Fractions and 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. Fractions, Arithmetic

Abstract. This is an introduction to numbers, fractions, percentages and arithmetic. We provide as well a brief introduction to modern number theory, at the end.

Preface

Number theory is the Queen of Mathematics, who has not wished do deal with numbers, in their 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.

Part I deals with numbers, fractions and percentages, all you need to know. Part II goes into basic arithmetics, all sorts of useful tricks and formulae, in the spirit of the above Theorem. Part III deals with real numbers, which are something more complicated, and

far-reaching. As for Part IV, that goes back to arithmetic, with more on the subject, notably on prime numbers, by benefiting from the knowledge of real numbers.

In the hope that you will find this book useful, and get to love numbers and their theory, and for more, we will provide some references at the end.

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 complex numbers, but I learned from them many things too.

Cergy, January 2025
Teo Banica

Contents

Part I

Numbers, fractions

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

1a. Numbers

We can talk about numbers $1, 2, 3, 4, \ldots$, in the obvious way, with the only issue being that of finding some correct symbols for designating them.

1b. Numeration bases

There is a long story here, notably with the choice, in modern times, of using 10 as numeration basis. We will be back to this, later in this chapter.

1c. Basic arithmetic

We say that b divides a, and write $b|a$, when there is a number c such that $a = bc$. In this case we also use the following notation, for designating this quotient number c:

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

These beasts, called "fractions", are subject to a number of simple formulae, which are all useful, in the real life. For addition and substraction, the formulae are:

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

As for multiplication and division, here the formulae are as follows:

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

And more on this, divisibility of numbers, and on fractions too, in the above sense, and in some generalized sense too, when $a \not| b$, later in this book.

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

DEFINITION 1.1. *We say that $a, b \in \mathbb{Z}$ are congruent modulo $c \in \mathbb{Z}$, and write*

$$a = b(c)$$

when c divides $b - a$.

A first interesting question concerns solving $a = 0(n)$, with n fixed and small. By writing $n = n_1 \ldots n_k$ with the factors n_i having no common divisior, we just have to solve this question for certain special values of n, excluding for instance $n = 6$, and this because $6|a$ is equivalent to $2|a$ and $3|a$. These special values of n are actually called "powers of primes", and many things can be said about them, and more on this later.

In practice, the first such numbers are $n = 2, 3, 4, 5, 8, 9, 11, 16$, and in what regards solving $a = 0(n)$, there are many tricks here, which can be summarized as follows:

THEOREM 1.2. *Given a positive integer $a = a_1 \ldots a_r$, we have:*

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

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

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

THEOREM 1.3. *Assuming that we have convinced mankind to change the numeration basis from* 10 *to* 14*, given a positive integer $a = a_1 \ldots a_r$, we have:*

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

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

In short, we have solved the $q = 7, 13$ problems, but as a caveat, we have now $q = 11$ not working. And is this worth it or not, up to you to decide, and launch an online petition if enthusiastic about it. Be said in passing, our Theorem 1.3 is a bit ill-formulated, mixing

things written in basis 10 and basis 14, and we will leave fixing all this, with a fully correct mathematical statement, as another instructive exercise for you.

1d. Prime numbers

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:

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

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 definition for the prime numbers, we have:

DEFINITION 1.4. *The prime numbers are the integers $p > 1$ satisfying*

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

with each of these properties uniquely determining them.

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

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

THEOREM 1.5. *Any integer $n > 1$ decomposes uniquely as*

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

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

PROOF. This is something that you certainly know, related to the equivalent conditions (1,2,3) in Definition 1.4, and exercise for you, to remember how all this works. Exercise as well, work out this for all integers $n \leq 100$, with no calculators allowed. □

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

THEOREM 1.6. *There is an infinity of prime numbers.*

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

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

THEOREM 1.7. *The set of prime numbers P can be obtained as follows:*

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

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

$$\begin{array}{cccccccccccccccccccc}
\underline{2} & 3 & \not{4} & 5 & \not{6} & 7 & \not{8} & 9 & \not{10} & 11 & \not{12} & 13 & \not{14} & 15 & \not{16} & 17 & \not{18} & 19 & \not{20} \\
 & \underline{3} & & 5 & & 7 & & \not{9} & & 11 & & 13 & & \not{15} & & 17 & & 19 & \\
 & & & \underline{5} & & 7 & & & & 11 & & 13 & & & & 17 & & 19 & \\
 & & & & & \underline{7} & & & & 11 & & 13 & & & & 17 & & 19 & \\
 & & & & & & & & & \underline{11} & & 13 & & & & 17 & & 19 & \\
 & & & & & & & & & & & \underline{13} & & & & 17 & & 19 & \\
 & & & & & & & & & & & \vdots & & & & & & &
\end{array}$$

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

1e. Exercises

Exercises:

EXERCISE 1.8.

EXERCISE 1.9.

EXERCISE 1.10.

EXERCISE 1.11.

EXERCISE 1.12.

EXERCISE 1.13.

EXERCISE 1.14.

EXERCISE 1.15.

Bonus exercise.

CHAPTER 2

Counting

2a. Sets, counting

Sets, counting. Many things can be said here.

Among others, we have the inclusion-exclusion principle, and its applications.

We will be back to this later, after discussing rational and real numbers. The problem indeed is that, with our integers, we don't have enough room, for doing many things.

2b. Binomial formula

As a first theorem now, solving a problem which often appears in real life, we have:

THEOREM 2.1. *The number of possibilities of choosing k objects among n objects is*

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

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

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

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

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

everything by the number M of orderings of the k objects that we chose:

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

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

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

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

As an important adding to Theorem 2.1, we should mention that, by definition, we must declare that $0! = 1$, as for the following computation to work:

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

Going ahead now with more mathematics and less philosophy, with Theorem 2.1 complemented by this convention being in final form, we have:

THEOREM 2.2. *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:

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

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

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

In order to finish, it remains to compute the coefficients C_k. But, according to our product formula, C_k is the number of choices for the k needed a variables among the n available a variables. Thus, according to Theorem 2.1, we have:

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

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

Theorem 2.2 is something quite interesting, so let us doublecheck it with some numerics. At small values of n we obtain the following formulae, which are all correct:

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

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

THEOREM 2.3. *The Pascal triangle, formed by the binomial coefficients* $\binom{n}{k}$,

$$1$$
$$1\ ,\ 1$$
$$1\ ,\ 2\ ,\ 1$$
$$1\ ,\ 3\ ,\ 3\ ,\ 1$$
$$1\ ,\ 4\ ,\ 6\ ,\ 4\ ,\ 1$$
$$1\ ,\ 5\ ,\ 10\ ,\ 10\ ,\ 5\ ,\ 1$$
$$\vdots$$

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

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

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

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

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

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

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

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

By using the binomial formula, this formula becomes:

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

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

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

There are many more things that can be said about binomial coefficients, with all sorts of interesting formulae, and we will be back to this, later in this book, on a regular basis, and with the idea being 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.

2c. Binomial coefficients

Binomial coefficients. Divisibility. Many things can be said here.

2d. Further counts

We would like to count now loops on graphs, with this being a quite interesting question. Think for instance percolation, when making coffee, each droplet of water will have to make its way through the coffee particles, and this is how making coffee works.

Generally speaking, counting loops on graphs can be a quite tricky question, and if you don't believe me, I challenge you to find a finite graph, having a reasonable number of vertices, say an arbitrary $N\in\mathbb{N}$ vertices, where this can be easily done.

Instead, let us try to count the length k paths on the graph $\mathbb{Z}$, based at 0.

At $k = 1$ we have 2 such paths, ending at -1 and 1, and the count results can be pictured as follows, with everything being self-explanatory:

$$\circ - \circ - \circ - \bullet - \circ - \circ - \circ$$
$$1 \qquad 1$$

At $k = 2$ now, we have 4 paths, one of which ends at -2, two of which end at 0, and one of which ends at 2. The results can be pictured as follows:

$$\circ - \circ - \circ - \bullet - \circ - \circ - \circ$$
$$1 \qquad 2 \qquad 1$$

At $k = 3$ now, we have 8 paths, the distribution of the endpoints being as follows:

$$\circ - \circ - \circ - \circ - \bullet - \circ - \circ - \circ - \circ$$
$$1 \qquad 3 \qquad 3 \qquad 1$$

As for $k = 4$, here we have 16 paths, the distribution of the endpoints being as follows:

$$\circ - \circ - \circ - \circ - \circ - \bullet - \circ - \circ - \circ - \circ - \circ$$
$$1 \qquad 4 \qquad 6 \qquad 4 \qquad 1$$

And good news, we can see in the above the Pascal triangle. Thus, eventually, we found the simplest graph ever, namely $\mathbb{Z}$, and we have the following result about it:

THEOREM 2.4. *The paths on $\mathbb{Z}$ are counted by the binomial coefficients. In particular, the $2k$-paths based at 0 are counted by the central binomial coefficients,*

$$\binom{2k}{k} \simeq \frac{4^k}{\sqrt{\pi k}}$$

with the estimate, in the $k \to \infty$ limit, coming from the Stirling formula.

PROOF. This basically follows from the above discussion, as follows:

(1) In what regards the count, we certainly have the Pascal triangle, as discovered above, and the rest is just a matter of finishing. There are many possible ways here, a straightforward one being that of arguing that the number C_k^l of length k loops $0 \to l$ is subject, due to the binary choice at the end, to the following recurrence relation:

$$C_k^l = C_{k-1}^{l-1} + C_{k-1}^{l+1}$$

But this is exactly the recurrence for the Pascal triangle, so done with the count.

(2) In what regards the estimate, this follows indeed from Stirling, as follows:

$$\begin{aligned}\binom{2k}{k} &= \frac{(2k)!}{k!k!}\\ &\simeq \left(\frac{2k}{e}\right)^{2k}\sqrt{4\pi k}\times\left(\frac{e}{k}\right)^{2k}\frac{1}{2\pi k}\\ &= \frac{4^k}{\sqrt{\pi k}}\end{aligned}$$

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

2e. Exercises

Exercises:

Exercise 2.5.

Exercise 2.6.

Exercise 2.7.

Exercise 2.8.

Exercise 2.9.

Exercise 2.10.

Exercise 2.11.

Exercise 2.12.

Bonus exercise.

CHAPTER 3

Fractions

3a. Fractions

Time now for some more complicated mathematics, going beyond what we know about the positive integers. We will denote as usual by $\mathbb{N}$ the set of positive integers, $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$, with $\mathbb{N}$ standing for "natural". Quite often we will need negative numbers too, and we denote by $\mathbb{Z}$ the set of all integers, $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$, with $\mathbb{Z}$ standing from "zahlen", which is German for "numbers".

We recall from chapter 1 that given an integer dividing another integer, $b|a$, we can talk about the corresponding quotient c, given by $a = bc$, which is denoted as follows:

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

The above beasts, called "fractions", are subject to a number of simple formulae, which are all useful, in the real life. For addition and substraction, the formulae are:

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

As for multiplication and division, here the formulae are as follows:

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

The point now is that we can talk about fractions even when $b|a$ does not hold, in the obvious way. And, with this convention, the above formulae still hold. Good to know.

3b. Rational numbers

Let us formulate the following definition, based on the above:

DEFINITION 3.1. *The rational numbers are the quotients of type*

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

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

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

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

Observe that we have inclusions $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q}$. The integers add and multiply according to the rules that you know well. As for the rational numbers, these add according to the usual rule for quotients, which is as follows, and never ever forget it:

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

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

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

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

As a basic result about the rational numbers, in relation with what we like to do the most, since the beginning of this book, namely counting, we have:

THEOREM 3.2. *$\mathbb{Q}$ is countable.*

PROOF. This can be proved by using a standard diagonal trick. Consider indeed the following table, containing all quotients of type a/b, with $a, b \in \mathbb{N}$:

$$\begin{array}{cccccccc}
\frac{1}{1} & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \cdots \\
\frac{2}{1} & \frac{2}{2} & \frac{2}{3} & \frac{2}{4} & \frac{2}{5} & \frac{2}{6} & \cdots \\
\frac{3}{1} & \frac{3}{2} & \frac{3}{3} & \frac{3}{4} & \frac{3}{5} & \frac{3}{6} & \cdots \\
\frac{4}{1} & \frac{4}{2} & \frac{4}{3} & \frac{4}{4} & \frac{4}{5} & \frac{4}{6} & \cdots \\
\frac{5}{1} & \frac{5}{2} & \frac{5}{3} & \frac{5}{4} & \frac{5}{5} & \frac{5}{6} & \cdots \\
\frac{6}{1} & \frac{6}{2} & \frac{6}{3} & \frac{6}{4} & \frac{6}{5} & \frac{6}{6} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

We can then snake our way inside this table, in the obvious way, starting from top left, and we count in this way $\mathbb{Q}_+$, with some redundancies. Thus, theorem proved. □

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

3c. Fields, algebra

Beware, some algebra and philosophy coming next. As a first result here, which is quite interesting, and you can call this science, philosophy, or even religion, we have:

THEOREM 3.3. *The positive integers $\mathbb{N}$ self-created themselves starting from the empty set $\emptyset$, according to the following scheme,*

$$|\emptyset| = 0$$

$$|\{\emptyset\}| = 1$$

$$|\{\emptyset, \{\emptyset\}\}| = 2$$

$$|\{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}| = 3$$

$$\vdots$$

and then $\mathbb{Z}, \mathbb{Q}$ naturally came after, by performing differences, and quotients.

PROOF. This is indeed something self-explanatory. Many other things can be said, as a continuation to this, notably in relation with ordinals. We will be back to this. □

As a further philosophical comment here, physicists tend to disagree with the creation operation from Theorem 3.3, with their explanation for the existence of $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ involving their so-called "fields", and a certain quite complicated operation, called Big Bang.

Switching topics now, but still remaining highly philosophical, about numbers, in more advanced mathematical terms, the basic operations on the rationals, namely sum, product and inversion, tell us that $\mathbb{Q}$ is a field, in the following abstract sense:

DEFINITION 3.4. *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, these are the field axioms, in the sense of mathematics, of course, and God thanks here, nothing to do with the fields of physicists, and with the above-mentioned Big Bang operation. Although, in recent times, there has been speculation that mathematical fields with 1 element, and don't ask me what this means, might be actually related to the fields of the physicists. Quite difficult questions here, probably worth a Fields medal.

Apparently, what we did so far, with our philosophical discussion regarding creation, $\diamondsuit \to \mathbb{N} \to \mathbb{Z} \to \mathbb{Q}$, was to construct the simplest possible field, $\mathbb{Q}$. However, this is not

exactly true, because, by a strange twist of fate, the numbers $0, 1$, whose presence in a field is mandatory, $0, 1 \in F$, can form themselves a field, with addition as follows:

$$1 + 1 = 0$$

To be more precise, according to our field axioms, we certainly must have:

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

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

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

$$\mathbb{F}_2 = \{0, 1\}$$

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

THEOREM 3.5. *The following happen:*

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

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

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

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

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

(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 $1 + \ldots + 1$, which must cycle modulo p, as shown by (2). □

Getting back now to our philosophical discussion regarding numbers, what we have in Theorem 3.5 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, and as an answer to this, we have:

ANSWER 3.6. *Ignoring what pure mathematics might say, and trusting instead physics and chemistry, we will choose to trust in $\mathbb{Q}$, as being the simplest field.*

In short, welcome to science, and with this being something quite natural for us, science being the topic of the present book.

3d. More arithmetic

Moving ahead now, with some more arithmetic, many things can be done with $\mathbb{Q}$, but getting straight to the point, one thing that fails is solving $x^2 = 2$:

THEOREM 3.7. *The field $\mathbb{Q}$ does not contain a square root of 2:*

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

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

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

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

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

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

We can talk if we want about fields like $\mathbb{Q}[\sqrt{2}]$, as follows:

PROPOSITION 3.8. *The following set, with $\sqrt{2}$ formally solving $x^2 = 2$, is a field,*

$$\mathbb{Q}[\sqrt{2}] = \left\{ a + b\sqrt{2} \middle| a, b \in \mathbb{Q} \right\}$$

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 nonzero, due to $a^2/b^2 \neq 2$, that we know from Theorem 3.7. As for the case of $\mathbb{Q}[\sqrt{n}]$, this is similar, again by using Theorem 3.7. □

We will be back to questions regarding square roots later in this book, first with more arithmetics, in relation with the existence and non-existence of square roots, and then later, when talking real numbers, with a suitable concrete definition for $\sqrt{2}$.

We will be back as well to fields, on regular occasions, in what follows.

3e. Exercises

Exercises:

Exercise 3.9.

Exercise 3.10.

Exercise 3.11.

Exercise 3.12.

Exercise 3.13.

Exercise 3.14.

Exercise 3.15.

Exercise 3.16.

Bonus exercise.

CHAPTER 4

Percentages

4a. Percentages

Percentages.

4b. Games, winning

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

THEOREM 4.1. *The probabilities at poker are as follows:*

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

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

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

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

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

Thus, the probability of having a pair is:

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

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

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

Thus, the probability of having two pairs is:

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

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

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

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

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

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

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

Thus, the probability of having full house is:

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

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

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

Thus, the probability of having a straight is:

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

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

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

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

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

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

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

Thus, the probability of having a flush is:

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

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

$$N = 4 \cdot 4$$

Thus, the probability of having a straight flush is:

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

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

4c. Flipping coins

Here is now a theorem about flipping coins:

THEOREM 4.2. *When flipping a coin k times what you can win are quantities of type $\$k - 2s$, with $s = 0, 1, \ldots, k$, with the probability for this to happen being:*

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

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

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

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

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

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

Shall we doublecheck this? Sure yes, doublecheking is the first thing to be done, when you come across a theorem, in your mathematics. As a first check, the sum of probabilities

that we found should be 1, which is intuitive, right, and 1 that is, as shown by:

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

But shall we really trust this. So, as second doublecheck, let us verify that, on average, what you win is exactly \$0, which is something very intuitive, the game itself obviously not favoring you, nor your partner. But this can be checked as follows:

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

Summarizing, done with all our checks, and we have now a good and valid theorem here, ready to be used in practice. □

Many more things can be said, as a continuation of the above, for instance by replacing coins with dice, or biased coins, or all sorts of other objects.

4d. Binomial laws

Let us discuss now the notion of independence, which is of key importance, when talking probability. We have here the following result:

THEOREM 4.3. *The following happen, in the context of a 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. Obviously, this is something a bit informal, but let us prove this as stated, and we will come back later to it, with precise definitions, theorems and everything. In what regards the first assertion, nothing to be said there, this is what life teaches us. As for the second assertion, the formula $\mu_{bin} = \mu_{ber}^{*k}$ there certainly looks like mathematics, so job for us to figure out what this exactly means. And, this can be done as follows:

(1) The first idea is to encapsulate the data from the coin game into the 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, the point is that, as we will soon discover abstractly, the mathematics there is that of the following formula, with $*$ standing for the convolution operation for the real measures, which is given by $\delta_x * \delta_y = \delta_{x+y}$ and linearity:

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

(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:

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

(4) Summarizing, save for some uncertainties regarding what independence exactly means, mathematically speaking, and more on this in a moment, theorem proved. □

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

4e. Exercises

Exercises:

Exercise 4.4.

Exercise 4.5.

Exercise 4.6.

Exercise 4.7.

Exercise 4.8.

Exercise 4.9.

Exercise 4.10.

Exercise 4.11.

Bonus exercise.

Part II

Basic arithmetic

Do you remember
Before the rain came down
You were so full of life
So bring that right back around

CHAPTER 5

Prime numbers

5a. Prime numbers

We already know a bit about prime numbers, from the above. Many other things can be said, as a continuation of this. Of particular interest is the sieve method.

5b. Euler formula

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

THEOREM 5.1. *We have the following formula, implying $|P| = \infty$:*

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

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

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

valid for any integer $N \geq 2$.

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

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

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

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

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

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

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

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

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

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

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

5c. Discussion

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

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

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

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

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

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

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

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

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

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

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

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

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

5d. Further results

Further results.

5e. Exercises

Exercises:

Exercise 5.2.

Exercise 5.3.

Exercise 5.4.

Exercise 5.5.

Exercise 5.6.

Exercise 5.7.

Exercise 5.8.

Exercise 5.9.

Bonus exercise.

CHAPTER 6

Basic arithmetic

6a. Basic arithmetic

Basic arithmetic.

6b. Some applications

Some applications.

6c. Further results

Further results.

6d. Advanced theory

Advanced theory.

6e. Exercises

Exercises:

Exercise 6.1.

Exercise 6.2.

Exercise 6.3.

Exercise 6.4.

Exercise 6.5.

Exercise 6.6.

Exercise 6.7.

Exercise 6.8.

Bonus exercise.

CHAPTER 7

Squares, residues

7a. Squares, residues

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

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

Many things can be said here, of various levels of difficulty.

7b. Legendre symbol

We have the following definition, putting everything on a solid basis:

DEFINITION 7.1. *The Legendre symbol is defined as follows,*

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

with $p \geq 3$ prime.

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

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

THEOREM 7.2. *The Legendre symbol is given by the formula*

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

called Euler formula for the Legendre symbol.

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

(1) We know from Fermat that we have $a^p = a(p)$, and leaving aside the case $a = 0(p)$, which is trivial, and therefore solved, this tells us that $a^{p-1} = 1(p)$. But since our prime

p was assumed to be odd, $p \geq 3$, we can write this formula as follows:

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

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

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

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

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

By comparing now with Definition 7.1, we obtain the formula in the statement. □

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

PROPOSITION 7.3. *We have the following formula, valid for any $a, b \in \mathbb{Z}$:*

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

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

PROOF. This is clear indeed from the Euler formula, because $a^{\frac{p-1}{2}}(p)$ is obviously multiplicative in $a \in \mathbb{Z}$. Alternatively, this can be proved as well directly, with no need for the Fermat formula used in the proof of Euler, just by thinking at what is quadratic residue and what is not in $\mathbb{F}_p$, along the lines of (2) in the proof of Theorem 7.2. □

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

PROPOSITION 7.4. *We have the following formula, telling us that modulo any prime number p, a product of non-squares is a square:*

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

Also, the Legendre symbol, regarded as a function

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

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

PROOF. The first asssertion is a consequence of Proposition 7.3, more or less equivalent to it, and with the remark that this formally holds at $p = 2$ too, as $\emptyset \implies \emptyset$. As for the second assertion, this is just a fancy reformulation of Proposition 7.3. □

7c. Quadratic reciprocity

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

THEOREM 7.5. *We have the quadratic reciprocity formula*

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

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

PROOF. This is something quite tricky, one proof being as follows:

(1) First we have a combinatorial formula for the Legendre symbol, called Gauss lemma. Given a prime number $q \geq 3$, and $a \neq 0(q)$, consider the following sequence:

$$a\ ,\ 2a\ ,\ 3a\ ,\ \ldots\ ,\ \frac{q-1}{2}\,a$$

The Gauss lemma tells us that if we look at these numbers modulo q, and denote by n the number of residues modulo q which are greater than $q/2$, then:

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

(2) In order to prove this lemma, the idea is to look at the following product:

$$Z = a \times 2a \times 3a \times \ldots \times \frac{q-1}{2}\,a$$

Indeed, on one hand we have the following formula, with Euler used at the end:

$$Z = a^{\frac{q-1}{2}}\left(\frac{q-1}{2}\right)! = \left(\frac{a}{q}\right)\left(\frac{q-1}{2}\right)!$$

(3) On the other hand, we can compute Z in more complicated way, but leading to a simpler answer. Indeed, let us define the following function:

$$|x| = \begin{cases} x & \text{if } 0 < x < q/2 \\ q - x & \text{if } q/2 < x < q \end{cases}$$

With this convention, our product Z is given by the following formula, with n being as in (1), namely the number of residues modulo q which are greater than $q/2$:

$$Z=(-1)^n\times|a|\times|2a|\times|3a|\times\ldots\times\left|\frac{q-1}{2}\,a\right|$$

(4) But, the numbers $|ra|$ appearing in the above formula are all distinct, so up to a permutation, these must be exactly the numbers $1,2,\ldots,\frac{q-1}{2}$. That is, we have:

$$\left\{|a|\,,\,|2a|\,,\,|3a|\,,\,\ldots\,,\,\left|\frac{q-1}{2}\,a\right|\right\}=\left\{1,2,3,\ldots,\frac{q-1}{2}\right\}$$

Now by multiplying all these numbers, we obtain, via the formula in (3):

$$Z=(-1)^n\left(\frac{q-1}{2}\right)!$$

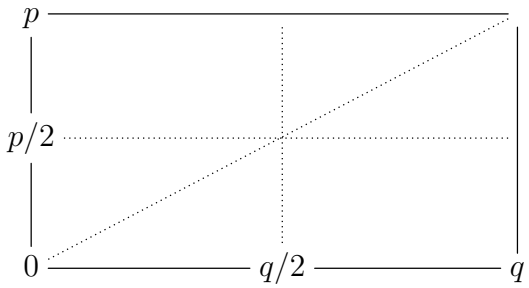
(5) But this is what we need, because when comparing with what we have in (2), we obtain the following formula, which is exactly the one claimed by the Gauss lemma:

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

(6) Next, we have a variation of this formula, due to Eisenstein. His formula for the Legendre symbol, this time involving a prime number numerator $p\geq 3$ in the symbol, is as follows, with the quantities on the right being integer parts, and with the proof being very similar to the proof of the Gauss lemma, that we will leave here as an exercise:

$$\left(\frac{p}{q}\right)=(-1)^n\quad,\quad n=\sum_{k=0}^{(q-1)/2}\left[\frac{2kp}{q}\right]$$

(7) The key point now is that, in this latter formula of Eisenstein, the number n itself counts the points of the lattice $\mathbb{Z}^2$ lying in the triangle $(0,0),(q,0),(q,p)$. So, based on this observation, let us draw a picture, as follows:



(8) We must count the points of $\mathbb{Z}^2$ lying in the triangle $(0,0),(q,0),(q,p)$, modulo 2. This triangle has 3 components, when split by the dotted lines above. Since the points at right, in the small rectangle, and in the small triangle above it, will cancel modulo 2,

we are left with the points at left, in the small triangle there, and the conclusion is that, if we denote by m the number of integer points there, we have the following formula:

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

(9) Now by flipping the diagram, we have as well the following formula, with r being the number of integer points in the small triangle above the small triangle in (8):

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

(10) But, since our two small triangles add up to a small rectangle, we have:

$$m + r = \frac{p-1}{2} \cdot \frac{q-1}{2}$$

Thus, by multiplying the formulae in (8) and (9), we are led to the result. □

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

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

In fact, when combining Theorem 7.5 with Proposition 7.3, it is quite clear that, no matter how big p is, if a has only small prime factors, we are saved.

Besides Proposition 7.3, the quadratic reciprocity formula comes accompanied by two other statements, which are very useful in practice. First, at $a = -1$, we have:

PROPOSITION 7.6. *We have the following formula,*

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

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

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

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

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

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

THEOREM 7.7. *We have the following formula,*

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

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

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

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

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

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

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

As a continuation of this, speaking Legendre symbol for small values of the upper variable, we can try to compute these for $a=\pm\,3,4,5,6,7,8,\ldots$ But by multiplicativity plus Proposition 7.6 plus Theorem 7.7 we are left with the case where $a=q$ is an odd prime, and we can solve the problem with quadratic reciprocity, so done.

Let us record however a few statements here, which can be useful in practice, and with this being mostly for illustration purposes, for Theorem 7.5. We first have:

PROPOSITION 7.8. *We have the following formula,*

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

valid for any prime $p\geq 5$.

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

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

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

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

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

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

PROPOSITION 7.9. *We have the following formula,*

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

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

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

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

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

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

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

7d. Jacobi and Kronecker

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

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

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

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

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

THEOREM 7.11. *The theory of Jacobi symbols can be further extended into a theory of Kronecker symbols, according to the formula*

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

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

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

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

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

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

The answer for the second symbol is as follows:

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

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

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

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

7e. Exercises

Exercises:

EXERCISE 7.12.

EXERCISE 7.13.

EXERCISE 7.14.

EXERCISE 7.15.

EXERCISE 7.16.

EXERCISE 7.17.

EXERCISE 7.18.

EXERCISE 7.19.

Bonus exercise.

CHAPTER 8

Higher equations

8a. Higher equations

Higher equations.

8b. Some tricks

Some tricks.

8c. Fermat equation

Fermat equation.

8d. Further results

Further results.

8e. Exercises

Exercises:

Exercise 8.1.

Exercise 8.2.

Exercise 8.3.

Exercise 8.4.

Exercise 8.5.

Exercise 8.6.

Exercise 8.7.

Exercise 8.8.

Bonus exercise.

Part III

Real 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

Real numbers

9a. Real numbers

We have certainly used real numbers in the above, as everyone does, but time now to get more in detail into their definition, and philosophy. Let us start with something well-known, and quite concerning, that you are surely aware of, namely:

FACT 9.1. *The real numbers $x \in \mathbb{R}$ can be certainly introduced via their decimal form, but with this, the field structure of $\mathbb{R}$ remains something quite unclear.*

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

DEFINITION 9.2. *The real numbers $x \in \mathbb{R}$ are formal cuts in the set of rationals,*

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

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

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

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

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

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

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

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

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

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

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

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

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

THEOREM 9.4. *The solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{R}$ are*

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

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

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

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

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

9b. Decimal writing

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

$$x = \pm a_1 \ldots a_n . b_1 b_2 b_3 \ldots \ldots$$

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

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

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

(2) Moreover, we have a second claim as well, namely that any expression of type $x = \pm a_1 \ldots a_n . b_1 b_2 b_3 \ldots\ldots$ corresponds to a real number $x \in \mathbb{R}$, and that with the convention $\ldots b999 \ldots = \ldots (b+1)000 \ldots$, the correspondence is bijective.

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

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

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

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 $b_1 \in \{0, \ldots, 9\}$, and it is clear that such an approximation exists, simply by comparing x with the numbers $0.0, 0.1, \ldots, 0.9$. Thus, we have our first digit b_1, and then we can construct the second digit b_2 as well, by comparing x with the numbers $0.b_1 0, 0.b_1 1, \ldots, 0.b_1 9$. 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 $x = 0.b_1 b_2 b_3 \ldots$ defines a real number $x \in [0, 1]$, simply by declaring that the corresponding cut $\mathbb{Q} = A_x \sqcup B_x$ comes from the following set, and its complement:

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

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

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

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

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

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

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

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

THEOREM 9.6. *A real number $r \in \mathbb{R}$ is rational precisely when*

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

that is, when its decimal writing is periodic.

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

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

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

At a more advanced level, passed the rationals, our problem remains the same, namely how to recognize the arithmetic properties of the real numbers $r \in \mathbb{R}$, as for instance being square roots of rationals, and so on, when written in decimal form.

9c. Analytic aspects

Getting back now to Theorem 9.5, that was definitely something quite difficult. Alternatively, we have the following definition for the real numbers:

THEOREM 9.7. *The field of real numbers $\mathbb{R}$ can be defined as well as the completion of $\mathbb{Q}$ with respect to the usual distance on the rationals, namely*

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

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

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

(1) Getting back to chapter 3, we know from there what the rational numbers are. But, as a continuation of the material there, we can talk about the distance between such rational numbers, as being given by the formula in the statement, namely:

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

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

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

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

(3) With this notion in hand, the idea will be to define the reals $x\in\mathbb{R}$ as being the limits of the Cauchy sequences $\{r_n\}\subset\mathbb{Q}$. But since these limits are not known yet to exist to us, precisely because they are real, we must employ a trick. So, let us define instead the reals $x\in\mathbb{R}$ as being the Cauchy sequences $\{r_n\}\subset\mathbb{Q}$ themselves.

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

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

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

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

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

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

$$(r_n)=(p_n)\iff d(r_n,p_n)\to0$$

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

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

THEOREM 9.8. *The real numbers $\mathbb{R}$ can be defined as well via the decimal form*

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

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

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

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

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

and then summing and multiplying, in the obvious way.

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

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

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

(2) Now by truncating, we can see that what we have here are certain Cauchy sequences of rationals, and with a bit more work, we conclude that the $\mathbb{R}$ that we constructed is precisely the $\mathbb{R}$ that we constructed in Theorem 9.7. Thus, we get the result.

(3) Alternatively, by getting back to Theorem 9.5 and its proof, we can argue, based on that, that the $\mathbb{R}$ that we constructed coincides with the old $\mathbb{R}$ from Definition 9.2, the one constructed via Dedekind cuts, and this gives again all the assertions. □

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

THEOREM 9.9. *The number e from analysis, given by*

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

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

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

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

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

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

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

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

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

9d. p-adic numbers

Let us discuss now some wild arithmetic tricks, for dealing with equations over the rationals, and with the rational numbers themselves, based on the notion of p-adic number. The idea will be very simple, namely that of completing $\mathbb{Q}$ with respect to a different norm, which privileges the prime number p that we have chosen in advance.

Before that, some motivational talk. The dream in arithmetics, usually concerned with solving equations $f=0$ over the rationals, is something very simple, namely:

DREAM 9.10. *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.

As a second piece of motivation, let us talk some analysis too. Everything in analytic number theory comes from the Euler formula from chapter 5, namely:

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

But this is again something of "local-global" type, with on the left the global quantity, that is, a usual number, which actually happens to be ∞, in our case, and on the right the "local" versions of this number, with respect to the various primes p.

Summarizing, our dream is something important, both from the algebraic and analytic perspective, and is definitely worth a second look, with the aim of fixing it. We are led in this way to the following update to it, which is a bit more modest:

HOPE 9.11. *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 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, let us further reformulate our dreams and hopes, as follows:

QUESTION 9.12. *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 $k>>0$ 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 9.13. *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 $e = 2.7182\ldots$ or $\pi = 3.1415\ldots$ or $1/\alpha = 137.0359\ldots$ 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 9.13 correct? That is, is $|r|$ indeed a norm? And here, is depends a bit on your background, with mathematicians being a bit dissapointed, 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 9.14. *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 arithmetics modulo p:

(1) That axiom clearly holds, with the remark that we forgot to say in Definition 9.13 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 $||.||$ 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. □

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 9.15. *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 9.14.

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 9.14, 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 9.14. 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. □

Very nice all this, and getting back now to what we have in Theorem 9.14, namely the modified norm axioms there, we can formulate, as a simple consequence:

PROPOSITION 9.16. *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. □

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 9.17. *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 $\mathbb{Q} \to \mathbb{Q}_p$ being very similar to the passage $\mathbb{Q} \to \mathbb{R}$, 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. □

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 9.18. *We can introduce the p-adic integers $\mathbf{Z}_p \subset \mathbb{Q}_p$ as being*

$$\mathbf{Z}_p = \left\{ x \in \mathbb{Q}_p \,\middle|\, |x| \leq 1 \right\}$$

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 9.14, 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 is 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 decency and modesty and common sense, the simplest notation, $\mathbb{Z}_p$, should be attributed to the cyclic group.

(6) And many other things can be said, about this. The fight continues to the present day, and if you ever see guerrilla groups inside your Math Department, in military fatigues and duly armed with AR-15 and AK-47 guns, they are probably fighting about $\mathbb{Z}_p$. □

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:

PROPOSITION 9.19. *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 $p^n \to 0$, with respect to the p-adic norm, we have:

$$\begin{aligned}
\sum_{k=0}^{n-1} p^k &= \frac{1-p^n}{1-p} \\
&= \frac{1}{1-p} - \frac{p^n}{1-p} \\
&\simeq \frac{1}{1-p} - \frac{0}{1-p} \\
&= \frac{1}{1-p}
\end{aligned}$$

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

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 doublecheking, this latter formula can be proved as follows:

$$\begin{aligned}\sum_{k=0}^{n-1} 2^k &= 2^n - 1\\ &\simeq 0 - 1\\ &= -1\end{aligned}$$

But we will not get scared by this. Moving ahead now with our general program, of understanding the Cauchy sequences and convergence in $\mathbb{Q}_p$, we have:

THEOREM 9.20. *Convergence in $\mathbb{Q}_p$, and corresponding picture of $\mathbb{Q}_p$.*

PROOF. This follows, as usual, from some elementary arithmetic modulo p, with the conclusion being that the arbitrary p-adic numbers $x \in \mathbb{Q}_p$ have, after all, a quite intuitive interpretation, when it comes to their decimal, or rather p-adic, expansion. □

Finally, again in the analogy with what we know about numbers, we have:

THEOREM 9.21. *The field of p-adic numbers $\mathbb{Q}_p$ can be further enlarged,*

$$\mathbb{Q}_p \subset \bar{\mathbb{Q}}_p$$

into an algebrically closed field $\bar{\mathbb{Q}}_p$, having many interesting properties.

PROOF. This follows indeed by using the general $F \to \bar{F}$ technology from Galois theory, and with this being quite similar to the construction $\mathbb{R} \to \mathbb{C}$. □

Getting back now to our original motivations, namely equations for the integers and rationals, and the local-global principle for them, that we are dreaming of, we have:

THEOREM 9.22. *Hasse local-global principle, and Hasse-Minkowski theorem.*

PROOF. Many things can be said here, but the proofs use a lot of non-trivial algebra. We will present here the main ideas, behind these proofs, with some details missing. □

So long for completions of $\mathbb{Q}$. We will be back to this, on several occasions.

9e. Exercises

Exercises:

Exercise 9.23.

Exercise 9.24.

Exercise 9.25.

Exercise 9.26.

Exercise 9.27.

Exercise 9.28.

Exercise 9.29.

Exercise 9.30.

Bonus exercise.

CHAPTER 10

Some calculus

10a. Some calculus

Many interesting things can be said about real functions and calculus, with the ultimate result on the subject, called Taylor formula, being as follows:

THEOREM 10.1. *Any function $f : \mathbb{R} \to \mathbb{R}$ can be locally approximated as*

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

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

PROOF. Consider the function to be approximated, namely:

$$\varphi(t) = f(x+t)$$

Let us try to best approximate this function at a given order $n \in \mathbb{N}$. We are therefore looking for a certain polynomial in t, of the following type:

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

The natural conditions to be imposed are those stating that P and φ should match at $t = 0$, at the level of the actual value, of the derivative, second derivative, and so on up the n-th derivative. Thus, we are led to the approximation in the statement:

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

In order to prove now that this approximation holds indeed, we can use L'Hôpital's rule, which states that the $0/0$ type limits can be computed as follows:

$$\frac{f(x)}{g(x)} \simeq \frac{f'(x)}{g'(x)}$$

Observe that this formula holds indeed, as an application of basic calculus. Now by using this, if we denote by $\varphi(t) \simeq P(t)$ the formula to be proved, we have:

$$\begin{aligned}
\frac{\varphi(t) - P(t)}{t^n} &\simeq \frac{\varphi'(t) - P'(t)}{nt^{n-1}} \\
&\simeq \frac{\varphi''(t) - P''(t)}{n(n-1)t^{n-2}} \\
&\vdots \\
&\simeq \frac{\varphi^{(n)}(t) - P^{(n)}(t)}{n!} \\
&= \frac{f^{(n)}(x) - f^{(n)}(x)}{n!} \\
&= 0
\end{aligned}$$

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

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

PROPOSITION 10.2. *For a polynomial of degree n, the Taylor approximation*

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

is an equality. The converse of this statement holds too.

PROOF. By linearity, it is enough to check the equality in question for the monomials $f(x) = x^p$, with $p \leq n$. But here, the formula to be proved is as follows:

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

We recognize the binomial formula, so our result holds indeed. As for the converse, this is clear, because the Taylor approximation is a polynomial of degree n. □

In relation with the local extrema, we have the following result:

THEOREM 10.3. *The one-variable smooth functions are subject to the Taylor formula*

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

which allows, via suitable truncations, to determine the local maxima and minima.

PROOF. This is a compact summary of what we know from the above, with everything being in fact quite technical, and with the idea being as follows:

(1) In order to compute the local maxima and minima, a first method is by using the following formula, which comes straight from the definition of the derivative:

$$f(x+t) \simeq f(x) + f'(x)t$$

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

(2) In relation with the problems left, the second derivative comes to the rescue. Indeed, we can use the following more advanced formula, coming via l'Hôpital's rule:

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

To be more precise, assume that we have $f'(x) = 0$, as required by the study in (1). Then this second order formula simply reads:

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

But this is something very useful, telling us that when $f''(x) < 0$, what we have is a local maximum, and when $f''(x) > 0$, what we have is a local minimum. As for the remaining case, that when $f''(x) = 0$, things here remain open.

(3) All this is very useful in practice, and with what we have in (1), complemented if needed with what we have in (2), we can in principle compute the local minima and maxima, without much troubles. However, if really needed, more tools are available. Indeed, we can use if we want the order 3 Taylor formula, which is as follows:

$$f(x+t) \simeq f(x) + f'(x)t + \frac{f''(x)}{2}\,t^2 + \frac{f'''(x)}{6}\,t^3$$

To be more precise, assume that we are in the case $f'(x) = f''(x) = 0$, which is where our joint algorithm coming from (1) and (2) fails. In this case, our formula becomes:

$$f(x+t) \simeq f(x) + \frac{f'''(x)}{6}\,t^3$$

But this solves the problem in the case $f'''(x) \neq 0$, because here we cannot have a local minimum or maximum, due to $t \to -t$ 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 formulate now an abstract result about this, we can use the Taylor formula at order n:

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

Indeed, assume that we started to compute the derivatives $f'(x), f''(x), f''(x), \ldots$ of our function at the point x, with the goal of finding the first such derivative which does not vanish, and we found this derivative, as being the order n one:

$$f'(x) = f''(x) = \ldots = f^{(n-1)}(x) = 0 \quad , \quad f^{(n)}(x) \neq 0$$

Then, the Taylor formula at x at order n takes the following form:

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

But this is exactly what we need, in order to fully solve our local extremum problem. Indeed, when n is even, if $f^{(n)}(x) < 0$ what we have is a local maximum, and if $f^{(n)}(x) > 0$, what we have is a local minimum. As for the case where n is odd, here we cannot have a local minimum or maximum, due to $t \to -t$ which switches growth. □

All the above, Theorem 10.3 and its proof, must be of course perfectly known, when looking for applications of such things. However, for theoretical purposes, let us record as well, in a very compact form, what is basically to be remembered:

THEOREM 10.4. *Given a differentiable function $f : \mathbb{R} \to \mathbb{R}$, we can always write*

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

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

PROOF. This was the conclusion of the proof of Theorem 10.3, and with the extra remark that local extremum means that n is even, with in this case $f^{(n)}(x) < 0$ corresponding to local maximum, and $f^{(n)}(x) > 0$ corresponding to local minimum. □

10b. Complex numbers

Getting back now to number theory, an obvious challenge there is that of solving the eqution $x^2 = -1$. And here, we have the following key result:

THEOREM 10.5. *The complex numbers, $z = a + ib$ with $a, b \in \mathbb{R}$ and with i being a formal number satisying $i^2 = -1$, form a field $\mathbb{C}$. Moreover:*

(1) *We have a field embedding $\mathbb{R} \subset \mathbb{C}$, given by $a \to a + 0 \cdot i$.*
(2) *Additively, we have $\mathbb{C} \simeq \mathbb{R}^2$, with $z = a + ib$ corresponding to (a, b).*
(3) *The length of vectors $r = |z|$, with $z = a + ib$, is given by $r = \sqrt{a^2 + b^2}$.*
(4) *With $z = r(\cos t + i \sin t)$, the products $z = z'z''$ are given by $r = r'r''$, $t = t' + t''$.*
(5) *We have the formula $e^{it} = \cos t + i \sin t$, so we can write $z = re^{it}$.*
(6) *There are N solutions to the equation $z^N = 1$, called N-th roots of unity.*
(7) *Any degree 2 equation with complex coefficients has both roots in $\mathbb{C}$.*

PROOF. We have indeed a field, with all the fields axioms being clear, and with the inversion operation being given by $z^{-1} = (a - ib)/(a^2 + b^2)$, and regarding the rest:

(1) This is clear.

(2) Again, this is clear.

(3) Again, this is clear. Observe also that we have $r^2 = z\bar{z}$, with $\bar{z} = a - ib$.

(4) We need here the formulae for the sines and cosines of sums, which are as follows, coming from some trigonometry, done the old way, with triangles in the plane:

$$\cos(s + t) = \cos s \cos t - \sin s \sin t$$

$$\sin(s + t) = \sin s \cos t + \cos s \sin t$$

Indeed, with these formulae in hand, we have the following computation, as desired:

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

(5) This follows from some heavy calculus, namely Taylor formula for $\exp, \sin, \cos$:

$$\begin{aligned} e^{it} &= \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \\ &= \sum_{l=0}^{\infty} \frac{(it)^{2l}}{(2l)!} + \sum_{l=0}^{\infty} \frac{(it)^{2l+1}}{(2l+1)!} \\ &= \sum_{l=0}^{\infty} (-1)^l \frac{t^{2l}}{(2l)!} + i \sum_{l=0}^{\infty} (-1)^l \frac{t^{2l+1}}{(2l+1)!} \\ &= \cos t + i \sin t \end{aligned}$$

(6) This is clear from (5), with $z = w^k$, with $w = e^{2\pi i/N}$ and $k = 0, 1, \ldots, N - 1$.

(7) This follows in the usual way, with $\sqrt{re^{it}} = \pm\sqrt{r}e^{it/2}$ at the end, using (5). □

We have in fact the following result, generalizing what we know in degree 2:

THEOREM 10.6. *Any polynomial $P \in \mathbb{C}[X]$ decomposes as*

$$P = c(X - a_1) \ldots (X - a_N)$$

with $c \in \mathbb{C}$ and with $a_1, \ldots, a_N \in \mathbb{C}$.

PROOF. The problem is that of proving that our polynomial has at least one root, because afterwards we can proceed by recurrence. We prove this by contradiction. So, assume that P has no roots, and pick a number $z \in \mathbb{C}$ where $|P|$ attains its minimum:

$$|P(z)| = \min_{x \in \mathbb{C}} |P(x)| > 0$$

Since $Q(t) = P(z+t) - P(z)$ is a polynomial which vanishes at $t = 0$, this polynomial must be of the form ct^k + higher terms, with $c \neq 0$, and with $k \geq 1$ being an integer. We obtain from this that, with $t \in \mathbb{C}$ small, we have the following estimate:

$$P(z+t) \simeq P(z) + ct^k$$

Now let us write $t = rw$, with $r > 0$ small, and with $|w| = 1$. Our estimate becomes:

$$P(z+rw) \simeq P(z) + cr^k w^k$$

Now recall that we have 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:

$$|P(z+rw)| \simeq |P(z) + cr^k w^k| = |P(z)|(1 - |c|r^k)$$

Now by choosing $r > 0$ small enough, as for the error in the first estimate to be small, and overcame by the negative quantity $-|c|r^k$, we obtain from this:

$$|P(z+rw)| < |P(z)|$$

But this contradicts our definition of $z \in \mathbb{C}$, as a point where $|P|$ attains its minimum. Thus P has a root, and by recurrence it has N roots, as stated. □

10c. The discriminant

In practice now, we already know how to solve the equations of degree 2. Getting now to degree 3 and higher, let us start with the following result:

THEOREM 10.7. *Given a monic polynomial $P \in \mathbb{C}[X]$, factorized as*

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

the following happen:

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

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

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

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

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

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

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

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

$$S_r = \sum_i a_i^r$$

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

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

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

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

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

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

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

THEOREM 10.8. *Given two polynomials $P, Q \in \mathbb{C}[X]$, written as*

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

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

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

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

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

(1) Given two polynomials $P, Q \in \mathbb{C}[X]$, we can certainly construct the quantity $R(P, Q)$ in the statement, with the role of the normalization factor $c^l d^k$ 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) You might perhaps say, nice, but why not doing things the other way around, that is, formulating our theorem with the explicit formula of $R(P, Q)$, in terms of the coefficients of P, Q, and then proving that we have $R(P, Q) = 0$, via roots and everything. Good point, but this is not exactly obvious, the formula of $R(P, Q)$ in terms of the coefficients of P, Q being something terribly complicated. In short, trust me, let us prove our theorem as stated, and for alternative formulae of $R(P, Q)$, we will see later.

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

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

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

As an illustration, consider a polynomial of degree 2, and a polynomial of degree 1:

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

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

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

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

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

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

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

Regarding now the explicit formula of the resultant $R(P,Q)$, we have here:

THEOREM 10.9. *The resultant of two polynomials, written as*

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

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

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

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

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

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

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

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

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

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

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

(4) Finally, our claim is that we have $\det\Phi = R(P,Q)$. But this follows from the uniqueness of the resultant, up to a scalar, and with this uniqueness property being elementary to establish, along the lines of the proofs of Theorems 10.7 and 10.8. □

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

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

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

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

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

THEOREM 10.10. *Given a polynomial $P \in \mathbb{C}[X]$, written as*

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

its discriminant, defined as being the following quantity,

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

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

PROOF. The fact that the discriminant $\Delta(P)$ is a polynomial in the coefficients of P, with integer coefficients, comes from Theorem 10.8, coupled with the fact that the division by the leading coefficient a is indeed possible, under $\mathbb{Z}$, as being shown by the following formula, which is of course a bit informal, coming from Theorem 10.9:

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

Also, the fact that we have $\Delta(P) = 0$ precisely when P has a double root is clear from Theorem 10.8. 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. □

As already mentioned, by using Theorem 10.9, we have an explicit formula for the discriminant, as the determinant of a certain matrix. There is a lot of theory here, and in order to get into this, let us first see what happens in degree 2. Here we have:

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

Thus, the resultant is given by the following formula:

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

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

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

Alternatively, we can use the formula in Theorem 10.9, and we obtain:

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

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

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

THEOREM 10.11. *The discriminant of a polynomial P is given by the formula*

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

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

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

(1) The first thought goes to the formula in Theorem 10.8, so let us see what that formula teaches us, in the case $Q = P'$. Let us write P, P' as follows:

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

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

According to Theorem 10.8, the resultant of P, P' is then given by:

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

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

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

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

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

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

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

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

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

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

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

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

(4) Time now to put everything together. By taking the formula in (2), making the normalizations in Theorem 10.7, and then using the formula found in (3), we obtain:

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

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

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

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

10d. Degree 3 and 4

As applications now, the formula in Theorem 10.11 is quite useful for the real polynomials $P\in\mathbb{R}[X]$ in small degree, because it allows to say when the roots are real, or complex, or at least have some partial information about this. For instance, we have:

PROPOSITION 10.12. *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 well, but let us see how this comes via the formula in Theorem 10.11, namely:

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

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

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

Thus $\Delta>0$ amounts in saying that we have $(r_1-r_2)^2>0$. Now since r_1,r_2 are conjugate, and with this being something trivial, meaning no need here for the computations in Theorem 10.8, 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 10.11, which in degree 3 reads:

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

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

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

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

Let us discuss now what happens in degree 3. Here the result is as follows:

THEOREM 10.13. *The discriminant of a degree* 3 *polynomial,*

$$P = aX^3 + bX^2 + cX + d$$

is the number $\Delta(P) = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd$.

PROOF. We have two methods available, based on Theorem 10.8 and Theorem 10.9, and both being instructive, we will try them both. The computations are as follows:

(1) Let us first go the pedestrian way, based on the definition of the resultant, from Theorem 10.8. 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:

$$\begin{aligned}R(P,Q) &= a^2e^3(p-s)(p-t)(q-s)(q-t)(r-s)(r-t)\\ &= a^2 \cdot e(p-s)(p-t) \cdot e(q-s)(q-t) \cdot e(r-s)(r-t)\\ &= a^2Q(p)Q(q)Q(r)\\ &= a^2(ep^2+fp+g)(eq^2+fq+g)(er^2+fr+g)\end{aligned}$$

By expanding, we obtain the following formula for this resultant:

$$\begin{aligned}\frac{R(P,Q)}{a^2} &= e^3p^2q^2r^2 + e^2f(p^2q^2r + p^2qr^2 + pq^2r^2)\\ &+ e^2g(p^2q^2 + p^2r^2 + q^2r^2) + ef^2(p^2qr + pq^2r + pqr^2)\\ &+ efg(p^2q + pq^2 + p^2r + pr^2 + q^2r + qr^2) + f^3pqr\\ &+ eg^2(p^2 + q^2 + r^2) + f^2g(pq + pr + qr)\\ &+ fg^2(p + q + r) + g^3\end{aligned}$$

Note in passing that we have 27 terms on the right, as we should, and with this kind of check being mandatory, when doing such computations. Next, we have:

$$p+q+r=-\frac{b}{a}\quad,\quad pq+pr+qr=\frac{c}{a}\quad,\quad pqr=-\frac{d}{a}$$

By using these formulae, we can produce some more, as follows:

$$p^2+q^2+r^2=(p+q+r)^2-2(pq+pr+qr)=\frac{b^2}{a^2}-\frac{2c}{a}$$

$$p^2q+pq^2+p^2r+pr^2+q^2r+qr^2=(p+q+r)(pq+pr+qr)-3pqr=-\frac{bc}{a^2}+\frac{3d}{a}$$

$$p^2q^2+p^2r^2+q^2r^2=(pq+pr+qr)^2-2pqr(p+q+r)=\frac{c^2}{a^2}-\frac{2bd}{a^2}$$

By plugging now this data into the formula of $R(P,Q)$, we obtain:

$$\begin{aligned}R(P,Q)&=a^2e^3\cdot\frac{d^2}{a^2}-a^2e^2f\cdot\frac{cd}{a^2}+a^2e^2g\left(\frac{c^2}{a^2}-\frac{2bd}{a^2}\right)+a^2ef^2\cdot\frac{bd}{a^2}\\&+a^2efg\left(-\frac{bc}{a^2}+\frac{3d}{a}\right)-a^2f^3\cdot\frac{d}{a}\\&+a^2eg^2\left(\frac{b^2}{a^2}-\frac{2c}{a}\right)+a^2f^2g\cdot\frac{c}{a}-a^2fg^2\cdot\frac{b}{a}+a^2g^3\end{aligned}$$

Thus, we have the following formula for the resultant:

$$\begin{aligned}R(P,Q)&=d^2e^3-cde^2f+c^2e^2g-2bde^2g+bdef^2-bcefg+3adefg\\&-adf^3+b^2eg^2-2aceg^2+acf^2g-abfg^2+a^2g^3\end{aligned}$$

Getting back now to our discriminant problem, with $Q=P'$, which corresponds to $e=3a$, $f=2b$, $g=c$, we obtain the following formula:

$$\begin{aligned}R(P,P')&=27a^3d^2-18a^2bcd+9a^2c^3-18a^2bcd+12ab^3d-6ab^2c^2+18a^2bcd\\&-8ab^3d+3ab^2c^2-6a^2c^3+4ab^2c^2-2ab^2c^2+a^2c^3\end{aligned}$$

By simplifying terms, and dividing by a, we obtain the following formula:

$$-\Delta(P)=27a^2d^2-18abcd+4ac^3+4b^3d-b^2c^2$$

But this gives the formula in the statement, namely:

$$\Delta(P)=b^2c^2-4ac^3-4b^3d-27a^2d^2+18abcd$$

(2) Let us see as well how the computation does, by using Theorem 10.9, 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 now Theorem 10.9 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 10.10, 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. □

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:

$$x^3 + 3px + 2q = 0$$

Regarding such equations, many things can be said, and to start with, we have the following famous result, dealing with real roots, due to Cardano:

THEOREM 10.14. *For a normalized degree 3 equation, namely*

$$x^3+3px+2q=0$$

the discriminant is $\Delta=-108(p^3+q^2)$. *Assuming* $p,q\in\mathbb{R}$ *and* $\Delta<0$, *the numbers*

$$z=w\sqrt[3]{-q+\sqrt{p^3+q^2}}+w^2\sqrt[3]{-q-\sqrt{p^3+q^2}}$$

with $w=1,e^{2\pi i/3},e^{4\pi i/3}$ *are the solutions of our equation.*

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

(1) The formula of Δ comes the theory of the discriminant, as developed above.

(2) With z as in the statement, by using $(a+b)^3=a^3+b^3+3ab(a+b)$, we have:

$$\begin{aligned}z^3&=\left(w\sqrt[3]{-q+\sqrt{p^3+q^2}}+w^2\sqrt[3]{-q-\sqrt{p^3+q^2}}\right)^3\\&=-2q+3\sqrt[3]{-q+\sqrt{p^3+q^2}}\cdot\sqrt[3]{-q-\sqrt{p^3+q^2}}\cdot z\\&=-2q+3\sqrt[3]{q^2-p^3-q^2}\cdot z\\&=-2q-3pz\end{aligned}$$

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

In degree 4 now, we first have the following result, dealing with the discriminant:

THEOREM 10.15. *The discriminant of* $P=ax^4+bx^3+cx^2+dx+e$ *is given by the following formula:*

$$\begin{aligned}\Delta&=256a^3e^3-192a^2bde^2-128a^2c^2e^2+144a^2cd^2e-27a^2d^4\\&+144ab^2ce^2-6ab^2d^2e-80abc^2de+18abcd^3+16ac^4e\\&-4ac^3d^2-27b^4e^2+18b^3cde-4b^3d^3-4b^2c^3e+b^2c^2d^2\end{aligned}$$

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 10.10, with the resultant computed via Theorem 10.9, 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. □

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 10.15:

PROPOSITION 10.16. *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 10.15, 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 10.15. □

We still have some work to do. Indeed, looking back at what we did in degree 3, the passage there from Theorem 10.13 to Theorem 10.14 was made of two operations, namely "depressing" the equation, that is, getting rid of the next-to-highest term, and then rescaling the coefficients, as for the formula of Δ to become as simple as possible.

In our present setting now, degree 4, with the depressing done as above, in Proposition 10.16, it remains to rescale the coefficients, as for the formula of Δ to become as simple as possible. And here, a bit of formula hunting, in relation with $2, 3$ powers, leads to:

THEOREM 10.17. *The discriminant of a normalized degree 4 polynomial, written as*

$$P = x^4 + 6px^2 + 4qx + 3r$$

is given by the following formula:

$$\Delta = 256 \times 27 \times \left(9p^4r - 2p^3q^2 - 6p^2r^2 + 6pq^2r - q^4 + r^3\right)$$

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 10.16, 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 10.15. □

Time now to get to the real thing, solving the equation. We have here:

THEOREM 10.18. *The roots of a normalized degree 4 equation, written as*

$$x^4 + 6px^2 + 4qx + 3r = 0$$

are as follows, with y satisfying the equation $(y^2 - 3r)(y - 3p) = 2q^2$,

$$x_1 = \frac{1}{\sqrt{2}}\left(-\sqrt{y - 3p} + \sqrt{-y - 3p + \frac{4q}{\sqrt{2y - 6p}}}\right)$$

$$x_2 = \frac{1}{\sqrt{2}}\left(-\sqrt{y - 3p} - \sqrt{-y - 3p + \frac{4q}{\sqrt{2y - 6p}}}\right)$$

$$x_3 = \frac{1}{\sqrt{2}}\left(\sqrt{y - 3p} + \sqrt{-y - 3p - \frac{4q}{\sqrt{2y - 6p}}}\right)$$

$$x_4 = \frac{1}{\sqrt{2}}\left(\sqrt{y - 3p} - \sqrt{-y - 3p - \frac{4q}{\sqrt{2y - 6p}}}\right)$$

and with y being computable via the Cardano formula.

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

(1) To start with, let us write our equation in the following form:

$$x^4 = -6px^2 - 4qx - 3r$$

Now assume that we have a number y satisfying the following equation:

$$(y^2 - 3r)(y - 3p) = 2q^2$$

With this magic number y in hand, our equation takes the following form:

$$\begin{aligned}(x^2+y)^2 &= x^4+2x^2y+y^2\\ &= -6px^2-4qx-3r+2x^2y+y^2\\ &= (2y-6p)x^2-4qx+y^2-3r\\ &= (2y-6p)x^2-4qx+\frac{2q^2}{y-3p}\\ &= \left(\sqrt{2y-6p}\cdot x-\frac{2q}{\sqrt{2y-6p}}\right)^2\end{aligned}$$

(2) Which looks very good, leading us to the following degree 2 equations:

$$x^2+y+\sqrt{2y-6p}\cdot x-\frac{2q}{\sqrt{2y-6p}}=0$$

$$x^2+y-\sqrt{2y-6p}\cdot x+\frac{2q}{\sqrt{2y-6p}}=0$$

Now let us write these two degree 2 equations in standard form, as follows:

$$x^2+\sqrt{2y-6p}\cdot x+\left(y-\frac{2q}{\sqrt{2y-6p}}\right)=0$$

$$x^2-\sqrt{2y-6p}\cdot x+\left(y+\frac{2q}{\sqrt{2y-6p}}\right)=0$$

(3) Regarding the first equation, the solutions there are as follows:

$$x_1=\frac{1}{2}\left(-\sqrt{2y-6p}+\sqrt{-2y-6p+\frac{8q}{\sqrt{2y-6p}}}\right)$$

$$x_2=\frac{1}{2}\left(-\sqrt{2y-6p}-\sqrt{-2y-6p+\frac{8q}{\sqrt{2y-6p}}}\right)$$

As for the second equation, the solutions there are as follows:

$$x_3=\frac{1}{2}\left(\sqrt{2y-6p}+\sqrt{-2y-6p-\frac{8q}{\sqrt{2y-6p}}}\right)$$

$$x_4=\frac{1}{2}\left(\sqrt{2y-6p}-\sqrt{-2y-6p-\frac{8q}{\sqrt{2y-6p}}}\right)$$

(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. □

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 10.18, is as follows:

THEOREM 10.19 (continuation). *The value of y in the previous theorem is*

$$y = t + p + \frac{a}{t}$$

where the number t is given by the formula

$$t = \sqrt[3]{b + \sqrt{b^2 - a^3}}$$

with $a = p^2 + r$ and $b = 2p^2 - 3pr + q^2$.

PROOF. The legend has it that this is what comes from Cardano, but depressing and normalizing and solving $(y^2 - 3r)(y - 3p) = 2q^2$ makes it for too many operations, so the most pragmatic way is to simply check this equation. With y as above, we have:

$$\begin{aligned} y^2 - 3r &= t^2 + 2pt + (p^2 + 2a) + \frac{2pa}{t} + \frac{a^2}{t^2} - 3r \\ &= t^2 + 2pt + (3p^2 - r) + \frac{2pa}{t} + \frac{a^2}{t^2} \end{aligned}$$

With this in hand, we have the following computation:

$$\begin{aligned} (y^2 - 3r)(y - 3p) &= \left(t^2 + 2pt + (3p^2 - r) + \frac{2pa}{t} + \frac{a^2}{t^2}\right)\left(t - 2p + \frac{a}{t}\right) \\ &= t^3 + (a - 4p^2 + 3p^2 - r)t + (2pa - 6p^3 + 2pr + 2pa) \\ &\quad + (3p^2a - ra - 4p^2a + a^2)\frac{1}{t} + \frac{a^3}{t^3} \\ &= t^3 + (a - p^2 - r)t + 2p(2a - 3p^2 + r) + a(a - p^2 - r)\frac{1}{t} + \frac{a^3}{t^3} \\ &= t^3 + 2p(-p^2 + 3r) + \frac{a^3}{t^3} \end{aligned}$$

Now by using the formula of t in the statement, this gives:

$$\begin{aligned} (y^2 - 3r)(y - 3p) &= b + \sqrt{b^2 - a^3} - 4p^2 + 6pr + \frac{a^3}{b + \sqrt{b^2 - a^3}} \\ &= b + \sqrt{b^2 - a^3} - 4p^2 + 6pr + b - \sqrt{b^2 - a^3} \\ &= 2b - 4p^2 + 6pr \\ &= 2(2p^2 - 3pr + q^2) - 4p^2 + 6pr \\ &= 2q^2 \end{aligned}$$

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

In degree 5 and more, things become fairly complicated, and we have:

THEOREM 10.20. *There is no general formula for the roots of polynomials of degree $N = 5$ and higher, with the reason for this, coming from Galois theory, being that the group S_5 is not solvable. The simplest numeric example is $P = X^5 - X - 1$.*

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

(1) Given a field F, assume that the roots of $P \in F[X]$ can be computed by using iterated roots, a bit as for the degree 2 equation, or the degree 3 and 4 equations. Then, algebrically speaking, this gives rise to a tower of fields as folows, with $F_0 = F$, and each F_{i+1} being obtained from F_i by adding a root, $F_{i+1} = F_i(x_i)$, with $x_i^{n_i} \in F_i$:

$$F_0 \subset F_1 \subset \ldots \subset F_k$$

(2) In order for Galois theory to apply to this situation, we must make all the extensions normal, which amounts in replacing each $F_{i+1} = F_i(x_i)$ by its extension $K_i(x_i)$, with K_i extending F_i by adding a n_i-th root of unity. Thus, with this replacement, we can assume that the tower in (1) in normal, meaning that all Galois groups are cyclic.

(3) Now by Galois theory, at the level of the corresponding Galois groups we obtain a tower of groups as follows as follows, which is a resolution of the last group G_k, the Galois group of P, in the sense of group theory, in the sense that all quotients are cyclic:

$$G_1 \subset G_2 \subset \ldots \subset G_k$$

As a conclusion, Galois theory tells us that if the roots of a polynomial $P \in F[X]$ can be computed by using iterated roots, then its Galois group $G = G_k$ must be solvable.

(4) In the generic case, the conclusion is that Galois theory tells us that, in order for all polynomials of degree 5 to be solvable, via square roots, the group S_5, which appears there as Galois group, must be solvable, in the sense of group theory. But this is wrong, because the alternating subgroup $A_5 \subset S_5$ is simple, and therefore not solvable. □

10e. Exercises

Exercises:

EXERCISE 10.21.

EXERCISE 10.22.

EXERCISE 10.23.

EXERCISE 10.24.

EXERCISE 10.25.

EXERCISE 10.26.

EXERCISE 10.27.

EXERCISE 10.28.

Bonus exercise.

CHAPTER 11

Gauss sums

11a. Gauss sums

Time for the roots of unity to strike again, this time with some non-trivial applications to the Legendre symbols. Going back to what we learned about these symbols, there were several mysterious things there, that we will attempt to elucidate now.

Let us start with the $a = 2$ case. The result here is as follows:

THEOREM 11.1. *We have the following formula,*

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

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

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

(1) As a first observation, the Euler formula at $a = 2$ is as follows, obviously well below the quality of the very precise formula in the statement:

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

As a second observation, the quadratic reciprocity formula, assuming that known, cannot help either, because in that formula $p, q \geq 3$ are odd primes.

(2) Thus, we must prove the result. As already mentioned before, the proof will come via the following formula, which is equivalent to the formula in the statement:

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

Finally, let us mention too that, despite 2 being an even prime, the problematics here is a bit similar to the one of the quadratic reciprocity formula, and the proof below will contain many good ideas, that we will use later in the proof of quadratic reciprocity.

(3) Getting started now, let us set $w = e^{\pi i/4}$, so that $w^2 = i$, do not ask me why, and then $t = w + w^{-1}$. We have of course $t = \sqrt{2}$, but it is better to forget this, and do formal

arithmetics instead, with integers as scalars, based on the following computation:

$$\begin{aligned} t^2 &= 2 + w^2 + w^{-2} \\ &= 2 + i - i \\ &= 2 \end{aligned}$$

Now by using the Euler formula for the Legendre symbol, we have:

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

(4) By multiplying now by t we obtain from this, in a formal sense, and I will leave it you to clarify all the details here, namely what this formal sense exactly means:

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

(5) On the other hand, by using the binomial formula, and the standard fact that all non-trivial binomial coefficients are multiples of p, we obtain, again formally:

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

(6) Now let us look at $w^p + w^{-p}$, as usual complex number. Since $w = e^{\pi i/4}$, this quantity will depend only on p modulo 8, and more precisely, we have:

$$w^p + w^{-p} = \begin{cases} w + w^{-1} & \text{if } p = \pm 1(8) \\ -w - w^{-1} & \text{if } p = \pm 3(8) \end{cases}$$

Thus $w^p + w^{-p} = \pm t$, with the sign depending on p modulo 8, and more specifically:

$$w^p + w^{-p} = (-1)^{\frac{p^2-1}{8}} t$$

(7) Time now to put everything together. By combining (4,5,6) we obtain:

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

By dividing by t, this gives the following formula:

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

But the mod p symbol can now be dropped, because our equality is between two ± 1 quantities, and we obtain the formula in the statement. □

11b. Reciprocity, revised

With the same idea, we can prove as well the quadratic reciprocity theorem:

THEOREM 11.2. *We have the quadratic reciprocity formula*

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

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

PROOF. This is something already advertised in the above, and we refer to the discussion there for the mighty power of this formula, and its enigmatic nature. However, thinking a bit, our $t = w + w^{-1}$ trick above can be adapted, as follows:

(1) To start with, we need an analogue of that $t = w + w^{-1}$ variable. For this purpose, let us set $w = e^{2\pi i/q}$, now that we have a prime $q \geq 3$ involved, and then:

$$t = \sum_{k=0}^{q-1} w^{k^2}$$

Observe that at $q = 2$, excluded by the statement, we have $w = -1$, and so $t = 1 + (-1) = 0$, instead of the $t = w + w^{-1}$ with $w = e^{\pi i/4}$ used before. However, believe me, this is due to some bizarre reasons, and the above t is the good variable, at $q \geq 3$.

(2) The above variable t is called Gauss sum, can be defined for any $q \in \mathbb{N}$, not necessarily prime, and can be explicitely computed, the formula being as follows:

$$t = \begin{cases} \sqrt{q} & \text{if } q = 1(4) \\ 0 & \text{if } q = 2(4) \\ \sqrt{q}\, i & \text{if } q = 3(4) \\ \sqrt{q}(1+i) & \text{if } q = 0(4) \end{cases}$$

In particular, assuming that q is odd, as is our $q \geq 3$ prime, we have:

$$t^2 = \begin{cases} q & \text{if } q = 1(4) \\ -q & \text{if } q = 3(4) \end{cases}$$

(3) In what follows we will only need this latter formula, for $q \geq 3$ prime, so let us prove this now, and with the comment that the proof of the first formula in (2) is something

quite complicated, and better avoid that. We have, by definition of our variable t:

$$\begin{aligned}|t|^2 &= \sum_{kl} w^{k^2-l^2}\\ &= \sum_{kl} w^{(k+l)(k-l)}\\ &= \sum_{lr} w^{r(2l+r)}\\ &= \sum_{r} w^{r^2}\sum_{l}(w^{2r})^l\\ &= q\end{aligned}$$

(4) On the other hand, it is easy to see that t^2 is real, so $t^2=\pm q$. With a bit more work it is possible to compute the sign too, $t^2=(-1)^{\frac{q-1}{2}}q$, but we will not need this here, because the sign will come for free at the end of the proof, via a symmetry argument. So, as a conclusion, we have a formula as follows, for a certain $e_q\in\{0,1\}$:

$$t^2=(-1)^{e_q}q$$

(5) With this done, let us turn to the proof of our theorem, by using the variable t a bit as before, in the proof of Theorem 11.1. By using the Euler formula, we have:

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

By multiplying now by t we obtain from this, in a formal sense:

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

(6) In order to compute now t^p by other means, observe first that, if we denote by $\mathbb{Z}_q-\{0\}=S\sqcup N$ the partition into squares and non-squares, we have:

$$\begin{aligned}t &= \sum_{k=0}^{q-1} w^{k^2}\\ &= 1+2\sum_{s\in S} w^s\\ &= \sum_{s\in S} w^s-\sum_{s\in N} w^s\\ &= \sum_{r=0}^{k-1}\left(\frac{r}{q}\right)w^r\end{aligned}$$

(7) By using now the multinomial formula, with the observation that all the non-trivial multinomial coefficients are multiples of p, we obtain, in a formal sense:

$$\begin{aligned}
t^p &= \left(\sum_r \left(\frac{r}{q}\right) w^r\right)^p \\
&= \sum_r \left(\frac{r}{q}\right) w^{rp} \ (p) \\
&= \sum_s \left(\frac{p^{-1}s}{q}\right) w^s \ (p) \\
&= \left(\frac{p^{-1}}{q}\right) \sum_s \left(\frac{s}{q}\right) w^s \ (p) \\
&= \left(\frac{p}{q}\right) t \ (p)
\end{aligned}$$

(8) Time now to put everything together. By combining (5,7) we obtain:

$$\left(\frac{t^2}{p}\right) t = \left(\frac{p}{q}\right) t \ (p)$$

We can divide by t, and then drop the modulo p symbol, because our new equality, without t, is between two ± 1 quantities, and we obtain:

$$\left(\frac{t^2}{p}\right) = \left(\frac{p}{q}\right)$$

Now by taking into account the formula found in (4), this reads:

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

By using the Euler formula for the symbol on the left, we obtain from this:

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

Now by symmetry we must have $e_q = \frac{q-1}{2}$, and this finishes the proof. □

11c. 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 11.3. *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 11.4. *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 11.5. *The first few quadratic Gauss sums are given by*

$$\begin{array}{c|cccc|cccc|} q & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8\\ t & 1 & 0 & \sqrt{3}\,i & 2(1+i) & \sqrt{5} & 0 & \sqrt{7}\,i & 2\sqrt{2}(1+i)\\ \end{array}$$

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 11.6. *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 that 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 11.2, 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 0, 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 11.4, 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 11.2:

$$t = \sum_{r=0}^{q-1} \left(\frac{r}{q}\right) w^r$$

(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 11.4, 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 11.4 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. Obviously, 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. □

11d. The Gauss sign

Computation of the missing sign.

11e. Exercises

Exercises:

Exercise 11.7.

Exercise 11.8.

Exercise 11.9.

Exercise 11.10.

Exercise 11.11.

Exercise 11.12.

Exercise 11.13.

Exercise 11.14.

Bonus exercise.

CHAPTER 12

Transcendence

12a. Weird numbers

Weird numbers.

12b. Transcendence of e

Time for some tough calculus. We first have the following result, about e:

THEOREM 12.1. *The number e from analysis, given by*

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

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

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

(1) To start with, there are several possible definitions for e, with the old style one, which is quite cool, and that you can still find in fine calculus books, being:

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

The definition in the statement is the modern one. Indeed, imagine that you are looking for a function exp, satisfying $\exp' = \exp$, and $\exp(0) = 1$. With $\exp(x) = \sum c_k x^k$, you must have $c_0 = 1$, then $c_1 = 1$, $c_2 = 1/2$, $c_3 = 1/6$ and so on, meaning:

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

But now, it is an easy exercice to show that $\exp(x + y) = \exp(x)\exp(y)$, which gives $\exp(x) = e^x$, for a certain number $e > 0$. Which number e can only be $e = \exp(1)$.

(2) Getting now to numerics, the series of e converges very fast, when compared to the old style sequence in (1), so if you are in a hurry, this series is for you. We have:

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

Thus, the error term in the approximation is really tiny, the estimate being:

$$\sum_{k=0}^{N}\frac{1}{k!}<e<\sum_{k=0}^{N}\frac{1}{k!}+\frac{1}{N\cdot N!}$$

(3) Now by using this, you can easily compute the decimals of e. Actually, you can't call yourself mathematician, or scientist, if you haven't done this by hand, just for the fun, but just in case, here is how the approximation goes, for small values of N:

$$N=2 \implies 2.5<e<2.75$$

$$N=3 \implies 2.666\ldots<e<2.722\ldots$$

$$N=4 \implies 2.70833\ldots<e<2.71875\ldots$$

$$N=5 \implies 2.71666\ldots<e<2.71833\ldots$$

$$N=6 \implies 2.71805\ldots<e<2.71828\ldots$$

$$N=7 \implies 2.71825\ldots<e<2.71828\ldots$$

Thus, first 4 decimals computed, $e=2.7182\ldots$, and I would leave the continuation to you. With the remark that, when carefully looking at the above, the estimate on the right works much better than the one on the left, so before getting into more serious numerics, try to find a better lower estimate for e, that can help you in your work.

(4) Getting now to irrationality, a look at $e=2.7182818284\ldots$ might suggest that the $81, 82, 84\ldots$ values might eventually, after some internal fight, decide for a winner, and so that e might be rational. However, this is wrong, and e is in fact irrational.

(5) So, let us prove now this, that e is irrational. Following Fourier, we will do this by contradiction. So, assume $e = m/n$, and let us look at the following number:

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

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

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

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

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

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

As a continuation, we have the following result, which is substantially harder:

THEOREM 12.2. *The number e is transcendental.*

PROOF. Assume by contradiction that e is algebraic, with this meaning that it is a root of a polynomial with integer coefficients, $c_i \in \mathbb{Z}$, as follows:

$$c_0 + c_1 e + \ldots + c_n e^n = 0$$

(1) Following Hermite, consider the following polynomials, and we will see later why:

$$f_k(x) = x^k \left[(x-1)\dots(x-n)\right]^{k+1}$$

Consider also the following quantities, that we will study more in detail later:

$$A_k = \int_0^{\infty} f_k(x) e^{-x} dx$$

By multiplying our equation for e by this quantity A_k, we obtain:

$$c_0\int_0^\infty f_k(x)e^{-x}dx+c_1\int_0^\infty f_k(x)e^{1-x}dx+\ldots c_n\int_0^\infty f_k(x)e^{n-x}dx=0$$

(2) Here comes the trick. Consider the following two quantities:

$$P=c_0\int_0^\infty f_k(x)e^{-x}dx+c_1\int_1^\infty f_k(x)e^{1-x}dx+\ldots+c_n\int_n^\infty f_k(x)e^{n-x}dx$$

$$Q=c_1\int_0^1 f_k(x)e^{-x}dx+\ldots+c_n\int_0^n f_k(x)e^{n-x}dx$$

In terms of these quantities, the formula that we found in (1) reads:

$$P+Q=0$$

(3) Now let us look at P. Our claim is that this is an integer, $P\in\mathbb{Z}$, and that there are arbitrarily large numbers $k>>0$ for which the following holds:

$$\frac{P}{k!}\in\mathbb{Z}-\{0\}$$

Indeed, according to our formula above defining P, we have:

$$\begin{aligned}P&=\sum_{r=0}^n c_r\int_r^\infty f_k(x)e^{r-x}dx\\&=\sum_{r=0}^n c_r\int_0^\infty f_k(x+r)e^{-x}dx\\&=\int_0^\infty\left(\sum_{r=0}^n c_rf_k(x+r)\right)e^{-x}dx\end{aligned}$$

On the other hand, integrating such functions is easy, according to:

$$\begin{aligned}\int_0^\infty x^se^{-x}dx&=\int_0^\infty\left(\frac{x^{s+1}}{s+1}\right)'e^{-x}dx\\&=\int_0^\infty\frac{x^{s+1}}{s+1}\,e^{-x}dx\\&=\frac{1}{s+1}\int_0^\infty x^{s+1}e^{-x}dx\end{aligned}$$

Thus, we are led by recurrence on $s\in\mathbb{N}$ to the following formula:

$$\int_0^\infty x^se^{-x}dx=s!$$

For a linear combination now, we are led to the following formula:

$$g(x)=\sum_s a_s x^s \implies \int_0^\infty g(x)e^{-x}dx=\sum_s a_s s!$$

But this shows that we have indeed $P\in\mathbb{Z}$, and also, via a bit of study based on the exact formula of f_k, from the beginning of (1), that we have in fact:

$$\frac{P}{k!}\in\mathbb{Z}$$

Finally, we still have to prove that we have $P\neq 0$, for arbitrarily large numbers $k>>0$. But the point here is that for $k+1>n,|c_0|$, chosen prime, a detailed study of our integral shows that we have $(k+1)\not| P$, and so $P\neq 0$ indeed, as desired.

(4) With this done, let us look now at Q. Our claim is that for $k>>0$ we have:

$$\left|\frac{Q}{k!}\right|<1$$

Indeed, by using the exact formula of f_k, from the beginning of (1), we have:

$$\begin{aligned}f_k(x)e^{-x}&=x^k\left[(x-1)\ldots(x-n)\right]^{k+1}e^{-x}\\&=\left[x(x-1)\ldots(x-n)\right]^k\times(x-1)\ldots(x-n)e^{-x}\end{aligned}$$

We conclude that for $x\in[0,n]$ we have an estimate as follows, with $G,H>0$ being certain constants, appearing as maxima of the two components appearing above:

$$|f_k(x)e^{-x}|<G^kH$$

Now by integrating, we obtain from this the following estimate for Q itself:

$$\begin{aligned}|Q|&=\left|c_1\int_0^1 f_k(x)e^{-x}dx+\ldots+c_ne^n\int_0^n f_k(x)e^{-x}dx\right|\\&\leq|c_1|\int_0^1|f_k(x)e^{-x}|dx+\ldots+|c_n|e^n\int_0^n|f_k(x)e^{-x}|dx\\&\leq|c_1|\cdot G^kH+\ldots+|c_n|e^n\cdot nG^kH\\&=(|c_1|e+\ldots+|c_n|e^n)\frac{n(n+1)}{2}G^kH\end{aligned}$$

But in this estimate the only term depending on k is the power G^k, and since since $k!$ grows much faster than this power G^k, this proves our claim:

$$\left|\frac{Q}{k!}\right|\approx\frac{G^k}{k!}\to 0$$

(5) And with this, done, because what we found in (3,4) contradicts the formula $P+Q=0$ from (2). Thus e is indeed transcendental, as claimed. □

12c. Transcendence of pi

Let us prove now, a bit as for e before, that π is irrational, and even transcendental. Let us start with:

THEOREM 12.3. *The number π is irrational.*

PROOF. This is indeed something quite routine, by using the same ideas as before for e, but with everything being now a bit more technical. □

As a continuation, we have the following result, which is substantially harder:

THEOREM 12.4. *The number π is transcendental.*

PROOF. Again, this is something quite routine, by using the same ideas as before for e, but with everything being now a bit more technical. □

12d. Field theory

Field theory.

12e. Exercises

Exercises:

EXERCISE 12.5.

EXERCISE 12.6.

EXERCISE 12.7.

EXERCISE 12.8.

EXERCISE 12.9.

EXERCISE 12.10.

EXERCISE 12.11.

EXERCISE 12.12.

Bonus exercise.

Part IV

Number theory

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

Primes, revised

13a. Euler estimates

Let us start now a more advanced study of the prime numbers, by improving the Euler formula, that we know well. We have here the following result, to start with:

THEOREM 13.1. *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* $\log(\pi^2/6)=0.49770..$

PROOF. This is something quite straightforward, as follows:

(1) By using the unique factorization $n=p_1^{a_1}\ldots p_k^{a_k}$, we have:

$$\begin{aligned}\prod_{p<N}\left(1-\frac{1}{p}\right)^{-1}&=\prod_{p<N}\left(1+\frac{1}{p}+\frac{1}{p^2}+\frac{1}{p^3}+\ldots\right)\\&>\sum_{n=1}^{N-1}\frac{1}{n}\\&>\int_1^N\frac{1}{x}\,dx\\&=\log N\end{aligned}$$

(2) But the product on the left can be estimated by using log, as follows:

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

(3) Thus, we are led to the estimate in the statement, namely:

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

(4) In order now to improve this, a quick look at what we did in (1) and (2) reveals four $<$ signs, that we can all improve, if we want to. However, we will leave this for later, when talking about Mertens and his theorems. In the meantime, we would like to present a slight improvement, coming via a different technique, which is quite instructive.

(5) The point indeed is that we have a rival method, based by using the factorization $n = p_1 \ldots p_k m^2$, with p_i distinct primes. This factorization gives:

$$\begin{aligned}
\sum_{n=1}^{N-1} \frac{1}{n} &< \prod_{p<N} \left(1 + \frac{1}{p}\right) \sum_{m=1}^{N} \frac{1}{m^2} \\
&< \prod_{p<N} \exp\left(\frac{1}{p}\right) \sum_{m=1}^{\infty} \frac{1}{(m-1/2)(m+1/2)} \\
&= \exp\left(\sum_{p<N} \frac{1}{p}\right) \sum_{m=1}^{\infty} \frac{1}{m-1/2} - \frac{1}{m+1/2} \\
&= 2\exp\left(\sum_{p<N} \frac{1}{p}\right)
\end{aligned}$$

We therefore obtain the following estimate, for our sum:

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

(6) However, $\log 2 = 0.69314..$ does not improve our $1/2$ constant, and we have to be more careful with our telescoping in (5). 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:

$$\sum_{m=1}^{\infty} \frac{1}{m^2} < 1 + \frac{1}{4} + \frac{1}{9} + \frac{2}{7} = \frac{415}{252} \quad , \quad \log\left(\frac{415}{252}\right) = 0.49884..$$

(7) In practice now, in order to finish this discussion, in a professional way, we can invoke the Basel formula, due to Euler, which is however something quite complicated:

$$\sum_{m=1}^{\infty} \frac{1}{m^2} = \frac{\pi^2}{6}$$

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

Although we will not need this here, with the above estimates to be soon improved by theorems of Mertens, let us prove however the formula that we used at the end:

THEOREM 13.2. *We have the following formula, due to Euler,*

$$\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{6}$$

answering the Basel problem, asking for the computation of this sum.

PROOF. This is something quite tricky. The original proof of Euler is as follows, making some manipulations on the Taylor series expansion of $\sin x/x$, based on the fact that the zeroes of this function appear at $x=k\pi$, with $k\in\mathbb{Z}$:

$$\begin{aligned}\frac{\sin x}{x}&=1-\frac{x^2}{3!}+\frac{x^4}{5!}-\frac{x^6}{7!}+\ldots\\&=\left(1-\frac{x}{\pi}\right)\left(1+\frac{x}{\pi}\right)\left(1-\frac{x}{2\pi}\right)\left(1+\frac{x}{2\pi}\right)\ldots\\&=\left(1-\frac{x^2}{\pi^2}\right)\left(1-\frac{x^2}{4\pi^2}\right)\left(1-\frac{x^2}{9\pi^2}\right)\ldots\\&=1-\frac{1}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\,x^2+\ldots\end{aligned}$$

In practice, all this needs a bit more justification, which can be obtained by taking the logarithm, or passing to complex numbers, of even passing to Fourier analysis, and getting the result from the Parseval formula. Exercise for you, to read all this. □

13b. Zeta function

Before moving ahead with the Mertens theorems, substantially improving the above, several comments are in order, with respect to the Euler method. Let us introduce:

DEFINITION 13.3. *Associated to any $s>1$ is the function*

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

called Riemann zeta function.

Observe that the above series converges indeed, as a Riemann sum approximation, by usual rectangles, of the following convergent integral:

$$\begin{aligned}\int_1^{\infty}\frac{1}{x^s}\,dx&=\left[\frac{x^{1-s}}{1-s}\right]_1^{\infty}\\&=0-\frac{1}{1-s}\\&=\frac{1}{s-1}\\&<\infty\end{aligned}$$

Based on this, we can further say that, more generally, the series converges for any $s \in \mathbb{C}$ satisfying $Re(s) > 1$. But more on this, later in this book.

As a first observation, the Basel formula, from Theorem 13.2, reformulates as:

THEOREM 13.4. *We have the following formula, coming from the Basel problem:*

$$\zeta(2) = \frac{\pi^2}{6}$$

More generally, any value $\zeta(2k)$ with $k \in \mathbb{N}$ is a rational multiple of π^{2k}.

PROOF. Here the formula of $\zeta(2)$ is what we have in Theorem 13.2, and the generalization to $\zeta(2k)$ with $k \in \mathbb{N}$ comes by further studying the Euler formula, namely:

$$\frac{\sin x}{x} = \left(1 - \frac{x^2}{\pi^2}\right)\left(1 - \frac{x^2}{4\pi^2}\right)\left(1 - \frac{x^2}{9\pi^2}\right)\ldots$$

To be more precise, after some combinatorial work, that we will not get into here, we are led to the following formula, with B_n being the Bernoulli numbers:

$$\zeta(2k) = (-1)^{k+1}\frac{(2\pi)^{2k}B_{2k}}{2 \cdot (2k)!}$$

In practice, this gives the following fomulae for the first few values $\zeta(2k)$:

$$\zeta(2) = \frac{\pi^2}{6} \quad , \quad \zeta(4) = \frac{\pi^4}{90} \quad , \quad \zeta(6) = \frac{\pi^6}{945} \quad , \quad \zeta(8) = \frac{\pi^8}{9450}$$

As usual, exercise for you to read more about this, as a continuation of the reading suggested in the proof of Theorem 13.2. All first-class mathematics, worth the effort. □

Many other things can be said about zeta, along the same lines, but it is not about this that we want to talk, in this chapter, with all this zeta material being deferred to chapter 15 below. What we want to discuss here is what happens to the Euler estimate from Theorem 13.1, when adding an exponent $s \in \mathbb{R}$ there. Let us start with:

PROPOSITION 13.5. *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 $n = p_1^{a_1} \ldots p_k^{a_k}$, as before, we have:

$$\begin{aligned}\prod_{p<N}\left(1-\frac{1}{p^s}\right)^{-1} &= \prod_{p<N}\left(1+\frac{1}{p^s}+\frac{1}{p^{2s}}+\frac{1}{p^{3s}}+\ldots\right)\\ &> \sum_{n=1}^{N-1}\frac{1}{n^s}\\ &> \int_1^N \frac{1}{x^s}\,dx\end{aligned}$$

But the product on the left can be estimated by using log, as follows:

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

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

In the case $s > 1$, which is the one of main interest, we obtain in this way:

THEOREM 13.6. *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 13.5 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. □

It is possible to futher build along the above lines, but we will leave this discussion for later, in chapter 15, when talking more in detail about the Riemann zeta function.

13c. Mertens theorems

Moving ahead now, the continuation of the story involves the work of Mertens, that we would like to discuss now. Let us start with some analysis conventions:

DEFINITION 13.7. *We use the following notations:*

(1) *We write $f \simeq g$ when $f - g \to 0$.*
(2) *We write $f \cong g$ when $f - g$ is bounded.*
(3) *We write $f \sim g$ when $f/g \to 1$.*
(4) *We write $f \approx g$ when f/g is bounded.*

Occasionaly, we will use as well the Landau $O(f), o(f)$ symbols, making it for 2 notations instead of 4. With these conventions, the formulae of Mertens are as follows:

FACT 13.8. *We have the following Mertens estimates, in the $N \to \infty$ limit,*

$$\sum_{p<N} \frac{\log p}{p} \cong \log N$$

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

$$\sum_{p<N} \log\left(1 - \frac{1}{p}\right) \simeq -\log\log N - \gamma$$

$M = 0.26149..$ and $\gamma = 0.57721..$ being the Mertens and Euler-Mascheroni constants.

Obviously, these formulae are related, and there are many things that can be said here. We will do this slowly. To start with, we would like to talk about the second formula, which improves our Euler estimates before. The precise result here is as follows:

THEOREM 13.9. *We have the following formula, with sum over primes,*

$$\sum_{p\leq N} \frac{1}{p} \simeq \log\log N + M$$

and with $M = 0.26149..$ being a constant, called Mertens constant.

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

(1) As a first comment, observe that we have switched in the statement from sums over primes $p < N$, to sums over primes $p \leq N$. The point is that sums of type $p < N$ were best adapted to the Euler summation, which eventually leads to an integral of $1/x$, that we want to be $\log N$ instead of $\log(N + 1)$. However, as we will see in a moment, the Mertens summation is best written with $p \leq N$. Of course, at the level of the final results, Theorem 13.1 and the present theorem, this does not matter, because:

$$\log\log N \simeq \log\log(N + 1)$$

(2) Getting now to the proof, this is based on the following formula, which comes as usual from the unique factorization of integers, $n = p_1^{a_1} \ldots p_k^{a_k}$, with the sum being over prime powers p^k, and with the exponent $[N/p^k]$ being an integer part:

$$N! = \prod_{p^k\leq N} p^{[N/p^k]}$$

(3) By talking the logarithm, we obtain from this the following estimate:

$$\begin{aligned}\log N! &= \sum_{p^k\leq N}\left[\frac{N}{p^k}\right]\log p\\ &= \sum_{p^k\leq N}\left(\frac{N}{p^k}+o(1)\right)\log p\\ &= N\sum_{p^k\leq N}\frac{\log p}{p^k}+o(1)\sum_{p^k\leq N}\log p\end{aligned}$$

(4) By dividing by N and using $\log N!=N\log N+O(N)$, this gives:

$$\begin{aligned}\sum_{p^k\leq N}\frac{\log p}{p^k} &= \frac{\log N!}{N}+o\left(\frac{1}{N}\right)\sum_{p^k\leq N}\log p\\ &= \log N+o(1)+o\left(\frac{1}{N}\right)\sum_{p^k\leq N}\log p\end{aligned}$$

(5) Now let us analyze the sum on the right. We have:

$$\begin{aligned}\sum_{p^k\leq N}\log p &\leq \sum_{p\in(N,2N]}\log p\\ &\leq \log\binom{2N}{N}\\ &= O(N)\end{aligned}$$

(6) We conclude that the estimate in (4) can be written as follows:

$$\sum_{p^k\leq N}\frac{\log p}{p^k}=\log N+o(1)$$

(7) 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:

$$\sum_{p\leq N}\frac{\log p}{p}=\log N+o(1)$$

(8) 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:

$$\sum_{p\leq N}\frac{1}{p}\simeq\log\log N+M+O\left(\frac{1}{\log N}\right)$$

Thus, we are led to the convergence conclusion in the statement, and of course with the precise numerics for the Mertens constant M remaining to be justified. □

Observe that the above proof crucially uses $\log N! = N \log N + O(N)$. Although we will not really need this, at this point, let us record the following famous result here:

THEOREM 13.10. *We have the Stirling formula*

$$N! \simeq \left(\frac{N}{e}\right)^N \sqrt{2\pi N}$$

valid in the $N \to \infty$ limit.

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

(1) Let us first see what we can get with Riemann sums. We have:

$$\begin{aligned} \log(N!) &= \sum_{k=1}^{N} \log k \\ &\approx \int_1^N \log x \, dx \\ &= N \log N - N + 1 \end{aligned}$$

By exponentiating, this gives the following estimate, which is not bad:

$$N! \approx \left(\frac{N}{e}\right)^N \cdot e$$

(2) We can improve our estimate by replacing the rectangles from the Riemann sum approach to the integrals by trapezoids. In practice, this gives the following estimate:

$$\begin{aligned} \log(N!) &= \sum_{k=1}^{N} \log k \\ &\approx \int_1^N \log x \, dx + \frac{\log 1 + \log N}{2} \\ &= N \log N - N + 1 + \frac{\log N}{2} \end{aligned}$$

By exponentiating, this gives the following estimate, which gets us closer:

$$N! \approx \left(\frac{N}{e}\right)^N \cdot e \cdot \sqrt{N}$$

(3) In order to conclude, we must take some kind of mathematical magnifier, and carefully estimate the error made in (2). Fortunately, this mathematical magnifier exists, called Euler-Maclaurin formula, and after some 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:

$$N!=\int_0^\infty x^Ne^{-x}dx$$

(5) Since the integrand is sharply peaked at $x=N$, as you can see by computing the derivative of $\log(x^Ne^{-x})$, this suggests writing $x=N+y$, and we obtain:

$$\begin{aligned}\log(x^Ne^{-x}) &= N\log x-x\\ &= N\log(N+y)-(N+y)\\ &= N\log N+N\log\left(1+\frac{y}{N}\right)-(N+y)\\ &\simeq N\log N+N\left(\frac{y}{N}-\frac{y^2}{2N^2}\right)-(N+y)\\ &= N\log N-N-\frac{y^2}{2N}\end{aligned}$$

(6) By exponentiating, we obtain from this the following estimate:

$$x^Ne^{-x}\simeq\left(\frac{N}{e}\right)^Ne^{-y^2/2N}$$

(7) Now by integrating, we obtain from this the following estimate:

$$\begin{aligned}N! &= \int_0^\infty x^Ne^{-x}dx\\ &\simeq \int_{-N}^N\left(\frac{N}{e}\right)^Ne^{-y^2/2N}\,dy\\ &\simeq \left(\frac{N}{e}\right)^N\int_{\mathbb{R}}e^{-y^2/2N}\,dy\\ &= \left(\frac{N}{e}\right)^N\sqrt{2N}\int_{\mathbb{R}}e^{-z^2}\,dz\\ &= \left(\frac{N}{e}\right)^N\sqrt{2\pi N}\end{aligned}$$

(8) Here we have used at the end the following key formula, due to Gauss:

$$\begin{aligned}\left(\int_{\mathbb{R}} e^{-z^2}dz\right)^2 &= \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-x^2-y^2}dxdy\\ &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2}rdrdt\\ &= 2\pi\int_0^{\infty}\left(-\frac{e^{-r^2}}{2}\right)'dr\\ &= 2\pi\left[0-\left(-\frac{1}{2}\right)\right]\\ &= \pi\end{aligned}$$

Thus, we have proved the Stirling formula, as formulated in the statement. □

Now back to the Mertens second theorem, the continuation of the story, involving Mertens, Meissel and others, is quite long. The Mertens proof can be of course improved, with some technical bounds for M, and for the rate of convergence too.

However, skipping this discussion, which is quite technical, and getting to the point, the Mertens constant M itself, there are several interesting formulae for it. According to Theorem 13.9, this constant appears by definition as follows:

$$M=\lim_{N\to\infty}\sum_{p<N}\frac{1}{p}-\log\log N$$

In order to further build on this, we will need the following standard result:

THEOREM 13.11. *The following limit converges,*

$$\gamma=\lim_{N\to\infty}\sum_{n=1}^{N}\frac{1}{n}-\log N$$

the result being the Euler-Mascheroni constant $\gamma=0.57721..$

PROOF. This is indeed something very standard, coming from basic calculus. In addition to the formula in the statement, there is a bewildering quantity of alternative formulae for γ, all being useful when doing number theory, which are as follows:

(1) First, we have the following alternative formula:

$$\gamma=-\int_0^{\infty}e^{-x}\log x\,dx$$

With a change of variables, this is equivalent to the following formula:

$$\gamma = -\int_0^1 \log\left(\log\frac{1}{x}\right)dx$$

(2) We have as well the following formula, with [.] being the integer part:

$$\gamma = \int_1^\infty \frac{1}{[x]} - \frac{1}{x}\,dx$$

Alternatively, in terms of the upper integer part [[.]], we have:

$$\gamma = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\left[\left[\frac{n}{k}\right]\right] - \frac{n}{k}$$

(3) In relation with the gamma function, we have the following formula:

$$\gamma = -\Gamma'(1)$$

Equivalently, still in terms of the gamma function, we have the following formula:

$$\gamma = \lim_{z\to 0}\frac{1}{z} - \Gamma(z)$$

As a third formula for γ, still in terms of the gamma function, we have:

$$\gamma = \lim_{z\to 0}\frac{1}{2z}\left(\frac{1}{\Gamma(1+z)} - \frac{1}{\Gamma(1-z)}\right)$$

(4) In relation now with the zeta function, we have the following formula:

$$\gamma = \sum_{n=2}^{\infty}(-1)^n\frac{\zeta(n)}{n}$$

Alternatively, still in terms of zeta, we have the following formula:

$$\gamma = \log\left(\frac{4}{\pi}\right) + \sum_{n=2}^{\infty}(-1)^n\frac{\zeta(n)}{2^{n-1}n}$$

(5) We have as well the following alternative formula:

$$\gamma = \lim_{s\to 1^+}\sum_{n=1}^{\infty}\frac{1}{n^s} - \frac{1}{s^n}$$

In terms of the zeta function, this latter formula simply reads:

$$\gamma = \lim_{s\to 1}\zeta(s) - \frac{1}{s-1}$$

Alternatively, still in terms of the zeta function around 1, this reads:

$$\gamma = \lim_{s\to 0}\frac{\zeta(1+s)+\zeta(1-s)}{2}$$

(6) And as usual, exercise for you to do the calculus for all this, or of course look it up, in case the calculus turns too complicated. □

Now back to the Mertens constant, we have the following formula for it:

THEOREM 13.12. *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 $\gamma=0.57721..$ *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 the Euler-Mascheroni constant is related as well to the primes, as follows:

$$\sum_{p<N}\log\left(1-\frac{1}{p}\right)\simeq-\log\log N-\gamma$$

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

Getting back now to the Mertens theorem, the above considerations eventually lead, via some more work, to the precise numeric figure from Theorem 13.9, namely:

$$M=0.26149..$$

Changing topics now, as already mentioned in the above, Mertens proved in fact three theorems regarding the prime numbers, with Theorem 13.9, the most famous one, being his second theorem. His first theorem is a related formula, as follows:

THEOREM 13.13. *We have the following formula,*

$$\sum_{p<N}\frac{\log p}{p}\cong\log N$$

with the sum being over primes.

PROOF. This is indeed something quite standard, 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 all this. □

As for the third theorem of Mertens, again related to all this, this is as follows:

THEOREM 13.14. *We have the following formula,*

$$\prod_{p<N}\left(1-\frac{1}{p}\right)\approx\frac{e^{-\gamma}}{\log N}$$

with the product being over primes.

PROOF. In order to establish the result, we can use the following formula:

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

Indeed, this gives the following formula for the product in the statement:

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

Now by inverting and applying the logarithm, we obtain:

$$\begin{aligned}
\log\left[\prod_{p<N}\left(1-\frac{1}{p}\right)^{-1}\right] &= \log\left[\prod_{p<N}\left(1-\frac{1}{p^2}\right)^{-1}\right]+\log\left[\prod_{p<N}\left(1-\frac{1}{p}\right)\right]\\
&= \log\left[\prod_{p<N}\left(1+\frac{1}{p^2}+\frac{1}{p^4}+\ldots\right)\right]+\sum_{p<N}\log\left(1-\frac{1}{p}\right)\\
&\simeq \log\left[\sum_{n=1}^{\infty}\frac{1}{n^2}\right]+\sum_{p<N}\log\left(1-\frac{1}{p}\right)\\
&= \frac{\pi^2}{6}+\sum_{p<N}\log\left(1-\frac{1}{p}\right)\\
&\simeq \frac{\pi^2}{6}-\log\log N-\gamma
\end{aligned}$$

Now by exponentiating, we are led to the conclusion in the statement: □

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

13d. Chebycheff estimates

Let us investigate now some related questions, again regarding the primes and their distribution, which look more intuitive and appealing, but which in the end, require more complicated techniques. We would like to estimate the following number:

DEFINITION 13.15. *We define the function* $\pi:\mathbb{N}\to\mathbb{N}$ *by*

$$\pi(N)=\#\Big\{p\leq N\text{ prime}\Big\}$$

the first few values being $0,0,1,2,2,3,3,4,4,4,4,5,5,6,6,6,6,\ldots$

Many things can be said here, especially now that we are already quite seriously into prime numbers, with the Euler estimates, and the theorems of Mertens, which can be converted into results about $\pi(N)$. However, according to our general policy for this opening chapter on analysis, let us do things slowly. To start with, we have:

PROPOSITION 13.16. *We have the following estimate,*

$$\pi(N) \geq \log\log N$$

coming from the unique factorization of integers, $n = p_1^{a_1} \ldots p_k^{a_k}$.

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:

$$p_{n+1} \leq p_1 \ldots p_n + 1$$

But this gives, by recurrence on n, the following estimate:

$$p_n \leq 2^{2^n}$$

In terms of the function π from Definition 13.15, 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. □

Next in line, we have the following estimate, heavily improving Proposition 13.16:

PROPOSITION 13.17. *We have the following estimate,*

$$\pi(N) \geq \frac{\log N}{\log 4}$$

coming from the unique factorization $n = p_1 \ldots p_k m^2$*, with* p_i *distinct.*

PROOF. This is again something that I learned from my algebra colleagues. Consider the first n primes, denoted $p_1, \ldots, p_n$, and let us try to compute the number $f(N)$ of integers $K \leq N$ all whose prime factors are among $\{p_1, \ldots, p_n\}$. 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. □

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 13.18. *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 13.19. *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:

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

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 13.18, this gives:

$$\theta(2n) - \theta(n) < \log 4 \cdot n$$

Now by summing, we are led to the formula in the statement. □

We can now formulate a first key theorem of Chebycheff, as follows:

THEOREM 13.20. *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 13.19, 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 13.21. *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 13.22. *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 13.20 and Theorem 13.21, 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. □

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.

13e. Exercises

Exercises:

Exercise 13.23.

Exercise 13.24.

Exercise 13.25.

Exercise 13.26.

Exercise 13.27.

Exercise 13.28.

Exercise 13.29.

Exercise 13.30.

Bonus exercise.

CHAPTER 14

Complex analysis

14a. Complex functions

We discuss in this chapter the theory of complex functions $f : \mathbb{C} \to \mathbb{C}$, in analogy with the theory of the real functions $f : \mathbb{R} \to \mathbb{R}$, that we will need later. We will see that many results that we know from the real setting extend to the complex setting, but there will be quite a number of new phenomena too. We will need, in order to get started:

DEFINITION 14.1. *The distance between two complex numbers is the usual distance in the plane between them, namely:*

$$d(x, y) = |x - y|$$

With this, we can talk about convergence, by saying that $x_n \to x$ when $d(x_n, x) \to 0$.

Here the fact that $d(x, y) = |x-y|$ is indeed the usual distance in the plane is clear for $y = 0$, because we have $d(x, 0) = |x|$, by definition of the modulus $|x|$. As for the general case, $y \in \mathbb{C}$, this comes from the fact that the distance in the plane is given by:

$$d(x, y) = d(x - y, 0) = |x - y|$$

Observe that in real coordinates, the distance formula is quite complicated, namely:

$$\begin{aligned} d(a + ib, c + id) &= |(a + ib) - (c + id)| \\ &= |(a - c) + i(b - d)| \\ &= \sqrt{(a - c)^2 + (b - d)^2} \end{aligned}$$

However, for most computations, we will not need this formula, and we can get away with the various tricks regarding complex numbers that we know. As a first result now, regarding $\mathbb{C}$ and its distance, that we will need in what follows, we have:

PROPOSITION 14.2. *The complex plane $\mathbb{C}$ is complete, in the sense that any Cauchy sequence converges.*

PROOF. Consider indeed a Cauchy sequence $\{x_n\}_{n\in\mathbb{N}} \subset \mathbb{C}$. If we write $x_n = a_n + ib_n$ for any $n \in \mathbb{N}$, then we have the following estimates:

$$|a_n - a_m| \leq \sqrt{(a_n - a_m)^2 + (b_n - b_m)^2} = |x_n - x_m|$$

$$|b_n - b_m| \leq \sqrt{(a_n - a_m)^2 + (b_n - b_m)^2} = |x_n - x_m|$$

Thus both the sequences $\{a_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ and $\{b_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ are Cauchy, and since we know that $\mathbb{R}$ itself is complete, we can consider the limits of these sequences:

$$a_n \to a \quad , \quad b_n \to b$$

With $x = a + ib$, our claim is that $x_n \to x$. Indeed, we have:

$$\begin{aligned} |x_n - x| &= \sqrt{(a_n - a)^2 + (b_n - b)^2} \\ &\leq |a_n - a| + |b_n - b| \end{aligned}$$

It follows that we have $x_n \to x$, as claimed, and this gives the result. □

Talking complex functions now, we have the following definition:

DEFINITION 14.3. *A complex function $f : \mathbb{C} \to \mathbb{C}$, or more generally $f : X \to \mathbb{C}$, with $X \subset \mathbb{C}$ being a subset, is called continuous when, for any $x_n, x \in X$:*

$$x_n \to x \implies f(x_n) \to f(x)$$

Also, we can talk about pointwise convergence of functions, $f_n \to f$, and about uniform convergence too, $f_n \to_u f$, exactly as for the real functions.

Observe that, since $x_n \to x$ in the complex sense means that $(a_n, b_n) \to (a, b)$ in the usual, real plane sense, a function $f : \mathbb{C} \to \mathbb{C}$ is continuous precisely when it is continuous when regarded as real function, $f : \mathbb{R}^2 \to \mathbb{R}^2$. But more on this later in this book. At the level of examples now, we first have the polynomials, $P \in \mathbb{C}[X]$. We already met such polynomials in the above, so let us recall from there that we have:

THEOREM 14.4. *Each polynomial $P \in \mathbb{C}[X]$ can be regarded as a continuous function $P : \mathbb{C} \to \mathbb{C}$. Moreover, we have the formula*

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

with $a \in \mathbb{C}$, and with the numbers $r_1, \ldots, r_n \in \mathbb{C}$ being the roots of P.

PROOF. This is something that we know well, the idea being that one root can be always constructed, by reasoning by contradiction, and doing some analysis around the minimum of $|P|$, and then a recurrence on the degree $n \in \mathbb{N}$ does the rest. □

Next in line, we have the rational functions, which are defined as follows:

THEOREM 14.5. *The quotients of complex polynomials $f = P/Q$ are called rational funtions. When written in reduced form, with P, Q prime to each other,*

$$f = \frac{P}{Q}$$

is well-defined and continuous outside the zeroes $P_f \subset \mathbb{C}$ of Q, called poles of f:

$$f : \mathbb{C} - P_f \to \mathbb{C}$$

In addition, the rational functions, regarded as algebraic expressions, are stable under summing, making products and taking inverses.

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

(1) First of all, we can surely talk about quotients of polynomials, $f = P/Q$, regarded as abstract algebraic expressions. Also, the last assertion is clear, because we can indeed perform sums, products, and take inverses, by using the following formulae:

$$\frac{P}{Q} + \frac{R}{S} = \frac{PS + QR}{QS} \quad , \quad \frac{P}{Q} \cdot \frac{R}{S} = \frac{PR}{QS} \quad , \quad \left(\frac{P}{Q}\right)^{-1} = \frac{Q}{P}$$

(2) The question is now, given a rational function f, can we regard it as a complex function? In general, we cannot say that we have $f : \mathbb{C} \to \mathbb{C}$, for instance because $f(x) = x^{-1}$ is not defined at $x = 0$. More generally, assuming $f = P/Q$ with $P, Q \in \mathbb{C}$, we cannot talk about $f(x)$ when x is a root of Q, unless of course we are in the special situation where x is a root of P too, and we can simplify the fraction.

(3) In view of this discussion, in order to solve our question, we must avoid the situation where the polynomials P, Q have common roots. But this can be done by writing our rational function f in reduced form, as follows, with $P, Q \in \mathbb{C}[X]$ prime to each other:

$$f = \frac{P}{Q}$$

(4) Now with this convention made, it is clear that f is well-defined, and continuous too, outside of the zeroes of f. Now since these zeroes can be obviously recovered from the knowledge of f itself, as being the points where "f explodes", we can call them poles of f, and so we have a function $f : \mathbb{C} - P_f \to \mathbb{C}$, as in the statement. □

As a comment here, the term "pole" does not come from the Poles who invented this, but rather from the fact that, when trying to draw the graph of f, or rather imagine that graph, which takes place in $2+2 = 4$ real dimensions, we are faced with some sort of tent, which is suspended by infinite poles, which lie, guess where, at the poles of f.

Getting back now to Theorem 14.5, this tells us that the rational functions form a field. This is quite interesting, and opposite to the general spirit of analysis and function spaces, which are in general not fields. Let us record this finding, as follows:

DEFINITION 14.6. *We denote by $\mathbb{C}(X)$ the field of rational functions*

$$f = \frac{P}{Q} \quad , \quad P, Q \in \mathbb{C}[X]$$

with the usual sum and product operations $+$ and $\times$ for the rational functions.

To be more precise, this is some sort of reformulation of Theorem 14.5, or rather of the algebraic content of Theorem 14.5, telling us that the rational functions form indeed a field. And to the question, how can a theorem suddenly become a definition, the answer is that this is quite commonplace in mathematics, and especially in algebra.

Back now to analysis, let us point out that, contrary to what the above might suggest, everything does not always extend trivially from the real to the complex case. For instance, we have the following result, that we already talked about a bit before:

PROPOSITION 14.7. *We have the following formula, valid for any $|x| < 1$,*

$$\frac{1}{1-x} = 1 + x + x^2 + \ldots$$

but, for $x \in \mathbb{C} - \mathbb{R}$, the geometric meaning of this formula is quite unclear.

PROOF. Here the formula in the statement holds indeed, by multiplying and cancelling terms, exactly as in the real case, with the convergence being justified by:

$$\left| \sum_{n=0}^{\infty} x^n \right| \leq \sum_{n=0}^{\infty} |x|^n = \frac{1}{1-|x|}$$

As for the last assertion, this is something rather informal, which hides however many interesting things, that we discussed in some detail in the above. □

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 14.8. *The functions* $\sin, \cos, \exp, \log$ *have complex extensions, given by*

$$\sin x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l+1}}{(2l+1)!} \quad , \quad \cos x = \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l}}{(2l)!}$$

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

with $|x| < 1$ needed for log*, which are continuous over their domain, and satisfy the formulae $e^{x+y} = e^x e^y$ and $e^{ix} = \cos x + i \sin x$.*

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

(1) We already know about e^x from before, the idea being that the convergence of the series, and then the continuity of e^x, come from the following estimate:

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

(2) Regarding $\sin x$, the same method works, with the following estimate:

$$|\sin x| \leq \sum_{l=0}^{\infty} \frac{|x|^{2l+1}}{(2l+1)!} \leq \sum_{k=0}^{\infty} \frac{|x|^k}{k!} = e^{|x|}$$

(3) The same goes for $\cos x$, the estimate here being as follows:

$$|\cos x| \leq \sum_{l=0}^{\infty} \frac{|x|^{2l}}{(2l)!} \leq \sum_{k=0}^{\infty} \frac{|x|^k}{k!} = e^{|x|}$$

(4) Regarding now the formulae satisfied by $\sin, \cos, \exp$, we already know from before that the exponential has the following property, exactly as in the real case:

$$e^{x+y} = e^x e^y$$

We also have the following formula, connecting $\sin, \cos, \exp$, again as before:

$$\begin{aligned} e^{ix} &= \sum_{k=0}^{\infty} \frac{(ix)^k}{k!} \\ &= \sum_{k=2l} \frac{(ix)^k}{k!} + \sum_{k=2l+1} \frac{(ix)^k}{k!} \\ &= \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l}}{(2l)!} + i \sum_{l=0}^{\infty} (-1)^l \frac{x^{2l+1}}{(2l+1)!} \\ &= \cos x + i \sin x \end{aligned}$$

(5) In order to discuss now the complex logarithm function log, let us first study some more the complex exponential function exp. By using $e^{x+y} = e^x e^y$ we obtain $e^x \neq 0$ for any $x \in \mathbb{C}$, so the complex exponential function is as follows:

$$\exp : \mathbb{C} \to \mathbb{C} - \{0\}$$

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 $\exp : \mathbb{C} \to \mathbb{C} - \{0\}$ is surjective. Also, again by using $e^{x+iy} = e^x e^{iy}$, we deduce that we have:

$$e^x = e^y \iff x - y \in 2\pi i \mathbb{Z}$$

(6) With these ingredients in hand, we can now talk about log. Indeed, let us fix a horizontal strip in the complex plane, having width 2π:

$$S = \left\{ x + iy \middle| x \in \mathbb{R}, y \in [a, a + 2\pi) \right\}$$

We know from the above that the restriction map $\exp : S \to \mathbb{C} - \{0\}$ is bijective, so we can define log as to be the inverse of this map:

$$\log = \exp^{-1} : \mathbb{C} - \{0\} \to S$$

(7) In practice now, the best is to choose for instance $a = 0$, or $a = -\pi$, as to have the whole real line included in our strip, $\mathbb{R} \subset S$. In this case on $\mathbb{R}_+$ we recover the usual logarithm, while on $\mathbb{R}_-$ we obtain complex values, as for instance $\log(-1) = \pi i$ in the case $a = 0$, or $\log(-1) = -\pi i$ in the case $a = -\pi$, coming from $e^{\pi i} = -1$.

(8) Finally, assuming $|x| < 1$, we can consider the following series, which converges:

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

We have then $e^{f(x)} = 1 + x$, and so $f(x) = \log(1 + x)$, when $1 + x \in S$. □

Moving ahead, Theorem 14.8 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:

PROPOSITION 14.9. *The following functions, called hyperbolic sine and cosine,*

$$\sinh x = \frac{e^x - e^{-x}}{2} \quad , \quad \cosh x = \frac{e^x + e^{-x}}{2}$$

are subject to the following formulae:

(1) $e^x = \cosh x + \sinh x$.
(2) $\sinh(ix) = i \sin x$, $\cosh(ix) = \cos x$, *for* $x \in \mathbb{R}$.
(3) $\sinh(x + y) = \sinh x \cosh y + \cosh x \sinh y$.
(4) $\cosh(x + y) = \cosh x \cosh y + \sinh x \sinh y$.
(5) $\sinh x = \sum_l \frac{x^{2l+1}}{(2l+1)!}$, $\cosh x = \sum_l \frac{x^{2l}}{(2l)!}$.

PROOF. The formula (1) follows from definitions. As for (2), this follows from:

$$\sinh(ix) = \frac{e^{ix} - e^{-ix}}{2} = \frac{\cos x + i \sin x}{2} - \frac{\cos x - i \sin x}{2} = i \sin x$$

$$\cosh(ix) = \frac{e^{ix} + e^{-ix}}{2} = \frac{\cos x + i \sin x}{2} + \frac{\cos x - i \sin x}{2} = \cos x$$

Regarding now (3,4), observe first that the formula $e^{x+y} = e^x + e^y$ reads:

$$\cosh(x + y) + \sinh(x + y) = (\cosh x + \sinh x)(\cosh y + \sinh y)$$

Thus, we have some good explanation for (3,4), and in practice, these formulae can be checked by direct computation, as follows:

$$\frac{e^{x+y} - e^{-x-y}}{2} = \frac{e^x - e^{-x}}{2} \cdot \frac{e^y + e^{-y}}{2} + \frac{e^x + e^{-x}}{2} \cdot \frac{e^y - e^{-y}}{2}$$

$$\frac{e^{x+y} + e^{-x-y}}{2} = \frac{e^x + e^{-x}}{2} \cdot \frac{e^y + e^{-y}}{2} + \frac{e^x - e^{-x}}{2} \cdot \frac{e^y - e^{-y}}{2}$$

Finally, (5) is clear from the definition of sinh, cosh, and from $e^x = \sum_k \frac{x^k}{k!}$. □

Finally, we can talk as well about powers, in the following way:

FACT 14.10. *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 before 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 the problems left open above, with bare hands. More later.

At the level of the general theory now, the main tool for dealing with the continuous functions $f : \mathbb{R} \to \mathbb{R}$ was the intermediate value theorem. In the complex setting, that of the functions $f : \mathbb{C} \to \mathbb{C}$, we do not have such a theorem, at least in its basic formulation, because there is no order relation for the complex numbers, or things like complex intervals. However, the intermediate value theorem in its advanced formulation, that with connected sets, extends of course, and we have the following result:

THEOREM 14.11. *Assuming that $f : X \to \mathbb{C}$ with $X \subset \mathbb{C}$ is continuous, if the domain X is connected, then so is its image $f(X)$.*

PROOF. This follows exactly as in the real case, with just a bit of discussion being needed, in relation with open and closed sets, and then connected sets, inside $\mathbb{C}$. □

14b. Holomorphic functions

Let us study now the differentiability of the complex functions $f : \mathbb{C} \to \mathbb{C}$. Things here are quite tricky, but let us start with a straightforward definition, as follows:

DEFINITION 14.12. *We say that a function $f : X \to \mathbb{C}$ is differentiable in the complex sense when the following limit is defined for any $x \in X$:*

$$f'(x) = \lim_{t \to 0} \frac{f(x + t) - f(x)}{t}$$

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:

$$\begin{aligned}
(x^n)' &= \lim_{t \to 0} \frac{(x + t)^n - x^n}{t} \\
&= \lim_{t \to 0} \frac{nx^{n-1}t + \binom{n}{2} x^{n-2}t^2 + \ldots + t^n}{t} \\
&= \lim_{t \to 0} \frac{nx^{n-1}t}{t} \\
&= nx^{n-1}
\end{aligned}$$

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 some theory. The general results from the real case extend well, as follows:

PROPOSITION 14.13. *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. □

As an obvious consequence of (1,2) above, any poynomial $P \in \mathbb{C}[X]$ 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, outsides 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 14.14. *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. □

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 14.15. *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 $t \to 0$.*

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 $t \to 0$, 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 $r \in \mathbb{R}$, $r \to 0$, we can obtain as limit any number on the unit circle:

$$\lim_{r \to 0} \frac{f(x + rw) - f(x)}{rw} = \lim_{r \to 0} \frac{r\bar{w}}{rw} = \bar{w}^2$$

(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 $r \to 0$ we obtain a certain multiple of $\bar{w}$, with the multiplication factor being computed as follows:

$$\begin{aligned} \frac{|x + rw| - |x|}{r} &= \frac{|x + rw|^2 - |x|^2}{(|x + rw| + |x|)r} \\ &\simeq \frac{xr\bar{w} + \bar{x}rw}{2|x|r} \\ &= Re\left(\frac{x\bar{w}}{|x|}\right) \end{aligned}$$

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. □

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 14.16. *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 14.15. So, based on that computations, let us formulate the following definition, coming as a complement to Definition 14.12:

DEFINITION 14.17. *A function $f : X \to \mathbb{C}$ is called differentiable:*

(1) *In the real sense, if the following two limits converge, for any $x \in X$:*

$$f_1'(x) = \lim_{t \in \mathbb{R} \to 0} \frac{f(x+t) - f(x)}{t} \quad , \quad f_i'(x) = \lim_{t \in i\mathbb{R} \to 0} \frac{f(x+t) - f(x)}{t}$$

(2) *In a radial sense, if the following limit converges, for any $x \in X$, and $w \in \mathbb{T}$:*

$$f_w'(x) = \lim_{t \in w\mathbb{R} \to 0} \frac{f(x+t) - f(x)}{t}$$

(3) *In the complex sense, if the following limit converges, for any $x \in X$:*

$$f'(x) = \lim_{t \to 0} \frac{f(x+t) - f(x)}{t}$$

If f is differentiable in the complex sense, we also say that f is holomorphic.

We can see now more clearly what is going on. We have (3) $\implies$ (2) $\implies$ (1) in general, and most of the functions that we know, namely the polynomials, the rational functions, and $\sin, \cos, \exp, \log$, satisfy (3). As for the functions $\bar{x}, |x|$, these do not satisfy (3), and do not satisfy (2) either, but they satisfy however (1). It is possible to say more about all this, and we will certainly come back to this topic, later in this book.

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 14.18. *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:

$$R = \frac{1}{C} \quad , \quad C = \limsup_{n \to \infty} \sqrt[n]{|c_n|}$$

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 1, which says that a series $\sum_n x_n$ converges if $c < 1$, and diverges if $c > 1$, where:

$$c = \limsup_{n \to \infty} \sqrt[n]{|x_n|}$$

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 $R = 1$. □

Back now to our questions regarding derivatives, we have:

THEOREM 14.19. *Assuming that a function $f : X \to \mathbb{C}$ 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 $\sqrt[n]{n} \to 1$:

$$\frac{1}{R'} = \limsup_{n \to \infty} \sqrt[n]{|nc_n|} = \limsup_{n \to \infty} \sqrt[n]{|c_n|} = \frac{1}{R}$$

Thus f' exists and is analytic, on the same domain, and this gives the result. □

14c. 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 14.20. *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. □

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 14.21. *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. □

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 14.22. *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 $\gamma : [a, b] \to \mathbb{C}$.

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 $\gamma, \eta : [a, b] \to \mathbb{C}$ 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 $\phi : [a, b] \to [a, b]$ 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. □

The main properties of the above integration method are as follows:

PROPOSITION 14.23. *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 14.22. □

Now by getting back to polynomials, we have the following result:

THEOREM 14.24. *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 14.21. 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:

$$\gamma : [0, 2\pi] \to \mathbb{C} \quad , \quad \gamma(t) = x + re^{it}$$

By using now the computation from the proof of Theorem 6.23, 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. □

All this is quite interesting, and obviously, we are now into some serious mathematics. Importantly, Theorem 14.20, Theorem 14.21 and Theorem 14.24 provide us with a path for proving the converse of Theorem 14.19. Indeed, if we manage to prove the Cauchy formula for any holomorphic function $f : X \to \mathbb{C}$, then it will follow that our function is analytic, and so infinitely differentiable. So, let us start with the following result:

THEOREM 14.25. *The Cauchy formula, namely*

$$f(x) = \frac{1}{2\pi i}\int_\gamma \frac{f(y)}{y-x}\,dy$$

holds for any holomorphic function $f : X \to \mathbb{C}$.

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:

$$\begin{aligned}\left|\int_{\Delta^{(n)}} f(x)dx\right| &= \left|\int_{\Delta^{(n)}} f(x) - f(z) - f'(z)(x-z)dx\right|\\ &\leq \int_{\Delta^{(n)}} \left|f(x) - f(z) - f'(z)(x-z)\right| dx\\ &\leq \int_{\Delta^{(n)}} \varepsilon|x-z|dx\\ &\leq \varepsilon(2^{-n}L)^2\end{aligned}$$

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 [80]. □

As a main application of the Cauchy formula, we have:

THEOREM 14.26. *The following conditions are equivalent, for a function $f : X \to \mathbb{C}$:*

(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 14.25.

(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 14.19. □

As another application of the Cauchy formula, we have:

THEOREM 14.27. *Any holomorphic function $f : X \to \mathbb{C}$ 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. □

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 14.28. *Given an holomorphic function $f : X \to \mathbb{C}$, 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 14.22 to integrals viewed as averages, which are those that we refer to, in the present statement. □

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 14.29. *An entire, bounded holomorphic function*

$$f : \mathbb{C} \to \mathbb{C} \quad , \quad |f| \leq M$$

must be constant. In particular, knowing $f \to 0$ with $z \to \infty$ 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 14.28, 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 $R >> 0$:

$$f(x) = \int_D f(z)dz \quad , \quad f(y) = \int_E f(z)dz$$

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. □

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

14d. Further results

Further results.

14e. Exercises

Exercises:

Exercise 14.30.

Exercise 14.31.

Exercise 14.32.

Exercise 14.33.

Exercise 14.34.

Exercise 14.35.

Exercise 14.36.

Exercise 14.37.

Bonus exercise.

CHAPTER 15

Zeta function

15a. Real zeta

We have already met the Riemann zeta function on several occasions, in the above, at values $s > 1$ of the parameter, with the conclusion every time that this function is intimately related to the primes. In this chapter we discuss a systematic approach to this phenomenon, by using complex analysis. As a first observation, we can talk without much pain about zeta at complex values of s as well, in the following way:

THEOREM 15.1. *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 $Re(z) > 1$.

PROOF. We have the following computation, assuming $s = r + it$ with $r > 1$:

$$\begin{aligned}
|\zeta(s)| &= \left|\sum_{n=1}^{\infty} \frac{1}{n^s}\right| \\
&\leq \sum_{n=1}^{\infty} \frac{1}{|n^s|} \\
&\leq \sum_{n=1}^{\infty} \frac{1}{n^r} \\
&< 1 + \int_1^{\infty} \frac{1}{x^r}\, dx \\
&= 1 + \left[\frac{x^{1-r}}{1-r}\right]_1^{\infty} \\
&= 1 + \frac{1}{r-1}
\end{aligned}$$

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

As a first result, we can write zeta as an Euler product, as follows:

PROPOSITION 15.2. *We have the following formula,*

$$\zeta(s)=\prod_p\left(1-\frac{1}{p^s}\right)^{-1}$$

valid for any exponent $s\in\mathbb{C}$ with $Re(s)>1$.

PROOF. We have the following computation, with everything converging:

$$\begin{aligned}\zeta(s)&=\sum_{n=1}^\infty\frac{1}{n^s}\\&=\prod_p\left(1+\frac{1}{p^s}+\frac{1}{p^{2s}}+\frac{1}{p^{3s}}+\ldots\right)\\&=\prod_p\left(1-\frac{1}{p^s}\right)^{-1}\end{aligned}$$

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

We have as well the following formula, which is elementary too:

PROPOSITION 15.3. *We have the following formula,*

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

with μ being the Möbius function, given by the formula

$$\mu(n)=\begin{cases}(-1)^k&\text{if }n=p_1\ldots p_k\\0&\text{if }n\text{ is not square}-\text{free}\end{cases}$$

valid for any exponent $s\in\mathbb{C}$ with $Re(s)>1$.

PROOF. We have the following computation, with everything converging:

$$\begin{aligned}\frac{1}{\zeta(s)}&=\prod_p\left(1-\frac{1}{p^s}\right)\\&=\sum_{k=0}^\infty(-1)^k\prod_{p_1\ldots p_k}\frac{1}{p_1^s\ldots p_k^s}\\&=\sum_{n=1}^\infty\frac{\mu(n)}{n^s}\end{aligned}$$

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

Along the same lines, as another elementary result, we have:

PROPOSITION 15.4. *The square of the zeta function is given by*

$$\zeta^2(s) = \sum_{n=1}^{\infty} \frac{\tau(n)}{n^s}$$

with $\tau(n)$ being the number of divisors of n, for any $s \in \mathbb{C}$ with $Re(s) > 1$.

PROOF. We have the following computation, with everything converging:

$$\zeta(s)^2 = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \frac{1}{(kl)^s} = \sum_{n=1}^{\infty} \frac{\tau(n)}{n^s}$$

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

In order to present now a more advanced result, we will need:

PROPOSITION 15.5. *We can talk about the gamma function*

$$\Gamma(s) = \int_0^{\infty} x^{s-1} e^{-x}\, dx$$

extending the usual factorial of integers, $\Gamma(s) = (s-1)!$.

PROOF. The integral converges indeed, and by partial integration we have:

$$\begin{aligned} \Gamma(s+1) &= \int_0^{\infty} x^s e^{-x}\, dx \\ &= \int_0^{\infty} s x^{s-1} e^{-x}\, dx \\ &= s\Gamma(s) \end{aligned}$$

Regarding now the case $s \in \mathbb{N}$, for the initial value $s = 1$ we have:

$$\Gamma(1) = \int_0^{\infty} e^{-x} dx = 1$$

Thus, for $s \in \mathbb{N}$ we have indeed $\Gamma(s) = (s-1)!$, as claimed. □

We can now formulate a key result about zeta, as follows:

THEOREM 15.6. *We have the following formula,*

$$\zeta(s) = \frac{1}{\Gamma(s)} \int_0^{\infty} \frac{x^{s-1}}{e^x - 1}\, dx$$

valid for any $s \in \mathbb{C}$ with $Re(s) > 1$.

PROOF. We have indeed the following computation:

$$\begin{aligned}
\int_0^\infty \frac{x^{s-1}}{e^x-1}\,dx &= \int_0^\infty \frac{x^{s-1}}{e^x}\cdot\frac{1}{1-e^{-x}}\,dx\\
&= \int_0^\infty x^{s-1}(e^{-x}+e^{-2x}+e^{-3x}+\ldots)\\
&= \sum_{n=1}^\infty \int_0^\infty x^{s-1}e^{-nx}\,dx\\
&= \sum_{n=1}^\infty \int_0^\infty \left(\frac{y}{n}\right)^{s-1} e^{-y}\,\frac{dy}{n}\\
&= \sum_{n=1}^\infty \frac{1}{n^s}\int_0^\infty y^{s-1}e^{-y}\,dy\\
&= \zeta(s)\Gamma(s)
\end{aligned}$$

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

15b. Special values

At a more advanced level, we can try to compute particular values of ζ. Things are quite tricky here, and we have the following result, briefly discussed before:

THEOREM 15.7. *We have the following formula, for the even integers $s=2k$,*

$$\zeta(2k)=(-1)^{k+1}\frac{(2\pi)^{2k}B_{2k}}{2\cdot(2k)!}$$

with B_n being the Bernoulli numbers, which in practice gives the formulae

$$\zeta(2)=\frac{\pi^2}{6}\quad,\quad \zeta(4)=\frac{\pi^4}{90}\quad,\quad \zeta(6)=\frac{\pi^6}{945}\quad,\quad \zeta(8)=\frac{\pi^8}{9450}\quad,\quad \ldots$$

generalizing the formula $\zeta(2)=\pi^2/6$ of Euler, solving the Basel problem.

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

(1) To start with, at $s=2$ the Euler computation, from before, was as follows:

$$\begin{aligned}\frac{\sin x}{x} &= 1-\frac{x^2}{3!}+\frac{x^4}{5!}-\frac{x^6}{7!}+\ldots\\ &= \left(1-\frac{x}{\pi}\right)\left(1+\frac{x}{\pi}\right)\left(1-\frac{x}{2\pi}\right)\left(1+\frac{x}{2\pi}\right)\ldots\\ &= \left(1-\frac{x^2}{\pi^2}\right)\left(1-\frac{x^2}{4\pi^2}\right)\left(1-\frac{x^2}{9\pi^2}\right)\ldots\\ &= 1-\frac{1}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\,x^2+\ldots\end{aligned}$$

It is possible to use the same idea for dealing with $\zeta(2k)$ with $k\in\mathbb{N}$, but this is quite complicated, and in addition the above method of Euler needs some justification, that we have not really provided before, so in short, not a path to be followed.

(2) Instead, we have the following luminous computation, based on Theorem 15.6:

$$\begin{aligned}\zeta(2k) &= \frac{1}{\Gamma(2k)}\int_0^{\infty}\frac{x^{2k-1}}{e^x-1}\,dx\\ &= \frac{1}{(2k-1)!}\int_0^{\infty}\frac{x^{2k-1}}{e^x-1}\,dx\\ &= \frac{1}{(2k-1)!}\int_0^{\infty}\frac{(2\pi t)^{2k-1}}{e^{2\pi t}-1}\,2\pi dt\\ &= \frac{(2\pi)^{2k}}{(2k-1)!}\int_0^{\infty}\frac{t^{2k-1}}{e^{2\pi t}-1}\,dt\end{aligned}$$

(3) But, we recognize on the right the integral giving rise to the even Bernoulli numbers, with one of the many definitions of these numbers being as follows:

$$B_{2k}=4k(-1)^{k+1}\int_0^{\infty}\frac{t^{2k-1}}{e^{2\pi t}-1}\,dt$$

Thus, we can finish our computation of the values $\zeta(2k)$ as follows:

$$\begin{aligned}\zeta(2k) &= \frac{(2\pi)^{2k}}{(2k-1)!}\cdot(-1)^{k+1}\frac{B_{2k}}{4k}\\ &= (-1)^{k+1}\frac{(2\pi)^{2k}B_{2k}}{2\cdot(2k)!}\end{aligned}$$

(4) Regarding now the Bernoulli numbers, there is a long story here. At the beginning, we have the following formula of Bernoulli, standing as a definition for them:

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

This leads to the following recurrence relation, which computes them:

$$B_m = -\frac{1}{m+1}\sum_{k=0}^{m-1}\binom{m+1}{k}B_k$$

In practice, we can see that the odd Bernoulli numbers all vanish, except for the first one, $B_1 = -1/2$, and that the even Bernoulli numbers are as follows:

$$\frac{1}{6}\quad,\quad -\frac{1}{30}\quad,\quad \frac{1}{42}\quad,\quad -\frac{1}{30}\quad,\quad \frac{5}{66}\quad,\quad -\frac{691}{2730}\quad,\quad \frac{7}{6}\quad,\quad \ldots$$

(5) For analytic purposes, the Bernoulli numbers are best viewed as follows, with this coming from the fact that the coefficients satisfy the above recurrence relation:

$$\begin{aligned}\frac{x}{e^x-1} &= \sum_{n=0}^{\infty}B_n\frac{x^n}{n!}\\ &= 1-\frac{1}{2}x+\frac{1}{6}\cdot\frac{x^2}{2!}-\frac{1}{30}\cdot\frac{x^4}{4!}+\frac{1}{42}\cdot\frac{x^6}{6!}-\frac{1}{30}\cdot\frac{x^8}{8!}+\ldots\end{aligned}$$

Observe that all this is related as well to the hyperbolic functions, via:

$$\frac{x}{2}\left(\coth\frac{x}{2}-1\right)=\frac{x}{e^x-1}=\sum_{n=0}^{\infty}B_n\frac{x^n}{n!}$$

The point now is that, in relation with our zeta business, the above analytic formulae give, after some calculus, the formula that we used in (3), namely:

$$B_{2k}=4k(-1)^{k+1}\int_0^{\infty}\frac{t^{2k-1}}{e^{2\pi t}-1}\,dt$$

(6) Finally, no discussion about the Bernoulli numbers would be complete without mentioning the Euler-Maclaurin formula, involving them, which is as follows:

$$\begin{aligned}\sum_{k=0}^{n-1}f(x) &\simeq \int_0^n f(x)dx-\frac{1}{2}(f(n)-f(0))\\ &\quad+\frac{1}{6}\cdot\frac{f'(n)-f'(0)}{2!}-\frac{1}{30}\cdot\frac{f^{(3)}(n)-f^{(3)}(0)}{4!}\\ &\quad+\frac{1}{42}\cdot\frac{f^{(5)}(n)-f^{(5)}(0)}{6!}-\frac{1}{30}\cdot\frac{f^{(7)}(n)-f^{(7)}(0)}{8!}+\ldots\end{aligned}$$

(7) And there is more coming from the complex extension of the zeta function, by analytic continuation, that we will discuss later. An an announcement here, the values of zeta at the negative integers $0,-1,-2,-3,\ldots$ will not be ∞, but rather given by:

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

Alternatively, we have the following formula for the Bernoulli numbers:

$$B_n = (-1)^{n-1} n\zeta(1-n)$$

(8) In any case, we are led to the various conclusions in the statement, both theoretical and numeric. And exercise for you of course to learn more about the Bernoulli numbers, and beware of the freakish notations used by mathematicians there. □

As a more digest form of Theorem 15.7, let us record as well:

THEOREM 15.8. *The generating function of the numbers $\zeta(2k)$ with $k \in \mathbb{N}$ is*

$$\sum_{k=0}^{\infty} \zeta(2k) x^{2k} = -\frac{\pi x}{2} \cot(\pi x)$$

and with this generalizing the formula involving Bernoulli numbers.

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

(1) A version of the recurrence formula for Bernoulli numbers is as follows:

$$B_{2n} = -\frac{1}{n+1/2} \sum_{k=1}^{n-1} \binom{2n}{2k} B_{2k} B_{2n-2k}$$

Now observe that this formula can be written in the following way:

$$\frac{B_{2n}}{(2n)!} = -\frac{1}{n+1/2} \sum_{k=1}^{n-1} \frac{B_{2k}}{(2k)!} \cdot \frac{B_{2n-2k}}{(2n-2k)!}$$

In view of Theorem 15.7, we obtain the following formula, valid at any $n > 1$:

$$\zeta(2n) = \frac{1}{n+1/2} \sum_{k=1}^{n-1} \zeta(2k)\zeta(2n-2k)$$

(2) But this allows the computation of the series in the statement, by squaring that series. Indeed, consider the following modified version of that series:

$$f(x) = 2\sum_{k=0}^{\infty} \zeta(2k) \left(\frac{x}{\pi}\right)^{2k}$$

By squaring, and using the recurrence formula for the numbers $\zeta(2n)$ found in (1), with some care at the values $n = 0, 1$, not covered by that formula, we obtain:

$$f^2 + f + x^2 = xf'$$

(3) But this is precisely the functional equation satisfied by $g(x) = -x\cot x$. Indeed, by using the well-known formula $\cot' = -\cot^2 - 1$, we have:

$$\begin{aligned} xg' &= x(-\cot x - x\cot' x) \\ &= x(-\cot x + x\cot^2 x + x) \\ &= g + g^2 + x^2 \end{aligned}$$

(4) We conclude that we have $f = g$, which reads:

$$2\sum_{k=0}^{\infty}\zeta(2k)\left(\frac{x}{\pi}\right)^{2k} = -x\cot x$$

Now by replacing $x \to \pi x$, we obtain the formula in the statement. □

Regarding now the values $\zeta(2k+1)$ with $k \in \mathbb{N}$, the story here is more complicated, with the first such number being the Apéry constant, given by:

$$\zeta(3) = \sum_{n=1}^{\infty}\frac{1}{n^3}$$

There has been a lot of work on this number, by Apéry and others, and on the higher $\xi(2k+1)$ values as well. Let us record here the following result, a bit of physics flavor:

THEOREM 15.9. *We have the following formula,*

$$\zeta(s) = \int_0^1\ldots\int_0^1\frac{dx_1\ldots dx_s}{1 - x_1\ldots x_s}$$

valid for any $s \in \mathbb{N}$, $s \geq 2$.

PROOF. This follows as usual from some calculus, the idea being as follows:

(1) At $s = 2$ we have indeed the following computation, using Theorem 15.6:

$$\begin{aligned} \int_0^1\int_0^1\frac{1}{1-xy}\,dxdy &= \int_0^1\left[-\frac{\log(1-xy)}{y}\right]_0^1 dy \\ &= -\int_0^1\frac{\log(1-y)}{y}\,dy \\ &= -\int_0^{\infty}\frac{\log(e^{-t})}{1-e^{-t}}\,e^{-t}dt \\ &= \int_0^{\infty}\frac{t}{e^t-1}\,dt \\ &= \zeta(2)\Gamma(2) \\ &= \zeta(2) \end{aligned}$$

In general the proof is similar, and we will leave this as an instructive exercise.

(2) Before leaving, however, let us see as well, out of mathematical curiosity, what happens at the exponent $s = 1$. Here the integral in the statement is:

$$\begin{aligned}\int_0^1 \frac{1}{1-x}\,dx &= [-\log(1-x)]_0^1\\ &= -\log(1-1)+\log(1-0)\\ &= \infty+0\\ &= \zeta(1)\end{aligned}$$

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, in a moment. □

Many other things can be said about ζ and its special values, as a continuation of the above, and check here any advanced number theory book. In what concerns us, we will rather head towards the analytic left half-plane $Re(s) \leq 1$, using complex analysis.

15c. Complex zeta

Quite remarkably, with a bit of complex analysis, we can have the zeta function working in the whole complex plane, via analytic continuation. However, analytic continuation being Devil's business, we will explain this slowly, by gradually going from the analytic right half-plane $Re(s) > 1$, that we understand well, to other parts of $\mathbb{C}$.

Getting started 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:

PROPOSITION 15.10. *We have the following formula,*

$$\lim_{s\to 1}(s-1)\zeta(s) = 1$$

showing that the complex zeta has a simple pole at $s = 1$, with residue 1.

PROOF. We have the following computation, using $\Gamma(1) = 1$:

$$\begin{aligned}\lim_{s\to 1}(s-1)\zeta(s) &= \lim_{s\to 1}(s-1)\int_0^\infty \frac{x^{s-1}}{e^x-1}\,dx\\ &= \lim_{t\to 0}\int_0^\infty \frac{tx^t}{e^x-1}\,dx\\ &= 1\end{aligned}$$

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,*

$$\lim_{s\to 1}\left|\zeta(s)-\frac{1}{s-1}\right|=\gamma$$

with γ being the Euler-Mascheroni constant.

PROOF. This is something more advanced, the idea being as follows:

(1) The Euler-Mascheroni constant is related to the zeta function by:

$$\gamma=\sum_{n=2}^{\infty}(-1)^n\frac{\zeta(n)}{n}$$

(2) On the other hand, we have we well the following formula:

$$\gamma=\lim_{s\to 1^+}\sum_{n=1}^{\infty}\frac{1}{n^s}-\frac{1}{s^n}$$

But in terms of the zeta function, this latter formula simply reads:

$$\gamma=\lim_{s\to 1^+}\zeta(s)-\frac{1}{s-1}$$

(3) Thus, we are led to the formula in the statement. Note that we have as well:

$$\gamma=\lim_{s\to 0}\frac{\zeta(1+s)+\zeta(1-s)}{2}$$

Indeed, this follows from the formula in the statement. □

Leaving aside now $s=1$, let us focus on the other points, $s=1+it$ with $t\neq 0$, of the boundary line $Re(s)=1$, between known and unknown. We have here:

THEOREM 15.12. *The Riemann zeta function, namely*

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

converges at any $s=1+it$ with $t\neq 0$.

PROOF. We have the following computation, to start with:

$$\begin{aligned}\zeta(1+it) &= \sum_{n=1}^{\infty}\frac{1}{n^{1+it}}\\ &= \sum_{n=1}^{\infty}\frac{1}{ne^{it\log n}}\\ &= \sum_{n=1}^{\infty}\frac{e^{-it\log n}}{n}\\ &= \sum_{n=1}^{\infty}\frac{\cos(t\log n)-i\sin(t\log n)}{n}\end{aligned}$$

And then, the convergence at $t\neq 0$ can be proved via some calculus. □

Let us get now into the true unknown, $Re(s)<1$, with our first objective being that of understanding what happens in the strip $0<Re(s)<1$. We first have here:

PROPOSITION 15.13. *Unlike the standard Riemann series, which diverges,*

$$\zeta(1)=1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\ldots=\infty$$

the signed version of this series, called standard Dirichlet series, converges,

$$\eta(1)=1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\frac{1}{5}-\frac{1}{6}+\ldots<\infty$$

and we can even compute its value, $\eta(1)=\log 2$.

PROOF. Here the convergence of the series $\eta(1)$ can be proved in a variety of ways, for instance by grouping terms and comparing to $\zeta(2)<\infty$:

$$\eta(1)=\frac{1}{2}+\frac{1}{12}+\frac{1}{30}+\frac{1}{56}+\ldots<\zeta(2)<\infty$$

As for the exact formula of $\eta(1)$, this follows from the Taylor formula for log:

$$\log(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\frac{x^5}{5}-\frac{x^6}{6}+\ldots$$

Indeed, by plugging in $x=1$, we obtain the formula in the statement. □

Thus, we have our idea, "forcing" zeta to converge in the strip $0<Re(s)<1$, by adding signs, and then recovering zeta, or rather its analytic continuation, in this same strip, by removing the signs. This leads to the following remarkable result:

THEOREM 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 $0<Re(s)<1$.

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 $0<Re(s)<1$.

(2) We must now connect ζ and η, at $Re(s)>1$, and this can be done as follows:

$$\begin{aligned}\zeta(s)+\eta(s)&=\sum_{n=1}^{\infty}\frac{1}{n^s}+\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^s}\\&=2\sum_{k=1}^{\infty}\frac{1}{(2k)^s}\\&=2^{1-s}\sum_{k=1}^{\infty}\frac{1}{k^s}\\&=2^{1-s}\zeta(s)\end{aligned}$$

(3) But this gives the following formula, valid at any exponent $s\in\mathbb{C}$ satisfying $Re(s)>1$, and which is the formula in the statement:

$$\zeta(s)=\frac{1}{1-2^{1-s}}\,\eta(s)$$

(4) In order now to conclude, we can invoke the theory of analytic continuation. Skipping some theoretical details here, and we refer for instance to Rudin [80] for all this, what we have in the statement is a formula for ζ in the whole right half-plane, $Re(s)>0$, which is analytic, and more specifically meromorphic, with a single pole, at $s=1$, and which coincides with the usual formula of ζ on the usual domain of definition, $Re(s)>1$. But, in this situation, the theory of analytic continuation tells us that we can redefine ζ all over the right half-plane, $Re(s)>0$, by the formula in the statement, and with this extension being unique, as per the general properties of the meromorphic functions.

(5) Finally, observe that our present result proves Theorem 15.12 as well. Thinking retrospectively, we were in need there precisely of a Dirichlet type idea. □

15d. Riemann formula

Getting now to the left half-plane, $Re(s) < 0$, many methods are available here, and with the main one, due to Riemann himself, which is something quite tough, but unavoidable for understanding the zeta function as a whole, being as follows:

THEOREM 15.15. *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 $0 < Re(s) < 1$, and can serve as definition for zeta in the left half-plane, $Re(s) < 0$, 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 $Re(s) > 1$ 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 $Re(s) < 0$, so that $Re(1-s) > 1$. 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 $Re(s) < 0$. 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 $Re(s) < 0$.

(4) Now let us have a look at the strip $0 < Re(s) < 1$. 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 $0 < Re(s) < 1$.

(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 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:

$$\Psi(x)=\sum_{n=1}^{\infty}e^{-n^2\pi x}dx$$

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 $s \to 1-s$, 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 this is equivalent to the Riemann symmetry formula in the statement.

(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. □

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 15.16. *We can talk about the Riemann zeta, as a meromorphic function $\zeta : \mathbb{C} \to \mathbb{C}$, with a single pole, at $s = 1$ with residue 1. At $Re(s) > 1$ we have*

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

and more generally at $Re(s) > 0$ we have the following formula:

$$\zeta(s) = \frac{1}{1-2^{1-s}}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^s}$$

Also, the values of zeta at any s and $1-s$ are related by the Riemann formula

$$\zeta(s) = 2^s\pi^{s-1}\sin\left(\frac{\pi s}{2}\right)\Gamma(1-s)\zeta(1-s)$$

with Γ being as usual the gamma function.

PROOF. This is a summary of our various findings from Theorems 15.14 and 15.15 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. □

Getting back now to the Riemann formula from Theorem 15.15, 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 $0 < Re(s) < 1$ is what matters, and that the vertical axis $Re(s) = 1/2$ is where interesting things should happen.

As a consequence of the Riemann formula, we have the following version of it:

THEOREM 15.17. *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 15.15, and is in fact what comes out from computations, when proving Theorem 15.15. □

In practice, the quantity in Theorem 15.17 is best normalized as follows:

THEOREM 15.18. *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 15.15. □

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:

$$\gamma_n=\lim_{m\to\infty}\left[\left(\sum_{k=1}^{m}\frac{(\log k)^n}{k}\right)-\frac{(\log m)^{n+1}}{n+1}\right]$$

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:

$$\sum_{k=0}^{\infty}\zeta(2k)x^{2k}=-\frac{\pi x}{2}\cot(\pi x)$$

In practice, we get 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 formula:

$$\zeta(-2k+1)=-\frac{B_{2k}}{2k}$$

In fact, by Riemann reflection, we have the following formula, for any $n\in\mathbb{N}$:

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

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:

$$\sum_{k=2}^{\infty}(\zeta(k)-1)=1\quad,\quad\sum_{k=1}^{\infty}(\zeta(2k)-1)=\frac{3}{4}$$

$$\sum_{k=1}^{\infty}(\zeta(2k+1)-1)=\frac{1}{4}\quad,\quad\sum_{k=2}^{\infty}(-1)^k(\zeta(k)-1)=\frac{1}{2}$$

Along the same lines, a second series of formulae is as follows:

$$\sum_{k=1}^{\infty}(-1)^k\frac{\zeta(k)}{k}=0\qquad,\qquad\sum_{k=1}^{\infty}\frac{\zeta(k)-1}{k}=0$$

$$\sum_{k=2}^{\infty}(-1)^k\frac{\zeta(k)}{k}=\gamma\qquad,\qquad\sum_{k=2}^{\infty}\frac{\zeta(k)-1}{k}=1-\gamma$$

And there are many more formulae computing or relating the values of zeta at positive integers, more specialized, quite often Ramanujan-looking.

Getting now to zeroes, as a consequence of Theorem 15.15, we have:

THEOREM 15.19. *We have the following formula, for any integer $k\geq 1$,*

$$\zeta(-2k)=0$$

with these being called the "trivial zeroes" of ζ.

PROOF. We recall that the Riemann symmetry formula from Theorem 15.15 is as follows, valid all over the complex plane, as an equality of meromorphic functions:

$$\zeta(s)=2^s\pi^{s-1}\sin\left(\frac{\pi s}{2}\right)\Gamma(1-s)\zeta(1-s)$$

By plugging in the value $s=-2k$, with $k\geq 1$ integer, we obtain:

$$\begin{aligned}\zeta(-2k)&=2^{-2k}\pi^{-2k-1}\sin(k\pi)\Gamma(1+2k)\xi(1+2k)\\&=0\end{aligned}$$

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

15e. Exercises

Exercises:

EXERCISE 15.20.

EXERCISE 15.21.

EXERCISE 15.22.

EXERCISE 15.23.

EXERCISE 15.24.

EXERCISE 15.25.

EXERCISE 15.26.

EXERCISE 15.27.

Bonus exercise.

CHAPTER 16

Riemann hypothesis

16a. Back to primes

Let us go back to the main result from chapter 13, namely the Chebycheff estimate there, which was as follows, with the function $\pi(x)$ counting the primes $p \leq x$:

$$\pi(x) \approx \frac{x}{\log x}$$

As mentioned in chapter 13, Hadamard and de la Vallée Poussin were able, using the Riemann zeta function, to prove the Prime Number Theorem, which states that:

$$\pi(x) \sim \frac{x}{\log x}$$

We will explain here this result, which is highly-non trivial, even by modern standards, following the original proof of Hadamard and de la Vallée Poussin.

We will comment as well on some other known proofs of the Prime Number Theorem, which are more modern, notably with the Selberg proof, not using zeta, and also with the Newman proof, not using zeta either, and being a bit shorter than Selberg's. And finally, we will discuss some further improvements of the above estimates.

So, this will be the plan for this chapter, and with a Theorem coming with 3 different proofs, which is highly unusual, you might think that we have something against the first proof, or against the zeta function in general. Quite the opposite, we love zeta. But the other proofs are instructive as well, revealing some things about prime numbers not necessarily captured by the mighty zeta, and we will present them too.

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 16.1. *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:

PROPOSITION 16.2. *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. □

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 16.3. *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. □

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 16.4. *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 16.1 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 16.3, 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 Proposition 16.2, 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:

$$\begin{aligned}\frac{\varphi(x)}{x^2} &= \frac{1}{2}\left(1-\frac{1}{x}\right)^2-\frac{1}{2\pi i}\int_{r-\infty i}^{r+\infty i}x^{s-1}f(x)\,ds\\ &= \frac{1}{2}\left(1-\frac{1}{x}\right)^2-\frac{1}{2\pi}\int_{-\infty}^{\infty}x^{r+it-1}f(r+it)\,dt\\ &= \frac{1}{2}\left(1-\frac{1}{x}\right)^2-\frac{x^{r-1}}{2\pi}\int_{-\infty}^{\infty}e^{it\log x}f(r+it)\,dt\end{aligned}$$

(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. □

Now remember from Proposition 16.2 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 16.4, 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 16.5. *The contribution to the modified Chebycheff function ψ of the trivial zeroes of zeta, namely $-2, -4, -6, \ldots$, 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 $x \to \infty$ 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. □

16b. Prime distribution

Regarding now the non-trivial zeroes of zeta, we know from chapter 15 that these lie inside the strip $0 \leq Re(s) \leq 1$, and as a first observation, we have:

PROPOSITION 16.6. *The contribution to the modified Chebycheff function ψ of the non-trivial zeroes of zeta lying in the strip $0 \leq Re(s) < 1$ satisfies*

$$\sum_{\zeta(s)=0} \frac{x^s}{s} = o(x)$$

so we are left with studying the zeroes on the line $Re(s) = 1$.

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 16.5. □

We are now getting to the core of the proof, with the key ingredient being:

THEOREM 16.7. *The Riemann zeta function has no zero on the line*

$$Re(s) = 1$$

and no zero on the line $Re(s) = 0$ *either.*

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

(1) To start with, the $Re(s) = 0$ result, that we will not need here for our current purposes, in view of Proposition 16.6, but which of course has great theoretical interest, follows from the $Re(s) = 1$ result, via the Riemann reflection formula from chapter 15.

(2) In order to study now the zeta function on the line $Re(s) = 1$, we use the Euler product formula for this function, namely:

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}$$

By taking the logarithm, we obtain from this the following formula:

$$\begin{aligned}\log\zeta(s) &= -\sum_p \log\left(1 - \frac{1}{p^s}\right)\\ &= \sum_p \sum_{k=0}^{\infty} \frac{1}{kp^{ks}}\end{aligned}$$

(3) Now with $s = r + it$ as usual, this formula reads:

$$\begin{aligned}\log\zeta(s) &= \sum_p \sum_{k=0}^{\infty} \frac{1}{kp^{k(r+it)}}\\ &= \sum_p \sum_{k=0}^{\infty} \frac{p^{-kit}}{kp^{kr}}\\ &= \sum_p \sum_{k=0}^{\infty} \frac{e^{-kit\log p}}{kp^{kr}}\\ &= \sum_p \sum_{k=0}^{\infty} \frac{\cos(kt\log p) - i\sin(kt\log p)}{kp^{kr}}\end{aligned}$$

(4) Now remember the following formula, for the complex exponentials:

$$|e^z|^2 = e^z \cdot \overline{e^z} = e^z e^{\bar{z}} = e^{z+\bar{z}} = e^{2Re(z)}$$

Thus we have $|e^z| = e^{Re(z)}$, and by using this with $z = \log \zeta(s)$, we get:

$$\begin{aligned} |\zeta(s)| &= |\exp(\log \zeta(s))| \\ &= \exp(Re(\log \zeta(s))) \\ &= \exp\left(\sum_p \sum_{k=0}^{\infty} \frac{\cos(kt \log p)}{kp^{kr}}\right) \end{aligned}$$

(5) 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 (4) the following estimate:

$$\begin{aligned} |\zeta(r)^3\zeta(r+it)^4\zeta(r+2it)| &= \exp\left(\sum_p \sum_{k=0}^{\infty} \frac{3 + 4\cos(kt\log p) + \cos(2kt\log p)}{kp^{kr}}\right) \\ &= \exp\left(\sum_p \sum_{k=0}^{\infty} \frac{2(1+\cos(kt\log p))^2}{kp^{kr}}\right) \\ &\geq 1 \end{aligned}$$

(6) 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 $r \to 1^+$ limit:

$$K = \zeta(r)^3\zeta(r+it)^4\zeta(r+2it)$$

What happens then in the $r \to 1^+$ limit is that we have $\zeta(r)^3 \to \infty$ with triple pole behavior, $\zeta(r+it)^4 \to 0$ with quadruple zero behavior, and $\zeta(r+2it) \to \zeta(2it)$ with analytic behavior. But since $3 < 4$ the quadruple zero will kill the triple pole, and so:

$$\lim_{r \to 1^+} K = 0$$

But this contradicts the estimate found in (5), and so our theorem is proved. □

By putting now everything together, we obtain:

THEOREM 16.8 (Prime Number Theorem). *We have*

$$\pi(x) \sim \frac{\log x}{x}$$

in the $x \to \infty$ limit.

PROOF. This follows by putting everything together, as follows:

(1) We know from Proposition 16.2 that $\pi(x) \sim x/\log x$ is equivalent to $\psi(x) \sim x$.

(2) We have in Theorem 16.4 a formula for $\psi(x)$, in terms of the zeroes of zeta.

(3) Most of these zeroes are taken care of by Proposition 16.5 and Proposition 16.6.

(4) As for the remaining zeroes, there are none, as shown by Theorem 16.7. □

16c. Riemann hypothesis

Riemann hypothesis.

16d. Further results

Further results.

16e. Exercises

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

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[19] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).

[20] A. Cottrell, An introduction to metallurgy, CRC Press (1997).

[21] C. Darwin, On the origin of species (1859).

[22] P.A. Davidson, Introduction to magnetohydrodynamics, Cambridge Univ. Press (2001).

[23] P.A.M. Dirac, Principles of quantum mechanics, Oxford Univ. Press (1930).

[24] S. Dodelson, Modern cosmology, Academic Press (2003).

[25] S.T. Dougherty, Combinatorics and finite geometry, Springer (2020).

[26] M. Dresher, The mathematics of games of strategy, Dover (1981).

[27] R. Durrett, Probability: theory and examples, Cambridge Univ. Press (1990).

[28] F. Dyson, Origins of life, Cambridge Univ. Press (1984).

[29] A. Einstein, Relativity: the special and the general theory, Dover (1916).

[30] L.C. Evans, Partial differential equations, AMS (1998).

[31] W. Feller, An introduction to probability theory and its applications, Wiley (1950).

[32] E. Fermi, Thermodynamics, Dover (1937).

[33] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics I: mainly mechanics, radiation and heat, Caltech (1963).

[34] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics II: mainly electromagnetism and matter, Caltech (1964).

[35] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics III: quantum mechanics, Caltech (1966).

[36] R.P. Feynman and A.R. Hibbs, Quantum mechanics and path integrals, Dover (1965).

[37] P. Flajolet and R. Sedgewick, Analytic combinatorics, Cambridge Univ. Press (2009).

[38] A.P. French, Special relativity, Taylor and Francis (1968).

[39] J.H. Gillespie, Population genetics, Johns Hopkins Univ. Press (1998).

[40] C. Godsil and G. Royle, Algebraic graph theory, Springer (2001).

[41] H. Goldstein, C. Safko and J. Poole, Classical mechanics, Addison-Wesley (1980).

[42] D.L. Goodstein, States of matter, Dover (1975).

[43] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).

[44] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).

[45] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).

[46] D.J. Griffiths, Revolutions in twentieth-century physics, Cambridge Univ. Press (2012).

[47] V.P. Gupta, Principles and applications of quantum chemistry, Elsevier (2016).

[48] W.A. Harrison, Solid state theory, Dover (1970).

[49] W.A. Harrison, Electronic structure and the properties of solids, Dover (1980).

[50] R.A. Horn and C.R. Johnson, Matrix analysis, Cambridge Univ. Press (1985).

[51] C.E. Housecroft and A.G. Sharpe, Inorganic chemistry, Pearson (2018).

[52] K. Huang, Introduction to statistical physics, CRC Press (2001).

[53] K. Huang, Fundamental forces of nature, World Scientific (2007).

[54] S. Huskey, The skeleton revealed, Johns Hopkins Univ. Press (2017).

[55] L. Hyman, Comparative vertebrate anatomy, Univ. of Chicago Press (1942).

[56] L.P. Kadanoff, Statistical physics: statics, dynamics and renormalization, World Scientific (2000).

[57] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).

[58] C. Kittel, Introduction to solid state physics, Wiley (1953).

[59] D.E. Knuth, The art of computer programming, Addison-Wesley (1968).

[60] M. Kumar, Quantum: Einstein, Bohr, and the great debate about the nature of reality, Norton (2009).

[61] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).

[62] L.D. Landau and E.M. Lifshitz, Mechanics, Pergamon Press (1960).

[63] L.D. Landau and E.M. Lifshitz, The classical theory of fields, Addison-Wesley (1951).

[64] L.D. Landau and E.M. Lifshitz, Quantum mechanics: non-relativistic theory, Pergamon Press (1959).

[65] S. Lang, Algebra, Addison-Wesley (1993).

[66] P. Lax, Linear algebra and its applications, Wiley (2007).

[67] P. Lax, Functional analysis, Wiley (2002).

[68] P. Lax and M.S. Terrell, Calculus with applications, Springer (2013).

[69] P. Lax and M.S. Terrell, Multivariable calculus with applications, Springer (2018).

[70] S. Ling and C. Xing, Coding theory: a first course, Cambridge Univ. Press (2004).

[71] J.P. Lowe and K. Peterson, Quantum chemistry, Elsevier (2005).

[72] S.J. Marshall, The story of the computer: a technical and business history, Create Space Publ. (2022).

[73] M.L. Mehta, Random matrices, Elsevier (2004).

[74] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge Univ. Press (2000).

[75] R.K. Pathria and and P.D. Beale, Statistical mechanics, Elsevier (1972).

[76] T.D. Pollard, W.C. Earnshaw, J. Lippincott-Schwartz and G. Johnson, Cell biology, Elsevier (2022).

[77] J. Preskill, Quantum information and computation, Caltech (1998).

[78] R. Rojas and U. Hashagen, The first computers: history and architectures, MIT Press (2000).

[79] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[80] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[81] W. Rudin, Functional analysis, McGraw-Hill (1973).

[82] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[83] B. Ryden and B.M. Peterson, Foundations of astrophysics, Cambridge Univ. Press (2010).

[84] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).

[85] R. Shankar, Fundamentals of physics I: mechanics, relativity, and thermodynamics, Yale Univ. Press (2014).

[86] R. Shankar, Fundamentals of physics II: electromagnetism, optics, and quantum mechanics, Yale Univ. Press (2016).

[87] N.J.A. Sloane and S. Plouffe, Encyclopedia of integer sequences, Academic Press (1995).

[88] A.M. Steane, Thermodynamics, Oxford Univ. Press (2016).

[89] S. Sternberg, Dynamical systems, Dover (2010).

[90] D.R. Stinson, Combinatorial designs: constructions and analysis, Springer (2006).

[91] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[92] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[93] J. von Neumann and O. Morgenstern, Theory of games and economic behavior, Princeton Univ. Press (1944).

[94] J. Watrous, The theory of quantum information, Cambridge Univ. Press (2018).

[95] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[96] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[97] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).

[98] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).

[99] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).

[100] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).

Index