p \\ &= \sum_{p < \sqrt{2n}} \left[\frac{\log 2n}{\log p} \right] \cdot \log p + \sum_{p > \sqrt{2n}} \left[\frac{\log 2n}{\log p} \right] \cdot \log p \\ &\leq \sqrt{2n} \log 2n + \theta(2n) \end{aligned}$$

(5) We conclude from this that we have the following estimate:

$$\theta(2n) \geq n \log 2 - \sqrt{2n} \log 2n$$

But this gives a constant c such that the following happens:

$$\theta(n) > cn$$

(6) In order to conclude now, observe that we have:

$$\theta(n) = \sum_{p \leq n} \log p \leq \pi(n) \log n$$

Thus, we obtain the following estimate, for the function π itself:

$$\pi(n) \geq \frac{\theta(n)}{\log n} \geq c \cdot \frac{n}{\log n}$$

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

We can now put the two Chebycheff theorems together, as follows:

THEOREM 7.25. *We have the following estimate for the π function,*

$$\pi(N) \approx \frac{N}{\log N}$$

in the sense that the quotient of these quantities is bounded from above, and below.

PROOF. According to Theorem 7.23 and Theorem 7.24, we have:

$$c \cdot \frac{N}{\log N} \leq \pi(N) \leq C \cdot \frac{N}{\log N}$$

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

In practice, the Chebycheff estimates are strong enough in order to prove the Bertrand postulate, stating that we should have a prime number as follows:

$$N < p < 2N$$

However, the story is not over here, because we have the following conjecture:

$$\pi(N) \sim \frac{N}{\log N}$$

And here, things become fairly complicated, with this formula being known to hold indeed, as the Prime Number Theorem, but with the proofs being all complicated. We will come back to this later, towards the end of the present book.

7e. Exercises

This was our first truly advanced arithmetic chapter, and as exercises, we have:

EXERCISE 7.26. *Read the full story, including proofs too, for $\sum_n 1/n^2 = \pi^2/6$.*

EXERCISE 7.27. *Learn also, more generally, about the numbers $\zeta(2k)$, with $k \in \mathbb{N}$.*

EXERCISE 7.28. *Read some other available proofs of the Stirling formula.*

EXERCISE 7.29. *Clarify what we said above about the Mertens constant.*

EXERCISE 7.30. *Read more about the Euler-Mascheroni constant.*

EXERCISE 7.31. *Read the proof of the first and third Mertens theorems.*

EXERCISE 7.32. *Work out some further elementary estimates for $\pi(N)$.*

EXERCISE 7.33. *Learn about the various applications of the Chebycheff estimates.*

As bonus exercise, get to know more about the Riemann zeta function.

CHAPTER 8

Algebra, fields

8a. Abstract algebra

With the basics of analytic number theory discussed, time now to get into the complementary topic, the basics of algebraic number theory. Our goals will be multiple, with the aim being that of further building on the algebraic material from chapters 3-4.

Before starting, a bit of philosophy, in relation with what we know so far about numbers, both of algebraic and analytic nature. As suggested by our feline collaborator in chapter 3, in advanced number theory, be that algebraic or analytic, it is all about localization at the prime numbers $p \in \mathbb{N}$. And, we have already met three instances of this general principle, in chapters 3, 4, 7, with our conclusions being as follows:

FACT 8.1. *Localization at primes $p \in \mathbb{N}$ is a powerful technique, leading to:*

- (1) *The structure theorem for finite abelian groups, via $G = \prod_p G_p$.*
- (2) *The computation of quadratic Jacobi symbols, via $\left(\frac{a}{n}\right) = \prod_p \left(\frac{a}{p}\right)^{s_p}$.*
- (3) *The Euler formula and its consequences, $\sum_{n=1}^{\infty} \frac{1}{n} = \prod_p \left(1 - \frac{1}{p}\right)^{-1}$.*

So, the question is, can we go further in this direction, algebraically? We would like to have some further results in arithmetic, based on this localization method.

In answer now, generally speaking, the dream in arithmetic, usually concerned with solving equations $f = 0$ over the rationals, is something very simple, namely:

DREAM 8.2. *I checked that my equation $f = 0$ has solutions modulo p , for any prime p , so my equation must have solutions over \mathbb{Q} .*

As a first observation, the dream holds when f is constant, $f = c$. Indeed, ignoring a bit the differences between integers and rationals, $c = 0(p)$ for any prime p means $c = 0$, so our equation is $c = 0$, having any rational number $x \in \mathbb{Q}$ as solution.

Along the same lines, there are some other examples of very simple equations $f = 0$ for which the dream holds. However, such equations are usually so simple, that we can solve them right away, and so our dream for them is not useful. In general, for more complicated equations, our dream remains wrong, and must be fine-tuned. We are led in this way to the following update to it, which is something more modest:

HOPE 8.3. *I checked that my equation $f = 0$ has solutions with respect to any prime p , in a suitable sense, so my equation must have solutions over \mathbb{Q} .*

So, this will be our plan for what follows, doing some mathematics, as for this hope to come true. We will see that this can indeed be done, with our vague wording above “with respect to any prime p , in a suitable sense” being replaced by something very precise and mathematical, namely “over the p -adics, for any prime p ”, and with the statement itself being a deep principle in number theory, called Hasse local-global principle.

Getting to work now, many things to be discussed, and here is a to-do list:

QUESTIONS 8.4. *In relation with our arithmetic dreams and hopes:*

- (1) *What are the abstract algebra tools needed for the equations $f = 0$?*
- (2) *What is the geometry of the spaces of solutions of equations $f = 0$?*
- (3) *What happens to geometry and arithmetic, when changing the ground field?*
- (4) *What are the p -adic numbers, making the local-global principle work?*

And good list of questions this is, with each of them being very interesting, and worth reading a whole book, about it. In what concerns us, with our present chapter being scheduled for 4 sections, having about 6 pages each, we will discuss (1-4) in these 4 sections, with a very basic introduction to this. In short, do not expect much, algebraic number theory is a huge construction, and we will just explore a bit the foundations.

Getting to work for good now.. excuse me, but our expert in arithmetic and localization techniques is here, meowing something. What is it, cat?

CAT 8.5. *In my opinion, complex numbers come first.*

Okay, thanks cat, there is some truth in what you say, because as simplest example of equation $f = 0$ that we would like to solve, we have $x^2 = -1$. This being said, I am pretty much sure that we can have some good learning on (1-4) above, even without knowing what the complex numbers are. But of course, keeping the complex numbers in mind, as a new item, (5), and we will have the whole Part III below, for understanding them.

Time perhaps to get started? In relation with abstract algebra, we first have:

DEFINITION 8.6. *We have notions of rings, modules and ideals, as follows:*

- (1) *A ring R is a set with operations $+$ and \times , satisfying the usual conditions for such operations, except for $ab = ba$, and for $a \neq 0 \implies \exists a^{-1}$.*
- (2) *A module V over a ring R is a vector space, but we will call it ring, and keep the name vector spaces for the modules over fields, $R = F$.*
- (3) *An ideal $I \subset R$ is a subgroup with the left ideal property $i \in I, r \in R \implies ir \in I$, or the right ideal property $i \in I, r \in R \implies ri \in I$, or both.*

This was a quite crowded statement, but you get the point, with (1) and (2) we are sort of trying to do field and vector space mathematics, over things which are not necessarily fields and vector spaces over them, and (3) is something technical, non-field specific. At the level of examples, these abound, and we have three important ones, as follows:

(1) The integers form a ring, $R = \mathbb{Z}$, which in addition is commutative, $ab = ba$. As obvious module over \mathbb{Z} , we have the lattice $V = \mathbb{Z}^N$. Finally, since $R = \mathbb{Z}$ is commutative, the 3 notions of ideals coincide, and these are the subsets $I = a\mathbb{Z}$, with $a \in \mathbb{Z}$.

(2) The matrices over the integers form a ring, $R = M_k(\mathbb{Z})$, which is noncommutative at $k > 1$. As obvious module over $M_k(\mathbb{Z})$, we have the lattice $V = \mathbb{Z}^k$. As for the ideals, things here are a bit more complicated, but since at $k = 2$ the matrices of type $\begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}$ form a left ideal which is not a right ideal, and the matrices of type $\begin{pmatrix} a & 0 \\ b & 0 \end{pmatrix}$ form a right ideal which is not a left ideal, at least we know that our 3 types of ideals make sense.

(3) In relation with arithmetic, the integers modulo $N \in \mathbb{N}$ form a ring $R = \mathbb{Z}_N$, which is commutative. Also, the matrices over these integers modulo N form a ring $R' = M_k(\mathbb{Z}_N)$, which is noncommutative at $k > 1$. As for a preliminary study of the modules and ideals, for these two rings, I will leave it to you, and enjoy.

The question that you surely have in mind is, what are ideals good for? Answer:

PROPOSITION 8.7. *For an additive subgroup $I \subset R$, the following are equivalent:*

- (1) *I is a two-sided ideal.*
- (2) *R/I is a ring.*

PROOF. This is something which requires some thinking, as follows:

(1) Since the additive group $(R, +)$ is abelian, given an additive subgroup $I \subset R$ we can form the quotient group R/I , which is abelian too, with addition as follows:

$$(a + I) + (b + I) = (a + b + I)$$

Observe that the unit is $(0 + I) = I$, and that inverses are given by $(-a + I)$.

(2) The question is now, can we turn this abelian group R/I into a ring? Normally the multiplication can only be as follows, and with this clarifying our statement, with the condition “ R/I is a ring” there meaning, with respect to this precise multiplication:

$$(a + I)(b + I) = (ab + I)$$

(3) But, will this work. As a first observation, there is a bit of analogy here with group theory, where $H \subset G$ must be normal in order for G/H to be a group. Thus, our claim is that the ideal condition is somehow the “analogue of normality, in the ring setting”.

(4) In practice now, it is quite clear, exactly as in the group theory setting, that everything will be fine, provided that our multiplication is well-defined. And for this multiplication to be well-defined, the following condition must be satisfied:

$$(a + I) = (a' + I) , (b + I) = (b' + I) \implies (ab + I) = (a'b' + I)$$

But this amounts in the following condition to be satisfied:

$$a - a' \in I , b - b' \in I \implies ab - a'b' \in I$$

(5) Now comes the math. We have the following identity, which shows that if $I \subset R$ is a two-sided ideal, then the above condition is satisfied, and so done:

$$ab - a'b' = a(b - b') + (a - a')b'$$

(6) Conversely now, if the condition in (4) is satisfied, we have in particular:

$$i - 0 \in I , r - r \in I \implies ir - 0r \in I$$

$$r - r \in I , i - 0 \in I \implies ri - r0 \in I$$

Thus $I \subset R$ must be a two-sided ideal, and this finishes the proof. \square

Many interesting things can be said about rings, modules and ideals, especially in the commutative case. As a key result on the subject, at the basic level, we have:

THEOREM 8.8. *Assuming that R is commutative and $I \subset R$ is a maximal ideal, in the sense that it is a proper ideal, $I \neq R$, and there is no bigger proper ideal*

$$I \subset J \subset R$$

the quotient ring $F = R/I$ is a field.

PROOF. Here is the proof, and with this being guaranteed to be useful learning:

(1) Before starting, a quick example. We know that over $R = \mathbb{Z}$, the ideals are the subsets $I = p\mathbb{Z}$ with $p \in \mathbb{N}$. But such an ideal is maximal precisely when p is prime, and this is the same as asking for the quotient ring $R/I = \mathbb{Z}_p$ to be a field.

(2) In general now, assume first that R/I is a field. This means that any nonzero element of R/I is invertible, and with our usual conventions for R/I , this reads:

$$\forall a \notin I , \exists b \in R , (ab + I) = (1 + I)$$

Now assume by contradiction that $I \subset R$ is not maximal, so that we have a bigger ideal $I \subset J \subset R$. If we pick $a \in J - I$, we obtain, by the above, the following:

$$a \in J - I , b \in R , ab = 1 + i , i \in I$$

But this is contradictory, because since J is an ideal, containing I , we must have $ab, i \in J$, so we conclude that we have $1 \in J$, and so $J = R$, contradiction.

(3) Conversely, assume now that I is maximal, and assume too, by contradiction, that R/I is not a field. Then we can find a zero divisor in R/I , which reads:

$$(a + I)(b + I) = (I) , \quad a, b \notin I$$

In other words, we can find $ab \in I$ with $a, b \notin I$. But then, let us look at:

$$I \subset I + aR \subset R$$

(4) What we have in the middle is an ideal, and it is also clear, from $a \notin I$, that the inclusion on the left is proper. As for the inclusion on the right, our claim is that this is proper too. Indeed, assuming otherwise, we would have a formula as follows:

$$i + ac = 1 , \quad i \in I$$

Now by multiplying everything by b , we obtain from this:

$$ib + acb = b , \quad i \in I$$

But this is contradictory, because on the left we have $ib \in I$ and $acb = (ab)c \in I$, which gives $b \in I$, contradicting the condition $b \notin I$. Thus, our claim is proved.

(5) But this is the end of the story, because what we just proved is that what we have in (3) is indeed a proper ideal, contradicting the maximality of I , as desired. \square

Still with me, I hope, after all these abstractions, and please believe me, Theorem 8.8 is something of key importance, be that for algebraic geometry, or for arithmetic, as we will soon discover, or even for analysis, in the context of the Banach algebras.

Going ahead now with our general abstract algebra program, let us formulate:

DEFINITION 8.9. *We have notions of fields, vector spaces and algebras, as follows:*

- (1) *A field F is a field F as we know them, with in algebra parlance these being the commutative rings R with each nonzero element being invertible.*
- (2) *A vector space V over a field F is a vector space as we know them, in algebra parlance these being the modules V over a field F .*
- (3) *An algebra A over a field F is a vector space over F , with a ring multiplication operation \times , compatible with the vector space structure.*

In relation with this, we already know of course about fields, and in what regards the vector spaces, we know about them since ever, and finally, regarding algebras, we know many algebras of functions from analysis. But, thinking well, from a purely algebraic perspective, all these objects have many operations, and this is why they came at last.

As basic examples now, passed the fields F and the vector spaces V that we know well, we are left with finding interesting examples of algebras A . And here the examples abound, with all being related to geometry or analysis of some sort, as follows:

(1) First we have algebra of polynomials $A = F[X]$. This is a very basic algebra, important for geometry, and with the extra feature that it is commutative, $PQ = QP$.

(2) More generally, we have the algebra of polynomials $A = F[X_1, \dots, X_N]$. Again, this algebra is important for algebraic geometry, and is commutative, $PQ = QP$.

(3) Still talking commutative algebras, we have many of them coming from analysis, the general principle being that “functions form algebras”. More on this in a moment.

(4) We have as well the algebra of matrices $A = M_N(F)$. Again this is a very basic example, that we know well, which this time is not commutative, $ST \neq TS$.

More in detail now, getting to the various algebras of functions, we have here the following key result, bringing among others some further light on Theorem 8.8 too:

THEOREM 8.10. *Given a compact space X , the following happen:*

- (1) *The continuous functions $f : X \rightarrow \mathbb{R}$ form a real algebra $C(X)$.*
- (2) *Given $x \in X$, the functions satisfying $f(x) = 0$ form an ideal $I \subset C(X)$.*
- (3) *This ideal is maximal, and any maximal ideal $I \subset C(X)$ appears in this way.*
- (4) *In this picture, the fact that the quotient is a field, $C(X)/I = \mathbb{R}$, is clear.*

PROOF. All this is self-explanatory, the idea being as follows:

- (1) This is clear. Observe that our algebra is commutative, $fg = gf$.
- (2) This is again clear, because $f(x) = 0$ implies $(fg)(x) = 0$.
- (3) This follows from basic topology, via a suitable open cover for X .
- (4) This is clear, because $C(X) \rightarrow C(X)/I$ maps $f \rightarrow f(x) \in \mathbb{R}$. □

There are many other examples of algebras of functions, along these lines. In fact, we can even trick, and view certain algebras, which are certainly not algebras of functions, as algebras of functions too. As an example, here is a wild physics speculation:

SPECULATION 8.11. *We can view the matrix algebra $M_2(\mathbb{R})$ as being the algebra of functions on some sort of quantum space M_2 , according to the following formula:*

$$M_2(\mathbb{R}) = C(M_2)$$

This quantum space M_2 formally has $|M_2| = 4$ points, and appears as a sort of twist of $\{1, 2, 3, 4\}$. Moreover, we can integrate over M_2 , according to the formula

$$\int_{M_2} T = \frac{T_{11} + T_{22}}{2}$$

with the underlying measure being positive and of mass 1.

To be more precise here, let us be crazy, and define M_2 according to the formula $C(M_2) = M_2(\mathbb{R})$, without really knowing what we are doing. Then, we have:

$$|M_2| = \dim_{\mathbb{R}} C(M_2) = \dim_{\mathbb{R}} M_2(\mathbb{R}) = 4$$

Next, since we have $M_2(\mathbb{R}) \simeq \mathbb{R}^4$ as vector spaces, which reads $C(M_2) \simeq C(1, 2, 3, 4)$, this suggests that we should have $M_2 \sim \{1, 2, 3, 4\}$, as some sort of twisting operation. But this can be given a mathematical formulation too, the idea being that at the level of standard bases of $C(M_2) \simeq C(1, 2, 3, 4)$, the multiplication gets twisted as follows:

$$e_{ij}e_{kl} = \delta_{jk}e_{il} \quad \longleftrightarrow \quad e_j e_k = \delta_{jk}e_j$$

Finally, in what regards the last assertion, this expresses the standard fact that the normalized trace of 2×2 matrices $tr = Tr/2$ is unital and positive, in the sense that:

$$tr(1) = 1 \quad , \quad T \geq 0 \implies tr(T) \geq 0$$

Excited about this? Such things come from quantum mechanics, as developed by Heisenberg, and the above space M_2 can be given a precise mathematical sense, and is the entry point to “noncommutative geometry”. For more on this, see Connes [17].

8b. Manifolds, ideals

With the above understood, time now to get to our second objective for this chapter, namely basic algebraic geometry. At the beginning of everything, we have:

DEFINITION 8.12. *A conic is a plane algebraic curve of the form*

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

with $P \in \mathbb{R}[x, y]$ being of degree ≤ 2 .

As basic examples of conics, we have the ellipses, parabolas and hyperbolas. The simplest examples of these are as follows, with the ellipse actually being a circle:

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

Observe that, due to our assumption $\deg P \leq 2$, we have as conics some degenerate curves as well, such as lines, \emptyset , and \mathbb{R}^2 itself, coming from $\deg P \leq 1$, as follows:

$$x = 0 \quad , \quad 1 = 0 \quad , \quad 0 = 0$$

This might suggest to replace our assumption $\deg P \leq 2$ by $\deg P = 2$, but we will not do so, because $\deg P = 2$ does not rule out degenerate situations, such as:

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

In fact, what we get here are \emptyset , points, lines, and pairs of lines, so in the end, assuming $\deg P = 2$ instead of $\deg P \leq 2$ would only rule out \mathbb{R}^2 itself, which is not worth it.

Summarizing, our notion of conic from Definition 8.12 looks quite reasonable, so let us agree on this notion. Getting now to classification matters, we first have:

PROPOSITION 8.13. *Up to non-degenerate linear transformations of the plane,*

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

with $\det A \neq 0$, the conics fall into two classes, as follows:

- (1) *Non-degenerate: circles, parabolas, hyperbolas.*
- (2) *Degenerate: \emptyset , points, lines, pairs of lines, \mathbb{R}^2 .*

PROOF. As a first observation, looks like we forgot the ellipses, but via linear transformations these become circles, so things fine. As for the proof, this goes as follows:

- (1) Consider an arbitrary conic, written as follows, with $a, b, c, d, e, f \in \mathbb{R}$:

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

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

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

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

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

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

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

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

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

$$dx + ey + f = 0$$

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

- (4) So, this was the study of an arbitrary conic, and by putting now everything together, we are led to the conclusions in the statement. \square

In order now to plainly classify the conics, we have to apply linear transformations to the curves that we found in Proposition 8.13. This leads to the following result:

PROPOSITION 8.14. *The conics fall into two classes, as follows:*

- (1) *Non-degenerate: ellipses, parabolas, hyperbolas.*
- (2) *Degenerate: \emptyset , points, lines, pairs of lines, \mathbb{R}^2 .*

Also, the compact conics are \emptyset , the points, and the ellipses.

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

(1) As said above, in order to get to such a classification result, we just need to apply linear transformations to the curves that we found in Proposition 8.13. But this leaves the list there unchanged, up to the circles becoming ellipses, as stated above.

(2) In what regards the last assertion, this is clear from the first one, but since this assertion is quite interesting, let us give it a quick, independent proof as well. Consider an arbitrary conic, written as follows, with $a, b, c, d, e, f \in \mathbb{R}$:

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

Compacity rules then out the case $c \neq 0$, and our conic must be in fact:

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

But then with $a, b \neq 0$ we must have by compacity $a, b > 0$ or $a, b < 0$, and we get an ellipsis, then with $a = 0, b \neq 0$ or $a \neq 0, b = 0$ we get by compacity either \emptyset or a point, and finally with $a = b = 0$ the compacity rules out again everything, except for \emptyset . \square

As a third main result now regarding the conics, also known since the ancient Greeks, and which justifies the name “conics”, coming from “cone”, we have:

PROPOSITION 8.15. *Up to some degenerate cases, the conics are exactly the curves which appear by cutting a 2-sided cone with a plane.*

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

(1) By suitably choosing our coordinate axes (x, y, z) , we can assume that our 2-sided 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 intersect this cone with an arbitrary plane, which has an equation as follows, with $(a, b, c) \neq (0, 0, 0)$:

$$ax + by + cz = d$$

(2) 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:

$$z = 0$$

(3) 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.

(4) Alternatively, and perhaps more concretely, we can use the original coordinates, with the cone being $x^2 + y^2 = kz^2$, and compute the intersection, with the conclusion that what we get, depending on the slope of the cone, and modulo degenerate cases, is an ellipsis, hyperbola or parabola. So, by invoking Proposition 8.14, we obtain the result.

(5) Summarizing, we have proved the result, modulo some details and interesting computations which are left to you, reader. Left to you as well is the full discussion concerning degree 2 curve degeneracy vs cone cutting degeneracy, with the remark that in what regards the cone cuts, the degenerate cases are very easy to identify and list, with the list consisting of \emptyset , the points, the lines, the pairs of lines, and \mathbb{R}^2 itself. \square

All this is very nice, and as a conclusion to what we have so far about conics, we have the following statement, containing all the needed essentials:

THEOREM 8.16. *The conics, which are the algebraic curves of degree 2 in the plane,*

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

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 indeed by putting together the above results, and with the discussion concerning degeneration being left, as usual, as an instructive exercise. \square

All this is very nice, but what is next? In answer, as a natural idea, in order to generalize the conics, we can look at the zeroes of arbitrary polynomials:

$$P(x, y) = 0 \quad , \quad P \in \mathbb{R}[x, y]$$

More generally, we can look at zeroes of polynomials in arbitrary N dimensions:

$$P(x_1, \dots, x_N) = 0 \quad , \quad P \in \mathbb{R}[x_1, \dots, x_N]$$

Observe that, at $N \geq 3$, what we have is not exactly a curve, but rather some sort of $(N - 1)$ -dimensional surface, called algebraic hypersurface. Due to this, in order to have a full collection of beasts, of all possible dimensions, we must intersect such algebraic hypersurfaces. We are led in this way to zeroes of families of polynomials, as follows:

DEFINITION 8.17. *An algebraic manifold is a space of the form*

$$X = \left\{ (x_1, \dots, x_N) \in \mathbb{R}^N \mid P_i(x_1, \dots, x_N) = 0, \forall i \right\}$$

with $P_i \in \mathbb{R}[x_1, \dots, x_N]$ being a family of polynomials.

The question is now, what can we say about such manifolds? In answer, many things, and as a natural continuation of our previous work on the conics, you can:

(1) Look at plane curves of higher degree, cubics, quartics and so on. And here, all sorts of interesting curves, such as sinusoidal spirals, polynomial lemniscates or stelloids, quite often appearing as field lines in physics, are waiting for your attention.

(2) Look at degree 2 hypersurfaces in \mathbb{R}^3 or higher, which are called quadrics. And here, again, many interesting beasts, quite often related to physics, with some appearing for instance in Einstein's relativity theory, are waiting for your attention.

In short, many things to be learned here. Getting now towards arithmetic, in view of the various equations that we would like to solve over \mathbb{Q} , or perhaps over some other fields F too, it makes sense to formulate the following extension of Definition 8.17:

DEFINITION 8.18. *An algebraic manifold over a field F is a space of the form*

$$X = \left\{ (x_1, \dots, x_N) \in F^N \mid P_i(x_1, \dots, x_N) = 0, \forall i \right\}$$

with $P_i \in F[x_1, \dots, x_N]$ being a family of polynomials.

Very nice, and with this done, which obviously is something useful for arithmetic, the same question as before appears, namely, what can we say about such manifolds?

In answer to this, at the very general level, that of Definition 8.18 as stated, a fruitful idea is that of trying to understand the correspondence between algebraic manifolds and polynomials. And in relation with this, we first have the following result:

PROPOSITION 8.19. *The algebraic manifolds are the sets of the form*

$$X = \left\{ x \in F^N \mid P(x) = 0, \forall P \in I \right\}$$

with $I \subset F[x_1, \dots, x_N]$ being a radical ideal, meaning a set satisfying:

- (1) *Ideal property:* $P_i \in I, Q_i \in F[x_1, \dots, x_N] \implies \sum_i P_i Q_i \in I$.
- (2) *Radicality property:* $P^k \in I \implies P \in I$.

PROOF. This is indeed something self-explanatory, by taking I to be the set of all polynomials $P \in F[x_1, \dots, x_N]$ vanishing on X , which certainly satisfies (1,2). \square

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

THEOREM 8.20 (Hilbert basis theorem). *Any ideal of polynomials*

$$I \subset F[x_1, \dots, x_N]$$

is finitely generated, $I = (P_1, \dots, P_k)$, for some $P_i \in F[x_1, \dots, x_N]$.

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

(1) Following Emmy Noether, let us call a ring R Noetherian when any ideal $I \subset R$ is finitely generated. Equivalently, any increasing sequence of ideals $I_1 \subset I_2 \subset \dots$ must stabilize, in the sense that we must have $I_n = I_{n+1} = \dots$, for some $n \in \mathbb{N}$.

(2) We want to prove that $F[x_1, \dots, x_N]$ is Noetherian, and we will do this by recurrence on N . Since $R = F$ is clearly Noetherian, as being a field, we are left with proving the recurrence step. And, for this purpose, we will prove something which is a bit more general, namely that if a ring R is Noetherian, then so is the ring $R[X]$.

(3) We do this by contradiction. So, assume that R is Noetherian, and that $R[X]$ is not Noetherian, so that we have an ideal $I \subset R[X]$ which is not finitely generated.

(4) In order to find a contradiction, let us pick $P_1 \in I$ of minimal degree $d_1 \in \mathbb{N}$, then $P_2 \in I/(P_1)$ of minimal degree $d_2 \in \mathbb{N}$, then $P_3 \in I/(P_1, P_2)$ of minimal degree $d_3 \in \mathbb{N}$, and so on. Since our ideal $I \subset R[X]$ was assumed to be not finitely generated, this procedure will not stop, and we obtain an increasing sequence, as follows:

$$d_1 \leq d_2 \leq d_3 \leq \dots$$

(5) Now let $a_i \in R$ be the leading coefficient of each P_i , and set $J = (a_1, a_2, \dots) \subset R$. Since R was assumed to be Noetherian, we can find $n \in \mathbb{N}$ such that $J = (a_1, \dots, a_n)$. Thus, we have a formula as follows, for certain scalars $\lambda_i \in R$:

$$a_{n+1} = \sum_{i=1}^n \lambda_i a_i$$

(6) With this done, consider the following polynomial, with $\lambda_i \in R$ as above:

$$Q = \sum_{i=1}^n \lambda_i X^{d_{n+1}-d_i} P_i$$

This polynomial satisfies then $Q \in (P_1, \dots, P_n)$, and has the same leading coefficient as $P_{n+1} \notin (P_1, \dots, P_n)$. Thus, the following polynomial has degree $< d_{n+1}$:

$$P_{n+1} - Q \in I/(P_1, \dots, P_n)$$

But this is a contradiction, as desired, and this finishes the proof. \square

In practice, Theorem 8.20 is best remembered geometrically, as follows:

THEOREM 8.21. *The algebraic manifolds $X \subset F^N$ are precisely the intersections*

$$X = S_1 \cap \dots \cap S_k$$

with $S_i \subset F^N$ being hypersurfaces.

PROOF. Indeed, given an algebraic manifold $X \subset F^N$, we can consider the ideal $I \subset F[x_1, \dots, x_N]$ of polynomials vanishing on X , then write $I = (P_1, \dots, P_k)$ with $k < \infty$, as in Theorem 8.20, and then set $S_i \subset F^N$ to be the set of zeroes of P_i . \square

Moving ahead now, let us further investigate the correspondence $X \leftrightarrow I$. We would like this to be bijective, but there is a clear obstruction to this, over $F = \mathbb{R}$, coming from the fact that at $N = 1$, the ideal $I = (x^2 + 1) \subset \mathbb{R}[x]$ produces the manifold $X = \emptyset$. Thus, we must trade \mathbb{R} for something bigger, say for the field $\mathbb{C} = \mathbb{R}[i]$, with $i^2 = -1$.

Not bad all this, and in the lack of any other clear obstruction to the bijectivity of the correspondence $X \leftrightarrow I$, let us formulate our findings as follows:

CONJECTURE 8.22 (Nullstellensatz). *We have a bijective correspondence*

$$\left(X \subset \mathbb{C}^N \right) \longleftrightarrow \left(I \subset \mathbb{C}[x_1, \dots, x_N] \right)$$

between algebraic manifolds in \mathbb{C}^N , and radical ideals of $\mathbb{C}[x_1, \dots, x_N]$.

Here the name is there for reminding us that such a statement would be as beautiful, powerful and reliable as a Mercedes or BMW car. And more on this, later in this book, after discussing more in detail the above-mentioned field $\mathbb{C} = \mathbb{R}[i]$, with $i^2 = -1$. We will see at that time, following Hilbert, that the Nullstellensatz holds indeed.

8c. Fields, extensions

Switching topics, and as a continuation of our original plan, let us discuss now field theory. Let us start with something that we know well since chapter 2, namely:

DEFINITION 8.23. *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 $a(b + c) = ab + ac$.*

In other words, a field satisfies what we can normally expect from “numbers”, and as basic examples, we have of course \mathbb{Q}, \mathbb{R} . There are many other examples of fields, along the same lines. We can talk for instance about fields like $\mathbb{Q}[\sqrt{2}]$, as follows:

PROPOSITION 8.24. *The following is an intermediate field $\mathbb{Q} \subset F \subset \mathbb{R}$,*

$$\mathbb{Q}[\sqrt{2}] = \left\{ a + b\sqrt{2} \mid a, b \in \mathbb{Q} \right\}$$

and the same happens for any $\mathbb{Q}[\sqrt{n}]$, with $n \neq m^2$ being not a square.

PROOF. All the field axioms are clearly satisfied, except perhaps for the inversion axiom. But this axiom is satisfied too, due to the following formula:

$$\frac{1}{a + b\sqrt{2}} = \frac{a - b\sqrt{2}}{a^2 - 2b^2}$$

Observe that the denominator is indeed nonzero, due to $a^2 \neq 2b^2$, which follows by reasoning modulo 2. As for the case of $\mathbb{Q}[\sqrt{n}]$ with $n \neq m^2$, this is similar. \square

The above result is quite interesting, obviously in relation with arithmetic, and suggests looking into the intermediate fields of numbers, as follows:

$$\mathbb{Q} \subset F \subset \mathbb{R}$$

As another observation now, complementary to this, with our field theory we are not away from geometry, quite the opposite. Indeed, while the usual spaces of functions are obviously not fields, geometry and analysis remain around the corner, due to:

THEOREM 8.25. *The quotients of real polynomials, called rational functions, when written in reduced form, as follows, with P, Q prime to each other,*

$$f = \frac{P}{Q}$$

are well-defined and continuous outside the zeroes $P_f \subset \mathbb{R}$ of Q , called poles of f :

$$f : \mathbb{R} - P_f \rightarrow \mathbb{R}$$

Also, these functions are stable under summing, making products and taking inverses,

$$\frac{P}{Q} + \frac{R}{S} = \frac{PS + QR}{QS} \quad , \quad \frac{P}{Q} \cdot \frac{R}{S} = \frac{PR}{QS} \quad , \quad \left(\frac{P}{Q}\right)^{-1} = \frac{Q}{P}$$

so they form a field $\mathbb{R}(X)$, called field of real rational functions.

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

(1) To start with, almost everything in the statement is clear from definitions, and with the extra remark that the rational functions are stable under composition too.

(2) Next, in what regards the term “pole” in the statement, this comes from the fact that, when trying to draw the graph of f , we are faced with some sort of tent, which is suspended by infinite poles, which lie, guess where, at the poles of f .

(3) Finally, many other things can be said about rational functions, but it is probably wiser here to wait first for the introduction of the complex numbers \mathbb{C} , where all polynomials, and in particular the denominator Q , have roots. So, more on this later. \square

Back now to arithmetic, there are many other interesting examples of fields, such as the integers modulo a prime p , which form a field \mathbb{F}_p , that we met in chapter 3, or the p -adic numbers, again with respect to a prime p , that we will discuss in a moment.

In short, field theory is certainly relevant to geometry, arithmetic, and many other things. Getting now to what can be said about fields, according to chapter 3 we have:

THEOREM 8.26. *Given a field F , define its characteristic $p = \text{char}(F)$ as being the smallest $p \in \mathbb{N}$ such that the following happens, and as $p = 0$, if this never happens:*

$$\underbrace{1 + \dots + 1}_{p \text{ times}} = 0$$

Then, assuming $p > 0$, this number p must be prime, we have a field embedding $\mathbb{F}_p \subset F$, and $q = |F|$ must be of the form $q = p^k$, with $k \in \mathbb{N}$. Also, we have the formulae

$$(a + b)^p = a^p + b^p \quad , \quad a^q = a$$

valid for any $a, b \in F$, and the Fermat polynomial $X^q - X$ factorizes as:

$$X^q - X = \prod_{a \in F} (X - a)$$

Also, regardless of p , any finite multiplicative subgroup $G \subset F - \{0\}$ must be cyclic.

PROOF. This is a very crowded statement, containing all sorts of things that we know from chapter 3, the idea with all this being as follows:

(1) The fact that $p > 0$ must be prime comes by contradiction, by using:

$$\underbrace{(1 + \dots + 1)}_{a \text{ times}} \times \underbrace{(1 + \dots + 1)}_{b \text{ times}} = \underbrace{1 + \dots + 1}_{ab \text{ times}}$$

Indeed, assuming that we have $p = ab$ with $a, b > 1$, the above formula corresponds to an equality of type $AB = 0$ with $A, B \neq 0$ inside F , which is impossible.

(2) Back to the general case, F has a smallest subfield $E \subset F$, called prime field, consisting of the various sums $1 + \dots + 1$, and their quotients. In the case $p = 0$ we obviously have $E = \mathbb{Q}$. In the case $p > 0$ now, the multiplication formula in (1) shows that the set $S = \{1 + \dots + 1\}$ is stable under taking quotients, and so $E = S$.

(3) Now with $E = S$ in hand, we obviously have $(E, +) = \mathbb{Z}_p$, and since the multiplication is given by the formula in (1), we conclude that we have $E = \mathbb{F}_p$, as a field. Thus, in the case $p > 0$, we have constructed an embedding $\mathbb{F}_p \subset F$, as claimed.

(4) In the context of the above embedding $\mathbb{F}_p \subset F$, we can say that F is a vector space over \mathbb{F}_p , and so we have $|F| = p^k$, with $k \in \mathbb{N}$ being the dimension of this space.

(5) The baby Fermat formula $(a + b)^p = a^p + b^p$ can be established as follows:

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

(6) As for the Fermat formula $a^q = a$ itself, which implies the assertion about $X^q - X$, this follows from the last assertion, which can be proved via some basic arithmetic inside F , and which for $G = F - \{0\}$ itself, with $|F| = q$, gives $a^{q-1} = 1$, for any $a \neq 0$. \square

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

THEOREM 8.27. *Given a field extension $E \subset F$, we can talk about its Galois group G , as the group of automorphisms of F fixing E . The intermediate fields*

$$E \subset K \subset F$$

are then in correspondence with the subgroups $H \subset G$, with such a field K corresponding to the subgroup H consisting of automorphisms $g \in G$ fixing K .

PROOF. This is something quite self-explanatory, and follows indeed from some algebra, under suitable assumptions, in order for that algebra to properly apply. There are many good books, that you can learn this technology from, such as Lang [63]. \square

Getting now towards polynomials and their roots, we have here:

THEOREM 8.28. *Given a field F and a polynomial $P \in F[X]$, we can talk about the abstract splitting field of P , where this polynomial decomposes as:*

$$P(X) = c \prod_i (X - a_i)$$

In particular, any field F has a certain algebraic closure \bar{F} , where all the polynomials $P \in F[X]$, and in fact all polynomials $P \in \bar{F}[X]$ too, have roots.

PROOF. This is again something self-explanatory, which follows from Theorem 8.27 and from some extra algebra, under suitable assumptions, in order for that extra algebra to properly apply. Regarding the construction at the end, as main example here we have $\mathbb{R} = \mathbb{C}$. However, as an interesting fact, $\bar{\mathbb{Q}} \subset \mathbb{C}$ is a proper subfield. See [63]. \square

Good news, with this in hand, we can now elucidate the structure of finite fields:

THEOREM 8.29. *For any prime power $q = p^k$ there is a unique field \mathbb{F}_q having q elements. At $k = 1$ this is the usual \mathbb{F}_p . In general, this is the splitting field of:*

$$P = X^q - X$$

Moreover, we can construct an explicit model for \mathbb{F}_q , at $q = p^2$ or higher, as

$$\mathbb{F}_q = \mathbb{F}_p[X]/(Q)$$

with $Q \in \mathbb{F}_p[X]$ being a suitable irreducible polynomial, of degree k .

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

(1) The first assertion, regarding the existence and uniqueness of \mathbb{F}_q , follows from Theorem 8.26 and Theorem 8.28. Indeed, we know from Theorem 8.26 that given a finite field, $|F| = q$ with $k \in \mathbb{N}$, the Fermat polynomial $P = X^q - X$ factorizes as:

$$X^q - X = \prod_{a \in F} (X - a)$$

But this shows, via the general theory from Theorem 8.28, that our field F must be the splitting field of P , and so is unique. As for the existence, this follows again from Theorem 8.28, telling us that the splitting field always exists.

(2) In what regards now the modeling of \mathbb{F}_q , at $q = p$ there is nothing to do, because we have our usual \mathbb{F}_p here. At $q = p^2$ and higher, we know from commutative algebra that we have an isomorphism as follows, whenever $Q \in \mathbb{F}_p[X]$ is taken irreducible:

$$\mathbb{F}_q = \mathbb{F}_p[X]/(Q)$$

(3) Regarding now the best choice of the irreducible polynomial $Q \in \mathbb{F}_p[X]$, providing us with a good model for the finite field \mathbb{F}_q , that we can use in practice, this question depends on the value of $q = p^k$, and many things can be said here. All in all, our models are quite similar to $\mathbb{C} = \mathbb{R}[i]$, with i being a formal number satisfying $i^2 = -1$.

(4) To be more precise, at the simplest exponent, $q = 4$, to start with, we can use $Q = X^2 + X + 1$, with this being actually the unique possible choice of a degree 2 irreducible polynomial $Q \in \mathbb{F}_2[X]$, and this leads to a model as follows:

$$\mathbb{F}_4 = \left\{ 0, 1, a, a + 1 \mid a^2 = a + 1 \right\}$$

To be more precise here, we assume of course that the characteristic of our model is $p = 2$, which reads $x + x = 0$ for any x , and so determines the addition table. As for the multiplication table, this is uniquely determined by $a^2 = -a - 1 = a + 1$.

(5) Next, at exponents of type $q = p^2$ with $p \geq 3$ prime, we can use $Q = X^2 - r$, with r being a non-square modulo p , and with $(p - 1)/2$ choices here. We are led to:

$$\mathbb{F}_{p^2} = \left\{ a + b\gamma \mid \gamma^2 = r \right\}$$

Here, as before with \mathbb{F}_4 , our formula is something self-explanatory. Observe the analogy with $\mathbb{C} = \mathbb{R}[i]$, with i being a formal number satisfying $i^2 = -1$.

(6) Finally, at $q = p^k$ with $k \geq 3$ things become more complicated, but the main idea remains the same. We have for instance models for \mathbb{F}_8 , \mathbb{F}_{27} using $Q = X^3 - X - 1$, and a model for \mathbb{F}_{16} using $Q = X^4 + X + 1$. Many other things can be said here. \square

As another application of the above, which motivated Galois, we have:

THEOREM 8.30. *There is no explicit formula for the roots of polynomials of degree $N = 5$ and higher, with the reason for this, coming from Galois theory, being that S_5 is not solvable. The simplest numeric example is $P = X^5 - X - 1$.*

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

(1) The first assertion, for generic polynomials, is due to Abel-Ruffini, but Galois theory helps in better understanding this, and comes with a number of bonus points too, namely the possibility of formulating a finer result, with Abel-Ruffini's original "generic", which was something algebraic, being now replaced by an analytic "generic", and also with the possibility of dealing with concrete polynomials, such as $P = X^5 - X - 1$.

(2) Regarding now the details of the Galois proof of the Abel-Ruffini theorem, assume that the roots of a polynomial $P \in F[X]$ can be computed by using iterated roots, a bit as for the degree 2 equation, or for the degree 3 and 4 equations, via Cardano. Then, algebraically speaking, this gives rise to a tower of fields as follows, with $F_0 = F$, and each F_{i+1} being obtained from F_i by adding a root, $F_{i+1} = F_i(x_i)$, with $x_i^{n_i} \in F_i$:

$$F_0 \subset F_1 \subset \dots \subset F_k$$

(3) In order for Galois theory to apply well to this situation, we must make all the extensions normal, which amounts in replacing each $F_{i+1} = F_i(x_i)$ by its extension $K_i(x_i)$, with K_i extending F_i by adding a n_i -th root of unity. Thus, with this replacement, we can assume that the tower in (2) is normal, meaning that all Galois groups are cyclic.

(4) Now by Galois theory, at the level of the corresponding Galois groups we obtain a tower of groups as follows, which is a resolution of the last group G_k , the Galois group of P , in the sense of group theory, in the sense that all quotients are cyclic:

$$G_1 \subset G_2 \subset \dots \subset G_k$$

As a conclusion, Galois theory tells us that if the roots of a polynomial $P \in F[X]$ can be computed by using iterated roots, then its Galois group $G = G_k$ must be solvable.

(5) In the generic case, the conclusion is that Galois theory tells us that, in order for all polynomials of degree 5 to be solvable, via square roots, the group S_5 , which appears there as Galois group, must be solvable, in the sense of group theory. But this is wrong, because the alternating subgroup $A_5 \subset S_5$ is simple, and therefore not solvable.

(6) Finally, regarding the polynomial $P = X^5 - X - 1$, some elementary computations here, based on arithmetic over $\mathbb{F}_2, \mathbb{F}_3$, and involving various cycles of length 2, 3, 5, show that its Galois group is S_5 . Thus, we have our counterexample.

(7) To be more precise, our polynomial factorizes over \mathbb{F}_2 as follows:

$$X^5 - X - 1 = (X^2 + X + 1)(X^3 + X^2 + 1)$$

We deduce from this the existence of an element $\tau\sigma \in G \subset S_5$, with $\tau \in S_5$ being a transposition, and with $\sigma \in S_5$ being a 3-cycle, disjoint from it. Thus, we have:

$$\tau = (\tau\sigma)^3 \in G$$

(8) On the other hand since $P = X^5 - X - 1$ is irreducible over \mathbb{F}_5 , we have as well available a certain 5-cycle $\rho \in G$. Now since $\langle \tau, \rho \rangle = S_5$, we conclude that the Galois group of P is full, $G = S_5$, and by (4) and (5) we have our counterexample.

(9) Finally, as mentioned in (1), all this shows as well that a random polynomial of degree 5 or higher is not solvable by square roots, and with this being an elementary consequence of the main result from (5), via some standard analysis arguments. \square

So long for the main ideas of Galois theory. In practice, what we said in the above was of course something very quick, and for further learning such things, which normally takes some time, we recommend any solid abstract algebra book, such as Lang [63].

8d. p-adic numbers

We would like to discuss now some arithmetic techniques, for dealing with equations over the rationals, and with the rational numbers themselves, based on the key notion of p -adic number. As a main question here, that we would like to solve, which was raised and explained in the beginning of this chapter, in some detail, we have:

QUESTION 8.31. *What are the p -adic numbers, defined with respect to a chosen prime number p , making the local-global principle work?*

In answer, let us temporarily forget about equations, and the local-global principle, and simply pick a prime number p , and look at the world from the perspective of p . So, imagining that we are p , both me and you, what we see is something as follows:

(1) First, we see all sorts of integers $a \in \mathbb{Z}$. Some appear friendly, namely those of the form $a \in p\mathbb{Z}$, while the others, of the form $a \notin p\mathbb{Z}$, appear bizarre and distant.

(2) Moreover, between friends $a \in p\mathbb{Z}$, those of the form $a \in p^2\mathbb{Z}$ appear particularly close. And among them, $a \in p^3\mathbb{Z}$ are truly very close friends. And so on.

(3) Then, we see all sorts of rationals, $r = a/b$, and again, some are close, some are distant, depending on the exact p^k factor, with $k \in \mathbb{Z}$, appearing inside r .

(4) In particular, the rationals of the form $r = 1/p^k$ with $k \gg 0$ appear really frightening. Fortunately they are very far away from us, we can barely see them.

(5) And finally, we can see some irrationals $x \notin \mathbb{Q}$ too, but these being uncountable, it is quite hard to figure out how they look like, and are distributed in space.

Very good, so getting back to Earth now, let us write down a definition, based on what we saw in our Prime Number Experience. By focusing on the integers, and more generally the rationals, and leaving the irrationals for later, we have:

DEFINITION 8.32. *Given p prime, we define the p -adic norm of $r \in \mathbb{Q}$ as being:*

$$|r| = p^{-k} \quad , \quad r = p^k \frac{a}{b} \quad , \quad a, b \neq 0(p)$$

Also, we call the integer $k \in \mathbb{Z}$ the p -adic valuation of r , and denote it $k = v(r)$.

As a comment here, $|r| = p^{-k}$ is the natural choice, because according to our Prime Number Experience, the bigger $k \in \mathbb{Z}$ is, the smaller $|r| > 0$ must be, and so we are looking for a formula of type $|r| = \beta^{-k}$ with $\beta > 1$, as for this to happen. Of course, there is still a question left, in regards with the value of $\beta > 1$. But, again coming from our Prime Number Experience, if I am for instance $p = 11$, why shall I use $\beta = 17$.

Of course you might argue here that there might be some mighty universal number, such as $e = 2.7182\dots$ or $\pi = 3.1415\dots$ or $1/\alpha = 137.0359\dots$ doing the job for all prime numbers p . But this cannot work, as we will see next, with some simple math.

Going ahead now with math, the question is, is our Definition 8.32 correct? That is, is $|r|$ indeed a norm? And here, it depends a bit on your background, with mathematicians being a bit dissatisfied, to the point of even choosing to stop calling $|r|$ a norm, but physicists and others being fully happy with it, the result being as follows:

THEOREM 8.33. *The p -adic norm $|r| = p^{-k}$ is not exactly a norm, but satisfies the following conditions, which are even better:*

- (1) *First axiom:* $|x| \geq 0$, with $|x| = 0$ when $x = 0$.
- (2) *Modified second axiom:* $|xy| = |x| \cdot |y|$.
- (3) *Strong triangle inequality:* $|x + y| \leq \max(|x|, |y|)$.

PROOF. All this follows indeed from some simple arithmetic modulo p :

(1) That axiom clearly holds, with the remark that we forgot to say in Definition 8.32 that $v(0) = \infty$, by definition, because any p^k , no matter how big $k \in \mathbb{N}$ is, divides 0.

(2) As a first observation, the usual second norm axiom, namely $|\lambda x| = ||\lambda|| \cdot |x|$, with $||\cdot||$ standing here for the usual absolute value of the numbers, definitely fails, and this because all the p -adic norms $|r|$ are by definition integer powers of p , and an arbitrary $\lambda \in \mathbb{Q}$ will mess up this. However, we have instead $|xy| = |x| \cdot |y|$, coming from:

$$v(xy) = v(x)v(y)$$

And is this good news or not. After some thinking, this modified second axiom is just as good as the failed usual second axiom, because who cares about arbitrary numbers $\lambda \in \mathbb{Q}$, not viewed from the perspective of p , I mean. More on this in a moment.

(3) Finally, let us look at sums $x + y$. Over the integers $p^k|x, y$ implies $p^k|x + y$, and with a bit of fractions arithmetic, that we will leave here as an easy exercise, the same holds for rationals, in the sense that we have, in terms of the p -adic valuation:

$$v(x + y) \geq \min(v(x), v(y))$$

Thus the p -adic norm itself, $|r| = p^{-v(r)}$, satisfies the following inequality:

$$|x + y| \leq \max(|x|, |y|)$$

Now, what does this inequality mean, geometrically? Good question, and as a first remark, since this is obviously something stronger than the usual triangle inequality satisfied by the norms, $|x + y| \leq |x| + |y|$, we will call it strong triangle inequality. \square

Before going ahead, let us further examine the strong triangle inequality found in the above. This is something new to us, and as a further result on it, we have:

PROPOSITION 8.34. *The strong triangle inequality implies*

$$|x| \neq |y| \implies |x + y| = \max(|x|, |y|)$$

and with this being valid for any modified norm, in the sense of Theorem 8.33.

PROOF. This is again something elementary, the idea being as follows:

(1) In what regards the p -adic norm, going back to (3) in the proof of Theorem 8.33, we can add there the observation that, trivially over the integers, and then over the rationals too, with a bit of fraction work, the p -adic valuation satisfies:

$$v(x) \neq v(y) \implies v(x + y) = \min(v(x), v(y))$$

Thus the p -adic norm itself satisfies the condition in the statement.

(2) More generally now, and with this being something quite interesting, our claim is that this phenomenon is valid for any generalized norm in the sense of Theorem 8.33. Indeed, assume that $|x| \geq 0$, with $|x| = 0$ when $x = 0$, as usual, and that:

$$|xy| = |x| \cdot |y| \quad , \quad |x + y| \leq \max(|x|, |y|)$$

In order to prove our result, assume $|x| > |y|$. We then have, trivially:

$$|x + y| \leq \max(|x|, |y|) = |x|$$

(3) In the other sense now, we have to work a bit. We have the following computation, with at the end the observation that the max cannot be $|y|$, because if that would be the case, the inequality that we would obtain would be $|x| \leq |y|$, contradicting $|x| > |y|$:

$$\begin{aligned} |x| &= |(x + y) - y| \\ &\leq \max(|x + y|, |y|) \\ &= |x + y| \end{aligned}$$

Thus, we have equality in the estimate in (2), as desired. \square

Very nice all this, and getting back now to what we have in Theorem 8.33, namely the modified norm axioms there, we can formulate, as a simple consequence:

PROPOSITION 8.35. *The p -adic norm $|r| = p^{-k}$ is not exactly a norm, but*

$$d(x, y) = |x - y|$$

is a distance. Thus, the rationals \mathbb{Q} become in this way a metric space.

PROOF. With the conditions satisfied by the p -norm $|r|$ in hand, it follows, trivially, that $d(x, y) = |x - y|$ is indeed a distance, making \mathbb{Q} a metric space. \square

Now let us turn to irrationals. The quite blurry picture that we saw during our Prime Number Experience, and with the blame at that time being on the uncountability of these beasts, in the lack of something better, can be now explained. Indeed, what we saw were not the “usual” irrationals $x \in \mathbb{R} - \mathbb{Q}$, but rather some irrationals $x \in \mathbb{Q}_p - \mathbb{Q}$ viewed from the perspective of p , constructed according to the following result:

THEOREM 8.36. *By completing \mathbb{Q} with respect to the p -adic distance*

$$d(x, y) = |x - y|$$

we obtain a certain field \mathbb{Q}_p , called field of p -adic numbers.

PROOF. This is something very standard, with the passage $\mathbb{Q} \rightarrow \mathbb{Q}_p$ being very similar to the passage $\mathbb{Q} \rightarrow \mathbb{R}$, that we are very familiar with. In fact, some things get even simpler for p -adics, due to the strong triangle inequality satisfied by the norm. \square

What is next? Many things, especially in relation with understanding what the p -adic irrationals $x \in \mathbb{Q}_p - \mathbb{Q}$ really are, concretely speaking. But before that, inspired by the theory of usual numbers, $\mathbb{Z} \subset \mathbb{Q}$, we can introduce the p -adic integers, as follows:

THEOREM 8.37. *We can introduce the p -adic integers $\mathbf{Z}_p \subset \mathbb{Q}_p$ as being*

$$\mathbf{Z}_p = \left\{ x \in \mathbb{Q}_p \mid |x| \leq 1 \right\}$$

not to be confused with \mathbb{Z}_p , and this is a ring, appearing as completion of $\mathbb{Z} \subset \mathbf{Z}_p$.

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

(1) We can certainly introduce a set $\mathbf{Z}_p \subset \mathbb{Q}_p$ by the condition in the statement, and the ring axioms are all clear from the modified norm conditions, from Theorem 8.33, the verifications of the fact that \mathbf{Z}_p is stable under sums and products being as follows:

$$|x|, |y| \leq 1 \implies |x + y| \leq \max(|x|, |y|) \leq 1$$

$$|x|, |y| \leq 1 \implies |xy| = |x| \cdot |y| \leq 1$$

(2) Next, since the valuation of a usual integer $x \in \mathbb{Z}$ satisfies $v(x) \geq 0$, the norm satisfies $|x| \leq 1$, and so we have an inclusion $\mathbb{Z} \subset \mathbf{Z}_p$, as in the statement.

(3) With a bit more work, we can see that \mathbf{Z}_p is closed with respect to the p -adic norm, and also, that it appears as the completion of its subring $\mathbb{Z} \subset \mathbf{Z}_p$.

(4) Finally, and getting now into hot stories and other funny facts, the ring of p -adic integers \mathbf{Z}_p is obviously not to be confused with the cyclic group \mathbb{Z}_p . There are actually two schools of thought here, with the other school denoting the p -adic integers by \mathbb{Z}_p , and using for the cyclic group all sorts of bizarre notations, such as C_p .

(5) In what regards our philosophy, that is very simple. If you need some sort of integers with respect to p , for your mathematics, this is a no-brainer, go with the remainders modulo p , or even better, with the p -th roots of unity, and that will solve your mathematical question, in 99% of the cases. And in the remaining 1% cases, what you need are probably the p -adic integers. So, assuming at least a little bit of modesty and common sense, the simplest notation, \mathbb{Z}_p , should be attributed to the cyclic group. \square

With this understood, let us get now to the irrationals, and non-integers, and the p -adic numbers in general, viewed as a whole. Obviously, in order to understand them, we must understand well the Cauchy sequences and convergence in \mathbb{Q}_p . But here, many surprises are waiting for us, as for instance the following notorious formula:

THEOREM 8.38. *We have the following formula,*

$$\sum_{k=0}^{\infty} p^k = \frac{1}{1-p}$$

with respect to the p -adic norm.

PROOF. By using $p^n \rightarrow 0$, with respect to the p -adic norm, we have:

$$\sum_{k=0}^{n-1} p^k = \frac{1-p^n}{1-p} \simeq \frac{1}{1-p} - \frac{0}{1-p} = \frac{1}{1-p}$$

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

Quite cool the above formula, we are learning new things here, aren't we, and even more spectacular is its $p = 2$ particular case, which reads:

$$\sum_{k=0}^{\infty} 2^k = -1$$

As a matter of doublechecking, this latter formula can be proved as follows:

$$\sum_{k=0}^{n-1} 2^k = 2^n - 1 \simeq 0 - 1 = -1$$

Moving ahead, the continuation of the story is more complicated, involving some non-trivial algebra, with the main result here, called Hasse local-global principle, stating more

or less that for certain questions in arithmetic, for instance in relation with the elliptic curves, our dreams and hopes expressed in the beginning of this chapter come true. For more on this, have a look at any advanced number theory book, such as Serre [79].

* * *

Finally, since this was our last pure algebra chapter in this book, with most of what is to follow being analysis, let us recommend as well some general algebra reading:

(1) Speaking algebra in general, first comes linear algebra, no doubt about this. There are many texts available here, and I can only recommend you my own book [9].

(2) Then, as mentioned on many occasions, for foundations and generalities, and everything you need to know, comes a solid abstract algebra book, such as Lang [63].

(3) Then comes commutative algebra, say learned from Atiyah and MacDonald [7], and algebraic geometry, say from Harris [46], Hartshorne [47] or Shafarevich [82].

(4) And then, you are free to learn what you want, from all the algebra greats who wrote great books, such as Atiyah, Connes, Grothendieck, Serre and others.

8e. Exercises

This was a truly advanced algebraic chapter, and as exercises, we have:

EXERCISE 8.39. *Learn a bit of commutative and noncommutative algebra.*

EXERCISE 8.40. *Clarify everything that we said, in relation with the conics.*

EXERCISE 8.41. *Also, check what Kepler and Newton said, about the same conics.*

EXERCISE 8.42. *Learn all the details in the proof of the Hilbert basis theorem.*

EXERCISE 8.43. *Think a bit about complex numbers, and the Nullstellensatz.*

EXERCISE 8.44. *Learn how to locally write the rational functions as series.*

EXERCISE 8.45. *Learn more about the classification and modeling of finite fields.*

EXERCISE 8.46. *Learn more about the p -adic numbers, and what they can do.*

As bonus exercise, start reading some algebraic geometry. All good learning.

Part III

Complex numbers

*No no limits, we'll reach for the sky
No valley too deep, no mountain too high
No no limits, won't give up the fight
We do what we want and we do it with pride*

CHAPTER 9

Complex numbers

9a. Complex numbers

We have already learned many interesting things, and in order to further upgrade our knowledge in arithmetic, we will discuss now the complex numbers. There is a lot of magic here, and we will carefully explain this material. Their definition is as follows:

DEFINITION 9.1. *The complex numbers are variables of the form*

$$x = a + ib$$

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

$$i^2 = -1$$

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

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

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

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

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

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

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

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

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

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

THEOREM 9.2. *The complex solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{R}$ are*

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

with the square root of negative real numbers being defined as

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

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

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

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

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

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

THEOREM 9.3. *The complex numbers, written as usual as*

$$x = a + ib$$

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

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

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

PROOF. Consider indeed two arbitrary complex numbers, as follows:

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

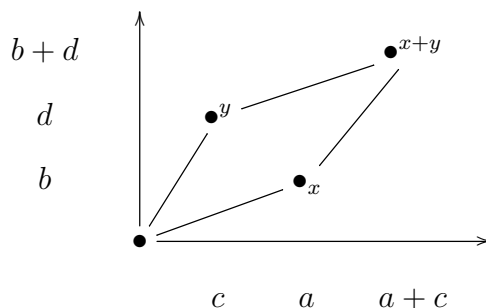
Their sum is then by definition the following complex number:

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

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

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

According to vector calculus, $x + y$ comes by forming a parallelogram, as follows:



In other words, the vector sum $x + y$ is given by the following formula:

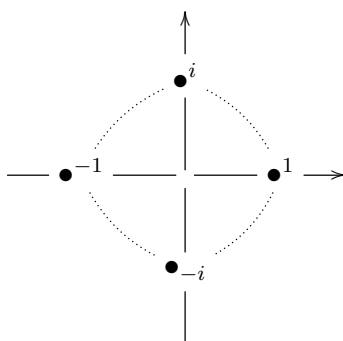
$$x + y = \begin{pmatrix} a + c \\ b + d \end{pmatrix}$$

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

Observe that in our geometric picture from Theorem 9.3, the real numbers correspond to the numbers on the Ox axis. As for the purely imaginary numbers, these lie on the Oy axis, with the number i itself being given by the following formula:

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

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



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

ADVICE 9.4. *When drawing complex numbers, always begin with the coordinate axes Ox, Oy , and with a copy of the unit circle.*

Back to equations now, it is pretty much clear that the computation in the proof of Theorem 9.2 will perfectly work in the case $a, b, c \in \mathbb{C}$, so in order to have that theorem working over \mathbb{C} , we need to extract square roots. And, this can be done as follows:

THEOREM 9.5. *Any complex number $x = a + ib$ has two square roots, given by*

$$\sqrt{x} = \pm \sqrt{\frac{a + \sqrt{a^2 + b^2}}{2}} \pm i \sqrt{\frac{-a + \sqrt{a^2 + b^2}}{2}}$$

with the signs being identical when $b > 0$, and opposite when $b < 0$.

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

(1) With $x = a + ib$ as in the statement, and $\sqrt{x} = c + id$, our equation is:

$$(c + id)^2 = a + ib$$

In terms of the real and imaginary parts, we have two equations, as follows:

$$c^2 - d^2 = a, \quad 2cd = b$$

(2) Let us first compute the number $u = c^2$. The equation for it is as follows:

$$u - \frac{b^2}{4u} = a$$

Thus, the number $u = c^2$ satisfies the following degree 2 equation:

$$u^2 - au - \frac{b^2}{4} = 0$$

But this latter equation has a unique positive solution, given by:

$$u = \frac{a + \sqrt{a^2 + b^2}}{2}$$

Thus, we are led to the formula of $c = \pm\sqrt{u}$ in the statement.

(3) Similarly, let us compute now $v = d^2$. The equation for it is as follows:

$$\frac{b^2}{4v} - v = a$$

Thus, the number $v = d^2$ satisfies the following degree 2 equation:

$$v^2 + av - \frac{b^2}{4} = 0$$

But this latter equation has a unique positive solution, given by:

$$v = \frac{-a + \sqrt{a^2 + b^2}}{2}$$

Thus, we are led to the formula of $d = \pm\sqrt{v}$ in the statement, and this gives the result, with the last assertion regarding signs being clear, coming from $2cd = b$. \square

With this being said, I don't know about you, but personally, for better sleeping at night, I would rather prefer to have this doublechecked. So, given two numbers $a, b \in \mathbb{R}$, consider the following numbers $c, d \in \mathbb{R}$, with the sign on the right being that of b :

$$c = \sqrt{\frac{a + \sqrt{a^2 + b^2}}{2}} \quad , \quad d = \pm \sqrt{\frac{-a + \sqrt{a^2 + b^2}}{2}}$$

We have then $(c + id)^2 = (c^2 - d^2) + 2icd$, whose real part is given by:

$$\begin{aligned} c^2 - d^2 &= \frac{a + \sqrt{a^2 + b^2}}{2} - \frac{-a + \sqrt{a^2 + b^2}}{2} \\ &= \frac{a}{2} + \frac{a}{2} \\ &= a \end{aligned}$$

As for the imaginary part, this can be computed as follows:

$$\begin{aligned} 2cd &= \pm 2 \sqrt{\frac{a + \sqrt{a^2 + b^2}}{2}} \cdot \frac{-a + \sqrt{a^2 + b^2}}{2} \\ &= \pm 2 \sqrt{\frac{-a^2 + a^2 + b^2}{4}} \\ &= \pm |b| \\ &= b \end{aligned}$$

Thus we have indeed $(c + id)^2 = a + ib$, as desired. Now by getting back to the degree 2 equations, we can formulate a new result regarding them, as follows:

THEOREM 9.6. *The complex solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{C}$ are*

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

with the square root of $b^2 - 4ac = p + iq$ being extracted as above, namely

$$\sqrt{p + iq} = \pm \sqrt{\frac{p + \sqrt{p^2 + q^2}}{2}} \pm i \sqrt{\frac{-p + \sqrt{p^2 + q^2}}{2}}$$

with the signs being identical when $q > 0$, and opposite when $q < 0$.

PROOF. This follows indeed from our old degree 2 computation, from the proof of Theorem 9.2, with the square roots being extracted as in Theorem 9.5. \square

And more on this, including some further generalizations, later in this chapter.

9b. Polar coordinates

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

DEFINITION 9.7. *The complex numbers $x = a + ib$ can be written in polar coordinates,*

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

with the connecting formulae being as follows,

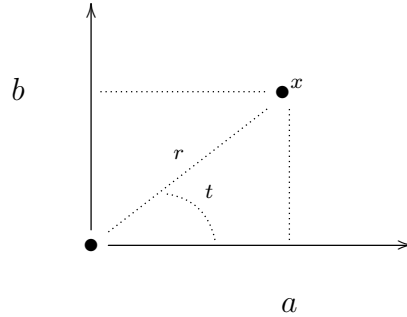
$$a = r \cos t \quad , \quad b = r \sin t$$

and in the other sense being as follows,

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

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

Observe that there is a clear relation here with the vector notation from Theorem 9.3, because r is the length of the vector, and t is the angle made by the vector with the Ox axis. To be more precise, the picture for what is going on above is as follows:



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

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

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

THEOREM 9.8. *Two complex numbers written in polar coordinates,*

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

multiply according to the following formula:

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

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

PROOF. This follows, and is in fact equivalent to, the following standard formulae for the addition of the angles, that we know well since chapter 5:

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

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

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

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

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

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

THEOREM 9.9. *The powers of a complex number, written in polar form,*

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

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

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

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

PROOF. Given a complex number x , written in polar form as above, and an exponent $k \in \mathbb{N}$, we have indeed the following computation, with k terms everywhere:

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

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

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

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

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

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

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

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

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

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

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

As a basic application now of Theorem 9.9, we have the following result:

PROPOSITION 9.10. *Each complex number, written in polar form,*

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

has two square roots, given by the following formula:

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

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

PROOF. The first assertion is clear indeed from the general formula in Theorem 9.9, at $k = 1/2$. As for its particular cases with $x \in \mathbb{R}$, these are clear from it. \square

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

With the above results in hand, and notably with the square root formula from Proposition 9.10, we can go back now to the degree 2 equations, and we have:

THEOREM 9.11. *The complex solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{C}$ are*

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

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

PROOF. This is something that we already know from Theorem 9.6, with the computations being the same as in the real case, namely:

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

However, and here comes the point, in order to extract the square root we can use Proposition 9.10, which is better than what we had before, in Theorem 9.5. \square

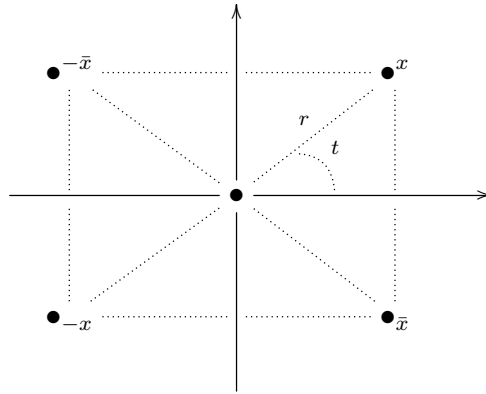
As a last general topic regarding the complex numbers, let us discuss conjugation. This is something quite tricky, complex number specific, as follows:

DEFINITION 9.12. *The complex conjugate of $x = a + ib$ is the following number,*

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

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

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



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

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

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

THEOREM 9.13. *The conjugation operation $x \rightarrow \bar{x}$ has the following properties:*

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

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

- (1) This is something that we already know, coming from definitions.
- (2) This is something clear too, because with $x = a + ib$ our equation $x = -\bar{x}$ reads $a + ib = -a + ib$, and so $a = 0$, which amounts in saying that x is purely imaginary.
- (3) This is a key formula, which can be proved as follows, with $x = a + ib$:

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

- (4) This is clear indeed from the picture following Definition 9.12.
- (5) This is something quite magic, which can be proved as follows:

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

- (6) This comes from the formula of the solutions, that we know from Theorem 9.2, but we can deduce this as well directly, without computations. Indeed, by using our assumption that the coefficients are real, $a, b, c \in \mathbb{R}$, we have:

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

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

And with this, good news, we have all the needed basic tools for dealing with complex numbers. There are actually some advanced tools too, and more on them, later.

9c. Exponential writing

Let us discuss now the theory of complex functions $f : \mathbb{C} \rightarrow \mathbb{C}$, in analogy with the theory of the real functions $f : \mathbb{R} \rightarrow \mathbb{R}$. We will see that many results that we know from the real setting extend to the complex setting. Before starting, two remarks on this:

(1) Most of the real functions $f : \mathbb{R} \rightarrow \mathbb{R}$ that we know, such as \sin, \cos, \exp, \log , extend into complex functions $f : \mathbb{C} \rightarrow \mathbb{C}$, and the study of these latter extensions brings some new light on the original real functions. Thus, what we will be doing here will be, in a certain sense, a refinement of the theory that we developed in chapter 6.

(2) On the other hand, since we have $\mathbb{C} \simeq \mathbb{R}^2$, the complex functions $f : \mathbb{C} \rightarrow \mathbb{C}$ that we will study here can be regarded as functions $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$. This is something quite subtle, but in any case, what we will be doing here will stand as well as an introduction to the functions of type $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$, which are something more complicated.

In short, one complex variable is something in between one real variable, and two or more real variables, and we can only expect to end up with a mysterious mixture of surprising and unsurprising results. Welcome to complex analysis. Let us start with:

DEFINITION 9.14. *A complex function $f : \mathbb{C} \rightarrow \mathbb{C}$, or more generally $f : X \rightarrow \mathbb{C}$, with $X \subset \mathbb{C}$ being a subset, is called continuous when, for any $x_n, x \in X$:*

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

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

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

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

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

Generally speaking, what we know about the real functions extends quite well to the complex functions, and more on this later. However, let us point out that not everything extends trivially from the real to the complex case. For instance, we have:

PROPOSITION 9.15. *We have the following formula, valid for any $|x| < 1$,*

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

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

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

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

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

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

More generally, for $x \in (-1, 1)$ the meaning of the formula in the statement is something quite clear and intuitive, by using a similar argument. However, when x is complex, and not real, we are led into a mysterious spiral there, and the only case where the formula is “obvious”, geometrically speaking, is that when $x = rw$, with $r \in [0, 1)$, and with $w^N = 1$, for some $N \in \mathbb{N}$. Indeed, in this case we have the following formula:

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

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

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

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

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

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

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

$$x = re^{it}$$

In order to establish this formula, let us first study the complex exponential:

THEOREM 9.16. *We can exponentiate the complex numbers, according to the formula*

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

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

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

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

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

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

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

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

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

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

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

As a consequence of the above formula $e^{x+y} = e^x e^y$, we have the following result:

PROPOSITION 9.17. *The exponential of complex numbers is given by*

$$e^{s+it} = e^s e^{it}$$

with e^s being a usual real exponential, and with e^{it} , in need to be computed.

PROOF. This is indeed something self-explanatory, coming from $e^{x+y} = e^x e^y$, and with the somewhat non-standard notation $x = s + it$ being something needed later. \square

Now let us get to the remaining problem, computation of e^{it} with $t \in \mathbb{R}$. Here are a few elementary observations, regarding the operation $t \rightarrow e^{it}$:

PROPOSITION 9.18. *For $t \in \mathbb{R}$ the number e^{it} belongs to the unit circle,*

$$e^{it} \in \mathbb{T}$$

and the operation $t \rightarrow e^{it}$ is subject to the following formulae,

$$e^{i(s+t)} = e^{is} e^{it} \quad , \quad e^{i0} = 1 \quad , \quad (e^{it})^{-1} = e^{-it}$$

telling us $t \rightarrow e^{it}$ is a group morphism $\mathbb{R} \rightarrow \mathbb{T}$.

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

(1) To start with, we have the following formula, valid for any $x \in \mathbb{C}$:

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

We have as well the following computation, again valid for any $x \in \mathbb{C}$:

$$e^x e^{-x} = e^{x-x} = e^0 = 1 \implies (e^x)^{-1} = e^{-x}$$

(2) But with these two formulae in hand, we can prove the first assertion. Indeed, the above two formulae, applied with $x = it$, with $t \in \mathbb{R}$, give the following equalities:

$$e^{-it} = \overline{e^{it}} \quad , \quad (e^{it})^{-1} = e^{-it}$$

We conclude that the complex number $z = e^{it}$ has the following property:

$$z^{-1} = \bar{z}$$

But this is exactly the equation of the unit circle \mathbb{T} , as desired.

(3) Regarding now the various formulae in the statement, for the operation $t \rightarrow e^{it}$, these are all trivial, coming from the multiplicativity formula $e^{x+y} = e^x e^y$. \square

Good news, we can now establish a key result, called Euler formula, as follows:

THEOREM 9.19. *We have the following formula,*

$$e^{it} = \cos t + i \sin t$$

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

PROOF. There are several possible proofs of this, the idea being as follows:

(1) Intuitive proof. We know from Proposition 9.18 that $t \rightarrow e^{it}$ is a continuous group morphism $\mathbb{R} \rightarrow \mathbb{T}$. But in view of this, and skipping some details, this operation can only appear by “wrapping”. That is, we must have a formula as follows, for some $\alpha \in \mathbb{R}$:

$$e^{it} = \cos(\alpha t) + i \sin(\alpha t)$$

In order now to find the value of $\alpha \in \mathbb{R}$, let us look at what happens around 0. We have the following elementary estimate, obtained by truncating the series of exp:

$$e^{it} \simeq 1 + it$$

On the other hand, we know from chapter 6 that we have $\sin t \simeq t$ and $\cos t \simeq 1 - t^2/2$, for $t \simeq 0$. We conclude that we have the following estimate, for $t \simeq 0$:

$$\cos(\alpha t) + i \sin(\alpha t) \simeq 1 + i\alpha t$$

Thus we must have $\alpha = 1$, and we are led to the Euler formula in the statement.

(2) Physics proof. We have indeed the following quick computation, and I will leave it to you to figure out what is wrong with it, and what the fix is:

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

(3) Calculus proof. This is something less intuitive, but rock-solid, which will stand as our official proof for the Euler formula. Consider the following function $f : \mathbb{R} \rightarrow \mathbb{C}$:

$$f(t) = \frac{\cos t + i \sin t}{e^{it}}$$

By using $\sin' = \cos$ and $\cos' = -\sin$, that we know since chapter 6, we have:

$$\begin{aligned} f'(t) &= (e^{-it}(\cos t + i \sin t))' \\ &= -ie^{-it}(\cos t + i \sin t) + e^{-it}(-\sin t + i \cos t) \\ &= e^{-it}(-i \cos t + \sin t) + e^{-it}(-\sin t + i \cos t) \\ &= 0 \end{aligned}$$

We conclude that $f : \mathbb{R} \rightarrow \mathbb{C}$ is constant, equal to $f(0) = 1$, as desired. \square

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

THEOREM 9.20. *We have the following formula,*

$$e^{\pi i} = -1$$

and we have $E = mc^2$ as well.

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

(1) The first formula, $e^{\pi i} = -1$, which is actually the main formula in mathematics, comes from Theorem 9.19, by setting $t = \pi$. Indeed, we obtain:

$$\begin{aligned} e^{\pi i} &= \cos \pi + i \sin \pi \\ &= -1 + i \cdot 0 \\ &= -1 \end{aligned}$$

(2) As for $E = mc^2$, which is the main formula in physics, this is something deep too. Although we will not really need it here, we recommend learning it as well, for symmetry reasons between math and physics, say from Feynman [34]. \square

Now back to our $x = re^{it}$ objectives, with the above theory in hand we can indeed use from now on this notation, the complete statement being as follows:

THEOREM 9.21. *The complex numbers $x = a + ib$ can be written in polar coordinates,*

$$x = re^{it}$$

with the connecting formulae being

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

and in the other sense being

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

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

PROOF. This is a reformulation of our previous Definition 9.7, by using the formula $e^{it} = \cos t + i \sin t$ from Theorem 9.19, and multiplying everything by r . \square

With this in hand, we can now go back to the basics, namely the addition and multiplication of the complex numbers. We have the following result:

THEOREM 9.22. *In polar coordinates, the complex numbers multiply as*

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

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

PROOF. This is something that we already know, from Theorem 9.8, reformulated by using the notations from Theorem 9.21. Observe that this follows as well directly, from the fact that we have $e^{x+y} = e^x e^y$, that we know from Theorem 9.16. \square

We can investigate as well more complicated operations, as follows:

THEOREM 9.23. *We have the following operations on the complex numbers, written in polar form, as above:*

- (1) *Inversion:* $(re^{it})^{-1} = r^{-1}e^{-it}$.
- (2) *Square roots:* $\sqrt{re^{it}} = \pm\sqrt{r}e^{it/2}$.
- (3) *Powers:* $(re^{it})^a = r^ae^{ita}$.
- (4) *Conjugation:* $\overline{re^{it}} = re^{-it}$.

PROOF. This is something that we already know, from Theorem 9.9, but we can now discuss all this, from a more conceptual viewpoint, the idea being as follows:

- (1) We have indeed the following computation, using Theorem 9.22:

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

- (2) Once again by using Theorem 9.22, we have:

$$(\pm\sqrt{r}e^{it/2})^2 = (\sqrt{r})^2 e^{i(t/2+t/2)} = re^{it}$$

- (3) Given an arbitrary number $a \in \mathbb{R}$, we can define, as stated:

$$(re^{it})^a = r^ae^{ita}$$

Due to Theorem 9.22, this operation $x \rightarrow x^a$ is indeed the correct one.

- (4) This comes from the fact, that we know from Theorem 9.13, that the conjugation operation $x \rightarrow \bar{x}$ keeps the modulus, and switches the sign of the argument. \square

9d. Equations, roots

Getting back now to algebra, recall from Theorem 9.11 that any degree 2 equation has 2 complex roots. We can in fact prove that any polynomial equation, of arbitrary degree $N \in \mathbb{N}$, has exactly N complex solutions, counted with multiplicities:

THEOREM 9.24. *Any polynomial $P \in \mathbb{C}[X]$ decomposes as*

$$P = c(X - a_1) \dots (X - a_N)$$

with $c \in \mathbb{C}$ and with $a_1, \dots, a_N \in \mathbb{C}$.

PROOF. The problem is that of proving that our polynomial has at least one root, because afterwards we can proceed by recurrence. We prove this by contradiction. So, assume that P has no roots, and pick a number $z \in \mathbb{C}$ where $|P|$ attains its minimum:

$$|P(z)| = \min_{x \in \mathbb{C}} |P(x)| > 0$$

Since $Q(t) = P(z+t) - P(z)$ is a polynomial which vanishes at $t = 0$, this polynomial must be of the form $ct^k + \text{higher terms}$, with $c \neq 0$, and with $k \geq 1$ being an integer. We obtain from this that, with $t \in \mathbb{C}$ small, we have the following estimate:

$$P(z+t) \simeq P(z) + ct^k$$

Now let us write $t = rw$, with $r > 0$ small, and with $|w| = 1$. Our estimate becomes:

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

Now recall that we assumed $P(z) \neq 0$. We can therefore choose $w \in \mathbb{T}$ such that cw^k points in the opposite direction to that of $P(z)$, and we obtain in this way:

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

Now by choosing $r > 0$ small enough, as for the error in the first estimate to be small, and overcome by the negative quantity $-|c|r^k$, we obtain from this:

$$|P(z+rw)| < |P(z)|$$

But this contradicts our definition of $z \in \mathbb{C}$, as a point where $|P|$ attains its minimum. Thus P has a root, and by recurrence it has N roots, as stated. \square

All this is very nice, and we will see applications in a moment. As a word of warning, however, we should mention that the above result remains something quite theoretical. Indeed, the proof is by contradiction, and there is no way of recycling the material there into something explicit, that can be used for effectively computing the roots.

We will be back to this in the next chapter, with a number of sharp results on the subject, extending to degree 3 and 4 what we know well in degree 2.

As a last topic now regarding the equations and roots, which is something really beautiful, we have the roots of unity. Let us start with the following fact:

THEOREM 9.25. *The equation $x^N = 1$ has N complex solutions, namely*

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

which are called roots of unity of order N .

PROOF. This follows from the general multiplication formula for complex numbers from Theorem 9.22. Indeed, with $x = re^{it}$ our equation reads:

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

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

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

$N = 1$. Here the unique root of unity is 1.

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

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

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

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

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

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

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

The roots of unity are very useful variables, and have many interesting properties. As a first application, we can now solve the ambiguity questions related to the extraction of N -th roots, from Theorem 9.9 and Theorem 9.23, the statement being as follows:

THEOREM 9.26. *Any $x = re^{it}$ has exactly N roots of order N , which appear as*

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

multiplied by the N roots of unity of order N .

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

$$z^N = y^N$$

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

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

PROPOSITION 9.27. *The roots of unity, $\{w^k\}$ with $w = e^{2\pi i/N}$, have the property*

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

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

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

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

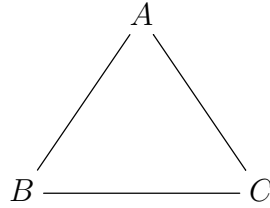
Now observe that in the case $N \nmid s$ our polygon P_s is non-degenerate, circling around the unit circle, and having center $C_s = 0$. As for the case $N \mid s$, here the polygon is degenerate, lying at 1, and having center $C_s = 1$. Thus, we have the following formula:

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

Thus, we obtain the formula in the statement. \square

In relation now with geometry, the roots of unity can be something very useful. In order to discuss some applications, using third roots of unity, let us start with:

PROPOSITION 9.28. *A triangle ABC , with A, B, C appearing counterclockwise*

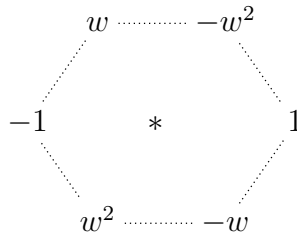


is equilateral precisely when its vertices, regarded as complex numbers, satisfy

$$A + wB + w^2C = 0$$

with $w = e^{2\pi i/3}$. When A, B, C appear clockwise, the same happens, with $w \rightarrow w^2$.

PROOF. The roots of unity of order 3, and their opposites, are as follows:



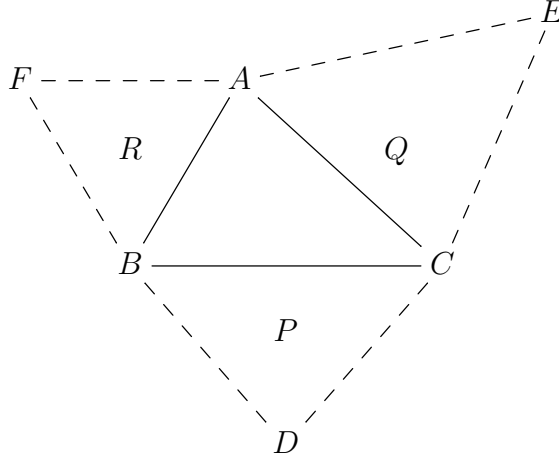
Thus the clockwise rotation by 60° is $P \rightarrow -wP$, and by using this, along with $1 + w + w^2$, coming from $w^3 = 1$, the condition for ABC to be equilateral reads:

$$\begin{aligned} A - C = -w(B - C) &\iff A + wB - (1 + w)C = 0 \\ &\iff A + wB + w^2C = 0 \end{aligned}$$

As for the last assertion, this follows from this, by interchanging $B \leftrightarrow C$. \square

As a main result now, following Napoleon, Fermat, Torricelli and others, we have:

THEOREM 9.29. *In the context of the Napoleon configuration, namely*



with equilateral triangles, and their barycenters drawn, the following happen,

- (1) *Napoleon theorem: the triangle PQR is equilateral.*
- (2) *Torricelli circles: the circles ABF, BCD, ACE are concurrent.*
- (3) *Torricelli point: AD, BE, CF cross, on this circle concurrence point.*
- (4) *More Torricelli: these lines AD, BE, CF cross at $60^\circ - 120^\circ$ angles.*
- (5) *Fermat point: the Torricelli point minimizes $AX + BX + CX$.*
- (6) *Napoleon point: AP, BQ, CR cross too, at the Napoleon point.*

with the assumption that all angles of ABC are $\leq 120^\circ$ being needed for (5).

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

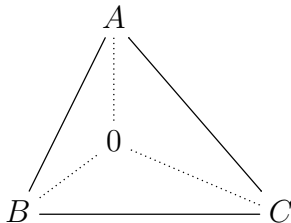
- (1) The Napoleon theorem comes, majestically, using Proposition 9.28, as follows:

$$\begin{aligned}
 P + wQ + w^2R &= \frac{B + C + D}{3} + w \cdot \frac{A + C + E}{3} + w^2 \cdot \frac{A + B + F}{3} \\
 &= \frac{B + wA + w^2F}{3} + \frac{C + wE + w^2A}{3} + \frac{D + wC + w^2B}{3} \\
 &= 0 + 0 + 0 \\
 &= 0
 \end{aligned}$$

(2,3,4) These assertions, which are all related, are all elementary, and follow from some angle hunting, without any major difficulty. We will leave them as exercises, for you.

(5) Let us define the Fermat point of a triangle ABC as being the point which minimizes $AX + BX + CX$, with the existence being clear, but with the uniqueness, not. Our claim is that when the triangle ABC has all angles $\leq 120^\circ$, this Fermat point is the

Torricelli point from (2,3,4), appearing as follows, with all angles around it being 120° :



In order to prove this, we use vector calculus. By fixing the origin 0 at the Torricelli point, as indicated above, we have the following estimate, for any point X in the plane, with i, j, k denoting the unit vectors along A, B, C , which satisfy $i + j + k = 0$:

$$\begin{aligned} \|A\| + \|B\| + \|C\| &= \langle A, i \rangle + \langle B, j \rangle + \langle C, k \rangle \\ &= \langle A - X, i \rangle + \langle B - X, j \rangle + \langle C - X, k \rangle \\ &\leq \|A - X\| + \|B - X\| + \|C - X\| \end{aligned}$$

Thus, claim proved. As for the case where one of the angles of ABC is $\geq 120^\circ$, here the Fermat point must be that vertex, and we will leave this as an exercise.

(6) Well, I must admit that I tried to prove this with my favorite plane geometry method, complex numbers, and failed, the computations being quite complicated. Moral of the story, not everyone is Napoleon, and I will leave this to you, as an exercise. \square

9e. Exercises

Welcome to the complex numbers, and as exercises about them, we have:

EXERCISE 9.30. *What happens to \mathbb{C} when representing it upside-down?*

EXERCISE 9.31. *What about drawing \mathbb{R} from right to left? Or doing both?*

EXERCISE 9.32. *Further explore the basic geometric series, for complex numbers.*

EXERCISE 9.33. *Study more the various possible proofs of $e^{it} = \cos t + i \sin t$.*

EXERCISE 9.34. *And, as mentioned above, do not forget to learn about $E = mc^2$ too.*

EXERCISE 9.35. *Study the real and complex solutions of $x_1 + \dots + x_N = 0$.*

EXERCISE 9.36. *Diagonalize the all-one matrix, over \mathbb{R} , and then over \mathbb{C} .*

EXERCISE 9.37. *Learn more about plane geometry, and the various triangle centers.*

As bonus exercise, you can start reading some complex analysis. All good learning.

CHAPTER 10

Polynomials

10a. Resultant, discriminant

We have seen that many questions lead us into computing roots of polynomials. Let us start with something that we know well, but is always good to remember:

PROPOSITION 10.1. *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. This is indeed something that we know well, coming from:

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

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

In degree 3 and higher, we would first like to understand what the analogue of the discriminant $\Delta = b^2 - 4ac$ is. In order to discuss this question, let us start with:

THEOREM 10.2. *Given a monic polynomial $P \in \mathbb{C}[X]$, factorized as*

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

the following happen:

- (1) *The coefficients of P are symmetric functions in a_1, \dots, a_k .*
- (2) *The symmetric functions in a_1, \dots, a_k are polynomials in the coefficients of P .*

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

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

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

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

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

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

(2) Conversely now, let us look at the symmetric functions in the roots a_1, \dots, 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, \dots, S_k . That is, the symmetric functions \mathcal{F} in our variables a_1, \dots, a_k , with integer coefficients, appear as follows:

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

(3) The point now is that, alternatively, the symmetric functions in our variables a_1, \dots, 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, \dots, f_k . That is, we have as well:

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

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

Next, we can decide whether two polynomials have a common root, as follows:

THEOREM 10.3. *Given two polynomials $P, Q \in \mathbb{C}[X]$, written as*

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

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 $c^l d^k$ 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) 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, \dots, 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.

(4) With this done, let us look now what happens with respect to the remaining variables b_1, \dots, b_l , which are the roots of Q . Once again what we have here are certain symmetric functions in these variables b_1, \dots, 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.

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

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:

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

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 10.4. *The resultant of two polynomials, written as*

$$P = p_k X^k + \dots + p_1 X + p_0 \quad , \quad Q = q_l X^l + \dots + q_1 X + q_0$$

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 quite clever, due to Sylvester, as follows:

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

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

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

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

(2) Our first claim is that with respect to the standard bases for all the vector spaces involved, namely those consisting of the monomials $1, X, X^2, \dots$, 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 10.2 and 10.3. \square

As an illustration for this, 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, regarding Δ , and we have:

THEOREM 10.5. *Given a polynomial $P \in \mathbb{C}[X]$, written as*

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

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 10.3, 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 written a bit informally, coming from Theorem 10.4:

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

Also, the fact that we have $\Delta(P) = 0$ precisely when P has a double root is clear from Theorem 10.3. 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 a first illustration, let us see what happens in degree 2. We have:

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

Thus, the discriminant is given by the following formula, as it should:

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

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 10.6. *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, \dots, 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 10.3, so let us see what that formula teaches us, in the case $Q = P'$. Let us write P, P' as follows:

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

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

According to Theorem 10.3, 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, \dots, 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:

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

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:

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

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

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

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 10.5, and then using the formula found in (3), we obtain:

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

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

As applications now, the formula in Theorem 10.6 is quite useful for the real polynomials $P \in \mathbb{R}[X]$ in small degree, because it allows us to say when the roots are real, or complex, or at least have some partial information about this. For instance, we have:

PROPOSITION 10.7. *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 we certainly know, coming from Proposition 10.1, but let us see how this comes via the formula in Theorem 10.6, 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 Proposition 10.1, 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 10.6, 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:

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

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

10b. Cardano formulae

Let us discuss now what exactly happens in degree 3. We must first compute the discriminant of arbitrary degree 3 polynomials, and the result here is as follows:

THEOREM 10.8. *The discriminant of a degree 3 polynomial,*

$$P = aX^3 + bX^2 + cX + d$$

is given by $\Delta(P) = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd$.

PROOF. We have two methods here, both instructive, and we will try them both:

(1) Let us first go the pedestrian way, based on the definition of the resultant, from Theorem 10.3. Consider two polynomials, of degree 3 and degree 2, written as follows:

$$P = aX^3 + bX^2 + cX + d$$

$$Q = eX^2 + fX + g = e(X - s)(X - t)$$

The resultant of these two polynomials is then given by:

$$\begin{aligned} R(P, Q) &= a^2e^3(p - s)(p - t)(q - s)(q - t)(r - s)(r - t) \\ &= a^2 \cdot e(p - s)(p - t) \cdot e(q - s)(q - t) \cdot e(r - s)(r - t) \\ &= a^2Q(p)Q(q)Q(r) \\ &= a^2(ep^2 + fp + g)(eq^2 + fq + g)(er^2 + fr + g) \end{aligned}$$

Next, we can use the following formulae, in order to get rid of p, q, r :

$$p + q + r = -\frac{b}{a} \quad , \quad pq + pr + qr = \frac{c}{a} \quad , \quad pqr = -\frac{d}{a}$$

Indeed, by using these formulae, we obtain, after some routine work:

$$\begin{aligned} R(P, Q) &= d^2e^3 - cde^2f + c^2e^2g - 2bde^2g + bdef^2 - bcefg + 3adefg \\ &\quad - adf^3 + b^2eg^2 - 2aceg^2 + acf^2g - abfg^2 + a^2g^3 \end{aligned}$$

Getting back now to our discriminant problem, with $Q = P'$, which corresponds to $e = 3a$, $f = 2b$, $g = c$, we obtain the following formula:

$$\begin{aligned} R(P, P') &= 27a^3d^2 - 18a^2bcd + 9a^2c^3 - 18a^2bcd + 12ab^3d - 6ab^2c^2 + 18a^2bcd \\ &\quad - 8ab^3d + 3ab^2c^2 - 6a^2c^3 + 4ab^2c^2 - 2ab^2c^2 + a^2c^3 \end{aligned}$$

By simplifying now terms, and dividing by $-a$, we get the formula in the statement.

(2) Let us see as well how the computation goes, by using Theorem 10.4, which is our most advanced tool, so far. Consider a polynomial of degree 3, and its derivative:

$$P = aX^3 + bX^2 + cX + d$$

$$P' = 3aX^2 + 2bX + c$$

By using Theorem 10.4 and computing the determinant, we obtain:

$$\begin{aligned} R(P, P') &= \begin{vmatrix} a & 3a & & & \\ b & a & 2b & 3a & \\ c & b & c & 2b & 3a \\ d & c & & c & 2b \\ & d & & & c \end{vmatrix} \\ &= \begin{vmatrix} a & & & & \\ b & a & -b & 3a & \\ c & b & -2c & 2b & 3a \\ d & c & -3d & c & 2b \\ & d & & & c \end{vmatrix} \\ &= a \begin{vmatrix} a & -b & 3a & & \\ b & -2c & 2b & 3a & \\ c & -3d & c & 2b & \\ d & & & c & \end{vmatrix} \\ &= -ad \begin{vmatrix} -b & 3a & & \\ -2c & 2b & 3a & \\ -3d & c & 2b & \end{vmatrix} + ac \begin{vmatrix} a & -b & 3a & \\ b & -2c & 2b & \\ c & -3d & c & \end{vmatrix} \\ &= -ad(-4b^3 - 27a^2d + 12abc + 3abc) \\ &\quad + ac(-2ac^2 - 2b^2c - 9abd + 6ac^2 + b^2c + 6abd) \\ &= a(4b^3d + 27a^2d^2 - 15abcd + 4ac^3 - b^2c^2 - 3abcd) \\ &= a(4b^3d + 27a^2d^2 - 18abcd + 4ac^3 - b^2c^2) \end{aligned}$$

Now according to Theorem 10.5, the discriminant of our polynomial is given by:

$$\begin{aligned}\Delta(P) &= -\frac{R(P, P')}{a} \\ &= -4b^3d - 27a^2d^2 + 18abcd - 4ac^3 + b^2c^2 \\ &= b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd\end{aligned}$$

Thus, we have again obtained the formula in the statement. \square

Still talking degree 3 equations, let us try now to solve such an equation $P = 0$, with $P = aX^3 + bX^2 + cX + d$ as above. By linear transformations we can assume $a = 1, b = 0$, and then it is convenient to write $c = 3p, d = 2q$. Thus, our equation becomes:

$$x^3 + 3px + 2q = 0$$

And, regarding such equations, we have the following famous result of Cardano:

THEOREM 10.9. *For a normalized degree 3 equation, namely*

$$x^3 + 3px + 2q = 0$$

the discriminant is $\Delta = -108(p^3 + q^2)$. Assuming $p, q \in \mathbb{R}$ and $\Delta < 0$, the numbers

$$x = w \sqrt[3]{-q + \sqrt{p^3 + q^2}} + w^2 \sqrt[3]{-q - \sqrt{p^3 + q^2}}$$

with $w = 1, e^{2\pi i/3}, e^{4\pi i/3}$ are the solutions of our equation.

PROOF. The formula of Δ comes from Theorem 10.8, with $108 = 4 \times 27$. Now with x as in the statement, by using $(a + b)^3 = a^3 + b^3 + 3ab(a + b)$, we have:

$$\begin{aligned}x^3 &= \left(w \sqrt[3]{-q + \sqrt{p^3 + q^2}} + w^2 \sqrt[3]{-q - \sqrt{p^3 + q^2}} \right)^3 \\ &= -2q + 3 \sqrt[3]{-q + \sqrt{p^3 + q^2}} \cdot \sqrt[3]{-q - \sqrt{p^3 + q^2}} \cdot x \\ &= -2q + 3 \sqrt[3]{q^2 - p^3 - q^2} \cdot x \\ &= -2q - 3px\end{aligned}$$

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

As a comment here, the formula for roots in Theorem 10.9 holds in the case $\Delta > 0$ too, and also when the coefficients are complex numbers, $p, q \in \mathbb{C}$. However, these extensions are usually not very useful, because when it comes to extract the above square and cubic roots, you can well end up with the initial question, the one that you started with.

In higher degree things become quite complicated. In degree 4, to start with, we first have the following result, dealing with the discriminant and its applications:

THEOREM 10.10. *The discriminant of $P = ax^4 + bx^3 + cx^2 + dx + e$ is given by:*

$$\begin{aligned}\Delta = & 256a^3e^3 - 192a^2bde^2 - 128a^2c^2e^2 + 144a^2cd^2e - 27a^2d^4 \\ & + 144ab^2ce^2 - 6ab^2d^2e - 80abc^2de + 18abcd^3 + 16ac^4e \\ & - 4ac^3d^2 - 27b^4e^2 + 18b^3cde - 4b^3d^3 - 4b^2c^3e + b^2c^2d^2\end{aligned}$$

In the case $\Delta < 0$ we have 2 real roots and 2 complex conjugate roots, and in the case $\Delta > 0$ the roots are either all real or all complex.

PROOF. The formula of Δ follows from the definition of the discriminant, from Theorem 10.5, with the resultant computed via Theorem 10.4, as follows:

$$\Delta = \frac{1}{a} \begin{vmatrix} a & & & 4a & & & \\ b & a & & 3b & 4a & & \\ c & b & a & 2c & 3b & 4a & \\ d & c & b & d & 2c & 3b & 4a \\ e & d & c & & d & 2c & 3b \\ & e & d & & & d & 2c \\ & & e & & & & d \end{vmatrix}$$

As for the last assertion, the study here is routine, a bit as in degree 3. \square

In practice, as in degree 3, we can do first some manipulations on our polynomials, as to have them in simpler form, and we have the following version of Theorem 10.10:

PROPOSITION 10.11. *The discriminant of $P = x^4 + cx^2 + dx + e$, normalized degree 4 polynomial, is given by the following formula:*

$$\Delta = 16c^4e - 4c^3d^2 - 128c^2e^2 + 144cd^2e - 27d^4 + 256e^3$$

As before, if $\Delta < 0$ we have 2 real roots and 2 complex conjugate roots, and if $\Delta > 0$ the roots are either all real or all complex.

PROOF. This is a consequence of Theorem 10.10, with $a = 1, b = 0$, but we can deduce this as well directly. Indeed, the formula of Δ follows, quite easily, from:

$$\Delta = \begin{vmatrix} 1 & & & 4 & & & \\ & 1 & & & 4 & & \\ c & & 1 & 2c & & 4 & \\ d & c & & d & 2c & & 4 \\ e & d & c & & d & 2c & \\ & e & d & & & d & 2c \\ & & e & & & & d \end{vmatrix}$$

As for the last assertion, this is something that we know, from Theorem 10.10. \square

Along the same lines, we can further simplify what we have, as follows:

THEOREM 10.12. *The discriminant of a normalized degree 4 polynomial, written as*

$$P = x^4 + 6px^2 + 4qx + 3r$$

is given by the following formula:

$$\Delta = 256 \times 27 \times (9p^4r - 2p^3q^2 - 6p^2r^2 + 6pq^2r - q^4 + r^3)$$

In the case $\Delta < 0$ we have 2 real roots and 2 complex conjugate roots, and in the case $\Delta > 0$ the roots are either all real or all complex.

PROOF. This follows from Proposition 10.11, with $c = 6p, d = 4q, e = 3r$, but we can deduce this as well directly. Indeed, the formula of Δ follows, quite easily, from:

$$\Delta = \begin{vmatrix} 1 & & & & 4 & & & & \\ & 1 & & & & 4 & & & \\ 6p & & 1 & 12p & & & 4 & & \\ 4q & 6p & & 4q & 12p & & & 4 & \\ 3r & 4q & 6p & & 4q & 12p & & & \\ & 3r & 4q & & & 4q & 12p & & \\ & & 3r & & & & 4q & & \end{vmatrix}$$

As for the last assertion, this is something that we know from Theorem 10.10. □

Time now to get to the real thing, solving the equation. We have here:

THEOREM 10.13. *The roots of a normalized degree 4 equation, written as*

$$x^4 + 6px^2 + 4qx + 3r = 0$$

are as follows, with y satisfying the equation $(y^2 - 3r)(y - 3p) = 2q^2$,

$$x_1 = \frac{1}{\sqrt{2}} \left(-\sqrt{y - 3p} + \sqrt{-y - 3p + \frac{4q}{\sqrt{2y - 6p}}} \right)$$

$$x_2 = \frac{1}{\sqrt{2}} \left(-\sqrt{y - 3p} - \sqrt{-y - 3p + \frac{4q}{\sqrt{2y - 6p}}} \right)$$

$$x_3 = \frac{1}{\sqrt{2}} \left(\sqrt{y - 3p} + \sqrt{-y - 3p - \frac{4q}{\sqrt{2y - 6p}}} \right)$$

$$x_4 = \frac{1}{\sqrt{2}} \left(\sqrt{y - 3p} - \sqrt{-y - 3p - \frac{4q}{\sqrt{2y - 6p}}} \right)$$

and with y being computable via the Cardano formula.

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

(1) To start with, let us write our equation in the following form:

$$x^4 = -6px^2 - 4qx - 3r$$

The idea will be that of adding a suitable common term, to both sides, as to make square on both sides, as to eventually end with a sort of double quadratic equation. For this purpose, our claim is that what we need is a number y satisfying:

$$(y^2 - 3r)(y - 3p) = 2q^2$$

Indeed, assuming that we have this number y , our equation becomes:

$$\begin{aligned} (x^2 + y)^2 &= x^4 + 2x^2y + y^2 \\ &= -6px^2 - 4qx - 3r + 2x^2y + y^2 \\ &= (2y - 6p)x^2 - 4qx + y^2 - 3r \\ &= (2y - 6p)x^2 - 4qx + \frac{2q^2}{y - 3p} \\ &= \left(\sqrt{2y - 6p} \cdot x - \frac{2q}{\sqrt{2y - 6p}} \right)^2 \end{aligned}$$

(2) Which looks very good, leading us to the following degree 2 equations:

$$x^2 + y + \sqrt{2y - 6p} \cdot x - \frac{2q}{\sqrt{2y - 6p}} = 0$$

$$x^2 + y - \sqrt{2y - 6p} \cdot x + \frac{2q}{\sqrt{2y - 6p}} = 0$$

Now let us write these two degree 2 equations in standard form, as follows:

$$x^2 + \sqrt{2y - 6p} \cdot x + \left(y - \frac{2q}{\sqrt{2y - 6p}} \right) = 0$$

$$x^2 - \sqrt{2y - 6p} \cdot x + \left(y + \frac{2q}{\sqrt{2y - 6p}} \right) = 0$$

(3) Regarding the first equation, the solutions there are as follows:

$$x_1 = \frac{1}{2} \left(-\sqrt{2y - 6p} + \sqrt{-2y - 6p + \frac{8q}{\sqrt{2y - 6p}}} \right)$$

$$x_2 = \frac{1}{2} \left(-\sqrt{2y - 6p} - \sqrt{-2y - 6p + \frac{8q}{\sqrt{2y - 6p}}} \right)$$

As for the second equation, the solutions there are as follows:

$$x_3 = \frac{1}{2} \left(\sqrt{2y - 6p} + \sqrt{-2y - 6p - \frac{8q}{\sqrt{2y - 6p}}} \right)$$

$$x_4 = \frac{1}{2} \left(\sqrt{2y - 6p} - \sqrt{-2y - 6p - \frac{8q}{\sqrt{2y - 6p}}} \right)$$

(4) Now by cutting a $\sqrt{2}$ factor from everything, this gives the formulae in the statement. As for the last claim, regarding the nature of y , this comes from Cardano. \square

We still have to compute the number y appearing in the above via Cardano, and the result here, adding to what we already have in Theorem 10.13, is as follows:

THEOREM 10.14 (continuation). *The value of y in the previous theorem is*

$$y = t + p + \frac{a}{t}$$

where the number t is given by the formula

$$t = \sqrt[3]{b + \sqrt{b^2 - a^3}}$$

with $a = p^2 + r$ and $b = 2p^2 - 3pr + q^2$.

PROOF. The legend has it that this is what comes from Cardano, but depressing and normalizing and solving $(y^2 - 3r)(y - 3p) = 2q^2$ makes it for too many operations, so the most pragmatic is to simply check this equation. With y as above, we have:

$$\begin{aligned} y^2 - 3r &= t^2 + 2pt + (p^2 + 2a) + \frac{2pa}{t} + \frac{a^2}{t^2} - 3r \\ &= t^2 + 2pt + (3p^2 - r) + \frac{2pa}{t} + \frac{a^2}{t^2} \end{aligned}$$

With this in hand, we have the following computation:

$$\begin{aligned} (y^2 - 3r)(y - 3p) &= \left(t^2 + 2pt + (3p^2 - r) + \frac{2pa}{t} + \frac{a^2}{t^2} \right) \left(t - 2p + \frac{a}{t} \right) \\ &= t^3 + (a - 4p^2 + 3p^2 - r)t + (2pa - 6p^3 + 2pr + 2pa) \\ &\quad + (3p^2a - ra - 4p^2a + a^2)\frac{1}{t} + \frac{a^3}{t^3} \\ &= t^3 + (a - p^2 - r)t + 2p(2a - 3p^2 + r) + a(a - p^2 - r)\frac{1}{t} + \frac{a^3}{t^3} \\ &= t^3 + 2p(-p^2 + 3r) + \frac{a^3}{t^3} \end{aligned}$$

Now by using the formula of t in the statement, this gives:

$$\begin{aligned}
 (y^2 - 3r)(y - 3p) &= b + \sqrt{b^2 - a^3} - 4p^2 + 6pr + \frac{a^3}{b + \sqrt{b^2 - a^3}} \\
 &= b + \sqrt{b^2 - a^3} - 4p^2 + 6pr + b - \sqrt{b^2 - a^3} \\
 &= 2b - 4p^2 + 6pr \\
 &= 2(2p^2 - 3pr + q^2) - 4p^2 + 6pr \\
 &= 2q^2
 \end{aligned}$$

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

Finally, in degree 5 and more, we already know from chapter 8, as a consequence of Galois theory, that there is no explicit formula for the roots. Thus, discussion closed about polynomials and their roots, at least at the general level. Very nice.

10c. Spirals, lemniscates

Moving on, no discussion about polynomials, be them real or complex, would be complete without a word on the various plane curves, and basic algebraic geometry in general. Let us start with the following notion, that we know well since chapter 8:

DEFINITION 10.15. *An algebraic curve in \mathbb{R}^2 is the vanishing set*

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

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

We already know a bit about such curves from chapter 8, where we discussed the conics, in some detail. So, as a first question now, what results from what we know about conics have a chance to be relevant to the arbitrary algebraic curves?

And here, normally none, because the conics are obviously very particular curves, having very particular properties. Let us record however a useful statement here, regarding the conics, which can be a useful source of inspiration later, as follows:

PROPOSITION 10.16. *The conics can be written in cartesian, polar, parametric or complex coordinates, with the equations for the unit circle being*

$$x^2 + y^2 = 1 \quad , \quad r = 1 \quad , \quad x = \cos t, y = \sin t \quad , \quad |z| = 1$$

and with the equations for ellipses, parabolas and hyperbolas being similar.

PROOF. The equations for the circle are clear, and we will leave as an instructive exercise working out those for ellipses, parabolas and hyperbolas. \square

As a true answer to our question now, coming this time from a very modest conic, namely $xy = 0$, that we dismissed in chapter 8 as being “degenerate”, we have:

THEOREM 10.17. *The following happen, for curves C defined by polynomials P :*

- (1) *In degree $d = 2$, curves can have singularities, such as $xy = 0$ at $(0, 0)$.*
- (2) *In general, assuming $P = P_1 \dots P_k$, we have $C = C_1 \cup \dots \cup C_k$.*
- (3) *A union of curves $C_i \cup C_j$ is generically non-smooth, unless disjoint.*
- (4) *Due to this, we say that C is non-degenerate when P is irreducible.*

PROOF. All this is self-explanatory, the details being as follows:

(1) This is something obvious, just the story of two lines crossing.

(2) This comes from the following trivial fact, with the notation $z = (x, y)$:

$$P_1 \dots P_k(z) = 0 \iff P_1(z) = 0, \text{ or } P_2(z) = 0, \dots, \text{ or } P_k(z) = 0$$

(3) This is something very intuitive, and it actually takes a bit of time to imagine a situation where $C_1 \cap C_2 \neq \emptyset$, $C_1 \not\subset C_2$, $C_2 \not\subset C_1$, but $C_1 \cup C_2$ is smooth. In practice now, “generically” has of course a mathematical meaning, in relation with probability, and our assertion does say something mathematical, that we are supposed to prove. But, we will not insist on this, and leave this as an instructive exercise, precise formulation of the claim, and its proof, in the case you are familiar with probability theory.

(4) This is just a definition, based on the above, that we will use in what follows. \square

With degree 1 and 2 investigated, and our conclusions recorded, let us get now to degree 3, see what new phenomena appear here. And here, to start with, we have the following remarkable curve, well-known from calculus, because 0 is not a maximum or minimum of the function $x \rightarrow y$, despite the derivative vanishing there:

$$x^3 = y$$

Also, in relation with set theory and logic, and with the foundations of mathematics in general, we have the following curve, which looks like the emptyset \emptyset :

$$(x - y)(x^2 + y^2 - 1) = 0$$

But, it is not about counterexamples to calculus, or about logic, that we want to talk about here. As a first truly remarkable degree 3 curve, or cubic, we have the cusp:

PROPOSITION 10.18. *The standard cusp, which is the cubic given by*

$$x^3 = y^2$$

has a singularity at $(0, 0)$, with only 1 tangent line at that singularity.

PROOF. The two branches of the cusp are indeed both tangent to Ox , because:

$$y' = \pm \frac{3}{2} \sqrt{x} \implies y'(0) = 0$$

Observe also that what happens for the cusp is different from what happens for $xy = 0$, precisely because we have 1 line tangent at the singularity, instead of 2. \square

As a second remarkable cubic, which gets the crown, and the right to have a Theorem about it, we have the Tschirnhausen curve, which is as follows:

THEOREM 10.19. *The Tschirnhausen cubic, given by the following equation,*

$$x^3 = x^2 - 3y^2$$

makes the dream of $xy = 0$ come true, by self-intersecting, and being non-degenerate.

PROOF. This is something self-explanatory, by drawing a picture, but there are several other interesting things that can be said about this curve, as follows:

(1) Let us start with the curve written in polar coordinates as follows:

$$r \cos^3 \left(\frac{\theta}{3} \right) = a$$

With $t = \tan(\theta/3)$, the equations of the coordinates are as follows:

$$x = a(1 - 3t^2) \quad , \quad y = at(3 - t^2)$$

Now by eliminating t , we reach to the following equation:

$$(a - x)(8a + x)^2 = 27ay^2$$

(2) By translating horizontally by $8a$, and changing signs of variables, we have:

$$x = 3a(3 - t^2) \quad , \quad y = at(3 - t^2)$$

Now by eliminating t , we reach to the following equation:

$$x^3 = 9a(x^2 - 3y^2)$$

But with $a = 1/9$ this is precisely the equation in the statement. □

In degree 4 now, quartics, we have enough dimensions for “improving” the cusp and the Tschirnhausen curve. First we have the cardioid, which is as follows:

PROPOSITION 10.20. *The cardioid, which is a quartic, given in polar coordinates by*

$$2r = a(1 - \cos \theta)$$

makes the dream of $x^3 = y^2$ come true, by being a closed curve, with a cusp.

PROOF. As before with the Tschirnhausen curve, this is something self-explanatory, by drawing a picture, but there are several things that must be said, as follows:

(1) The cardioid appears by definition by rolling a circle of radius $c > 0$ around another circle of same radius $c > 0$. With θ being the rolling angle, we have:

$$x = 2c(1 - \cos \theta) \cos \theta \quad , \quad y = 2c(1 - \cos \theta) \sin \theta$$

(2) Thus, in polar coordinates we get the equation in the statement, with $a = 4c$:

$$r = 2c(1 - \cos \theta)$$

(3) Finally, in cartesian coordinates, the equation is as follows:

$$(x^2 + y^2)^2 + 4cx(x^2 + y^2) = 4c^2y^2$$

Thus, what we have is indeed a degree 4 curve, as claimed. \square

Still in degree 4, the crown gets to the Bernoulli lemniscate, which is as follows:

THEOREM 10.21. *The Bernoulli lemniscate, a quartic, which is given by*

$$r^2 = a^2 \cos 2\theta$$

makes the dream of $x^3 = x^2 - 3y^2$ come true, by being closed, and self-intersecting.

PROOF. As usual, this is something self-explanatory, by drawing a picture, which looks like ∞ , but there are several other things that must be said, as follows:

(1) In cartesian coordinates, the equation is as follows, with $a^2 = 2c^2$:

$$(x^2 + y^2)^2 = c^2(x^2 - y^2)$$

(2) Also, we have the following nice complex reformulation of this equation:

$$|z + c| \cdot |z - c| = c^2$$

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

In degree 5, in the lack of any spectacular quintic, let us record:

THEOREM 10.22. *Unlike in degree 3, 4, where equations can be solved, by the Cardano formula, in degree 5 this generically does not happen, an example being*

$$x^5 - x - 1 = 0$$

having Galois group S_5 , not solvable. Geometrically, this tells us that the intersection of the quintic $y = x^5 - x - 1$ with the line $y = 0$ cannot be computed.

PROOF. Obviously off-topic, but with no good quintic available, and still a few more minutes before the bell ringing, I had to improvise a bit, and tell you about this. \square

Back now to our usual business, in degree 6, sextics, we first have here:

PROPOSITION 10.23. *The trefoil sextic, or Kiepert curve, which is given by*

$$r^3 = a^3 \cos 3\theta$$

looks like a trefoil, closed curve, with a triple self-intersection.

PROOF. As before, drawing a picture is mandatory. With $z = re^{i\theta}$ we have:

$$\begin{aligned}
 r^3 = a^3 \cos 3\theta &\iff r^3 \cos 3\theta = \left(\frac{r^2}{a}\right)^3 \\
 &\iff z^3 + \bar{z}^3 = 2\left(\frac{z\bar{z}}{a}\right)^3 \\
 &\iff (x+iy)^3 + (x-iy)^3 = 2\left(\frac{x^2+y^2}{a}\right)^3 \\
 &\iff x^3 - 3xy^2 = \left(\frac{x^2+y^2}{a}\right)^3 \\
 &\iff (x^2+y^2)^3 = a^3(x^3 - 3xy^2)
 \end{aligned}$$

Thus, we have indeed a sextic, as claimed. \square

We also have in degree 6 the most beautiful of curves them all, the Cayley sextic:

THEOREM 10.24. *The Cayley sextic, given in polar coordinates by*

$$r = a \cos^3\left(\frac{\theta}{3}\right)$$

makes the dream of everyone come true, by looking like a self-intersecting heart.

PROOF. As before, picture mandatory. With $z = re^{i\theta}$ and $u = z^{1/3}$ we have:

$$\begin{aligned}
 r = a \cos^3\left(\frac{\theta}{3}\right) &\iff ar \cos^3\left(\frac{\theta}{3}\right) = r^2 \\
 &\iff a\left(\frac{u+\bar{u}}{2}\right)^3 = r^2 \\
 &\iff a(u^3 + \bar{u}^3 + 3u\bar{u}(u+\bar{u})) = 8r^2 \\
 &\iff 3au\bar{u} \cdot \frac{u+\bar{u}}{2} = 4r^2 - ax \\
 &\iff 27a^3r^6 \cdot \frac{r^2}{a} = (4r^2 - ax)^3 \\
 &\iff 27a^2(x^2+y^2)^2 = (4x^2+4y^2-ax)^3
 \end{aligned}$$

Thus, we have indeed a sextic, as claimed. \square

And with this, end of our study regarding the low degree plane curves. We have learned many interesting things, and we insist, spend some more time on all this, by drawing pictures for all the above curves, and learning more, about each of them.

Quite remarkably, most of the above curves are sinusoidal spirals, in the following sense, and with actually the term “sinusoidal spiral” being a bit unfortunate:

THEOREM 10.25. *The sinusoidal spirals, which are as follows,*

$$r^n = a^n \cos n\theta$$

with $a \neq 0$ and $n \in \mathbb{Q} - \{0\}$, include the following curves:

- (1) $n = -1$ line.
- (2) $n = 1$ circle, $n = -1/2$ parabola, $n = -2$ hyperbola.
- (3) $n = -3$ Humbert cubic, $n = -1/3$ Tschirnhausen curve.
- (4) $n = 1/2$ cardioid, $n = 2$ Bernoulli lemniscate.
- (5) $n = 3$ Kiepert trefoil, $n = 1/3$ Cayley sextic.

PROOF. We first have to prove that the sinusoidal spirals are indeed algebraic curves. But this is best done by using the complex coordinate $z = re^{i\theta}$, as follows:

$$\begin{aligned} r^n = a^n \cos n\theta &\iff r^n \cos n\theta = \left(\frac{r^2}{a}\right)^n \\ &\iff z^n + \bar{z}^n = 2\left(\frac{z\bar{z}}{a}\right)^n \\ &\iff (x + iy)^n + (x - iy)^n = 2\left(\frac{x^2 + y^2}{a}\right)^n \end{aligned}$$

As a first observation now, in the case $n \in \mathbb{N}$ we can simply use the binomial formula, and we get an algebraic equation of degree $2n$, as follows:

$$\sum_{k=0}^{[n/2]} (-1)^k \binom{n}{2k} x^{n-2k} y^{2k} = \left(\frac{x^2 + y^2}{a}\right)^n$$

In general, things are a bit more complicated, as shown for instance by our computation for the Cayley sextic. However, the same idea as there applies, and we are led in this way to the equation of an algebraic curve, as claimed. Regarding now the examples:

- (1) At $n = -1$ the equation is as follows, producing a line:

$$r \cos \theta = a \iff x = a$$

- (2) At $n = 1$ the equation is as follows, producing a circle:

$$r = a \cos \theta \iff r^2 = ax \iff x^2 + y^2 = ax$$

- (3) At $n = -1/2$ the equation is as follows, producing a parabola:

$$a = r \cos^2(\theta/2) \iff r + x = 2a \iff y^2 = 4a(a - x)$$

- (4) At $n = -2$ the equation is as follows, producing a hyperbola:

$$a^2 = r \cos^2 2\theta \iff a^2 = 2x^2 - r^2 \iff (x + y)(x - y) = a^2$$

(5) At $n = -3$ the equation is as follows, producing a curve with 3 components, which looks like some sort of “trivalent hyperbola”, called Humbert cubic:

$$r^3 \cos 3\theta = a^3 \iff z^3 + \bar{z}^3 = 2a^3 \iff x^3 - 3xy^2 = a^3$$

(6) As for the other curves, this follows from our various formulae above. \square

As another twist to the story, the sinusoidal spirals with $n \in \mathbb{N}$ are also part of another remarkable family of plane algebraic curves, going back to Cassini, as follows:

THEOREM 10.26. *The polynomial lemniscates, which are as follows,*

$$|P(z)| = b^n$$

with $P \in \mathbb{C}[X]$ having n distinct roots, and $b > 0$, include the following curves:

- (1) *The sinusoidal spirals with $n \in \mathbb{N}$, including the $n = 1$ circle, $n = 2$ Bernoulli lemniscate, and $n = 3$ Kiepert trefoil.*
- (2) *The Cassini ovals, which are the quartics given by $|z + c| \cdot |z - c| = b^2$, covering too the Bernoulli lemniscate, appearing at $b = c$.*

PROOF. This is something quite self-explanatory, the details being as follows:

(1) Regarding the sinusoidal spirals with $n \in \mathbb{N}$, their equation is, with $a^n = 2c^n$:

$$\begin{aligned} a^n \frac{z^n + \bar{z}^n}{2} = (z\bar{z})^n &\iff c^n(z^n + \bar{z}^n) = (z\bar{z})^n \\ &\iff (z^n - c^n)(\bar{z}^n - c^n) = c^{2n} \\ &\iff |z^n - c^n| = c^n \end{aligned}$$

(2) Regarding the Cassini ovals, these correspond to the case where the polynomial $P \in \mathbb{C}[X]$ has degree 2. In general, the equation for the Cassini ovals is:

$$\begin{aligned} |z + c| \cdot |z - c| = b^2 &\iff |z^2 - c^2| = b^2 \\ &\iff (z^2 - c^2)(\bar{z}^2 - c^2) = b^4 \\ &\iff (z\bar{z})^2 - c^2(z^2 + \bar{z}^2) + c^4 = b^4 \\ &\iff (x^2 + y^2)^2 - c^2(x^2 - y^2) + c^4 = b^4 \\ &\iff (x^2 + y^2)^2 = c^2(x^2 - y^2) + b^4 - c^4 \end{aligned}$$

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

10d. Nullstellensatz

Getting now to arbitrary fields, and arbitrary dimensions too, let us recall from chapter 8 that the fundamental question, establishing a bijection between ideals $I \subset F[x_1, \dots, x_N]$ and algebraic manifolds $X \subset F^N$, basically reduces to the question of deciding whether, for an ideal $I \subset F[x_1, \dots, x_N]$, the following inclusion is an equality or not:

$$\sqrt{I} \subset I_{X_I}$$

We will see that when F is algebraically closed, this inclusion is indeed an equality, with the result being called Hilbert's Nullstellensatz theorem. Getting started now, let us first establish a weak version of the Nullstellensatz, as follows:

THEOREM 10.27 (Weak Nullstellensatz). *If F is algebraically closed, we have*

$$X_I \neq \emptyset$$

for any proper ideal $I \subset F[x_1, \dots, x_N]$.

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

(1) As a first observation, we have indeed here a Weak Nullstellensatz, because when assuming that the above-mentioned Nullstellensatz holds, we have:

$$\begin{aligned} X_I = \emptyset &\implies I_{X_I} = F[x_1, \dots, x_N] \\ &\implies \sqrt{I} = F[x_1, \dots, x_N] \\ &\implies I = F[x_1, \dots, x_N] \end{aligned}$$

(2) As a second observation, the assumption that F is algebraically closed is really needed, because otherwise we can come with polynomials of type $P = X^2 + 1$, say when $F = \mathbb{R}$, having no zeroes, and so with ideals of type $I = (P) \in F[X]$, with $X_I = \emptyset$.

(3) As a third and last observation, our assumption that F is algebraically closed tells us that any $P \in F[X]$ has zeroes, and based on this, we want to prove that any $I \subset F[x_1, \dots, x_N]$ has zeroes, $X_I \neq \emptyset$. Which sounds like a quite plausible claim.

(4) Getting to work now, our precise claim, which will prove our theorem, simply by replacing $I \subset F[x_1, \dots, x_N]$ with a maximal ideal containing it, is that the maximal ideals $I \subset F[x_1, \dots, x_N]$ are precisely those of the following form, with $a_1, \dots, a_N \in F$:

$$I = (x_1 - a_1, \dots, x_N - a_N)$$

(5) In order to prove this latter claim, let us pick a maximal ideal $I \subset F[x_1, \dots, x_N]$, and consider the following quotient, that we know to be a field:

$$K = F[x_1, \dots, x_N]/I$$

Our claim in (4), namely $I = (x_1 - a_1, \dots, x_N - a_N)$, is then equivalent to:

$$K \simeq F$$

Now since F was assumed to be algebraically closed, proving this amounts in proving that K is algebraic over F . And this is what we will prove, by contradiction.

(6) So, assume that K is purely transcendental over F . By reordering the variables x_1, \dots, x_N , we can assume that $x_1, \dots, x_k \in K$ are algebraically independent over F , and that $x_{k+1}, \dots, x_N \in K$ are algebraic over the following subfield:

$$L = K(x_1, \dots, x_k) \subset K$$

Observe now that K is finitely generated as a L -module. Our claim, based on this, and which will easily prove the theorem, is that L is finitely generated, as a F -algebra.

(7) In short, we are in need here of some commutative algebra input. Inspired by the above, consider a Noetherian ring R , and an intermediate ring as follows:

$$R \subset S \subset R[x_1, \dots, x_N]$$

Our claim is that if $R[x_1, \dots, x_N]$ is finitely generated as S -module, then S is finitely generated as S -algebra. Observe that this will prove indeed our claim in (6).

(8) So, let us prove this. For this purpose, let us pick a family of S -module generators $y_1, \dots, y_m \in R[x_1, \dots, x_N]$, and write formulae as follows, with $a_{ij}, b_{ijk} \in S$:

$$x_i = \sum_j a_{ij} y_j \quad , \quad y_i y_j = \sum_k b_{ijk} y_k$$

Now if we set $T = \langle a_{ij}, b_{ijk} \rangle$, this ring being finitely generated over R , it is Noetherian, and since a submodule of a finitely generated module over a Noetherian ring is finitely generated, with this being something general, and elementary, it follows that S is a finitely generated T -module, and so is a finitely generated R -algebra, as claimed.

(9) With this in hand, let us get back to our proof of the Weak Nullstellensatz. Our claim at the end of (6) is now proved, so let us pick algebra generators $z_1, \dots, z_l \in K$, and write these generators as quotients of polynomials, as follows:

$$z_i = \frac{P_i}{Q_i}$$

(10) Now observe that given any irreducible polynomial $P \in F[x_1, \dots, x_k]$, the quotient $1/P$ must be a polynomial in the rational functions z_i , and so P must divide at least one Q_i . Thus, we can only have finitely many irreducible polynomials $P \in F[x_1, \dots, x_k]$, and with this being wrong at $k \geq 1$, we have reached to a contradiction, as desired. \square

Still with me I hope, after all this algebra. We can now formulate a main result, namely the Hilbert Nullstellensatz, in its general form, as follows:

THEOREM 10.28 (Nullstellensatz). *If F is algebraically closed, we have a bijection*

$$\left(X \subset F^N \right) \longleftrightarrow \left(I \subset F[x_1, \dots, x_N] \right)$$

between algebraic manifolds in F^N , and radical ideals of $F[x_1, \dots, x_N]$.

PROOF. We want to prove that we have $I_{X_I} = \sqrt{I}$, for any ideal $I \subset F[x_1, \dots, x_N]$, and this can be deduced from the Weak Nullstellensatz, as follows:

(1) To start with, let us first recall that we trivially have $\sqrt{I} \subset I_{X_I}$, and also that what we want to prove is stronger than the Weak Nullstellensatz.

(2) In practice, we want to prove that given an ideal $I \subset F[x_1, \dots, x_N]$, any polynomial $P \in F[x_1, \dots, x_N]$ vanishing on X_I has the property $P^m \in I$, for some $m \in \mathbb{N}$.

(3) For this purpose, we add 1 dimension, and we consider the following ideal:

$$J = \langle I, x_{N+1}P(x_1, \dots, x_N) - 1 \rangle$$

Since we have $X_J = \emptyset$, the Weak Nullstellensatz applies, and shows that J is trivial.

(4) In order to best interpret this finding, consider the following algebra:

$$F[x_1, \dots, x_N][P^{-1}] = F[x_1, \dots, x_{N+1}]/(x_{N+1}P - 1)$$

The triviality of J gives then a formula of the following type, with $f_i \in I$:

$$1 = f_0 + f_1x_{N+1} + \dots + f_mx_{N+1}^m$$

(5) Now by multiplying by P^m , we obtain from this the following formula:

$$P^m = P^mf_0 + P^{m-1}f_1 + \dots + f_m$$

Thus we have $P^m \in I$, as desired, which finishes the proof of the theorem. \square

And for more on all this, we refer to any standard algebraic geometry book.

10e. Exercises

This was a quite tricky, and very useful chapter. As exercises, we have:

EXERCISE 10.29. *Learn more tricks for degree 2 equations, which yes, do exist.*

EXERCISE 10.30. *Clarify what we said, in relation with symmetric polynomials.*

EXERCISE 10.31. *Compute more resultants, using both the available methods.*

EXERCISE 10.32. *Clarify the various normalizations, in the definition of $\Delta(P)$.*

EXERCISE 10.33. *Learn more about degree 3, including trigonometric aspects.*

EXERCISE 10.34. *Learn more about degree 4 too, including trigonometric aspects.*

EXERCISE 10.35. *Learn more about sinusoidal spirals, and their properties.*

EXERCISE 10.36. *Learn more about polynomial lemniscates, and stelloids too.*

As bonus exercise, find and start reading a nice, old-style algebraic geometry book.

CHAPTER 11

Gauss sums

11a. Gauss sums

Time for the complex numbers to strike again, with some non-trivial applications to the Legendre symbols. We recall from chapter 4 that these symbols, motivated by the equation $a = b^2(p)$ over the integers, with p prime, were constructed as follows:

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.

We have learned many things about these symbols, in chapter 4, with the summary of what we know from there about them being as follows:

THEOREM 11.2. *The Legendre symbols are given by the Euler formula:*

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

They are subject to upper multiplicativity and quadratic reciprocity,

$$\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right) \left(\frac{b}{p}\right) \quad , \quad \left(\frac{p}{q}\right) \left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2} \cdot \frac{q-1}{2}}$$

and together with the Euler formula at $a = -1$, and with the extra formula

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

this fully computes them, in practice.

PROOF. This is something that we know from chapter 4, the idea being as follows:

(1) The Euler formula follows from Fermat, $a^{p-1} = 1(p)$ for $a \neq 0(p)$, by factorizing, $(a^{\frac{p-1}{2}} - 1)(a^{\frac{p-1}{2}} + 1) = 0(p)$, and doing some counting work inside \mathbb{F}_p .

(2) The multiplicativity of the Legendre symbol in its upper variable is clear from the Euler formula, because $a^{\frac{p-1}{2}}(p)$ is obviously multiplicative in $a \in \mathbb{Z}$.

(3) The quadratic reciprocity formula is something non-trivial, but we eventually managed to prove that in chapter 4, following Gauss, Eisenstein and others.

(4) Regarding now the explicit computation of the Legendre symbol, the Euler formula remains something theoretical, except at $a = -1$, where this formula gives:

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

(5) In general, the Legendre symbol can be computed by recurrence, using multiplicativity and quadratic reciprocity, with a sample computation here being as follows:

$$\begin{aligned} \left(\frac{15}{37}\right) &= \left(\frac{3}{37}\right) \left(\frac{5}{37}\right) \\ &= (-1)^{\frac{3-1}{2} \cdot \frac{37-1}{2}} \left(\frac{37}{3}\right) (-1)^{\frac{5-1}{2} \cdot \frac{37-1}{2}} \left(\frac{37}{5}\right) \\ &= \left(\frac{37}{3}\right) \left(\frac{37}{5}\right) \\ &= \left(\frac{1}{3}\right) \left(\frac{2}{5}\right) \\ &= -1 \end{aligned}$$

(6) However, and as illustrated by the above example, when doing the recurrence we need as input an explicit formula at $a = 2$. And the formula here, which is the one in the statement, can be obtained by using some wild tricks, explained in chapter 4. \square

Summarizing, we have a quite complete theory for the Legendre symbols, ready to be used in practice, but when looking at the above proofs, some of them remain a bit mysterious, and we are led in this way to the following theoretical question:

QUESTION 11.3. *Can we have a better understanding of the $a = 2$ formula? And, why not, of the quadratic reciprocity formula too?*

These two questions are in fact related, as we will soon discover, and we will discuss them here. Let us start with the $a = 2$ formula. The result here is as follows:

THEOREM 11.4. *We have the following formula, as announced above,*

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

providing the missing piece for the full computation of the Legendre symbols.

PROOF. This is something from chapter 4, coming via a quite mysterious proof, but now that we know about complex numbers, we can slightly improve that, as follows:

(1) Let us set $w = e^{\pi i/4}$, so that $w^2 = i$, and $t = w + w^{-1}$. We have of course $t = \sqrt{2}$, but it is better to forget this, and do formal arithmetic instead. We first have:

$$t^2 = 2 + w^2 + w^{-2} = 2$$

By using the Euler formula for the Legendre symbol, we obtain:

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

Now by multiplying by t we obtain from this, again in a formal sense:

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

(2) On the other hand, by using the binomial formula we have, again formally:

$$t^p = (w + w^{-1})^p = w^p + w^{-p}(p)$$

Since $w^8 = 1$, this will depend only on p modulo 8. More precisely, we have:

$$w^p + w^{-p} = (-1)^{\frac{p^2-1}{8}}t$$

Now by putting everything together, and dividing by t , we obtain:

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

But this gives the formula in the statement, as desired. \square

The above proof still remains a bit mysterious, but do not worry, we will learn in a moment more about all this. Getting back now to Question 11.3, the point is that with the same idea, we can prove as well the quadratic reciprocity theorem, as follows:

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 $p, q \geq 3$, as announced above.

PROOF. This is something from chapter 4, coming there via a quite ad-hoc proof. However, thinking a bit, our $t = w + w^{-1}$ trick above can be adapted, as follows:

(1) To start with, we need an analogue of that $t = w + w^{-1}$ variable. For this purpose, let us set $w = e^{2\pi i/q}$, now that we have a prime $q \geq 3$ involved, and then:

$$t = \sum_{k=0}^{q-1} w^{k^2}$$

Observe that at $q = 2$, excluded by the statement, we have $w = -1$, and so $t = 1 + (-1) = 0$, instead of the $t = w + w^{-1}$ with $w = e^{\pi i/4}$ used before. However, believe me, this is due to some bizarre reasons, and the above t is the good variable, at $q \geq 3$.

(2) The above variable t is called Gauss sum, can be defined for any $q \in \mathbb{N}$, not necessarily prime, and can be explicitly computed, the formula being as follows:

$$t = \begin{cases} \sqrt{q} & \text{if } q \equiv 1(4) \\ 0 & \text{if } q \equiv 2(4) \\ \sqrt{q}i & \text{if } q \equiv 3(4) \\ \sqrt{q}(1+i) & \text{if } q \equiv 0(4) \end{cases}$$

In particular, assuming that q is odd, as is our $q \geq 3$ prime, we have:

$$t^2 = \begin{cases} q & \text{if } q \equiv 1(4) \\ -q & \text{if } q \equiv 3(4) \end{cases}$$

(3) In what follows we will only need this latter formula, for $q \geq 3$ prime, so let us prove this now, and with the comment that the proof of the first formula in (2) is something quite complicated, and better avoid that. We have, by definition of our variable t :

$$\begin{aligned} |t|^2 &= \sum_{kl} w^{k^2-l^2} \\ &= \sum_{kl} w^{(k+l)(k-l)} \\ &= \sum_{lr} w^{r(2l+r)} \\ &= \sum_r w^{r^2} \sum_l (w^{2r})^l \\ &= q \end{aligned}$$

(4) On the other hand, it is easy to see that t^2 is real, so $t^2 = \pm q$. With a bit more work it is possible to compute the sign too, $t^2 = (-1)^{\frac{q-1}{2}} q$, but we will not need this here, because the sign will come for free at the end of the proof, via a symmetry argument. So, as a conclusion, we have a formula as follows, for a certain $e_q \in \{0, 1\}$:

$$t^2 = (-1)^{e_q} q$$

(5) With this done, let us turn to the proof of our theorem, by using the variable t a bit as before, in the proof of Theorem 11.4. By using the Euler formula, we have:

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

By multiplying now by t we obtain from this, in a formal sense:

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

(6) In order to compute now t^p by other means, observe first that, if we denote by $\mathbb{Z}_q - \{0\} = S \sqcup N$ the partition into squares and non-squares, we have:

$$\begin{aligned} t &= \sum_{k=0}^{q-1} w^{k^2} \\ &= 1 + 2 \sum_{s \in S} w^s \\ &= \sum_{s \in S} w^s - \sum_{s \in N} w^s \\ &= \sum_{r=0}^{q-1} \left(\frac{r}{q}\right) w^r \end{aligned}$$

(7) By using now the multinomial formula, with the observation that all the non-trivial multinomial coefficients are multiples of p , we obtain, in a formal sense:

$$\begin{aligned} t^p &= \left(\sum_r \left(\frac{r}{q}\right) w^r \right)^p \\ &= \sum_r \left(\frac{r}{q}\right) w^{rp} \ (p) \\ &= \sum_s \left(\frac{p^{-1}s}{q}\right) w^s \ (p) \\ &= \left(\frac{p^{-1}}{q}\right) \sum_s \left(\frac{s}{q}\right) w^s \ (p) \\ &= \left(\frac{p}{q}\right) t \ (p) \end{aligned}$$

(8) Time now to put everything together. By combining (5,7) we obtain:

$$\left(\frac{t^2}{p}\right) t = \left(\frac{p}{q}\right) t \ (p)$$

We can divide by t , and then drop the modulo p symbol, because our new equality, without t , is between two ± 1 quantities, and we obtain:

$$\left(\frac{t^2}{p}\right) = \left(\frac{p}{q}\right)$$

Now by taking into account the formula found in (4), this reads:

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

By using the Euler formula for the symbol on the left, we obtain from this:

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

Now by symmetry we must have $e_q = \frac{q-1}{2}$, and this finishes the proof. \square

11b. Further summing

We have seen in the above that the quadratic reciprocity theorem can be established via Gauss sums t , and this is certainly excellent news. However, we have mentioned in step (2) of our proof above a very nice, powerful and final formula for the Gauss sum t itself, and this even in the general case, where $q \in \mathbb{N}$ is not necessarily prime.

Time now to discuss all this. So, we want to solve the following question:

QUESTION 11.6. *What is the value of the Gauss quadratic sum*

$$t = \sum_{k=0}^{q-1} w^{k^2}$$

where $w = e^{2\pi i/q}$, with $q \in \mathbb{N}$?

Let us begin with some experiments, at small values of q . We have here:

PROPOSITION 11.7. *The first few Gauss sums are as follows:*

- (1) At $q = 1$ we have $t = 1$.
- (2) At $q = 2$ we have $t = 0$.
- (3) At $q = 3$ we have $t = \sqrt{3}i$.
- (4) At $q = 4$ we have $t = 2(1 + i)$.
- (5) At $q = 5$ we have $t = \sqrt{5}$.
- (6) At $q = 6$ we have $t = 0$.
- (7) At $q = 7$ we have $t = \sqrt{7}i$.
- (8) At $q = 8$ we have $t = 2\sqrt{2}(1 + i)$.

PROOF. The computations are as follows, with $w = e^{2\pi i/q}$:

- (1) At $q = 1$ we have $w = 1$, and $t = 1$.
- (2) At $q = 2$ we have $w = -1$, and $t = 1 + (-1) = 0$

(3) At $q = 3$ we have $w = e^{2\pi i/3}$, and the computation goes as follows:

$$\begin{aligned} t &= 1 + w + w^4 \\ &= 1 + 2w \\ &= 1 + 2 \left(-\frac{1}{2} + \frac{\sqrt{3}}{2} i \right) \\ &= \sqrt{3} i \end{aligned}$$

(4) At $q = 4$ we have $w = i$, and the computation goes as follows:

$$\begin{aligned} t &= 1 + i + i^4 + i^9 \\ &= 1 + i + 1 + i \\ &= 2 + 2i \\ &= 2(1 + i) \end{aligned}$$

(5) At $q = 5$ we have $w = e^{2\pi i/5}$, and the computation goes as follows:

$$\begin{aligned} t &= 1 + w + w^4 + w^9 + w^{16} \\ &= 1 + w + w^4 + w^4 + w \\ &= 1 + 2(w + w^4) \\ &= 1 + 4 \cos \left(\frac{2\pi}{5} \right) \\ &= \sqrt{5} \end{aligned}$$

Here we have used some crazy trigonometry at the end, which can be avoided, or rather proved, when thinking well, at where this trigonometry comes from, as follows:

$$\begin{aligned} t^2 &= (1 + 2w + 2w^4)^2 \\ &= 1 + 4w^2 + 4w^3 + 4w + 4w^4 + 8 \\ &= 5 + 4(1 + w + w^2 + w^3 + w^4) \\ &= 5 \end{aligned}$$

Observe that there is actually still some work to be done here, when extracting the square root of $t^2 = 5$. But the picture shows that the root is positive, $t = \sqrt{5}$.

(6) At $q = 6$ it is most convenient to use $w = e^{2\pi i/3}$ as variable, as it is customary, and with this convention our root of unity is $e^{2\pi i/6} = -w^2$, and we have:

$$\begin{aligned} t &= 1 - w^2 + w^8 - w^{18} + w^{32} - w^{50} \\ &= 1 - w^2 + w^2 - 1 + w^2 - w^2 \\ &= 0 \end{aligned}$$

(7) At $q = 7$ we have $w = e^{2\pi i/7}$, and the computation goes as follows:

$$\begin{aligned}
 t &= 1 + w + w^4 + w^9 + w^{16} + w^{25} + w^{36} \\
 &= 1 + w + w^4 + w^2 + w^2 + w^4 + w \\
 &= 1 + 2(w + w^2 + w^4) \\
 &= \sqrt{7}i
 \end{aligned}$$

Here again we have used some crazy trigonometry, the justification being as follows, and with the correct root of $t^2 = -7$, among $t = \pm\sqrt{7}i$, being $t = \sqrt{7}i$, as shown by the picture, with the components w, w^2, w^4 of our sum t tending to lie North-West:

$$\begin{aligned}
 t^2 &= (1 + 2w + 2w^2 + 2w^4)^2 \\
 &= 1 + 4w^2 + 4w^4 + 4w \\
 &\quad + 4w + 4w^2 + 4w^4 \\
 &\quad + 8w^3 + 8w^5 + 8w^6 \\
 &= 1 + 8(w + w^2 + w^3 + w^4 + w^5 + w^6) \\
 &= -7 + 8(1 + w + w^2 + w^3 + w^4 + w^5 + w^6) \\
 &= -7
 \end{aligned}$$

(8) At $q = 8$ we have $w = e^{\pi i/4}$, and the computation goes as follows:

$$\begin{aligned}
 t &= 1 + w + w^4 + w^9 + w^{16} + w^{25} + w^{36} + w^{49} \\
 &= 1 + w - 1 + w + 1 + w - 1 + w \\
 &= 4w \\
 &= 2\sqrt{2}(1 + i)
 \end{aligned}$$

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

All the above is quite interesting, and we can formulate our conclusion as follows:

CONCLUSION 11.8. *The first few quadratic Gauss sums are given by*

q		1	2	3	4		5	6	7	8	
t		1	0	$\sqrt{3}i$	$2(1+i)$		$\sqrt{5}$	0	$\sqrt{7}i$	$2\sqrt{2}(1+i)$	

with everything coming from easy algebra, except for the signs.

Moving ahead now with the general case, there is some obvious periodicity in the above table, of order 4, and with everything working fine, I mean with the dependence on q being clear in all cases modulo 4, we are led to the following statement:

THEOREM 11.9. *We have the following formula for the Gauss sums,*

$$t = \begin{cases} \sqrt{q} & \text{if } q = 1(4) \\ 0 & \text{if } q = 2(4) \\ \sqrt{q}i & \text{if } q = 3(4) \\ \sqrt{q}(1+i) & \text{if } q = 0(4) \end{cases}$$

valid for any $q \in \mathbb{N}$, not necessarily prime.

PROOF. This is straightforward, except for the signs, the idea being as follows:

(1) To start with, let us compute $|t|^2$. This is something that we did in the proof of Theorem 11.5, for $q \geq 3$ prime, and the computation there can be recycled, as follows:

$$\begin{aligned} |t|^2 &= \sum_{kl} w^{k^2-l^2} = \sum_{kl} w^{(k+l)(k-l)} \\ &= \sum_{lr} w^{r(2l+r)} = \sum_r w^{r^2} \sum_l (w^{2r})^l \\ &= \sum_r w^{r^2} \times \delta_{2|2r} q = q \sum_{q|2r} w^{r^2} \end{aligned}$$

(2) We have some cases here. For q odd we get q , and for q even, we have:

$$\begin{aligned} |t|^2 &= q(1 + (w^{(q/2)^2}) \\ &= q(1 + (w^{q/2})^{q/2}) \\ &= q(1 + (-1)^{q/2}) \end{aligned}$$

(3) We are therefore led to the following formula, for our variable $|t|^2$:

$$|t|^2 = \begin{cases} q & \text{if } q = 1(4) \\ 0 & \text{if } q = 2(4) \\ q & \text{if } q = 3(4) \\ 2q & \text{if } q = 0(4) \end{cases}$$

(4) Now by extracting the square root, we have the following formula, for $|t|$:

$$|t| = \begin{cases} \sqrt{q} & \text{if } q = 1(4) \\ 0 & \text{if } q = 2(4) \\ \sqrt{q} & \text{if } q = 3(4) \\ \sqrt{2q} & \text{if } q = 0(4) \end{cases}$$

(5) The question is now, shall we go ahead and compute t , or be less greedy, and compute t^2 first. And let us be modest, of course, and go with t^2 first. But here, it is pretty much clear, from the computations in the proof of Proposition 11.7, that we can

get away with some simple algebra, I mean with algebra a hair more complicated than that in (1,2) above. For this purpose, the best is to go with the following alternative definition of the Gauss sums, that we already met in the proof of Theorem 11.5:

$$t = \sum_{r=0}^{q-1} \left(\frac{r}{q} \right) w^r$$

(6) Now by taking the square of this quantity, and then working out what exactly happens at $q = 1, 2, 3, 0(4)$, exactly as in the proof of Proposition 11.7, and we will leave this as an instructive exercise, we are led to the following formula:

$$t^2 = \begin{cases} q & \text{if } q = 1(4) \\ 0 & \text{if } q = 2(4) \\ -q & \text{if } q = 3(4) \\ 2qi & \text{if } q = 0(4) \end{cases}$$

(7) In what regards now t itself, by taking the square root, we must have:

$$t = \begin{cases} \pm\sqrt{q} & \text{if } q = 1(4) \\ 0 & \text{if } q = 2(4) \\ \pm\sqrt{q}i & \text{if } q = 3(4) \\ \pm\sqrt{q}(1+i) & \text{if } q = 0(4) \end{cases}$$

(8) So, almost done, but thinking a bit, in fact we just got started. Indeed, remember from Proposition 11.7 that the computation of the signs is tricky business, done on pictures, more specifically at $q = 5$ by arguing that the components of t tend to pull it East, and at $q = 7$, by arguing that these components tend to pull it North-West.

(9) So, what kind of question is this, geography or something? Well, in answer, such things are called mathematical analysis. Normally, what we need are some estimates, with ε and everything, as to decide what is the approximate direction of the pull of the components of t , as to compute that missing sign. And, more on this in a moment. \square

11c. The Gauss sign

Time to get into the real thing, which is the computation of the missing sign, for the Gauss sums t . As a first observation, since $t = 0$ for $q = 2(4)$, the problem is solved in this case, at least we know one thing. Thus, we are left with the cases $q = 0, 1, 3(4)$.

In order to deal with these questions, let us introduce the following notion:

DEFINITION 11.10. *The Fourier matrix of order $q \in \mathbb{N}$ is*

$$F_q = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & w & w^2 & \dots & w^{q-1} \\ 1 & w^2 & w^4 & \dots & w^{2(q-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w^{q-1} & w^{2(q-1)} & \dots & w^{(q-1)^2} \end{pmatrix}$$

with $w = e^{2\pi i/q}$. That is, $F_q = (w^{kl})$, with indices $k, l = 0, 1, \dots, q-1$.

These matrices are well-known in discrete Fourier analysis, and more on this later, at the end of the present chapter. In the meantime, the interest in relation with our Gauss sum questions is quite obvious, coming from the following formula:

$$\text{Tr}(F_q) = \sum_k w^{k^2} = t$$

We are therefore led to the following question, whose answer would solve everything in relation with the Gauss sums, and in particular, our sign problem:

QUESTION 11.11. *What are the eigenvalues of F_q , and their multiplicities?*

So, this is the problem that we would like to solve. For certain reasons, that will become clear later, it is convenient to study, more generally, the following matrices:

DEFINITION 11.12. *Associated to any $p, q \in \mathbb{N}$ with $(p, q) = 1$ is the matrix*

$$F_{pq} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & z & z^2 & \dots & z^{q-1} \\ 1 & z^2 & z^4 & \dots & z^{2(q-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & z^{q-1} & z^{2(q-1)} & \dots & z^{(q-1)^2} \end{pmatrix}$$

with $z = e^{2\pi ip/q}$. That is, $F_{pq} = (w^{pkl})$, with indices $k, l = 0, 1, \dots, q-1$.

Observe that we have $F_{1q} = F_q$. In what follows we will be mainly interested in this case, $p = 1$, but as mentioned above, it is technically convenient to deal with F_{pq} . As a first result now regarding F_{pq} , providing the key to its algebra, we have:

PROPOSITION 11.13. *The square of the matrix F_{pq} is given by*

$$F_{pq}^2 = \begin{pmatrix} q & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & q \\ 0 & 0 & \dots & q & 0 \\ & & \vdots & & \\ 0 & q & \dots & 0 & 0 \end{pmatrix}$$

and its fourth power is $F_{pq}^4 = q^2 1_q$.

PROOF. The first formula can be proved as follows, using $(p, q) = 1$, which shows that $z = e^{2\pi ip/q}$ is a primitive q -root of unity, so that the usual barycenter trick applies:

$$\begin{aligned} (F_{pq}^2)_{kl} &= \sum_m (F_{pq})_{km} (F_{pq})_{ml} \\ &= \sum_m z^{km} z^{ml} \\ &= \sum_m z^{m(k+l)} \\ &= q\delta_{k+l,0} \end{aligned}$$

As for the second formula, $F_{pq}^4 = q^2 1_q$, this comes from this, with the matrix representing F_{pq}^2 in the statement being q times the symmetry $e_k \leftrightarrow e_{-k}$. \square

Getting now to diagonalization, we first have the following result:

PROPOSITION 11.14. *The following happen, in relation with the matrix F_{pq} :*

- (1) F_{pq}^4 has unique eigenvalue q^2 .
- (2) F_{pq}^2 has eigenvalues $q, -q$.
- (3) F_{pq} has eigenvalues $\pm\sqrt{q}, \pm i\sqrt{q}$.

PROOF. This is indeed something self-explanatory, with (1) coming from the equality $F_{pq}^4 = q^2 1_q$, then with (2) coming from (1), and then with (3) coming from (2). \square

We can in fact say more, with a word on multiplicities. First we have:

PROPOSITION 11.15. *The multiplicity of q as eigenvalue of F_{pq}^2 is*

$$m_q = \begin{cases} \frac{q+1}{2} & (q \text{ odd}) \\ \frac{q}{2} + 1 & (q \text{ even}) \end{cases}$$

and the multiplicity of $-q$ as eigenvalue of F_{pq}^2 is

$$m_{-q} = \begin{cases} \frac{q-1}{2} & (q \text{ odd}) \\ \frac{q}{2} - 1 & (q \text{ even}) \end{cases}$$

with these numbers summing up to q .

PROOF. This follows from the formula for F_{pq}^2 from Proposition 11.13, namely:

$$F_{pq}^2 = \begin{pmatrix} q & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & q \\ 0 & 0 & \dots & q & 0 \\ & & \vdots & & \\ 0 & q & \dots & 0 & 0 \end{pmatrix}$$

Indeed, we are led in this way to the multiplicities from the statement. \square

In what regards now the multiplicities for F_{pq} , we have here:

THEOREM 11.16. *The multiplicities a, b, c, d of the eigenvalues of F_{pq} ,*

$$\underbrace{\sqrt{q}}_a, \underbrace{-\sqrt{q}}_b, \underbrace{i\sqrt{q}}_c, \underbrace{-i\sqrt{q}}_d$$

are subject to the following equations, when q is odd,

$$a + b = \frac{q+1}{2} \quad , \quad c + d = \frac{q-1}{2}$$

and are subject to the following equations, when q is even:

$$a + b = \frac{q}{2} + 1 \quad , \quad c + d = \frac{q}{2} - 1$$

Also, $Tr(F_{pq}) = \sqrt{q}[(a-b) + (c-d)i]$, with at $p=1$ this being the Gauss sum t .

PROOF. According to Proposition 11.14, the multiplicities of the eigenvalues for the matrices F_{pq} and F_{pq}^2 are related by the following formulae:

$$a + b = m_q \quad , \quad c + d = m_{-q}$$

Now by using the formulae of m_q, m_{-q} from Proposition 11.15, this gives the formulae in the statement. Regarding now the last assertion, we have indeed:

$$\begin{aligned} Tr(F_{pq}) &= \sqrt{q} \times a - \sqrt{q} \times b + i\sqrt{q} \times c - i\sqrt{q} \times d \\ &= \sqrt{q}[(a-b) + (c-d)i] \end{aligned}$$

As for the very last statement, which is there for everything to be complete, this is something that we already know, coming from $Tr(F_q) = \sum_k w^{k^2} = t$. \square

Let us further investigate the trace of F_{pq} . We have here the following result:

PROPOSITION 11.17. *The trace of the matrix F_{pq} satisfies*

$$|Tr(F_{pq})| = \begin{cases} \sqrt{q} & \text{if } q \equiv 1(4) \\ 0 & \text{if } q \equiv 2(4) \\ \sqrt{q} & \text{if } q \equiv 3(4) \\ \sqrt{2q} & \text{if } q \equiv 0(4) \end{cases}$$

for any $(p, q) = 1$.

PROOF. This is something that we basically know, from our previous investigations of the Gauss sums, but here is that computation again, in terms of our present formalism

and notations, which are a bit more general, than what we were doing before:

$$\begin{aligned}
 |Tr(F_{pq})|^2 &= \sum_{kl} z^{k^2-l^2} = \sum_{kl} z^{(k+l)(k-l)} \\
 &= \sum_{lr} z^{r(2l+r)} = \sum_r z^{r^2} \sum_l (z^{2r})^l \\
 &= \sum_r z^{r^2} \times \delta_{2|2r} q = q \sum_{q|2r} z^{r^2}
 \end{aligned}$$

We have some cases here. For q odd we get q , and for q even, we have:

$$\begin{aligned}
 |Tr(F_{pq})|^2 &= q(1 + (z^{(q/2)^2}) \\
 &= q(1 + (z^{q/2})^{q/2}) \\
 &= q(1 + (-1)^{q/2})
 \end{aligned}$$

We are therefore led to the following formula, for our variable $|Tr(F_{pq})|^2$:

$$|Tr(F_{pq})|^2 = \begin{cases} q & \text{if } q = 1(4) \\ 0 & \text{if } q = 2(4) \\ q & \text{if } q = 3(4) \\ 2q & \text{if } q = 0(4) \end{cases}$$

Now by extracting the square root, we are led to the formula in the statement. \square

We can in fact say more about the trace, by using Theorem 11.16, as follows:

THEOREM 11.18. *The trace of the matrix F_{pq} satisfies*

$$Tr(F_{pq}) = \begin{cases} \pm\sqrt{q} & \text{if } q = 1(4) \\ 0 & \text{if } q = 2(4) \\ \pm\sqrt{q}i & \text{if } q = 3(4) \\ \pm\sqrt{q}(1+i) & \text{if } q = 0(4) \end{cases}$$

for any $(p, q) = 1$.

PROOF. Again, this is something that we basically know, from our previous investigations of the Gauss sums, and in the present setting, which is a bit more general, this can be obtained either by recycling the computation there, or by merging what we found in Theorem 11.16 and Proposition 11.17. We will leave this as an exercise. \square

As a last elementary result, regarding the traces, we have:

PROPOSITION 11.19. *We have the following multiplicativity formula,*

$$Tr(F_{p,qr}) = Tr(F_{pq,r})Tr(F_{pr,q})$$

valid for any numbers p, q, r satisfying $(p, qr) = 1$ and $(q, r) = 1$.

PROOF. We have indeed the following computation, with $w = e^{2\pi i/qr}$:

$$\begin{aligned} \text{Tr}(F_{pq,r})\text{Tr}(F_{pr,q}) &= \sum_k (w^q)^{pqk^2} \sum_l (w^r)^{prl^2} \\ &= \sum_{kl} w^{p(q^2k^2+r^2l^2)} \end{aligned}$$

By some basic quadratic residue arithmetic, this equals the following quantity:

$$\text{Tr}(F_{p,qr}) = \sum_k w^{pk^2}$$

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

This was for the elementary part of the proof, connecting our problems regarding the Gauss sums t to the algebra of the Fourier matrices F_{pq} . The continuation of the story, bringing some truly new results, involves the computation of $\det(F_{1q})$, as follows:

THEOREM 11.20. *We have the following formula,*

$$\det(F_{1q}) = \begin{cases} i^{\binom{q}{2}} q^{q/2} & \text{for } q \text{ odd} \\ i^{\binom{q}{2}+1} q^{q/2} & \text{for } q \text{ even} \end{cases}$$

for the determinant of the Fourier matrix, $F_{1q} = (w^{kl})$ with $w = e^{2\pi i/q}$.

PROOF. To start with, the Fourier matrix in the statement looks as follows:

$$F_{1q} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & w & w^2 & \dots & w^{q-1} \\ 1 & w^2 & w^4 & \dots & w^{2(q-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w^{q-1} & w^{2(q-1)} & \dots & w^{(q-1)^2} \end{pmatrix}$$

But this is a Vandermonde matrix, so we have the following formula:

$$\det(F_{1q}) = \prod_{k>l} (w^k - w^l)$$

On the other hand, recall from Proposition 11.13 that we have:

$$F_{1q}^2 = \begin{pmatrix} q & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & q \\ 0 & 0 & \dots & q & 0 \\ & \vdots & & & \\ 0 & q & \dots & 0 & 0 \end{pmatrix}$$

Thus we have $\det(F_{1q}^2) = \pm q^q$, and a more careful study gives:

$$\det(F_{1q}^2) = \begin{cases} (-1)^{\binom{q}{2}} q^q & \text{for } q \text{ odd} \\ (-1)^{\binom{q}{2}+1} q^q & \text{for } q \text{ even} \end{cases}$$

Now by extracting the square root, we obtain the following formula:

$$\det(F_{1q}) = \begin{cases} \pm i^{\binom{q}{2}} q^{q/2} & \text{for } q \text{ odd} \\ \pm i^{\binom{q}{2}+1} q^{q/2} & \text{for } q \text{ even} \end{cases}$$

In order to compute the sign, we can use the previous Vandermonde approach, which gives, in terms of the root of unity $x = e^{\pi i/q}$, satisfying $x^2 = w$:

$$\begin{aligned} \det(F_{1q}) &= \prod_{k>l} (x^{2k} - x^{2l}) \\ &= \prod_{k>l} x^{k+l} (x^{k-l} - x^{l-k}) \\ &= \prod_{k>l} x^{k+l} \prod_{k>l} 2i \sin\left(\frac{(k-l)\pi}{q}\right) \\ &= (2i)^{\binom{q}{2}} \prod_{k>l} x^{k+l} \prod_{k>l} \sin\left(\frac{(k-l)\pi}{q}\right) \\ &= (2i)^{\binom{q}{2}} \cdot i^{(q-1)^2} \cdot \prod_{k>l} \sin\left(\frac{(k-l)\pi}{q}\right) \end{aligned}$$

Here we have used the following computation, with $x = e^{\pi i/q}$, which is something elementary, and that we will leave here as an instructive exercise:

$$\prod_{k>l} x^{k+l} = i^{(q-1)^2}$$

Now observe that $2^{\binom{q}{2}}$ is positive, then $i^{(q-1)^2}$ is 1 when q is odd, and 0 when q is even, and the product of sines on the right, which are all positive, is positive too. We are therefore led to the conclusion that the missing sign that we were computing is +:

$$\det(F_{1q}) = \begin{cases} i^{\binom{q}{2}} q^{q/2} & \text{for } q \text{ odd} \\ i^{\binom{q}{2}+1} q^{q/2} & \text{for } q \text{ even} \end{cases}$$

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

We will need as well a key result regarding the traces, as follows:

THEOREM 11.21. *For q odd and $(p, q) = 1$ we have*

$$\text{Tr}(F_{pq}) = \left(\frac{p}{q}\right) \text{Tr}(F_{1q})$$

with on the right a Jacobi symbol.

PROOF. This is something quite tricky, coming in two steps, as follows:

(1) Let us first discuss the case where $q \geq 3$ is prime. Here we have, with $w = e^{2\pi i/q}$, by definition of the Legendre symbol, and since the sum of roots of unity vanishes:

$$\begin{aligned} \text{Tr}(F_{pq}) &= \sum_k w^{k^2 p} \\ &= \sum_l \left[\left(\frac{l}{q}\right) + 1 \right] w^{lp} \\ &= \sum_l \left(\frac{l}{q}\right) w^{lp} + \sum_l w^{lp} \\ &= \sum_l \left(\frac{l}{q}\right) w^{lp} \end{aligned}$$

Now by using the multiplicativity of the Legendre symbol in its upper variable, we can deduce from this the formula in the statement, by using the following trick:

$$\begin{aligned} \text{Tr}(F_{pq}) &= \left(\frac{p}{q}\right)^2 \sum_l \left(\frac{l}{q}\right) w^{lp} \\ &= \left(\frac{p}{q}\right) \sum_l \left(\frac{lp}{q}\right) w^{lp} \\ &= \left(\frac{p}{q}\right) \sum_k \left(\frac{k}{q}\right) w^k \\ &= \left(\frac{p}{q}\right) \text{Tr}(F_{1q}) \end{aligned}$$

(2) In the general case now, observe first that we have a formula as follows, with the product being over all prime factors of q , each repeated with its multiplicity:

$$\text{Tr}(F_{1q}) = \pm \prod_{r|q} \text{Tr}(F_{1r})$$

By applying to this the automorphism $z \rightarrow z^p$ we have, more generally:

$$\text{Tr}(F_{pq}) = \pm \prod_{r|q} \text{Tr}(F_{pr})$$

But with this formula in hand we obtain, by using (1) for all factors $r|q$:

$$\begin{aligned}
 \text{Tr}(F_{pq}) &= \pm \prod_{r|q} \left(\frac{p}{r}\right) \text{Tr}(F_{1r}) \\
 &= \prod_{r|q} \left(\frac{p}{r}\right) \cdot \left(\pm \prod_{r|q} \text{Tr}(F_{1r}) \right) \\
 &= \left(\frac{p}{q}\right) \text{Tr}(F_{1q})
 \end{aligned}$$

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

And with this, good news, done with the preliminaries, and we can now prove:

THEOREM 11.22. *We have the following formula for the Gauss sums,*

$$t = \begin{cases} \sqrt{q} & \text{if } q = 1(4) \\ 0 & \text{if } q = 2(4) \\ \sqrt{q}i & \text{if } q = 3(4) \\ \sqrt{q}(1+i) & \text{if } q = 0(4) \end{cases}$$

valid for any $q \in \mathbb{N}$, not necessarily prime.

PROOF. This follows indeed by using our various results above, as follows:

(1) To start with, let us recall that we have the following formula:

$$t = \begin{cases} \pm\sqrt{q} & \text{if } q = 1(4) \\ 0 & \text{if } q = 2(4) \\ \pm\sqrt{q}i & \text{if } q = 3(4) \\ \pm\sqrt{q}(1+i) & \text{if } q = 0(4) \end{cases}$$

(2) When q is odd, it follows from Theorem 11.20 that the sign is $+$. Thus, we know one thing, and with this being the main one, q odd being the case that matters.

(3) When q is even, and more specifically in the case left, $q = 0(4)$, the study is a bit more complicated, involving the general Fourier matrices F_{pq} , and Proposition 11.19 and a recurrence argument, based on Theorem 11.21. Our claim here is that we have:

$$\text{Tr}(F_{pq}) = \left(\frac{q}{p}\right) (1 + i^p) \sqrt{q}$$

In order to prove this claim, let us write $q = 2^a b$ with $a \geq 2$ and b odd. By using Proposition 11.19, then an elementary computation with the roots of unity of order 2^a , coupled with Theorem 11.21, then the multiplicativity property of the Jacobi symbols,

then the trace formula in the odd case, coming from (2), and finally the multiplicativity of Jacobi symbols again, coupled with quadratic reciprocity, we have:

$$\begin{aligned}
Tr(F_{pq}) &= Tr(F_{p,2^a b}) \\
&= Tr(F_{bp,2^a})Tr(F_{2^a p,b}) \\
&= \left(\frac{2^a}{bp}\right) (1 + i^{bp})\sqrt{2^a} \times \left(\frac{2^a p}{b}\right) Tr(F_{1b}) \\
&= \sqrt{2^a} \left(\frac{2^a}{bp}\right) \left(\frac{2^a p}{b}\right) (1 + i^{bp})Tr(F_{1b}) \\
&= \sqrt{2^a} \left(\frac{2^a}{b}\right) \left(\frac{2^a}{p}\right) \left(\frac{2^a}{b}\right) \left(\frac{p}{b}\right) (1 + i^{bp})Tr(F_{1b}) \\
&= \sqrt{2^a} \left(\frac{2^a}{p}\right) \left(\frac{p}{b}\right) (1 + i^{bp})Tr(F_{1b}) \\
&= \sqrt{2^a} \left(\frac{2^a}{p}\right) \left(\frac{p}{b}\right) (1 + i^{bp})i^{(b-1)^2/4}\sqrt{b} \\
&= \left(\frac{2^a}{p}\right) \left(\frac{p}{b}\right) (1 + i^{bp})i^{(b-1)^2/4}\sqrt{q} \\
&= \pm \left(\frac{2^a b}{p}\right) \cdot [\pm (1 + i^p)]\sqrt{q} \\
&= \left(\frac{q}{p}\right) (1 + i^p)\sqrt{q}
\end{aligned}$$

Thus, claim proved, and with $p = 1$ we are led to the formula in the statement. \square

11d. Some applications

Let us go back now to the Hadamard matrix questions, discussed chapter 4. As a further twist to the plot, bringing some sort of solution to that questions, we have:

THEOREM 11.23. *When enlarging the attention to the complex Hadamard matrices, $H \in M_N(\mathbb{T})$ having the rows pairwise orthogonal, the Fourier matrix,*

$$F_N = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & w & w^2 & \dots & w^{N-1} \\ 1 & w^2 & w^4 & \dots & w^{2(N-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w^{N-1} & w^{2(N-1)} & \dots & w^{(N-1)^2} \end{pmatrix}$$

with $w = e^{2\pi i/N}$, provides an example of such a matrix, at any $N \in \mathbb{N}$. Thus, the Hadamard Conjecture problematics dissapears, in the complex setting.

PROOF. By using the standard fact that the averages of complex numbers correspond to barycenters, we conclude that the scalar products between the rows of F_N are:

$$\langle R_a, R_b \rangle = \sum_j w^{(a-b)j} = N\delta_{ab}$$

Thus F_N is indeed a complex Hadamard matrix, as claimed. \square

In view of this, let us study more in detail the complex Hadamard matrices. Many examples can be constructed, quite often by using the combinatorics of roots of unity, and as a basic example here, we have the tensor products of Fourier matrices:

THEOREM 11.24. *Given a finite abelian group G , with dual group $\widehat{G} = \{\chi : G \rightarrow \mathbb{T}\}$, consider the Fourier coupling $\mathcal{F}_G : G \times \widehat{G} \rightarrow \mathbb{T}$, given by $(i, \chi) \rightarrow \chi(i)$.*

- (1) *Via the standard isomorphism $G \simeq \widehat{\widehat{G}}$, this Fourier coupling can be regarded as a square matrix, $F_G \in M_G(\mathbb{T})$, which is a complex Hadamard matrix.*
- (2) *In the case of the cyclic group $G = \mathbb{Z}_N$ we obtain in this way, via the standard identification $\mathbb{Z}_N = \{1, \dots, N\}$, the Fourier matrix F_N .*
- (3) *In general, when using a decomposition $G = \mathbb{Z}_{N_1} \times \dots \times \mathbb{Z}_{N_k}$, the corresponding Fourier matrix is given by $F_G = F_{N_1} \otimes \dots \otimes F_{N_k}$.*

PROOF. This follows indeed from some basic facts from group theory:

(1) With the identification $G \simeq \widehat{\widehat{G}}$ made our matrix is given by $(F_G)_{i\chi} = \chi(i)$, and the scalar products between the rows are then, as desired:

$$\langle R_i, R_j \rangle = \sum_{\chi} \chi(i - j) = |G| \cdot \delta_{ij}$$

(2) This follows from the well-known and elementary fact that, via the identifications $\mathbb{Z}_N = \widehat{\widehat{\mathbb{Z}_N}} = \{1, \dots, N\}$, the Fourier coupling here is $(i, j) \rightarrow w^{ij}$, with $w = e^{2\pi i/N}$.

(3) By decomposing G as a product of cyclic groups, as in the statement, and using (2) for the cyclic components, we obtain the formula in the statement. \square

As an application of this, in relation with the matrices from chapter 4, we have:

PROPOSITION 11.25. *The Walsh matrix, W_N with $N = 2^n$, which is given by*

$$W_N = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{\otimes n}$$

is the Fourier matrix of the finite abelian group $K_N = \mathbb{Z}_2^n$.

PROOF. We have indeed the following computation, for any $N = 2^n$:

$$W_N = F_2^{\otimes n} = F_{K_2}^{\otimes n} = F_{K_2^n} = F_{K_N}$$

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

Summarizing, the complex Hadamard matrices are quite interesting objects. Getting now to classification matters, we first have the following result:

THEOREM 11.26. *The complex Hadamard matrices are as follows, up to equivalence coming by permuting rows or columns, or multiplying them by numbers in \mathbb{T} :*

- (1) *At $N = 2$ we only have the Fourier matrix F_2 .*
- (2) *At $N = 3$ we only have the Fourier matrix F_3 .*
- (3) *At $N = 4$ we only have F_4 , and its deformations.*
- (4) *At $N = 5$ we only have the Fourier matrix F_5 .*

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

- (1) This is something trivial, coming from definitions.
- (2) This is again elementary, that we will leave as an instructive exercise.
- (3) Here the matrices are up to equivalence as follows, with $q \in \mathbb{T}$:

$$F_4^q = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & q & -1 & -q \\ 1 & -q & -1 & q \end{pmatrix}$$

As for the proof, this is again elementary, and we will leave it as an exercise.

(4) This is something which is remarkably difficult, due to Haagerup. Consider indeed an Hadamard matrix $H \in M_5(\mathbb{T})$, chosen to be dephased, as follows:

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & a & x & * & * \\ 1 & y & b & * & * \\ 1 & * & * & * & * \\ 1 & * & * & * & * \end{pmatrix}$$

The point is that some study, which remains something quite formal and mysterious, shows that the numbers a, b, x, y must satisfy in fact the following equation:

$$(x - y)(x - ab)(y - ab) = 0$$

But with this in hand, which is something powerful, the continuation is routine, and we are led to roots of unity, and to the conclusion in the statement. See [9]. \square

At $N = 6$ things explode, with a summary of what happens here being as follows:

THEOREM 11.27. *The complex Hadamard matrices at $N = 6$ are as follows:*

- (1) *As basic examples, we have F_6 and its two-parameter deformations.*
- (2) *We have as well another deformable matrix H_6 , and an isolated one T_6 .*
- (3) *The matrices F_6, H_6, T_6 and their deformations are the only regular ones.*
- (4) *However, we have many exotic examples too, which are not classified yet.*

PROOF. The idea here is that the matrices which are regular, in the sense that the vanishing scalar products between their rows come from roots of unity, can be classified, as said in (1,2,3), and that passed this, things are quite complicated, as said in (4). As an illustration for (4), here is a beast found by Björck and Fröberg:

$$BF_6 = \begin{pmatrix} 1 & ia & -a & -i & -\bar{a} & i\bar{a} \\ i\bar{a} & 1 & ia & -a & -i & -\bar{a} \\ -\bar{a} & i\bar{a} & 1 & ia & -a & -i \\ -i & -\bar{a} & i\bar{a} & 1 & ia & -a \\ -a & -i & -\bar{a} & i\bar{a} & 1 & ia \\ ia & -a & -i & -\bar{a} & i\bar{a} & 1 \end{pmatrix}$$

Indeed, assuming that $a \in \mathbb{T}$ is one of the roots of $a^2 + (\sqrt{3} - 1)a + 1 = 0$, this matrix is Hadamard. For more on all this, $N = 6$ and higher, you can check my book [9]. \square

In view of the above examples and counterexamples, and thinking a bit at number theory, where primes are the kings, we are led to the question of finding the complex Hadamard matrices which are “isolated”, in a geometric sense. And here, skipping some details, we have an interesting construction due to McNulty and Weigert, that we would like to explain now. This construction is based on the following simple fact:

THEOREM 11.28. *Assuming that $K \in M_N(\mathbb{C})$ is Hadamard, so is the matrix*

$$H_{ia,jb} = \frac{1}{\sqrt{Q}} K_{ij} (L_i^* R_j)_{ab}$$

provided that $\{L_1, \dots, L_N\} \subset \sqrt{Q}U_Q$ and $\{R_1, \dots, R_N\} \subset \sqrt{Q}U_Q$ are such that

$$\frac{1}{\sqrt{Q}} L_i^* R_j \in \sqrt{Q}U_Q$$

with $i, j = 1, \dots, N$, are complex Hadamard.

PROOF. The check of the unitarity of the matrix in the statement can be done as follows, by using our various assumptions on the various matrices involved:

$$\begin{aligned} \langle H_{ia}, H_{kc} \rangle &= \frac{1}{Q} \sum_{jb} K_{ij} (L_i^* R_j)_{ab} \bar{K}_{kj} \overline{(L_k^* R_j)_{cb}} \\ &= \sum_j K_{ij} \bar{K}_{kj} (L_i^* L_k)_{ac} \\ &= N \delta_{ik} (L_i^* L_k)_{ac} \\ &= NQ \delta_{ik} \delta_{ac} \end{aligned}$$

The entries of our matrix being in addition on the unit circle, we are done. \square

The above construction is of course something quite abstract, but as a very concrete input for it, we can use the following well-known Fourier analysis construction:

PROPOSITION 11.29. *For $q \geq 3$ prime, the family of matrices*

$$\{F_q, DF_q, \dots, D^{q-1}F_q\}$$

where F_q is the Fourier matrix, and where D is the diagonal matrix given by

$$D = \text{diag} \left(1, 1, w, w^3, w^6, w^{10}, \dots, w^{\frac{q^2-1}{8}}, \dots, w^{10}, w^6, w^3, w \right)$$

*with $w = e^{2\pi i/q}$, are such that $\frac{1}{\sqrt{q}}E_i^*E_j$ is complex Hadamard, for any $i \neq j$.*

PROOF. With by definition $0, 1, \dots, q-1$ as indices for our matrices, as usual in a Fourier analysis context, the formula of the above matrix D is:

$$D_c = w^{0+1+\dots+(c-1)} = w^{\frac{c(c-1)}{2}}$$

Since we have $\frac{1}{\sqrt{q}}E_i^*E_j \in \sqrt{q}U_q$, we just need to check that these matrices have entries belonging to \mathbb{T} , for any $i \neq j$. With $k = j - i$, these entries are given by:

$$\frac{1}{\sqrt{q}}(E_i^*E_j)_{ab} = \frac{1}{\sqrt{q}}(F_q^*D^kF_q)_{ab} = \frac{1}{\sqrt{q}} \sum_c w^{c(b-a)} D_c^k$$

But an elementary computation with roots of unity, that we will leave here as an exercise, shows that these entries are on the unit circle, as desired. \square

Next, we have the following result, making use of Gauss sums:

PROPOSITION 11.30. *The matrices $G_k = \frac{1}{\sqrt{q}}F_q^*D^kF_q$, with $D = \text{diag} \left(w^{\frac{c(c-1)}{2}} \right)$, and with $k \neq 0$ are circulant, their first row vectors V^k being given by*

$$V_i^k = \delta_q \left(\frac{k/2}{q} \right) w^{\frac{q^2-1}{8} \cdot k} \cdot w^{-\frac{i}{k} \left(\frac{i}{k} - 1 \right)}$$

where $\delta_q = 1$ if $q = 1(4)$ and $\delta_q = i$ if $q = 3(4)$, and with all inverses being taken in \mathbb{Z}_q .

PROOF. The above matrices G_k are indeed circulant, their first vectors being:

$$V_i^k = \frac{1}{\sqrt{q}} \sum_c w^{\frac{c(c-1)}{2} \cdot k + ic}$$

But this is a Gauss sum, and by computing the square, we obtain:

$$(V_i^k)^2 = \delta_q^2 \cdot w^{\frac{q^2-1}{4} \cdot k} \cdot w^{-\frac{i}{k} \left(\frac{i}{k} - 1 \right)}$$

By extracting now the square root, we obtain a formula as follows:

$$V_i^k = \pm \delta_q \cdot w^{\frac{q^2-1}{8} \cdot k} \cdot w^{-\frac{i}{k} \left(\frac{i}{k} - 1 \right)}$$

And with Theorem 11.22 computing the sign, this leads to the above formula. \square

Let us combine now the above results. We obtain the following statement:

THEOREM 11.31. *Let $q \geq 3$ be prime, consider subsets $S, T \subset \{0, 1, \dots, q-1\}$ satisfying the conditions $|S| = |T|$ and $S \cap T = \emptyset$, and write:*

$$S = \{s_1, \dots, s_N\} \quad , \quad T = \{t_1, \dots, t_N\}$$

Then, with the matrix V being as above, the following matrix,

$$H_{ia,jb} = K_{ij} V_{b-a}^{t_j - s_i}$$

is complex Hadamard, provided that the matrix $K \in M_N(\mathbb{C})$ is complex Hadamard.

PROOF. This follows indeed by using the general construction in Theorem 11.28, with input coming from Proposition 11.29 and Proposition 11.30. \square

The above construction covers many interesting examples of Hadamard matrices, known to be isolated, such as the Tao matrix, which is as follows, with $w = e^{2\pi i/3}$:

$$T_6 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & w & w & w^2 & w^2 \\ 1 & w & 1 & w^2 & w^2 & w \\ 1 & w & w^2 & 1 & w & w^2 \\ 1 & w^2 & w^2 & w & 1 & w \\ 1 & w^2 & w & w^2 & w & 1 \end{pmatrix}$$

For more on such matrices, there are many texts available, including my book [9].

11e. Exercises

This was a tricky number theory chapter, and as exercises on this, we have:

EXERCISE 11.32. *Compute more Gauss sums, as many as you can.*

EXERCISE 11.33. *When stuck with higher sums, develop more trigonometry.*

EXERCISE 11.34. *Learn about various generalizations of the Gauss sums.*

EXERCISE 11.35. *Fill in the details, in the computation of the Gauss sum sign.*

EXERCISE 11.36. *Learn as well some further proofs of quadratic reciprocity.*

EXERCISE 11.37. *Learn more about discrete Fourier analysis, in general.*

EXERCISE 11.38. *Learn about the classification of complex Hadamard matrices.*

EXERCISE 11.39. *Read about the other applications of the Gauss sums.*

As bonus exercise, review your trigonometry knowledge, and read more, if needed.

CHAPTER 12

Transcendence

12a. More about e

We have recently learned many interesting things about numbers, both of algebraic and analytic nature, and the question is, what is next? Not very clear, and it is probably a good idea to try to intercept the cat, see what he thinks about this.

So, let us do this. Contrary to pure mathematics, which can require a lot of thinking and everything, in applied mathematics, where there are no rules, and all methods are allowed, sometimes simple things like opening a fish can will do. And, with the fish can opened, here is what we quickly get, as answer to our mathematical questions:

CAT 12.1. *Advanced number theory needs a good knowledge of e and π . By the way, you should learn more calculus, and some complex analysis too.*

Okay, thanks very much, so let us do this, more study of e and π . Not very clear which of these beasts is simpler, and with cat being already gone, for a digestive sleep I guess, we will have to improvise. So, going alphabetically, let us start with e .

As a first objective, I don't know about you, but personally I would like to have as well a combinatorial interpretation of e . In order to discuss this, we will need:

PROPOSITION 12.2. *We have the following formula,*

$$\left| \left(\bigcup_i A_i \right)^c \right| = |A| - \sum_i |A_i| + \sum_{i < j} |A_i \cap A_j| - \sum_{i < j < k} |A_i \cap A_j \cap A_k| + \dots$$

called inclusion-exclusion principle.

PROOF. This is indeed quite clear, by thinking a bit, as follows:

- (1) In order to count $(\cup_i A_i)^c$, we certainly have to start with $|A|$.
- (2) Then, we obviously have to remove each $|A_i|$, and so remove $\sum_i |A_i|$.
- (3) But then, we have to put back each $|A_i \cap A_j|$, and so put back $\sum_{i < j} |A_i \cap A_j|$.
- \vdots
- (4) And so on, which leads to the formula in the statement. □

With this discussed, here is now the application of the inclusion-exclusion principle that we were having in mind, making appear e , in a nice combinatorial way:

THEOREM 12.3. *The probability for a random permutation $\sigma \in S_N$ to be a derangement, that is, to have no fixed points, is given by the following formula:*

$$P = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \dots + (-1)^N \frac{1}{N!}$$

Thus we have the following asymptotic formula, in the $N \rightarrow \infty$ limit,

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

with $e = 2.71828\dots$ being the usual constant from analysis.

PROOF. This is something very classical, which is best viewed by using the inclusion-exclusion principle. Consider indeed the following sets of permutations:

$$S_N^i = \left\{ \sigma \in S_N \mid \sigma(i) = i \right\}$$

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

$$X_N = \left(\bigcup_i S_N^i \right)^c$$

By inclusion-exclusion, the probability that we are interested in is given by:

$$\begin{aligned} & P(\sigma \in X_N) \\ &= \frac{1}{N!} \left(|S_N| - \sum_i |S_N^i| + \sum_{i < j} |S_N^i \cap S_N^j| - \dots + (-1)^N \sum_{i_1 < \dots < i_N} |S_N^{i_1} \cap \dots \cap S_N^{i_N}| \right) \\ &= \frac{1}{N!} \sum_{k=0}^N (-1)^k \sum_{i_1 < \dots < i_k} (N-k)! \\ &= \frac{1}{N!} \sum_{k=0}^N (-1)^k \binom{N}{k} (N-k)! \\ &= \frac{1}{N!} \sum_{k=0}^N (-1)^k \frac{N!}{k!(N-k)!} (N-k)! \\ &= \sum_{k=0}^N \frac{(-1)^k}{k!} \end{aligned}$$

Since at the end we have the expansion of $1/e$, we obtain the result. \square

Let us discuss now, more generally, what happens when counting permutations having exactly k fixed points. The result here, extending Theorem 12.3, is as follows:

THEOREM 12.4. *The probability for a random permutation $\sigma \in S_N$ to have exactly k fixed points is given by the following formula:*

$$P = \frac{1}{k!} \left(1 - \frac{1}{1!} + \frac{1}{2!} - \dots + (-1)^{N-1} \frac{1}{(N-1)!} + (-1)^N \frac{1}{N!} \right)$$

Thus we have the formula $P \simeq 1/(ek!)$, in the $N \rightarrow \infty$ limit.

PROOF. We already know, from Theorem 12.3, that this formula holds at $k = 0$. In the general case now, we have to count the permutations $\sigma \in S_N$ having exactly k points. Since having such a permutation amounts in choosing k points among $1, \dots, N$, and then permuting the $N - k$ points left, without fixed points allowed, we have:

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

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

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

By using now the computation at $k = 0$, from Theorem 12.3, it follows that with $N \rightarrow \infty$ we have the following estimate, χ being the number of fixed points:

$$P(\chi = k) \simeq \frac{1}{k!} \cdot P(\chi = 0) \simeq \frac{1}{k!} \cdot \frac{1}{e}$$

Thus, we are led to the conclusion in the statement. □

In order to further build on this, let us formulate the following key definition, coming as a continuation of our basic discussion on probability, from chapter 2:

DEFINITION 12.5. *The Poisson law of parameter 1 is the following measure,*

$$p_1 = \frac{1}{e} \sum_{k \in \mathbb{N}} \frac{\delta_k}{k!}$$

and the Poisson law of parameter $t > 0$ is the following measure,

$$p_t = e^{-t} \sum_{k \in \mathbb{N}} \frac{t^k}{k!} \delta_k$$

with the letter “p” standing for Poisson.

We are using here some simplified notations for these laws. Observe that our laws have indeed mass 1, as they should, due to the following formula:

$$e^t = \sum_{k \in \mathbb{N}} \frac{t^k}{k!}$$

We will see in a moment why these measures are important, the reasons for this coming from the Poisson Limit Theorem (PLT), which is closely related to our previous investigations from chapter 2, regarding the Bernoulli and binomial laws. For the moment, let us first develop some general theory, for these Poisson laws. We first have:

PROPOSITION 12.6. *The mean and variance of the Poisson law p_t are*

$$E = V = t$$

for any $t > 0$. In particular, at $t = 1$ we have $E = V = 1$.

PROOF. In what regards the mean of the Poisson law p_t , this is given by:

$$\begin{aligned} E &= e^{-t} \sum_{k \geq 1} \frac{t^k k}{k!} \\ &= e^{-t} \sum_{k \geq 1} \frac{t^k}{(k-1)!} \\ &= e^{-t} \times t e^t \\ &= t \end{aligned}$$

Let us compute now the second moment, $M_2 = E(f^2)$. We can do this as follows:

$$\begin{aligned} M_2 &= e^{-t} \sum_{k \geq 1} \frac{t^k k^2}{k!} \\ &= e^{-t} \left(\sum_{k \geq 1} \frac{t^k (k-1)}{(k-1)!} + \sum_{k \geq 1} \frac{t^k}{(k-1)!} \right) \\ &= e^{-t} (t^2 e^t + t e^t) \\ &= t^2 + t \end{aligned}$$

Thus, the variance is $V = (t^2 + t) - t^2 = t$, as claimed. □

At a more advanced level now, we first have the following result:

THEOREM 12.7. *We have the following formula, for any $s, t > 0$,*

$$p_s * p_t = p_{s+t}$$

so the Poisson laws form a convolution semigroup.

PROOF. By using $\delta_k * \delta_l = \delta_{k+l}$ and the binomial formula, we obtain:

$$\begin{aligned}
 p_s * p_t &= e^{-s} \sum_k \frac{s^k}{k!} \delta_k * e^{-t} \sum_l \frac{t^l}{l!} \delta_l \\
 &= e^{-s-t} \sum_n \delta_n \sum_{k+l=n} \frac{s^k t^l}{k! l!} \\
 &= e^{-s-t} \sum_n \frac{(s+t)^n}{n!} \delta_n \\
 &= p_{s+t}
 \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Next in line, we have the following result, which is fundamental as well:

THEOREM 12.8. *The Poisson laws appear as formal exponentials*

$$p_t = \sum_k \frac{t^k (\delta_1 - \delta_0)^{*k}}{k!}$$

with respect to the convolution of measures $*$.

PROOF. By using the binomial formula, the measure on the right is:

$$\begin{aligned}
 \mu &= \sum_k \frac{t^k}{k!} \sum_{r+s=k} (-1)^s \frac{k!}{r! s!} \delta_r \\
 &= \sum_k t^k \sum_{r+s=k} (-1)^s \frac{\delta_r}{r! s!} \\
 &= \sum_r \frac{t^r \delta_r}{r!} \sum_s \frac{(-1)^s t^s}{s!} \\
 &= \frac{1}{e^t} \sum_r \frac{t^r \delta_r}{r!} \\
 &= p_t
 \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

In order to say more, we will need a key technical result, as follows:

THEOREM 12.9. *Assuming that $f, g : X \rightarrow \mathbb{R}$ are independent, we have*

$$F_{f+g}(x) = F_f(x) F_g(x)$$

where $F(x) = E(e^{ifx})$ if the Fourier transform. The converse of this holds, too.

PROOF. Assuming indeed that $f, g : X \rightarrow \mathbb{R}$ are independent, we have:

$$\begin{aligned}
 F_{f+g}(x) &= \int_{\mathbb{R}} e^{ixz} d\mu_{f+g}(z) \\
 &= \int_{\mathbb{R}} e^{ixz} d(\mu_f * \mu_g)(z) \\
 &= \int_{\mathbb{R} \times \mathbb{R}} e^{ix(y+t)} d\mu_f(y) d\mu_g(t) \\
 &= \int_{\mathbb{R}} e^{ixy} d\mu_f(y) \int_{\mathbb{R}} e^{ixt} d\mu_g(t) \\
 &= F_f(x) F_g(x)
 \end{aligned}$$

As for the converse, this follows also from the above computation. \square

For the Poisson laws, the formula of the Fourier transform is as follows:

THEOREM 12.10. *The Fourier transform of p_t is given by*

$$F_{p_t}(x) = \exp((e^{ix} - 1)t)$$

for any $t > 0$.

PROOF. We have indeed the following computation, coming from definitions:

$$\begin{aligned}
 F_{p_t}(x) &= e^{-t} \sum_k \frac{t^k}{k!} F_{\delta_k}(x) \\
 &= e^{-t} \sum_k \frac{t^k}{k!} e^{ikx} \\
 &= e^{-t} \sum_k \frac{(e^{ix}t)^k}{k!} \\
 &= \exp(-t) \exp(e^{ix}t) \\
 &= \exp((e^{ix} - 1)t)
 \end{aligned}$$

Thus, we have reached to the formula in the statement. \square

Observe that we obtain in this way another proof for the convolution semigroup property of the Poisson laws, from Theorem 12.7, by using Theorem 12.9.

Good news, we can now establish the Poisson Limit Theorem (PLT), as follows:

THEOREM 12.11. *We have the following convergence, in moments,*

$$\left(\left(1 - \frac{t}{n} \right) \delta_0 + \frac{t}{n} \delta_1 \right)^{*n} \rightarrow p_t$$

for any $t > 0$.

PROOF. Let us denote by μ_n the measure under the convolution sign:

$$\mu_n = \left(1 - \frac{t}{n}\right) \delta_0 + \frac{t}{n} \delta_1$$

We have then the following computation, for the law in the statement:

$$\begin{aligned} F_{\delta_r}(x) = e^{irx} &\implies F_{\mu_n}(x) = \left(1 - \frac{t}{n}\right) + \frac{t}{n} e^{ix} \\ &\implies F_{\mu_n^{*n}}(x) = \left(\left(1 - \frac{t}{n}\right) + \frac{t}{n} e^{ix}\right)^n \\ &\implies F_{\mu_n^{*n}}(x) = \left(1 + \frac{(e^{ix} - 1)t}{n}\right)^n \\ &\implies F(x) = \exp((e^{ix} - 1)t) \end{aligned}$$

Thus, we obtain the Fourier transform of p_t , as desired. \square

Getting back now to permutations, we have the following result:

THEOREM 12.12. *The number of fixed points of permutations,*

$$\chi : S_N \rightarrow \mathbb{N} \quad , \quad \chi(\sigma) = \# \left\{ i \in \{1, \dots, N\} \mid \sigma(i) = i \right\}$$

follows a Poisson law, $\chi \sim p_1$, in the $N \rightarrow \infty$ limit.

PROOF. This is indeed a straightforward reformulation of Theorem 12.4, which tells us that with $N \rightarrow \infty$ we have the following estimate:

$$P(\chi = k) \simeq \frac{1}{e k!}$$

To be more precise, according to our definition of the Poisson laws, this formula tells us that the asymptotic law of the variable χ is Poisson (1), as stated. \square

As an important comment now on the above result, we have:

COMMENT 12.13. *When regarding $S_N \subset O_N$, via permutation matrices, we have:*

$$\chi(\sigma) = \sum_{i=1}^N \sigma_{ii}$$

Thus, $\chi : S_N \rightarrow \mathbb{N}$ is the main character, in the representation theory sense.

To be more precise here, as a continuation of the general group theory discussion from chapter 3, at a more advanced level, it is possible to talk about group representations $u : G \rightarrow U_N$, and their characters $\chi_u(g) = \text{Tr}(u(g))$. And for the standard representation of S_N , given by $\sigma(e_i) = e_{\sigma(i)}$, the character is the number of fixed points, as stated.

Summarizing, quite interesting all this, making a link between advanced group theory, and advanced probability. Moreover, the story is not over here, because we have:

THEOREM 12.14. *The number of partial fixed points of permutations,*

$$\chi_t : S_N \rightarrow \mathbb{N} \quad , \quad \chi(\sigma) = \# \left\{ i \in \{1, \dots, [tN]\} \mid \sigma(i) = i \right\}$$

follows a Poisson law, $\chi_t \sim p_t$, in the $N \rightarrow \infty$ limit, for any $t \in (0, 1]$.

PROOF. We already know from Theorem 12.12 that the result holds at $t = 1$. In general, the proof is similar, the idea being as follows:

(1) Consider indeed the following sets, as in the proof of Theorem 12.3:

$$S_N^i = \left\{ \sigma \in S_N \mid \sigma(i) = i \right\}$$

As before in the proof of Theorem 12.3, we obtain by inclusion-exclusion that:

$$\begin{aligned} P(\chi_t = 0) &= \frac{1}{N!} \sum_{k=0}^{[tN]} (-1)^k \sum_{i_1 < \dots < i_k < [tN]} |S_N^{i_1} \cap \dots \cap S_N^{i_k}| \\ &= \frac{1}{N!} \sum_{k=0}^{[tN]} (-1)^k \sum_{i_1 < \dots < i_k < [tN]} (N - k)! \\ &= \frac{1}{N!} \sum_{k=0}^{[tN]} (-1)^k \binom{[tN]}{k} (N - k)! \\ &= \sum_{k=0}^{[tN]} \frac{(-1)^k}{k!} \cdot \frac{[tN]! (N - k)!}{N! ([tN] - k)!} \end{aligned}$$

With $N \rightarrow \infty$, we obtain from this the following estimate:

$$\begin{aligned} P(\chi_t = 0) &\simeq \sum_{k=0}^{[tN]} \frac{(-1)^k}{k!} \cdot t^k \\ &= \sum_{k=0}^{[tN]} \frac{(-t)^k}{k!} \\ &\simeq e^{-t} \end{aligned}$$

(2) More generally now, by counting the permutations $\sigma \in S_N$ having exactly k fixed points among $1, \dots, [tN]$, as in the proof of Theorem 12.4, our claim is that we get:

$$P(\chi_t = k) \simeq \frac{t^k}{k! e^t}$$

We already know from (1) that this formula holds at $k = 0$. In the general case now, we have to count the permutations $\sigma \in S_N$ having exactly k fixed points among $1, \dots, [tN]$. Since having such a permutation amounts in choosing k points among $1, \dots, [tN]$, and then permuting the $N - k$ points left, without fixed points among $1, \dots, [tN]$ allowed, we obtain the following formula, where $s \in (0, 1]$ is such that $[s(N - k)] = [tN] - k$:

$$\begin{aligned} \#\{\sigma \in S_N \mid \chi_t(\sigma) = k\} &= \binom{[tN]}{k} \#\{\sigma \in S_{N-k} \mid \chi_s(\sigma) = 0\} \\ &= \frac{[tN]!}{k!([tN] - k)!} \#\{\sigma \in S_{N-k} \mid \chi_s(\sigma) = 0\} \\ &= \frac{1}{k!} \times \frac{[tN]!(N - k)!}{([tN] - k)!} \times \frac{\#\{\sigma \in S_{N-k} \mid \chi_s(\sigma) = 0\}}{(N - k)!} \end{aligned}$$

Now by dividing everything by $N!$, we obtain from this the following formula:

$$\frac{\#\{\sigma \in S_N \mid \chi_t(\sigma) = k\}}{N!} = \frac{1}{k!} \times \frac{[tN]!(N - k)!}{N!([tN] - k)!} \times \frac{\#\{\sigma \in S_{N-k} \mid \chi_s(\sigma) = 0\}}{(N - k)!}$$

By using now the computation at $k = 0$, that we already have, from (1) above, it follows that with $N \rightarrow \infty$ we have the following estimate:

$$\begin{aligned} P(\chi_t = k) &\simeq \frac{1}{k!} \times \frac{[tN]!(N - k)!}{N!([tN] - k)!} \cdot P(\chi_s = 0) \\ &\simeq \frac{t^k}{k!} \cdot P(\chi_s = 0) \\ &\simeq \frac{t^k}{k!} \cdot \frac{1}{e^s} \end{aligned}$$

Now recall that the parameter $s \in (0, 1]$ was chosen in the above such that:

$$[s(N - k)] = [tN] - k$$

Thus in the $N \rightarrow \infty$ limit we have $s = t$, and so we obtain, as claimed:

$$P(\chi_t = k) \simeq \frac{t^k}{k!} \cdot \frac{1}{e^t}$$

It follows that we obtain in the limit a Poisson law of parameter t , as stated. \square

As before with Theorem 12.12, there is a comment to be made here, as follows:

COMMENT 12.15. *When regarding $S_N \subset O_N$, via permutation matrices, we have:*

$$\chi_t(\sigma) = \sum_{i=1}^{[tN]} \sigma_{ii}$$

Thus, the variables $\chi_t : S_N \rightarrow \mathbb{N}$ are the standard truncations of the main character.

To be more precise, we previously talked, in Comment 12.13 and afterwards, about the group representations $u : G \rightarrow U_N$, and their characters $\chi_u(g) = \text{Tr}(u(g))$. The point now is that, in relation with certain advanced questions in representation theory, and their probabilistic and random matrix ramifications, the truncated characters, as defined above, naturally come into play. And for the standard representation of S_N , given by $\sigma(e_i) = e_{\sigma(i)}$, the truncated characters count the partial fixed points, as stated.

Summarizing, good idea that we had, of getting into combinatorial aspects of the number e . All this makes indeed an interesting link between advanced group theory, and advanced probability, and for more on this, you can check my book [10].

12b. Transcendence of e

Now that we are done with advanced algebra and combinatorics, time for some tough calculus, hopefully of the same technical level, still regarding the number e .

We have at least three interesting questions to be discussed, namely approximation of e , irrationality of e , and transcendentality of e . In what regards the first two questions, which are quite elementary, and related to each other, the result is as follows:

THEOREM 12.16. *The number e from analysis, given by*

$$e = \sum_{k=0}^{\infty} \frac{1}{k!}$$

which numerically means $e = 2.7182818284\dots$, is irrational.

PROOF. Many things can be said here, as follows:

(1) To start with, there are several possible definitions for e , with the old style one, which is quite cool, and that we used in this book too, being as follows:

$$\left(1 + \frac{1}{n}\right)^n \rightarrow e$$

The definition in the statement is the modern one. Indeed, imagine that you are looking for a function \exp , satisfying $\exp' = \exp$, and $\exp(0) = 1$. With $\exp(x) = \sum_k c_k x^k$, you must have $c_0 = 1$, then $c_1 = 1$, $c_2 = 1/2$, $c_3 = 1/6$ and so on, meaning:

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

But now, it is an easy exercise to show that $\exp(x+y) = \exp(x)\exp(y)$, which gives $\exp(x) = e^x$, for a certain number $e > 0$. Which number e can only be $e = \exp(1)$.

(2) Getting now to numerics, the series of e converges very fast, when compared to the old style sequence in (1), so if you are in a hurry, this series is for you. We have:

$$\begin{aligned}
 e &= \sum_{k=0}^{N-1} \frac{1}{k!} + \frac{1}{N!} \left(1 + \frac{1}{N+1} + \frac{1}{(N+1)(N+2)} + \dots \right) \\
 &< \sum_{k=0}^{N-1} \frac{1}{k!} + \frac{1}{N!} \left(1 + \frac{1}{N+1} + \frac{1}{(N+1)^2} + \dots \right) \\
 &= \sum_{k=0}^{N-1} \frac{1}{k!} + \frac{1}{N!} \left(1 + \frac{1}{N} \right) \\
 &= \sum_{k=0}^N \frac{1}{k!} + \frac{1}{N \cdot N!}
 \end{aligned}$$

Thus, the error term in the approximation is really tiny, the estimate being:

$$\sum_{k=0}^N \frac{1}{k!} < e < \sum_{k=0}^N \frac{1}{k!} + \frac{1}{N \cdot N!}$$

(3) Now by using this, you can easily compute the decimals of e . Actually, you can't call yourself mathematician, or scientist, if you haven't done this by hand, just for the fun, but just in case, here is how the approximation goes, for small values of N :

$$N = 2 \implies 2.5 < e < 2.75$$

$$N = 3 \implies 2.666\dots < e < 2.722\dots$$

$$N = 4 \implies 2.70833\dots < e < 2.71875\dots$$

$$N = 5 \implies 2.71666\dots < e < 2.71833\dots$$

$$N = 6 \implies 2.71805\dots < e < 2.71828\dots$$

$$N = 7 \implies 2.71825\dots < e < 2.71828\dots$$

Thus, first 4 decimals computed, $e = 2.7182\dots$, and I would leave the continuation to you. With the remark that, when carefully looking at the above, the estimate on the right works much better than the one on the left, so before getting into more serious numerics, try to find a better lower estimate for e , that can help you in your work.

(4) Getting now to irrationality, a look at $e = 2.7182818284\dots$ might suggest that the 81, 82, 84... values might eventually, after some internal fight, decide for a winner, and so that e might be rational. However, this is wrong, and e is in fact irrational.

(5) So, let us prove now this, that e is irrational. Following Fourier, we will do this by contradiction. So, assume $e = m/n$, and let us look at the following number:

$$x = n! \left(e - \sum_{k=0}^n \frac{1}{k!} \right)$$

As a first observation, x is an integer, as shown by the following computation:

$$\begin{aligned} x &= n! \left(\frac{m}{n} - \sum_{k=0}^n \frac{1}{k!} \right) \\ &= m(n-1)! - \sum_{k=0}^n n(n-1) \dots (n-k+1) \\ &\in \mathbb{Z} \end{aligned}$$

(6) On the other hand $x > 0$, and we have as well the following estimate:

$$\begin{aligned} x &= n! \sum_{k=n+1}^{\infty} \frac{1}{k!} \\ &= \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \dots \\ &< \frac{1}{n+1} + \frac{1}{(n+1)^2} + \dots \\ &= \frac{1}{n} \end{aligned}$$

Thus $x \in (0, 1)$, which contradicts our previous finding $x \in \mathbb{Z}$, as desired. \square

As a continuation, we have the following result, which is substantially harder:

THEOREM 12.17. *The number e is transcendental.*

PROOF. Assume by contradiction that e is algebraic, with this meaning that it is a root of a polynomial with integer coefficients, $c_i \in \mathbb{Z}$, as follows:

$$c_0 + c_1 e + \dots + c_n e^n = 0$$

(1) Following Hermite, consider the following polynomials, and we will see later why:

$$f_k(x) = x^k [(x-1) \dots (x-n)]^{k+1}$$

Consider also the following quantities, that we will study more in detail later:

$$A_k = \int_0^{\infty} f_k(x) e^{-x} dx$$

By multiplying our equation for e by this quantity A_k , we obtain:

$$c_0 \int_0^\infty f_k(x) e^{-x} dx + c_1 \int_0^\infty f_k(x) e^{1-x} dx + \dots + c_n \int_0^\infty f_k(x) e^{n-x} dx = 0$$

(2) Here comes the trick. Consider the following two quantities:

$$P = c_0 \int_0^\infty f_k(x) e^{-x} dx + c_1 \int_1^\infty f_k(x) e^{1-x} dx + \dots + c_n \int_n^\infty f_k(x) e^{n-x} dx$$

$$Q = c_1 \int_0^1 f_k(x) e^{-x} dx + \dots + c_n \int_0^n f_k(x) e^{n-x} dx$$

In terms of these quantities, the formula that we found in (1) reads:

$$P + Q = 0$$

(3) Now let us look at P . Our claim is that this is an integer, $P \in \mathbb{Z}$, and that there are arbitrarily large numbers $k \gg 0$ for which the following holds:

$$\frac{P}{k!} \in \mathbb{Z} - \{0\}$$

Indeed, according to our formula above defining P , we have:

$$\begin{aligned} P &= \sum_{r=0}^n c_r \int_r^\infty f_k(x) e^{r-x} dx \\ &= \sum_{r=0}^n c_r \int_0^\infty f_k(x+r) e^{-x} dx \\ &= \int_0^\infty \left(\sum_{r=0}^n c_r f_k(x+r) \right) e^{-x} dx \end{aligned}$$

On the other hand, integrating such functions is easy, according to:

$$\begin{aligned} \int_0^\infty x^s e^{-x} dx &= \int_0^\infty \left(\frac{x^{s+1}}{s+1} \right)' e^{-x} dx \\ &= \int_0^\infty \frac{x^{s+1}}{s+1} e^{-x} dx \\ &= \frac{1}{s+1} \int_0^\infty x^{s+1} e^{-x} dx \end{aligned}$$

Thus, we are led by recurrence on $s \in \mathbb{N}$ to the following formula:

$$\int_0^\infty x^s e^{-x} dx = s!$$

For a linear combination now, we are led to the following formula:

$$g(x) = \sum_s a_s x^s \implies \int_0^\infty g(x) e^{-x} dx = \sum_s a_s s!$$

But this shows that we have indeed $P \in \mathbb{Z}$, and also, via a bit of study based on the exact formula of f_k , from the beginning of (1), that we have in fact:

$$\frac{P}{k!} \in \mathbb{Z}$$

Finally, we still have to prove that we have $P \neq 0$, for arbitrarily large numbers $k \gg 0$. But the point here is that for $k+1 > n, |c_0|$, chosen prime, a detailed study of our integral shows that we have $(k+1) \nmid P$, and so $P \neq 0$ indeed, as desired.

(4) With this done, let us look now at Q . Our claim is that for $k \gg 0$ we have:

$$\left| \frac{Q}{k!} \right| < 1$$

Indeed, by using the exact formula of f_k , from the beginning of (1), we have:

$$\begin{aligned} f_k(x) e^{-x} &= x^k [(x-1) \dots (x-n)]^{k+1} e^{-x} \\ &= [x(x-1) \dots (x-n)]^k \times (x-1) \dots (x-n) e^{-x} \end{aligned}$$

We conclude that for $x \in [0, n]$ we have an estimate as follows, with $G, H > 0$ being certain constants, appearing as maxima of the two components appearing above:

$$|f_k(x) e^{-x}| < G^k H$$

Now by integrating, we obtain from this the following estimate for Q itself:

$$\begin{aligned} |Q| &= \left| c_1 \int_0^1 f_k(x) e^{-x} dx + \dots + c_n e^n \int_0^n f_k(x) e^{-x} dx \right| \\ &\leq |c_1| \int_0^1 |f_k(x) e^{-x}| dx + \dots + |c_n| e^n \int_0^n |f_k(x) e^{-x}| dx \\ &\leq |c_1| \cdot G^k H + \dots + |c_n| e^n \cdot n G^k H \\ &= (|c_1| e + \dots + |c_n| e^n) \frac{n(n+1)}{2} G^k H \end{aligned}$$

But in this estimate the only term depending on k is the power G^k , and since since $k!$ grows much faster than this power G^k , this proves our claim:

$$\left| \frac{Q}{k!} \right| \approx \frac{G^k}{k!} \rightarrow 0$$

(5) And with this, done, because what we found in (3,4) contradicts the formula $P + Q = 0$ from (2). Thus e is indeed transcendental, as claimed. \square

12c. Approximating pi

As a continuation of the above, we would like to prove now, a bit as for e before, that π is irrational, and even transcendental. Before that, however, again as before for e , let us start with some approximation work. We first have the following result:

PROPOSITION 12.18. *We have the following approximations of π ,*

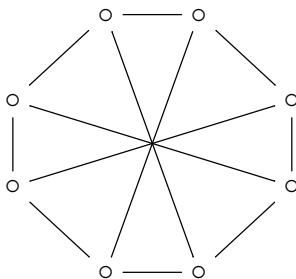
$$2.828 < \pi < 4$$

$$3 < \pi < 3.464$$

$$3.061 < \pi < 3.314$$

obtained respectively by using squares, hexagons and octagons.

PROOF. Leaving the squares and hexagons for you, respectively chess and badminton, let me get directly to MMA questions, in relation with the octagon:



The octagons inscribed and circumscribed to the unit circle have edges as follows, coming by looking at the respective “pizza slices”, which are both $45-67.5-67.5$ triangles, the inner ones having radial edge 1, and the outer ones having radial altitude 1:

$$e = 2 \sin(22.5^\circ) \quad , \quad E = 2 \tan(22.5^\circ)$$

Thus by looking at half-perimeters, we obtain the following estimate for π :

$$4e < \pi < 4E \quad \implies \quad 8 \sin(22.5^\circ) < \pi < 8 \tan(22.5^\circ)$$

We have as well an estimate for π coming by looking at areas, which is as follows, using the fact that the area of a $45-67.5-67.5$ triangle is $\cot(22.5^\circ) \times \text{edge}^2/4$:

$$\begin{aligned} 2 \cot(22.5^\circ) e^2 &< \pi < 2 \cot(22.5^\circ) E^2 \\ \implies 8 \sin(22.5^\circ) \cos(22.5^\circ) &< \pi < 8 \tan(22.5^\circ) \\ \implies 4 \sin(45^\circ) &< \pi < 8 \tan(22.5^\circ) \end{aligned}$$

In order to reach now to some concrete numeric estimates, we first have:

$$\cos(22.5^\circ) = \sqrt{\frac{1 + \cos 45^\circ}{2}} = \frac{\sqrt{2 + \sqrt{2}}}{2}$$

For the sine we can use Pythagoras, $\sin^2 + \cos^2 = 1$, and we obtain:

$$\sin(22.5^\circ) = \sqrt{1 - \frac{2 + \sqrt{2}}{4}} = \frac{\sqrt{2 - \sqrt{2}}}{2}$$

Finally, by taking the quotient we obtain a formula for the tangent, as follows:

$$\tan(22.5^\circ) = \sqrt{\frac{2 - \sqrt{2}}{2 + \sqrt{2}}} = \sqrt{2} - 1$$

Time now for the final numerics. The half-perimeter estimate for π reads:

$$\begin{aligned} 8 \sin(22.5^\circ) < \pi < 8 \tan(22.5^\circ) &\implies 4\sqrt{2 - \sqrt{2}} < \pi < 8(\sqrt{2} - 1) \\ &\implies 3.061 < \pi < 3.314 \end{aligned}$$

As for the area estimate for π , also found above, this reads:

$$\begin{aligned} 4 \sin(45^\circ) < \pi < 8 \tan(22.5^\circ) &\implies 4/\sqrt{2} < \pi < 8(\sqrt{2} - 1) \\ &\implies 2.828 < \pi < 3.314 \end{aligned}$$

Thus, we are led to the conclusions in the statement, and with the remark that the half-perimeter estimate beats the area estimate, which is good to know. \square

The above is quite encouraging, and getting now to the general case, that of the arbitrary polygons, we have the following result, further building on what we have:

THEOREM 12.19. *We have the following estimates for π , obtained by computing the half perimeter of an inscribed and circumscribed regular N -gon:*

$$N \sin\left(\frac{180^\circ}{N}\right) < \pi < N \tan\left(\frac{180^\circ}{N}\right)$$

By using the inscribed and circumscribed regular hexadecagon, these estimates give

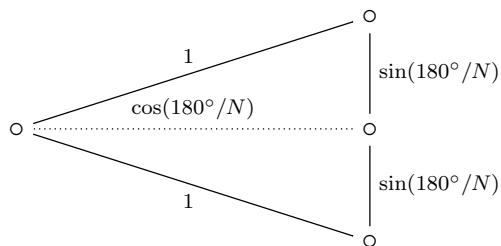
$$3.121 < \pi < 3.183$$

and with a bit more work, with $N = 2^n$ with $n \gg 0$, we obtain $\pi = 3.14159 \dots$

PROOF. We use the same method as before, the idea being as follows:

(1) In order to compute the edge of the inscribed N -gon, let us look at the corresponding “pizza slices”. These are isosceles triangles with angle $360^\circ/N$ and edge 1, so when

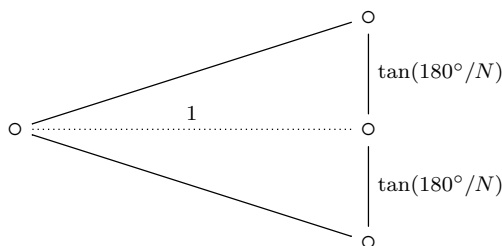
drawing an altitude and computing the lengths, the picture is as follows:



Thus, the edge and half perimeter of the inscribed N -gon are as follows:

$$e = 2 \sin\left(\frac{180^\circ}{N}\right) \quad , \quad p = N \sin\left(\frac{180^\circ}{N}\right)$$

(2) Getting now to the circumscribed N -gon, the “pizza slices” here are again isosceles triangles with angle $360^\circ/N$, but this time with altitude 1. Thus when drawing these altitudes and computing the relevant lengths, the picture is as follows:



Thus, the edge and half perimeter of the circumscribed N -gon are as follows:

$$E = 2 \tan\left(\frac{180^\circ}{N}\right) \quad , \quad P = N \tan\left(\frac{180^\circ}{N}\right)$$

(3) Time now to derive some conclusions, from our study. With our half perimeter method, we obtain the estimate from the statement, namely:

$$N \sin\left(\frac{180^\circ}{N}\right) < \pi < N \tan\left(\frac{180^\circ}{N}\right)$$

Note in passing that, as an alternative idea, we can try to use instead areas. But this gives an estimate which is weaker than the one above, as follows:

$$\frac{N}{2} \sin\left(\frac{360^\circ}{N}\right) < \pi < N \tan\left(\frac{180^\circ}{N}\right)$$

(4) In view of numerics, let us begin with some general trigonometry, regarding the sines and tangents of the halves of angles. For the sines, we have:

$$\begin{aligned}
 \cos(2t) = 1 - 2\sin^2 t &\implies (1 - 2\sin^2 t)^2 = 1 - \sin^2(2t) \\
 &\implies 1 - 2\sin^2 t = \sqrt{1 - \sin^2(2t)} \\
 &\implies 2\sin^2 t = 1 - \sqrt{1 - \sin^2(2t)} \\
 &\implies \sin t = \sqrt{\frac{1 - \sqrt{1 - \sin^2(2t)}}{2}}
 \end{aligned}$$

As for the tangents, we have here the following computation, and with exercise here for you to figure out why we choose the sign plus, for the square root:

$$\begin{aligned}
 \tan(2t) = \frac{2\tan t}{1 - \tan^2 t} &\implies \tan(2t)\tan^2 t + 2\tan t - \tan(2t) = 0 \\
 &\implies \tan t = \frac{-2 + \sqrt{4 + 4\tan^2(2t)}}{2\tan(2t)} \\
 &\implies \tan t = \frac{\sqrt{1 + \tan^2(2t)} - 1}{\tan(2t)}
 \end{aligned}$$

(5) Now let us see what we get at $N = 16$. With $t = 11.25^\circ$, and with trigonometric input concerning $2t = 22.5^\circ$ from the proof of Proposition 12.18, we obtain:

$$\sin(11.25^\circ) = \sqrt{\frac{1 - \sqrt{1 - \frac{2-\sqrt{2}}{4}}}{2}} = \frac{\sqrt{2 - \sqrt{2 + \sqrt{2}}}}{2}$$

As for the tangent, again with $t = 11.25^\circ$, and with trigonometric input concerning $2t = 22.5^\circ$ from the proof of Proposition 12.18, we obtain:

$$\tan(11.25^\circ) = \frac{\sqrt{1 + (\sqrt{2} - 1)^2} - 1}{\sqrt{2} - 1} = \frac{\sqrt{4 - 2\sqrt{2}} - 1}{\sqrt{2} - 1}$$

(6) At the level of the numerics, the formulae that we obtain are as follows:

$$\sin(11.25^\circ) = 0.915\dots, \quad \tan(11.25^\circ) = 0.198\dots$$

By multiplying by 16, as required by the estimate that we found, we have:

$$16\sin(11.25^\circ) = 3.121\dots, \quad \tan(11.25^\circ) = 3.182\dots$$

Thus, we are led to the hexadecagon estimate in the statement, namely:

$$3.121 < \pi < 3.183$$

(7) Regarding now better estimates, the method indicated in the statement, namely using $N = 2^n$ with $n \gg 0$, normally works, with the remark however that the convergence is very slow, and with the extra remark, which is concerning too, that the numeric extraction of square roots, required by our formulae in (4), is no easy business.

(8) In short, and contrary to what we saw before for e , approximating π is no easy business. However, with some patience, we are led to the known figure for π , namely:

$$\pi = 3.14159 \dots$$

By the way, as a young nerd, make sure that you get to know more decimals of π than I do. So, exercise for you to do some computations, and memorize what you found. \square

On the same topic, and at a more advanced level now, we have:

THEOREM 12.20 (Leibnitz). *We have the following formula,*

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \dots$$

which can be used in order to find the decimals of π .

PROOF. Many things can be said here, the idea being as follows:

(1) To start with, let us recall from chapter 6 that we have the following formula:

$$(\arctan x)' = \frac{1}{1+x^2}$$

On the other hand, the function on the right expands as a series as:

$$\frac{1}{1+x^2} = \sum_{k=0}^{\infty} (-1)^k x^{2k}$$

Now by integrating, we obtain the following formula for the arctangent:

$$\arctan x = \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} x^{2k+1}$$

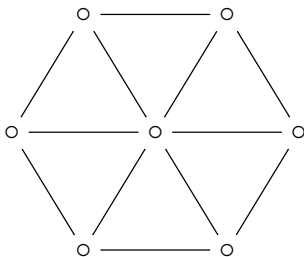
(2) But this gives, with $x = 1$, the Leibnitz formula in the statement, namely:

$$\frac{\pi}{4} = \arctan(1) = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \dots$$

And in particular, with this, we can compute the decimals of π , as stated.

(3) This being said, regarding this, and the approximation of π in general, there is a long story here, involving many mathematicians, and their ideas, worth to be told.

(4) To start with, forgetting about calculus and everything advanced, that is, fast forward to the ancient times, we know that π appears as the semiperimeter of the circle having radius 1. And we also know, by drawing a hexagon, that $\pi > 3$, not by much:



(5) But this suggests approximating π by computing the perimeter of various inscribed and circumscribed regular polygons. In modern terms, as explained in the proof of Theorem 12.19, and save for the old angle conventions used there, which must be updated according to $90^\circ = \pi/2$, what we get from a N -gon is the following estimate:

$$N \sin\left(\frac{\pi}{N}\right) < \pi < N \tan\left(\frac{\pi}{N}\right)$$

(6) So, let us see how this formula works. In view of our halving formulae for \sin, \cos, \tan , it makes sense to do the computations for $N = 2^s$. And for $N = 4, 8, 16$, what we get, again as explained in the proof of Theorem 12.19, is as follows:

$$\begin{aligned} 2\sqrt{2} &< \pi < 4 \\ 4\sqrt{2 - \sqrt{2}} &< \pi < 8(\sqrt{2} - 1) \\ 8\sqrt{2 - \sqrt{2 + \sqrt{2}}} &< \pi < 16\left(\sqrt{4 + 2\sqrt{2}} - \sqrt{2} - 1\right) \end{aligned}$$

(7) Numerically now, we obtain in this way the following estimates:

$$\begin{aligned} 2.828 &< \pi < 4 \\ 3.061 &< \pi < 3.314 \\ 3.121 &< \pi < 3.183 \end{aligned}$$

Which does not look great, lots of work for doing all this, especially in extracting the square roots, and this, for not that many decimals, by the end of the day.

(8) This being said, with this being the only method available, let us see what we get at higher $N = 2^s$. And here, forgetting about the tangent, and focusing on the sine, we are led to the following formula for π , coming via the above method, with $k = s - 1$:

$$\pi = \lim_{k \rightarrow \infty} 2^k \underbrace{\sqrt{2 - \sqrt{2 + \sqrt{2 + \dots + \sqrt{2}}}}}_{k \text{ square roots}}$$

(9) So, this was the old method for computing π , but in practice this has not prevented the ancients from doing lots of computations, and approximating π to a fair amount of decimals, a few dozens, which is just perfect for usual engineering purposes:

$$\pi = 3.14159265358979323846 \dots$$

(10) The continuation of the story involves calculus, with the formula in the statement by Leibnitz, and several modifications of this Leibnitz formula. Among others, we have here the following beautiful formula of Euler, that we already met in this book:

$$\frac{\pi^2}{6} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \dots$$

However, in practice, while these formulae are certainly easier to evaluate than the old ones, their convergence is very slow too, leading to a few hundred decimals for π .

(11) And then, in more modern times, there was the following key formula by Ramanujan, coming from advanced arithmetic, which converges remarkably fast:

$$\frac{1}{\pi} = \frac{2\sqrt{2}}{99^2} \sum_{k=0}^{\infty} \frac{(4k)!}{k!^4} \cdot \frac{26390k + 1103}{396^{4k}}$$

And it is this formula, and its versions, those which are presently used. □

We will be back to this, and notably to the Euler formula in (10), in Part IV.

12d. Transcendence of pi

With this discussed, let us get now to algebraic questions. The situation for π is a bit similar to that for e , discussed before in this chapter, with however a few notable differences appearing, and at the level of the main results, these are as follows:

THEOREM 12.21. *The number $\pi = 3.14159\dots$ has the following properties:*

- (1) *It is irrational.*
- (2) *It is transcendental.*

PROOF. This is indeed something quite routine, by using the same ideas as before for e , but with everything being now a bit more technical, the idea being as follows:

(1) In what regards the irrationality of π , no simple argument as for e is available, so we rather have to take our inspiration from the Hermite proof of the transcendence of e , given above. With this idea in mind, consider the following quantities:

$$I_n(t) = \int_{-1}^1 (1 - x^2)^n \cos(xt) dx$$

By double partial integration we obtain the following formula:

$$\begin{aligned}
I_n(t) &= \int_{-1}^1 (1-x^2)^n \cos(xt) dx \\
&= \int_{-1}^1 2nx(1-x^2)^{n-1} \frac{\sin(xt)}{t} dx \\
&= \frac{2n}{t} \int_{-1}^1 x(1-x^2)^{n-1} \sin(xt) dx \\
&= \frac{2n}{t} \int_{-1}^1 [(1-x^2)^{n-1} - 2(n-1)x^2(1-x^2)^{n-2}] \frac{\cos(xt)}{t} dx \\
&= \frac{2n}{t^2} \int_{-1}^1 (1-x^2)^{n-2} [1-x^2 - 2(n-1)x^2] \cos(xt) dx \\
&= \frac{2n}{t^2} \int_{-1}^1 (1-x^2)^{n-2} [1 - (2n-1)x^2] \cos(xt) dx \\
&= \frac{2n}{t^2} \int_{-1}^1 (1-x^2)^{n-2} [(2n-1)(1-x^2) - (2n-2)] \cos(xt) dx \\
&= \frac{2n}{t^2} [(2n-1)I_{n-1}(t) - (2n-2)I_{n-2}(t)]
\end{aligned}$$

Thus, we have the following recurrence relation for our quantities:

$$t^2 I_n(t) = 2n(2n-1)I_{n-1}(t) - 4n(n-1)I_{n-2}(t)$$

In terms of $J_n(t) = t^{2n+1}I_n(t)$, this recurrence formula becomes:

$$J_n(t) = 2n(2n-1)J_{n-1}(t) - 4n(n-1)t^2 J_{n-2}(t)$$

Regarding now the initial data, for this latter recurrence, this is as follows:

$$J_0(t) = 2 \sin t \quad , \quad J_1(t) = -4t \cos t + 4 \sin t$$

We conclude from this that we must have a formula as follows, with P_n, Q_n being certain polynomials of degree $\leq n$, with integer coefficients:

$$J_n(t) = n!(P_n(t) \sin t + Q_n(t) \cos t)$$

Now observe that with $t = \pi/2$, we obtain from this the following formula:

$$\left(\frac{\pi}{2}\right)^{2n+1} I_n\left(\frac{\pi}{2}\right) = n! P_n\left(\frac{\pi}{2}\right)$$

Assume now by contradiction that π is rational, so that $\pi/2 = a/b$ with $a, b \in \mathbb{N}$. We can rewrite the formula found above in the following more convenient way:

$$\frac{a^{2n+1}}{n!} I_n\left(\frac{a}{b}\right) = b^{2n+1} P_n\left(\frac{a}{b}\right)$$

But, by definition of the integrals I_n , we know that we have:

$$I_n\left(\frac{a}{b}\right) = I_n\left(\frac{\pi}{2}\right) \in (0, 2)$$

Thus with $n \gg 0$ the number on the left belongs to $(0, 1)$, which is contradictory, because the number on the right is an integer. And so π is irrational, as claimed.

(2) Regarding now the transcendence of π , again it is possible to adapt the ideas of Hermite for e , from the proof of Theorem 12.17, but this remains something quite technical. Instead, it is better to have an algebraic look at this, by using the Lindemann-Weierstrass theorem, which states that if $a_1, \dots, a_n \in \mathbb{C}$ are algebraically independent over \mathbb{Q} , then e^{a_1}, \dots, e^{a_n} are algebraically independent too over \mathbb{Q} . To be more precise:

– To start with, observe that the Lindemann-Weierstrass theorem shows with $n = 1$ that $e = e^1$ is transcendental. Thus, this is definitely something non-trivial.

– However, this is something that can be proved, with some knowledge of algebra and field theory, in the spirit of what we did in chapter 8, when discussing Galois theory.

– And, in relation with π , we can use again $n = 1$, but this time in conjunction with the Euler formula $e^{\pi i} = -1$, and we obtain that π is transcendental. \square

Summarizing, many things to be learned here, and for getting introduced to all this, you first need to read Galois theory in detail, say from Lang [63], and then go with a reasonably advanced number theory book, and there are many good choices here.

As a last piece of mathematics, in relation with trigonometry and approximations, we have the following remarkable result, concerning π itself, due to Buffon:

THEOREM 12.22. *The probability for a needle of length 1, when thrown on a grid of parallel 1-spaced lines, to intersect one line, is:*

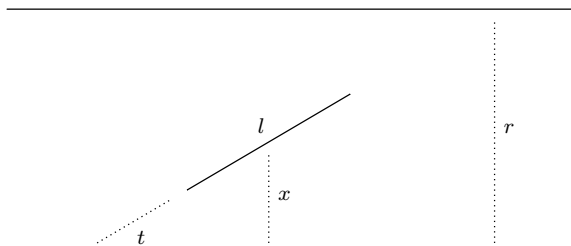
$$P = \frac{2}{\pi}$$

Moreover, we have generalizations of this result, with needles of arbitrary length, thrown over a grid of parallel lines, with arbitrary spacing.

PROOF. This is something quite tricky, and mandatory for properly learning probability theory, and science in general, because there are several possible modelings of the problem, leading, quite surprisingly, to different values of P . And, obviously, only one such modeling can be the correct one. So, here is how the correct computation goes:

(1) Let us examine, right from the beginning, the general question involving throwing a needle of length l on a grid of parallel r -spaced lines. We denote by $x \in [0, r/2]$ the

distance from the middle of the needle to the closest line, and by $t \in [0, \pi/2]$ the acute angle formed by the needle and that line, according to the following picture:



(2) Now since the distance $x \in [0, r/2]$ varies uniformly, with density $2/r$, and the angle $t \in [0, \pi/2]$ varies uniformly too, with density $2/\pi$, and these variables are independent, in the case $l \leq r$ the probability for our needle to cross a line is given by:

$$\begin{aligned} P\left(x \leq \frac{l \sin t}{2}\right) &= \int_0^{\pi/2} \int_0^{l \sin t/2} \frac{4}{\pi r} dx dt \\ &= \frac{4}{\pi r} \int_0^{\pi/2} \frac{l \sin t}{2} dt \\ &= \frac{2l}{\pi r} \end{aligned}$$

(3) In particular, with $l = r = 1$ the crossing probability is $P = 2/\pi$, as stated. \square

12e. Exercises

This was a quite technical and exciting chapter, and as exercises on this, we have:

EXERCISE 12.23. *Learn more about the Poisson laws, and their properties.*

EXERCISE 12.24. *Read also about group representations, and their characters.*

EXERCISE 12.25. *Compute decimals of e , as many as you can.*

EXERCISE 12.26. *Learn some other proofs, for the transcendence of e .*

EXERCISE 12.27. *Compute decimals of π , as many as you can.*

EXERCISE 12.28. *Learn about modular forms, and the Ramanujan formula for π .*

EXERCISE 12.29. *Clarify what we said, in relation with the transcendence of π .*

EXERCISE 12.30. *Have more fun with the Buffon needle, and its mathematics.*

As bonus exercise, read a bit about the history of e , π and mathematics.

Part IV

Analytic aspects

Because the night belongs to lovers
Because the night belongs to lust
Because the night belongs to lovers
Because the night belongs to us

CHAPTER 13

Complex analysis

13a. Complex functions

In order to further advance on our various questions regarding the primes, we are in need of more calculus input. And here, passed many ad-hoc things that can be said, we are mostly in need of complex analysis. For instance, as we will soon discover, the zeta function naturally lives over \mathbb{C} , so the ground field for analytic number theory is \mathbb{C} .

Be said in passing, as a remarkable coincidence, \mathbb{C} is also the ground field for many other interesting theories, such as quantum mechanics, and advanced physics in general. So, with complex analysis, we will be learning here very useful mathematics.

Getting started now, we already met complex functions $f : \mathbb{C} \rightarrow \mathbb{C}$ in chapter 9, when talking complex numbers, so let us begin with a quick review of that material. The simplest such functions are the polynomials, and to be known about them is:

THEOREM 13.1. *Each polynomial $P \in \mathbb{C}[X]$ can be regarded as a complex function $P : \mathbb{C} \rightarrow \mathbb{C}$, which is continuous. Also, we can write*

$$P(x) = a(x - r_1) \dots (x - r_n)$$

with $a \in \mathbb{C}$, and with the numbers $r_1, \dots, r_n \in \mathbb{C}$ being the roots of P .

PROOF. This was something that we discussed in chapter 9, the idea being that a root can be constructed by contradiction, and a recurrence does the rest. \square

Next in complexity come the rational functions. And here, as a first surprise, in what regards the basic geometric series, things are quite complicated over \mathbb{C} , as follows:

PROPOSITION 13.2. *We have the following formula, valid for any $|x| < 1$,*

$$\frac{1}{1-x} = 1 + x + x^2 + \dots$$

but, unlike in the real case, the geometric meaning of this formula is quite unclear.

PROOF. This is again something from chapter 9, and we refer to the material there for a discussion about this, mathematics of the spiral which appears, over \mathbb{C} . \square

Then, we have the general rational functions, to be discussed. We already talked about these, over \mathbb{R} , in chapter 8, and over \mathbb{C} the basics are similar, as follows:

PROPOSITION 13.3. *The quotients of complex polynomials, called rational functions, when written in reduced form, as follows, with P, Q prime to each other,*

$$f = \frac{P}{Q}$$

are well-defined and continuous outside the zeroes $P_f \subset \mathbb{C}$ of Q , called poles of f :

$$f : \mathbb{C} - P_f \rightarrow \mathbb{C}$$

Also, these functions are stable under summing, making products and taking inverses,

$$\frac{P}{Q} + \frac{R}{S} = \frac{PS + QR}{QS} \quad , \quad \frac{P}{Q} \cdot \frac{R}{S} = \frac{PR}{QS} \quad , \quad \left(\frac{P}{Q}\right)^{-1} = \frac{Q}{P}$$

so they form a field $\mathbb{C}(X)$, called field of real rational functions.

PROOF. As already mentioned, this is something from chapter 8, now formulated over \mathbb{C} , and with everything being self-explanatory, and trivial from definitions. \square

Still regarding the rational functions, by taking now advantage of the fact that we are over \mathbb{C} , where all polynomials have roots, we can say more about them, as follows:

THEOREM 13.4. *The complex rational functions can be written as follows,*

$$f(x) = \sum_i \frac{A_i(x)}{(r_i - x)^{n_i}}$$

with $A_i \in \mathbb{C}[X]$, and $r_i \in \mathbb{C}$ being the poles. Also, we have the following formula,

$$\frac{1}{(r - x)^n} = \frac{1}{r^n} \sum_{k=0}^{\infty} \binom{n+k-1}{n-1} \left(\frac{x}{r}\right)^k$$

valid for $|x| < r$, which computes these rational functions, in practice.

PROOF. Consider indeed a rational function $f = P/Q$, with $P, Q \in \mathbb{C}[X]$ chosen prime to each other. By factorizing Q we have a formula as follows, with $P(r_i) \neq 0$:

$$f(x) = \frac{P(x)}{(r_1 - x)^{n_1} \dots (r_k - x)^{n_k}}$$

Now comes the trick. Assuming that $S, T \in \mathbb{C}[X]$ are prime to each other, we can find, a bit as for the usual numbers, by performing successive divisions, polynomials $A, B \in \mathbb{C}[X]$ such that $AT + BS = 1$, and so that the following happens:

$$\frac{1}{ST} = \frac{A}{S} + \frac{B}{T}$$

Thus, we are led to the first formula in the statement. As for the second formula, this is something that we know since chapter 6, binomial formula with negative exponent. \square

Getting now to more complicated functions, such as \sin , \cos , \exp , \log , again many things extend well from real to complex, the basic theory here being as follows:

THEOREM 13.5. *The functions \sin , \cos , \exp , \log have complex extensions, given by*

$$\sin x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l+1}}{(2l+1)!} \quad , \quad \cos x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l}}{(2l)!}$$

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \quad , \quad \log(1+x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k}$$

with $|x| < 1$ needed for \log , which are continuous over their domain, and satisfy the formulae $e^{x+y} = e^x e^y$ and $e^{ix} = \cos x + i \sin x$.

PROOF. This is a mixture of trivial and non-trivial results, as follows:

(1) We already know about e^x from chapter 9, the idea being that the convergence of the series, and then the continuity of e^x , come from the following estimate:

$$|e^x| \leq \sum_{k=0}^{\infty} \frac{|x|^k}{k!} = e^{|x|} < \infty$$

Regarding $\sin x$, the same method works, with the following estimate:

$$|\sin x| \leq \sum_{l=0}^{\infty} \frac{|x|^{2l+1}}{(2l+1)!} \leq \sum_{k=0}^{\infty} \frac{|x|^k}{k!} = e^{|x|}$$

The same goes for $\cos x$, the estimate here being as follows:

$$|\cos x| \leq \sum_{l=0}^{\infty} \frac{|x|^{2l}}{(2l)!} \leq \sum_{k=0}^{\infty} \frac{|x|^k}{k!} = e^{|x|}$$

Finally, regarding the basic formulae satisfied by \sin , \cos , \exp , we already know from chapter 9 that the exponential has indeed the following properties:

$$e^{x+y} = e^x e^y \quad , \quad e^{ix} = \cos x + i \sin x$$

(2) In order to discuss now the complex logarithm function \log , let us first study some more the complex exponential function \exp . By using $e^{x+y} = e^x e^y$ we obtain $e^x \neq 0$ for any $x \in \mathbb{C}$, so the complex exponential function is as follows:

$$\exp : \mathbb{C} \rightarrow \mathbb{C} - \{0\}$$

Now since we have $e^{x+iy} = e^x e^{iy}$ for $x, y \in \mathbb{R}$, with e^x being surjective onto $(0, \infty)$, and with e^{iy} being surjective onto the unit circle \mathbb{T} , we deduce that $\exp : \mathbb{C} \rightarrow \mathbb{C} - \{0\}$ is surjective. Also, again by using $e^{x+iy} = e^x e^{iy}$, we deduce that we have:

$$e^x = e^y \iff x - y \in 2\pi i \mathbb{Z}$$

(3) With these ingredients in hand, we can now talk about \log . Indeed, let us fix a horizontal strip in the complex plane, having width 2π :

$$S = \left\{ x + iy \mid x \in \mathbb{R}, y \in [a, a + 2\pi) \right\}$$

We know from the above that the restriction map $\exp : S \rightarrow \mathbb{C} - \{0\}$ is bijective, so we can define \log as to be the inverse of this map:

$$\log = \exp^{-1} : \mathbb{C} - \{0\} \rightarrow S$$

(4) In practice now, the best is to choose for instance $a = 0$, or $a = -\pi$, as to have the whole real line included in our strip, $\mathbb{R} \subset S$. In this case on \mathbb{R}_+ we recover the usual logarithm, while on \mathbb{R}_- we obtain complex values, as for instance $\log(-1) = \pi i$ in the case $a = 0$, or $\log(-1) = -\pi i$ in the case $a = -\pi$, coming from $e^{\pi i} = -1$.

(5) Finally, assuming $|x| < 1$, we can consider the following series, which converges:

$$f(1+x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k}$$

As explained in chapter 6, a direct computation shows that we have:

$$f((1+x)(1+y)) = f(1+x) + f(1+y)$$

Thus f satisfies the functional equation of \log , and so $f(x) = \log x$, as claimed. \square

Moving on, Theorem 13.5 leads us into the question on whether the other formulae that we know about \sin, \cos , such as the values of these functions on sums $x + y$, or on doubles $2x$, extend to the complex setting. Things are quite tricky here, and in relation with this, we have the following result, which is something of general interest:

THEOREM 13.6. *The following functions, called hyperbolic sine and cosine,*

$$\sinh x = \frac{e^x - e^{-x}}{2}, \quad \cosh x = \frac{e^x + e^{-x}}{2}$$

are subject to the following formulae:

- (1) $e^x = \cosh x + \sinh x$.
- (2) $\sinh(ix) = i \sin x$, $\cosh(ix) = \cos x$, for $x \in \mathbb{R}$.
- (3) $\sinh(x+y) = \sinh x \cosh y + \cosh x \sinh y$.
- (4) $\cosh(x+y) = \cosh x \cosh y + \sinh x \sinh y$.
- (5) $\sinh x = \sum_l \frac{x^{2l+1}}{(2l+1)!}$, $\cosh x = \sum_l \frac{x^{2l}}{(2l)!}$.

PROOF. The formula (1) follows from definitions. As for (2), this follows from:

$$\begin{aligned} \sinh(ix) &= \frac{e^{ix} - e^{-ix}}{2} = \frac{\cos x + i \sin x}{2} - \frac{\cos x - i \sin x}{2} = i \sin x \\ \cosh(ix) &= \frac{e^{ix} + e^{-ix}}{2} = \frac{\cos x + i \sin x}{2} + \frac{\cos x - i \sin x}{2} = \cos x \end{aligned}$$

Regarding now (3,4), observe first that the formula $e^{x+y} = e^x + e^y$ reads:

$$\cosh(x+y) + \sinh(x+y) = (\cosh x + \sinh x)(\cosh y + \sinh y)$$

Thus, we have some good explanation for (3,4), and in practice, these formulae can be checked by direct computation, as follows:

$$\begin{aligned}\frac{e^{x+y} - e^{-x-y}}{2} &= \frac{e^x - e^{-x}}{2} \cdot \frac{e^y + e^{-y}}{2} + \frac{e^x + e^{-x}}{2} \cdot \frac{e^y - e^{-y}}{2} \\ \frac{e^{x+y} + e^{-x-y}}{2} &= \frac{e^x + e^{-x}}{2} \cdot \frac{e^y + e^{-y}}{2} + \frac{e^x - e^{-x}}{2} \cdot \frac{e^y - e^{-y}}{2}\end{aligned}$$

Finally, (5) is clear from the definition of \sinh , \cosh , and from $e^x = \sum_k \frac{x^k}{k!}$. \square

Finally, we can talk as well about powers, in the following way:

FACT 13.7. *Under suitable assumptions, we can talk about x^y with $x, y \in \mathbb{C}$, and in particular about the complex functions a^x and x^a , with $a \in \mathbb{C}$.*

To be more precise, in what regards x^y , we already know from basic calculus that things are quite tricky, even in the real case. In the complex case the same problems appear, along with some more, but these questions can be solved by using the above theory of \exp , \log . To be more precise, in order to solve the first question, we can set:

$$x^y = e^{y \log x}$$

We will be back to these functions later, when we will have more tools for studying them. In fact, all of a sudden, we are now into quite complicated mathematics, and we cannot really deal with such questions, with bare hands. More later.

13b. Holomorphic functions

Let us study now the differentiability of the complex functions $f : \mathbb{C} \rightarrow \mathbb{C}$. Things here are quite tricky, but let us start with a straightforward definition, as follows:

DEFINITION 13.8. *We say that a function $f : X \rightarrow \mathbb{C}$ is differentiable in the complex sense when the following limit is defined for any $x \in X$:*

$$f'(x) = \lim_{t \rightarrow 0} \frac{f(x+t) - f(x)}{t}$$

In this case, we also say that f is holomorphic, and we write $f \in H(X)$.

As basic examples, we have the power functions $f(x) = x^n$. Indeed, the derivative of such a power function can be computed exactly as in the real case, and we get:

$$(x^n)' = \lim_{t \rightarrow 0} \frac{(x+t)^n - x^n}{t} = \lim_{t \rightarrow 0} \frac{nx^{n-1}t}{t} = nx^{n-1}$$

We will see later more computations of this type, similar to those from the real case. To summarize, our definition of differentiability seems to work nicely.

So, let us start developing now some general theory for the complex derivatives, in the sense of Definition 13.8. The basic results from the real case extend well, as follows:

PROPOSITION 13.9. *We have the following results:*

- (1) $(f + g)' = f' + g'$.
- (2) $(\lambda f)' = \lambda f'$.
- (3) $(fg)' = f'g + fg'$.
- (4) $(f \circ g)' = f'(g)g'$.

PROOF. These formulae are all clear from definitions, following exactly as in the real case. Thus, we are led to the conclusions in the statement. \square

As an obvious consequence of (1,2) above, any polynomial $P \in \mathbb{C}[X]$ is differentiable, with its derivative being given by the same formula as in the real case, namely:

$$P(x) = \sum_{k=0}^n c_k x^k \implies P'(x) = \sum_{k=1}^n k c_k x^{k-1}$$

More generally, any rational function $f \in \mathbb{C}(X)$ is differentiable on its domain, that is, outside its poles, because if we write $f = P/Q$ with $P, Q \in \mathbb{C}[X]$, we have:

$$f' = \left(\frac{P}{Q} \right)' = \frac{P'Q - PQ'}{Q^2}$$

Let us record these conclusions in a statement, as follows:

PROPOSITION 13.10. *The following happen:*

- (1) *Any polynomial $P \in \mathbb{C}[X]$ is holomorphic, and in fact infinitely differentiable in the complex sense, with all its derivatives being polynomials.*
- (2) *Any rational function $f \in \mathbb{C}(X)$ is holomorphic, and in fact infinitely differentiable, with all its derivatives being rational functions.*

PROOF. This follows indeed from the above discussion. \square

Let us look now into more complicated complex functions that we know. And here, surprise, things are quite tricky, the result being as follows:

THEOREM 13.11. *The following happen:*

- (1) *\sin, \cos, \exp, \log are holomorphic, and in fact are infinitely differentiable, with their derivatives being given by the same formulae as in the real case.*
- (2) *However, functions like \bar{x} or $|x|$ are not holomorphic, and this because the limit defining $f'(x)$ depends on the way we choose $t \rightarrow 0$.*

PROOF. There are several things going on here, the idea being as follows:

(1) Here the first assertion is standard, because our functions \sin, \cos, \exp, \log have Taylor series that we know, and the derivative can be therefore computed by using the same rule as in the real case, similar to the one for polynomials, namely:

$$f(x) = \sum_{k=0}^{\infty} c_k x^k \implies f'(x) = \sum_{k=1}^{\infty} k c_k x^{k-1}$$

(2) Regarding now the function $f(x) = \bar{x}$, the point here is that we have:

$$\frac{f(x+t) - f(x)}{t} = \frac{\bar{x} + \bar{t} - \bar{x}}{t} = \frac{\bar{t}}{t}$$

But this limit does not converge with $t \rightarrow 0$, for instance because with $t \in \mathbb{R}$ we obtain 1 as limit, while with $t \in i\mathbb{R}$ we obtain -1 as limit. In fact, with $t = rw$ with $|w| = 1$ fixed and $r \in \mathbb{R}$, $r \rightarrow 0$, we can obtain as limit any number on the unit circle:

$$\lim_{r \rightarrow 0} \frac{f(x+rw) - f(x)}{rw} = \lim_{r \rightarrow 0} \frac{r\bar{w}}{rw} = \bar{w}^2$$

(3) The situation for the function $f(x) = |x|$ is similar. To be more precise, we have:

$$\frac{f(x+rw) - f(x)}{rw} = \frac{|x+rw| - |x|}{r} \cdot \bar{w}$$

Thus with $|w| = 1$ fixed and $r \rightarrow 0$ we obtain a certain multiple of \bar{w} , with the multiplication factor being computed as follows:

$$\begin{aligned} \frac{|x+rw| - |x|}{r} &= \frac{|x+rw|^2 - |x|^2}{(|x+rw| + |x|)r} \\ &\simeq \frac{xr\bar{w} + \bar{x}rw}{2|x|r} \\ &= \operatorname{Re} \left(\frac{x\bar{w}}{|x|} \right) \end{aligned}$$

Now by making w vary on the unit circle, as in (2) above, we can obtain in this way limits pointing in all possible directions, so our limit does not converge, as stated. \square

The above result is quite surprising, because we are so used, from the real case, to the notion of differentiability to correspond to some form of “smoothness” of the function, and to be more precisely, “smoothness at first order”. Or, if you prefer, to correspond to the “non-bumpiness” of the function. So, we are led to the following dilemma:

DILEMMA 13.12. *It's either that \bar{x} and $|x|$ are smooth, as the intuition suggests, and we are wrong with our definition of differentiability. Or that \bar{x} and $|x|$ are bumpy, while this being not very intuitive, and we are right with our definition of differentiability.*

And we won't get discouraged by this. After all, this is just some empty talking, and if there is something to rely upon, mathematics and computations, these are the computations from the proof of Theorem 13.11. So, based on that computations, let us formulate the following definition, coming as a complement to Definition 13.8:

DEFINITION 13.13. *A function $f : X \rightarrow \mathbb{C}$ is called differentiable:*

- (1) *In the real sense, if the following two limits converge, for any $x \in X$:*

$$f'_1(x) = \lim_{t \in \mathbb{R} \rightarrow 0} \frac{f(x+t) - f(x)}{t}, \quad f'_i(x) = \lim_{t \in i\mathbb{R} \rightarrow 0} \frac{f(x+t) - f(x)}{t}$$

- (2) *In a radial sense, if the following limit converges, for any $x \in X$, and $w \in \mathbb{T}$:*

$$f'_w(x) = \lim_{t \in w\mathbb{R} \rightarrow 0} \frac{f(x+t) - f(x)}{t}$$

- (3) *In the complex sense, if the following limit converges, for any $x \in X$:*

$$f'(x) = \lim_{t \rightarrow 0} \frac{f(x+t) - f(x)}{t}$$

If f is differentiable in the complex sense, we also say that f is holomorphic.

With this, we have (3) \implies (2) \implies (1), and most of the functions that we know, namely the polynomials, the rational functions, and sin, cos, exp, log, satisfy (3). However, the functions \bar{x} , $|x|$ satisfy (2), which is not bad, but do not satisfy (3).

Back to business now, all the examples of holomorphic functions that we have are infinitely differentiable, and this raises the question of finding a function such that f' exists, while f'' does not exist. Quite surprisingly, we will see that such functions do not exist. In order to get into this latter phenomenon, let us start with:

THEOREM 13.14. *Each power series $f(x) = \sum_n c_n x^n$ has a radius of convergence $R \in [0, \infty]$*

which is such that f converges for $|x| < R$, and diverges for $|x| > R$. We have:

$$R = \frac{1}{C}, \quad C = \limsup_{n \rightarrow \infty} \sqrt[n]{|c_n|}$$

Also, in the case $|x| = R$ the function f can either converge, or diverge.

PROOF. This follows from the Cauchy criterion for series, from chapter 5, which says that a series $\sum_n x_n$ converges if $c < 1$, and diverges if $c > 1$, where:

$$c = \limsup_{n \rightarrow \infty} \sqrt[n]{|x_n|}$$

Indeed, with $x_n = |c_n x^n|$ we obtain that the convergence radius $R \in [0, \infty]$ exists, and is given by the formula in the statement. Finally, for the examples and counterexamples at the end, when $|x| = R$, the simplest here is to use $f(x) = \sum_n x^n$, where $R = 1$. \square

Back now to our questions regarding derivatives, we have:

THEOREM 13.15. *Assuming that a function $f : X \rightarrow \mathbb{C}$ is analytic, in the sense that it is a series, around each point $x \in X$,*

$$f(x+t) = \sum_{n=0}^{\infty} c_n t^n$$

it follows that f is infinitely differentiable, in the complex sense. In particular, f' exists, and so f is holomorphic in our sense.

PROOF. Assuming that f is analytic, as in the statement, we have:

$$f'(x+t) = \sum_{n=1}^{\infty} n c_n t^{n-1}$$

Moreover, the radius of convergence is the same, as shown by the following computation, using the Cauchy formula for the convergence radius, and $\sqrt[n]{n} \rightarrow 1$:

$$\frac{1}{R'} = \limsup_{n \rightarrow \infty} \sqrt[n]{|n c_n|} = \limsup_{n \rightarrow \infty} \sqrt[n]{|c_n|} = \frac{1}{R}$$

Thus f' exists and is analytic, on the same domain, and this gives the result. \square

13c. Cauchy formula

Our goal in what follows will be that of proving that any holomorphic function is analytic. This is something quite subtle, which cannot be proved with bare hands, and requires lots of preliminaries. Getting to these preliminaries now, our claim is that a lot of useful knowledge, in order to deal with the holomorphic functions, can be gained by further studying the analytic functions, and even the usual polynomials $P \in \mathbb{C}[X]$.

So, let us further study the polynomials $P \in \mathbb{C}[X]$, and other analytic functions. We already know from before that in the polynomial case, $P \in \mathbb{C}[X]$, some interesting things happen, because any such polynomial has a root, and even $\deg(P)$ roots, after a recurrence. Keeping looking at polynomials, with the same methods, we are led to:

THEOREM 13.16. *Any polynomial $P \in \mathbb{C}[X]$ satisfies the maximum principle, in the sense that given a domain D , with boundary γ , we have:*

$$\exists x \in \gamma \quad , \quad |P(x)| = \max_{y \in D} |P(y)|$$

That is, the maximum of $|P|$ over a domain is attained on its boundary.

PROOF. In order to prove this, we can split D into connected components, and then assume that D is connected. Moreover, we can assume that D has no holes, and so is homeomorphic to a disk, and even assume that D itself is a disk. But with this assumption made, the result follows from by contradiction, by using the same arguments as in the

proof of the existence of a root. To be more precise, assume $\deg P \geq 1$, and that the maximum of $|P|$ is attained at the center of a disk $D = D(z, r)$:

$$|P(z)| = \max_{x \in D} |P(x)|$$

We can write then $P(z+t) \simeq P(z) + ct^k$ with $c \neq 0$, for t small, and by suitably choosing the argument of t on the unit circle we conclude, as in the proof for the existence of the roots, that the function $|P|$ cannot have a local maximum at z , as stated. \square

A good explanation for the fact that the maximum principle holds for polynomials $P \in \mathbb{C}[X]$ could be that the values of such a polynomial inside a disk can be recovered from its values on the boundary. And fortunately, this is indeed the case, and we have:

THEOREM 13.17. *Given a polynomial $P \in \mathbb{C}[X]$, and a disk D , with boundary γ , we have the following formulae, with the integrations being the normalized, mass 1 ones:*

- (1) *P satisfies the plain mean value formula $P(x) = \int_D P(y) dy$.*
- (2) *P satisfies the boundary mean value formula $P(x) = \int_\gamma P(y) dy$.*

PROOF. As a first observation, the two mean value formulae in the statement are equivalent, by restricting the attention to disks D , having as boundaries circles γ , and using annuli and polar coordinates for the proof of the equivalence. As for the formulae themselves, these can be checked by direct computation for a disk D , with the formulation in (2) being the most convenient. Indeed, for a monomial $P(x) = x^n$ we have:

$$\begin{aligned} \int_\gamma y^n dy &= \frac{1}{2\pi} \int_0^{2\pi} (x + re^{it})^n dt \\ &= \frac{1}{2\pi} \int_0^{2\pi} \sum_{k=0}^n \binom{n}{k} x^k (re^{it})^{n-k} dt \\ &= \sum_{k=0}^n \binom{n}{k} x^k r^{n-k} \frac{1}{2\pi} \int_0^{2\pi} e^{i(n-k)t} dt \\ &= \sum_{k=0}^n \binom{n}{k} x^k r^{n-k} \delta_{kn} \\ &= x^n \end{aligned}$$

Here we have used the following key identity, valid for any exponent $m \in \mathbb{Z}$:

$$\begin{aligned} \frac{1}{2\pi} \int_0^{2\pi} e^{imt} dt &= \frac{1}{2\pi} \int_0^{2\pi} \cos(mt) + i \sin(mt) dt \\ &= \delta_{m0} + i \cdot 0 \\ &= \delta_{m0} \end{aligned}$$

Thus, we have the result for monomials, and the general case follows by linearity. \square

All the above is very nice, but we can in fact do even better, with a more powerful integration formula. Let us start with some preliminaries. We first have:

PROPOSITION 13.18. *We can integrate functions f over curves γ by setting*

$$\int_{\gamma} f(x)dx = \int_a^b f(\gamma(t))\gamma'(t)dt$$

with this quantity being independent on the parametrization $\gamma : [a, b] \rightarrow \mathbb{C}$.

PROOF. We must prove that the quantity in the statement is independent on the parametrization. In other words, we must prove that if we pick two different parametrizations $\gamma, \eta : [a, b] \rightarrow \mathbb{C}$ of our curve, then we have the following formula:

$$\int_a^b f(\gamma(t))\gamma'(t)dt = \int_a^b f(\eta(t))\eta'(t)dt$$

But for this purpose, let us write $\gamma = \eta\phi$, with $\phi : [a, b] \rightarrow [a, b]$ being a certain function, that we can assume to be bijective, via an elementary cut-and-paste argument. By using the chain rule for derivatives, and the change of variable formula, we have:

$$\begin{aligned} \int_a^b f(\gamma(t))\gamma'(t)dt &= \int_a^b f(\eta\phi(t))(\eta\phi)'(t)dt \\ &= \int_a^b f(\eta\phi(t))\eta'(\phi(t))\phi'(t)dt \\ &= \int_a^b f(\eta(t))\eta'(t)dt \end{aligned}$$

Thus, we are led to the conclusions in the statement. \square

The main properties of the above integration method are as follows:

PROPOSITION 13.19. *We have the following formula, for a union of paths:*

$$\int_{\gamma \cup \eta} f(x)dx = \int_{\gamma} f(x)dx + \int_{\eta} f(x)dx$$

Also, when reversing the path, the integral changes its sign.

PROOF. Here the first assertion is clear from definitions, and the second assertion comes from the change of variable formula, by using Proposition 13.18. \square

Now by getting back to polynomials, we have the following result:

THEOREM 13.20. *Any polynomial $P \in \mathbb{C}[X]$ satisfies the Cauchy formula*

$$P(x) = \frac{1}{2\pi i} \int_{\gamma} \frac{P(y)}{y - x} dy$$

with the integration over γ being constructed as above.

PROOF. This follows by using abstract arguments and computations similar to those in the proof of Theorem 13.17. Indeed, by linearity we can assume $P(x) = x^n$. Also, by using a cut-and-paste argument, we can assume that we are on a circle:

$$\gamma : [0, 2\pi] \rightarrow \mathbb{C} \quad , \quad \gamma(t) = x + re^{it}$$

By using now the computation from the proof of Theorem 13.17, we obtain:

$$\begin{aligned} \int_{\gamma} \frac{y^n}{y-x} dy &= \int_0^{2\pi} \frac{(x + re^{it})^n}{re^{it}} rie^{it} dt \\ &= i \int_0^{2\pi} (x + re^{it})^n dt \\ &= i \cdot 2\pi x^n \end{aligned}$$

Thus, we are led to the formula in the statement. \square

All this is quite interesting, and obviously, we are now into some serious mathematics. Importantly, Theorem 13.16, Theorem 13.17 and Theorem 13.20 provide us with a path for proving the converse of Theorem 13.15. Indeed, if we manage to prove the Cauchy formula for any holomorphic function $f : X \rightarrow \mathbb{C}$, then it will follow that our function is analytic, and so infinitely differentiable. So, let us start with the following result:

THEOREM 13.21. *The Cauchy formula, namely*

$$f(x) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(y)}{y-x} dy$$

holds for any holomorphic function $f : X \rightarrow \mathbb{C}$.

PROOF. This is something standard, which can be proved as follows:

(1) Our first claim is that given $f \in H(X)$, with $f' \in C(X)$, the integral of f' vanishes on any path. Indeed, by using the change of variable formula, we have:

$$\begin{aligned} \int_{\gamma} f'(x) dx &= \int_a^b f'(\gamma(t)) \gamma'(t) dt \\ &= f(\gamma(b)) - f(\gamma(a)) \\ &= 0 \end{aligned}$$

(2) Our second claim is that given $f \in H(X)$ and a triangle $\Delta \subset X$, we have:

$$\int_{\Delta} f(x) dx = 0$$

Indeed, let us call $\Delta = ABC$ our triangle. Now consider the midpoints A', B', C' of the edges BC, CA, AB , and then consider the following smaller triangles:

$$\Delta_1 = AC'B' \quad , \quad \Delta_2 = BA'C' \quad , \quad \Delta_3 = CB'A' \quad , \quad \Delta_4 = A'B'C'$$

These smaller triangles partition then Δ , and due to our above conventions for the vertex ordering, which produce cancellations when integrating over them, we have:

$$\int_{\Delta} f(x)dx = \sum_{i=1}^4 \int_{\Delta_i} f(x)dx$$

Thus we can pick, among the triangles Δ_i , a triangle $\Delta^{(1)}$ such that:

$$\left| \int_{\Delta} f(x)dx \right| \leq 4 \left| \int_{\Delta^{(1)}} f(x)dx \right|$$

In fact, by repeating the procedure, we obtain triangles $\Delta^{(n)}$ such that:

$$\left| \int_{\Delta} f(x)dx \right| \leq 4^n \left| \int_{\Delta^{(n)}} f(x)dx \right|$$

(3) Now let z be the limiting point of these triangles $\Delta^{(n)}$, and fix $\varepsilon > 0$. By using the fact that the functions $1, x$ integrate over paths up to 0, coming from (1), we obtain the following estimate, with $n \in \mathbb{N}$ being big enough, and L being the perimeter of Δ :

$$\begin{aligned} \left| \int_{\Delta^{(n)}} f(x)dx \right| &= \left| \int_{\Delta^{(n)}} f(x) - f(z) - f'(z)(x - z)dx \right| \\ &\leq \int_{\Delta^{(n)}} |f(x) - f(z) - f'(z)(x - z)| dx \\ &\leq \int_{\Delta^{(n)}} \varepsilon |x - z| dx \\ &\leq \varepsilon (2^{-n}L)^2 \end{aligned}$$

Now by combining this with the estimate in (2), this proves our claim.

(4) The rest is quite routine. First, we can pass from triangles to boundaries of convex sets, in a straightforward way, with the same conclusion as in (2), namely:

$$\int_{\gamma} f(x)dx = 0$$

Getting back to what we want to prove, namely the Cauchy formula for an arbitrary holomorphic function $f \in H(X)$, let $x \in X$, and consider the following function:

$$g(y) = \begin{cases} \frac{f(y)-f(x)}{y-x} & (y \neq x) \\ f'(x) & (y = x) \end{cases}$$

Now assuming that γ encloses a convex set, we can apply what we found, namely vanishing of the integral, to this function g , and we obtain the Cauchy formula for f .

(5) Finally, the extension to general curves is standard, and standard as well is the discussion of what exactly happens at x , in the above proof. See Rudin [74]. \square

As a main application now of the Cauchy formula, we have:

THEOREM 13.22. *The following conditions are equivalent, for a function $f : X \rightarrow \mathbb{C}$:*

- (1) *f is holomorphic.*
- (2) *f is infinitely differentiable.*
- (3) *f is analytic.*
- (4) *The Cauchy formula holds for f .*

PROOF. This is routine from what we have, the idea being as follows:

(1) \implies (4) is non-trivial, but we know this from Theorem 13.21.

(4) \implies (3) is something trivial, because we can expand the series in the Cauchy formula, and we conclude that our function is indeed analytic.

(3) \implies (2) \implies (1) are both elementary, known from Theorem 13.15. \square

As another application of the Cauchy formula, we have:

THEOREM 13.23. *Any holomorphic function $f : X \rightarrow \mathbb{C}$ satisfies the maximum principle, in the sense that given a domain D , with boundary γ , we have:*

$$\exists x \in \gamma \quad , \quad |f(x)| = \max_{y \in D} |f(y)|$$

That is, the maximum of $|f|$ over a domain is attained on its boundary.

PROOF. This follows indeed from the Cauchy formula. Observe that the converse is not true, for instance because functions like \bar{x} satisfy too the maximum principle. We will be back to this later, when talking about harmonic functions. \square

As before with polynomials, a good explanation for the fact that the maximum principle holds could be that the values of our function inside a disk can be recovered from its values on the boundary. And fortunately, this is indeed the case, and we have:

THEOREM 13.24. *Given an holomorphic function $f : X \rightarrow \mathbb{C}$, and a disk D , with boundary γ , the following happen:*

- (1) *f satisfies the plain mean value formula $f(x) = \int_D f(y)dy$.*
- (2) *f satisfies the boundary mean value formula $f(x) = \int_\gamma f(y)dy$.*

PROOF. As usual, this follows from the Cauchy formula, with of course some care in passing from integrals constructed as in Proposition 13.18 to integrals viewed as averages, which are those that we refer to, in the present statement. \square

Finally, as yet another application of the Cauchy formula, which is something nice-looking and conceptual, we have the following statement, called Liouville theorem:

THEOREM 13.25. *An entire, bounded holomorphic function*

$$f : \mathbb{C} \rightarrow \mathbb{C} \quad , \quad |f| \leq M$$

must be constant. In particular, knowing $f \rightarrow 0$ with $z \rightarrow \infty$ gives $f = 0$.

PROOF. This follows as usual from the Cauchy formula, namely:

$$f(x) = \frac{1}{2\pi i} \int_{\gamma} \frac{f(y)}{y - x} dy$$

Alternatively, we can view this as a consequence of Theorem 13.24, because given two points $x \neq y$, we can view the values of f at these points as averages over big disks centered at these points, say $D = D_x(R)$ and $E = D_y(R)$, with $R \gg 0$:

$$f(x) = \int_D f(z) dz \quad , \quad f(y) = \int_E f(z) dz$$

Indeed, the point is that when the radius goes to ∞ , these averages tend to be equal, and so we have $f(x) \simeq f(y)$, which gives $f(x) = f(y)$ in the limit. \square

13d. Harmonic functions

With the Cauchy formula proved, and applied, done with complex analysis, you might say? You must be kidding, that was just the tip of the iceberg, and many things remain to be discussed. In order to reach to a continuation of the above, let us formulate:

DEFINITION 13.26. *The Laplace operator in 2 dimensions is:*

$$\Delta f = \frac{d^2 f}{dx^2} + \frac{d^2 f}{dy^2}$$

A function $f : \mathbb{R}^2 \rightarrow \mathbb{C}$ satisfying $\Delta f = 0$ will be called harmonic.

Here the Laplace operator is something very standard, coming from virtually all branches of physics, and its presence here remains to be explained, yes I know. So, let us try to find the functions $f : \mathbb{R}^2 \rightarrow \mathbb{C}$ which are harmonic. And here, as a good surprise, we have an interesting link with the holomorphic functions:

THEOREM 13.27. *Any holomorphic function $f : \mathbb{C} \rightarrow \mathbb{C}$, when regarded as function*

$$f : \mathbb{R}^2 \rightarrow \mathbb{C}$$

is harmonic. Moreover, the conjugates \bar{f} of holomorphic functions are harmonic too.

PROOF. The first assertion follows from the following computation, for the power functions $f(z) = z^n$, with the usual notation $z = x + iy$:

$$\begin{aligned}\Delta z^n &= \frac{d^2 z^n}{dx^2} + \frac{d^2 z^n}{dy^2} \\ &= \frac{d(nz^{n-1})}{dx} + \frac{d(inz^{n-1})}{dy} \\ &= n(n-1)z^{n-2} - n(n-1)z^{n-2} \\ &= 0\end{aligned}$$

As for the second assertion, this follows from $\Delta \bar{f} = \overline{\Delta f}$, which is clear from definitions, and which shows that if f is harmonic, then so is its conjugate \bar{f} . \square

All this is quite interesting, and the idea in what follows will be that of developing the theory of harmonic functions, as a generalization of the theory that we know for the holomorphic functions, but covering as well functions of type \bar{z} .

As a first goal, in order to understand the harmonic functions, we can try to find the homogeneous polynomials $P \in \mathbb{R}[x, y]$ which are harmonic. In order to do so, the most convenient is to use the variable $z = x + iy$, and think of these polynomials as being homogeneous polynomials $P \in \mathbb{R}[z, \bar{z}]$. With this convention, the result is as follows:

THEOREM 13.28. *The degree n homogeneous polynomials $P \in \mathbb{R}[x, y]$ which are harmonic are precisely the linear combinations of*

$$P = z^n \quad , \quad P = \bar{z}^n$$

with the usual identification $z = x + iy$.

PROOF. As explained above, any homogeneous polynomial $P \in \mathbb{R}[x, y]$ can be regarded as an homogeneous polynomial $P \in \mathbb{R}[z, \bar{z}]$, with the change of variables $z = x + iy$, and in this picture, the degree n homogeneous polynomials are as follows:

$$P(z) = \sum_{k+l=n} c_{kl} z^k \bar{z}^l$$

In order to solve now the Laplace equation $\Delta P = 0$, we must compute the quantities $\Delta(z^k \bar{z}^l)$, for any k, l . But the computation here is routine. We first have the following formula, with the derivatives being computed with respect to the variable x :

$$\begin{aligned}\frac{d(z^k \bar{z}^l)}{dx} &= (z^k)' \bar{z}^l + z^k (\bar{z}^l)' \\ &= k z^{k-1} \bar{z}^l + l z^k \bar{z}^{l-1}\end{aligned}$$

By taking one more time the derivative with respect to x , we obtain:

$$\begin{aligned}
\frac{d^2(z^k \bar{z}^l)}{dx^2} &= k(z^{k-1} \bar{z}^l)' + l(z^k \bar{z}^{l-1})' \\
&= k[(z^{k-1})' \bar{z}^l + z^{k-1}(\bar{z}^l)'] + l[(z^k)' \bar{z}^{l-1} + z^k(\bar{z}^{l-1})'] \\
&= k[(k-1)z^{k-2} \bar{z}^l + lz^{k-1} \bar{z}^{l-1}] + l[kz^{k-1} \bar{z}^{l-1} + (l-1)z^k \bar{z}^{l-2}] \\
&= k(k-1)z^{k-2} \bar{z}^l + 2klz^{k-1} \bar{z}^{l-1} + l(l-1)z^k \bar{z}^{l-2}
\end{aligned}$$

With respect to the variable y , the computations are similar, but some $\pm i$ factors appear, due to $z' = i$ and $\bar{z}' = -i$, coming from $z = x + iy$. We first have:

$$\begin{aligned}
\frac{d(z^k \bar{z}^l)}{dy} &= (z^k)' \bar{z}^l + z^k(\bar{z}^l)' \\
&= ikz^{k-1} \bar{z}^l - ilz^k \bar{z}^{l-1}
\end{aligned}$$

By taking one more time the derivative with respect to y , we obtain:

$$\begin{aligned}
\frac{d^2(z^k \bar{z}^l)}{dy^2} &= ik(z^{k-1} \bar{z}^l)' - il(z^k \bar{z}^{l-1})' \\
&= ik[(z^{k-1})' \bar{z}^l + z^{k-1}(\bar{z}^l)'] - il[(z^k)' \bar{z}^{l-1} + z^k(\bar{z}^{l-1})'] \\
&= ik[i(k-1)z^{k-2} \bar{z}^l - ilz^{k-1} \bar{z}^{l-1}] - il[ikz^{k-1} \bar{z}^{l-1} - i(l-1)z^k \bar{z}^{l-2}] \\
&= -k(k-1)z^{k-2} \bar{z}^l + 2klz^{k-1} \bar{z}^{l-1} - l(l-1)z^k \bar{z}^{l-2}
\end{aligned}$$

We can now sum the formulae that we found, and we obtain:

$$\begin{aligned}
\Delta(z^k \bar{z}^l) &= \frac{d^2(z^k \bar{z}^l)}{dx^2} + \frac{d^2(z^k \bar{z}^l)}{dy^2} \\
&= k(k-1)z^{k-2} \bar{z}^l + 2klz^{k-1} \bar{z}^{l-1} + l(l-1)z^k \bar{z}^{l-2} \\
&\quad - k(k-1)z^{k-2} \bar{z}^l + 2klz^{k-1} \bar{z}^{l-1} - l(l-1)z^k \bar{z}^{l-2} \\
&= 4klz^{k-1} \bar{z}^{l-1}
\end{aligned}$$

In other words, we have reached to the following nice formula:

$$f = z^k \bar{z}^l \implies \Delta f = \frac{4klf}{|z|^2}$$

Now let us get back to our homogeneous polynomial P , written as follows:

$$P(z) = \sum_{k+l=n} c_{kl} z^k \bar{z}^l$$

By using the above formula, it follows that the Laplacian of P is given by:

$$\Delta P(z) = \frac{4}{|z|^2} \sum_{k+l=n} kl c_{kl} z^k \bar{z}^l$$

We conclude that the Laplace equation for P takes the following form:

$$\begin{aligned}\Delta P = 0 &\iff klc_{kl} = 0, \forall k, l \\ &\iff [k, l \neq 0 \implies c_{kl} = 0] \\ &\iff P = c_{n0}z^n + c_{0n}\bar{z}^n\end{aligned}$$

Thus, we are led to the conclusion in the statement. And with the observation that the real formulation of the final result is something quite complicated, and so, for one more time, the use of the complex variable $z = x + iy$ is something very useful. \square

As a conclusion to what we have so far, we know that the holomorphic functions, and so their real and imaginary parts too, are harmonic. That is, if f is holomorphic, then the following function is harmonic, for any values of the parameters $\alpha, \beta \in \mathbb{C}$:

$$f_{\alpha\beta} = \alpha \operatorname{Re}(f) + \beta \operatorname{Im}(f)$$

Observe that this situation covers all the examples that we have so far, for instance with the function \bar{z} , that we know to be harmonic, appearing as follows:

$$\bar{z} = \operatorname{Re}(z) - i\operatorname{Im}(z)$$

And more on this phenomenon, later in this chapter. As a next objective, let us try now to find the harmonic functions which are radial, in the following sense:

$$f(z) = \varphi(|z|)$$

However, things are quite tricky here, involving a blowup phenomenon at the dimension value $N = 2$, which is precisely the one that we are interested in. In view of this phenomenon, it makes sense to move now to arbitrary $N \in \mathbb{N}$ dimensions. So, let us introduce the Laplace operator, acting on the functions $f : \mathbb{R}^N \rightarrow \mathbb{C}$, as follows:

$$\Delta f = \sum_{i=1}^N \frac{d^2 f}{dx_i^2}$$

As before in 2 dimensions, we can say that a function $f : \mathbb{R}^N \rightarrow \mathbb{C}$ is harmonic when $\Delta f = 0$. With these conventions, the result about radial harmonics is as follows:

THEOREM 13.29. *The fundamental radial solutions of $\Delta f = 0$ are*

$$f(x) = \begin{cases} ||x||^{2-N} & (N \neq 2) \\ \log ||x|| & (N = 2) \end{cases}$$

with the log at $N = 2$ basically coming from $\log' = 1/x$.

PROOF. Consider indeed a radial function, defined outside the origin $x = 0$. This function can be written as follows, with $\varphi : (0, \infty) \rightarrow \mathbb{C}$ being a certain function:

$$f : \mathbb{R}^N - \{0\} \rightarrow \mathbb{C} \quad , \quad f(x) = \varphi(||x||)$$

Our first goal will be that of reformulating the Laplace equation $\Delta f = 0$ in terms of the one-variable function $\varphi : (0, \infty) \rightarrow \mathbb{C}$. For this purpose, observe that we have:

$$\begin{aligned} \frac{d||x||}{dx_i} &= \frac{d\sqrt{\sum_{i=1}^N x_i^2}}{dx_i} \\ &= \frac{1}{2} \cdot \frac{1}{\sqrt{\sum_{i=1}^N x_i^2}} \cdot \frac{d\left(\sum_{i=1}^N x_i^2\right)}{dx_i} \\ &= \frac{1}{2} \cdot \frac{1}{||x||} \cdot 2x_i \\ &= \frac{x_i}{||x||} \end{aligned}$$

By using this formula, we have the following computation:

$$\begin{aligned} \frac{df}{dx_i} &= \frac{d\varphi(||x||)}{dx_i} \\ &= \varphi'(||x||) \cdot \frac{d||x||}{dx_i} \\ &= \varphi'(||x||) \cdot \frac{x_i}{||x||} \end{aligned}$$

By differentiating one more time, we obtain the following formula:

$$\begin{aligned} \frac{d^2 f}{dx_i^2} &= \frac{d}{dx_i} \left(\varphi'(||x||) \cdot \frac{x_i}{||x||} \right) \\ &= \frac{d\varphi'(||x||)}{dx_i} \cdot \frac{x_i}{||x||} + \varphi'(||x||) \cdot \frac{d}{dx_i} \left(\frac{x_i}{||x||} \right) \\ &= \left(\varphi''(||x||) \cdot \frac{x_i}{||x||} \right) \cdot \frac{x_i}{||x||} + \varphi'(||x||) \cdot \frac{||x|| - x_i \cdot x_i/||x||}{||x||^2} \\ &= \varphi''(||x||) \cdot \frac{x_i^2}{||x||^2} + \varphi'(||x||) \cdot \frac{||x||^2 - x_i^2}{||x||^3} \end{aligned}$$

Now by summing over $i \in \{1, \dots, N\}$, this gives the following formula:

$$\begin{aligned} \Delta f &= \sum_{i=1}^N \varphi''(||x||) \cdot \frac{x_i^2}{||x||^2} + \sum_{i=1}^N \varphi'(||x||) \cdot \frac{||x||^2 - x_i^2}{||x||^3} \\ &= \varphi''(||x||) \cdot \frac{||x||^2}{||x||^2} + \varphi'(||x||) \cdot \frac{(N-1)||x||^2}{||x||^3} \\ &= \varphi''(||x||) + \varphi'(||x||) \cdot \frac{N-1}{||x||} \end{aligned}$$

Thus, with $r = ||x||$, the Laplace equation $\Delta f = 0$ can be reformulated as follows:

$$\varphi''(r) + \frac{(N-1)\varphi'(r)}{r} = 0$$

Equivalently, the equation that we want to solve is as follows:

$$r\varphi'' + (N-1)\varphi' = 0$$

Now observe that we have the following formula:

$$\begin{aligned} (r^{N-1}\varphi')' &= (N-1)r^{N-2}\varphi' + r^{N-1}\varphi'' \\ &= r^{N-2}((N-1)\varphi' + r\varphi'') \end{aligned}$$

Thus, the equation to be solved can be simply written as follows:

$$(r^{N-1}\varphi')' = 0$$

We conclude that $r^{N-1}\varphi'$ must be a constant K , and so, that we must have:

$$\varphi' = Kr^{1-N}$$

But the fundamental solutions of this latter equation are as follows:

$$\varphi(r) = \begin{cases} r^{2-N} & (N \neq 2) \\ \log r & (N = 2) \end{cases}$$

Thus, we are led to the conclusion in the statement. \square

Back now to the general theory of harmonic functions, the above passage to arbitrary $N \in \mathbb{N}$ dimensions, that we will adopt, proves to be something fruitful, allowing us to see many things obscured by various $N = 2$ phenomena. Among others, we have:

THEOREM 13.30. *The harmonic functions in N dimensions obey to the same general principles as the holomorphic functions, namely:*

- (1) *The plain mean value formula.*
- (2) *The boundary mean value formula.*
- (3) *The maximum modulus principle.*
- (4) *The Liouville theorem.*

PROOF. This is something quite straightforward, the idea being as follows:

(1) Regarding the plain mean value formula, here the statement is that given an harmonic function $f : X \rightarrow \mathbb{C}$, and a ball B , the following happens:

$$f(x) = \int_B f(y) dy$$

In order to prove this formula, we can assume that our ball B is centered at 0, say of radius $r > 0$. Now if we denote by χ_r the characteristic function of this ball, normalized as to integrate up to 1, we want to prove that we have the following formula:

$$f = f * \chi_r$$

To be more precise, here $*$ is the standard convolution operation, given by:

$$(f * g)(x) = \int_{\mathbb{R}^N} f(x - y)g(y)dy$$

For proving the above formula, pick $0 < s < r$, and a solution w of the following equation, on B , which can be constructed explicitly, say as a radial function:

$$\Delta w = \chi_r - \chi_s$$

By using the basic properties of the convolution operation $*$, we have:

$$\begin{aligned} f * \chi_r - f * \chi_s &= f * (\chi_r - \chi_s) \\ &= f * \Delta w \\ &= \Delta f * w \\ &= 0 \end{aligned}$$

Thus $f * \chi_r = f * \chi_s$, and by letting now $s \rightarrow 0$, we get $f * \chi_r = f$, as desired.

(2) Regarding the boundary mean value formula, here the statement is that given an harmonic function $f : X \rightarrow \mathbb{C}$, and a ball B , with boundary γ , the following happens:

$$f(x) = \int_{\gamma} f(y)dy$$

But this follows as a consequence of the plain mean value formula in (1), with our two mean value formulae, the one there and the one here, being in fact equivalent, by using annuli and radial integration for the proof of the equivalence, in the obvious way.

(3) Regarding the maximum modulus principle, the statement here is that any holomorphic function $f : X \rightarrow \mathbb{C}$ has the property that the maximum of $|f|$ over a domain is attained on its boundary. That is, given a domain D , with boundary γ , we have:

$$\exists x \in \gamma \quad , \quad |f(x)| = \max_{y \in D} |f(y)|$$

But this is something which follows again from the mean value formula in (1), first for the balls, and then in general, by using a standard division argument.

(4) Finally, regarding the Liouville theorem, the statement here is that an entire, bounded harmonic function must be constant:

$$f : \mathbb{R}^N \rightarrow \mathbb{C} \quad , \quad \Delta f = 0 \quad , \quad |f| \leq M \quad \implies \quad f = \text{constant}$$

As a slightly weaker statement, again called Liouville theorem, we have the fact that an entire harmonic function which vanishes at ∞ must vanish globally:

$$f : \mathbb{R}^N \rightarrow \mathbb{C} \quad , \quad \Delta f = 0 \quad , \quad \lim_{x \rightarrow \infty} f(x) = 0 \quad \implies \quad f = 0$$

But can view these as a consequence of the mean value formula in (1), because given two points $x \neq y$, we can view the values of f at these points as averages over big balls centered at these points, say $B = B_x(R)$ and $C = B_y(R)$, with $R \gg 0$:

$$f(x) = \int_B f(z) dz \quad , \quad f(y) = \int_C f(z) dz$$

Indeed, the point is that when the radius goes to ∞ , these averages tend to be equal, and so we have $f(x) \simeq f(y)$, which gives $f(x) = f(y)$ in the limit, as desired. \square

Getting back now to 2D, as a useful complement to Theorem 13.30, we have:

THEOREM 13.31. *The real harmonic functions $f : X \rightarrow \mathbb{R}$ with $X \subset \mathbb{C}$ are locally the real parts of the holomorphic functions $g : X \rightarrow \mathbb{C}$.*

PROOF. The precise claim here is that a real harmonic function on the unit disk $f : D \rightarrow \mathbb{R}$ must equal the real part of the following function, which is holomorphic:

$$g(z) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{it} + z}{e^{it} - z} f(e^{it}) dt$$

Indeed, observe for instance that this holds for $f(z) = \operatorname{Re}(z^n)$. However, the proof in general is non-trivial, and we refer to Rudin [74] for details on this. \square

13e. Exercises

This was a quite standard calculus chapter, and as exercises on this, we have:

EXERCISE 13.32. *Experiment some more with the geometric series, over \mathbb{C} .*

EXERCISE 13.33. *Learn more about the logarithm, as a complex function.*

EXERCISE 13.34. *Learn as well about a^x and x^a , as complex functions.*

EXERCISE 13.35. *Learn more about hyperbolic functions, and the need for them.*

EXERCISE 13.36. *Clarify the relation between the various notions of differentiability.*

EXERCISE 13.37. *Fill in the various details, in the proof of the Cauchy formula.*

EXERCISE 13.38. *Learn more about harmonic functions, and their properties.*

EXERCISE 13.39. *Read the proof of the last theorem, $f = \operatorname{Re}(g)$, locally.*

As bonus exercise, read some quantum mechanics, the scalars there being \mathbb{C} too.

CHAPTER 14

Zeta function

14a. Zeta function

We already met the Riemann zeta function $\zeta(s) = \sum_n 1/n^s$ in chapter 5, at values $s > 1$ of the parameter, with its convergence there being one of the basic results in real analysis, and then later in chapter 7, still at values $s > 1$ of the parameter, with the conclusion there that this function is intimately related to the primes.

In what follows we discuss a systematic approach to this phenomenon, by using complex analysis. As a first observation, we can talk without much pain about zeta at complex values of s as well, under the assumption $\operatorname{Re}(s) > 1$, in the following way:

THEOREM 14.1. *We can talk about the Riemann zeta function*

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

at any $s \in \mathbb{C}$ with $\operatorname{Re}(s) > 1$.

PROOF. We have indeed the following computation, with $s = r + it$:

$$|\zeta(s)| = \left| \sum_{n=1}^{\infty} \frac{1}{n^s} \right| \leq \sum_{n=1}^{\infty} \frac{1}{|n^s|} = \sum_{n=1}^{\infty} \frac{1}{n^r}$$

But the sum on the right converges for $r > 1$, as we know well since chapter 5, with an explicit bound, obtained via calculus, being as follows:

$$\sum_{n=1}^{\infty} \frac{1}{n^r} < 1 + \int_1^{\infty} \frac{1}{x^r} dx = 1 + \left[\frac{x^{1-r}}{1-r} \right]_1^{\infty} = 1 + \frac{1}{r-1}$$

Thus, we are led to the conclusion in the statement. □

As a first result now, we can write zeta as an Euler product, as follows:

THEOREM 14.2. *We have the following formula, with product over the primes,*

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s} \right)^{-1}$$

valid for any exponent $s \in \mathbb{C}$ with $\operatorname{Re}(s) > 1$.

PROOF. We have indeed the following computation, coming as usual from the unique factorization of the integers, $n = p_1^{a_1} \dots p_k^{a_k}$, and with everything converging:

$$\begin{aligned}\zeta(s) &= \sum_{n=1}^{\infty} \frac{1}{n^s} \\ &= \prod_p \left(1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \frac{1}{p^{3s}} + \dots \right) \\ &= \prod_p \left(1 - \frac{1}{p^s} \right)^{-1}\end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

We refer to chapter 7 for more on all this, including some elementary estimates for zeta, inspired by the standard estimates in relation with the Euler formula, at $s = 1$.

Some further useful formulae for zeta come in terms of the Möbius function. Let us first recall that the Möbius function is defined, in general, as follows:

DEFINITION 14.3. *The Möbius function of a lattice L is given by*

$$\mu(a, b) = \begin{cases} 1 & \text{if } a = b \\ -\sum_{a \leq c < b} \mu(a, c) & \text{if } a < b \\ 0 & \text{if } a \not\leq b \end{cases}$$

with this construction being performed by recurrence.

As an illustration for this, consider the lattice of positive integers $L = \mathbb{N}$, with the order $a \leq b$ when $a|b$. Let us try to compute the following function:

$$\mu(n) = \mu(1, n)$$

For this function, the recurrence relation from Definition 14.3 reads:

$$\mu(n) = \begin{cases} 1 & \text{if } n = 1 \\ -\sum_{m|n, m < n} \mu(m) & \text{if } n > 1 \end{cases}$$

We have the following computations, to start with, for any prime number p :

$$\begin{aligned}\mu(1) &= 1 \\ \mu(p) &= -\mu(1) = -1 \\ \mu(p^2) &= -\mu(1) - \mu(p) = -1 + 1 = 0 \\ \mu(p^3) &= -\mu(1) - \mu(p) - \mu(p^2) = -1 + 1 - 0 = 0 \\ \mu(p^4) &= -\mu(1) - \mu(p) - \mu(p^2) - \mu(p^3) = -1 + 1 - 0 - 0 = 0 \\ &\vdots\end{aligned}$$

Thus, by recurrence, for any prime p , we have the following formula:

$$\mu(p^k) = \begin{cases} 1 & \text{if } k = 0 \\ -1 & \text{if } k = 1 \\ 0 & \text{if } k \geq 2 \end{cases}$$

Which does not look very good, but let us do some more computations. When coming with a second prime number q , different from p , we have the following computation:

$$\mu(pq) = -\mu(1) - \mu(p) - \mu(q) = -1 + 1 + 1 = 1$$

Which looks quite nice, at least we don't get a 0. So let us add now to the picture a third prime r , distinct from p, q . We have the following computation:

$$\begin{aligned} \mu(pqr) &= -\mu(1) - \mu(p) - \mu(q) - \mu(r) - \mu(pq) - \mu(pr) - \mu(qr) \\ &= -1 + 1 + 1 + 1 - 1 - 1 - 1 \\ &= -1 \end{aligned}$$

When adding a fourth prime number s to be picture, distinct from p, q, r , we have:

$$\begin{aligned} \mu(pqrs) &= -\mu(1) - \mu(p) - \mu(q) - \mu(r) - \mu(s) \\ &\quad - \mu(pq) - \mu(pr) - \mu(ps) - \mu(qr) - \mu(qs) - \mu(rs) \\ &\quad - \mu(pqr) - \mu(pqs) - \mu(prs) - \mu(qrs) \\ &= -1 + 4 - 6 + 4 \\ &= 1 \end{aligned}$$

In general now, by recurrence, for any p_1, \dots, p_k distinct primes, we have:

$$\mu(1, p_1 \dots p_k) = (-1)^k$$

Which looks good, and the next question is, what is the formula unifying what we have, for $n = p^k$, and $n = p_1 \dots p_k$? And here the answer is quite obvious, as follows:

THEOREM 14.4. *The Möbius function $\mu(n) = \mu(1, n)$ is given by*

$$\mu(n) = \begin{cases} (-1)^k & \text{if } n = p_1 \dots p_k \\ 0 & \text{if } n \text{ is not square-free} \end{cases}$$

and its two-variable version is given by $\mu(a, b) = \delta_{a|b} \mu(b/a)$.

PROOF. We already have some good evidence for the first formula in the statement, but let us do some more computations. For p, q distinct primes, we have:

$$\begin{aligned} \mu(pq^2) &= -\mu(1) - \mu(p) - \mu(pq) - \mu(q) \\ &= -1 + 1 - 1 + 1 \\ &= 0 \end{aligned}$$

Similarly, we have as well the following computation:

$$\begin{aligned}\mu(pq^3) &= -\mu(1) - \mu(p) - \mu(pq) - \mu(pq^2) - \mu(q) - \mu(q^2) \\ &= -1 + 1 - 1 - 0 + 1 - 0 \\ &= 0\end{aligned}$$

Which makes it pretty clear that we have $\mu(pq^k) = 0$ for any $k \geq 2$, by recurrence. Next, when adding a third prime r to the picture, distinct from p, q , we have:

$$\begin{aligned}\mu(pqr^2) &= -\mu(1) - \mu(p) - \mu(pq) - \mu(pr) - \mu(pqr) - \mu(pr^2) \\ &\quad - \mu(q) - \mu(qr) - \mu(qr^2) - \mu(r) - \mu(r^2) \\ &= -1 + 1 - 1 - 1 + 1 - 0 \\ &\quad + 1 - 1 - 0 + 1 - 0 \\ &= 0\end{aligned}$$

And so on, we certainly have very solid evidence now for our formula, and we will leave its proof in general, by recurrence, as an instructive exercise. As for the second formula in the statement, regarding the general function $\mu(a, b)$, this follows from this. \square

Back to theory, the main interest in the Möbius function comes from the Möbius inversion formula, which in linear algebra terms can be stated and proved as follows:

THEOREM 14.5. *We have the following implication,*

$$f(b) = \sum_{a \leq b} g(a) \implies g(b) = \sum_{a \leq b} \mu(a, b) f(a)$$

valid for any two functions $f, g : L \rightarrow \mathbb{C}$.

PROOF. Consider the adjacency matrix of L , given by the following formula:

$$K_{ab} = \begin{cases} 1 & \text{if } a \leq b \\ 0 & \text{if } a \not\leq b \end{cases}$$

Our claim is that the inverse of this matrix is the Möbius matrix of L , given by:

$$M_{ab} = \mu(a, b)$$

Indeed, the above matrix K is upper triangular, and when trying to invert it, we are led to the recurrence in Definition 14.3, so to the Möbius matrix M . Thus we have:

$$M = K^{-1}$$

Thus, in practice, we are led to the inversion formula in the statement. \square

Getting back now to the Möbius function for the integers, from Theorem 14.4, many interesting things can be said about it. First, we have the following formula:

$$\sum_{d|n} \mu(d) = \begin{cases} 1 & \text{if } n = 1 \\ 0 & \text{if } n \geq 2 \end{cases}$$

Next, we have the following formula, which is something elementary too:

$$\sum_{n=1}^{\infty} \frac{\mu(n)}{n} = 0$$

We have as well the following formula, which is elementary too:

$$\sum_{n=1}^{\infty} \frac{\mu(n) \log n}{n} = -1$$

Next, we have the following formula, with $\gamma = 0.57721..$ being the Euler-Mascheroni constant, that we met in chapter 7, when discussing the Mertens theorems:

$$\sum_{n=1}^{\infty} \frac{\mu(n) \log^2 n}{n} = -2\gamma$$

Finally, we have the following related formula, valid for any $|q| < 1$:

$$\sum_{n=1}^{\infty} \frac{\mu(n) q^n}{1 - q^n} = q$$

We will leave the proofs of the above formulae, which can be useful for various purposes, as an exercise. Getting back now to the zeta function, we have:

THEOREM 14.6. *We have the following formula,*

$$\frac{1}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\mu(n)}{n^s}$$

with μ being the Möbius function, given by the formula

$$\mu(n) = \begin{cases} (-1)^k & \text{if } n = p_1 \dots p_k \\ 0 & \text{if } n \text{ is not square-free} \end{cases}$$

valid for any exponent $s \in \mathbb{C}$ with $\operatorname{Re}(s) > 1$.

PROOF. We have the following computation, with everything converging:

$$\begin{aligned}\frac{1}{\zeta(s)} &= \prod_p \left(1 - \frac{1}{p^s}\right) \\ &= \sum_{k=0}^{\infty} (-1)^k \prod_{p_1 \dots p_k} \frac{1}{p_1^s \dots p_k^s} \\ &= \sum_{n=1}^{\infty} \frac{\mu(n)}{n^s}\end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

Along the same lines, as another elementary result about zeta, we have:

THEOREM 14.7. *The square of the zeta function is given by*

$$\zeta^2(s) = \sum_{n=1}^{\infty} \frac{\tau(n)}{n^s}$$

with $\tau(n)$ being the number of divisors of n , for any $s \in \mathbb{C}$ with $\operatorname{Re}(s) > 1$.

PROOF. We have the following computation, with everything converging:

$$\zeta(s)^2 = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \frac{1}{(kl)^s} = \sum_{n=1}^{\infty} \frac{\tau(n)}{n^s}$$

Thus, we are led to the conclusion in the statement. \square

There are some further formulae of the same type, again involving the zeta function, and the above function $\tau(n)$, counting the number of divisors of n . First, we have:

$$\frac{\zeta^3(s)}{\zeta(2s)} = \sum_{n=1}^{\infty} \frac{\tau(n^2)}{n^s}$$

Along the same lines, we have as well the following formula:

$$\frac{\zeta^4(s)}{\zeta(2s)} = \sum_{n=1}^{\infty} \frac{\tau(n)^2}{n^s}$$

Many other things can be said, along these lines, and we will leave some further study here, including the proofs of the above two formulae, as an exercise.

14b. Basel formula

At a more advanced level now, we can try to compute particular values of ζ . Things are quite tricky here, and we first have the following remarkable result:

THEOREM 14.8. *We have the following formula of Euler,*

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \dots = \frac{\pi^2}{6}$$

which reads $\zeta(2) = \pi^2/6$, solving the Basel problem.

PROOF. Many things can be said here, the idea being as follows:

(1) The original proof by Euler was as follows, making some clever manipulations on the Taylor series expansion of the function $\sin x/x$, based on the observation that the zeroes of this function appear precisely at the points $x = k\pi$, with $k \in \mathbb{Z}$:

$$\begin{aligned} \frac{\sin x}{x} &= 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \dots \\ &= \left(1 - \frac{x}{\pi}\right) \left(1 + \frac{x}{\pi}\right) \left(1 - \frac{x}{2\pi}\right) \left(1 + \frac{x}{2\pi}\right) \dots \\ &= \left(1 - \frac{x^2}{\pi^2}\right) \left(1 - \frac{x^2}{4\pi^2}\right) \left(1 - \frac{x^2}{9\pi^2}\right) \dots \\ &= 1 - \frac{1}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} x^2 + \dots \end{aligned}$$

(2) Which is obviously nice and remarkable, and accurate too, but which is far from being rigorous. Euler himself struggled for some time, with finding the fix, and many other mathematicians at that time produced afterwards their own proofs too.

(3) From a modern perspective, the fix for the above proof comes from a result of Weierstrass, which shows that the above factorization of $\sin x/x$ holds indeed, as a general property of the analytic functions. However, this is something quite advanced. \square

In what follows, our plan will be to understand the Basel formula via Fourier analysis, and the Parseval formula there. Let us start with some generalities:

DEFINITION 14.9. *A Hilbert space is a complex vector space H , with a map*

$$\langle, \rangle: H \times H \rightarrow \mathbb{C}$$

called scalar product, subject to the following conditions:

- (1) $\langle x, y \rangle$ is linear in x , and antilinear in y .
- (2) $\overline{\langle x, y \rangle} = \langle y, x \rangle$, for any x, y .
- (3) $\langle x, x \rangle \geq 0$, for any $x \neq 0$.
- (4) H is complete with respect to the norm $\|x\| = \sqrt{\langle x, x \rangle}$.

We will see examples of such spaces in a moment, but before that, some explanations in regards with the condition (4). Assuming that (1,2,3) are satisfied, given two vectors $x, y \in H$, consider the following degree 2 function of $t \in \mathbb{R}$, depending on $w \in \mathbb{T}$:

$$f(t) = ||wx + ty||^2 = ||x||^2 + 2t\operatorname{Re}(w \langle x, y \rangle) + t^2||y||^2$$

This function being positive, its discriminant must be negative, which gives:

$$|\langle x, y \rangle| \leq ||x|| \cdot ||y||$$

But this gives in turn $||x + y|| \leq ||x|| + ||y||$, and so we have indeed a norm, as said in (4), so we can talk about the completeness of H with respect to this norm.

At the level of examples of Hilbert spaces, we have the following result:

THEOREM 14.10. *Given a measured space X , the following space, with the convention that functions are identified up to equality almost everywhere,*

$$L^2(X) = \left\{ f : X \rightarrow \mathbb{C} \mid \int_X |f(x)|^2 dx < \infty \right\}$$

is a vector space, and the following is a scalar product on it, making it a Hilbert space:

$$\langle f, g \rangle = \int_X f(x) \overline{g(x)} dx$$

In the discrete case, where X is set, with its counting measure, we obtain the sequence spaces $l^2(X)$, which generalize \mathbb{C}^N with its usual scalar product, $\langle u, v \rangle = \sum_i u_i \bar{v}_i$.

PROOF. This is indeed something self-explanatory, with everything being elementary when X is a set, and with the generalization to the case of measured spaces being quite straightforward. We will leave clarifying all this as a long, instructive exercise. \square

Getting now to the real thing, geometry in the Hilbert space setting, which can be something very useful in the functional analytic context, we first have:

PROPOSITION 14.11. *Given a Hilbert space H , the following happen:*

- (1) *Norm formula:* $||x|| = \sqrt{\langle x, x \rangle}$.
- (2) *Cauchy-Schwarz:* $|\langle x, y \rangle| \leq ||x|| \cdot ||y||$.
- (3) *Parallelogram identity:* $||x + y||^2 + ||x - y||^2 = 2(||x||^2 + ||y||^2)$.
- (4) *Polarization:* $4\langle x, y \rangle = ||x + y||^2 - ||x - y||^2 + i||x + iy||^2 - i||x - iy||^2$.

PROOF. Here (1) was part of the axioms, (2) is something that we already know, which was explained above, and (3) comes from the following computation:

$$\begin{aligned} & ||x + y||^2 + ||x - y||^2 \\ &= \langle x + y, x + y \rangle + \langle x - y, x - y \rangle \\ &= ||x||^2 + ||y||^2 + \langle x, y \rangle + \langle y, x \rangle + ||x||^2 + ||y||^2 - \langle x, y \rangle - \langle y, x \rangle \\ &= 2(||x||^2 + ||y||^2) \end{aligned}$$

As for (4), this comes from a similar computation, as follows:

$$\begin{aligned}
& \|x + y\|^2 - \|x - y\|^2 + i\|x + iy\|^2 - i\|x - iy\|^2 \\
= & \|x\|^2 + \|y\|^2 - \|x\|^2 - \|y\|^2 + i\|x\|^2 + i\|y\|^2 - i\|x\|^2 - i\|y\|^2 \\
& + 2\operatorname{Re}\langle x, y \rangle + 2\operatorname{Re}\langle x, y \rangle + 2i\operatorname{Im}\langle x, y \rangle + 2i\operatorname{Im}\langle x, y \rangle \\
= & 4\langle x, y \rangle
\end{aligned}$$

Finally, observe that (1,4) show that the norm is uniquely determined by the scalar product, and vice versa. This is something very useful, and good to know. \square

As a first piece of true geometry now, in the Hilbert space setting, we have:

THEOREM 14.12. *Given a Hilbert space H , we can talk about:*

- (1) *Bounded operators $T : H \rightarrow H$, for which $\|T\| = \sup_{\|x\|=1} \|Tx\|$ is bounded.*
- (2) *Adjoint of such bounded operators, given by $\langle T^*x, y \rangle = \langle x, Ty \rangle$.*
- (3) *Unitaries $U : H \rightarrow H$, which must satisfy the condition $U^* = U^{-1}$.*
- (4) *Projections $P : H \rightarrow H$, which must satisfy the condition $P^2 = P^* = P$.*

PROOF. This is obviously something quite advanced, generalizing some key facts from linear algebra, with the corresponding matrices being now possibly infinite. So, as before with other things, take this as something which is good to know, making the point with geometry, and whose details you can get into later, when really needing such things. \square

Moving on, as a key feature of the Hilbert spaces, we have bases, and we will discuss this in some detail. At the general level, we first have the following result:

THEOREM 14.13. *Any Hilbert space H has an orthonormal basis $\{e_i\}_{i \in I}$, which is by definition a set of vectors whose span is dense in H , and which satisfy*

$$\langle e_i, e_j \rangle = \delta_{ij}$$

with δ being a Kronecker symbol. The cardinality $|I|$ of the index set, which can be finite, countable, or uncountable, depends only on H , and is called dimension of H . We have

$$H \simeq l^2(I)$$

in the obvious way, mapping $\sum \lambda_i e_i \rightarrow (\lambda_i)$. The Hilbert spaces with $\dim H = |I|$ being countable, such as $l^2(\mathbb{N})$, are all isomorphic, and are called separable.

PROOF. We have many assertions here, the idea being as follows:

(1) In finite dimensions an orthonormal basis $\{e_i\}_{i \in I}$ can be constructed by starting with any vector space basis $\{f_i\}_{i \in I}$, and using the well-known Gram-Schmidt procedure. As for the other assertions, these are all clear, from basic linear algebra.

(2) In general, the same method works, namely Gram-Schmidt, with a subtlety coming from the fact that the basis $\{e_i\}_{i \in I}$ will not span in general the whole H , but just a dense subspace of it, as it is in fact obvious by looking at the standard basis of $l^2(\mathbb{N})$.

(3) And there is a second subtlety as well, coming from the fact that the recurrence procedure needed for Gram-Schmidt must be replaced by some sort of “transfinite recurrence”, using standard tools from logic, and more specifically the Zorn lemma.

(4) Finally, everything at the end, regarding our notion of separability for the Hilbert spaces, is clear from definitions, and from our various results above. \square

As a continuation of this, we have the following result, dealing with separability:

THEOREM 14.14. *The following happen, in relation with separability:*

- (1) *The Hilbert space $H = L^2[-1, 1]$ is separable, with orthonormal basis coming by applying Gram-Schmidt to the basis $\{x^k\}_{k \in \mathbb{N}}$, coming from Weierstrass.*
- (2) *In fact, $H = L^2(X)$ with $X \subset \mathbb{R}$ and $d\mu(x) = f(x)dx$ is separable, with standard basis $\{P_k\}_{k \in \mathbb{N}}$ formed by the orthogonal polynomials with respect to μ .*
- (3) *More generally, given a separable abstract measured space X , the associated Hilbert space of square-summable functions $H = L^2(X)$ is separable.*

PROOF. Many things can be said here, the idea being as follows:

(1) The fact that $H = L^2[-1, 1]$ is separable is clear indeed from the Weierstrass density theorem, which provides us with the algebraic basis $f_k = x^k$, which can be orthogonalized by using the Gram-Schmidt procedure, as explained in Theorem 14.13.

(2) This is a straightforward generalization of (1), with the polynomials $\{P_k\}_{k \in \mathbb{N}}$ coming from the Weierstrass basis $\{x^k\}_{k \in \mathbb{N}}$, via Gram-Schmidt with respect to the measure $d\mu(x) = f(x)dx$, being called the orthogonal polynomials with respect to μ .

(3) As for the last assertion, regarding the spaces of type $H = L^2(X)$, which generalizes what we have in (1,2), this comes as a consequence of general measure theory. As before with other such things, we will leave some learning here as an exercise. \square

Getting now to more concrete things, which will eventually lead to applications, we have the following key fact, which is the starting point for Fourier analysis:

CLAIM 14.15. *The space of square-summable functions on the unit circle,*

$$L^2(\mathbb{T}) = \left\{ f : \mathbb{T} \rightarrow \mathbb{C} \mid \int_{\mathbb{T}} |f(z)|^2 dz < \infty \right\}$$

with respect to the usual mass 1 measure, has $\{z^n\}_{n \in \mathbb{Z}}$ as orthonormal basis.

As a first observation, this reminds a bit what we said above regarding the Weierstrass basis $\{x^n\}_{n \in \mathbb{N}}$ for the various spaces of functions $f : X \rightarrow \mathbb{C}$, with $X \subset \mathbb{R}$, and the related notion of orthogonal polynomials. That is, what Claim 14.15 says is that we have some sort of Weierstrass approximation theorem on the circle, but with respect to the 2-norm, and with the standard basis which is used, $\{z^n\}_{n \in \mathbb{Z}}$, being orthonormal for free.

In practice now, Claim 14.15 remains something which, while being certainly simple, beautiful and understandable, is a bit compact and abstract. For most purposes, it is better to replace it by the following equivalent formulation, in terms of real functions:

THEOREM 14.16. *The space of 2π -periodic square-summable functions on \mathbb{R} ,*

$$L^2(\mathbb{R})_{\text{per}} = \left\{ f : \mathbb{R} \rightarrow \mathbb{C} \mid f(t) = f(t + 2\pi), \int_{-\pi}^{\pi} |f(t)|^2 dt < \infty \right\}$$

has $\{e^{int}\}_{n \in \mathbb{Z}}$ as orthonormal basis, with respect to the normalized mass 1 measure.

PROOF. This is something quite tricky, which came as a big surprise at the time of its discovery by Fourier, the idea with all this being as follows:

(1) As a first observation, which is philosophical, as already mentioned above, this is a real function reformulation of Claim 14.15, and for various reasons, in relation with both the proof and the future applications, this is the version that we will prefer.

(2) Next, still talking generalities, according to what the statement says at the end, the scalar product on $L^2(\mathbb{R})_{\text{per}}$ is by definition given by the following formula:

$$\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \overline{g(t)} dt$$

As for the corresponding norm on $L^2(\mathbb{R})_{\text{per}}$, this is given by the following formula:

$$\|f\| = \frac{1}{2\pi} \int_{-\pi}^{\pi} |f(t)|^2 dt$$

Observe that we could have used $[0, 2\pi]$ for integrating, or more generally, any interval $I \subset \mathbb{R}$ having length 2π . For certain technical reasons, we prefer to use $I = [-\pi, \pi]$.

(3) Getting now to the proof, as a first basic computation that we can do, coming from the 2π -periodicity of $e^{it} = \cos t + i \sin t$, we have the following formula:

$$\langle e^{int}, e^{imt} \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i(n-m)t} dt = \delta_{nm}$$

Thus $\{e^{int}\}_{n \in \mathbb{Z}}$ is as orthonormal system, and our theorem says that this is a basis.

(4) But, how to prove this? The first thought goes to the Weierstrass approximation theorem, mentioned before, but after thinking a bit, that is not of much use. Indeed, Weierstrass approximates the functions $f : [-\pi, \pi] \rightarrow \mathbb{C}$ by polynomials $\sum_n c_n t^n$, and what Fourier says is that the same functions $f : [-\pi, \pi] \rightarrow \mathbb{C}$ can be approximated by trigonometric polynomials $\sum_n c_n e^{int}$, which is a different story.

(5) In fact, the difference between Fourier and Weierstrass is even more visible with Theorem 14.14 in mind. Indeed, up to a rescaling, Theorem 14.14 tells us that what we get from Weierstrass is the basis of $L^2(\mathbb{R})_{\text{per}}$ formed by Legendre polynomials $\{L_k\}_{k \in \mathbb{N}}$, and the Fourier basis $\{e^{int}\}_{n \in \mathbb{Z}}$ is obviously something of different nature.

(6) Time now for the proof? For reasons that will become clear in a moment, consider the following trigonometric polynomials, with $c_k \in \mathbb{R}$ being chosen for having mass 1:

$$E_k(t) = c_k \left(\frac{1 + \cos t}{2} \right)^k$$

Observe that c_k can be computed explicitly, by using $1 + \cos t = 2 \cos^2(t/2)$ and the Wallis formulae from chapter 7, but in what follows, we will not need this. Our claim, which is the reason for introducing these functions E_k , is that for any $\delta > 0$ we have:

$$\lim_{k \rightarrow \infty} \sup_{\delta < |t| < \pi} E_k(t) = 0$$

In other words, our claim is that $E_k \rightarrow 0$ uniformly on any $[-\pi, -\delta] \cup [\delta, \pi]$.

(7) So, let us prove this claim. As mentioned, c_k can be computed explicitly, but in what follows we will only need the following elementary estimate:

$$\begin{aligned} 1 &= \frac{c_k}{\pi} \int_0^\pi \left(\frac{1 + \cos t}{2} \right)^k dt \\ &> \frac{c_k}{\pi} \int_0^\pi \left(\frac{1 + \cos t}{2} \right)^k \sin t dt \\ &= \frac{c_k}{\pi} \left[-\frac{2}{k+1} \left(\frac{1 + \cos t}{2} \right)^{k+1} \right]_0^\pi \\ &= \frac{2c_k}{\pi(k+1)} \end{aligned}$$

Now since E_k is decreasing on $[0, \pi]$, we obtain from this, for $\delta < |t| < \pi$:

$$E_k(t) < E_k(\delta) < \frac{\pi(k+1)}{2} \left(\frac{1 + \cos \delta}{2} \right)^k$$

But this proves our claim, because for $\delta > 0$ we have $(1 + \cos \delta)/2 < 1$, as needed.

(8) Getting now to what we wanted to do, we must prove that $\{e^{int}\}_{n \in \mathbb{Z}}$ spans a dense subset of $L^2(\mathbb{R})_{per}$. Since $C(\mathbb{R})_{per} \subset L^2(\mathbb{R})_{per}$ is dense, it is enough to prove that any $f \in C(\mathbb{R})_{per}$ can be approximated by trigonometric polynomials $\sum_n c_n e^{int}$. Moreover, since $\|\cdot\|_2 \leq \|\cdot\|_\infty$, it is enough to prove our approximation with respect to $\|\cdot\|_\infty$.

(9) All in all, it remains to prove that given a function $f \in C(\mathbb{R})_{per}$ and a number $\varepsilon > 0$, we can always come with a trigonometric polynomial $\sum_n c_n e^{int}$, such that:

$$\left| f(t) - \sum_n c_n e^{int} \right| < \varepsilon, \quad \forall t \in [-\pi, \pi]$$

(10) But for this, we can use the polynomials E_k from (6). Let us set indeed:

$$Q_k(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t-s)E_k(s) ds$$

As a first observation, with the change of variables $s \rightarrow t-s$, we have the following alternative formula, which shows that $Q_k(t)$ are indeed trigonometric polynomials:

$$Q_k(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(s)E_k(t-s) ds$$

(11) Now given $\varepsilon > 0$, let us prove that the estimate in (9) holds indeed, with the trigonometric polynomial there being $Q_k(t)$, for $k \gg 0$ large enough. For this purpose, we use the uniform continuity of f , which tells us that we can find $\delta > 0$ such that:

$$|s-t| < \delta \implies |f(s) - f(t)| < \varepsilon$$

Indeed, by using this, we have the following estimate, for the error in (9):

$$\begin{aligned} |Q_k(t) - f(t)| &= \frac{1}{2\pi} \left| \int_{-\pi}^{\pi} (f(t-s) - f(t))E_k(s) ds \right| \\ &\leq \frac{1}{2\pi} \int_{-\pi}^{\pi} |f(t-s) - f(t)|E_k(s) ds \end{aligned}$$

(12) Now let us split the last integral into three parts, according to:

$$[-\pi, \pi] = [-\pi, -\delta] \cup [-\delta, \delta] \cup [\delta, \pi]$$

On the middle part the integrand is $< \varepsilon$, so the middle integral is $< \varepsilon$. As for the other two integrals, on $[-\pi, -\delta] \cup [\delta, \pi]$, we can use here (6), telling us that $E_k(t) \rightarrow 0$ uniformly, on that domain. Indeed, with $k \gg 0$ big enough the other two integrals are $< \varepsilon$ too, so we have obtained (9) as desired, with $\varepsilon \rightarrow 3\varepsilon$, which finishes the proof. \square

Still with me I hope, after all these computations. In practice now, there are several useful versions of Theorem 14.16, which can be of great use. We notably have here:

THEOREM 14.17. *We have an isomorphism $L^2(\mathbb{R})_{\text{per}} \simeq l^2(\mathbb{Z})$, as follows:*

(1) *Associated to $f \in L^2(\mathbb{R})_{\text{per}}$ are its Fourier coefficients, given by:*

$$\widehat{f}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)e^{int} dt$$

(2) *Associated to $g \in l^2(\mathbb{Z})$ is the series $S_g(t) = \sum_{n \in \mathbb{Z}} g(n)e^{-int}$.*

PROOF. This is something self-explanatory, based on the general orthonormal basis theory from Theorem 14.13, with the Fourier coefficients of $f \in L^2(\mathbb{R})_{\text{per}}$ being its coefficients $\widehat{f}(n) = \langle f, e^{-int} \rangle$ with respect to the basis $\{e^{int}\}$ from Theorem 14.16. \square

As yet another reformulation, or rather consequence of what we have, we have the Parseval formula, which is something very concrete and useful, as follows:

THEOREM 14.18. *The Fourier coefficients $\widehat{f}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)e^{int} dt$ satisfy*

$$\sum_{n \in \mathbb{Z}} \widehat{f}(n) \overline{\widehat{g}(n)} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \overline{g(t)} dt$$

for any $f, g \in L^2(\mathbb{R})_{\text{per}}$, and in particular satisfy the formula

$$\sum_{n \in \mathbb{Z}} |\widehat{f}(n)|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |f(t)|^2 dt$$

for any $f \in L^2(\mathbb{R})_{\text{per}}$, with this being called Parseval formula.

PROOF. This is indeed yet another reformulation of what we have, coming from the fact that the scalar products and norms are invariant under $L^2(\mathbb{R})_{\text{per}} \simeq l^2(\mathbb{Z})$. \square

With this discussed, time for some applications? And here, there are countless of them, because the above technology can be used in order to decompose various signals, such as mechanical, electromagnetic, seismic or acoustic waves, or even solutions of more complicated differential equations, somewhat of wave type, into sinusoids.

In what concerns us, we will be interested in applications to arithmetic:

THEOREM 14.19. *We have the following formula of Euler,*

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \dots = \frac{\pi^2}{6}$$

computing $\zeta(2)$, and solving the Basel problem.

PROOF. In order to prove this formula, we can use our Fourier series knowledge. Indeed, the nonzero Fourier coefficients of the function $f(t) = t$ on $[-\pi, \pi]$ are:

$$\begin{aligned} \widehat{f}(n) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} te^{int} dt \\ &= \frac{1}{2\pi} \left[\frac{1 - int}{n^2} e^{int} \right]_{-\pi}^{\pi} \\ &= \frac{(-1)^{n+1}}{n} i \end{aligned}$$

Thus, the Parseval formula for the function $f(t) = t$ gives:

$$\sum_{n \in \mathbb{Z}^*} \frac{1}{n^2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} t^2 dt = \frac{1}{2\pi} \cdot \frac{2\pi^3}{3} = \frac{\pi^2}{3}$$

And we therefore solved the Basel problem, just like that. Amazing. \square

14c. Bernoulli numbers

Still following Euler, our aim now will be that of generalizing the formula $\zeta(2) = \pi^2/6$, by computing the numbers $\zeta(2k)$, with $k \in \mathbb{N}$. The precise claim is as follows:

CLAIM 14.20. *We have the following formula, for the even integers $s = 2k$,*

$$\zeta(2k) = (-1)^{k+1} \frac{(2\pi)^{2k} B_{2k}}{2 \cdot (2k)!}$$

with B_n being the Bernoulli numbers, generalizing the Basel formula $\zeta(2) = \pi^2/6$.

To be more precise here, the Bernoulli numbers B_n are those that we met in chapter 6, in relation with computing the Riemann sums for $f(x) = x^p$. So, let us start with a discussion regarding these. Following the material in chapter 6, we first have:

DEFINITION 14.21. *The Bernoulli numbers $B_n \in \mathbb{Q}$ are defined by the formula*

$$\sum_{n=0}^m \binom{m+1}{n} B_n = \delta_{m0}$$

which fully computes them, by recurrence. Numerically, this formula gives

$$B_0 = 1, \quad B_1 = -\frac{1}{2}, \quad B_2 = \frac{1}{6}, \quad B_3 = 0, \quad B_4 = -\frac{1}{30}, \quad B_5 = 0, \quad B_6 = \frac{1}{42}$$

suggesting that all odd Bernoulli numbers, save for $B_1 = -1/2$, vanish.

As explained in chapter 6, the Bernoulli numbers naturally appear when looking at sums of type $1^p + \dots + N^p$. More on this in a moment, but before that, let us formulate the following result, which is something very useful, making the link with analysis, and which proves right away the above conjecture, regarding the odd Bernoulli numbers:

THEOREM 14.22. *We have the following formula, valid for any $x \in \mathbb{R}$,*

$$\frac{x}{e^x - 1} = \sum_{n=0}^{\infty} B_n \frac{x^n}{n!}$$

which can stand as an alternative definition for the Bernoulli numbers B_n .

PROOF. This is something very standard, the idea being as follows:

(1) To start with, the function in the statement expands as follows:

$$\frac{x}{e^x - 1} = \frac{1}{1 + \frac{x}{2} + \frac{x^2}{6} + \frac{x^3}{24} + \dots} = 1 - \frac{x}{2} + \frac{x^2}{12} - \frac{x^4}{720} + \dots$$

To be more precise, the first coefficient $-1/2$ is needed for x^2 , and the next coefficient, $1/12$, comes according to $1/6 - 1/4 + 1/12 = 0$, which is needed for x^3 . And so on.

(2) In order to prove now the result, let us write our function as follows:

$$\frac{x}{e^x - 1} = \sum_{n=0}^{\infty} b_n \frac{x^n}{n!}$$

For finding the coefficients, we multiply both sides by $(e^x - 1)/x$. By using the usual series of the exponential, this gives the following formula, involving our coefficients:

$$\begin{aligned} 1 &= \frac{e^x - 1}{x} \sum_{n=0}^{\infty} b_n \frac{x^n}{n!} \\ &= \sum_{k=1}^{\infty} \frac{x^{k-1}}{k!} \sum_{n=0}^{\infty} b_n \frac{x^n}{n!} \\ &= \sum_{m=0}^{\infty} x^m \sum_{n=0}^m \frac{b_n}{(m-n+1)!n!} \\ &= \sum_{m=0}^{\infty} \frac{x^m}{(m+1)!} \sum_{n=0}^m \binom{m+1}{n} b_n \end{aligned}$$

(3) By looking now at the coefficient of x^m , we obtain the following formula:

$$\delta_{m0} = \frac{1}{(m+1)!} \sum_{n=0}^m \binom{m+1}{n} b_n$$

But the factorial on the right can be erased, since at $m \geq 1$ we are dealing with a $0 = 0$ equation, and we conclude that we have $b_n = B_n$, as desired.

(4) Finally, let us prove what was claimed before the statement, namely that the odd Bernoulli numbers, save for $B_1 = -1/2$, vanish. With $f(x) = x/(e^x - 1)$ we have:

$$\begin{aligned} f(x) - f(-x) &= \frac{x}{e^x - 1} + \frac{x}{e^{-x} - 1} \\ &= \frac{x}{e^x - 1} - \frac{xe^x}{e^x - 1} \\ &= -x \end{aligned}$$

Thus the odd coefficients vanish indeed, except for that first one, which is $-1/2$. \square

Getting now to the computation of the sums $1^p + \dots + N^p$, we first have:

PROPOSITION 14.23. *We have the following formula,*

$$1^p + 2^p + \dots + (n-1)^p = \frac{1}{p+1} \sum_{k=0}^p \binom{p+1}{k} B_k n^{p+1-k}$$

valid for any integer exponent $p \geq 1$.

PROOF. We use the following formula, coming from the definition of exp:

$$e^{kt} = \sum_{p=0}^{\infty} \frac{(kt)^p}{p!}$$

By using this formula at $k = 0, 1, \dots, n-1$ and summing, we get:

$$1 + e^t + \dots + e^{(n-1)t} = \sum_{p=0}^{\infty} (1^p + 2^p + \dots + (n-1)^p) \frac{t^p}{p!}$$

Now observe that the term on the left is given by the following formula:

$$\begin{aligned} 1 + e^t + \dots + e^{(n-1)t} &= \frac{e^{nt} - 1}{e^t - 1} \\ &= \frac{e^{nt} - 1}{t} \cdot \frac{t}{e^t - 1} \\ &= \sum_{s=1}^{\infty} n^s \frac{t^{s-1}}{s!} \sum_{k=0}^{\infty} B_k \frac{t^k}{k!} \\ &= \sum_{p=0}^{\infty} t^p \sum_{k=0}^p n^{p-k+1} \frac{B_k}{(p-k+1)!k!} \\ &= \sum_{p=0}^{\infty} \frac{t^p}{(p+1)!} \sum_{k=0}^p \binom{p+1}{k} B_k n^{p+1-k} \\ &= \sum_{p=0}^{\infty} \frac{t^p}{p!} \cdot \frac{1}{p+1} \sum_{k=0}^p \binom{p+1}{k} B_k n^{p+1-k} \end{aligned}$$

Thus, we are led to the formula in the statement. □

As an interesting remark now, the above formula fails at $p = 0$, due to:

$$1^0 + \dots + (n-1)^0 = n-1 \neq n = \binom{1}{0} B_0 n$$

Alternatively, however, we have the following result, which works at $p = 0$ too:

THEOREM 14.24. *We have the following formula,*

$$1^p + 2^p + \dots + N^p = \frac{1}{p+1} \sum_{k=0}^p (-1)^k \binom{p+1}{k} B_k N^{p+1-k}$$

valid for any integer exponent $p \geq 0$.

PROOF. As before, we can use the series of the exponential, namely:

$$e^{-kt} = \sum_{p=0}^{\infty} \frac{(-kt)^p}{p!}$$

Indeed, by using this formula at $k = 1, 2, \dots, N$ and summing, we get:

$$e^{-t} + e^{-2t} + \dots + e^{-Nt} = \sum_{p=0}^{\infty} (1^p + 2^p + \dots + N^p) \frac{(-t)^p}{p!}$$

Now observe that the term on the left is given by the following formula:

$$\begin{aligned} e^{-t} + e^{-2t} + \dots + e^{-Nt} &= e^{-Nt}(1 + e^t + \dots + e^{Nt}) \\ &= e^{-Nt} \frac{e^{Nt} - 1}{e^t - 1} \\ &= \frac{1 - e^{-Nt}}{e^t - 1} \\ &= \frac{1 - e^{-Nt}}{t} \cdot \frac{t}{e^t - 1} \\ &= \sum_{s=1}^{\infty} (-1)^{s-1} N^s \frac{t^{s-1}}{s!} \sum_{k=0}^{\infty} B_k \frac{t^k}{k!} \\ &= \sum_{p=0}^{\infty} t^p \sum_{k=0}^p (-1)^{p-k} N^{p+1-k} \frac{B_k}{(p+1-k)!k!} \\ &= \sum_{p=0}^{\infty} \frac{(-t)^p}{(p+1)!} \sum_{k=0}^p (-1)^k \binom{p+1}{k} B_k N^{p+1-k} \\ &= \sum_{p=0}^{\infty} \frac{(-t)^p}{p!} \cdot \frac{1}{p+1} \sum_{k=0}^p (-1)^k \binom{p+1}{k} B_k N^{p+1-k} \end{aligned}$$

Thus, we have our formula. And as a verification, at $p = 0$ we have indeed:

$$1^0 + \dots + N^0 = N = \binom{1}{0} B_0 N$$

Thus, we are led to the conclusions in the statement. \square

Summarizing, we reviewed the material from chapter 6, regarding the Bernoulli numbers, with full proofs this time, for everything, and with a number of extra results, and notably Theorem 14.22, making the link with analysis, coming as a bonus.

Finally, no general discussion about the Bernoulli numbers would be complete without mentioning the Euler-Maclaurin formula, involving them, which is as follows:

$$\begin{aligned} \sum_{k=0}^{n-1} f(x) &\simeq \int_0^n f(x)dx - \frac{1}{2}(f(n) - f(0)) \\ &\quad + \frac{1}{6} \cdot \frac{f'(n) - f'(0)}{2!} - \frac{1}{30} \cdot \frac{f^{(3)}(n) - f^{(3)}(0)}{4!} \\ &\quad + \frac{1}{42} \cdot \frac{f^{(5)}(n) - f^{(5)}(0)}{6!} - \frac{1}{30} \cdot \frac{f^{(7)}(n) - f^{(7)}(0)}{8!} + \dots \end{aligned}$$

And exercise of course for you, to learn more about this, and about Bernoulli numbers in general. Needless to say, all this is very useful knowledge, for many purposes.

14d. Special values

Getting now towards zeta, and what we wanted to do, namely prove Claim 14.20, we must develop some further analytic theory for the Bernoulli numbers. We already know from Theorem 14.22 that, up to a rescaling, the Bernoulli numbers B_n appear as the Taylor coefficients of $x/(e^x - 1)$. But, this latter function reminds the hyperbolic cotangent function, $\coth = \cosh / \sinh$, and with this idea in mind, we are led to:

PROPOSITION 14.25. *The Taylor series of \coth is given by*

$$\coth x = \sum_{k=0}^{\infty} \frac{4^k B_{2k}}{(2k)!} x^{2k-1}$$

exact formula, with B_n being the Bernoulli numbers.

PROOF. We have indeed the following manipulations, on the function \coth :

$$\begin{aligned} \coth x = \frac{e^x + e^{-x}}{e^x - e^{-x}} &\implies \coth x = \frac{e^{2x} + 1}{e^{2x} - 1} \\ &\implies \coth \frac{x}{2} = \frac{e^x + 1}{e^x - 1} \\ &\implies \coth \frac{x}{2} - 1 = \frac{2}{e^x - 1} \\ &\implies \frac{x}{2} \left(\coth \frac{x}{2} - 1 \right) = \frac{x}{e^x - 1} \end{aligned}$$

But this makes the link with Theorem 14.22, which applies and gives:

$$\frac{x}{2} \left(\coth \frac{x}{2} - 1 \right) = \sum_{n=0}^{\infty} \frac{B_n}{n!} x^n$$

Thus, we are led to the formula for \coth in the statement. □

By using now the standard tricks for passing from usual functions to hyperbolic ones, and vice versa, we can convert Proposition 14.25 into something about \cot , as follows:

THEOREM 14.26. *The Taylor series of \cot is given by*

$$\cot x = \sum_{k=0}^{\infty} (-1)^k \frac{4^k B_{2k}}{(2k)!} x^{2k-1}$$

exact formula, with B_n being the Bernoulli numbers.

PROOF. Recall indeed the following basic formulae, for the cosine and sine:

$$\cos x = \frac{e^{ix} + e^{-ix}}{2}, \quad \sin x = \frac{e^{ix} - e^{-ix}}{2i}$$

By using these formulae, we obtain the following formula, for $\cot x$:

$$\cot x = i \cdot \frac{e^{ix} + e^{-ix}}{e^{ix} - e^{-ix}} = i \coth(ix)$$

But now, Proposition 14.25 applies, and gives the following formula:

$$\cot x = i \sum_{k=0}^{\infty} \frac{4^k B_{2k}}{(2k)!} (ix)^{2k-1} = \sum_{k=0}^{\infty} (-1)^k \frac{4^k B_{2k}}{(2k)!} x^{2k-1}$$

Thus, we are led to the formula for \cot in the statement. \square

All the above certainly makes a link with Claim 14.20, but remains something quite elementary. So, time for Fourier analysis to strike. We have the following key result:

THEOREM 14.27. *We have the following formula for \cot ,*

$$\cot x = \frac{1}{x} - 2 \sum_{n=1}^{\infty} \frac{x}{\pi^2 n^2 - x^2}$$

coming from Fourier analysis.

PROOF. Consider the function $f(x) = \cos(\alpha t)$. Its Fourier coefficients are:

$$\begin{aligned} \widehat{f}(n) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\alpha t) e^{int} dt \\ &= \frac{1}{4\pi} \int_{-\pi}^{\pi} e^{i(n+\alpha)t} + e^{i(n-\alpha)t} dt \\ &= \frac{1}{4\pi i} \left[\frac{e^{i(n+\alpha)t}}{n+\alpha} + \frac{e^{i(n-\alpha)t}}{n-\alpha} \right]_{-\pi}^{\pi} \\ &= \frac{1}{2\pi} \left(\frac{\sin((n+\alpha)\pi)}{n+\alpha} + \frac{\sin((n-\alpha)\pi)}{n-\alpha} \right) \end{aligned}$$

We can further process the above formula, as follows:

$$\begin{aligned}\widehat{f}(n) &= \frac{1}{2\pi} \left(\frac{(-1)^n \sin(\alpha\pi)}{n + \alpha} - \frac{(-1)^n \sin(-\alpha\pi)}{n - \alpha} \right) \\ &= (-1)^n \frac{\sin(\alpha\pi)}{2\pi} \left(\frac{1}{n + \alpha} - \frac{1}{n - \alpha} \right) \\ &= (-1)^{n+1} \frac{\alpha \sin(\alpha\pi)}{\pi(n^2 - \alpha^2)}\end{aligned}$$

By the Fourier inversion theorem, we conclude from this that we have:

$$\cos(\alpha t) = \sum_{n \in \mathbb{Z}} (-1)^{n+1} \frac{\alpha \sin(\alpha\pi)}{\pi(n^2 - \alpha^2)} e^{-int}$$

Now with $t = \pi$, so that $e^{-int} = (-1)^n$, we obtain the following formula:

$$\cos(\alpha\pi) = - \sum_{n \in \mathbb{Z}} \frac{\alpha \sin(\alpha\pi)}{\pi(n^2 - \alpha^2)}$$

By dividing now both sides by $\sin(\alpha\pi)$, this gives the following formula:

$$\cot(\alpha\pi) = - \sum_{n \in \mathbb{Z}} \frac{\alpha}{\pi(n^2 - \alpha^2)}$$

Finally, in terms of the new variable $x = \alpha\pi$, this formula becomes:

$$\cot x = - \sum_{n \in \mathbb{Z}} \frac{x}{\pi^2 n^2 - x^2}$$

We are therefore led to the formula in the statement. □

We can go back now to the zeta function, and following Euler, we have:

THEOREM 14.28. *We have the following formula, for the even integers $s = 2k$,*

$$\zeta(2k) = (-1)^{k+1} \frac{(2\pi)^{2k} B_{2k}}{2 \cdot (2k)!}$$

with B_n being the Bernoulli numbers, which in practice gives the formulae

$$\zeta(2) = \frac{\pi^2}{6} \quad , \quad \zeta(4) = \frac{\pi^4}{90} \quad , \quad \zeta(6) = \frac{\pi^6}{945} \quad , \quad \zeta(8) = \frac{\pi^8}{9450} \quad , \quad \dots$$

generalizing the formula $\zeta(2) = \pi^2/6$ of Euler, solving the Basel problem.

PROOF. This follows by comparing Theorems 14.26 and 14.27, but it is instructive to do some further manipulations on the formulae found there, as follows:

(1) To start with, we can rewrite the formula in Theorem 14.27 as follows:

$$\begin{aligned}
 \cot x = \frac{1}{x} - 2 \sum_{n=1}^{\infty} \frac{x}{\pi^2 n^2 - x^2} &\implies x \cot x = 1 - 2 \sum_{n=1}^{\infty} \frac{x^2}{\pi^2 n^2 - x^2} \\
 &\implies \frac{1 - x \cot x}{2} = \sum_{n=1}^{\infty} \frac{x^2}{\pi^2 n^2 - x^2} \\
 &\implies \frac{1 - \pi x \cot(\pi x)}{2} = \sum_{n=1}^{\infty} \frac{x^2}{n^2 - x^2}
 \end{aligned}$$

(2) Now by expanding the series on the right, we are led to the following remarkable formula, for the generating series of the numbers $\zeta(2k)$, with $k \geq 1$:

$$\begin{aligned}
 \frac{1 - \pi x \cot(\pi x)}{2} &= \sum_{n=1}^{\infty} \frac{(x/n)^2}{1 - (x/n)^2} \\
 &= \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \left(\frac{x}{n}\right)^{2k} \\
 &= \sum_{k=1}^{\infty} \zeta(2k) x^{2k}
 \end{aligned}$$

(3) On the other hand, we can rewrite the formula in Theorem 14.26 as follows:

$$\begin{aligned}
 \cot x = \sum_{k=0}^{\infty} (-1)^k \frac{4^k B_{2k}}{(2k)!} x^{2k-1} &\implies x \cot x = \sum_{k=0}^{\infty} (-1)^k \frac{4^k B_{2k}}{(2k)!} x^{2k} \\
 &\implies \frac{1 - x \cot x}{2} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{4^k B_{2k}}{2 \cdot (2k)!} x^{2k} \\
 &\implies \frac{1 - \pi x \cot(\pi x)}{2} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{(2\pi)^{2k} B_{2k}}{2 \cdot (2k)!} x^{2k}
 \end{aligned}$$

Thus, we are led to the various conclusions in the statement. \square

Good to have this done. Regarding now the values $\zeta(2k+1)$ with $k \in \mathbb{N}$, the story here is more complicated, with the first such number being the Apéry constant:

$$\zeta(3) = \sum_{n=1}^{\infty} \frac{1}{n^3}$$

There has been a lot of work on this number, by Apéry and others, and on the higher $\zeta(2k+1)$ values as well. As a very basic result on this topic, let us record:

THEOREM 14.29. *We have the following formula,*

$$\zeta(s) = \frac{1}{(s-1)!} \int_0^\infty \frac{x^{s-1}}{e^x - 1} dx$$

valid for any integer $s \geq 2$.

PROOF. We have indeed the following computation, for the above integral:

$$\begin{aligned} \int_0^\infty \frac{x^{s-1}}{e^x - 1} dx &= \int_0^\infty x^{s-1} (e^{-x} + e^{-2x} + e^{-3x} + \dots) \\ &= \sum_{n=1}^\infty \int_0^\infty x^{s-1} e^{-nx} dx \\ &= \sum_{n=1}^\infty \int_0^\infty \left(\frac{y}{n}\right)^{s-1} e^{-y} \frac{dy}{n} \\ &= \sum_{n=1}^\infty \frac{1}{n^s} \int_0^\infty y^{s-1} e^{-y} dy \\ &= (s-1)! \zeta(s) \end{aligned}$$

Here we have used a standard formula for the last integral, obtained by iterated partial integration, that we know from chapter 7. Thus, we have our result. \square

Let us record as well the following statement, a bit of physics flavor:

THEOREM 14.30. *We have the following formula,*

$$\zeta(s) = \int_0^1 \dots \int_0^1 \frac{dx_1 \dots dx_s}{1 - x_1 \dots x_s}$$

valid for any integer $s \geq 2$.

PROOF. This follows as usual from some calculus, the idea being as follows:

(1) At $s = 2$ we have indeed the following computation, using Theorem 14.29:

$$\begin{aligned} \int_0^1 \int_0^1 \frac{1}{1 - xy} dx dy &= \int_0^1 \left[-\frac{\log(1 - xy)}{y} \right]_0^1 dy \\ &= - \int_0^1 \frac{\log(1 - y)}{y} dy \\ &= - \int_0^\infty \frac{\log(e^{-t})}{1 - e^{-t}} e^{-t} dt \\ &= \int_0^\infty \frac{t}{e^t - 1} dt \\ &= \zeta(2) \end{aligned}$$

(2) In the general case, $s \in \mathbb{N}$, the best is to start with the following formula:

$$\frac{1}{1 - x_1 \dots x_s} = \sum_{n=0}^{\infty} (x_1 \dots x_s)^n$$

Thus, the integral in the statement is given by the following formula:

$$\int_0^1 \dots \int_0^1 \frac{dx_1 \dots dx_s}{1 - x_1 \dots x_s} = \sum_{n=0}^{\infty} \int_0^1 \dots \int_0^1 (x_1 \dots x_s)^n dx_1 \dots dx_s$$

But this eventually leads to the formula in the statement, say exercise for you.

(3) Before leaving, let us see as well, out of mathematical curiosity, what happens at the exponent $s = 1$. Here the integral in the statement is:

$$\begin{aligned} \int_0^1 \frac{1}{1-x} dx &= [-\log(1-x)]_0^1 \\ &= -\log(1-1) + \log(1-0) \\ &= \infty + 0 \\ &= \zeta(1) \end{aligned}$$

Not a big deal, you would say, but as an interesting remark, since $\log(1-x) \simeq -x$, we are led to the conclusion that ζ , when suitably extended by analytic continuation, should have a simple pole at $s = 1$, with residue 1. We will be back to this, later. \square

Many other things can be said about ζ and its special values, as a continuation of the above, and for more about this, you can check any advanced number theory book.

14e. Exercises

Welcome to the zeta function, and no way back. As exercises about it, we have:

EXERCISE 14.31. *Learn more about the Möbius function, and its properties.*

EXERCISE 14.32. *Establish the arithmetic formulae involving ζ^3 and ζ^4 .*

EXERCISE 14.33. *Learn more about Fourier analysis, as much as you can.*

EXERCISE 14.34. *Learn about the Weierstrass factorization of analytic functions.*

EXERCISE 14.35. *Learn more about the Bernoulli numbers, and their story.*

EXERCISE 14.36. *Learn about the Euler-Maclaurin formula, and its applications.*

EXERCISE 14.37. *Compute the Taylor series of \csc , csch , and of \tan , \tanh too.*

EXERCISE 14.38. *Work out the details, for the formula for $\zeta(s)$, $s \in \mathbb{N}$ at the end.*

As standard bonus exercise, spend more time with ζ , the more the better.

CHAPTER 15

Complex zeta

15a. Gamma function

We have learned many interesting things about the zeta function in chapter 14, but that study was mostly at real exponents $s > 1$, not really in relation with complex analysis, where the zeta function is really useful, in relation with questions from arithmetic.

So, time to discuss this now, what happens for complex exponents s , how the zeta function can be constructed there, and what its properties are. Let us start with:

THEOREM 15.1. *We can talk about the Riemann zeta function*

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

when $\operatorname{Re}(s) > 1$, and we have the following formula, with product over primes:

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}$$

Also, the generating function of the numbers $\zeta(2k)$ with $k \in \mathbb{N}$ is given by

$$\sum_{k=1}^{\infty} \zeta(2k)x^{2k} = \frac{1 - \pi x \cot(\pi x)}{2}$$

which gives the following formula, with B_n being the Bernoulli numbers,

$$\zeta(2k) = (-1)^{k+1} \frac{(2\pi)^{2k} B_{2k}}{2 \cdot (2k)!}$$

which generalizes the Euler formula $\zeta(2) = \pi^2/6$, solving the Basel problem.

PROOF. In what concerns the first assertion, convergence at $\operatorname{Re}(s) > 1$, we have indeed the following computation, assuming $s = r + it$, with $r > 1$:

$$|\zeta(s)| \leq \sum_{n=1}^{\infty} \frac{1}{|n^s|} = \sum_{n=1}^{\infty} \frac{1}{n^r} < 1 + \int_1^{\infty} \frac{1}{x^r} dx = 1 + \frac{1}{r-1}$$

The second assertion follows from the prime factorization of the integers. As for the computation of $\zeta(2k)$, this was something quite tricky, explained in chapter 14. \square

We refer to chapter 14 for a number of supplementary facts regarding zeta, notably with the following inversion formula, where μ is the Möbius function of the integers:

$$\frac{1}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\mu(n)}{n^s}$$

In order to present now a number of more advanced results regarding zeta, the idea will be that of relating it to another well-known function from complex analysis, the gamma one. Let us start with the following fact, which is something elementary:

THEOREM 15.2. *The following integral converges absolutely when $\operatorname{Re}(s) > 0$,*

$$\Gamma(s) = \int_0^{\infty} x^{s-1} e^{-x} dx$$

satisfies $\Gamma(s+1) = s\Gamma(s)$, and extends the usual factorial of integers, $\Gamma(s) = (s-1)!$.

PROOF. There are several assertions here, the idea being as follows:

(1) Regarding the convergence claim, we have the following trivial estimate, with $s = r + it$, coming from the formula $|x^{s-1}| = |x^{r+it-1}| = x^{r-1}$:

$$|\Gamma(s)| \leq \int_0^{\infty} |x^{s-1}| e^{-x} dx = \int_0^{\infty} x^{r-1} e^{-x} dx$$

Now regarding the integral on the right, which is $\Gamma(r)$, that converges at ∞ , due to the presence of the exponential, and converges at 0 too, because x^{r-1} with $r > 0$ is integrable at 0. Thus, we have indeed absolute convergence at $\operatorname{Re}(s) > 0$, as stated.

(2) Next, by partial integration we have the following formula, as claimed:

$$\begin{aligned} \Gamma(s+1) &= \int_0^{\infty} x^s e^{-x} dx \\ &= \int_0^{\infty} s x^{s-1} e^{-x} dx \\ &= s \Gamma(s) \end{aligned}$$

(3) Regarding now the case $s \in \mathbb{N}$, for the initial value $s = 1$ we have:

$$\Gamma(1) = \int_0^{\infty} e^{-x} dx = 1$$

Thus, by recurrence we obtain $\Gamma(s) = (s-1)!$ for any $s \in \mathbb{N}$, as claimed. \square

The above result is something quite magic, because it allows to talk about the factorial of any complex number z with $\operatorname{Re}(z) > -1$, according to the following formula:

$$z! = \Gamma(z+1)$$

More on this in a moment, but before that, let us point out the fact that the assumption $\operatorname{Re}(z) > -1$ can be in fact removed, due to the following extension of Theorem 15.2:

THEOREM 15.3 (update). *We can talk about the gamma function*

$$\Gamma(s) = \int_0^\infty x^{s-1} e^{-x} dx$$

with the integral converging absolutely when $\operatorname{Re}(s) > 0$, which satisfies the formula

$$\Gamma(s+1) = s\Gamma(s)$$

which can stand as a definition for Γ , for $s \notin -\mathbb{N}$. Also, $\Gamma(s) = (s-1)!$ for $s \in \mathbb{N}$.

PROOF. This is indeed something self-explanatory, based on Theorem 15.2, the idea being that once we have $\Gamma(s)$ via the integral in the statement, for $\operatorname{Re}(s) > 0$, we can successively extend it to the case $\operatorname{Re}(s) > -1$, then $\operatorname{Re}(s) > -2$ and so on, by using:

$$\Gamma(s) = \frac{\Gamma(s+1)}{s}$$

To be more precise, this recursive procedure will certainly work, providing us with a formula for $\Gamma(s)$ at any $s \in \mathbb{C}$, provided that the various denominators which are involved do not vanish, leading to the assumption $s \notin -\mathbb{N}$ from the statement. \square

As a comment now, the above result obviously involves some trickery, because what we are saying, in the end, is that the divergent integrals at $\operatorname{Re}(s) < 0$ can be regarded as producing bounded quantities, via the formula $\Gamma(s+1) = s\Gamma(s)$, used as indicated. You might probably say that this sounds like physics, and in answer, this is in fact mathematics, called “analytic continuation”. More on such things later.

Many interesting things can be said about the gamma function, notably with an explicit computation at the half-integers, which leads to the following result:

THEOREM 15.4. *The gamma function is given at half-integers by*

$$\Gamma(n) = (n-1)! \quad , \quad \Gamma\left(n + \frac{1}{2}\right) = \frac{(2n)!!}{2^n} \sqrt{\pi}$$

and we have the following almost uniform formula for it, valid for any $N \in \mathbb{N}$,

$$\Gamma\left(\frac{N}{2}\right) = \frac{(N-1)!!}{2^{(N-1)/2}} c$$

with $M!! = (M-1)(M-3)\dots$ as usual, and $c = \sqrt{2}, \sqrt{\pi}$ for N even, odd.

PROOF. There are several things going on here, the idea being as follows:

(1) Regarding the first formula, values of the gamma function at the positive integers, we know indeed, from Theorem 15.2, that for $n \in \mathbb{N}$ we have:

$$\Gamma(n) = (n-1)!$$

(2) Regarding now the half-integers, we first have the following computation, using at the end the formula of the Gauss integral, that we learned in chapter 7:

$$\begin{aligned}
 \Gamma\left(\frac{1}{2}\right) &= \int_0^\infty x^{-1/2} e^{-x} dx \\
 &= \int_0^\infty y^{-1} e^{-y^2} 2y dy \\
 &= 2 \int_0^\infty e^{-y^2} dy \\
 &= 2 \times \frac{\sqrt{\pi}}{2} \\
 &= \sqrt{\pi}
 \end{aligned}$$

Next, by using the formula $\Gamma(s+1) = s\Gamma(s)$, we have the following computations:

$$\Gamma\left(\frac{3}{2}\right) = \frac{1}{2} \Gamma\left(\frac{1}{2}\right) = \frac{1}{2} \sqrt{\pi}$$

$$\Gamma\left(\frac{5}{2}\right) = \frac{3}{2} \Gamma\left(\frac{3}{2}\right) = \frac{3}{4} \sqrt{\pi}$$

$$\Gamma\left(\frac{7}{2}\right) = \frac{5}{2} \Gamma\left(\frac{5}{2}\right) = \frac{15}{8} \sqrt{\pi}$$

\vdots

Thus, we can solve the problem by recurrence, and we obtain in this way, with our usual convention $N!! = (N-1)(N-3)(N-5)\dots$ for the double factorials:

$$\begin{aligned}
 \Gamma\left(n + \frac{1}{2}\right) &= \frac{2n-1}{2} \cdot \frac{2n-3}{2} \dots \frac{5}{2} \cdot \frac{3}{2} \cdot \frac{1}{2} \Gamma\left(\frac{1}{2}\right) \\
 &= \frac{1 \cdot 3 \cdot 5 \dots (2n-1)}{2^n} \sqrt{\pi} \\
 &= \frac{(2n)!!}{2^n} \sqrt{\pi}
 \end{aligned}$$

(3) Regarding now the unification of the formulae that we have in (1) and (2), let us first rewrite the formula that we just found in (2), in the following way:

$$\Gamma\left(\frac{2n+1}{2}\right) = \frac{(2n)!!}{2^n} \sqrt{\pi}$$

Which looks good and in final form, no questions about this, so for a unification we are left with refurbishing the nice formula found in (1), in the following way:

$$\begin{aligned}
 \Gamma\left(\frac{2n}{2}\right) &= \Gamma(n) \\
 &= (n-1)! \\
 &= \frac{2 \cdot 4 \cdot 6 \cdots (2n-2)}{2^{n-1}} \\
 &= \frac{(2n-1)!!}{2^{n-1}} \\
 &= \frac{(2n-1)!!}{2^{n-1/2}} \sqrt{2}
 \end{aligned}$$

And with this, job done, we are led to the uniform formula in the statement. \square

As an interesting comment now, the above formulae remind those for the volumes of spheres, from chapter 7. In relation with this, we have the following result:

THEOREM 15.5. *The volume and area of the unit sphere in \mathbb{R}^N , given by*

$$V = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!} \quad , \quad A = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N-1)!!}$$

can be expressed in terms of the gamma function, the formulae being

$$V = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^{(N-1)/2} c}{\Gamma(N/2 + 1)} \quad , \quad A = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^{(N+1)/2} c}{\Gamma(N/2)}$$

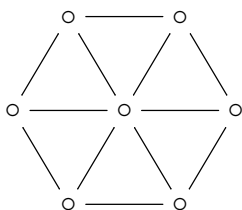
with $c = \sqrt{2}, \sqrt{\pi}$ for N even, odd, as before.

PROOF. There are several things going on here, the idea being as follows:

(1) To start with, the first formula of V is something that we know from chapter 7, obtained via multiple integrals, and the Wallis formula. It is of course possible to use the same method for computing A , but there is in fact no need for this, because we have:

$$A = N \cdot V$$

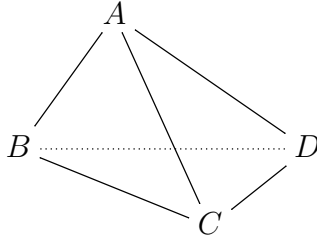
(2) So, question now, why does this latter formula hold? In answer, recall from chapter 5, when first talking about the number π , that at $N = 2$ we can cut the unit disk, or rather a regular M -gon approximating it, with $M \gg 0$, as a pizza, as follows:



The volume 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 $A = MS$ is the total length of the sides:

$$V = M \times \frac{HS}{2} = \frac{HP}{2} \simeq \frac{A}{2}$$

(3) Let us see as well what happens in $N = 3$ dimensions. In order to compute three-dimensional volumes, consider a tetrahedron, in 3D space:



The volume of this tetrahedron is then given by the following formula, coming for instance by constructing a triangular prism, out of 3 copies of this tetrahedron:

$$\text{volume} = \frac{1}{3} \times \text{basis area} \times \text{height}$$

But with this in hand, the pizza argument from (2) will apply as well in 3D, with a 3 factor, giving the following formula, connecting the volume and area of the sphere:

$$V = \frac{A}{3}$$

(4) And so on, the idea being that in N dimensions, the same pizza argument applies, and gives $A = NV$. But with this, we can compute the area of the sphere, as follows:

$$A = N \cdot \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!} = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N-1)!!}$$

(5) Getting now to the second series of formulae, involving gamma, recall from Theorem 15.4 that we have the following formula, with $c = \sqrt{2}, \sqrt{\pi}$ for N even, odd:

$$\Gamma\left(\frac{N}{2}\right) = \frac{(N-1)!!}{2^{(N-1)/2}} c$$

Thus, the double factorials can be expressed as follows, in terms of gamma:

$$(N-1)!! = \frac{2^{(N-1)/2}}{c} \Gamma\left(\frac{N}{2}\right)$$

(6) But with this in hand, the previous volume formula for the sphere becomes:

$$\begin{aligned} V &= \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!} \\ &= \left(\frac{\pi}{2}\right)^{[N/2]} 2^N \cdot \frac{c}{2^{(N+1)/2}} \cdot \frac{1}{\Gamma(N/2+1)} \\ &= \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^{(N-1)/2} c}{\Gamma(N/2+1)} \end{aligned}$$

As for the formula for the area of the sphere, this becomes, again as claimed:

$$\begin{aligned} A &= \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N-1)!!} \\ &= \left(\frac{\pi}{2}\right)^{[N/2]} 2^N \cdot \frac{c}{2^{(N-1)/2}} \cdot \frac{1}{\Gamma(N/2)} \\ &= \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^{(N+1)/2} c}{\Gamma(N/2)} \end{aligned}$$

(7) Time perhaps for some numerics? At $N = 2$ we have $c = \sqrt{2}$, and we get:

$$V = \frac{\pi}{2} \cdot \frac{\sqrt{2} \cdot \sqrt{2}}{\Gamma(2)} = \pi \quad , \quad A = \frac{\pi}{2} \cdot \frac{2\sqrt{2} \cdot \sqrt{2}}{\Gamma(1)} = 2\pi$$

Also, at $N = 3$ we have $c = \sqrt{\pi}$, and we obtain, again as we should:

$$V = \frac{\pi}{2} \cdot \frac{2 \cdot \sqrt{\pi}}{3\sqrt{\pi}/4} = \frac{4\pi}{3} \quad , \quad A = \frac{\pi}{2} \cdot \frac{4 \cdot \sqrt{\pi}}{\sqrt{\pi}/2} = 4\pi$$

(8) Thus, formulae proved and doublechecked. And with the remark, however, that the original formulae, in terms of double factorials, are quite often better in practice. \square

As a last piece of mathematics regarding the gamma function, again in relation with what we did in chapter 7 for the spheres, we have the following Stirling estimates:

THEOREM 15.6. *The gamma function is given at half-integers by*

$$\Gamma(n) \simeq \left(\frac{n}{e}\right)^n \sqrt{\frac{2\pi}{n}} \quad , \quad \Gamma\left(n + \frac{1}{2}\right) \simeq \left(\frac{n}{e}\right)^n \sqrt{2\pi}$$

and we have the following uniform approximation formula for it,

$$\Gamma\left(\frac{N}{2}\right) \simeq \left(\frac{N}{2e}\right)^{N/2} \sqrt{\frac{4\pi}{N}}$$

valid for $N \in \mathbb{N}$, in the $N \rightarrow \infty$ limit.

PROOF. This is very standard, based on the formulae in Theorem 15.4:

(1) At the usual integers $n \in \mathbb{N}$ we have the following Stirling estimate:

$$\begin{aligned}
 \Gamma(n) &= (n-1)! \\
 &\simeq \left(\frac{n-1}{e}\right)^{n-1} \sqrt{2\pi(n-1)} \\
 &= \left(\frac{n}{e}\right)^{n-1} \left(\frac{n-1}{n}\right)^{n-1} \sqrt{2\pi(n-1)} \\
 &\simeq \left(\frac{n}{e}\right)^{n-1} \frac{\sqrt{2\pi n}}{e} \\
 &= \left(\frac{n}{e}\right)^n \sqrt{\frac{2\pi}{n}}
 \end{aligned}$$

(2) At the half-integers, $n + 1/2$ with $n \in \mathbb{N}$, we have the following estimate:

$$\begin{aligned}
 \Gamma\left(n + \frac{1}{2}\right) &= \frac{(2n)!!}{2^n} \sqrt{\pi} \\
 &= \frac{(2n)!}{4^n n!} \sqrt{\pi} \\
 &\simeq \left(\frac{2n}{e}\right)^{2n} \sqrt{4\pi n} \left(\frac{e}{n}\right)^n \frac{1}{\sqrt{2\pi n}} \cdot \sqrt{\pi} \\
 &= \left(\frac{n}{e}\right)^n \sqrt{2\pi}
 \end{aligned}$$

(3) Regarding the uniform formula, with $N = 2n$ the estimate in (1) reads:

$$\Gamma\left(\frac{N}{2}\right) \simeq \left(\frac{N}{2e}\right)^{N/2} \sqrt{\frac{4\pi}{N}}$$

(4) Also, with $N = 2n + 1$, the estimate that we found in (2) reads:

$$\begin{aligned}
 \Gamma\left(\frac{N}{2}\right) &\simeq \left(\frac{N-1}{2e}\right)^{(N-1)/2} \sqrt{2\pi} \\
 &= \left(\frac{N}{2e}\right)^{(N-1)/2} \left(\frac{N-1}{N}\right)^{(N-1)/2} \sqrt{2\pi} \\
 &\simeq \left(\frac{N}{2e}\right)^{(N-1)/2} \sqrt{\frac{2\pi}{e}} \\
 &\simeq \left(\frac{N}{2e}\right)^{N/2} \sqrt{\frac{4\pi}{N}}
 \end{aligned}$$

We are therefore led to the uniform formula in the statement.

(5) Finally, let us mention that the uniform approximation formula that we found works in fact in general, with the formula being as follows, for $|s| \gg 0$:

$$\Gamma(s) \simeq \left(\frac{s}{e}\right)^s \sqrt{\frac{2\pi}{s}}$$

And we will leave some further learning here, as a continuation of our discussion from chapter 7, regarding the Stirling formula for the usual factorials, as an exercise. \square

Many other things can be said about the gamma function, some in relation with the usual factorial of the integers, and some other being complex analysis specific. We will be back to this, in what follows. For the moment, what we have in the above will do.

15b. Complex zeta

Getting now to what we wanted to do, studying the zeta function in the complex plane, we can formulate a key result, making the connection with gamma, as follows:

THEOREM 15.7. *We have the following formula,*

$$\zeta(s) = \frac{1}{\Gamma(s)} \int_0^\infty \frac{x^{s-1}}{e^x - 1} dx$$

valid for any $s \in \mathbb{C}$ with $\operatorname{Re}(s) > 1$.

PROOF. We have indeed the following computation:

$$\begin{aligned} \int_0^\infty \frac{x^{s-1}}{e^x - 1} dx &= \int_0^\infty \frac{x^{s-1}}{e^x} \cdot \frac{1}{1 - e^{-x}} dx \\ &= \int_0^\infty x^{s-1} (e^{-x} + e^{-2x} + e^{-3x} + \dots) \\ &= \sum_{n=1}^\infty \int_0^\infty x^{s-1} e^{-nx} dx \\ &= \sum_{n=1}^\infty \int_0^\infty \left(\frac{y}{n}\right)^{s-1} e^{-y} \frac{dy}{n} \\ &= \sum_{n=1}^\infty \frac{1}{n^s} \int_0^\infty y^{s-1} e^{-y} dy \\ &= \zeta(s) \Gamma(s) \end{aligned}$$

Thus, we are led to the formula in the statement. \square

The above result is something quite subtle, relating three interesting quantities, none of which is something trivial, namely zeta, gamma, and the integral on the right. Generally

speaking, the idea in what follows will be that of using this in order to get information about zeta, from the knowledge of gamma, and of the integral on the right.

This being said, let us start with a piece of reverse technology. We know from Theorem 15.1 the values of zeta at the even positive integers, $\zeta(2k)$ with $k \in \mathbb{N}$, and we also know from Theorem 15.2 the values of gamma at these integers. Thus, Theorem 15.7 can only provide us, at $s = 2k$, with an interesting analysis result. Here that result is:

PROPOSITION 15.8. *We have the following formula for the Bernoulli numbers,*

$$B_{2k} = 4k(-1)^{k+1} \int_0^\infty \frac{t^{2k-1}}{e^{2\pi t} - 1} dt$$

coming from our previous computation of the values $\zeta(2k)$, with $k \in \mathbb{N}$.

PROOF. We have indeed the following computation, using Theorem 15.7:

$$\begin{aligned} \zeta(2k) &= \frac{1}{\Gamma(2k)} \int_0^\infty \frac{x^{2k-1}}{e^x - 1} dx \\ &= \frac{1}{(2k-1)!} \int_0^\infty \frac{x^{2k-1}}{e^x - 1} dx \\ &= \frac{1}{(2k-1)!} \int_0^\infty \frac{(2\pi t)^{2k-1}}{e^{2\pi t} - 1} 2\pi dt \\ &= \frac{(2\pi)^{2k}}{(2k-1)!} \int_0^\infty \frac{t^{2k-1}}{e^{2\pi t} - 1} dt \end{aligned}$$

On the other hand, recall from Theorem 15.1 that we have the following formula:

$$\zeta(2k) = (-1)^{k+1} \frac{(2\pi)^{2k} B_{2k}}{2 \cdot (2k)!}$$

Thus, forgetting now about zeta, we have the following equality:

$$(-1)^{k+1} \frac{(2\pi)^{2k} B_{2k}}{2 \cdot (2k)!} = \frac{(2\pi)^{2k}}{(2k-1)!} \int_0^\infty \frac{t^{2k-1}}{e^{2\pi t} - 1} dt$$

We are therefore led to the formula in the statement. □

As a comment here, the above result, or rather the computation in its proof, raises the possibility of recovering our previous computation of $\zeta(2k)$, with $k \in \mathbb{N}$, done the hard way in chapter 14, using Fourier analysis, by computing the real integral in Proposition 15.8, by whatever means, and getting as answer that Bernoulli number, as indicated.

However, and here comes the point, this is not really possible, at least with elementary methods, because as we perfectly know from chapter 14, what we have at $k = 1$ is the Basel formula, which notoriously cannot be proved with bare hands. Thus, nice try, and in the end, our approach to these questions, since chapter 14, remains the best one.

Moving on, still with inspiration from Theorem 15.7, let us formulate:

QUESTION 15.9. *In analogy with what happens for the gamma function, can we have the zeta function defined, over the whole complex plane?*

We will see in what follows that the answer to this question is yes. However, things will be substantially more complicated for zeta than for gamma, and we will have to rely on delicate complex analysis, and in particular, on the theory of analytic continuation.

Now with analytic continuation being somehow Devil's business, our plan in what follows will be to explain all this slowly, by gradually going from the analytic right half-plane $\operatorname{Re}(s) > 1$, where we understand well the zeta function, to other parts of \mathbb{C} .

Getting started now with our exploratory trip West, and make sure that you have enough food, water and weapons, let us first see what happens at $s = 1$. Here we have:

THEOREM 15.10. *We have the following formula,*

$$\lim_{s \rightarrow 1} (s - 1)\zeta(s) = 1$$

showing that the complex zeta has a simple pole at $s = 1$, with residue 1.

PROOF. We have the following computation, using $\Gamma(1) = 1$:

$$\begin{aligned} \lim_{s \rightarrow 1} (s - 1)\zeta(s) &= \lim_{s \rightarrow 1} (s - 1) \int_0^\infty \frac{x^{s-1}}{e^x - 1} dx \\ &= \lim_{t \rightarrow 0} \int_0^\infty \frac{tx^t}{e^x - 1} dx \\ &= 1 \end{aligned}$$

Thus, we are led to the conclusions in the statement. □

As a more advanced result now, on the same topic, we have:

THEOREM 15.11. *We have the following formula,*

$$\lim_{s \rightarrow 1} \left| \zeta(s) - \frac{1}{s - 1} \right| = \gamma$$

with $\gamma = 0.57721\dots$ being the Euler-Mascheroni constant.

PROOF. This is something quite advanced, well beyond what we have in Theorem 15.10, and that we will not explain here in detail, the idea with this being as follows:

(1) To start with, let us recall from chapter 7 that the Euler-Mascheroni constant, that we met there in relation with the Mertens theorems, is defined as follows:

$$\gamma = \lim_{N \rightarrow \infty} \sum_{n=1}^N \frac{1}{n} - \log N$$

(2) Alternatively, as also explained in chapter 7, we have the following formula for γ , with $[.]$ being the integer part, which proves right away the convergence:

$$\gamma = \int_1^{\infty} \left(\frac{1}{[x]} - \frac{1}{x} \right) dx$$

(3) The point now is that the Euler-Mascheroni constant, as constructed above, can be shown to be related to the zeta function by the following remarkable formula:

$$\gamma = \sum_{n=2}^{\infty} (-1)^n \frac{\zeta(n)}{n}$$

(4) Next, and getting now to what we wanted to do, namely the study of zeta around $s = 1$, the point is that we have as well the following formula:

$$\gamma = \lim_{s \rightarrow 1^+} \sum_{n=1}^{\infty} \frac{1}{n^s} - \frac{1}{s-1}$$

(5) But in terms of the zeta function, this latter formula simply reads:

$$\gamma = \lim_{s \rightarrow 1^+} \zeta(s) - \frac{1}{s-1}$$

(6) Thus, we are led to the formula in the statement. Note that we have as well:

$$\gamma = \lim_{s \rightarrow 0} \frac{\zeta(1+s) + \zeta(1-s)}{2}$$

Indeed, this follows from the formula in the statement, applied twice. \square

Leaving aside now $s = 1$, let us focus on the other points, $s = 1 + it$ with $t \neq 0$, of the boundary line $\operatorname{Re}(s) = 1$, between known and unknown. And here, we have:

CLAIM 15.12. *The Riemann zeta function, namely*

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

converges at any $s = 1 + it$ with $t \neq 0$.

To be more precise here, this claim, which is something quite remarkable, states that the following series of complex numbers must converge, at any $t \neq 0$:

$$\zeta(1 + it) = \sum_{n=1}^{\infty} \frac{1}{n^{1+it}}$$

As a first comment on this, this series will obviously not converge absolutely, and this because the associated series of absolute values is $\zeta(1) = \infty$. However, what happens is that the n^{it} factors, which vary on the unit circle \mathbb{T} , will slow down the divergence of the series $\zeta(1) = \infty$, up to the point of making the sum converge, as stated.

To be more precise, we have the following computation, to start with:

$$\begin{aligned}
 \zeta(1+it) &= \sum_{n=1}^{\infty} \frac{1}{n^{1+it}} \\
 &= \sum_{n=1}^{\infty} \frac{1}{n e^{it \log n}} \\
 &= \sum_{n=1}^{\infty} \frac{e^{-it \log n}}{n} \\
 &= \sum_{n=1}^{\infty} \frac{\cos(t \log n) - i \sin(t \log n)}{n}
 \end{aligned}$$

Thus, we can see that our series seems indeed to converge, at $t \neq 0$. However, in practice, proving such things is certainly possible, but not exactly trivial, and we prefer to defer the discussion here, for a bit later, when we will know more things.

With this discussed, let us get now into the true unknown, $\operatorname{Re}(s) < 1$. Our first objective will be that of understanding what happens in the strip $0 < \operatorname{Re}(s) < 1$. And here, as a main source of inspiration, we have the following very classical result:

PROPOSITION 15.13. *Unlike the standard Riemann series, which diverges,*

$$\zeta(1) = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \dots = \infty$$

the signed version of this series, called standard Dirichlet series, converges,

$$\eta(1) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \dots < \infty$$

and we can even compute its value, $\eta(1) = \log 2$.

PROOF. This is indeed something very classical, that we know since chapter 5, the point being that the convergence of the series $\eta(1)$ can be proved in a variety of ways, all instructive, as for instance by grouping terms and comparing to $\zeta(2) < \infty$:

$$\eta(1) = \frac{1}{2} + \frac{1}{12} + \frac{1}{30} + \frac{1}{56} + \dots < \zeta(2) < \infty$$

As for the exact formula of $\eta(1)$, this follows from the Taylor formula for \log :

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \frac{x^6}{6} + \dots$$

Indeed, by plugging in $x = 1$, we obtain the formula in the statement. □

Now with Proposition 15.13 in hand, we have our idea, namely “forcing” the zeta function to converge in the strip $0 < \operatorname{Re}(s) < 1$, by adding signs, and then recovering zeta, or rather its analytic continuation, in this same strip, by removing the signs.

In practice, this simple idea leads to the following remarkable result:

THEOREM 15.14. *We have the following formula,*

$$\zeta(s) = \frac{1}{1 - 2^{1-s}} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s}$$

which can stand as definition for ζ , in the strip $0 < \operatorname{Re}(s) < 1$.

PROOF. This is something elementary, known since Dirichlet and Euler, but of key importance, and with many consequences, the idea being as follows:

(1) To start with, we can define the Dirichlet function η as being the signed version of ζ , exactly as we did in Proposition 15.13 at $s = 1$, as follows:

$$\eta(s) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s}$$

Observe that this function converges indeed in the strip $0 < \operatorname{Re}(s) < 1$.

(2) We must now connect ζ and η , at $\operatorname{Re}(s) > 1$, and this can be done as follows:

$$\begin{aligned} \zeta(s) + \eta(s) &= \sum_{n=1}^{\infty} \frac{1}{n^s} + \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s} \\ &= 2 \sum_{k=1}^{\infty} \frac{1}{(2k)^s} \\ &= 2^{1-s} \sum_{k=1}^{\infty} \frac{1}{k^s} \\ &= 2^{1-s} \zeta(s) \end{aligned}$$

(3) But this gives the following formula, valid at any exponent $s \in \mathbb{C}$ satisfying $\operatorname{Re}(s) > 1$, and which is the formula in the statement:

$$\zeta(s) = \frac{1}{1 - 2^{1-s}} \eta(s)$$

(4) In order now to conclude, we can invoke the theory of analytic continuation. Skipping some details here, and we refer for instance to Rudin [74] for this, what we have in the statement is a formula for ζ in the whole right half-plane, $\operatorname{Re}(s) > 0$, which is analytic, and more specifically meromorphic, with a single pole, at $s = 1$, and which coincides with the usual formula of ζ on the usual domain of definition, $\operatorname{Re}(s) > 1$.

(5) But, in this situation, the theory of analytic continuation tells us that we can redefine ζ all over the right half-plane, $\operatorname{Re}(s) > 0$, by the formula in the statement, and with this extension being unique, as per the general properties of the meromorphic functions. Thus, we are led to the various conclusions in the statement. \square

Getting back now to Claim 15.12, with basically the same technology we can prove that too. The result, coming as a useful complement to Theorem 15.14, is as follows:

THEOREM 15.15. *The Riemann zeta function, namely*

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

converges at any $s = 1 + it$ with $t \neq 0$.

PROOF. As mentioned, we can have this proved using our technology, as follows:

(1) Let us first go back to Theorem 15.14 and its proof. The method there will work at $\operatorname{Re}(s) = 1$ too, provided that the denominator appearing there does not vanish:

$$1 - 2^{1-s} \neq 0$$

In order to discuss this, observe that with $s = r + it$ we have the following formula:

$$2^{1-s} = 2^{1-r-it} = 2^{1-r} 2^{-it} = 2^{1-r} e^{-it \log 2}$$

We conclude from this that we have the following equivalence:

$$2^{1-s} = 1 \iff s = 1 + i \frac{2k\pi}{\log 2}, \quad k \in \mathbb{Z}$$

(2) Thus, there are no problems indeed in the strip $0 < \operatorname{Re}(s) < 1$, and in relation with what happens at $\operatorname{Re}(s) = 1$, things are fine as long as $s = 1 + it$ is not as above.

(3) As for the remaining case, namely $t = 2k\pi/\log 2$ with $k \in \mathbb{Z} - \{0\}$, here our method does not apply as stated, but with a bit of extra work and modifications, we can have this working too. We will leave some study and learning here as an exercise. \square

Summarizing, we have zeta up and working at $\operatorname{Re}(s) > 0$. We will see later that, in relation with certain delicate number theory questions, this extension is in fact enough, with that number theory questions taking place in the strip $0 < \operatorname{Re}(s) < 1$.

15c. Riemann formula

Getting now to the left half-plane, $\operatorname{Re}(s) < 0$, things will be quite complicated here, and we will start our discussion with a warning, as follows:

WARNING 15.16. *Contrary to what we have been doing so far in this chapter, understanding zeta over the whole \mathbb{C} requires an excellent knowledge of complex analysis.*

In short, many things to be learned, and there will be some technical details missing, in what comes next, and in case you are not familiar with all the needed preliminaries, which is probably the case, well, relax, take this as a physics class, and enjoy.

Getting now to the problem that we have, to be solved, namely extending zeta to the left half-plane, $\operatorname{Re}(s) < 0$, there are many methods available here, and with the main method, due to Riemann himself, which is something quite tough, but unavoidable for understanding the zeta function as a whole, over \mathbb{C} , being as follows:

THEOREM 15.17. *We have the following formula of Riemann, relating the values of zeta at s and $1 - s$,*

$$\zeta(s) = 2^s \pi^{s-1} \sin\left(\frac{\pi s}{2}\right) \Gamma(1-s) \zeta(1-s)$$

which holds on the strip $0 < \operatorname{Re}(s) < 1$, and can serve as definition for zeta in the left half-plane, $\operatorname{Re}(s) < 0$, by analytic continuation.

PROOF. This is something subtle, with even understanding the statement being non-trivial business, and with the proof being complicated too, the idea being as follows:

(1) To start with, let us check our formula for mistakes. With $\operatorname{Re}(s) > 1$ our formula tells us that the familiar $\zeta(s)$ can be expressed in terms of some virtual number $\zeta(1-s)$, which remains to be defined later, and normally no problem with this.

(2) However, looking more carefully, there might be a problem coming from the sine, which vanishes at $s = 2k$ with $k \in \mathbb{N}$. But, the point is that $\Gamma(1-s)$ has a pole at $s = 2k$, compensating for this vanishing of the sine. So, as a conclusion here, not only we avoided the contradictory $\zeta(2k) = 0$, but also know that, later when it will come to discuss $\zeta(1-2k)$, that will be a usual complex number, with no need for a pole there.

(3) Conversely now, let us plug in numbers with $\operatorname{Re}(s) < 0$, so that $\operatorname{Re}(1-s) > 1$. Here what our formula tells us is that the familiar $\zeta(1-s)$, when multiplied by the quantities in the statement, produces a candidate $\zeta(s)$ for the analytic continuation in the left half-plane $\operatorname{Re}(s) < 0$. So, very good, no contradiction whatsoever here, and in addition this tells us, confirming the finding in (2), that zeta will have no poles at $\operatorname{Re}(s) < 0$.

(4) Now let us have a look at the strip $0 < \operatorname{Re}(s) < 1$. Here our function ζ is already existent, thanks to Theorem 15.14, and we have something to prove, namely that the Riemann formula in the statement holds indeed, in this strip $0 < \operatorname{Re}(s) < 1$.

(5) But this is something that can be proved indeed, via some non-trivial calculus, done by Riemann a long time ago, and which has been barely simplified, since. In order to get started, we can use the following formula for the gamma function:

$$\Gamma\left(\frac{s}{2}\right) = n^s \pi^{\frac{s}{2}} \int_0^\infty x^{\frac{s}{2}-1} e^{-n^2 \pi x} dx$$

(6) Thus, we are led to the following formula for the zeta function:

$$\begin{aligned}\Gamma\left(\frac{s}{2}\right)\zeta(s) &= \pi^{\frac{s}{2}} \sum_{n=1}^{\infty} \int_0^{\infty} x^{\frac{s}{2}-1} e^{-n^2\pi x} dx \\ &= \pi^{\frac{s}{2}} \int_0^{\infty} x^{\frac{s}{2}-1} \sum_{n=1}^{\infty} e^{-n^2\pi x} dx\end{aligned}$$

(7) Now let us call Ψ the function appearing on the right, namely:

$$\Psi(x) = \sum_{n=1}^{\infty} e^{-n^2\pi x}$$

With this convention, the formula that we found can be written as follows:

$$\pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \zeta(s) = \int_0^{\infty} x^{\frac{s}{2}-1} \Psi(x) dx$$

(8) Now let us have a look at the function Ψ . By Poisson summation we obtain:

$$\sum_{n=-\infty}^{\infty} e^{-n^2\pi x} = \frac{1}{\sqrt{x}} \sum_{n=-\infty}^{\infty} e^{-\frac{n^2\pi}{x}}$$

We conclude that our function Ψ satisfies the following equation:

$$2\Psi(x) + 1 = \frac{1}{\sqrt{x}} \left(2\Psi\left(\frac{1}{x}\right) + 1 \right)$$

(9) With this equation in hand, let us go back to the formula for zeta in (7). We can further process that formula, in the following way:

$$\begin{aligned}\pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \zeta(s) &= \int_0^{\infty} x^{\frac{s}{2}-1} \Psi(x) dx \\ &= \int_0^1 x^{\frac{s}{2}-1} \Psi(x) dx + \int_1^{\infty} x^{\frac{s}{2}-1} \Psi(x) dx \\ &= \int_0^1 x^{\frac{s}{2}-1} \left(\frac{1}{\sqrt{x}} \Psi\left(\frac{1}{x}\right) + \frac{1}{2\sqrt{2}} - \frac{1}{2} \right) dx + \int_1^{\infty} x^{\frac{s}{2}-1} \Psi(x) dx \\ &= \frac{1}{s-1} + \frac{1}{s} + \int_0^1 x^{\frac{s-3}{2}} \Psi\left(\frac{1}{x}\right) dx + \int_1^{\infty} x^{\frac{s}{2}-1} \Psi(x) dx\end{aligned}$$

(10) We conclude from this that we have the following formula:

$$\pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \zeta(s) = \frac{1}{s(s-1)} + \int_1^{\infty} \left(x^{-\frac{s+1}{2}} + x^{\frac{s}{2}-1} \right) \Psi(x) dx$$

Now since the expression on the right is invariant under $s \rightarrow 1 - s$, we obtain:

$$\pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \zeta(s) = \pi^{-\frac{1-s}{2}} \Gamma\left(\frac{1-s}{2}\right) \zeta(1-s)$$

But with this we are done, because this latter formula is equivalent to the Riemann symmetry formula in the statement, namely:

$$\zeta(s) = 2^s \pi^{s-1} \sin\left(\frac{\pi s}{2}\right) \Gamma(1-s) \zeta(1-s)$$

(11) Next, there is some discussion at the border of the strip too, with the formula relating the values at $Re(s) = 1$, all finite except for a pole at $s = 1$, to the values at $Re(s) = 0$, which all follow to be finite, thanks to the mechanism explained in (2).

(12) Now with this done, we can take the formula in the statement as a definition for zeta in the left half-plane, $Re(s) < 0$, and with the general theory of analytic continuation telling us, a bit like before, at the end of the proof of Theorem 15.14, that this continuation is unique, thanks to the general properties of the meromorphic functions.

(13) So, this was for the idea of the proof, and in practice, there are of course many details still in need to be checked, and we will leave this as an instructive exercise. \square

Observe that, in what regards the Riemann formula itself, this remains a key symmetry formula of our newly defined zeta function, as a meromorphic function over \mathbb{C} .

All the above starts to be a bit heavy, and as a summary of all this, we have:

THEOREM 15.18. *We can talk about the Riemann zeta, as a meromorphic function $\zeta : \mathbb{C} \rightarrow \mathbb{C}$, with a single pole, at $s = 1$ with residue 1. At $Re(s) > 1$ we have*

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

and more generally at $Re(s) > 0$ we have the following formula:

$$\zeta(s) = \frac{1}{1-2^{1-s}} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s}$$

Also, the values of zeta at any s and $1-s$ are related by the Riemann formula

$$\zeta(s) = 2^s \pi^{s-1} \sin\left(\frac{\pi s}{2}\right) \Gamma(1-s) \zeta(1-s)$$

with Γ being as usual the gamma function.

PROOF. This is a summary of our various findings from Theorems 15.14 and 15.17 and their proofs, and with the thing to be always kept in mind, when dealing with all this, being that the formula at $Re(s) > 0$ generalizes indeed the formula at $Re(s) > 1$, thanks to a trivial computation, explained in the proof of Theorem 15.14. \square

Getting back now to the Riemann formula from Theorem 15.17, passed the technical difficulties for establishing it, this is something very beautiful and useful, with a lot of symmetry in it, making it clear that the strip $0 < \operatorname{Re}(s) < 1$ is what matters, and that the vertical axis $\operatorname{Re}(s) = 1/2$ is where interesting things should happen.

As a consequence of the Riemann formula, we have the following version of it:

THEOREM 15.19. *We have the following version of the Riemann formula,*

$$\pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \zeta(s) = \pi^{-\frac{1-s}{2}} \Gamma\left(\frac{1-s}{2}\right) \zeta(1-s)$$

symmetric in $s, 1-s$, which is in fact equivalent to it.

PROOF. The above formula is indeed equivalent to the one in Theorem 15.17, and is in fact what comes out from computations, when proving Theorem 15.17. \square

In practice, the quantity in Theorem 15.19 is best normalized as follows:

THEOREM 15.20. *The following function, called ξ function,*

$$\xi(s) = \frac{s(s-1)}{2} \pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \zeta(s)$$

satisfies $\xi(s) = \xi(1-s)$.

PROOF. Again, the above Riemann formula is equivalent to the previous ones, with the function ξ being what is used in computations, when proving Theorem 15.17. \square

So long for the story of the Riemann reflection formula. The above presentation was of course quite brief, and for more, the starting point is the general theory of analytic continuation read in detail, say from Rudin [74], and then the particular case of zeta to be further learned, from any standard, reasonably advanced number theory book.

15d. Further results

We have zeta up and working in the full complex plane \mathbb{C} , as a meromorphic function with a single pole at 1, and this gives rise to many interesting questions. To start with, regarding the analytic continuation, by other means, the situation is as follows:

(1) A first formula, due to Hasse, which works at any $s \neq 1$, is as follows:

$$\zeta(s) = \frac{1}{1-2^{1-s}} \sum_{n=0}^{\infty} \frac{1}{2^{n+1}} \sum_{k=0}^n \binom{n}{k} \frac{(-1)^k}{(k+1)^s}$$

(2) A second formula, due to Hasse too, which again works at any $s \neq 1$, is:

$$\zeta(s) = \frac{1}{s-1} \sum_{n=0}^{\infty} \frac{1}{n+1} \sum_{k=0}^n \binom{n}{k} \frac{(-1)^k}{(k+1)^{s-1}}$$

(3) We also have the following version, nicer, but working only at $Re(s) > 0$:

$$\zeta(s) = \frac{1}{s-1} \sum_{n=1}^{\infty} \left(\frac{n}{(n+1)^s} - \frac{n-s}{n^s} \right)$$

(4) But we can modify this latter formula as follows, as to have it at $Re(s) > -1$:

$$\zeta(s) = \frac{1}{s-1} \sum_{n=1}^{\infty} \frac{n(n+1)}{2} \left(\frac{2n+3+s}{(n+1)^{s+2}} - \frac{2n-1-s}{n^{s+2}} \right)$$

(5) And so on, the idea being that we can conquer the whole left half-plane $Re(s) < 0$ in this way, step by step, with at each step a more complicated formula being needed.

Getting now to a second question, other general formulae satisfied by zeta, there are many of them. To start with, we can write a Laurent series expansion, as follows:

$$\zeta(s) = \frac{1}{s-1} + \sum_{n=0}^{\infty} \frac{\gamma_n}{n!} (1-s)^n$$

The Laurent coefficients are the Euler-Mascheroni constant $\gamma_0 = \gamma$, and:

$$\gamma_n = \lim_{m \rightarrow \infty} \left[\left(\sum_{k=1}^m \frac{(\log k)^n}{k} \right) - \frac{(\log m)^{n+1}}{n+1} \right]$$

We also have the following formula, involving generalized binomial coefficients:

$$\frac{\zeta(s)}{s} = \frac{1}{s-1} - \sum_{n=1}^{\infty} \binom{n+s-1}{n+1} (\zeta(s+n) - 1)$$

Getting now to a third question, special values of zeta, we have already seen the formulae of $\zeta(2k)$ with $k \in \mathbb{N}$, the idea being these can be recaptured from:

$$\sum_{k=1}^{\infty} \zeta(2k) x^{2k} = \frac{1 - \pi x \cot(\pi x)}{2}$$

In practice, we have the following formula, with B_n being the Bernoulli numbers:

$$\zeta(2k) = (-1)^{k+1} \frac{(2\pi)^{2k} B_{2k}}{2 \cdot (2k)!}$$

Now by Riemann reflection, we obtain from this the following remarkable result:

THEOREM 15.21. *We have the following formula, for any $n \in \mathbb{N}$,*

$$\zeta(-n) = (-1)^n \frac{B_{n+1}}{n+1}$$

with B_n being as usual the Bernoulli numbers.

PROOF. We recall that the Riemann reflection formula from Theorem 15.18 was as follows, valid all over the complex plane, as an equality of meromorphic functions:

$$\zeta(s) = 2^s \pi^{s-1} \sin\left(\frac{\pi s}{2}\right) \Gamma(1-s) \zeta(1-s)$$

(1) By plugging in the value $s = -2k$, with $k \geq 1$ integer, we obtain:

$$\begin{aligned} \zeta(-2k) &= 2^{-2k} \pi^{-2k-1} \sin(-k\pi) \Gamma(1+2k) \zeta(1+2k) \\ &= 0 \end{aligned}$$

Now since we have $B_{2k+1} = 0$, this gives the formula in the statement, at $n = 2k$.

(2) By plugging in the value $s = 1 - 2k$, with $k \geq 1$ integer, we obtain:

$$\begin{aligned} \zeta(1-2k) &= 2^{1-2k} \pi^{-2k} \sin\left(k\pi + \frac{\pi}{2}\right) \Gamma(2k) \zeta(2k) \\ &= \frac{2}{(2\pi)^{2k}} (-1)^k (2k-1)! (-1)^{k+1} \frac{(2\pi)^{2k} B_{2k}}{2 \cdot (2k)!} \\ &= -\frac{B_{2k}}{2k} \end{aligned}$$

Thus, we obtain the formula in the statement, at $n = 2k - 1$. \square

Regarding now the values $\zeta(2k+1)$ with $k \in \mathbb{N}$, things here are quite complicated, starting with the Apéry constant, which is as follows, not computable:

$$\zeta(3) = 1.20205..$$

However, there are many interesting formulae relating the numbers $\zeta(2k+1)$, or more generally the numbers $\zeta(n)$, between themselves. We first have:

$$\begin{aligned} \sum_{k=2}^{\infty} (\zeta(k) - 1) &= 1 \quad , \quad \sum_{k=1}^{\infty} (\zeta(2k) - 1) = \frac{3}{4} \\ \sum_{k=1}^{\infty} (\zeta(2k+1) - 1) &= \frac{1}{4} \quad , \quad \sum_{k=2}^{\infty} (-1)^k (\zeta(k) - 1) = \frac{1}{2} \end{aligned}$$

Along the same lines, a second series of formulae is as follows:

$$\begin{aligned} \sum_{k=1}^{\infty} (-1)^k \frac{\zeta(k)}{k} &= 0 \quad , \quad \sum_{k=1}^{\infty} \frac{\zeta(k) - 1}{k} = 0 \\ \sum_{k=2}^{\infty} (-1)^k \frac{\zeta(k)}{k} &= \gamma \quad , \quad \sum_{k=2}^{\infty} \frac{\zeta(k) - 1}{k} = 1 - \gamma \end{aligned}$$

And there are many more such formulae computing or relating the values of zeta at positive integers, more specialized, and quite often Ramanujan-looking.

Getting now to the zeroes of zeta, as a consequence of Theorem 15.18, we have:

THEOREM 15.22. *We have the following formula, for any integer $k \geq 1$,*

$$\zeta(-2k) = 0$$

and these zeroes, called trivial, are the only ones outside the strip $0 \leq \operatorname{Re}(s) \leq 1$.

PROOF. This follows indeed from the Riemann reflection formula, namely:

$$\zeta(s) = 2^s \pi^{s-1} \sin\left(\frac{\pi s}{2}\right) \Gamma(1-s) \zeta(1-s)$$

(1) To start with, zeta has no zeroes at $\operatorname{Re}(s) > 1$, with this coming from:

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}$$

(2) But this shows, via Riemann reflection, that zeta has no zeroes either at $\operatorname{Re}(s) < 0$, except perhaps at the values $s = -2k$ with $k \geq 1$, where the sine vanishes.

(3) Finally, we have indeed $\zeta(-2k) = 0$, trivially, as we already know from the proof of Theorem 15.21. Thus, we are led to the conclusions in the statement. \square

Which brings us to the question, what are the zeroes of zeta in the critical strip. Many things can be said here, and as a key result, useful for many purposes, we have:

THEOREM 15.23. *The Riemann zeta function has no zero on the line*

$$\operatorname{Re}(s) = 1$$

and no zero on the line $\operatorname{Re}(s) = 0$ either.

PROOF. This is something quite tricky, the idea being as follows:

(1) To start with, the $\operatorname{Re}(s) = 0$ result is equivalent to the $\operatorname{Re}(s) = 1$ result, via the Riemann reflection formula. In what follows, we will prove the $\operatorname{Re}(s) = 1$ result.

(2) In order to study now the zeta function on the line $\operatorname{Re}(s) = 1$, we can use the Euler product formula for this function, coming from Theorem 15.1, namely:

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}$$

(3) Indeed, by taking the logarithm, we obtain from this the following formula:

$$\begin{aligned} \log \zeta(s) &= - \sum_p \log \left(1 - \frac{1}{p^s}\right) \\ &= \sum_p \sum_{k=0}^{\infty} \frac{1}{kp^{ks}} \end{aligned}$$

(4) Now with $s = r + it$ as usual, this formula takes the following form:

$$\begin{aligned}
 \log \zeta(s) &= \sum_p \sum_{k=0}^{\infty} \frac{1}{kp^{k(r+it)}} \\
 &= \sum_p \sum_{k=0}^{\infty} \frac{p^{-kit}}{kp^{kr}} \\
 &= \sum_p \sum_{k=0}^{\infty} \frac{e^{-kit \log p}}{kp^{kr}} \\
 &= \sum_p \sum_{k=0}^{\infty} \frac{\cos(kt \log p) - i \sin(kt \log p)}{kp^{kr}}
 \end{aligned}$$

(5) Now remember the following formula, for the complex exponentials:

$$|e^z|^2 = e^z \cdot \overline{e^z} = e^z e^{\bar{z}} = e^{z+\bar{z}} = e^{2\operatorname{Re}(z)}$$

Thus we have $|e^z| = e^{\operatorname{Re}(z)}$, and by using this with $z = \log \zeta(s)$, we get:

$$\begin{aligned}
 |\zeta(s)| &= |\exp(\log \zeta(s))| \\
 &= \exp(\operatorname{Re}(\log \zeta(s))) \\
 &= \exp\left(\sum_p \sum_{k=0}^{\infty} \frac{\cos(kt \log p)}{kp^{kr}}\right)
 \end{aligned}$$

(6) In order to get an estimate, we use the following formula, valid for any $\alpha \in \mathbb{R}$:

$$\begin{aligned}
 2(1 + \cos \alpha)^2 &= 2 + 4 \cos \alpha + 2 \cos^2 \alpha \\
 &= 3 + 4 \cos \alpha + \cos(2\alpha)
 \end{aligned}$$

Indeed, by using this, we obtain from the formula in (5) the following estimate:

$$\begin{aligned}
 |\zeta(r)^3 \zeta(r+it)^4 \zeta(r+2it)| &= \exp\left(\sum_p \sum_{k=0}^{\infty} \frac{3 + 4 \cos(kt \log p) + \cos(2kt \log p)}{kp^{kr}}\right) \\
 &= \exp\left(\sum_p \sum_{k=0}^{\infty} \frac{2(1 + \cos(kt \log p))^2}{kp^{kr}}\right) \\
 &\geq 1
 \end{aligned}$$

(7) But with this, we can now finish. Assume indeed by contradiction $\zeta(1+it) = 0$, for some $t \neq 0$, and let us look at the following quantity, in the $r \rightarrow 1^+$ limit:

$$K = \zeta(r)^3 \zeta(r+it)^4 \zeta(r+2it)$$

What happens then in the $r \rightarrow 1^+$ limit is that we have $\zeta(r)^3 \rightarrow \infty$ with triple pole behavior, $\zeta(r + it)^4 \rightarrow 0$ with quadruple zero behavior, and $\zeta(r + 2it) \rightarrow \zeta(2it)$ with analytic behavior. But since $3 < 4$ the quadruple zero will kill the triple pole, and so:

$$\lim_{r \rightarrow 1^+} K = 0$$

But this contradicts the estimate found in (6), and so our theorem is proved. \square

In relation with the above, we have the following conjecture:

CONJECTURE 15.24 (Riemann hypothesis). *The zeroes of zeta in the critical strip*

$$0 < \operatorname{Re}(s) < 1$$

can only appear on the critical line, $\operatorname{Re}(s) = 1/2$.

And we will be back to this conjecture, with more comments, in the next chapter. The point indeed is that, as we will soon discover, all this is related to the primes, and their distribution, with for instance Theorem 15.23 being a key ingredient there.

15e. Exercises

Tough analytic chapter that we had here, and as exercises on this, we have:

EXERCISE 15.25. *Learn also more about the gamma function, and its properties.*

EXERCISE 15.26. *In particular, work out the general version of Stirling, for gamma.*

EXERCISE 15.27. *Experiment with the integrals giving rise to the Bernoulli numbers.*

EXERCISE 15.28. *Learn more about the nature of the simple pole of zeta, at $s = 1$.*

EXERCISE 15.29. *Learn more about the Dirichlet function, and its properties.*

EXERCISE 15.30. *Read, with full details, the theory of analytic continuation.*

EXERCISE 15.31. *Clarify all details, in relation with the analytic continuation of ζ .*

EXERCISE 15.32. *Learn as well about the work of Hasse and others, on ζ .*

As bonus exercise, reiterated, spend more time with ζ , the more the better.

CHAPTER 16

Prime distribution

16a. Back to primes

Time to end this book, and we will discuss here, and no surprise, the primes and their distribution. We already know a number of things regarding this, from chapter 7, following Euler, Mertens and Chebycheff. We will substantially improve this here.

To be more precise, we will see in this chapter, following Hadamard and de la Vallée Poussin, how the Riemann zeta function has its say on all this, by producing a very powerful result, called Prime Number Theorem, which is as follows:

CLAIM 16.1 (Prime Number Theorem). *We have the estimate*

$$\pi(x) \sim \frac{x}{\log x}$$

with $\pi(x)$ being the function counting the primes $p \leq x$.

So, this will be the theorem that we will be proving in this chapter, which is something remarkable and difficult, even by modern standards. We will comment as well on some other known proofs of the Prime Number Theorem, notably with the Selberg proof, not using the zeta function, but which is not trivial either. And finally, we will comment on the Riemann hypothesis, which is related to all this, and to many other things.

Getting to work now, for putting things in context, it is perhaps instructive to start with a brief reminder of what we know from chapter 7. Following Euler, we first have:

THEOREM 16.2 (Euler). *We have the following formula, P being the primes,*

$$\sum_{n=1}^{\infty} \frac{1}{n} = \prod_{p \in P} \left(1 - \frac{1}{p}\right)^{-1}$$

which by standard analysis gives the following estimate,

$$\sum_{p \in P} \frac{1}{p} + \frac{1}{2} > \log \left(\sum_{n=1}^{\infty} \frac{1}{n} \right) = \infty$$

which in turn implies $|P| = \infty$.

PROOF. This is something that we know from chapter 7, but always good to talk about it again. The Euler formula comes from $n = p_1^{a_1} \dots p_k^{a_k}$, as follows:

$$\sum_{n=1}^{\infty} \frac{1}{n} = \prod_{p \in P} \left(1 + \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^3} + \dots \right) = \prod_{p \in P} \left(1 - \frac{1}{p} \right)^{-1}$$

Now observe that the log of the product on the right can be estimated as follows:

$$\begin{aligned} - \sum_{p \in P} \log \left(1 - \frac{1}{p} \right) &= \sum_{p \in P} \frac{1}{p} + \frac{1}{2p^2} + \frac{1}{3p^3} + \frac{1}{4p^4} + \dots \\ &< \sum_{p \in P} \frac{1}{p} + \frac{1}{2p^2} + \frac{1}{2p^3} + \frac{1}{2p^4} + \dots \\ &= \sum_{p \in P} \frac{1}{p} + \frac{1}{2} \sum_{p \in P} \left(\frac{1}{p-1} - \frac{1}{p} \right) \\ &< \sum_{p \in P} \frac{1}{p} + \frac{1}{2} \end{aligned}$$

Thus, we are led to the conclusions in the statement. □

Next, still following Euler, we have the following update of the above result:

THEOREM 16.3 (update). *We have the following formula, with sum over primes,*

$$\sum_{p < N} \frac{1}{p} > \log \log N - C$$

with the constant on the right being as follows, depending on your skills:

- (1) *Basic level:* $C = 1/2$.
- (2) *Intermediate:* $C = \log 2 = 0.69314\dots$
- (3) *Advanced:* $C = \log(\pi^2/6) = 0.49770\dots$

PROOF. We know all this from chapter 7 too, the idea being as follows:

(1) This comes indeed by truncating all the computations in proof of Theorem 16.2, which was based on the factorization $n = p_1^{a_1} \dots p_k^{a_k}$, in a straightforward way.

(2) This comes by using the rival factorization $n = p_1 \dots p_k m^2$, with p_i distinct primes. Observe that this estimate, supposedly at the intermediate level, is weaker than (1).

(3) This follows by further building on (2), by replacing the telescoping estimate $\sum_m 1/m^2 < 2$ used there by exact formula $\sum_m 1/m^2 = \pi^2/6$, also due to Euler. □

Next in chapter 7, we had the three Mertens theorems, which were as follows:

THEOREM 16.4 (Mertens). *We have the following estimates, in the $N \rightarrow \infty$ limit,*

$$\begin{aligned}\sum_{p < N} \frac{\log p}{p} &\cong \log N \\ \sum_{p < N} \frac{1}{p} &\simeq \log \log N + M \\ \sum_{p < N} \log \left(1 - \frac{1}{p}\right) &\simeq -\log \log N - \gamma\end{aligned}$$

$M = 0.26149\dots$ and $\gamma = 0.57721\dots$ being the Mertens and Euler-Mascheroni constants.

PROOF. This is something more technical, improving the Euler estimates, that we discussed in some detail in chapter 7. To be more precise, we have seen there the proof of the second formula, which is the main one. As for the first and third formulae, we have talked a bit about them in chapter 7, and in what regards their formal proof, this can be certainly established directly, via some work, as Mertens did. Alternatively, these estimates follow from the Prime Number Theorem, that we will prove in what follows. \square

As a last thing that we learned in chapter 7, we have the Chebycheff theorems:

THEOREM 16.5 (Chebycheff). *We have the following estimate for the π function,*

$$\pi(N) \approx \frac{N}{\log N}$$

in the sense that the quotient of these quantities is bounded from above, and below.

PROOF. This was something tricky too, still mixing combinatorics and calculus, in the spirit of Mertens, and with both the upper and lower bound being non-trivial, one being the first Chebycheff theorem, and the other being the second Chebycheff theorem. \square

Getting to work now, our tools for proving the Prime Number Theorem, following Hadamard and de la Vallée Poussin, will be, besides the Riemann zeta function ζ , the modified Chebycheff function ψ and the von Mangoldt function Λ . We have:

DEFINITION 16.6. *The modified Chebycheff and von Mangoldt functions are*

$$\psi(x) = \sum_{p^k \leq x} \log p \quad , \quad \Lambda(n) = \begin{cases} \log p & \text{if } n = p^k \\ 0 & \text{otherwise} \end{cases}$$

related by the formulae $\psi(x) = \sum_{n \leq x} \Lambda(n)$ and $\Lambda(n) = \psi(n) - \psi(n-)$.

You might of course ask, why using two functions instead of one. Good point, and in answer, we will see a bit later that, in the context of certain delicate questions, the Chebycheff function and the von Mangoldt function are not exactly the same thing.

In relation with the Prime Number Theorem, that we want to prove, we have:

THEOREM 16.7. *We have the following equivalence,*

$$\pi(x) \sim \frac{x}{\log x} \iff \psi(x) \sim x$$

with the condition on the left being the Prime Number Theorem one.

PROOF. This is something elementary, coming from two estimates, as follows:

(1) In one sense, we have the following basic estimate:

$$\begin{aligned} \psi(x) &= \sum_{p^k \leq x} \log p \\ &= \sum_{p \leq x} \log p \left[\frac{\log x}{\log p} \right] \\ &\leq \sum_{p \leq x} \log x \\ &= \pi(x) \log x \end{aligned}$$

(2) In the other sense, we have the following estimate, valid for any $\varepsilon > 0$:

$$\begin{aligned} \psi(x) &= \sum_{p^k \leq x} \log p \\ &\geq \sum_{x^{1-\varepsilon} \leq p \leq x} \log p \\ &\geq \sum_{x^{1-\varepsilon} \leq p \leq x} (1 - \varepsilon) \log x \\ &= (1 - \varepsilon)(\pi(x) + O(x^{1-\varepsilon})) \log x \end{aligned}$$

Thus, we are led to the equivalence in the statement. □

16b. Prime distribution

In order to estimate now the Chebycheff function ψ , we would need an analytic formula for it. However, finding such a formula is not obvious with bare hands, so let us examine instead the same question for the von Mangoldt function Λ , with the hope that we do have an analytic formula for Λ , that can be translated afterwards in terms of ψ .

And good news, our plan works, with the formula for Λ being as follows:

PROPOSITION 16.8. *The von Mangoldt function satisfies*

$$\sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s} = -(\log \zeta(s))'$$

with ζ being the Riemann zeta function.

PROOF. We use the Euler product formula for zeta, namely:

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}$$

By taking the logarithm, we obtain from this the following formula:

$$\log \zeta(s) = - \sum_p \log \left(1 - \frac{1}{p^s}\right)$$

Now by differentiating, we obtain the following formula:

$$\begin{aligned} (\log \zeta(s))' &= - \sum_p \left(1 - \frac{1}{p^s}\right)^{-1} \frac{d(1 - p^{-s})}{ds} \\ &= \sum_p \left(1 - \frac{1}{p^s}\right)^{-1} \frac{dp^{-s}}{ds} \\ &= - \sum_p \left(1 - \frac{1}{p^s}\right)^{-1} p^{-s} \log p \\ &= - \sum_p \frac{p^s}{p^s - 1} \cdot \frac{1}{p^s} \log p \\ &= - \sum_p \frac{\log p}{p^s - 1} \end{aligned}$$

On the other hand, the sum on the left in the statement is given by:

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s} &= \sum_{n=p^k} \frac{\log p}{n^s} \\ &= \sum_p \log p \sum_{k=1}^{\infty} \frac{1}{p^{ks}} \\ &= \sum_p \log p \cdot \frac{1}{p^s} \left(1 - \frac{1}{p^s}\right)^{-1} \\ &= \sum_p \frac{\log p}{p^s - 1} \end{aligned}$$

Thus, we are led to the equality in the statement. \square

Now let us turn to the second part of our plan, namely reformulating the formula for Λ that we found in terms of ψ . This is something more delicate, leading to:

THEOREM 16.9. *The modified Chebycheff function is given by*

$$\psi(x) = x - \log(2\pi) - \sum_{\zeta(s)=0} \frac{x^s}{s}$$

for $x \notin \mathbb{Z}$, with the sum being over all the zeroes of zeta.

PROOF. This follows via some complex analysis and tricks, as follows:

(1) To start with, we know from Definition 16.6 that the functions ψ and Λ are related by the following conversion formulae, which are both trivial:

$$\psi(x) = \sum_{n \leq x} \Lambda(n) \quad , \quad \Lambda(n) = \psi(n) - \psi(n-)$$

The problem now is to use these conversion formulae, in order to reformulate in terms of ψ the formula for Λ that we found in Proposition 16.8, namely:

$$\sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s} = -(\log \zeta(s))'$$

(2) As a first step, we have the following computation, with at the beginning the $n = 1$ term ignored, and at the end, the $n = 1$ term added, because these vanish anyway:

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s} &= \sum_{n=2}^{\infty} \frac{\psi(n) - \psi(n-)}{n^s} \\ &= \sum_{n=2}^{\infty} \frac{\psi(n) - \psi(n-1)}{n^s} \\ &= \sum_{n=1}^{\infty} \psi(n) \left(\frac{1}{n^s} - \frac{1}{(n+1)^s} \right) \end{aligned}$$

(3) Thus, we have the following equation, in terms of the function ψ :

$$\sum_{n=1}^{\infty} \psi(n) \left(\frac{1}{n^s} - \frac{1}{(n+1)^s} \right) = -(\log \zeta(s))'$$

(4) The problem is now, how to fine-tune this, into something truly analytical, involving the function $\psi(x)$ with real argument, $x > 1$. For this purpose, it is convenient to further modify the Chebycheff step function ψ , by making it continuous, as follows:

$$\varphi(x) = \int_1^x \psi(t) dt$$

(5) Observe that this latter function can be expressed in terms of Λ , as follows:

$$\varphi(x) = \sum_{n \leq x} (x - n) \Lambda(n)$$

Also, as another remark, in relation with Theorem 16.7, we have:

$$\psi(x) \sim x \iff \varphi(x) \sim \frac{x^2}{2}$$

Thus, we can normally do everything with φ instead of ψ . However, for our purposes here, φ will be a secondary object, with our main function remaining ψ .

(6) The point now is that we have the following formula, as a contour integral, with $r > 1$, coming via some manipulations involving the Cauchy formula:

$$\frac{\varphi(x)}{x^2} = \frac{1}{2\pi i} \int_{r-\infty i}^{r+\infty i} \frac{x^{s-1}}{s(s+1)} \sum_{n=1}^{\infty} \psi(n) \left(\frac{1}{n^s} - \frac{1}{(n+1)^s} \right) ds$$

(7) We recognize on the right the sum from (3), and by plugging that in, we get:

$$\begin{aligned} \frac{\varphi(x)}{x^2} &= -\frac{1}{2\pi i} \int_{r-\infty i}^{r+\infty i} \frac{x^{s-1}}{s(s+1)} (\log \zeta(s))' ds \\ &= -\frac{1}{2\pi i} \int_{r-\infty i}^{r+\infty i} \frac{x^{s-1}}{s(s+1)} \cdot \frac{\zeta'(s)}{\zeta(s)} ds \end{aligned}$$

(8) Now since the function $\zeta'(s)/\zeta(s)$ has a simple pole at 1, with residue -1 , we can separate the contribution of that pole, and we get, again with $r > 1$:

$$\frac{\varphi(x)}{x^2} = \frac{1}{2} \left(1 - \frac{1}{x} \right)^2 - \frac{1}{2\pi i} \int_{r-\infty i}^{r+\infty i} \frac{x^{s-1}}{s(s+1)} \left(\frac{\zeta'(s)}{\zeta(s)} + \frac{1}{s-1} \right) ds$$

(9) In order to simplify notation, let us introduce the following function:

$$f(s) = \frac{1}{s(s+1)} \left(\frac{\zeta'(s)}{\zeta(s)} + \frac{1}{s-1} \right)$$

In terms of this function, the formula that we found above reads:

$$\begin{aligned} \frac{\varphi(x)}{x^2} &= \frac{1}{2} \left(1 - \frac{1}{x} \right)^2 - \frac{1}{2\pi i} \int_{r-\infty i}^{r+\infty i} x^{s-1} f(s) ds \\ &= \frac{1}{2} \left(1 - \frac{1}{x} \right)^2 - \frac{1}{2\pi} \int_{-\infty}^{\infty} x^{r+it-1} f(r+it) dt \\ &= \frac{1}{2} \left(1 - \frac{1}{x} \right)^2 - \frac{x^{r-1}}{2\pi} \int_{-\infty}^{\infty} e^{it \log x} f(r+it) dt \end{aligned}$$

(10) Thus, getting back now to the usual Chebycheff function ψ , we have:

$$\frac{1}{x^2} \int_1^x \psi(t) dt = \frac{1}{2} \left(1 - \frac{1}{x} \right)^2 - \frac{x^{r-1}}{2\pi} \int_{-\infty}^{\infty} e^{it \log x} f(r+it) dt$$

By multiplying both sides by x^2 , we have the following formula:

$$\int_1^x \psi(t) dt = \frac{(x-1)^2}{2} - \frac{x^{r+1}}{2\pi} \int_{-\infty}^{\infty} e^{it \log x} f(r+it) dt$$

(11) Now by taking the derivative with respect to x , this formula gives:

$$\begin{aligned} \psi(x) &= \frac{d}{dx} \left[\frac{(x-1)^2}{2} - \frac{x^{r+1}}{2\pi} \int_{-\infty}^{\infty} e^{it \log x} f(r+it) dt \right] \\ &= x - 1 + \frac{d}{dx} \left[\frac{x^{r+1}}{2\pi} \int_{-\infty}^{\infty} e^{it \log x} f(r+it) dt \right] \end{aligned}$$

(12) The point now is that, by computing the derivative on the right, we get:

$$\psi(x) = x - \log(2\pi) - \sum_{\zeta(s)=0} \frac{x^s}{s}$$

Thus, we are led to the conclusion in the statement. \square

Now remember from Theorem 16.7 that what we want to do is to estimate ψ , with the following estimate, proving the Prime Number theorem, being our goal:

$$\psi(x) \sim x$$

Looking at the formula in Theorem 16.9, the x is already there, $\log(2\pi)$ does not matter, and what is left to prove that the sum over zeroes of ζ does not matter either:

$$\sum_{\zeta(s)=0} \frac{x^s}{s} = o(x)$$

In what regards the trivial zeroes, things are easily settled here, as follows:

PROPOSITION 16.10. *The contribution to the modified Chebycheff function ψ of the trivial zeroes of zeta, namely $-2, -4, -6, \dots$, is given by*

$$\sum_{k=1}^{\infty} \frac{x^{-2k}}{2k} = -\frac{1}{2} \log \left(1 - \frac{1}{x^2} \right)$$

and this quantity vanishes in the $x \rightarrow \infty$ limit.

PROOF. We have indeed the following computation:

$$\sum_{k=1}^{\infty} \frac{x^{-2k}}{2k} = \sum_{k=1}^{\infty} \frac{1}{2kx^{2k}} = -\log \left(1 - \frac{1}{x^2} \right)$$

Thus, we are led to the conclusion in the statement. \square

Regarding now the non-trivial zeroes of zeta, we know from chapter 15 that these lie inside the strip $0 \leq \operatorname{Re}(s) \leq 1$, and as a first observation, we have:

PROPOSITION 16.11. *The contribution to the modified Chebycheff function ψ of the non-trivial zeroes of zeta lying in the strip $0 \leq \operatorname{Re}(s) < 1$ satisfies*

$$\sum_{\zeta(s)=0} \frac{x^s}{s} = o(x)$$

so we are left with studying the zeroes on the line $\operatorname{Re}(s) = 1$.

PROOF. This is something quite self-explanatory, with some care needed however when summing all the $o(x)$ quantities associated to the zeroes in question. As for the final conclusion, this comes by combining our finding with Proposition 16.10. \square

But now, in order to finish, recall the following theorem from chapter 15:

THEOREM 16.12. *The Riemann zeta function has no zero on the line*

$$\operatorname{Re}(s) = 1$$

and no zero on the line $\operatorname{Re}(s) = 0$ either.

PROOF. This is something that we learned the hard way in chapter 15, obtained via some non-trivial estimates, and a fine knowledge of analysis, the idea being as follows:

(1) By using the Euler formula, and taking the logarithm, we obtain:

$$\log \zeta(s) = - \sum_p \log \left(1 - \frac{1}{p^s} \right) = \sum_p \sum_{k=0}^{\infty} \frac{1}{kp^{ks}}$$

(2) Next, with $s = r + it$ as usual, this formula takes the following form:

$$\log \zeta(s) = \sum_p \sum_{k=0}^{\infty} \frac{\cos(kt \log p) - i \sin(kt \log p)}{kp^{kr}}$$

(3) By using now the standard formula $|e^z| = e^{\operatorname{Re}(z)}$, with $z = \log \zeta(s)$, we get:

$$|\zeta(s)| = \exp \left(\sum_p \sum_{k=0}^{\infty} \frac{\cos(kt \log p)}{kp^{kr}} \right)$$

(4) Next, some crazy trigonometry gives the following estimate:

$$|\zeta(r)^3 \zeta(r + it)^4 \zeta(r + 2it)| \geq 1$$

(5) But with this, we can now finish. Assume indeed by contradiction $\zeta(1 + it) = 0$, for some $t \neq 0$, and let us look at the following quantity, in the $r \rightarrow 1^+$ limit:

$$K = \zeta(r)^3 \zeta(r + it)^4 \zeta(r + 2it)$$

What happens then in the $r \rightarrow 1^+$ limit is that we have $\zeta(r)^3 \rightarrow \infty$ with triple pole behavior, $\zeta(r + it)^4 \rightarrow 0$ with quadruple zero behavior, and $\zeta(r + 2it) \rightarrow \zeta(2it)$ with analytic behavior. But since $3 < 4$ the quadruple zero will kill the triple pole, and so:

$$\lim_{r \rightarrow 1^+} K = 0$$

But this contradicts the estimate found in (4), and so our theorem is proved. \square

By putting now everything together, we obtain:

THEOREM 16.13 (Prime Number Theorem). *We have*

$$\pi(x) \sim \frac{\log x}{x}$$

in the $x \rightarrow \infty$ limit.

PROOF. This follows by putting everything together, as follows:

- (1) We know from Theorem 16.7 that $\pi(x) \sim x / \log x$ is equivalent to $\psi(x) \sim x$.
- (2) We have in Theorem 16.9 a formula for $\psi(x)$, in terms of the zeroes of zeta.
- (3) Most of these zeroes are taken care of by Propositions 16.10 and 16.11.
- (4) As for the remaining zeroes, there are none, as shown by Theorem 16.12. \square

As already mentioned on several occasions, the Prime Number Theorem established above is something quite powerful, often beating the other estimates about primes that we have been talking about, in this book, starting from chapter 7.

It is actually instructive at this point to go back to chapter 7, and review the material there, with simplified proofs for the estimates that we found, and with proofs for the estimates that we did not prove, by using the Prime Number Theorem.

16c. Selberg formula

As mentioned in the beginning of the present chapter, there are as well some other known proofs of the Prime Number Theorem, which are more modern, notably:

- (1) The now classical Selberg proof, which does not use the zeta function.
- (2) The Newman proof, not using zeta either, and which is a bit shorter.

Both these proofs are instructive, having advantages and disadvantages with respect to the original proof by Hadamard and de la Vallée Poussin, and we will briefly discuss this here. We will be quite short, by leaving many computations as exercises.

Getting started, let us first recall that we have the following result:

PROPOSITION 16.14. *The Prime Number theorem is equivalent to the estimate*

$$\sum_{n \leq x} \Lambda(n) \sim x$$

for the von Mangoldt function Λ .

PROOF. This is something that we already know, from earlier in this chapter, the idea being as follows. Consider the modified Chebycheff and von Mangoldt functions:

$$\psi(x) = \sum_{p^k \leq x} \log p, \quad \Lambda(n) = \begin{cases} \log p & \text{if } n = p^k \\ 0 & \text{otherwise} \end{cases}$$

These functions are related by the following formulae:

$$\psi(x) = \sum_{n \leq x} \Lambda(n), \quad \Lambda(n) = \psi(n) - \psi(n-)$$

Our claim is that we have the following equivalence, with the condition on the left corresponding by definition to the Prime Number theorem, and with the condition on the right being, in view of the above conversion formulae, the one in the statement:

$$\pi(x) \sim \frac{x}{\log x} \iff \psi(x) \sim x$$

But this is something elementary, coming from two estimates explained earlier in this chapter. In one sense, we have indeed the following basic estimate:

$$\psi(x) \leq \sum_{p \leq x} \log x = \pi(x) \log x$$

As for the other sense, here we have the following estimate, valid for any $\varepsilon > 0$:

$$\psi(x) \geq \sum_{x^{1-\varepsilon} \leq p \leq x} (1 - \varepsilon) \log x = (1 - \varepsilon)(\pi(x) + O(x^{1-\varepsilon})) \log x$$

Thus, we are led to the equivalence in the statement. \square

In order to deal now with the estimate from Proposition 16.14, as a starting point, we will use the following elementary estimate for Λ , with different weights:

THEOREM 16.15. *We have the following estimate,*

$$\sum_{n \leq x} \Lambda(n) \left[\frac{x}{n} \right] = x \log x - x + O(\log x)$$

for any $x \geq 2$.

PROOF. This is something quite standard, the idea being as follows:

(1) To start with, we have the following formula, for the quantity on the left:

$$\begin{aligned} \sum_{n \leq x} \Lambda(n) \left[\frac{x}{n} \right] &= \sum_{n \leq x} \sum_{d|n} \Lambda(d) \\ &= \sum_{n \leq x} \log n \\ &= \log[x]! \end{aligned}$$

(2) In order to estimate now the quantity that we found above, we can use the Euler summation formula, which is something elementary, as follows, with the usual convention $x = [x] + \{x\}$ for the integer and fractionary parts of $x \in \mathbb{R}$:

$$\sum_{y < n \leq x} f(n) = \int_y^x f(t) dt + \int_y^x \{t\} f'(t) dt - \{x\} f(x) + \{y\} f(y)$$

To be more precise, in the case $y = a$ and $x = a + b$ with a, b integers, the Euler summation formula can be deduced by integrating by parts, as follows:

$$\begin{aligned} \int_a^{a+b} \{t\} f'(t) dt &= \sum_{k=0}^{b-1} \int_{a+k}^{a+k+1} \{t\} f'(t) dt \\ &= \sum_{k=0}^{b-1} \int_{a+k}^{a+k+1} (t - a - k) f'(t) dt \\ &= \sum_{k=0}^{b-1} \left[(t - a - k) f(t) \right]_{a+k}^{a+k+1} - \int_{a+k}^{a+k+1} f(t) dt \\ &= \sum_{k=0}^{b-1} f(a + k + 1) - \int_{a+k}^{a+k+1} f(t) dt \\ &= \sum_{k=0}^{b-1} f(a + k + 1) - \int_a^{a+b} f(t) dt \end{aligned}$$

Indeed, this gives the following formula, which is the Euler one, in this case:

$$\sum_{k=0}^{b-1} f(a + k + 1) = \int_a^{a+b} \{t\} f'(t) dt + \int_a^{a+b} f(t) dt$$

As for the general case, where the bounds $x, y \in \mathbb{R}$ are arbitrary, this follows by making some adjustments in the above computations, both at left and at right, which make appear the extra term $-\{x\}f(x) + \{y\}f(y)$ on the right, as stated.

(3) Getting back now to what we wanted to prove, by using the Euler summation formula for the function $f(t) = \log t$, we obtain the following formula:

$$\begin{aligned}
 \log[x]! &= \sum_{n \leq x} \log n \\
 &= \int_1^x \log t \, dt + \int_1^x \frac{\{t\}}{t} \, dt - \{x\} \log x \\
 &= x \log x - x + 1 + \int_1^x \frac{\{t\}}{t} \, dt + O(\log x) \\
 &= x \log x - x + O\left(\int_1^x \frac{1}{t} \, dt\right) + O(\log x) \\
 &= x \log x - x + O(\log x)
 \end{aligned}$$

Thus, we are led to the estimate in the statement. \square

Now let us have a look at the estimate that we have, from Theorem 16.15, and at the estimate that we want, from Proposition 16.14. The one that we have is:

$$\sum_{n \leq x} \Lambda(n) \left[\frac{x}{n} \right] = x \log x - x + O(\log x)$$

And the one that we want is a bit similar, but with different weights, as follows:

$$\sum_{n \leq x} \Lambda(n) \sim x$$

In view of this, the idea will be that of establishing first some general results, connecting such type of estimates, and then applying them to our situation.

Such general results, connecting sums of a series, with different weights, are called Tauberian theorems, and the one that we need here, due to Shapiro, is as follows:

THEOREM 16.16. *Consider a series $a_n \geq 0$ satisfying the condition*

$$\sum_{n \leq x} a_n \left[\frac{x}{n} \right] = x \log x + O(x)$$

for any $x \geq 1$. We have then the following estimate, for any $x \geq 1$:

$$\sum_{n \leq x} \frac{a_n}{n} = \log x + O(1)$$

Also, we can find constants $A, B > 0$ such that the following estimates hold,

$$Ax \leq \sum_{n \leq x} a_n \leq Bx$$

on the left for $x \gg 0$, and on the right for any $x \geq 1$.

PROOF. This is something quite technical, the idea being as follows:

(1) Consider the following sums, with $a_n \geq 0$ being as in the statement:

$$S(x) = \sum_{n \leq x} a_n \quad , \quad T(x) = \sum_{n \leq x} a_n \left[\frac{x}{n} \right]$$

Our first goal will be that of proving the inequality $S(x) \leq Bx$ at the end. For this purpose, observe first that we have the following estimate:

$$\begin{aligned} T(x) - 2T\left(\frac{x}{2}\right) &= \sum_{n \leq x} a_n \left[\frac{x}{n} \right] - 2 \sum_{n \leq x/2} a_n \left[\frac{x}{2n} \right] \\ &= \sum_{n \leq x/2} a_n \left(\left[\frac{x}{n} \right] - 2 \left[\frac{x}{2n} \right] \right) + \sum_{x/2 < n \leq x} a_n \left[\frac{x}{n} \right] \\ &\geq \sum_{x/2 < n \leq x} a_n \left[\frac{x}{n} \right] \\ &= \sum_{x/2 < n \leq x} a_n \\ &= S(x) - S\left(\frac{x}{2}\right) \end{aligned}$$

Now recall that, in terms of $T(x)$, our assumption on the series a_n was as follows:

$$T(x) = x \log x + O(x)$$

But this gives the following estimate, for the quantity considered above:

$$\begin{aligned} T(x) - 2T\left(\frac{x}{2}\right) &= x \log x + O(x) - 2 \left(\frac{x}{2} \log \frac{x}{2} + O\left(\frac{x}{2}\right) \right) \\ &= x \log x + O(x) - x \log x + O(x) \\ &= O(x) \end{aligned}$$

We conclude that, in terms of $S(x)$, we have the following estimate:

$$S(x) - S\left(\frac{x}{2}\right) = O(x)$$

Now this tells us that we can find a constant $C > 0$ such that, for any $x \geq 1$:

$$S(x) - S\left(\frac{x}{2}\right) \leq Cx$$

By replacing $x \rightarrow x/2$, we conclude that the following estimate must hold too:

$$S\left(\frac{x}{2}\right) - S\left(\frac{x}{4}\right) \leq \frac{Cx}{2}$$

By replacing again $x \rightarrow x/2$, the following estimate must hold too:

$$S\left(\frac{x}{4}\right) - S\left(\frac{x}{8}\right) \leq \frac{Cx}{4}$$

Now by summing all these estimates we obtain, as desired, with $B = 2C$:

$$S(x) \leq Cx + \frac{Cx}{2} + \frac{Cx}{4} + \dots = 2Cx$$

(2) With this done, let us turn now to the proof of the first assertion in the statement. By using the estimate $S(x) \leq Bx$ that we just proved, we have:

$$\begin{aligned} T(x) &= \sum_{n \leq x} a_n \left[\frac{x}{n} \right] \\ &= \sum_{n \leq x} a_n \left(\frac{x}{n} + O(1) \right) \\ &= x \sum_{n \leq x} \frac{a_n}{n} + O\left(\sum_{n \leq x} a_n \right) \\ &= x \sum_{n \leq x} \frac{a_n}{n} + O(S(x)) \\ &= x \sum_{n \leq x} \frac{a_n}{n} + O(x) \end{aligned}$$

But this gives the first assertion, by using our assumption on the series a_n in the statement, which reads $T(x) = x \log x + O(x)$, in the following way:

$$\begin{aligned} \sum_{n \leq x} \frac{a_n}{n} &= \frac{T(x) - O(x)}{x} \\ &= \frac{T(x)}{x} + O(1) \\ &= \frac{x \log x + O(x)}{x} + O(1) \\ &= \log x + O(1) \end{aligned}$$

(3) It remains to prove now the last assertion in the statement, namely the estimate $S(x) \geq Ax$ for $x \gg 0$. For this purpose, consider the following sums:

$$A(x) = \sum_{n \leq x} \frac{a_n}{n}$$

What we just found in (2) can be written as follows, with $E(x) = O(1)$:

$$A(x) = \log x + E(x)$$

Now observe that for any number $c \in (0, 1)$ we have the following estimate, with $M > 0$ being an upper bound for the absolute value of the function $E(x) = O(1)$, and with the variable x being subject to the condition $cx \geq 1$:

$$\begin{aligned} A(x) - A(cx) &= \log x + E(x) - \log(cx) - E(cx) \\ &= E(x) - E(cx) - \log c \\ &\geq -2M - \log c \end{aligned}$$

If we choose the number $c \in (0, 1)$ as for the quantity on the right to be 1, we conclude that we have the following estimate, valid for any $x \geq 1/c$:

$$A(x) - A(cx) \geq 1$$

On the other hand, the quantity on the left is subject to the following estimate:

$$\begin{aligned} A(x) - A(cx) &= \sum_{n \leq x} \frac{a_n}{n} - \sum_{n \leq cx} \frac{a_n}{n} \\ &= \sum_{cx < n \leq x} \frac{a_n}{n} \\ &\leq \frac{1}{cx} \sum_{n \leq x} a_n \\ &= \frac{S(x)}{cx} \end{aligned}$$

We conclude that we have the following estimate, for $x \geq 1/c$:

$$\frac{S(x)}{cx} \geq 1$$

Thus, we are led to the estimate in the statement, $S(x) \geq Ax$, with $A = c$, valid for $x \gg 0$, and more specifically valid for $x \geq 1/c$, with $c \in (0, 1)$ being as above. \square

As a first application of Theorem 16.16, in relation with our questions, we have:

THEOREM 16.17. *We have the following estimate, for any $x \geq 1$:*

$$\sum_{n \leq x} \frac{\Lambda(n)}{n} = \log x + O(1)$$

Also, we can find constants $A, B > 0$ such that the following estimates hold,

$$Ax \leq \sum_{n \leq x} \psi(x) \leq Bx$$

on the left for $x \gg 0$, and on the right for any $x \geq 1$.

PROOF. As a consequence of Theorem 16.15, we have the following estimate:

$$\begin{aligned}\sum_{n \leq x} \Lambda(n) \left[\frac{x}{n} \right] &= x \log x - x + O(\log x) \\ &= x \log x + O(x)\end{aligned}$$

Thus Theorem 16.16 applies with $a_n = \Lambda(n)$ and gives the first formula, namely:

$$\sum_{n \leq x} \frac{\Lambda(n)}{n} = \log x + O(1)$$

Also by Theorem 16.16, we obtain as well estimates as follows, with $A, B > 0$:

$$Ax \leq \sum_{n \leq x} \Lambda(n) \leq Bx$$

Now since we have $\psi(x) = \sum_{n \leq x} \Lambda(n)$, this gives the second assertion. \square

As a second application of Theorem 16.16, also of interest to us, we have:

THEOREM 16.18. *We have the following estimate, for any $x \geq 1$:*

$$\sum_{p \leq x} \frac{\log p}{p} = \log x + O(1)$$

Also, we can find constants $A, B > 0$ such that the following estimates hold,

$$Ax \leq \sum_{n \leq x} \theta(n) \leq Bx$$

on the left for $x \gg 0$, and on the right for any $x \geq 1$.

PROOF. This is something a bit more tricky. It is routine to prove that:

$$\sum_{p \leq x} \log p \left[\frac{x}{p} \right] = x \log x + O(x)$$

In view of this, consider the following version of the von Mangoldt function:

$$\Theta(n) = \begin{cases} \log p & \text{if } n = p \text{ prime} \\ 0 & \text{otherwise} \end{cases}$$

In terms of this function Θ , the above estimate takes the following form:

$$\sum_{n \leq x} \Theta(n) \left[\frac{x}{n} \right] = x \log x + O(x)$$

Thus Theorem 16.16 applies with $a_n = \Theta(n)$ and gives the following formula:

$$\sum_{n \leq x} \frac{\Theta(n)}{n} = \log x + O(1)$$

But this gives the formula in the statement, namely:

$$\sum_{p \leq x} \frac{\log p}{p} = \log x + O(1)$$

Also by Theorem 16.16, we obtain as well estimates as follows, with $A, B > 0$:

$$Ax \leq \sum_{n \leq x} \Theta(n) \leq Bx$$

Now since we have $\theta(x) = \sum_{n \leq x} \Theta(n)$, this gives the second assertion. \square

As yet another application of all this, again of interest to us, we have:

THEOREM 16.19. *We have the following estimate,*

$$\sum_{n \leq x} \psi\left(\frac{x}{n}\right) = x \log x - x + O(\log x)$$

as well as the following estimate,

$$\sum_{n \leq x} \theta\left(\frac{x}{n}\right) = x \log x + O(\log x)$$

both valid for any $x \geq 1$.

PROOF. These results are quite elementary, the idea being as follows:

(1) We recall from Theorem 16.15 that we have the following estimate:

$$\sum_{n \leq x} \Lambda(n) \left[\frac{x}{n} \right] = x \log x - x + O(\log x)$$

But this gives the first formula in the statement, by using:

$$\psi(x) = \sum_{n \leq x} \Lambda(n)$$

(2) As explained in the proof of Theorem 16.18, we have the following estimate:

$$\sum_{n \leq x} \Theta(n) \left[\frac{x}{n} \right] = x \log x + O(x)$$

But this gives the first formula in the statement, by using:

$$\theta(x) = \sum_{n \leq x} \Theta(n)$$

Thus, we are led to the conclusions in the statement. \square

Getting now to what we wanted to prove, the Prime Number theorem, the continuation of the story, due to Selberg, is more complicated. We first have the following result:

PROPOSITION 16.20. *Given a function $G : (0, \infty) \rightarrow \mathbb{C}$, appearing as*

$$G(x) = \log x \sum_{n \leq x} F\left(\frac{x}{n}\right)$$

with $F : (0, \infty) \rightarrow \mathbb{C}$, we have the following equality,

$$F(x) \log x + \sum_{n \leq x} F\left(\frac{x}{n}\right) \Lambda(n) = \sum_{d \leq x} \mu(d) G\left(\frac{x}{d}\right)$$

with μ and Λ being the Möbius and von Mangoldt functions.

PROOF. This is something elementary, coming from the basic properties of μ and Λ . Indeed, we first have the following elementary equality, valid for any function F :

$$F(x) \log x = \sum_{n \leq x} F\left(\frac{x}{n}\right) \log\left(\frac{x}{n}\right) \sum_{d|n} \mu(d)$$

On the other hand, recall that the von Mangoldt function Λ satisfies:

$$\Lambda(n) = \sum_{d|n} \mu(d) \log \frac{n}{d}$$

Thus, we have as well the following equality, again valid for any function F :

$$\sum_{n \leq x} F\left(\frac{x}{n}\right) \Lambda(n) = \sum_{n \leq x} F\left(\frac{x}{n}\right) \sum_{d|n} \mu(d) \log\left(\frac{n}{d}\right)$$

Now by adding the two equalities that we have, we obtain the following formula:

$$\begin{aligned} & F(x) \log x + \sum_{n \leq x} F\left(\frac{x}{n}\right) \Lambda(n) \\ &= \sum_{n \leq x} F\left(\frac{x}{n}\right) \log\left(\frac{x}{n}\right) \sum_{d|n} \mu(d) + \sum_{n \leq x} F\left(\frac{x}{n}\right) \sum_{d|n} \mu(d) \log\left(\frac{n}{d}\right) \\ &= \sum_{n \leq x} F\left(\frac{x}{n}\right) \sum_{d|n} \mu(d) \left[\log\left(\frac{x}{n}\right) + \log\left(\frac{n}{d}\right) \right] \\ &= \sum_{n \leq x} F\left(\frac{x}{n}\right) \sum_{d|n} \mu(d) \log\left(\frac{x}{d}\right) \\ &= \sum_{n \leq x} \sum_{d|n} F\left(\frac{x}{n}\right) \mu(d) \log\left(\frac{x}{d}\right) \\ &= \sum_{d|x} \mu(d) \log\left(\frac{x}{d}\right) \sum_{m \leq x/d} F\left(\frac{x}{md}\right) \end{aligned}$$

On the other hand, according to our definition of G , we have:

$$G\left(\frac{x}{d}\right) = \log\left(\frac{x}{d}\right) \sum_{m \leq x/d} F\left(\frac{x}{md}\right)$$

Thus, in terms of G , the formula established above reads:

$$F(x) \log x + \sum_{n \leq x} F\left(\frac{x}{n}\right) \Lambda(n) = \sum_{d|x} \mu(d) G\left(\frac{x}{d}\right)$$

But this is exactly the formula in the statement, as desired. \square

Next, we have the following key formula, due to Selberg:

THEOREM 16.21. *We have the following formula,*

$$\psi(x) \log x + \sum_{n \leq x} \Lambda(n) \psi\left(\frac{x}{n}\right) = 2x \log x + O(x)$$

valid for any $x > 0$.

PROOF. This is something quite tricky, the idea being as follows:

(1) Let us first apply the formula in Proposition 16.20 to the function $F_1(x) = \psi(x)$. The function G_1 constructed there is subject to the following estimate:

$$\begin{aligned} G_1(x) &= \log x \sum_{n \leq x} \psi\left(\frac{x}{n}\right) \\ &= x \log^2 x - x \log x + O(\log^2 x) \end{aligned}$$

As for the formula obtained by applying Proposition 16.20, this is as follows:

$$\psi(x) \log x + \sum_{n \leq x} \psi\left(\frac{x}{n}\right) \Lambda(n) = \sum_{d \leq x} \mu(d) G_1\left(\frac{x}{d}\right)$$

(2) Now let us first apply as well the formula in Proposition 16.20 to the function $F_2(x) = x - \gamma - 1$. The function G_2 there is subject to the following estimate:

$$\begin{aligned} G_2(x) &= \log x \sum_{n \leq x} \left(\frac{x}{n} - \gamma - 1\right) \\ &= x \log^2 x - x \log x + O(\log x) \end{aligned}$$

As for the formula obtained by applying Proposition 16.20, this is as follows:

$$(x - \gamma - 1) \log x + \sum_{n \leq x} \left(\frac{x}{n} - \gamma - 1\right) \Lambda(n) = \sum_{d \leq x} \mu(d) G_2\left(\frac{x}{d}\right)$$

(3) Let us compare now what we have in (1,2). As a first observation, we have:

$$G_1(x) - G_2(x) = O(\log^2 x)$$

Next, let us subtract the formulae coming from Proposition 16.20. By using the above estimate for the difference $G_1 - G_2$, we obtain the following estimate:

$$\begin{aligned}
 & \psi(x) \log x + \sum_{n \leq x} \psi\left(\frac{x}{n}\right) \Lambda(n) - (x - \gamma - 1) \log x - \sum_{n \leq x} \left(\frac{x}{n} - \gamma - 1\right) \Lambda(n) \\
 &= \sum_{d \leq x} \mu(d) G_1\left(\frac{x}{d}\right) - \sum_{d \leq x} \mu(d) G_2\left(\frac{x}{d}\right) \\
 &= \sum_{d \leq x} \mu(d) \left[G_1\left(\frac{x}{d}\right) - G_2\left(\frac{x}{d}\right) \right] \\
 &= O(x)
 \end{aligned}$$

(4) And with this, we are almost done. Indeed, rearranging the terms gives:

$$\begin{aligned}
 & \psi(x) \log x + \sum_{n \leq x} \Lambda(n) \psi\left(\frac{x}{n}\right) \\
 &= (x - \gamma - 1) \log x + \sum_{n \leq x} \left(\frac{x}{n} - \gamma - 1\right) \Lambda(n) + O(x) \\
 &= 2x \log x + O(x)
 \end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

Getting now to the Prime Number theorem, this can be now proved as follows:

THEOREM 16.22. *The Selberg formula established above, namely*

$$\psi(x) \log x + \sum_{n \leq x} \Lambda(n) \psi\left(\frac{x}{n}\right) = 2x \log x + O(x)$$

can be used in order to prove the Prime Number theorem.

PROOF. This is something quite tricky, that we will not explain here in detail, the idea being as follows. Consider the following function:

$$\sigma(x) = e^{-x} \psi(e^x) - 1$$

In terms of this function, the Selberg formula gives an estimate as follows:

$$|\sigma(x)| x^2 \leq 2 \int_0^x \int_0^y |\sigma(t)| dt dy + O(x)$$

Consider now the following quantity, with the function σ being as above:

$$C = \limsup_{x \rightarrow \infty} |\sigma(x)|$$

By reasoning by contradiction, the Selberg formula ultimately gives $C = 0$. But this latter estimate proves the Prime Number theorem, as desired. So, this was for the idea, and in practice, exercise of course for you, to read more about this. \square

So long for the Selberg proof of the Prime Number theorem. We should mention that there is as well a third known proof, due to Newman, which is more recent than Selberg's, also not using zeta, and a bit shorter. As usual, exercise for you, to read about this.

16d. Riemann hypothesis

Time to end this book, and with no book on number theory avoiding the Riemann hypothesis, we will talk of course about this, the Riemann hypothesis.

So, let us go back to the classical proof of the Prime Number Theorem, by Hadamard and de la Vallée Poussin, explained before in this chapter. This proof was crucially based on the following fact, which technically was our Theorem 16.12 above:

FACT 16.23. *The Riemann zeta function, analytically continuing*

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

has no zeroes on the lines $\operatorname{Re}(s) = 0$ and $\operatorname{Re}(s) = 1$.

The point now is that this can be regarded as a particular case of the Riemann hypothesis, stating that the zeroes of zeta in the strip $0 \leq \operatorname{Re}(s) \leq 1$ must satisfy:

$$\operatorname{Re}(s) = \frac{1}{2}$$

So, this is the famous Riemann hypothesis, and with the above Fact 16.23 standing as a first piece of motivation for it, and with the proof of the Prime Number theorem using it, discussed earlier in this chapter, explaining the relation with the prime numbers.

In practice now, many things can be said about the Riemann hypothesis, and it is beyond our scope here, at the end of this book, which was meant to be an introduction to arithmetic in general, and to the zeta function in particular, to get into this.

* * *

However, one thing that we can do, now that we know what the main problem is, is to review the material from chapters 14-15, or perhaps from this whole book, with this idea in mind, explaining in the simplest possible words what the main problem is.

At the beginning of everything, assuming a bit of calculus and complex numbers known, as per the level of chapter 9 in this book, we have the following fact:

THEOREM 16.24. *We can talk about the Riemann zeta function*

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

at any $s \in \mathbb{C}$ with $\operatorname{Re}(s) > 1$.

PROOF. We have indeed the following computation, with $s = r + it$ and $r > 1$:

$$\begin{aligned} |\zeta(s)| &\leq \sum_{n=1}^{\infty} \frac{1}{n^r} \\ &< 1 + \int_1^{\infty} \frac{1}{x^r} dx \\ &= 1 + \frac{1}{r-1} \end{aligned}$$

Thus, we are led to the conclusion in the statement. \square

Next, we have the following result, which was something more tricky:

THEOREM 16.25. *We have the following formula,*

$$\zeta(s) = \frac{1}{1 - 2^{1-s}} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s}$$

which can stand as definition for ζ , in the strip $0 < \operatorname{Re}(s) < 1$.

PROOF. We can define the Dirichlet function η as being the signed version of ζ , in the following way, and with the convergence for $\operatorname{Re}(s) > 0$ being elementary:

$$\eta(s) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s}$$

Now observe that ζ and η are connected at $\operatorname{Re}(s) > 1$, in the following way:

$$\begin{aligned} \zeta(s) + \eta(s) &= \sum_{n=1}^{\infty} \frac{1}{n^s} + \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^s} \\ &= 2 \sum_{k=1}^{\infty} \frac{1}{(2k)^s} \\ &= 2^{1-s} \zeta(s) \end{aligned}$$

But this gives the following formula, which is the one in the statement:

$$\zeta(s) = \frac{1}{1 - 2^{1-s}} \eta(s)$$

And with a bit of complex analysis, we are led to the above conclusions. \square

And with this, good news, skipping now many other interesting things that can be said about the zeta function, that we learned the hard way in chapters 14-15, and in the present chapter too, we can formulate the main problem in arithmetic, as follows:

CONJECTURE 16.26 (Riemann hypothesis). *The zeroes of zeta in the critical strip*

$$0 < \operatorname{Re}(s) < 1$$

can only appear on the critical line, $\operatorname{Re}(s) = 1/2$.

As a first comment, feel of course free to believe that this is correct, or not. I mean, if there is no proof yet for this, there should be a reason for that.

To be more precise, barring any bizarre phenomena from logic, this conjecture being wrong would more or less amount in having a certain complex number z in the critical strip, saying to us “here I am, the mighty $z \in \mathbb{C}$, all mathematics being about me”.

And can this be plausible, or not? Not very clear, but speaking “monsters” in mathematics, there are actually quite a few of them, especially in representation theory.

In practice now, many things can be said, about the Riemann hypothesis, for all tastes, and for more on this, there are many good books available. Enjoy.

16e. Exercises

Congratulations for having read this book, and no exercises for this final chapter. But of course, if looking for some difficult number theory exercises, there are many of them. So, have a look at the various books referenced below, read some of them as to get to research level, on a topic that you like, and then start solving problems.

Bibliography

- [1] E. Abe, Hopf algebras, Cambridge Univ. Press (1980).
- [2] T.M. Apostol, Introduction to analytic number theory, Springer (1976).
- [3] V.I. Arnold, Ordinary differential equations, Springer (1973).
- [4] V.I. Arnold, Catastrophe theory, Springer (1974).
- [5] M.F. Atiyah, K-theory, CRC Press (1964).
- [6] M.F. Atiyah, The geometry and physics of knots, Cambridge Univ. Press (1990).
- [7] M.F. Atiyah and I.G. MacDonald, Introduction to commutative algebra, Addison-Wesley (1969).
- [8] T. Banica, The study of functions (2025).
- [9] T. Banica, Advanced linear algebra (2025).
- [10] T. Banica, Invitation to finite groups (2025).
- [11] R.J. Baxter, Exactly solved models in statistical mechanics, Academic Press (1982).
- [12] N. Berline, E. Getzler and M. Vergne, Heat kernels and Dirac operators, Springer (2004).
- [13] B. Blackadar, K-theory for operator algebras, Cambridge Univ. Press (1986).
- [14] S.J. Blundell and K.M. Blundell, Concepts in thermal physics, Oxford Univ. Press (2006).
- [15] S.M. Carroll, Spacetime and geometry, Cambridge Univ. Press (2004).
- [16] V. Chari and A. Pressley, A guide to quantum groups, Cambridge Univ. Press (1994).
- [17] A. Connes, Noncommutative geometry, Academic Press (1994).
- [18] A. Connes and M. Marcolli, Noncommutative geometry, quantum fields and motives, AMS (2008).
- [19] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).
- [20] H.S.M. Coxeter, Regular polytopes, Dover (1948).
- [21] H. Davenport, Multiplicative number theory, Springer (1980).
- [22] W. de Launey and D. Flannery, Algebraic design theory, AMS (2011).

- [23] P.A.M. Dirac, *Principles of quantum mechanics*, Oxford Univ. Press (1930).
- [24] M.P. do Carmo, *Differential geometry of curves and surfaces*, Dover (1976).
- [25] M.P. do Carmo, *Riemannian geometry*, Birkhäuser (1992).
- [26] S.K. Donaldson, *Riemann surfaces*, Oxford Univ. Press (2004).
- [27] R. Durrett, *Probability: theory and examples*, Cambridge Univ. Press (1990).
- [28] A. Einstein, *Relativity: the special and the general theory*, Dover (1916).
- [29] P. Etingof, S. Gelaki, D. Nikshych and V. Ostrik, *Tensor categories*, AMS (2016).
- [30] L.C. Evans, *Partial differential equations*, AMS (1998).
- [31] B. Eynard, *Counting surfaces*, Birkhäuser (2016).
- [32] W. Feller, *An introduction to probability theory and its applications*, Wiley (1950).
- [33] E. Fermi, *Thermodynamics*, Dover (1937).
- [34] R.P. Feynman, R.B. Leighton and M. Sands, *The Feynman lectures on physics*, Caltech (1963).
- [35] P. Flajolet and R. Sedgewick, *Analytic combinatorics*, Cambridge Univ. Press (2009).
- [36] W. Fulton, *Algebraic topology*, Springer (1995).
- [37] W. Fulton and J. Harris, *Representation theory*, Springer (1991).
- [38] M.B. Green, J.H. Schwarz and E. Witten, *Superstring theory*, Cambridge Univ. Press (2012).
- [39] D.J. Griffiths, *Introduction to electrodynamics*, Cambridge Univ. Press (2017).
- [40] D.J. Griffiths and D.F. Schroeter, *Introduction to quantum mechanics*, Cambridge Univ. Press (2018).
- [41] D.J. Griffiths, *Introduction to elementary particles*, Wiley (2020).
- [42] P. Griffiths and J. Harris, *Principles of algebraic geometry*, Wiley (1994).
- [43] A. Grothendieck and J. Dieudonné, *Éléments de géométrie algébrique*, IHES (1967).
- [44] A. Grothendieck et al., *Séminaire de géométrie algébrique*, IHES (1972).
- [45] G.H. Hardy and E.M. Wright, *An introduction to the theory of numbers*, Oxford Univ. Press (1938).
- [46] J. Harris, *Algebraic geometry*, Springer (1992).
- [47] R. Hartshorne, *Algebraic geometry*, Springer (1977).
- [48] K.J. Horadam, *Hadamard matrices and their applications*, Princeton Univ. Press (2007).
- [49] L. Hörmander, *The analysis of linear partial differential operators*, Springer (1983).
- [50] R.A. Horn and C.R. Johnson, *Matrix analysis*, Cambridge Univ. Press (1985).
- [51] J.E. Humphreys, *Introduction to Lie algebras and representation theory*, Springer (1972).

- [52] J.E. Humphreys, Linear algebraic groups, Springer (1975).
- [53] K. Ireland and M. Rosen, A classical introduction to modern number theory, Springer (1982).
- [54] H. Iwaniec and E. Kowalski, Analytic number theory, AMS (2004).
- [55] N. Jacobson, Basic algebra, Dover (1974).
- [56] V.F.R. Jones, Subfactors and knots, AMS (1991).
- [57] L.P. Kadanoff, Statistical physics: statics, dynamics and renormalization, World Scientific (2000).
- [58] M. Karoubi, K-theory: an introduction, Springer (1978).
- [59] C. Kassel, Quantum groups, Springer (1995).
- [60] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).
- [61] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).
- [62] G. Landi, An introduction to noncommutative spaces and their geometry, Springer (1997).
- [63] S. Lang, Algebra, Addison-Wesley (1993).
- [64] S. Lang, Abelian varieties, Dover (1959).
- [65] P. Lax, Linear algebra and its applications, Wiley (2007).
- [66] P. Lax, Functional analysis, Wiley (2002).
- [67] F. Lusztig, Introduction to quantum groups, Birkhäuser (1993).
- [68] S. Majid, Foundations of quantum group theory, Cambridge Univ. Press (1995).
- [69] Y.I. Manin, Quantum groups and noncommutative geometry, Springer (2018).
- [70] M.L. Mehta, Random matrices, Elsevier (2004).
- [71] J. Neukirch, Algebraic number theory, Springer (1999).
- [72] P. Petersen, Riemannian geometry, Springer (2006).
- [73] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).
- [74] W. Rudin, Real and complex analysis, McGraw-Hill (1966).
- [75] W. Rudin, Fourier analysis on groups, Dover (1974).
- [76] H.J. Ryser, Combinatorial mathematics, Wiley (1963).
- [77] W. Schlag, A course in complex analysis and Riemann surfaces, AMS (2014).
- [78] J. Seberry and M. Yamada, Hadamard matrices, Wiley (2020).
- [79] J.P. Serre, A course in arithmetic, Springer (1973).
- [80] J.P. Serre, Linear representations of finite groups, Springer (1977).

- [81] J.P. Serre, Local fields, Springer (1979).
- [82] I.R. Shafarevich, Basic algebraic geometry, Springer (1974).
- [83] J.H. Silverman, The arithmetic of elliptic curves, Springer (1986).
- [84] J.H. Silverman and J.T. Tate, Rational points on elliptic curves, Springer (2015).
- [85] B. Singh, Basic commutative algebra, World Scientific (2011).
- [86] D.R. Stinson, Combinatorial designs: constructions and analysis, Springer (2006).
- [87] M.E. Sweedler, Hopf algebras, W.A. Benjamin (1969).
- [88] T. Tao, Topics in random matrix theory, AMS (2012).
- [89] T. Tao and V.H. Vu, Additive combinatorics, Cambridge Univ. Press (2016).
- [90] C.H. Taubes, Differential geometry, Oxford Univ. Press (2011).
- [91] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).
- [92] D.V. Voiculescu, K.J. Dykema and A. Nica, Free random variables, AMS (1992).
- [93] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).
- [94] L.C. Washington, Introduction to cyclotomic fields, Springer (1982).
- [95] A. Weil, Basic number theory, Springer (1967).
- [96] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).
- [97] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).
- [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

- abelian p-group, 69
- algebraic closure, 192
- algebraic curve, 183, 239
- algebraic manifold, 186
- algebraically closed, 192
- alternating series, 116, 118
- analytic continuation, 360
- analytic function, 307, 312
- approximation, 127
- arbitrary field, 191
- arctan, 137
- area of sphere, 349
- associativity, 38
- average of function, 149

- barycenter, 221
- Bernoulli lemniscate, 242
- Bernoulli numbers, 354
- binomial coefficient, 25, 27, 31, 66, 77
- binomial formula, 26, 147
- boundary of domain, 307
- bounded sequence, 112

- Cardano formula, 234, 236, 242
- cardioid, 241
- cartesian coordinates, 239
- Cassini ovals, 245
- Catalan numbers, 30
- Cauchy criterion, 306
- Cauchy formula, 309, 310, 312
- Cauchy sequence, 114
- Cauchy theorem, 67
- Cayley embedding, 60
- Cayley sextic, 243
- chain rule, 135
- change of variable, 152

- character, 91
- characteristic of field, 72, 76, 191
- characteristic zero, 72
- Chebycheff function, 173
- Chebycheff psi function, 371
- Chebycheff theorem, 175
- Chebycheff theta function, 173
- circulant matrix, 103
- common roots, 226
- comparison of functions, 167
- complete space, 114
- complex conjugate, 304
- complex coordinates, 239
- complex cosine, 301
- complex exponential, 301
- complex function, 213, 303
- complex logarithm, 301
- complex number, 203, 204
- complex power, 303
- complex power function, 303
- complex roots, 219, 231
- complex sine, 301
- complex zeta, 362
- concave function, 143
- congruence, 21
- conic, 183
- continuous function, 129, 213
- convergent sequence, 111
- convergent series, 114
- convex function, 143
- convolution, 55
- convolution exponential, 277
- convolution semigroup, 276
- cos, 126, 135, 147, 301
- cosh, 302

- countable set, 39
- critical strip, 366
- crossed product, 61
- cubic, 241
- cuspidal, 240, 241
- cyclic group, 58, 77

- decimal form, 108
- decreasing sequence, 112
- Dedekind cut, 107
- degenerate curve, 239
- degree 2 equation, 107, 108, 183, 204, 210, 225
- degree 3 equation, 234
- degree 3 polynomial, 232
- degree 4 equation, 236
- degree 4 polynomial, 234
- degree 5 polynomial, 193
- depressed cubic, 234
- depressed quartic, 235
- derivative, 133, 134
- diagonal trick, 39
- differentiable function, 133, 303
- digits, 108
- dihedral group, 62
- Dirac mass, 54
- Dirichlet eta, 358
- Dirichlet series, 357
- discrete convolution, 55
- discrete integration, 54
- discrete law, 54
- discrete measure, 54
- discrete probability, 54
- discriminant, 108, 228, 232
- discriminant formula, 230
- disjoint union, 239
- distributivity, 38
- divergent sum, 199
- divisibility, 21
- double factorial, 164, 347
- double root, 228

- e , 119, 121
- ellipse, 184
- entire function, 312
- Euler constant, 169
- Euler formula, 90, 94, 153
- Euler-Maclaurin formula, 161
- Euler-Mascheroni constant, 169, 170

- exp, 135, 147, 301
- exponential, 121, 215

- factorials, 25
- Fermat polynomial, 79, 191
- Fermat theorem, 31, 66, 77, 78, 191
- field, 38, 76, 189, 191
- field addition, 38
- field character, 91
- field extension, 192, 193
- field inversion, 38
- field multiplication, 38
- finite abelian group, 71
- finite field, 72, 76, 78, 79, 191, 192
- formal cut, 107
- Fourier transform, 278
- fraction, 33, 136
- function, 129

- Galois theorem, 192
- Galois theory, 193, 242
- gamma constant, 169
- gamma function, 346
- Gauss integral, 160
- Gauss sign, 256
- Gauss sum, 251, 256
- generalized binomial formula, 147
- generalized factorial, 346
- generalized Stirling formula, 351
- geometric series, 114, 199
- group of units, 77

- half-integers, 347
- harmonic function, 312, 313
- heart, 243
- higher derivatives, 145
- Hilbert basis theorem, 187
- Hilbert symbol, 98
- holomorphic function, 303, 312
- hyperbola, 184
- hyperbolic cosine, 302
- hyperbolic function, 302
- hyperbolic sine, 302

- i , 203
- increasing sequence, 112
- independence, 55
- infinitely differentiable, 304, 307, 312

- infinity of primes, 24
- integral over curve, 309
- intermediate value, 133
- irrationality of e , 282
- irrationality of π , 293
- isomorphism theorem, 63

- Jacobi symbol, 97
- Jensen inequality, 143

- Kiepert curve, 242
- Klein group, 268
- Kronecker symbol, 98

- L'Hôpital's rule, 141, 145
- Lambda function, 371
- Landau symbols, 167
- Laplace equation, 316
- Laplace method, 161
- Laplace operator, 313
- Legendre symbol, 89
- lemniscate, 242, 245
- \liminf , 113
- \limsup , 113
- limit of sequence, 111
- limit of series, 114
- Liouville theorem, 312
- local maximum, 137, 143
- local minimum, 137, 143
- local-global principle, 177
- locally affine, 134
- locally quadratic, 141
- log, 135, 147, 301

- Möbius function, 322
- main value formula, 308, 312
- maximum, 133, 137
- maximum principle, 307, 312
- meromorphic function, 362
- Mertens constant, 170
- Mertens estimates, 170
- minimum, 133, 137
- missing sign, 256
- modified Chebycheff function, 371
- modulus, 133, 304
- moments, 55
- multiplication of complex numbers, 218
- multiplicative group, 191

- Noetherian ring, 187
- non-degenerate curve, 239
- normal subgroup, 63
- Nullstellensatz, 247
- numeration basis, 22

- order of element, 66
- orthogonal polynomials, 330

- p -adic absolute value, 195
- p -adic distance, 195
- p -adic field, 195, 198
- p -adic geometric series, 199
- p -adic integers, 198
- p -adic norm, 195, 196
- p -adic number, 195
- p -adic rationals, 198
- p -adic valuation, 196
- Paley matrix, 102
- parabola, 184
- parallelogram rule, 204
- parametric coordinates, 239
- partial integration, 152
- Pascal triangle, 27
- percentages, 44
- perfect square, 89
- periodic decimal form, 110
- permutation group, 60
- π , 124
- π function, 172
- plane curve, 239
- PLT, 278
- Poisson law, 278
- Poisson limit theorem, 278
- poker, 44
- polar coordinates, 218, 239
- polar writing, 216
- pole of function, 190, 299
- polynomial, 146, 299
- polynomial lemniscate, 245
- power function, 134, 303
- power series, 306
- prime factors, 23
- prime field, 76, 191
- prime number, 23
- prime number theorem, 378
- principal ideal, 246
- probability, 44

- product of non-squares, 91
- product of polynomials, 239
- psi function, 371
- Pythagoras theorem, 81
- Pythagoras' theorem, 126

- quadratic field, 189
- quadratic Gauss sum, 254, 256, 271
- quadratic reciprocity, 251
- quadratic residue, 90
- quartic, 241
- quaternion units, 103
- quintic, 242
- quotient, 33
- quotient of polynomials, 113

- radial function, 316
- radial harmonic, 316
- radial limit, 306
- radius of convergence, 306
- rational function, 190, 299, 304
- rational number, 33, 110
- real number, 107
- real numbers, 108
- real roots, 231
- reflection formula, 360
- resultant, 226, 227
- Riemann formula, 360
- Riemann integration, 149
- Riemann reflection formula, 360
- Riemann series, 115
- Riemann sums, 161
- Riemann xi function, 363
- Riemann zeta function, 321, 345
- right angle, 81
- right triangle, 81
- root of polynomial, 192
- roots, 193
- roots of polynomial, 219, 299
- roots of unity, 58, 220, 221

- second derivative, 141
- self-intersection, 241
- separable extension, 192, 193
- sequence, 111
- series, 114
- sextic, 242, 243
- sieve, 24

- simple pole, 355
- simplest field, 72
- sin, 126, 135, 147, 301
- single roots, 228
- singularity, 239
- sinh, 302
- solvable group, 193, 242
- sparse matrix, 227
- splitting field, 192
- square root, 107, 204, 206, 210, 225
- Stirling formula, 161, 351
- strict partial sum, 153
- strong triangle inequality, 196
- subsequence, 112, 113
- sum of vectors, 204
- Sylvester determinant, 227
- symbol multiplicativity, 90
- symmetric function, 225

- tan, 137
- Taylor formula, 141, 145–147
- theta function, 173
- tower of extensions, 193
- trapezoids method, 161
- trefoil, 242
- triangle inequality, 196
- trigonometry, 254
- trivial zeroes, 365
- Tschirnhausen curve, 241

- uncountable, 39
- union of curves, 239
- unique factorization, 23
- uniqueness of finite fields, 79, 192

- valuation, 195
- vector, 204
- volume of sphere, 164, 349
- von Mangoldt function, 371

- Walsh matrix, 102, 268
- Weak Nullstellensatz, 246
- Weierstrass basis, 330
- Williamson matrix, 103

- xi function, 363

- zeta function, 321, 345