Basic number theory

Teo Banica

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

2010 *Mathematics Subject Classification.* 11A41

Key words and phrases. Number theory, Prime number

Abstract. This is an introduction to number theory, for mathematicians, physicists and other scientists, assuming calculus and some basic algebra known. We first discuss the elementary aspects of the theory, mostly by focusing on algebraic equations, and prime numbers. Then we discuss modern algebraic and analytic methods, again by focusing on algebraic equations, and prime numbers. Finally, we discuss various aspects of the Riemann zeta function, mixing algebra, analysis, physics and more.

Preface

You certainly know well calculus, and perhaps a little bit of algebra, geometry and probability too, which can only help with the calculus business. And good news, these are the foundations of mathematics, that you will need, time and again, no matter what kind of mathematics, physics, or other science or engineering you will be doing.

My question, however, is very simple: how far have you gone into your practice of calculus, have you met for instance, on a regular basis, all sorts of numbers which factorize, and with that factorizations being very useful for your computations?

Let me explain. First of all, calculus is serious business, and can be done at several levels. You were certainly awarded a white belt at the end of your 1st year, and you will surely win the black one too, at the end of your graduate studies. But then, let us not forget this, for true professionals, there are also Dans, ranging from 1 to 7. So, there is a long story with learning calculus, and all this learning is of course worth it, because the more calculus you know, the sharper and more original your science will be.

Getting now to the point, numbers, these appear all the time in calculus, when performed at the professional level. Trust me, and here is the situation:

(1) You surely know a bit about this, because you have binomials and factorials in the Taylor formula. Also, more concretely, in relation with the basic functions that you use all the time, namely $\sin, \cos, \exp, \log$. And not to forget here the modest power function x^n, which actually produces all these binomials and factorials, via $(x+y)^n$.

(2) At a more advanced level, typically you will want to compute a function f. And after days of work a number like 731 will show up, at key places, in your listings. But then you will say wait, $731 = 43 \times 17$ is an old friend of mine, and this suggests a decomposition of type $f = gh$. And good news, $f = gh$ indeed, with both g, h computable.

So, hope you got my point, knowing numbers, how they factorize, is something extremely useful. Along the same lines, knowing too what's power 2^k and what's not, what's factorial $n!$ and what's not, what's binomial $\binom{n}{k}$ and what's not, and so on, all sorts of other number things, all this is knowledge is golden, when doing advanced calculus.

Which gets us, as a continuation of my first question, raised above, to a second question, which is even more intriguing, I hope: how many numbers do you know?

Here by "knowing" I mean really knowing, like old friends. For instance you certainly know well $2, 3, 4, \ldots$, all these magical small numbers. However, in what regards 42 for instance, you would say yes $42 = 6 \times 7$, but the point is that this is a Catalan number too, $C_5 = 42$, and with this being something very useful, that must be known too.

Again, hope you got my point, and forgetting now about the previous belts and Dans, there is something even simpler for measuring your "calculus level", and that is the amount of numbers that you know. And here, the situation is more or less as follows:

(1) Knowing $1 - 10$: undergraduate level.

(2) Knowing $1 - 100$: graduate level.

(3) Knowing $1 - 1000$: good analyst.

(4) Knowing $1 - 10000$: good algebraist.

(5) Knowing $1 - 1000000$: Ramanujan and others.

In the hope that all this will motivate you for learning some number theory, perhaps a bit in advance with respect to what you will really need, in your scientific life, and the present book will be here for that. We will be discussing here all sorts of things, all useful and beautiful mathematics, that are good to know, in relation with the numbers.

But you might perhaps be genuinely interested in number theory. In this case no need of course for presentations, you know as I do that number theory is the Queen of Mathematics, and our sweet love. And in the hope that you will not be deceived by this book, which is for you too, quickly explaining the basics, and then dealing with both algebraic and analytic number theory, and ending with some weird zeta stuff.

It is a pleasure to thank everyone, collaborators, for our joint work on all sorts of probability and special function computations. As a grad student I originally wanted to do arithmetic geometry, but ended up in quantum physics, which is fun guarantee, too. Finally, many thanks go to my cats, are they algebraic or analytic about numbers, I will never know, but what is sure is that their computations are extremely fast.

Contents

Part I

Number theory

Ramona, come closer
Shut softly your watery eyes
The pangs of your sadness
Will pass as your senses will rise

CHAPTER 1

Prime numbers

1a. Number fields

What are numbers? You surely know that all sorts of numbers, be them rational, real, complex and so on, can be constructed starting from $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$, and more on this in a moment, but where does $\mathbb{N}$ itself come from? Not an easy question:

(1) The ancients did not bother much with this, simply assuming that $\mathbb{N}$ was invented by God, and then enjoying this, and starting doing their math. Which looks like a wise and sane attitude, but with the remark however that, also on religious grounds, but more extreme, the number 0 was at certain times in history considered blasphemous, its usage forbidden, and with this slowing down the development of arithmetic.

(2) According to modern mathematics, at the beginning was the empty set $\emptyset$. Indeed, the cardinality of this set is zero, $|\emptyset| = 0$. On the other hand, this empty set does exist, and in mathematical parlance, we can write $|\{\emptyset\}| = 1$. But then, if we look at the subsets of $\{\emptyset\}$, there are two of them, $|\{\emptyset, \{\emptyset\}\}| = 2$. And then, we can go ahead and define the number 3 as well, as being given by $|\{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}| = 3$, and so on.

(3) According to modern physics, at the beginning was absolutely nothing, but then came the Big Bang, bringing life as we know it, including $\mathbb{N}$. With the point here being that the original "nothing" really meant nothing, not to be confused for instance with the vacuum that we know well, which allows the propagation of gravity, electrostatic forces, electromagnetic waves and so on, and is probably smarter than both me and you.

Quite philosophical all this, and as a comment, both the modern mathematics in (2) and modern physics in (3) require, for proper functioning, a mysterious operation of type "let me see what's inside me", which can only be of divine nature, so here we are back at (1). But are we here for talking philosophy, or doing number theory. Number theorists use a version of (1) for their understanding of $\mathbb{N}$, and of the world, as follows:

FACT 1.1. *God created the prime numbers* $2, 3, 5, 7, \ldots$*, and these prime numbers started to multiply, maybe they are still multiplying, and produced* $\mathbb{N}$.

Observe that there is a slight bug here with 0 and 1, but it is better to leave things like this, and technically assume that 0 and 1 are not prime. More on this later.

In any case, what we have here is a bright definition, which is alive and vibrant, and this is more or less the definition and philosophy that we will use, in this book. Of course, our definition is something quite subtle and advanced, privileging the multiplication operation $\times$ over the addition operation $+$, and it will take us some time, and practice of basic number theory, in order to really agree with it. But no worries, we are here for that.

Finally, for ending this philosophical discussion, observe the stark contrast between our vibrant definition, and the previously mentioned abstract views on $\mathbb{N}$, coming from general modern mathematics and physics. So, are we here ahead of modern mathematics, modern physics, and perhaps other modern branches of science too, in what regards our attitude with respect to abstraction? And yes we are, the answer to this being:

FACT 1.2. *Numbers are the most concrete thing in the whole universe, mathematics and all other sciences accounted for. Study them, and you are into real.*

Which explains why so many people can be passioned by numbers, and with number theory itself being called Queen of Mathematics. It is all about being real, away from all sorts of abstractions. Numbers are alive and real, and this is why we love them.

Getting to work now, Fact 1.1 suggests looking into the prime numbers. But this is a quite dificult problem, let us leave that for later. We are in need of some tools for dealing with numbers, be them prime or not, and the simplest tools are, more numbers.

So, starting with the positive integers $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ that we know well, or perhaps that we want to better understand, let us construct more numbers. And here, things are quite straightforward. Indeed, once you have $\mathbb{N}$, solving the equation $a + b = c$ naturally leads you to negative numbers, and so to the set $\mathbb{Z}$ of all integers:

$$\mathbb{Z} = \mathbb{N} \cup (-\mathbb{N})$$

Then, once you have $\mathbb{Z}$, solving the equation $ab = c$ naturally leads you to fractions, so to the set of all these fractions $\mathbb{Q}$, also called rational numbers. So, this will be our starting point, for our mathematics in this book, the rationals $\mathbb{Q}$, defined as follows:

DEFINITION 1.3. *The rational numbers are the quotients $r = a/b$, with $a, b \in \mathbb{Z}$ and $b \neq 0$, identified according to the usual rule for quotients, namely:*

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

These quotients add, multiply and invert according to the following formulae:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \quad , \quad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd} \quad , \quad \left(\frac{a}{b}\right)^{-1} = \frac{b}{a}$$

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

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

DEFINITION 1.4. *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 an inverse $a^{-1}\in F$, satisfying $aa^{-1}=1$.*
(3) *The sum and product are compatible via $a(b+c)=ab+ac$.*

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 structure as follows:

$$1+1=0$$

To be more precise, according to our field axioms, all operations of type $a*b$ with $a,b=0,1$ are uniquely determined, except for $1+1$. You would say that we must 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, as follows:

PROPOSITION 1.5. *$\mathbb{Q}$ is the simplest field having the property $1+\ldots+1\neq 0$, in the sense that any field F satisfying this condition must contain $\mathbb{Q}$:*

$$\mathbb{Q}\subset F$$

However, when dropping the assumption $1+\ldots+1\neq 0$, the above conclusion fails, for instance for the field $\mathbb{F}_2=\{0,1\}$, with addition $1+1=0$.

PROOF. Here the first assertion 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 $\mathbb{Q}\subset F$. As for the second assertion, this follows from the above discussion. □

Thus, good news, we have some mathematics going on, eventually, with the above result being our first one in this book. As bad news, however, the above result is not exactly reassuring, because in our arithmetic quest, shall we study $\mathbb{Q}$ or $\mathbb{F}_2$ first, which one is the simplest. Not an easy question, but based on our previous philosophical discussion regarding $\mathbb{N}$, we can somehow conclude that fields with $|F|\simeq 0$, including those present before the Big Bang, are dangerous physics business, so we will leave them for later.

Many things can be done with $\mathbb{Q}$, but one thing that fails is solving $x^2=2$:

THEOREM 1.6. *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 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. □

As a conclusion, in order to advance with our mathematics, we are now in need to introduce the field of real numbers $\mathbb{R}$. This can be done in several ways, as follows:

THEOREM 1.7. *The field of real numbers $\mathbb{R}$ can be introduced as follows:*

(1) *As the field of Dedekind cuts of rationals, $x = \mathbb{Q}_{\leq x} \sqcup \mathbb{Q}_{>x}$.*
(2) *As the completion of $\mathbb{Q}$ with respect to $d(a, b) = |a - b|$.*
(3) *In decimal form, with detailed algorithms for $+$ and $\times$.*

PROOF. This is something quite non-trivial, the formal equivalence of what comes out of (1,2,3) I mean, and there is of course a philosophical discussion too, on which definition is the best, and which side, as number theorists, shall we be on. We will be rather preferring here (2), with (1) being a bit too algebraic, and (3), too analytic. □

Let us see now what the passage $\mathbb{Q} \to \mathbb{R}$ teaches us. Many things of course, to be explored later, but surprise, we run right away into a difficult question, namely:

QUESTION 1.8. *The reals are uncountable, so that we have*

$$|\mathbb{R}| > |\mathbb{Q}| = \infty$$

and is there anything in between.

As already mentioned, this is a difficult question. So, getting away now from this, obviously wrong way, let us stay with numbers and equations, and arithmetic, and beauty. Based on our previous experience with $\sqrt{2}$, which after all motivated our introduction of $\mathbb{R}$, we can try to look for the intermediate fields of numbers, as follows:

$$\mathbb{Q} \subset F \subset \mathbb{R}$$

Indeed, we can talk for instance about fields like $\mathbb{Q}[\sqrt{2}]$, as follows:

PROPOSITION 1.9. *The following is an intermediate field $\mathbb{Q} \subset F \subset \mathbb{R}$,*

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

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 1.6. As for the case of $\mathbb{Q}[\sqrt{n}]$, this is similar, again by using Theorem 1.6. □

However, in relation with extensions of $\mathbb{Q}$, the first question that comes to mind, in relation with square roots, is solving the following equation:

$$x^2 = -1$$

So, let us introduce the complex numbers $\mathbb{C}$. Many interesting things can be said here, and you certainly know all of them, with their summary being as follows:

THEOREM 1.10. *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) *With $z = r(\cos t + i \sin t)$, the products $z = z'z''$ are given by $r = r'r''$, $t = t' + t''$.*
(4) *We have $e^{it} = \cos t + i \sin t$, so we can write $z = re^{it}$.*
(5) *There are N solutions to the equation $z^N = 1$, called N-th roots of unity.*
(6) *Any degree 2 equation with complex coefficients has both roots in $\mathbb{C}$.*

PROOF. All this is very standard, and you surely know this, the idea being that (1,2) are both clear, then (3) follows from some trigonometry, done the old way, in the plane, then (4) follows from some heavy calculus, namely Taylor formula for $\exp, \sin, \cos$, and finally (5,6) both follow easily from (4), which is something very powerful. □

Quite remarkably, we have in fact the following result, generalizing what we know in degree 2, and telling us that with $\mathbb{C}$ we are safe, no need to look for more:

THEOREM 1.11. *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. □

With this discussed, let us go back to the question that we raised before, namely finding the intermediate fields $\mathbb{Q} \subset F \subset \mathbb{R}$. Obviously, the good question is as follows:

QUESTION 1.12. *What are the intermediate fields*

$$\mathbb{Q} \subset F \subset \mathbb{C}$$

between the rational numbers, and the complex numbers?

There is a lot of interesting theory that can be developed here, following Galois and others. However, we will leave this for later, chapter 4 below, because most of these fields F appear by adding various roots of polynomials, such as $\sqrt{2}$ or i, to the field $\mathbb{Q}$. And so, in order to develop Galois theory, we need some know-how, regarding the polynomials $P \in \mathbb{Q}[X]$ and their roots, and reaching to that knowledge level will take some time.

As a philosophical comment now, observe that Question 1.12 does not really help in relation with Question 1.8, because the fields that we can obtain from $\mathbb{Q}$ by adding various roots of polynomials $P \in \mathbb{Q}[X]$ are obviously all countable. Nevermind.

Still talking philosophy, a word for the analytic reader. With our field theory we are not at all away from analysis, quite the opposite. Indeed, while the usual spaces of functions are obviously not fields, analysis remains around the corner, due to:

THEOREM 1.13. *The quotients of complex polynomials, called rational funtions, when written in reduced form, as follows, with P, Q prime to each other,*

$$f = \frac{P}{Q}$$

are well-defined and continuous outside the zeroes $P_f \subset \mathbb{C}$ of Q, called poles of f:

$$f : \mathbb{C} - P_f \to \mathbb{C}$$

Also, these functions are stable under summing, making products and taking inverses,

$$\frac{P}{Q} + \frac{R}{S} = \frac{PS + QR}{QS} \quad , \quad \frac{P}{Q} \cdot \frac{R}{S} = \frac{PR}{QS} \quad , \quad \left(\frac{P}{Q}\right)^{-1} = \frac{Q}{P}$$

so they form a field $\mathbb{C}(X)$, called field of rational functions.

PROOF. Almost everything here is clear from definitions, and with the comment that, in what regards the term "pole", this 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. □

Still speaking fields, it is quite remarkable that both $\mathbb{R}$ and $\mathbb{R}^2$ have field structures. This fails for $\mathbb{R}^3$, but then for $\mathbb{R}^4$ something can be done, as follows:

THEOREM 1.14. *In contrast with $\mathbb{R}$, and with $\mathbb{R}^2 = \mathbb{C}$, which are fields:*

(1) *The vector space $\mathbb{R}^3$ does not have a multiplication, making it a field.*
(2) *For $\mathbb{R}^4$ however, something can be done, of rather physics flavor.*

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

(1) Some philosophy first. I can hear you screaming what's wrong with this book, and with this book author. Good point, and in answer, it's just that we want to talk about the above questions, which are both very beautiful, but also difficult, to the point that pulling out a formal proof, or even a formal statement, would be quite difficult.

(2) So, shall we give up? Not with me at the helm. So, in order to find a solution to our mathematical ethics questions, remember physics classes? These basically consist of a physics professor, reporting, via a mixture of clean and dirty math, on what mankind found about the various fields in physics: gravitational, electric, magnetic and so on.

(3) And guess what, we will do the same here. So, the present theorem and proof will be a sort of physics class, with me reporting, exactly as a physics professor in school,

via a mixture of clean and dirty math, on what mankind found about the various field structures, mathematically speaking of course, on $\mathbb{R}, \mathbb{R}^2, \mathbb{R}^3, \mathbb{R}^4$ and so on.

(4) Feel free of course to disagree with this, and pass to section 1b below. But better stay with me, many interesting things to be learned. Also, importantly, as previously said, on the occasion of Fact 1.1 and Fact 1.2, number theorists tend to regard their objects, be them numbers, prime numbers, and fields too, as being "alive and real". So, after all, we are some sort of physicists, in number theory, try getting along with this.

(5) And for ending this discussion, a quote from Hermann Weyl, who was one of the greatests mathematicians ever: "Among the correct and the beautiful, I always chose the beautiful". There is something really deep about this quote, profoundly going into what is mathematics, what is physics, what is life, what are we humans, what is modesty, what is humanity. Think of this from time to time, and with some practice of mathematics helping, you will certainly end up, one day, understanding what Weyl was saying.

(6) Getting to math now, let us first examine the field structures on $\mathbb{R}^N$, with $N \in \mathbb{N}$ arbitrary. A first idea, which is very natural, is that any multiplication on $\mathbb{R}^N$ must come by linearity from a multiplication on the unit sphere $S_{\mathbb{R}}^{N-1} \subset \mathbb{R}^N$. That is, once we know how to multiply the norm one vectors $x, y \in S_{\mathbb{R}}^{N-1}$, we can set, by linearity:

$$(\lambda x) * (\mu y) = (\lambda\mu)(x * y)$$

At the level of examples, this is certainly what happens at $N = 1, 2$, where the corresponding unit spheres are as follows, and with the multiplication on $\mathbb{R}^N$ itself appearing as above, from the obvious multiplication on these unit spheres, by linearity:

$$S_{\mathbb{R}}^0 = \{-1, 1\} \quad , \quad S_{\mathbb{R}}^1 = \mathbb{T}$$

(7) In practice now, such ideas require first proving that $||x|| = ||y|| = 1$ implies $||x * y|| = 1$, with $||x|| = \sqrt{\sum x_i^2}$ being the usual norm, and while not exactly obvious, this can be done indeed. As another remark, getting back now to $N = 1, 2$, while the possible multiplication on $S_{\mathbb{R}}^0 = \{-1, 1\}$ is unique, $(-1)^2 = 1$, in what regards the possible multiplications on $S_{\mathbb{R}}^1 = \mathbb{T}$ things are more complicated, of topology flavor. So, as conclusion, it is pretty much clear that all this leads us into geometry, and topology.

(8) Moving now to $\mathbb{R}^3$, you would say that the vector product $x \times y$ does the job, but this is wrong, because $x \sim y$ implies $x \times y = 0$, so definitely wrong way. However, thinking well, a multiplication on $\mathbb{R}^3$ would induce a multiplication on the unit sphere $S_{\mathbb{R}}^2 \subset \mathbb{R}^3$, as explained above, and the point is that there is a topological obstruction to this. However, this obstruction is a bit difficult to explain, involving some advanced mathematics, and our result at $N = 3$ being negative anyway, we will not further insist on this.

(9) Getting now to $\mathbb{R}^4$, as a good surprise here, the unit sphere $S^3_{\mathbb{R}} \subset \mathbb{R}^4$ is naturally a group, $S^3_{\mathbb{R}} = SU_2$. Indeed, solving $U^* = U^{-1}$ under the assumption $\det U = 1$ gives:

$$SU_2 = \left\{ \begin{pmatrix} a & b \\ -\bar{b} & a \end{pmatrix} \Big| |a|^2 + |b|^2 = 1 \right\}$$

Here we use complex numbers, and in real number notation, the result is:

$$SU_2 = \left\{ \begin{pmatrix} x+iy & z+it \\ -z+it & x-iy \end{pmatrix} \Big| x^2 + y^2 + z^2 + t^2 = 1 \right\}$$

(10) But this is obviously good news, we more or less solved our problem, and it remains to work out the details. So, consider the following matrices:

$$c_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad , \quad c_2 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \quad , \quad c_3 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad , \quad c_4 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$$

In terms of these matrices, which by the way are called Pauli spin matrices, discovered by Pauli in relation with quantum mechanics, but let us not get into this here, maybe later, towards the end of the present book, our result above reads:

$$SU_2 = \left\{ c_1 x + c_2 y + c_3 z + c_4 t \Big| x^2 + y^2 + z^2 + t^2 = 1 \right\}$$

In order to figure out how the resulting multiplication on $\mathbb{R}^4$ looks like, we must first multiply the Pauli matrices. Their products are given by the following formulae:

$$c_2^2 = c_3^2 = c_4^2 = -1$$

$$c_2 c_3 = -c_3 c_2 = c_4$$

$$c_3 c_4 = -c_4 c_3 = c_2$$

$$c_4 c_2 = -c_2 c_4 = c_3$$

Thus, we are led in this way to a multiplication on $\mathbb{R}^4$, as stated.

(11) Alternatively, we have the following real matrices, multiplying quite similarly:

$$1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad , \quad i = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$j = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad , \quad k = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Thus, one way or another, we are led to the conclusion in the statement.

(12) Finally, there is as well a purely algebraic approach to this, using formal numbers $1, i, j, k$, called quaternions, with $1, i$ being the $1, i$ that we know, and with j, k being

constructed similarly, a bit like i was, formally via $i^2 = -1$, when introducing $\mathbb{C}$. To be more precise, the multiplication rules for i, j, k, found by Hamilton, are as follows:

$$i^2 = j^2 = k^2 = ijk = -1$$

Observe that these are precisely the multiplication rules for the Pauli matrices, from (10) above. Thus, we are again led to the conclusion in the statement. □

All this way very nice, but was probably more confusing than enlightening, as interesting physics is supposed to be. Adding to the plot, some of the above discoveries were followed by euphoria, first with Hamilton who discovered (12) when walking on a street of Dublin, and was so happy that he carved that i, j, k formula into the nearest stone bridge, then with Pauli who, well, was Pauli and did nothing spectacular when finding (10), and then later with Dirac, who came upon a key version of the Pauli matrices, related to (11), and claimed that he found these by watching logs burning in the fireplace.

And there was probably some more euphoria later too, when Pauli and Dirac were both awarded the Nobel Prize in Physics, for their discoveries, and with that prize, coming as you probably know, with a good sum of money, from the works of Alfred Nobel.

Getiting back now to math, let us record, as a conclusion to the above discussion, and as a formal replacement of Theorem 1.14, featuring more math and less physics:

THEOREM 1.15. *In analogy with $\mathbb{R}^2 = \mathbb{R}[i]$, with $i^2 = -1$, which is a field, we can talk about $\mathbb{R}^4 = \mathbb{R}[i, j, k]$, with the following multiplication rules for i, j, k,*

$$i^2 = j^2 = k^2 = ijk = -1$$

called quaternion multiplication rules, and with this being a noncommutative field, in the sense that all the field axioms are satisfied, except for $ab = ba$.

PROOF. This follows indeed from the discussion in the proof of Theorem 1.14, and with the comment of course that, for proper understanding, all that discussion is needed. □

Needless to say, many other things can be said about $\mathbb{R}^2, \mathbb{R}^3, \mathbb{R}^4$, along the above lines, and also about $\mathbb{R}^N$ with arbitrary $N \in \mathbb{N}$, including $\mathbb{R}^8$ which is special too, all mixing geometry and physics. But this is quite advanced material, and we will stop here.

1b. Prime numbers

Time now to get into prime numbers, which will be a main theme of discussion, for 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 already talked about prime numbers in the above, and even used some of their basic properties, that you were certainly very familiar with, but time now to review all this, on a more systematic basis. First, as definition for them, we have:

DEFINITION 1.16. *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 1.17. *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 1.16, 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 1.18. *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 1.19. *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:

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

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

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

CONJECTURE 1.20. *A good analyst, physicist and engineer would probably have no troubles in elucidating everything about primes, using the sieve method.*

Along the same lines, analysis, this time making the connection with calculus, we have the following famous formula of Euler, improving Theorem 1.18:

THEOREM 1.21. *We have the following formula, implying $|P| = \infty$:*

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

Moreover, the strict partial sums of this series satisfy $S_N > \log\log N - 1/2$.

PROOF. Here is the original proof, due to Euler. The idea is to use Theorem 1.17, 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:

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

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 P_N of all primes $p<N$. We obtain in this way the following estimate, and with exercise for you, to work out the details:

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

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

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

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

$$\sum_{p\in P_N} \frac{1}{p} \simeq \log\log N$$

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 in Part III of the present 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:

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

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. More on all this also in Part III.

(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 all over this book, in Part II, Part III, and Part IV too.

Finally, getting now back to Earth, there are many other things that can be said about the prime numbers, at the elementary level. But no hurry with this, at least for the moment. We will be back to prime numbers, on a regular basis, throughout this book.

1c. Congruence, tricks

With the above discussed, prime numbers sort of dragging us into analysis, let us turn now to algebra, which will be our main occupation in the present Part I of this book. To start with, we will be mostly interested in congruence questions, based on:

DEFINITION 1.22. *We say that $a, b \in \mathbb{Z}$ are congruent modulo $c \in \mathbb{Z}$, and write $a = b(c)$, when c divides $b - a$.*

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

PROPOSITION 1.23. *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:

PROPOSITION 1.24. *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 Proposition 1.24 is a bit ill-formulated, mixing things written in basis 10 and basis 14, and we will leave fixing all this, with a fully correct mathematical statement, as another instructive exercise for you.

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

THEOREM 1.25. *We have the following congruence, for any prime p,*

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

called Fermat's little theorem.

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

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

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

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

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

The Fermat theorem is particularly interesting when extended from the integers to the arbitrary field case, and can be used in order to elucidate the structure of finite fields. In order to discuss this question, let us start with some basic facts, as follows:

THEOREM 1.26. *Given a field F, define its characteristic $p = char(F)$ as being the smallest $p \in \mathbb{N}$ such that the following happens, and as $p = 0$, if this never happens:*

$$\underbrace{1 + \ldots + 1}_{p\ times} = 0$$

Then, assuming $p > 0$, this characteristic p must be a prime number, we have a field embedding $\mathbb{F}_p \subset F$, and $q = |F|$ must be of the form $q = p^k$, with $k \in \mathbb{N}$.

PROOF. Very crowded statement that we have here, the idea being as follows:

(1) The fact that $p > 0$ must be prime comes by contradiction, by using:

$$(\underbrace{1 + \ldots + 1}_{a\ times}) \times (\underbrace{1 + \ldots + 1}_{b\ times}) = \underbrace{1 + \ldots + 1}_{ab\ times}$$

Indeed, assuming that we have $p = ab$ with $a, b > 1$, the above formula corresponds to an equality of type $AB = 0$ with $A, B \neq 0$ inside F, which is impossible.

(2) Back to the general case, F has a smallest subfield $E \subset F$, called prime field, consisting of the various sums $1 + \ldots + 1$, and their quotients. In the case $p = 0$ we obviously have $E = \mathbb{Q}$. In the case $p > 0$ now, the multiplication formula in (1) shows that the set $S = \{1 + \ldots + 1\}$ is stable under taking quotients, and so $E = S$.

(3) Now with $E = S$ in hand, we obviously have $(E, +) = \mathbb{Z}_p$, and since the multiplication is given by the formula in (1), we conclude that we have $E = \mathbb{F}_p$, as a field. Thus, in the case $p > 0$, we have constructed an embedding $\mathbb{F}_p \subset F$, as claimed.

(4) In the context of the above embedding $\mathbb{F}_p \subset F$, we can say that F is a vector space over $\mathbb{F}_p$, and so we have $|F| = p^k$, with $k \in \mathbb{N}$ being the dimension of this space. □

In relation with Fermat, we can extend the trick in the proof there, as follows:

PROPOSITION 1.27. *In a field F of characteristic $p > 0$ we have*

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

for any two elements $a, b \in F$.

PROOF. We have indeed the computation, exactly as in the proof of Fermat, by using the fact that the non-trivial binomial coefficients are all multiples of p:

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

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

Observe that we can iterate the Fermat formula, and we obtain $(a + b)^r = a^r + b^r$ for any power $r = p^s$. In particular we have, with $q = |F|$, the following formula:

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

But this is something quite interesting, showing that the following subset of F, which is closed under multiplication, is closed under addition too, and so is a subfield:

$$E = \left\{ a \in F \middle| a^q = a \right\}$$

So, what is this subfield $E \subset F$? In the lack of examples, or general theory for subfields $E \subset F$, we are a bit in the dark here, but it seems quite reasonable to conjecture that we

have $E = F$. Thus, our conjecture would be that we have the following formula, for any $a \in F$, and with this being the field extension of the Fermat theorem itself:

$$a^q = a$$

Now that we have our conjecture, let us think at a potential proof. And here, by looking at the proof of the Fermat theorem, the recurrence method from there, based on $a \to a + 1$, cannot work as such, and must be suitably fine-tuned.

Thinking a bit, the recurrence from the proof of Fermat somehow rests on the fact that the additive group $\mathbb{Z}$ is singly generated, by $1 \in \mathbb{Z}$. Thus, we need some sort of field extension of this single generation result, and in the lack of something additive here, the following theorem, which is something multiplicative, comes to the rescue:

THEOREM 1.28. *Given a field F, any finite subgroup of its multiplicative group*

$$G \subset F - \{0\}$$

must be cyclic.

PROOF. This can be done via some standard arithmetics, as follows:

(1) Let us pick an element $g \in G$ of highest order, $n = ord(g)$. Our claim, which will easily prove the result, is that the order $m = ord(h)$ of any $h \in G$ satisfies $m|n$.

(2) In order to prove this claim, let $d = (m, n)$, write $d = am + bn$ with $a, b \in \mathbb{Z}$, and set $k = g^a h^b$. We have then the following computations:

$$k^m = g^{am} h^{bm} = g^{am} = g^{d-bn} = g^d$$

$$k^n = g^{an} h^{bn} = h^{bn} = h^{d-am} = h^d$$

By using either of these formulae, say the first one, we obtain:

$$k^{[m,n]} = k^{mn/d} = (k^m)^{n/d} = (g^d)^{n/d} = g^n = 1$$

Thus $ord(k)|[m, n]$, and our claim is that we have in fact $ord(k) = [m, n]$.

(3) In order to prove this latter claim, assume first that we are in the case $d = 1$. But here the result is clear, because the formulae in (2) read $g = k^m, h = g^n$, and since $n = ord(g), m = ord(g)$ are prime to each other, we conclude that we have $ord(k) = mn$, as desired. As for the general case, where d is arbitrary, this follows from this.

(4) Summarizing, we have proved our claim in (2). Now since the order $n = ord(g)$ was assumed to be maximal, we must have $[m, n]|n$, and so $m|n$. Thus, we have proved our claim in (1), namely that the order $m = ord(h)$ of any $h \in G$ satisfies $m|n$.

(5) But with this claim in hand, the result follows. Indeed, since the polynomial $x^n - 1$ has all the elements $h \in G$ as roots, its degree must satisfy $n \geq |G|$. On the other hand, from $n = ord(g)$ with $g \in G$, we have $n||G|$. We therefore conclude that we have $n = |G|$, which shows that G is indeed cyclic, generated by the element $g \in G$. □

We can now extend the Fermat theorem to the finite fields, as follows:

THEOREM 1.29. *Given a finite field F, with $q = |F|$ we have*

$$a^q = a$$

for any $a \in F$.

PROOF. According to Theorem 1.28 the multiplicative group $F - \{0\}$ is cyclic, of order $q - 1$. Thus, the following formula is satisfied, for any $a \in F - \{0\}$:

$$a^{q-1} = 1$$

Now by multiplying by a, we are led to the conclusion in the statement, with of course the remark that the formula there trivially holds for $a = 0$. □

Observe that our proof of Theorem 1.29 finally did not use Proposition 1.27. The reason behind this is most likely that Theorem 1.28 is something quite powerful.

1d. Finite fields

The Fermat polynomial $X^p - X$ is something very useful, and its field generalization $X^q - X$, with $q = p^k$ prime power, can be used in order to elucidate the structure of finite fields. In order to discuss this question, let us start with a basic fact, as follows:

PROPOSITION 1.30. *Given a finite field F, we have*

$$X^q - X = \prod_{a \in F} (X - a)$$

with $q = |F|$.

PROOF. We know from the Fermat theorem above that we have $a^q = a$, for any $a \in F$. We conclude from this that all the elements $a \in F$ are roots of the polynomial $X^q - X$, and so this polynomial must factorize as in the statement. □

The continuation of the story is more complicated, as follows:

THEOREM 1.31. *For any prime power $q = p^k$ there is a unique field $\mathbb{F}_q$ having q elements. At $k = 1$ this is the usual $\mathbb{F}_p$, and in general, this is the field making*

$$X^q - X = \prod_{a \in F} (X - a)$$

happen, in some abstract algebraic sense.

PROOF. We are punching here a bit above our weight, the idea being as follows:

(1) At $k = 1$ there is nothing much to be said, because the prime field embedding $\mathbb{F}_p \subset F$ found in Theorem 1.26 must be an isomorphism. Thus, done with this.

(2) At $k \geq 2$ however, both the construction and uniqueness of $\mathbb{F}_q$ are non-trivial. However, the idea is not that complicated. Indeed, instead of struggling first with finding a model for $\mathbb{F}_q$, and then struggling some more with proving the uniqueness, the point is that we can solve both these problems, at the same time, by looking at $X^q - X$.

(3) To be more precise, this polynomial $X^q - X$ must have some sort of abstract, minimal "splitting field", and this is how $\mathbb{F}_q$ comes, both existence and uniqueness.

(4) However, all this is a bit technical, and we will defer the details here to chapter 4 below, when investigating more systematically field theory, and such questions. □

As an application, let us discuss projective geometry over finite fields. Here are some mathematical axioms, coming as usual from the ancient Greeks, to start with:

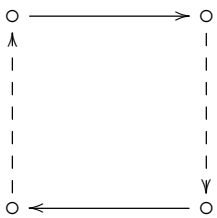
DEFINITION 1.32. *A projective space is a space consisting of points and lines, subject to the following conditions:*

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

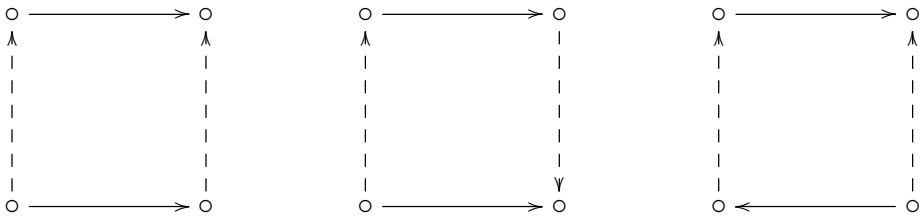
As a basic example, we can define the projective space $P_{\mathbb{R}}^{N-1}$ as being the space of lines in $\mathbb{R}^N$ passing through the origin. In small dimensions, the situation is as follows:

(1) $P_{\mathbb{R}}^1$ is the usual circle. Indeed, a line in $\mathbb{R}^2$ passing through the origin corresponds to 2 opposite points on the unit circle $\mathbb{T} \subset \mathbb{R}^2$, and so $P_{\mathbb{R}}^1$ corresponds to the upper semicircle of $\mathbb{T}$, with the endpoints identified, which is a circle, $P_{\mathbb{R}}^1 = \mathbb{T}$.

(2) $P_{\mathbb{R}}^2$ is some sort of twisted sphere. Indeed, $P_{\mathbb{R}}^2$ corresponds to the upper hemisphere of the sphere $S_{\mathbb{R}}^2 \subset \mathbb{R}^3$, with the points on the equator identified via $x = -x$. But we can deform if we want the hemisphere into a square, with the equator becoming the boundary of this square, and in this picture, the $x = -x$ identification corresponds to a "identify opposite edges, with opposite orientations" folding method for the square:



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



But these give the torus, the Klein bottle, and the Klein bottle again. So, as a conclusion, $P^2_{\mathbb{R}}$ is some sort of twisted sphere, reminding these, and which lives in $\mathbb{R}^4$.

Getting now to the finite fields, we can do here something similar, as follows:

THEOREM 1.33. *Given a field F, we can talk about the projective plane P^2_F, as being the space of lines in F^3 passing through the origin, having cardinality*

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

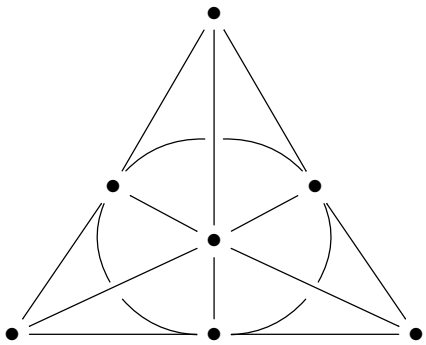
where $q = |F|$, in the case where our field F is finite.

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

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

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

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



Here the circle in the middle is by definition a line, and with this convention, the basic projective geometry axioms in Definition 1.32 are satisfied, in the sense that any two points determine a line, and any two lines determine a point. And isn't this magic.

As a conclusion, we have seen that a lot of exciting number theory can be developed, at the elementary level. However, there is a bit of a philosophical problem with our theory, because, after all, is that something algebraic, or analytic, or geometric. Not an easy question, and in what follows we will first look into algebra, in the remainder of this Part I, and in Part II too, and then in Part III and Part IV we will go with analysis.

1e. Exercises

This was a very dense chapter, and up to you to think what you want of this, depending on your previous experience with numbers. Anyway, as exercises, we have:

EXERCISE 1.34. *Review the various definitions of* $\mathbb{R}$.

EXERCISE 1.35. *Review the definition and basic properties of* $\mathbb{C}$.

EXERCISE 1.36. *Learn more about quaternions, and about octonions too.*

EXERCISE 1.37. *Design a mechanical sieve machine, producing primes.*

EXERCISE 1.38. *Work out the details for the Euler formula, in* $\log\log N$ *form.*

EXERCISE 1.39. *Learn more about finite fields, and their classification.*

EXERCISE 1.40. *Do some geometry, affine or projective, over the finite fields.*

EXERCISE 1.41. *Work out examples of infinite fields of nonzero characteristic.*

As bonus exercise, study a bit the Riemann zeta function ζ, that will come back only later, in this book. And this because, all number theorists spend some time with ζ.

CHAPTER 2

Equations, roots

2a. Resultant, discriminant

We have seen in the previous chapter that many number theory questions lead us into computing roots of polynomials $P \in \mathbb{Q}[X]$. We will investigate here such questions, first in the present chapter with a detailed study of the arbitrary polynomials $P \in \mathbb{C}[X]$, and their roots, often by using analytic methods, and then in chapter 4 below, following Galois, mostly for the polynomials $P \in \mathbb{Q}[X]$, using advanced algebraic methods.

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

THEOREM 2.1. *The solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{C}$ are*

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

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

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

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

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

Very nice all this, and you would probably say that the story is over here with degree 2, matters to be relegated to the elementary school. However, not really. Have you noticed that at the university, professors are usually faster than students, in dealing with degree 2? I am probably not supposed to talk about our secrets, but here are our tricks:

TRICKS 2.2. *The following happen:*

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

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

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

In order to find the eigenvalues r, s, you are certainly very used to compute the characteristic polynomial, then apply Theorem 2.1. But my point is that this characteristic polynomial is of the form $x^2 - ax + b$, with $a = Tr(a)$ and $b = \det A$, so we can normally apply the trick in (1), provided of course that r, s are simple numbers, say integers.

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

WARNING 2.3. *The above tricks work in pure mathematics, where the numbers r, s that we can meet are usually integers, or rationals. In applied mathematics, however, the numbers that we meet are integers or rationals with probability $P = 0$, so no tricks.*

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

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

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

THEOREM 2.4. *Given a monic polynomial $P \in \mathbb{C}[X]$, factorized as*

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

the following happen:

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

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

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

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

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

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

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

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

$$S_r=\sum_i a_i^r$$

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

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

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

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

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

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

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

THEOREM 2.5. *Given two polynomials $P,Q\in\mathbb{C}[X]$, written as*

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

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

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

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. □

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

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

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

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

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

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

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

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

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

THEOREM 2.6. *The resultant of two polynomials, written as*

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

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

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

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 2.4 and 2.5. □

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

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

its discriminant, defined as being the following quantity,

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

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

PROOF. The fact that the discriminant $\Delta(P)$ is a polynomial in the coefficients of P, with integer coefficients, comes from Theorem 2.5, coupled with the fact that the division by the leading coefficient a is indeed possible, under $\mathbb{Z}$, as being shown by the following formula, which is written of course a bit informally, coming from Theorem 2.6:

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

Also, the fact that we have $\Delta(P) = 0$ precisely when P has a double root is clear from Theorem 2.5. Finally, let us mention that the sign $(-1)^{\binom{N}{2}}$ is there for various reasons, including the compatibility with some well-known formulae, at small values of $N \in \mathbb{N}$, such as $\Delta(P) = b^2 - 4ac$ in degree 2, that we will discuss in a moment. □

As already mentioned, by using Theorem 2.6, we have an explicit formula for the discriminant, as the determinant of a certain matrix. There is a lot of theory here, and in order to get into this, let us first see what happens in degree 2. Here we have:

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

Thus, the resultant is given by the following formula:

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

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

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

Alternatively, we can use the formula in Theorem 2.6, and we obtain:

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

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

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

THEOREM 2.8. *The discriminant of a polynomial P is given by the formula*

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

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

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

(1) The first thought goes to the formula in Theorem 2.5, so let us see what that formula teaches us, in the case $Q = P'$. Let us write P, P' as follows:

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

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

According to Theorem 2.5, the resultant of P, P' is then given by:

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

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

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

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

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

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

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

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

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

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

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

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

(4) Time now to put everything together. By taking the formula in (2), making the normalizations in Theorem 2.7, and then using the formula found in (3), we obtain:

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

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

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

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

As applications now, the formula in Theorem 2.8 is quite useful for the real polynomials $P\in\mathbb{R}[X]$ in small degree, because it allows to say when the roots are real, or complex, or at least have some partial information about this. For instance, we have:

PROPOSITION 2.9. *Consider a polynomial with real coefficients, $P\in\mathbb{R}[X]$, assumed for simplicity to have nonzero discriminant, $\Delta\neq 0$.*

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

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

(1) The first assertion is something that you certainly know, coming from Theorem 2.1, but let us see how this comes via the formula in Theorem 2.8, namely:

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

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

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

Thus $\Delta>0$ amounts in saying that we have $(r_1-r_2)^2>0$. Now since r_1,r_2 are conjugate, and with this being something trivial, meaning no need here for the computations in Theorem 2.1, we conclude that $\Delta>0$ means that r_1,r_2 are real, as stated.

(2) In degree $N=3$ now, we know from analysis that P has at least one real root, and the problem is whether the remaining 2 roots are real, or complex conjugate. For this purpose, we can use the formula in Theorem 2.8, which in degree 3 reads:

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

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

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

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

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

Finally, as another application of all this, worth mentioning, we have:

THEOREM 2.10. *The diagonalizable matrices are dense.*

PROOF. As a first observation, this is something extremely useful, more or less allowing you in practice to assume that any matrix $A \in M_N(\mathbb{C})$ is diagonalizable, but of course do not try this at home, unless you know what you're doing. As for the proof, this is non-trivial, and there are actually two standard proofs, both non-trivial, as follows:

(1) Via the pedestrian way, by using the Jordan form. Here you have to learn well the Jordan form, and good luck with that, and once that done, you can argue that by perturbing the Jordan blocks, in the obvious way, you can arrange up to epsilon as for your matrix to have distinct eigenvalues, and so to be diagonalizable.

(2) As a geometry king, using the discriminant. Indeed, for a matrix $A \in M_N(\mathbb{C})$, with characteristic polynomial P_A, having distinct eigenvalues means:

$$\Delta(P_A) \neq 0$$

But this is the complement of a hypersurface, which is dense, and since all these matrices are diagonalizable, the diagonalizable matrices are dense too. Just like that. □

2b. Cardano formula

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

THEOREM 2.11. *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 2.5 and Theorem 2.6, 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 2.5. 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 2.6, 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 2.6 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 2.7, 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 2.12. *For a normalized degree 3 equation, namely*

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

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

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

is a real solution of our equation.

PROOF. The formula of Δ is clear from definitions, and with $108=4\times 27$. Now with x as in the statement, by using $(a+b)^3=a^3+b^3+3ab(a+b)$, we have:

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

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

Regarding the other roots, we know from Proposition 2.9 that these are both real when $\Delta<0$, and complex conjugate when $\Delta<0$. Thus, in the context of Theorem 2.12, the other two roots are complex conjugate, the formula for them being as follows:

PROPOSITION 2.13. *For a normalized degree 3 equation, namely*

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

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

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

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

with $w=e^{2\pi i/3}$ are the complex conjugate solutions of our equation.

PROOF. As before, 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. □

As a conclusion, we have the following statement, unifying the above:

THEOREM 2.14. *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*

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

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

PROOF. This follows indeed from Theorem 2.12 and Proposition 2.13. Alternatively, we can redo the computation in their proof, which was nearly identical anyway, in the present setting, with x being given by the above formula, by using $w^3=1$. □

As a comment here, the formula in Theorem 2.14 holds of course in the case $\Delta>0$ too, and also when the coefficients are complex numbers, $p,q\in\mathbb{C}$, and this due to the fact that the proof rests on the nearly trivial computation from the proof of Theorem 2.12, or of Proposition 2.13. However, these extensions are quite often not very useful, because when it comes to extract all the above square and cubic roots, for complex numbers, you can well end up with the initial question, the one that you started with.

Thus, as a conclusion to this, Theorem 2.14 as formulated above is what can be best said about the degree 3 equations. There are of course many versions of it, and slight generalizations, but in practice, Theorem 2.14 is what mostly matters.

2c. Higher degree

In higher degree things become quite complicated. In degree 4, to start with, we first have the following result, dealing with the discriminant and its applications:

THEOREM 2.15. *The discriminant of* $P=ax^4+bx^3+cx^2+dx+e$ *is given by the following formula:*

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

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 2.7, with the resultant computed via Theorem 2.6, 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 2.15:

PROPOSITION 2.16. *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 2.15, 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 2.15. □

We still have some work to do. Indeed, looking back at what we did in degree 3, the passage there from Theorem 2.11 to Theorem 2.12 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 2.16, 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 2.17. *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 2.16, 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 2.15. □

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

THEOREM 2.18. *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:

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

The idea will be that of adding a suitable common term, to both sides, as to make square on both sides, as to eventually end with a sort of double quadratic equation. For this purpose, our claim is that what we need is a number y satisfying:

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

Indeed, assuming that we have this number y, our equation becomes:

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

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

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

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

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

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

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

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

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

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

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

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

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

(4) Now by cutting a $\sqrt{2}$ factor from everything, this gives the formulae in the statement. As for the last claim, regarding the nature of y, this comes from Cardano. □

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 2.18, is as follows:

THEOREM 2.19 (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 goes that this is what comes from Cardano, but depressing and normalizing and solving $(y^2-3r)(y-3p)=2q^2$ makes it for too many operations, so the most pragmatic is to simply check this equation. With y as above, we have:

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

With this in hand, we have the following computation:

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

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

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

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

In degree 5 and more, things become fairly complicated, and we will be back to this phenomenon, with some explanations, in chapter 4 below, when doing Galois theory.

However, in higher degree we do have some arithmetic tricks, for computing the integer or rational roots of polynomials having integer of rational coefficients. There are a lot of analytic tricks too, which can be both of real and complex analysis nature.

2d. Roots of unity

We would like to end this chapter with something geometric and refreshing, namely the roots of unity. These are arguably the most famous roots of polynomials ever, and many things can be said about them. At the beginning of everything, we have:

PROPOSITION 2.20. *The roots of unity, $\{w^k\}$ with $w=e^{2\pi i/N}$, have the property*

$$\sum_{k=0}^{N-1}(w^k)^s=N\delta_{N|s}$$

for any exponent $s\in\mathbb{N}$, where on the right we have a Kronecker symbol.

PROOF. The numbers in the statement, when written more conveniently as $(w^s)^k$ with $k=0,\ldots,N-1$, form a certain regular polygon in the plane P_s. Thus, if we denote by C_s the barycenter of this polygon, we have the following formula:

$$\frac{1}{N}\sum_{k=0}^{N-1}w^{ks}=C_s$$

Now observe that in the case $N\not|\, s$ our polygon P_s is non-degenerate, circling around the unit circle, and having center $C_s=0$. As for the case $N|s$, here the polygon is degenerate, lying at 1, and having center $C_s=1$. Thus, we have the following formula:

$$C_s=\delta_{N|s}$$

Thus, we obtain the formula in the statement. □

With the help of roots of unity, we can construct the Fourier matrix, as follows:

PROPOSITION 2.21. *The Fourier matrix F_N, which is as follows, with $w = e^{2\pi i/N}$,*

$$F_N = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & w & w^2 & \ldots & w^{N-1} \\ 1 & w^2 & w^4 & \ldots & w^{2(N-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w^{N-1} & w^{2(N-1)} & \ldots & w^{(N-1)^2} \end{pmatrix}$$

has the property that $F_N/\sqrt{N}$ is unitary.

PROOF. Here the fact that the rows are pairwise orthogonal follows from Proposition 2.20, and after rescaling by $\sqrt{N}$, these rows become of norm 1 too, as desired. □

As a comment, the name comes from the fact that F_N is the matrix of the Fourier transform over the cyclic group $\mathbb{Z}_N$, also called "discrete Fourier transform". We will be back to this in chapter 4 below, with details, when discussing group theory.

However, even before that, we can still do some interesting discrete Fourier analysis, with some ad-hoc statements and proofs. For this purpose, it is convenient, as in discrete Fourier analysis, to make the convention that our matrix indices are $i, j \in \{0, 1, \ldots, N-1\}$, taken modulo N. With this convention, the formula of F_N simply becomes:

$$F_N = (w^{ij})_{ij}$$

As a first result, the diagonalization problem for the all-one, or flat matrix $\mathbb{I}_N$ can be solved by using the Fourier matrix F_N, in the following elegant way:

THEOREM 2.22. *The flat matrix $\mathbb{I}_N$ diagonalizes as follows,*

$$\begin{pmatrix} 1 & \ldots & \ldots & 1 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 1 & \ldots & \ldots & 1 \end{pmatrix} = \frac{1}{N} F_N \begin{pmatrix} N & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} F_N^*$$

with $F_N = (w^{ij})_{ij}$ being the Fourier matrix.

PROOF. Since $P_N = \mathbb{I}_N/N$ is the projection on the all-one vector, we are left with finding the 0-eigenvectors of $\mathbb{I}_N$, which amounts in solving the following equation:

$$x_0 + \ldots + x_{N-1} = 0$$

For this purpose, we use the formula in Proposition 2.20. That formula shows that for any $j = 1, \ldots, N-1$, the vector $v_j = (w^{ij})_i$ is a 0-eigenvector of our matrix. Moreover, these vectors are pairwise orthogonal, because we have:

$$< v_j, v_k > = \sum_i w^{ij-ik} = N\delta_{jk}$$

Thus, we have our basis $\{v_1, \ldots, v_{N-1}\}$ of 0-eigenvectors, and since the N-eigenvector is $\xi = v_0$, the passage matrix P that we are looking is given by:

$$P = \begin{bmatrix} v_0 & v_1 & \ldots & v_{N-1} \end{bmatrix}$$

But this is precisely the Fourier matrix, $P = F_N$, and we are done. □

Moving ahead, observe that the flat matrix $\mathbb{I}_N$ is circulant, in the following sense:

DEFINITION 2.23. *A real or complex matrix M is called circulant if*

$$M_{ij} = \xi_{j-i}$$

for a certain vector ξ, with the indices taken modulo N.

In other words, a matrix is called circulant when its entries "circulate" downwards and to the right. As an example, at $N = 4$, the matrix must be as follows:

$$M = \begin{pmatrix} a & b & c & d \\ d & a & b & c \\ c & d & a & b \\ b & c & d & a \end{pmatrix}$$

The circulant matrices are certainly beautiful mathematical objects, but they appear in many serious problems as well. For instance, the Sylvester discriminant matrices, that we have struggled a bit with, in the above, are not far from being circulant.

The point now is that, while certainly gently looking, the circulant matrices can be quite diabolic, when it comes to diagonalization, and other problems. Fortunately the complex numbers and the Fourier matrices come to the rescue, and we have:

THEOREM 2.24. *For a matrix $M \in M_N(\mathbb{C})$, the following are equivalent:*

(1) *M is circulant, $M_{ij} = \xi_{j-i}$, for a certain vector $\xi \in \mathbb{C}^N$.*
(2) *M is Fourier-diagonal, $M = F_N Q F_N^*$, for a certain diagonal matrix Q.*

Moreover, if these conditions hold, then $\xi = F_N^* q$, where $q = (Q_{11}, \ldots, Q_{NN})$.

PROOF. This follows from some basic computations with roots of unity, as follows:

(1) $\implies$ (2) Assuming $M_{ij} = \xi_{j-i}$, the matrix $Q = F_N^* M F_N$ is indeed diagonal, as shown by the following computation:

$$\begin{aligned} Q_{ij} &= \sum_{kl} w^{-ik} M_{kl} w^{lj} \\ &= \sum_{kl} w^{jl-ik} \xi_{l-k} \\ &= \sum_{kr} w^{j(k+r)-ik} \xi_r \\ &= \sum_r w^{jr} \xi_r \sum_k w^{(j-i)k} \\ &= N\delta_{ij} \sum_r w^{jr} \xi_r \end{aligned}$$

(2) $\implies$ (1) Assuming $Q = diag(q_1, \ldots, q_N)$, the matrix $M = F_N Q F_N^*$ is indeed circulant, as shown by the following computation:

$$M_{ij} = \sum_k w^{ik} Q_{kk} w^{-jk} = \sum_k w^{(i-j)k} q_k$$

To be more precise, in this formula the last term depends only on $j - i$, and so shows that we have $M_{ij} = \xi_{j-i}$, with ξ being the following vector:

$$\xi_i = \sum_k w^{-ik} q_k = (F_N^* q)_i$$

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

The above result is something quite powerful, and very useful, and suggests doing everything in Fourier, when dealing with circulant matrices. And we can use here:

THEOREM 2.25. *The various basic sets of $N \times N$ circulant matrices are as follows, with the convention that associated to any $q \in \mathbb{C}^N$ is the matrix $Q = diag(q_1, \ldots, q_N)$:*

(1) *The set of all circulant matrices is:*

$$M_N(\mathbb{C})^{circ} = \left\{ F_N Q F_N^* \middle| q \in \mathbb{C}^N \right\}$$

(2) *The set of all circulant unitary matrices is:*

$$U_N^{circ} = \left\{ \frac{1}{N} F_N Q F_N^* \middle| q \in \mathbb{T}^N \right\}$$

(3) *The set of all circulant orthogonal matrices is:*

$$O_N^{circ} = \left\{ \frac{1}{N} F_N Q F_N^* \middle| q \in \mathbb{T}^N, \bar{q}_i = q_{-i}, \forall i \right\}$$

In addition, in this picture, the first row vector of $F_N Q F_N^*$ is given by $\xi = F_N^* q$.

PROOF. All this follows from Theorem 2.24, as follows:

(1) This assertion, along with the last one, is Theorem 2.24 itself.

(2) This is clear from (1), and from the fact that the rescaled matrix $F_N/\sqrt{N}$ is unitary, because the eigenvalues of a unitary matrix must be on the unit circle $\mathbb{T}$.

(3) This follows from (2), because the matrix is real when $\xi_i = \bar{\xi}_i$, and in Fourier transform, $\xi = F_N^* q$, this corresponds to the condition $\bar{q}_i = q_{-i}$. □

There are many other things that can be said about the circulant matrices, along these lines. Importantly, all our results can be generalized to the setting of the matrices which are $(N_1, \ldots, N_k)$ patterned, in a certain technical sense, and the matrix which does the job here is the corresponding generalized Fourier matrix, namely:

$$F_{N_1,\ldots,N_k} = F_{N_1} \otimes \ldots \otimes F_{N_k}$$

We will be back to this in chapter 4, when doing group theory.

2e. Exercises

Here are some exercises, in relation with what we did in this chapter, for the most regarding the further clarification of certain results, that we did quite quickly:

EXERCISE 2.26. *Clarify the symmetric function theory needed for resultants.*

EXERCISE 2.27. *Rewrite the theory of resultants, with Sylvester coming first.*

EXERCISE 2.28. *Learn further degree* 3 *formulae, and more degree* 4 *too.*

EXERCISE 2.29. *Learn some arithmetic tricks for roots of polynomials.*

EXERCISE 2.30. *Learn complex analysis tricks for roots of polynomials.*

EXERCISE 2.31. *Learn about the various types of discrete Fourier transforms.*

EXERCISE 2.32. *Apply discrete Fourier to the Sylvester resultant matrices.*

EXERCISE 2.33. *Work out the theory of tensor products of Fourier matrices.*

As bonus exercise, learn some Fourier analysis, over the abelian groups. As the saying goes, there is only one tool in mathematics, and that is the Fourier transform.

CHAPTER 3

Legendre symbol

3a. Euler, Legendre

Let us go back to what we did in chapter 1 with congruences. Our aim here will be that of further building on some of the theorems there. To be more precise, we will be interested in solving the following ubiquitous equation, over the integers:

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

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

DEFINITION 3.1. *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 3.2. *The Legendre symbol is given by the formula*

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

called Euler formula for the Legendre symbol.

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

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

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

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

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

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

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

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

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

PROPOSITION 3.3. *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 3.2. □

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

PROPOSITION 3.4. *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 3.3, 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 3.3. □

It is possible to say some further conceptual things, some sounding very fancy, in relation with Proposition 3.3 and Proposition 3.4. But remember that, according to the plan made in the beginning of this chapter, we are here for the kill, namely computing the Legendre symbol, no matter what, and with no prisoners taken.

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

THEOREM 3.5. *We have the quadratic reciprocity formula*

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

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

PROOF. This is obviously something tough, because how on Earth, you would say, the above two Legendre symbols can be related to each other. Good point, and in answer, I do not have any simple explanation to offer, at this stage of things, but:

(1) We will see a proof later in this chapter, by using some calculus with the roots of unity, and more specifically, with beasts called quadratic Gauss sums.

(2) We will see as well a proof of this in chapter 4, using algebraic methods that time, coming from field theory, relating somehow questions over $\mathbb{F}_p$ and over $\mathbb{F}_q$. □

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 3.5 with Proposition 3.3, it is quite clear that, no matter how big p is, if a has only small prime factors, we are saved.

Besides Proposition 3.3, the quadratic reciprocity formula comes accompanied by two other statements, which are very useful in practice. First, at $a=-1$, we have:

PROPOSITION 3.6. *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 3.7. *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 3.6 plus Theorem 3.7 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 3.5. We first have:

PROPOSITION 3.8. *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 3.9. *We have the following formula,*

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

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

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

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

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

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

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

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

THEOREM 3.10. *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 3.11. *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. □

As a further plot to the story, the theory of Kronecker symbols can be further generalized into a theory of Hilbert symbols. But then, you guessed it right, Hilbert was not the last one in the series, which contains some other illustrous mathematicians.

So bad these times are over. More recently, Prince made an attempt to join the series, with a very interesting symbol, which was however not accepted by the community.

3b. Gauss sums

This chapter is far from being over, and 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 so far about these symbols, there were two gaps there, namely a big gap at the quadratic reciprocity, with us being still clueless about that statement, and then a smaller gap too, at the $a = 2$ companion statement. We will fill here both these gaps.

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

THEOREM 3.12. *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:

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

(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:

$$w^p+w^{-p}=\begin{cases}w+w^{-1} & \text{if } p=\pm1(8)\\ -w-w^{-1} & \text{if } p=\pm3(8)\end{cases}$$

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. □

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

THEOREM 3.13. *We have the quadratic reciprocity formula*

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

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

PROOF. This is something 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 3.12. 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. □

3c. More 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 3.14. *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 3.15. *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 3.16. *The first few quadratic Gauss sums are given by*

q	1	2	3	4	5	6	7	8
t	1	0	$\sqrt{3}\,i$	$2(1+i)$	$\sqrt{5}$	0	$\sqrt{7}\,i$	$2\sqrt{2}(1+i)$

with everything coming from easy algebra, except for the signs.

Moving ahead now with the general case, there is some obvious periodicity in the above table, of order 4, and with everything working fine, I mean with the dependence on q being clear in all cases modulo 4, we are led to the following statement:

Theorem 3.17. *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 3.13, 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 3.15, 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 3.13:

$$t = \sum_{r=0}^{q-1} \left(\frac{r}{q}\right) w^r$$

(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 3.15, 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 3.15 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, what we have left, 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.

(10) And we will stop here, with my apologies to you, and to mathematics well done, in general. For the story, Gauss himself struggled quite a bit with this question, and there have been countless other victims, afterwards. Incuding myself, once I got into this, in my research, not realized that this is the Gauss sign, that I'm looking for, and spent a few days with it, with the conclusion that the question is guaranteed undoable. □

So, this was for the story of Gauss sums. Still a gap left, but we will have the whole Part III of this book for doing analysis, remind me at that time about that sign.

3d. Some applications

Time for some applications, now that we are well into number theory. And here, regarding applications of basic number theory, there are so many of them, to the point that even the most lugubrious mathematicians, and there are quite a few of them, have to forget about their rigor, and be subjective, and talk about, well, what they love.

So, here is my own story. With the folks in quantum physics we got increasingly interested in the Hadamard matrices, which are the matrices $H\in M_N(\pm 1)$ all those rows

are pairwise orthogonal. Here is a basic example, called first Walsh matrix:

$$W_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

This matrix is quite trivial, of size 2×2, but by taking tensor powers of it, we have as examples the higher Walsh matrices as well, having size $2^k \times 2^k$, given by:

$$W_{2^k} = W_2^{\otimes k}$$

What happens then in arbitrary size $N \times N$? It is clear that we must have $2|N$, and along the same lines, it is easy to see, by playing around with the first rows, that once your matrix has $N \geq 3$ rows, we must have $4|N$, the precise result being as follows:

PROPOSITION 3.18. *The size of an Hadamard matrix $H \in M_N(\pm 1)$ must satisfy*

$$N \in \{2\} \cup 4\mathbb{N}$$

with this coming from the orthogonality condition between the first 3 rows.

PROOF. By permuting the rows and columns or by multiplying them by -1, as to rearrange the first 3 rows, we can always assume that our matrix looks as follows:

$$H = \begin{pmatrix} \underbrace{1 \ldots\ldots 1}_{x} & \underbrace{1 \ldots\ldots 1}_{y} & \underbrace{1 \ldots\ldots 1}_{z} & \underbrace{1 \ldots\ldots 1}_{t} \\ 1 \ldots\ldots 1 & 1 \ldots\ldots 1 & -1 \ldots -1 & -1 \ldots -1 \\ 1 \ldots\ldots 1 & -1 \ldots -1 & 1 \ldots\ldots 1 & -1 \ldots -1 \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \end{pmatrix}$$

Now if we denote by x, y, z, t the sizes of the block columns, as indicated, the orthogonality conditions between the first 3 rows give the following system of equations:

$$\begin{aligned} (1 \perp 2) \quad &: \quad x + y = z + t \\ (1 \perp 3) \quad &: \quad x + z = y + t \\ (2 \perp 3) \quad &: \quad x + t = y + z \end{aligned}$$

The numbers x, y, z, t being such that the average of any two equals the average of the other two, and so equals the global average, the solution of our system is $x = y = z = t$. Thus the matrix size $N = x + y + z + t$ must be a multiple of 4, as claimed. □

The above result is something quite interesting, and the point is that a similar analysis with 4 rows or more does not give any further restriction on the possible values of the size $N \in \mathbb{N}$. In fact, we are led in this way to the following famous conjecture:

CONJECTURE 3.19 (Hadamard). *There is an Hadamard matrix of order N,*

$$H \in M_N(\pm 1)$$

for any $N \in 4\mathbb{N}$.

Normally this is an analytic question, because in practice the number of Hadamard matrices grows exponentially with N, and so in order to prove the conjecture, you just need a modest lower estimate on this number. But, no one knows how to do this, and this despite the Hadamard conjecture being open for more than 100 years.

This being said, what we can do with our number theory methods is to verify at least the Hadamard conjecture at small values of $N \in 4\mathbb{N}$. And here, with $N = 4, 8$ being solved by the Walsh matrices, we are faced with constructing a matrix at $N = 12$.

In order to solve this question, let $q = p^k$ be an odd prime power, and set:

$$\chi(a) = \begin{cases} 0 & \text{if } a = 0 \\ 1 & \text{if } a = b^2, b \neq 0 \\ -1 & \text{otherwise} \end{cases}$$

Then set $Q_{ab} = \chi(b - a)$, with indices in $\mathbb{F}_q$. With these conventions, the Paley construction of Hadamard matrices, which works well at $N = 12$, is as follows:

THEOREM 3.20. *Given an odd prime power $q = p^k$, construct $Q_{ab} = \chi(b-a)$ as above. We have then constructions of Hadamard matrices, as follows:*

(1) *Paley 1: if $q = 3(4)$ we have a matrix of size $N = q + 1$, as follows:*

$$P_N^1 = 1 + \begin{pmatrix} 0 & 1 & \ldots & 1 \\ -1 & & & \\ \vdots & & Q & \\ -1 & & & \end{pmatrix}$$

(2) *Paley 2: if $q = 1(4)$ we have a matrix of size $N = 2q + 2$, as follows:*

$$P_N^2 = \begin{pmatrix} 0 & 1 & \ldots & 1 \\ 1 & & & \\ \vdots & & Q & \\ 1 & & & \end{pmatrix} \quad : \quad 0 \to \begin{pmatrix} 1 & -1 \\ -1 & -1 \end{pmatrix} \quad , \quad \pm 1 \to \pm \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

These matrices are skew-symmetric ($H + H^t = 2$), respectively symmetric ($H = H^t$).

PROOF. In order to simplify the presentation, we denote by 1 all the identity matrices, of any size, and by $\mathbb{I}$ all the rectangular all-one matrices, of any size as well. It is elementary to check that the matrix $Q_{ab} = \chi(a - b)$ has the following properties:

$$QQ^t = q1 - \mathbb{I} \quad , \quad Q\mathbb{I} = \mathbb{I}Q = 0$$

In addition, we have the following formulae, which are elementary as well, coming from the fact that -1 is a square in $\mathbb{F}_q$ precisely when $q = 1(4)$:

$$q = 1(4) \implies Q = Q^t \quad , \quad q = 3(4) \implies Q = -Q^t$$

With these observations in hand, the proof goes as follows:

(1) With our above conventions for 1 and $\mathbb{I}$, the matrix in the statement is:

$$P_N^1 = \begin{pmatrix} 1 & \mathbb{I} \\ -\mathbb{I} & 1+Q \end{pmatrix}$$

With this formula in hand, the Hadamard matrix condition follows from:

$$\begin{aligned} P_N^1(P_N^1)^t &= \begin{pmatrix} 1 & \mathbb{I} \\ -\mathbb{I} & 1+Q \end{pmatrix}\begin{pmatrix} 1 & -\mathbb{I} \\ \mathbb{I} & 1-Q \end{pmatrix} \\ &= \begin{pmatrix} N & 0 \\ 0 & \mathbb{I}+1-Q^2 \end{pmatrix} \\ &= \begin{pmatrix} N & 0 \\ 0 & N \end{pmatrix} \end{aligned}$$

(2) If we denote by G, F the 2×2 matrices in the statement, which replace respectively the $0, 1$ entries, then we have the following formula for our matrix:

$$P_N^2 = \begin{pmatrix} 0 & \mathbb{I} \\ \mathbb{I} & Q \end{pmatrix} \otimes F + 1 \otimes G$$

With this formula in hand, the Hadamard matrix condition follows from:

$$\begin{aligned} (P_N^2)^2 &= \begin{pmatrix} 0 & \mathbb{I} \\ \mathbb{I} & Q \end{pmatrix}^2 \otimes F^2 + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes G^2 + \begin{pmatrix} 0 & \mathbb{I} \\ \mathbb{I} & Q \end{pmatrix} \otimes (FG+GF) \\ &= \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix} \otimes 2 + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes 2 + \begin{pmatrix} 0 & \mathbb{I} \\ \mathbb{I} & Q \end{pmatrix} \otimes 0 \\ &= \begin{pmatrix} N & 0 \\ 0 & N \end{pmatrix} \end{aligned}$$

Finally, the last assertion is clear, from the above formulae relating Q, Q^t. □

In practice, with Walsh and Paley, the next problem is at $N = 92$. But here, we have:

THEOREM 3.21. *Assuming that $A, B, C, D \in M_K(\pm 1)$ are circulant, symmetric, pairwise commute and satisfy the condition*

$$A^2 + B^2 + C^2 + D^2 = 4K$$

the following $4K \times 4K$ matrix is Hadamard, called of Williamson type:

$$H = \begin{pmatrix} A & B & C & D \\ -B & A & -D & C \\ -C & D & A & -B \\ -D & -C & B & A \end{pmatrix}$$

Moreover, matrices A, B, C, D as above exist at $K = 23$, where $4K = 92$.

PROOF. Consider the quaternion units $1, i, j, k \in M_4(0, 1)$, that we met in chapter 1, when discussing $\mathbb{R}^4$. These matrices describe the positions of the A, B, C, D entries in the matrix H from the statement, and so this matrix can be written as follows:

$$H = A \otimes 1 + B \otimes i + C \otimes j + D \otimes k$$

Assuming now that A, B, C, D are symmetric, we have:

$$\begin{aligned} HH^t &= (A \otimes 1 + B \otimes i + C \otimes j + D \otimes k) \\ &\quad (A \otimes 1 - B \otimes i - C \otimes j - D \otimes k) \\ &= (A^2 + B^2 + C^2 + D^2) \otimes 1 - ([A, B] - [C, D]) \otimes i \\ &\quad -([A, C] - [B, D]) \otimes j - ([A, D] - [B, C]) \otimes k \end{aligned}$$

Now assume that our matrices A, B, C, D pairwise commute, and satisfy the condition in the statement. In this case, it follows from the above formula that we have:

$$HH^t = 4K$$

Thus, we obtain indeed an Hadamard matrix, as claimed. However, finding such matrices is in general a difficult task, and this is where Williamson's extra assumption in the statement, that A, B, C, D should be taken circulant, comes from. Finally, regarding the $K = 23$ and $N = 92$ example, this comes via a computer search. □

At higher N things become more technical, and more complicated constructions, along the lines of those of Paley and Williamson, are needed. Quite curiously, as of now, early 21th century, the human knowledge stops at the number of the beast, namely:

$$\mathfrak{N} = 666$$

That is, explicit examples of Hadamard matrices have been constructed for all multiples of four $N \leq 664$, but no such matrix is known so far at $N = 668$.

But hey, the story is not over here. We will not let the Devil win, and as a further twist to the plot, bringing some sort of solution to this, we have:

THEOREM 3.22. *When enlarging the attention to the complex Hadamard matrices, $H \in M_N(\mathbb{T})$ having the rows pairwise orthogonal, the Fourier matrix,*

$$F_N = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & w & w^2 & \ldots & w^{N-1} \\ 1 & w^2 & w^4 & \ldots & w^{2(N-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w^{N-1} & w^{2(N-1)} & \ldots & w^{(N-1)^2} \end{pmatrix}$$

with $w = e^{2\pi i/N}$, provides an example of such a matrix, at any $N \in \mathbb{N}$. Thus, the Hadamard Conjecture problematics dissapears, in the complex setting.

PROOF. We have seen in chapter 2 that the rescaling $U = F_N/\sqrt{N}$ is unitary. Thus the rows of U are pairwise orthogonal, and so follow to be the rows of F_N. □

In view of the above result, and of some interesting quantum physics questions too, that we will not get into here, let us study more in detail the complex Hadamard matrices. Many examples can be constructed, quite often by using the combinatorics of roots of unity, and as a basic example here, we have the tensor products of Fourier matrices:

$$F_{N_1,\ldots,N_k} = F_{N_1} \otimes \ldots \otimes F_{N_k}$$

Of course, not all examples of complex Hadamard matrices come from roots of unity, as shown by the following quite exotic looking result, due to Björck and Fröberg:

PROPOSITION 3.23. *The following is a complex Hadamard matrix,*

$$BF_6 = \begin{pmatrix} 1 & ia & -a & -i & -\bar{a} & i\bar{a} \\ i\bar{a} & 1 & ia & -a & -i & -\bar{a} \\ -\bar{a} & i\bar{a} & 1 & ia & -a & -i \\ -i & -\bar{a} & i\bar{a} & 1 & ia & -a \\ -a & -i & -\bar{a} & i\bar{a} & 1 & ia \\ ia & -a & -i & -\bar{a} & i\bar{a} & 1 \end{pmatrix}$$

where $a \in \mathbb{T}$ is one of the roots of $a^2 + (\sqrt{3} - 1)a + 1 = 0$.

PROOF. The matrix in the statement is circulant, in the sense that the rows appear by cyclically permuting the first row. Thus, we only have to check that the first row is orthogonal to the other 5 rows. But this follows from $a^2 + (\sqrt{3} - 1)a + 1 = 0$. □

Leaving aside such monsters, we have as well deformations, as shown by:

THEOREM 3.24. *The only Hadamard matrices at $N = 4$ are, up to equivalence*

$$F_4^q = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & q & -1 & -q \\ 1 & -q & -1 & q \end{pmatrix}$$

with $q \in \mathbb{T}$, which appear as suitable deformations of $W_4 = F_2 \otimes F_2$.

PROOF. This is something quite self-explanatory, and we will leave working out all this, namely finding the correct meaning of the equivalence relation, and of the deformation notion used, along of course with the proof, as an instructive exercise. □

In view of the above examples and counterexamples, and thinking a bit at number theory, where primes are the kings, we are led to the question of finding the complex Hadamard matrices which are "isolated", in a geometric sense. And here, skipping some details, we have an interesting construction due to McNulty and Weigert, that we would like to explain now. This construction is based on the following simple fact:

THEOREM 3.25. *Assuming that $K \in M_N(\mathbb{C})$ is Hadamard, so is the matrix*

$$H_{ia,jb} = \frac{1}{\sqrt{Q}} K_{ij}(L_i^* R_j)_{ab}$$

provided that $\{L_1, \ldots, L_N\} \subset \sqrt{Q}U_Q$ and $\{R_1, \ldots, R_N\} \subset \sqrt{Q}U_Q$ are such that

$$\frac{1}{\sqrt{Q}} L_i^* R_j \in \sqrt{Q}U_Q$$

with $i, j = 1, \ldots, N$, are complex Hadamard.

PROOF. The check of the unitarity of the matrix in the statement can be done as follows, by using our various assumptions on the various matrices involved:

$$\begin{aligned} < H_{ia}, H_{kc} > &= \frac{1}{Q} \sum_{jb} K_{ij}(L_i^* R_j)_{ab} \bar{K}_{kj} \overline{(L_k^* R_j)}_{cb} \\ &= \sum_j K_{ij}\bar{K}_{kj}(L_i^* L_k)_{ac} \\ &= N\delta_{ik}(L_i^* L_k)_{ac} \\ &= NQ\delta_{ik}\delta_{ac} \end{aligned}$$

The entries of our matrix being in addition on the unit circle, we are done. □

The above construction is of course something quite abstract, but as a very concrete input for it, we can use the following well-known Fourier analysis construction:

PROPOSITION 3.26. *For $q \geq 3$ prime, the matrices $\{F_q, DF_q, \ldots, D^{q-1}F_q\}$, where F_q is the Fourier matrix, and where*

$$D = diag\left(1, 1, w, w^3, w^6, w^{10}, \ldots, w^{\frac{q^2-1}{8}}, \ldots, w^{10}, w^6, w^3, w\right)$$

with $w = e^{2\pi i/q}$, are such that $\frac{1}{\sqrt{q}} E_i^* E_j$ is complex Hadamard, for any $i \neq j$.

PROOF. With by definition $0, 1, \ldots, q-1$ as indices for our matrices, as usual in a Fourier analysis context, the formula of the above matrix D is:

$$D_c = w^{0+1+\ldots+(c-1)} = w^{\frac{c(c-1)}{2}}$$

Since we have $\frac{1}{\sqrt{q}} E_i^* E_j \in \sqrt{q}U_q$, we just need to check that these matrices have entries belonging to $\mathbb{T}$, for any $i \neq j$. With $k = j - i$, these entries are given by:

$$\frac{1}{\sqrt{q}}(E_i^* E_j)_{ab} = \frac{1}{\sqrt{q}}(F_q^* D^k F_q)_{ab} = \frac{1}{\sqrt{q}} \sum_c w^{c(b-a)} D_c^k$$

Now observe that with $s=b-a$, we have the following formula:

$$\begin{aligned}\left|\sum_c w^{cs}D_c^k\right|^2 &= \sum_{cd} w^{cs-ds}w^{\frac{c(c-1)}{2}\cdot k-\frac{d(d-1)}{2}\cdot k}\\ &= \sum_{cd} w^{(c-d)\left(\frac{c+d-1}{2}\cdot k+s\right)}\\ &= \sum_{de} w^{e\left(\frac{2d+e-1}{2}\cdot k+s\right)}\\ &= \sum_e\left(w^{\frac{e(e-1)}{2}\cdot k+es}\sum_d w^{edk}\right)\\ &= \sum_e w^{\frac{e(e-1)}{2}\cdot k+es}\cdot q\delta_{e0}\\ &= q\end{aligned}$$

Thus the entries are on the unit circle, and we are done. □

Next, we have the following result, making use of Gauss sums:

PROPOSITION 3.27. *The matrices $G_k=\frac{1}{\sqrt{q}}F_q^*D^kF_q$, with $D=diag\left(w^{\frac{c(c-1)}{2}}\right)$, and with $k\neq 0$ are circulant, their first row vectors V^k being given by*

$$V_i^k=\delta_q\left(\frac{k/2}{q}\right)w^{\frac{q^2-1}{8}\cdot k}\cdot w^{-\frac{i}{k}\left(\frac{i}{k}-1\right)}$$

where $\delta_q=1$ if $q=1(4)$ and $\delta_q=i$ if $q=3(4)$, and with all inverses being taken in $\mathbb{Z}_q$.

PROOF. The above matrices G_k are indeed circulant, their first vectors being:

$$V_i^k=\frac{1}{\sqrt{q}}\sum_c w^{\frac{c(c-1)}{2}\cdot k+ic}$$

But this is a Gauss sum, and by computing the square, we obtain:

$$(V_i^k)^2=\delta_q^2\cdot w^{\frac{q^2-1}{4}\cdot k}\cdot w^{-\frac{i}{k}\left(\frac{i}{k}-1\right)}$$

By extracting now the square root, we obtain a formula as follows:

$$V_i^k=\pm\delta_q\cdot w^{\frac{q^2-1}{8}\cdot k}\cdot w^{-\frac{i}{k}\left(\frac{i}{k}-1\right)}$$

And with Gauss computing for us the sign, this leads to the above formula. □

Let us combine now all the above results. We obtain the following statement:

THEOREM 3.28. *Let $q \geq 3$ be prime, consider subsets $S, T \subset \{0, 1, \ldots, q-1\}$ satisfying the conditions $|S| = |T|$ and $S \cap T = \emptyset$, and write:*

$$S = \{s_1, \ldots, s_N\} \quad , \quad T = \{t_1, \ldots, t_N\}$$

Then, with the matrix V being as above, the following matrix,

$$H_{ia,jb} = K_{ij} V_{b-a}^{t_j - s_i}$$

is complex Hadamard, provided that the matrix $K \in M_N(\mathbb{C})$ is complex Hadamard.

PROOF. This follows indeed by using the general construction in Theorem 3.25, with input coming from Proposition 3.26 and Proposition 3.27. □

The above construction covers many interesting examples of Hadamard matrices, known to be isolated, such as the Tao matrix, which is as follows, with $w = e^{2\pi i/3}$:

$$T_6 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & w & w & w^2 & w^2 \\ 1 & w & 1 & w^2 & w^2 & w \\ 1 & w & w^2 & 1 & w & w^2 \\ 1 & w^2 & w^2 & w & 1 & w \\ 1 & w^2 & w & w^2 & w & 1 \end{pmatrix}$$

For more on Hadamard matrices, real and complex, there are many texts available, including Horadam [48], Seberry and Yamada [78], or my book [8]. Also, for design theory, which is the math discipline behind all this, appearing as a ramification of number theory, you have de Launey and Flannery [22], Ryser [76] and Stinson [86].

3e. Exercises

Many interesting things going on in this chapter, of both algebraic and analytic nature, and as exercises on all this, exciting too, we hope, we have:

EXERCISE 3.29. *Work out the basic consequences of Euler, with and without Fermat.*

EXERCISE 3.30. *Learn more about field characters, and their properties.*

EXERCISE 3.31. *Work out the Legendre $a = 3, 5$ results, and do $a = 7, 11, 13$ too.*

EXERCISE 3.32. *Clarify the Jacobi and Kronecker symbols, why not Hilbert too.*

EXERCISE 3.33. *Learn about generalized Gauss sums, and their properties.*

EXERCISE 3.34. *Compute that missing sign, for the Gauss sums.*

EXERCISE 3.35. *Try finding the Williamson $N = 92$ matrix by hand.*

EXERCISE 3.36. *Based on the above, find the formula of the Tao matrix entries.*

As bonus exercise, learn some design theory and Hadamard matrices. There are countless interesting questions there, quite often of number theory flavor.

CHAPTER 4

Algebra tools

4a. Abstract algebra

We have seen some interesting mathematics in this book so far, which was of quite different nature, but having however as an important common point the various types of polynomials, $P \in \mathbb{Z}[X]$, or $P \in \mathbb{Q}[X]$, or $P \in \mathbb{R}[X]$, or $P \in \mathbb{C}[X]$, and their roots. In this chapter, following Galois, we will investigate more in depth the question of finding the roots of polynomials, by using this time tools from abstract algebra.

The idea of Galois is very simple. Remember the polynomial $P = x^2 - 2$, regarded as $P \in \mathbb{Q}[X]$, which was at the beginning of everything, motivating the introduction of $\sqrt{2}$, and of $\mathbb{R}$ itself? When thinking at this, from an algebraic perspective, looking at P is more or less the same thing as looking at the field $\mathbb{Q}[\sqrt{2}]$, where P factorizes. Thus, we can see a relation between our question, regarding polynomials and their roots, and the seemingly unrelated question of understanding the intermediate fields, as follows:

$$\mathbb{Q} \subset F \subset \mathbb{C}$$

And the point is that, as strange as this might seem, this latter question, which looks at the first glance a bit abstract, and scary, is in fact simpler than the original one, with abstract algebra providing some serious tools for dealing with it, and going well beyond what we know about polynomials and their roots, from what we learned so far.

So long for Galois and his idea, and more on this later. By the way, let me mention too that Galois found all this when he was 18, making the whole story even more amazing. So, up to you to decide now if you want to continue reading this book, or start right away doing some intense research, as Galois was doing, when he was younger than you.

Getting started now, we first need to learn some abstract algebra, and we have:

QUESTION 4.1. *What are the various beasts and deities in abstract algebra, namely groups, rings, modules, ideals, fields, vector spaces and algebras?*

Quite interestingly, the fields that we already know about come quite lately in the list, and the vector spaces, that we know well about, since ages, come second to last. Well, time now to forget all this, which is analytic, or geometric, or even number theoretical philosophy, and learn things how abstract algebraists do, in the above precise order.

For purely pedagogical purposes, we will need someone to take credit for inventing all this, and God being taken since chapter 1, with the invention of $\mathbb{N}$, we will blame the Devil for inventing abstract algebra. First, the Devil created the groups:

DEFINITION 4.2. *A group is a set G with a multiplication operation $(g, h) \to gh$, which must satisfy the following conditions:*

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

When the multiplication is commutative, $gh = hg$, we say that G is abelian.

Let us first look at the abelian groups. Here as basic examples we have $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ with the addition operation $+$. However, we have as well as $\mathbb{Q}, \mathbb{R}, \mathbb{C}$, and the unit circle $\mathbb{T}$, with the multiplication operation $\times$. In relation with this, observe that we made a choice in Definition 4.2, namely that of privileging the multiplicative notations $gh, 1, g^{-1}$ over of the additive ones $g + h, 0, -g$. More on this choice in a moment.

Still speaking abelian groups, let us look into the finite group case, $|G| < \infty$. Here as basic examples we have the cyclic groups, constructed as follows:

PROPOSITION 4.3. *The following constructions produce the same group, denoted $\mathbb{Z}_N$, which is finite and abelian, and is called cyclic group of order N:*

(1) *$\mathbb{Z}_N$ is the set of remainders modulo N, with operation $+$.*
(2) *$\mathbb{Z}_N \subset \mathbb{T}$ is the group of N-th roots of unity, with operation $\times$.*

PROOF. Here the equivalence between (1) and (2) is obvious. More complicated, however, is the question of finding the good philosophy and notation for this group. In what concerns us, we will be rather geometers, as usual, and in what regards the philosophy, we will often prefer the interpretation (2). As for the notation, here we will be physicists, or engineers, also as usual, and we will use $\mathbb{Z}_N$, which is very natural. □

As a basic thing to be known, about the abelian groups, still in the finite case, we can construct further examples of such groups by making products between various cyclic groups $\mathbb{Z}_N$. Quite remarkably, we obtain in this way all the finite abelian groups:

THEOREM 4.4. *The finite abelian groups are precisely the products of cyclic groups:*

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

Moreover, there are technical extensions of this result, going beyond the finite case.

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

(1) In order to prove our result, assume that G is finite and abelian. For any prime number $p \in \mathbb{N}$, let us define $G_p \subset G$ to be the subset of elements having as order a power

of p. Equivalently, this subset $G_p \subset G$ can be defined as follows:

$$G_p = \left\{ g \in G \middle| \exists k \in \mathbb{N}, g^{p^k} = 1 \right\}$$

(2) It is then routine to check, based on definitions, that each G_p is a subgroup. Our claim now is that we have a direct product decomposition as follows:

$$G = \prod_p G_p$$

(3) Indeed, by using the fact that our group G is abelian, we have a morphism as follows, with the order of the factors when computing $\prod_p g_p$ being irrelevant:

$$\prod_p G_p \to G \quad , \quad (g_p) \to \prod_p g_p$$

Moreover, it is routine to check that this morphism is both injective and surjective, via some simple manipulations, so we have our group decomposition, as in (2).

(4) Thus, we are left with proving that each component G_p decomposes as a product of cyclic groups, having as orders powers of p, as follows:

$$G_p = \mathbb{Z}_{p^{r_1}} \times \ldots \times \mathbb{Z}_{p^{r_s}}$$

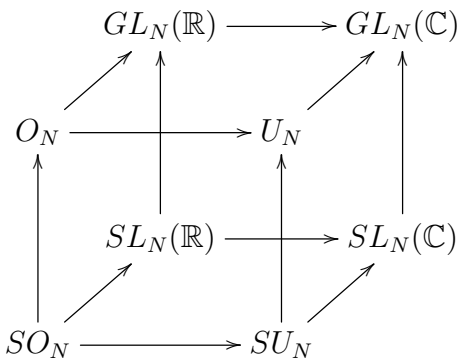
But this is something that can be checked by recurrence on $|G_p|$, via some routine computations, and we are led to the conclusion in the statement.

(5) Finally, for full details on all this, and for some technical extensions to the infinite groups as well, we recommend a solid algebra book, such as Lang [63]. □

Moving forward now, let us look as well into the general, non-abelian case. The first thought goes here to the $N \times N$ matrices with their multiplication, but these do not form a group, because we must assume $\det A \neq 0$ in order for our matrix to be invertible.

So, let us call $GL_N(\mathbb{C})$ the group formed by these latter matrices, with nonzero determinant, with GL standing here for "general linear". By further imposing the condition $\det A = 1$ we obtain a subgroup $SL_N(\mathbb{C})$, with SL standing for "special linear", and then we can talk as well about the real versions of these groups, and also intersect everything with the group of unitary matrices U_N. We obtain in this way 8 groups, as follows:

THEOREM 4.5. *We have groups of invertible matrices as follows,*



with S standing here for "special", meaning having determinant 1.

PROOF. This is clear indeed from the above discussion. As a comment, we can talk in fact about $GL_N(F)$ and $SL_N(F)$, once we have a ground field F, but in what regards the corresponding orthogonal and unitary groups, things here are more complicated. We will certainly have an exercise about this, at the end of this chapter. □

There are many other groups of matrices, besides the above ones, as for instance the symplectic groups $Sp_N \subset U_N$, appearing at $N \in 2\mathbb{N}$. Generally speaking, the theory of Lie groups and algebras is in charge with the classification of such beasts.

Finally, a word about the finite non-abelian groups. As basic example here you have the symmetric group S_N, and its various subgroups. Let us record here:

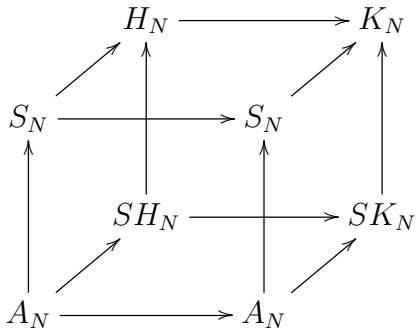
PROPOSITION 4.6. *We have finite non-abelian groups, as follows:*

(1) *S_N, the group of permutations of $\{1, \ldots, N\}$.*
(2) *$A_N \subset S_N$, the permutations having signature 1.*
(3) *$D_N \subset S_N$, the group of symmetries of the regular N-gon.*

PROOF. The fact that we have indeed groups is clear from definitions, and the non-abelianity of these groups is clear as well, provided of course that in each case N is chosen big enough, and with exercise for you to work out all this, with full details. □

For constructing further examples of finite non-abelian groups, the best is to "look up", by regarding S_N as being the permutation group of the N coordinate axes of $\mathbb{R}^N$. Indeed, this suggests looking at the symmetry groups of all sorts of geometric beasts inside $\mathbb{R}^N$, or even $\mathbb{C}^N$, and we end with a whole menagery of groups, as follows:

THEOREM 4.7. *We have groups of unitary matrices as follows,*



for the most finite, and non-abelian, called complex reflection groups.

PROOF. The above statement is of course something informal, and here are explanations on all this, including definitions for all the groups involved:

(1) To start with, S_N is the symmetric group S_N that we know, but regarded now as permutation group of the N coordinate axes of $\mathbb{R}^N$, and so as subgroup $S_N \subset O_N$.

(2) Similarly, A_N is the alternating group A_N that we know, but coming now geometrically, as $A_N = S_N \cap SO_N$, with the intersection being computed inside O_N.

(3) Regarding $H_N \subset O_N$, this is a famous group, called hyperoctahedral group, appearing as the symmetry group of the hypercube $\square_N \subset \mathbb{R}^N$.

(4) Regarding $K_N \subset U_N$, this is the complex analogue of H_N, consisting of the unitary matrices $U \in U_N$ having exactly one nonzero entry, on each row and each column.

(5) We have as well on our diagram the groups SH_N, SK_N, with S standing as usual for "special", that is, consisting of the matrices in H_N, K_N having determinant 1.

(6) In what regards now the diagram itself, sure I can see that S_N, A_N appear twice, but nothing can be done here, after thinking a bit, at how the diagram works.

(7) Let us mention too that the groups $\mathbb{Z}_N, D_N$ have their place here, in N-dimensional geometry, but not exactly on our diagram, as being the symmetry groups of the oriented cycle, and unoriented cycle, with vertices at the simplex $X_N = \{e_i\} \subset \mathbb{R}^N$.

(8) Finally, in what regards finiteness, non-abelianity, and also the name "complex reflection groups", many things to be checked here, left to you as an exercise. □

Very nice all this. Let us summarize this group theory discussion as follows:

CONCLUSION 4.8. *All groups, or almost, are best seen as being groups of matrices. And even as groups of unitary matrices, in most cases.*

Observe that this justifies our choice in Definition 4.2, for the group operation to be denoted multiplicatively, $\times$. Indeed, in most cases, that is a matrix multiplication.

As a second conclusion, not bad all this, for an invention of the Devil. The continuation of the story involves a second batch of objects invented by the Devil, namely:

DEFINITION 4.9. *We have notions of rings, modules and ideals, as follows:*

(1) *A ring R is a set with operations $+$ and $\times$, satisfying the usual conditions for such operations, except for $ab = ba$, and for $a \neq 0 \implies \exists a^{-1}$.*
(2) *A module V over a ring R is a vector space, but we will call it ring, and keep the name vector spaces for the modules over fields, $R = F$.*
(3) *An ideal $I \subset R$ is a subgroup with the left ideal property $i \in I, r \in R \implies ir \in I$, or the right ideal property $i \in I, r \in R \implies ri \in I$, or both.*

This was a quite crowded statement, but you get the point, with (1) and (2) we are sort of trying to do field and vector space mathematics, over things which are not necessarily fields and vector spaces over them, and (3) is something technical, non-field specific. At the level of examples, these abound, and we have two important ones, as follows:

(1) The integers form a ring, $R = \mathbb{Z}$, which in addition is commutative, $ab = ba$. As obvious module over $\mathbb{Z}$, we have the lattice $V = \mathbb{Z}^N$. Finally, since $R = \mathbb{Z}$ is commutative, the 3 notions of ideals coincide, and these are the subsets $I = a\mathbb{Z}$, with $a \in \mathbb{Z}$.

(2) The matrices over the integers form a ring, $R = M_N(\mathbb{Z})$, which is noncommutative at $N \geq 1$. As obvious module over $M_N(\mathbb{Z})$, we have the lattice $V = \mathbb{Z}^N$. As for the ideals, things here are a bit more complicated, but since at $N = 2$ the matrices of type $\left(\begin{smallmatrix} a & b \\ 0 & 0 \end{smallmatrix}\right)$ form a left ideal which is not a right ideal, and the matrices of type $\left(\begin{smallmatrix} a & 0 \\ b & 0 \end{smallmatrix}\right)$ form a right ideal which is not a left ideal, at least we know that our 3 types of ideals make sense.

The question that you surely have in mind is, what are ideals good for? Answer:

PROPOSITION 4.10. *For a subgroup $I \subset R$, the following are equivalent:*

(1) *I is a two-sided ideal.*
(2) *R/I is a ring.*

PROOF. This is something which requires some thinking, as follows:

(1) Since the additive group $(R, +)$ is abelian, given an additive subgroup $I \subset R$ we can form the quotient group R/I, which is abelian too, with addition as follows:

$$(a + I) + (b + I) = (a + b + I)$$

Observe that the unit is $(0 + I) = I$, and that inverses are given by $(-a + I)$.

(2) The question is now, can we turn this abelian group R/I into a ring? Normally the multiplication can only be as follows, and with this clarifying our statement, with the condition "R/I is a ring" there meaning, with respect to this precise multiplication:

$$(a+I)(b+I)=(ab+I)$$

(3) But, will this work. As a first observation, there is a bit of analogy here with group theory, where $H\subset G$ must be normal in order for G/H to be a group. Thus, our claim is that the ideal condition is somehow the "analogue of normality, in the ring setting".

(4) In practice now, it is quite clear, exactly as in the group theory setting, that everything will be fine, provided that our multiplication is well-defined. And for this multiplication to be well-defined, the following condition must be satisfied:

$$(a+I)=(a'+I)\ ,\ (b+I)=(b'+I)\quad\implies\quad(ab+I)=(a'b'+I)$$

But this amounts in the following condition to be satisfied:

$$a-a'\in I\ ,\ b-b'\in I\quad\implies\quad ab-a'b'\in I$$

(5) Now comes the math. We have the following identity, which shows that if $I\subset R$ is a two-sided ideal, then the above condition is satisfied, and so done:

$$ab-a'b'=a(b-b')+(a-a')b'$$

(6) Conversely now, if the condition in (4) is satisfied, we have in particular:

$$i-0\in I\ ,\ r-r\in I\quad\implies\quad ir-0r\in I$$

$$r-r\in I\ ,\ i-0\in I\quad\implies\quad ri-r0\in I$$

Thus $I\subset R$ must be a two-sided ideal, and this finishes the proof. □

As a conclusion to this, we have now some intuition on the ideals, whose name is quite unfortunate, and whose definition is quite opaque. With a remaining problem, however, regarding the precise significance of the left or right ideals, which are not two-sided ideals. To which I would answer, for simplifying, let us forget about those.

Many things can be said about rings, modules and ideals, especially in the commutative case, and we will back to this with some results in chapter 5 below, when doing algebraic geometry. For formulating however a theorem on the subject, we have:

THEOREM 4.11. *Assuming that R is commutative and $I\subset R$ is a maximal ideal, in the sense that it is a proper ideal, $I\neq R$, and there is no bigger proper ideal*

$$I\subset J\subset R$$

the quotient ring $F=R/I$ is a field.

PROOF. This is actually cool stuff, and we will see a geometric illustration in a moment, and many more illustrations in chapter 5, when doing algebraic geometry. So, here is the proof, and with this being guaranteed to be useful learning:

(1) Before starting, a quick example. We know that over $R = \mathbb{Z}$, the ideals are the subsets $I = p\mathbb{Z}$ with $p \in \mathbb{N}$. But such an ideal is maximal precisely when p is prime, and this is the same as asking for the quotient ring $R/I = \mathbb{Z}_p$ to be a field.

(2) In general now, assume first that R/I is a field. This means that any nonzero element of R/I is invertible, and with our usual conventions for R/I, this reads:

$$\forall a \notin I \ , \ \exists b \in R \ , \ (ab + I) = (1 + I)$$

Now assume by contradiction that $I \subset R$ is not maximal, so that we have a bigger ideal $I \subset J \subset R$. If we pick $a \in J - I$, we obtain, by the above, the following:

$$a \in J - I \ , \ b \in R \ , \ ab = 1 + i \ , \ i \in I$$

But this is contradictory, because since J is an ideal, containing I, we must have $ab, i \in J$, so we conclude that we have $1 \in J$, and so $J = R$, contradiction.

(3) Conversely, assume now that I is maximal, and assume too, by contradiction, that R/I is not a field. Then we can find a zero divisor in R/I, which reads:

$$(a + I)(b + I) = (I) \ , \ a, b \notin I$$

In other words, we can find $ab \in I$ with $a, b \notin I$. But then, let us look at:

$$I \subset I + aR \subset R$$

(4) What we have in the middle is an ideal, and it is also clear, from $a \notin I$, that the inclusion on the left is proper. As for the inclusion on the right, our claim is that this is proper too. Indeed, assuming otherwise, we would have a formula as follows:

$$i + ac = 1 \ , \ i \in I$$

Now by multiplying everything by b, we obtain from this:

$$ib + acb = b \ , \ i \in I$$

But this is contradictory, because on the left we have $ib \in I$ and $acb = (ab)c \in I$, which gives $b \in I$, contradicting the condition $b \notin I$. Thus, our claim is proved.

(5) But this is the end of the story, because what we just proved is that what we have in (3) is indeed a proper ideal, contradicting the maximality of I, as desired. □

Still with me I hope, and by reiterating my claim that Theorem 4.11 is really cool stuff, and we will see an illustration in a moment, and many more later.

As already mentioned, more on all this, rings, modules and ideals, in chapter 5 below, when doing algebraic geometry. Going ahead now with our general abstract algebra program, as a third and last batch of objects invented by the Devil, we have:

DEFINITION 4.12. *We have notions of fields, vector spaces and algebras, as follows:*

(1) *A field F is a field F as we know them, with in algebra parlance these being the commutative rings R with each nonzero element being invertible.*
(2) *A vector space V over a field F is a vector space as we know them, in algebra parlance these being the modules V over a field F.*
(3) *An algebra A over a field F is a vector space over F, with a ring multiplication operation $\times$, compatible with the vector space structure.*

As previously mentioned, we already know of course about fields, and in what regards the vector spaces, we know about them since ever, and finally, regarding algebras, we know many algebras of functions from analysis. But, thinking well, from a purely algebraic perspective, all these objects have many operations, and this is why they come at last.

As basic examples now, passed the fields F and the vector spaces V that we know well, we are left with finding interesting examples of algebras A. And here the examples abound, with this being actually easy to believe, due to the name "algebras" that algebraists chose for these beasts, and among them, we have two main examples, as follows:

(1) The algebra of polynomials $A = F[X]$. This is a very nice and important algebra, with the extra feature that it is commutative, $PQ = QP$.

(2) The algebra of matrices $A = M_N(F)$. Again this is a very basic example, that we know well, which this time is not commutative, $PQ \neq QP$.

As an illustration for all this, providing us with a third basic class of algebras, and bringing some light too on Theorem 4.11, we have the following basic result:

THEOREM 4.13. *Given a compact space X, the following happen:*

(1) *The continuous functions $f : X \to \mathbb{C}$ form a complex algebra $C(X)$.*
(2) *Given $x \in X$, the functions satisfying $f(x) = 0$, form an ideal $I \subset C(X)$.*
(3) *This ideal is maximal, and any maximal ideal $I \subset C(X)$ appears in this way.*
(4) *In this picture, the fact that the quotient is a field, $C(X)/I = \mathbb{C}$, is clear.*

PROOF. All this is self-explanatory, the idea being as follows:

(1) This is clear. Observe that our algebra is commutative, $fg = gf$.

(2) This is again clear, because $f(x) = 0$ implies $(fg)(x) = 0$.

(3) This follows from basic topology, via a suitable open cover for X.

(4) This is clear, because $C(X) \to C(X)/I$ maps $f \to f(x) \in \mathbb{C}$. □

And good news, that is all. Done with abstract algebra, and good learning that was, that we can use later when needed, and we can now turn to more concrete things.

4b. Galois theory

We are now ready for Galois theory. Let us start with a reminder of what we know about fields, from the previous chapters. Barring all sorts of interesting examples of fields, and barring as well some announcements, coming without proof, plus various things about the quadratic fields, coming from the theory of the Legendre symbol, which are quite technical, our knowledge so far about fields can be summarized as follows:

THEOREM 4.14. *Given a field F, define its characteristic $p = char(F)$ as being the smallest $p \in \mathbb{N}$ such that the following happens, and as $p = 0$, if this never happens:*

$$\underbrace{1 + \ldots + 1}_{p \ times} = 0$$

Then, assuming $p > 0$, this number p must be prime, we have a field embedding $\mathbb{F}_p \subset F$, and $q = |F|$ must be of the form $q = p^k$, with $k \in \mathbb{N}$. Also, we have the formulae

$$(a + b)^p = a^p + b^p \quad , \quad a^q = a$$

valid for any $a, b \in F$, and the Fermat poynomial $X^q - X$ factorizes as:

$$X^q - X = \prod_{a \in F} (X - a)$$

Also, regardless of p, any finite multiplicative subgroup $G \subset F - \{0\}$ must be cyclic.

PROOF. This is a very crowded statement, that we basically know from chapter 1, the idea being that all this comes from some elementary arithmetic, as follows:

(1) The various assertions in the beginning, regarding the characteristic $p = char(F)$ and its basic properties, all follow from definitions and from some quick thinking, based on formulae of type $(1 + \ldots + 1)(1 + \ldots + 1) = 1 + \ldots + 1$, inside the field F.

(2) We can also see that the sums $1 + \ldots + 1$, and their quotients, form a minimal subfield $E \subset F$, called prime field. At $p = 0$ we have $E = \mathbb{Q}$. At $p > 0$ we have $E = \mathbb{F}_p$, and $q = |F|$ is given by $q = p^k$, with $k \in \mathbb{N}$ being the dimension of F over E.

(3) The baby Fermat formula $(a + b)^p = a^p + b^p$, which reminds the Fermat little theorem, $a^p = a(p)$ over $\mathbb{Z}$, follows in the same way, namely from the binomial formula, because all the non-trivial binomial coefficients $\binom{p}{s}$ are multiples of p.

(4) As for the Fermat formula $a^q = a$ itself, which implies the assertion about $X^q - X$, this follows from the last assertion, which can be proved via some basic arithmetic inside F, and which for $G = F - \{0\}$ itself, with $|F| = q$, gives $a^{q-1} = 1$, for any $a \neq 0$. □

The above result raises a lot of questions. First, we have the question of understanding the finite fields, $|F| = q < \infty$, with $q = p^k$. Then, in connection with algebraic equations and field extensions of $\mathbb{Q}$ via roots of polynomials $P \in \mathbb{Q}[X]$, we must understand what

fails in degree $N = 5$ and higher. And finally, in the quadratic case, we are still in need of better understanding the quadratic reciprocity formula, say via field theory.

We will answer here all these questions. Since most of them seem to have something to do with field extensions, let us start by discussing this. We first have:

THEOREM 4.15. *Given a field extension $E \subset F$, we can talk about its Galois group G, as the group of automorphisms of F fixing E. The intermediate fields*

$$E \subset K \subset F$$

are then in correspondence with the subgroups $H \subset G$, with such a field K corresponding to the subgroup H consisting of automorphisms $g \in G$ fixing K.

PROOF. This is something self-explanatory, and follows indeed from some algebra, under suitable assumptions, in order for that algebra to properly apply. □

Getting now towards polynomials and their roots, we have here:

THEOREM 4.16. *Given a field F and a polynomial $P \in F[X]$, we can talk about the abstract splitting field of P, where this polynomial decomposes as:*

$$P(X) = c\prod_i (X - a_i)$$

In particular, any field F has a certain algebraic closure $\bar{F}$, where all the polynomials $P \in F[X]$, and in fact all polynomials $P \in \bar{F}[X]$ too, have roots.

PROOF. This is again something self-explanatory, which follows from Theorem 4.15 and from some extra algebra, under suitable assumptions, in order for that extra algebra to properly apply. Regarding the construction at the end, as main example here we have $\bar{\mathbb{R}} = \mathbb{C}$. However, as an interesting fact, $\bar{\mathbb{Q}} \subset \mathbb{C}$ is a proper subfield. □

Good news, with this in hand, we can now elucidate the structure of finite fields:

THEOREM 4.17. *For any prime power $q = p^k$ there is a unique field $\mathbb{F}_q$ having q elements. At $k = 1$ this is the usual $\mathbb{F}_p$. In general, this is the splitting field of:*

$$P = X^q - X$$

Moreover, we can construct an explicit model for $\mathbb{F}_q$, at $q = p^2$ or higher, as

$$\mathbb{F}_q = \mathbb{F}_p[X]/(Q)$$

with $Q \in \mathbb{F}_p[X]$ being a suitable irreducible polynomial, of degree k.

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

(1) The first assertion, regarding the existence and uniqueness of $\mathbb{F}_q$, follows from Theorem 4.14 and Theorem 4.16. Indeed, we know from Theorem 4.14 that given a finite field, $|F| = q$ with $k \in \mathbb{N}$, the Fermat polynomial $P = X^q - X$ factorizes as follows:

$$X^q - X = \prod_{a \in F} (X - a)$$

But this shows, via the general theory from Theorem 4.16, that our field F must be the splitting field of P, and so is unique. As for the existence, this follows again from Theorem 4.16, telling us that the splitting field always exists.

(2) In what regards now the modeling of $\mathbb{F}_q$, at $q = p$ there is nothing to do, because we have our usual $\mathbb{F}_p$ here. At $q = p^2$ and higher, we can use Theorem 4.11, which tells us that we have an isomorphism as follows, whenever $Q \in \mathbb{F}_p[X]$ is taken irreducible:

$$\mathbb{F}_q = \mathbb{F}_p[X]/(Q)$$

(3) Regarding now the best choice of the irreducible polynomial $Q \in \mathbb{F}_p[X]$, providing us with a good model for the finite field $\mathbb{F}_q$, that we can use in practice, this question depends on the value of $q = p^k$, and many things can be said here. All in all, our models are quite similar to $\mathbb{C} = \mathbb{R}[i]$, with i being a formal number satisfying $i^2 = -1$.

(4) To be more precise, at the simplest exponent, $q = 4$, to start with, we can use $Q = X^2 + X + 1$, with this being actually the unique possible choice of a degree 2 irreducible polynomial $Q \in \mathbb{F}_2[X]$, and this leads to a model as follows:

$$\mathbb{F}_4 = \left\{ 0, 1, a, a + 1 \,\middle|\, a^2 = a + 1 \right\}$$

To be more precise here, we assume of course that the characteristic of our model is $p = 2$, which reads $x + x = 0$ for any x, and so determines the addition table. As for the multiplication table, this is uniquely determined by $a^2 = -a - 1 = a + 1$.

(5) Next, at exponents of type $q = p^2$ with $p \geq 3$ prime, we can use $Q = X^2 - r$, with r being a non-square modulo p, and with $(p - 1)/2$ choices here. We are led to:

$$\mathbb{F}_{p^2} = \left\{ a + b\gamma \,\middle|\, \gamma^2 = r \right\}$$

Here, as before with $\mathbb{F}_4$, our formula is something self-explanatory. Observe the analogy with $\mathbb{C} = \mathbb{R}[i]$, with i being a formal number satisfying $i^2 = -1$.

(6) Finally, at $q = p^k$ with $k \geq 3$ things become more complicated, but the main idea remains the same. We have for instance models for $\mathbb{F}_8$, $\mathbb{F}_{27}$ using $Q = X^3 - X - 1$, and a model for $\mathbb{F}_{16}$ using $Q = X^4 + X + 1$. Many other things can be said here. □

As another application of the above, which motivated Galois, we have:

THEOREM 4.18. *Unlike in degree $N \leq 4$, there is no formula for the roots of polynomials of degree $N = 5$ and higher, with the reason for this, coming from Galois theory, being that S_5 is not solvable. The simplest numeric example is $P = X^5 - X - 1$.*

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

(1) The first assertion, for generic polynomials, is due to Abel-Ruffini, but Galois theory helps in better understanding this, and comes with a number of bonus points too, namely the possibility of formulating a finer result, with Abel-Ruffini's original "generic", which was something algebraic, being now replaced by an analytic "generic", and also with the possibility of dealing with concrete polynomials, such as $P = X^5 - X - 1$.

(2) Regarding now the details of the Galois proof of the Abel-Ruffini theorem, assume that the roots of a polynomial $P \in F[X]$ can be computed by using iterated roots, a bit as for the degree 2 equation, or for the degree 3 and 4 equations, via Cardano. Then, 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 $x_i^{n_i} \in F_i$:

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

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

(4) Now by Galois theory, at the level of the corresponding Galois groups we obtain a tower of groups as follows 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.

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

(6) Finally, regarding the polynomial $P = X^5 - X - 1$, some elementary computations here, based on arithmetic over $\mathbb{F}_2, \mathbb{F}_3$, and involving various cycles of length $2, 3, 5$, show that its Galois group is S_5. Thus, we have our counterexample.

(7) To be more precise, our polynomial factorizes over $\mathbb{F}_2$ as follows:

$$X^5 - X - 1 = (X^2 + X + 1)(X^3 + X^2 + 1)$$

We deduce from this the existence of an element $\tau\sigma \in G \subset S_5$, with $\tau \in S_5$ being a transposition, and with $\sigma \in S_5$ being a 3-cycle, disjoint from it. Thus, we have:

$$\tau = (\tau\sigma)^3 \in G$$

(8) On the other hand since $P = X^5 - X - 1$ is irreducible over $\mathbb{F}_5$, we have as well available a certain 5-cycle $\rho \in G$. Now since $< \tau, \rho >= S_5$, we conclude that the Galois group of P is full, $G = S_5$, and by (4) and (5) we have our counterexample.

(9) Finally, as mentioned in (1), all this shows as well that a random polynomial of degree 5 or higher is not solvable by square roots, and with this being an elementary consequence of the main result from (5), via some standard analysis arguments. □

Many other things can be said about Galois theory, and its applications. We will be back to this in Part II, when doing systematic algebra, in the number theory context.

4c. Squares, again

Let us go back now to the quadratic reciprocity theorem, established in chapter 3 by using Gauss sums, with the aim of better understanding the algebra behind. We first have the following result, due to Eisenstein, based on a previous proof of Gauss, not exactly related to the Gauss sums and proof from chapter 3, which is something standard:

THEOREM 4.19. *The quadratic reciprocity formula*

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

can be established via lattice counting in the plane.

PROOF. This is something very standard, the idea 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:

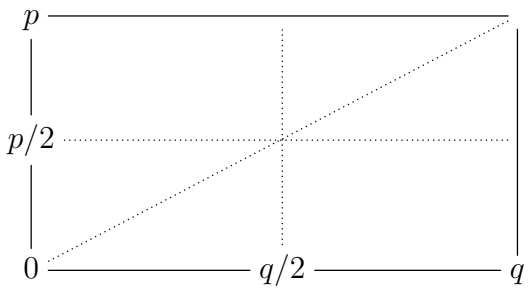
$$\left(\frac{a}{q}\right)=(-1)^n$$

(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. □

We have as well the following result, which is more advanced:

THEOREM 4.20. *The quadratic reciprocity formula*

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

can be established as well via algebra, with this leading to several key extensions of it.

PROOF. Here the first assertion is something quite standard, by using $\mathbb{Q}(w)$ with $w = e^{2\pi i/p}$, with the remark however that the algebra is quite tough, and is no match for the proof of Gauss from chapter 3, or for the proof of Eisenstein above. However, in what regards the second assertion, extensions, this is something very interesting, and generally speaking, this is where the algebraic approach to quadratic reciprocity really shines. □

Many other things can be said about all this. We will be back to this in Part II, when doing systematic algebra, in the number theory context.

4d. Sums of roots

As another application of abstract algebra, let us go back to the complex Hadamard matrices, introduced at the end of chapter 3. Following Butson, let us formulate:

DEFINITION 4.21. *An Hadamard matrix is called of Butson type if its entries are roots of unity of finite order. The Butson class $H_N(l)$ consists of the Hadamard matrices*

$$H \in M_N(\mathbb{Z}_l)$$

where $\mathbb{Z}_l$ is the group of the l-th roots of unity. The level of a Butson matrix $H \in M_N(\mathbb{T})$ is the smallest integer $l \in \mathbb{N}$ such that $H \in H_N(l)$.

As a basic illustration, the real Hadamard matrices $H \in M_N(\pm 1)$ that we were struggling with before are Butson matrices, $H \in M_N(2)$. The Fourier matrix is a Butson matrix too, $F_N \in H_N(N)$. As already mentioned, many other examples can be constructed, via all sorts of arithmetic tricks, a bit similar to those of Paley and Williamson.

Generally speaking, the main question regarding the Butson matrices is that of understanding when $H_N(l) \neq 0$, via a theorem providing obstructions, and then a result or conjecture stating that these obstructions are the only ones. Let us begin with:

PROPOSITION 4.22 (Sylvester obstruction). *The following holds,*

$$H_N(2) \neq \emptyset \implies N \in \{2\} \cup 4\mathbb{N}$$

due to the orthogonality of the first 3 rows.

PROOF. This is indeed something that we know well, from chapter 3. □

Our purpose now will be that of finding analogous statements at $l \geq 3$. At very small values of l this is certainly possible, and in what regards the needed obstructions, we can get away with the following simple fact, coming from basic number theory:

PROPOSITION 4.23. *For a prime power $l = p^a$, the vanishing sums of roots of unity*

$$\lambda_1 + \ldots + \lambda_N = 0 \quad , \quad \lambda_i \in \mathbb{Z}_l$$

appear as formal sums of rotated full sums of p-th roots of unity.

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

(1) Consider indeed the full sum of p-th roots of unity, taken in a formal sense:

$$S = \sum_{k=1}^{p} (e^{2\pi i/p})^k$$

Let also $w = e^{2\pi i/l}$, and for $r \in \{1, 2, \ldots, l/p\}$ let us denote by $S_p^r = w^r \cdot S$ the above formal sum of roots of unity, rotated by w^r:

$$S_p^r = \sum_{k=1}^{p} w^r (e^{2\pi i/p})^k$$

(2) We must show that any vanishing sum of l-th roots of unity appears as a sum of such quantities S_p^r. For this purpose, consider the following map, which assigns to the abstract elements of the ring $\mathbb{Z}[\mathbb{Z}_l]$ their precise numeric values, inside $\mathbb{Z}(w) \subset \mathbb{C}$:

$$\Phi : \mathbb{Z}[\mathbb{Z}_l] \to \mathbb{Z}(w)$$

Our claim is that the elements $\{S_p^r\}$ form a basis of the vector space $\ker \Phi$.

(3) In order to prove this claim, observe first that we have $S_p^r \in \ker \Phi$. Also, the elements S_p^r are linearly independent, because the support of S_p^r contains a unique element of the subset $\{1, 2, \ldots, p^{a-1}\} \subset \mathbb{Z}_l$, namely the element $r \in \mathbb{Z}_l$, so all the coefficients of a vanishing linear combination of such sums S_p^r must vanish.

(4) Thus, we are left with proving that $\ker \Phi$ is spanned by the elements $\{S_p^r\}$. For this purpose, let us recall from chapter 1 that the minimal polynomial of w is:

$$\frac{X^{p^a} - 1}{X^{p^{a-1}} - 1} = 1 + X^{p^{a-1}} + X^{2p^{a-1}} + \ldots + X^{(p-1)p^{a-1}}$$

We conclude that the dimension of $\ker \Phi$ is given by:

$$\dim(\ker \Phi) = p^a - (p^a - p^{a-1}) = p^{a-1}$$

(5) Now since this is exactly the number of the sums S_p^r, this finishes the proof of our claim. Thus, any vanishing sum of l-th roots of unity must be of the form $\sum \pm S_p^r$, and the above support considerations show the coefficients must be positive, as desired. □

We can now formulate a result in the spirit of Proposition 4.22, as follows:

PROPOSITION 4.24 (Butson obstruction). *The following holds,*

$$H_N(p^a) \neq \emptyset \implies N \in p\mathbb{N}$$

due to the orthogonality of the first 2 *rows.*

PROOF. This follows indeed from Proposition 4.23, because the scalar product between the first 2 rows of our matrix is a vanishing sum of l-th roots of unity. □

Let us discuss now a generalization of the above obstruction. We first have:

DEFINITION 4.25. *A cycle is a full sum of roots of unity, possibly rotated by a scalar,*

$$C = q \sum_{k=1}^{l} w^k \quad , \quad w = e^{2\pi i/l} \quad , \quad q \in \mathbb{T}$$

and taken in a formal sense. A sum of cycles is a formal sum of cycles.

The actual sum of a cycle, or of a sum of cycles, is of course 0. This is why the word "formal" is there, for reminding us that we are working with formal sums. As an example, here is a sum of cycles, with $w = e^{2\pi i/6}$, and with $|q| = 1$:

$$1 + w^2 + w^4 + qw + qw^4 = 0$$

We know from Proposition 4.23 that any vanishing sum of l-th roots of unity must be a sum of cycles, at least when $l = p^a$ is a prime power. However, this is not the case in general, the simplest counterexample being as follows, with $w = e^{2\pi i/30}$:

$$w^5 + w^6 + w^{12} + w^{18} + w^{24} + w^{25} = 0$$

Indeed, this sum is obviously not a sum a cycles. However, this sum vanishes indeed, as shown by the following computation:

$$\begin{aligned} w^5 + w^6 + w^{12} + w^{18} + w^{24} + w^{25} &= w^5 + w^{15} + w^{25} \\ &+ w^0 + w^6 + w^{12} + w^{18} + w^{24} \\ &- w^0 - w^{15} \\ &= 0 + 0 - 0 \\ &= 0 \end{aligned}$$

It is convenient at this point to formulate an abstract definition, as follows:

DEFINITION 4.26. *Let $l \in \mathbb{N}$ be a number, and $K \subset \mathbb{C}$ be a subset.*

(1) *A K-sum of l-roots is a sum of l-roots, with coefficients in K.*
(2) *A K-trivial sum of l-roots is a sum of sums S_p^r, with coefficients in K.*

The question is, if a K-sum of l-roots vanishes, is it K-trivial? We discuss here this question, in the cases $K = \mathbb{Q}, \mathbb{Z}, \mathbb{N}$, and then we will come back to the complex Hadamard matrices, and the Butson obstruction. To start with, in the case $l = p^a$, we have:

PROPOSITION 4.27. *Given a prime power $l = p^a$, a sum of l-roots of unity with rational coefficients vanishes if and only if it is $\mathbb{Q}$-trivial.*

PROOF. This is a version of Proposition 4.23. Let $w = e^{2\pi i/l}$, and consider the following map, which assigns to the elements of the algebra $\mathbb{Q}[\mathbb{Z}_l]$ their numeric values:

$$\Phi : \mathbb{Q}[\mathbb{Z}_l] \to \mathbb{Q}(w)$$

In order to prove the result, we have to show is that the $\mathbb{Q}$-trivial sums exhaust $\ker \Phi$, and for this purpose it is enough to show that the sums S_p^r with $r \in \{1, 2, \ldots, l/p\}$ form a basis of $\ker \Phi$. But this follows as in the proof of Proposition 4.23. □

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

THEOREM 4.28. *Given a number $l \in \mathbb{N}$, a sum of l-roots of unity with rational coefficients vanishes if and only if it is $\mathbb{Q}$-trivial.*

PROOF. This can be established by using similar arguments, as follows:

(1) We let $w_l = e^{2\pi i/l}$, and we consider the corresponding evaluation map, from abstract sums of l-roots to the complex numbers:

$$\Phi_l : \mathbb{Q}[\mathbb{Z}_l] \to \mathbb{Q}(w_l)$$

By arguing like in the previous proof, we have to show that the basic sums S_p^r with $p|l$ prime and $r \in \{1, 2, \ldots, l/p\}$ form a basis of $\ker \Phi_l$.

(2) For this purpose, we write $l = l_1 \ldots l_s$, where $l_i = p_i^{a_i}$ are the various prime power factors of l. We have the following commuting diagram:

$$\begin{array}{ccc} \mathbb{Q}[\mathbb{Z}_l] & \simeq & \mathbb{Q}[\mathbb{Z}_{l_1}] \otimes \ldots \otimes \mathbb{Q}[\mathbb{Z}_{l_s}] \\ \Phi_l \downarrow & & \downarrow \Phi_{l_1} \otimes \ldots \otimes \Phi_{l_s} \\ \mathbb{Q}(w_l) & \simeq & \mathbb{Q}(w_{l_1}) \otimes \ldots \otimes \mathbb{Q}(w_{l_s}) \end{array}$$

Here the upper isomorphism is the canonical one, induced by the group isomorphism $\mathbb{Z}_l \simeq \mathbb{Z}_{l_1} \times \ldots \times \mathbb{Z}_{l_s}$, along with the fact that the $\times$ operation at level of groups corresponds to the $\otimes$ operation at level of group algebras, with all this being very standard.

(3) We want to show that $\ker \Phi_l$ is spanned by the following set:

$$K_l = \bigcup_i \left\{ S_{p_i}^r \middle| r = 1, 2, \ldots, l/p_i \right\}$$

In order to prove this, we regard each exponent r as being an element of the cyclic group $\mathbb{Z}_l$, and we write it $r = r_1 \ldots r_s$, with $r_i \in \{1, 2, \ldots, l_i/p_i\}$ and $r_j \in \{1, 2, \ldots, l_j\}$ for $j \neq i$. With this convention, the above set becomes:

$$K_l = \bigcup_i \left\{ S_{p_i}^{r_1 \ldots r_s} \middle| r_i = 1, \cdots, l_i/p_i, r_j = 1, \cdots, l_j \text{ for } j \neq i \right\}$$

(4) Now this is by definition a subset of the group algebra $\mathbb{Q}[\mathbb{Z}_l]$. By regarding it as subset of the tensor product of the group algebras $\mathbb{Q}[\mathbb{Z}_{l_i}]$, its elements are:

$$S_{p_i}^{r_1 \ldots r_s} = \{r_1\} \otimes \ldots \otimes \{r_{i-1}\} \otimes S_{p_i}^{r_i} \otimes \{r_{i+1}\} \otimes \ldots \otimes \{r_s\}$$

In other words, in terms of tensor products, we have:

$$K_l = \bigcup_i \mathbb{Z}_{l_1} \otimes \ldots \otimes \mathbb{Z}_{l_{i-1}} \otimes K_{l_i} \otimes \mathbb{Z}_{l_{i+1}} \otimes \ldots \otimes \mathbb{Z}_{l_s}$$

Now since each K_{l_i} is a basis of $\ker(\Phi_{l_i})$ and each $\mathbb{Z}_{l_i}$ is a basis of $\mathbb{Q}[\mathbb{Z}_{l_i}]$, it follows that K_l is a basis for $\ker(\Phi_{l_1} \otimes \ldots \otimes \Phi_{l_s})$, as desired. □

We have the following positive answer to our triviality question:

THEOREM 4.29. *Given a number $l \in \mathbb{N}$, a sum of l-roots of unity with integer coefficients vanishes if and only if it is $\mathbb{Z}$-trivial.*

PROOF. This can be established by using the above results, as follows:

(1) Let S be a vanishing sum of l-roots, with integer coefficients. In the case $l = p^a$, we regard S as a sum with rational coefficients. By applying Proposition 4.27 we get that S is $\mathbb{Q}$-trivial. This means that we have a formula as follows, with $\lambda_1, \ldots, \lambda_{l/p} \in \mathbb{Q}$:

$$S = \sum_{r=1}^{l/p} \lambda_r S_p^r$$

The sums S_p^r being disjoint, each λ_r can be regarded as multiplicity of a given root of unity $w_r \in S_p^r$ inside the total sum S. Thus these scalars are integers, and we are done.

(2) In the general case $l = p_1^{a_1} \ldots p_s^{a_s}$, we can follow the proof of Theorem 4.28, by changing the scalars. So, let $w_l = e^{2\pi i/l}$ and consider the evaluation map $\Psi_l : \mathbb{Z}[\mathbb{Z}_l] \to \mathbb{Z}(w_l)$. We have to show that $\ker \Psi_l$ is spanned by the following set:

$$K_l = \bigcup_i \left\{ S_{p_i}^r \middle| r = 1, 2, \ldots, l/p_i \right\}$$

With the notation $l_i = p_i^{a_i}$, we have the following commuting diagram:

$$\begin{array}{ccc} \mathbb{Z}[\mathbb{Z}_l] & \simeq & \mathbb{Z}[\mathbb{Z}_{l_1}] \otimes \ldots \otimes \mathbb{Z}[\mathbb{Z}_{l_s}] \\ \Psi_l \downarrow & & \downarrow \Psi_{l_1} \otimes \ldots \otimes \Psi_{l_s} \\ \mathbb{Z}(w_l) & \simeq & \mathbb{Z}(w_{l_1}) \otimes \ldots \otimes \mathbb{Z}(w_{l_s}) \end{array}$$

As in the proof of Theorem 4.28, we can use these canonical isomorphisms in order to decompose K_l, in the following way:

$$\begin{aligned} K_l &= \bigcup_i \left\{ S_{p_i}^{r_1 \ldots r_s} \mid r_i = 1, 2, \ldots, l_i/p_i,\ r_j = 1, \cdots, l_j \text{ for } j \neq i \right\} \\ &= \bigcup_i \Big\{ \{r_1\} \otimes \ldots \otimes \{r_{i-1}\} \otimes S_{p_i}^{r_i} \otimes \{r_{i+1}\} \otimes \ldots \otimes \{r_s\} \mid \\ &\qquad r_i = 1, 2, \ldots, l_i/p_i,\ r_j = 1, \cdots, l_j \text{ for } j \neq i \Big\} \\ &= \bigcup_i \mathbb{Z}_{l_1} \otimes \ldots \otimes \mathbb{Z}_{l_{i-1}} \otimes K_{l_i} \otimes \mathbb{Z}_{l_{i+1}} \otimes \ldots \otimes \mathbb{Z}_{l_s} \end{aligned}$$

Now we know from the $l = p^a$ case that each K_{l_i} spans $\ker(\Psi_{l_i})$. Together with the fact that each $\mathbb{Z}_{l_i}$ spans $\mathbb{Z}[\mathbb{Z}_{l_i}]$, this shows that K_l spans $\ker(\Psi_{l_1} \otimes \ldots \otimes \Psi_{l_s})$, as desired. □

It is convenient at this point to start using a simpler terminology, by calling "sum" the usual sums, and "trivial sum" the $\mathbb{N}$-trivial sums. With this convention, we have:

THEOREM 4.30. *For a number l, the following are equivalent:*

(1) *If a sum of l-roots vanishes, then it is trivial.*
(2) *l has at most two prime factors.*

PROOF. This follows from what we have so far, as follows:

(1) We already know that this cannot happen at $l = 30$, and the same construction works for any number l, having at least three prime factors. We may assume that we have $l = pqr$, with p, q, r relatively prime, and a counterexample is as follows:

$$\begin{aligned} S &= S_p^{pq} + S_p^{2pq} + \ldots + S_p^{(r-1)pq} \\ &\quad + S_q^{qr} + S_q^{2qr} + \ldots + S_q^{(p-1)qr} \\ &\quad - S_r^{qr} - S_r^{2qr} - \ldots - S_r^{(p-1)qr} \end{aligned}$$

(2) Now for the converse, assume that l has at most two prime factors, and let S be a vanishing sum of l-roots. We have to prove that S is trivial. By applying Theorem 4.29 we can write $S = U - V$, where both U, V are trivial sums. We choose such a decomposition of S, having the property that the number of terms of V is minimal.

(3) Assume that this number of terms, say n, is nonzero. In the case $l = p^a$ we get from $S + V = U$ that any basic component S_p^r of the sum V has to appear in U. By removing one such component from both U, V we get a decomposition of type $S = U' - V'$ with V' having less terms than V, contradiction. Thus our assumption $n \neq 0$ is wrong.

(4) In the case $l = p^a q^b$ we can use a similar argument. We pick a basic component of V, say S_p^r. Then if S_p^r appears in U, we can remove it, and as before we get a new decomposition $S = U' - V'$ with less terms, which is a contradiction.

(5) Otherwise, this means that the elements of S_p^r have to appear in U as part of sums of type S_q^s. But there are p such elements, and the union of the p sums of type S_q^s which contain them can be rewritten as a union of q sums of type S_p^v, with one of these sums being precisely S_p^r. Thus this case reduces in fact to the first one, and so done. □

Going now towards a generalized Butson obstruction, we have:

PROPOSITION 4.31. *For two numbers n and $l = p^a q^b$, the following are equivalent:*

(1) *There is a n-sum of l-roots which vanishes.*
(2) *$n \in p\mathbb{N} + q\mathbb{N}$.*

PROOF. When n and $l = p^a q^b$ are such that there is a n-sum of l-roots which vanishes, Theorem 4.30 applies, and shows that the sum is trivial. In particular the number of terms, which is n, must be a sum of numbers of the form $\#S_p^r = p$ and $\#S_q^r = q$, so we have indeed $n \in p\mathbb{N} + q\mathbb{N}$. As for the converse, if we have $n = pa + qb$ with $a, b \in \mathbb{N}$, we can consider the sum formed by a copies of S_p and b copies of S_q, and we are done. □

We will see later that the above result holds as well for exponents of type $l=p^aq^br^c$, and even in general. Let us record here the fact that this is true for the only example of non-trivial sum that we have so far, namely the one in proof of Theorem 4.30:

PROPOSITION 4.32. *The number of terms of the sum*

$$\begin{aligned}S &= S_p^{pq}+S_p^{2pq}+\ldots+S_p^{(r-1)pq}\\ &\quad +S_q^{qr}+S_q^{2qr}+\ldots+S_q^{(p-1)qr}\\ &\quad -S_r^{qr}-S_r^{2qr}-\ldots-S_r^{(p-1)qr}\end{aligned}$$

satisfies $n\in p\mathbb{N}+q\mathbb{N}+r\mathbb{N}$.

PROOF. The number of terms of the sum is given by:

$$\begin{aligned}n &= (r-1)p+(p-1)q-(p-1)r\\ &= rp-p+pq-q-pr+r\\ &= pq-p-q+r\\ &= (p-1)(q-1)+(r-1)\end{aligned}$$

In the case $r>p$, we obtain the following estimate:

$$n>(p-1)(q-1)+(p-1)=(p-1)q$$

But this shows that the numbers $n,n-q,n-2q,\ldots,n-(p-1)q$ are all positive. Now since p must divide one of them, we get $n\in p\mathbb{N}+q\mathbb{N}$, which is the desired result. As for the other case, where $p>r$, here we get the estimate $n>(r-1)q$, and by using the same argument we obtain $n\in q\mathbb{N}+r\mathbb{N}$, so we are done as well. □

The following non-trivial result on the subject is due to Lam and Leung:

THEOREM 4.33. *Let* $l=p_1^{a_1}\ldots p_k^{a_k}$, *and assume that* $\lambda_i\in\mathbb{Z}_l$ *satisfy:*

$$\lambda_1+\ldots+\lambda_N=0$$

(1) $\sum\lambda_i$ *is a sum of cycles, with* $\mathbb{Z}$ *coefficients.*
(2) *If* $k\leq 2$ *then* $\sum\lambda_i$ *is a sum of cycles, with* $\mathbb{N}$ *coefficients.*
(3) *If* $k\geq 3$ *then* $\sum\lambda_i$ *might not decompose as a sum of cycles.*
(4) $\sum\lambda_i$ *has the same length as a sum of cycles:* $N\in p_1\mathbb{N}+\ldots+p_k\mathbb{N}$.

PROOF. This is something technical, the idea of the proof being as follows:

(1) This is a well-known result, which follows from basic number theory, by using arguments in the spirit of those in the proof of Proposition 4.23.

(2) This is something that we already know at $k=1$, from Proposition 4.23. At $k=2$ the proof is more technical, along the same lines, as explained above.

(3) The smallest possible l potentially producing a counterexample is $l=2\cdot 3\cdot 5=30$, and we have here indeed the sum given above, with $w=e^{2\pi i/30}$.

(4) This is a deep result, due to Lam and Leung, relying on advanced number theory techniques, and we refer to their work for the proof. □

As a consequence of the above result, we have the following generalization of the Butson obstruction, which is something final and optimal on this subject:

THEOREM 4.34 (Lam-Leung obstruction). *Assuming that we have*

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

the following must hold, due to the orthogonality of the first 2 rows:

$$H_N(l) \neq \emptyset \implies N \in p_1\mathbb{N} + \ldots + p_k\mathbb{N}$$

In the case $k \geq 2$, the latter condition is automatically satisfied at $N >> 0$.

PROOF. Here the first assertion, which generalizes the $l = p^a$ obstruction from Proposition 4.24, comes from Theorem 4.33 (4), applied to the vanishing sum of l-th roots of unity coming from the scalar product between the first 2 rows. As for the second assertion, this is something well-known, coming from basic number theory. □

4e. Exercises

Tough chapter that we had here, with a lot of algebra, and even myself I feel a bit dizzy, after typing it in. As exercises on all this, we have more algebra, as follows:

EXERCISE 4.35. *Learn, with full details, the classification of finite abelian groups.*

EXERCISE 4.36. *Have a look as well into Lie groups, and into reflection groups.*

EXERCISE 4.37. *Find out what left and right ideals, not two-sided, are good for.*

EXERCISE 4.38. *List all the examples of complex algebras that you know.*

EXERCISE 4.39. *Clarify under which exact assumptions Galois theory works.*

EXERCISE 4.40. *Clarify all details for splitting fields, and algebraic closures.*

EXERCISE 4.41. *Prove the Eisenstein formula for the Legendre symbol.*

EXERCISE 4.42. *Have some fun with finite fields, and what can be done with them.*

As bonus exercise, learn full Galois theory, the hard way, from an old book of your choice, with all theorems read, details understood, and a lot of exercises done too.

Part II

Algebraic methods

See for me that her hair's hanging down
It curls and falls all down her breast
See for me that her hair's hanging down
That's the way I remember her best

CHAPTER 5

Algebraic geometry

5a. Curves, surfaces

Welcome to geometry. At the beginning were the ellipses, motivated by the trajectory of the Sun around the Earth, or perhaps vice versa, their math being as follows:

THEOREM 5.1. *The ellipses, taken centered at the origin* 0*, and squarely oriented with respect to* Oxy*, can be defined in* 4 *possible ways, as follows:*

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

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

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

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

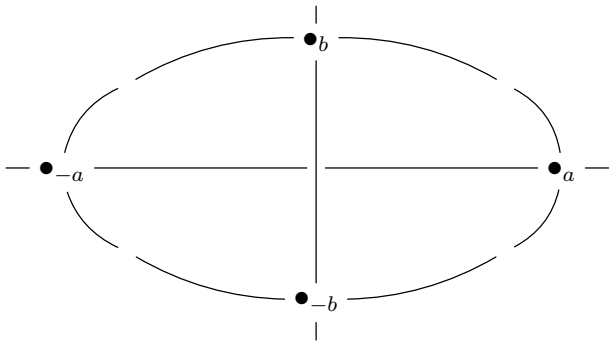
(3) *As the curves appearing when drawing a circle, from various perspectives:*

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

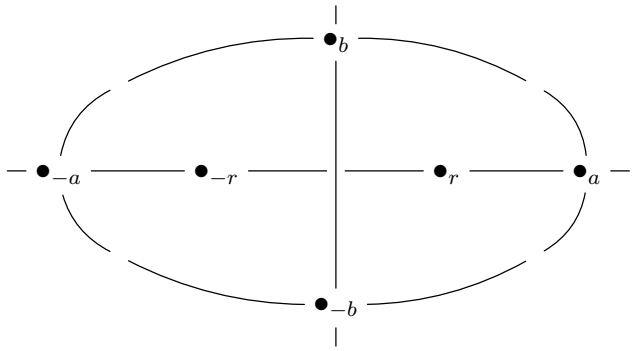
(4) *As the closed non-degenerate curves appearing by cutting a cone with a plane.*

PROOF. We have to prove that the above constructions (1-4) give rise indeed to the same class of curves, and this can be done as follows:

(1) To start with, let us draw a picture from what comes out of (1), which will be our main definition for the ellipses, in what follows:



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



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

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

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

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

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

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

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

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

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

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

All this is very nice, but before getting into physics, with some explanations for the fact that planets travel indeed on ellipses, let us settle as well the question of wandering asteroids. Observations show that these can travel on parabolas and hyperbolas, so what we need as mathematics is a unified theory of ellipses, parabolas and hyperbolas. And fortunately, this theory exists, also since the ancient Greeks, summarized as follows:

THEOREM 5.2. *The conics, which are the algebraic curves of degree 2 in the plane,*

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

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

PROOF. This follows by further building on Theorem 5.1, as follows:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$dx+ey+f=0$$

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

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

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

Ready for some physics? And here, good news, not only what we did in the above is relevant, but is actually at the origin of the whole modern physics, thanks to:

THEOREM 5.3. *Planets and other celestial bodies move around the Sun on conics,*

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

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

PROOF. This is something quite long, due to Kepler and Newton, as follows:

(1) The force of attraction between two bodies of masses M,m is given by:

$$||F||=G\cdot\frac{Mm}{d^2}$$

Here d is the distance between the two bodies, and $G\simeq 6.674\times 10^{-11}$ is a constant. Now assuming that M is fixed at $0\in\mathbb{R}^3$, the force exterted on m positioned at $x\in\mathbb{R}^3$, regarded as a vector $F\in\mathbb{R}^3$, is given by the following formula:

$$F=-||F||\cdot\frac{x}{||x||}=-\frac{GMm}{||x||^2}\cdot\frac{x}{||x||}=-\frac{GMmx}{||x||^3}$$

But $F=ma=m\ddot{x}$, with $a=\ddot{x}$ being the acceleration, second derivative of the position, so the equation of motion of m, assuming that M is fixed at 0, is:

$$\ddot{x}=-\frac{GMx}{||x||^3}$$

Obviously, the problem happens in 2 dimensions, and you can even find, as an exercise, a formal proof of that, based on the above equation, if your really want to. Now here the most convenient is to use standard x,y coordinates, and denote our point as $z=(x,y)$. With this change made, and by setting $K=GM$, the equation of motion becomes:

$$\ddot{z}=-\frac{Kz}{||z||^3}$$

(2) The idea now is that the problem can be solved via some calculus. Let us write indeed our vector $z=(x,y)$ in polar coordinates, as follows:

$$x=r\cos\theta\quad,\quad y=r\sin\theta$$

We have then $||z||=r$, and our equation of motion becomes:

$$\ddot{z}=-\frac{Kz}{r^3}$$

Let us differentiate now x, y. By using the standard calculus rules, we have:

$$\dot{x} = \dot{r}\cos\theta - r\sin\theta\cdot\dot{\theta}$$

$$\dot{y} = \dot{r}\sin\theta + r\cos\theta\cdot\dot{\theta}$$

Differentiating one more time gives the following formulae:

$$\ddot{x} = \ddot{r}\cos\theta - 2\dot{r}\sin\theta\cdot\dot{\theta} - r\cos\theta\cdot\dot{\theta}^2 - r\sin\theta\cdot\ddot{\theta}$$

$$\ddot{y} = \ddot{r}\sin\theta + 2\dot{r}\cos\theta\cdot\dot{\theta} - r\sin\theta\cdot\dot{\theta}^2 + r\cos\theta\cdot\ddot{\theta}$$

Consider now the following two quantities, appearing as coefficients in the above:

$$a = \ddot{r} - r\dot{\theta}^2 \quad , \quad b = 2\dot{r}\dot{\theta} + r\ddot{\theta}$$

In terms of these quantities, our second derivative formulae read:

$$\ddot{x} = a\cos\theta - b\sin\theta$$

$$\ddot{y} = a\sin\theta + b\cos\theta$$

(3) We can now solve the equation of motion from (1). Indeed, with the formulae that we found for $\ddot{x}, \ddot{y}$, our equation of motion takes the following form:

$$a\cos\theta - b\sin\theta = -\frac{K}{r^2}\cos\theta$$

$$a\sin\theta + b\cos\theta = -\frac{K}{r^2}\sin\theta$$

But these two formulae can be written in the following way:

$$\left(a + \frac{K}{r^2}\right)\cos\theta = b\sin\theta \quad , \quad \left(a + \frac{K}{r^2}\right)\sin\theta = -b\cos\theta$$

By making now the product, and assuming that we are in a non-degenerate case, where the angle θ varies indeed, we obtain by positivity that we must have:

$$a + \frac{K}{r^2} = b = 0$$

(4) Let us first examine the second equation, $b = 0$. This can be solved as follows:

$$\begin{aligned}
b = 0 &\iff 2\dot{r}\dot{\theta} + r\ddot{\theta} = 0\\
&\iff \frac{\ddot{\theta}}{\dot{\theta}} = -2\frac{\dot{r}}{r}\\
&\iff (\log\dot{\theta})' = (-2\log r)'\\
&\iff \log\dot{\theta} = -2\log r + c\\
&\iff \dot{\theta} = \frac{\lambda}{r^2}
\end{aligned}$$

As for the first equation the we found, namely $a + K/r^2 = 0$, this becomes:

$$\ddot{r} - \frac{\lambda^2}{r^3} + \frac{K}{r^2} = 0$$

As a conclusion to all this, in polar coordinates, $x = r\cos\theta$, $y = r\sin\theta$, our equations of motion are as follows, with λ being a constant, not depending on t:

$$\ddot{r} = \frac{\lambda^2}{r^3} - \frac{K}{r^2} \quad , \quad \dot{\theta} = \frac{\lambda}{r^2}$$

Even better now, by writing $K = \lambda^2/c$, these equations read:

$$\ddot{r} = \frac{\lambda^2}{r^2}\left(\frac{1}{r} - \frac{1}{c}\right) \quad , \quad \dot{\theta} = \frac{\lambda}{r^2}$$

(5) In order to study the first equation, we use a trick. Let us write:

$$r(t) = \frac{1}{f(\theta(t))}$$

Abbreviated, and by reminding that f takes $\theta = \theta(t)$ as variable, this reads:

$$r = \frac{1}{f}$$

With the convention that dots mean as usual derivatives with respect to t, and that the primes will denote derivatives with respect to $\theta = \theta(t)$, we have:

$$\dot{r} = -\frac{f'\dot{\theta}}{f^2} = -\frac{f'}{f^2} \cdot \frac{\lambda}{r^2} = -\lambda f'$$

By differentiating one more time with respect to t, we obtain:

$$\ddot{r} = -\lambda f''\dot{\theta} = -\lambda f'' \cdot \frac{\lambda}{r^2} = -\frac{\lambda^2}{r^2} f''$$

On the other hand, our equation for $\ddot{r}$ found in (4) above reads:

$$\ddot{r} = \frac{\lambda^2}{r^2}\left(\frac{1}{r} - \frac{1}{c}\right) = \frac{\lambda^2}{r^2}\left(f - \frac{1}{c}\right)$$

Thus, in terms of $f = 1/r$ as above, our equation for $\ddot{r}$ simply reads:

$$f'' + f = \frac{1}{c}$$

But this latter equation is elementary to solve. Indeed, both functions $\cos t, \sin t$ satisfy $g" + g = 0$, so any linear combination of them satisfies as well this equation. But the solutions of $f'' + f = 1/c$ being those of $g'' + g = 0$ shifted by $1/c$, we obtain:

$$f = \frac{1 + \varepsilon\cos\theta + \delta\sin\theta}{c}$$

Now by inverting, we obtain the following formula:

$$r = \frac{c}{1 + \varepsilon \cos\theta + \delta \sin\theta}$$

(6) But this leads to the conclusion that the trajectory is a conic. Indeed, in terms of the parameter θ, the formulae of the coordinates are:

$$x = \frac{c\cos\theta}{1 + \varepsilon \cos\theta + \delta \sin\theta} \quad , \quad y = \frac{c\sin\theta}{1 + \varepsilon \cos\theta + \delta \sin\theta}$$

Now observe that these two functions x, y satisfy the following formula:

$$x^2 + y^2 = \frac{c^2(\cos^2\theta + \sin^2\theta)}{(1 + \varepsilon \cos\theta + \delta \sin\theta)^2} = \frac{c^2}{(1 + \varepsilon \cos\theta + \delta \sin\theta)^2}$$

On the other hand, these two functions satisfy as well the following formula:

$$\begin{aligned}(\varepsilon x + \delta y - c)^2 &= \frac{c^2\big(\varepsilon \cos\theta + \delta \sin\theta - (1 + \varepsilon \cos\theta + \delta \sin\theta)\big)^2}{(1 + \varepsilon \cos\theta + \delta \sin\theta)^2} \\ &= \frac{c^2}{(1 + \varepsilon \cos\theta + \delta \sin\theta)^2}\end{aligned}$$

We conclude that our coordinates x, y satisfy the following equation:

$$x^2 + y^2 = (\varepsilon x + \delta y - c)^2$$

But what we have here is an equation of a conic, and we are done. □

As a conclusion, the conics are the simplest curves, both in mathematics and physics. In fact, this type of phenomenon, that what is simplest in mathematics is what is simplest in physics too, has been the guiding idea since, both in mathematics and physics.

5b. Algebraic manifolds

All the above was very nice, all good old things, but the continuation of the story is more complicated, because beyond conics, things ramify. A first idea, in order to generalize the conics, is to look at the smooth manifolds, in the following sense:

DEFINITION 5.4. *A smooth manifold is a space X which is locally isomorphic to $\mathbb{R}^N$. To be more precise, this space X must be covered by charts, bijectively mapping open pieces of it to open pieces of $\mathbb{R}^N$, with the changes of charts being C^∞ functions.*

As a basic example, we have $\mathbb{R}^N$ itself, or any open subset $X \subset \mathbb{R}^N$. Another example is the circle, or curves like ellipses and so on, for obvious reasons. To be more precise, the unit circle can be covered by 2 charts as above, by using polar coordinates, in the obvious way, and then by applying dilations, translations and other such transformations, namely bijections which are smooth, we obtain a whole menagery of circle-looking manifolds.

Here is a more precise statement in this sense, covering the conics:

THEOREM 5.5. *The following are smooth manifolds, in the plane:*

(1) *The circles.*
(2) *The ellipses.*
(3) *The non-degenerate conics.*
(4) *Smooth deformations of these.*

PROOF. All this is quite intuitive, the idea being as follows:

(1) Consider the unit circle, $x^2 + y^2 = 1$. We can write then $x = \cos t$, $y = \sin t$, with $t \in [0, 2\pi)$, and we seem to have here the solution to our problem, just using 1 chart. But this is of course wrong, because $[0, 2\pi)$ is not open, and we have a problem at 0. In practice we need to use 2 such charts, say with the first one being with $t \in (0, 3\pi/2)$, and the second one being with $t \in (\pi, 5\pi/2)$. As for the fact that the change of charts is indeed smooth, this comes by writing down the formulae, or just thinking a bit, and arguing that this change of chart being actually a translation, it is automatically linear.

(2) This follows from (1), by pulling the circle in both the Ox and Oy directions, and the formulae here, based on Theorem 5.1, are left to you reader.

(3) We already have the ellipses, and the case of the parabolas and hyperbolas is elementary as well, and in fact simpler than the case of the ellipses. Indeed, a parablola is clearly homeomorphic to $\mathbb{R}$, and a hyperbola, to two copies of $\mathbb{R}$.

(4) This is something which is clear too, depending of course on what exactly we mean by "smooth deformation", and by using a bit of multivariable calculus if needed. □

In higher dimensions now, as basic examples here, we have the unit sphere in $\mathbb{R}^N$, and smooth deformations of it, once again, somehow by obvious reasons. In case you are wondering on how to construct explicit charts for the sphere, the answer comes from:

THEOREM 5.6. *We have spherical coordinates in N dimensions,*

$$\begin{cases} x_1 &= r\cos t_1\\ x_2 &= r\sin t_1\cos t_2\\ \vdots\\ x_{N-1} &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\cos t_{N-1}\\ x_N &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\sin t_{N-1}\end{cases}$$

with the corresponding Jacobian being given by the following formula,

$$J(r,t) = r^{N-1}\sin^{N-2}t_1\sin^{N-3}t_2\ \ldots\ \sin^2 t_{N-3}\sin t_{N-2}$$

and with this guaranteeing that the sphere is indeed a smooth manifold.

PROOF. The fact that we have indeed spherical coordinates is clear. Regarding the Jacobian, by developing over the last column, we have:

$$\begin{aligned} J_N &= r\sin t_1\ldots\sin t_{N-2}\sin t_{N-1}\times\sin t_{N-1}J_{N-1}\\ &+ r\sin t_1\ldots\sin t_{N-2}\cos t_{N-1}\times\cos t_{N-1}J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}(\sin^2 t_{N-1}+\cos^2 t_{N-1})J_{N-1}\\ &= r\sin t_1\ldots\sin t_{N-2}J_{N-1}\end{aligned}$$

Thus, we obtain the formula in the statement, by recurrence. Finally, in what regards the last assertion, smooth manifold, this can proved a bit like for the circle, as we did in the proof of Theorem 5.5 (1), basically by cutting the sphere into 2^N parts. □

We have the stereographic projection as well, which was an operation dear to Riemann, the founding father of modern geometry, and which works more directly, as follows:

THEOREM 5.7. *The stereographic projection is given by inverse maps*

$$\Phi:\mathbb{R}^N\to S^N_{\mathbb{R}}-\{\infty\}\quad,\quad \Psi:S^N_{\mathbb{R}}-\{\infty\}\to\mathbb{R}^N$$

given by the following formulae,

$$\Phi(v)=(1,0)+\frac{2}{1+||v||^2}(-1,v)\quad,\quad \Psi(c,x)=\frac{x}{1-c}$$

with the convention $\mathbb{R}^{N+1}=\mathbb{R}\times\mathbb{R}^N$, and with the coordinate of $\mathbb{R}$ denoted x_0, and with the coordinates of $\mathbb{R}^N$ denoted $x_1,\ldots,x_N$.

PROOF. We are looking for the formulae of the isomorphism $\mathbb{R}^N\simeq S^N_{\mathbb{R}}-\{\infty\}$, obtained by identifying $\mathbb{R}^N=\mathbb{R}^N\times\{0\}\subset\mathbb{R}^{N+1}$ with the unit sphere $S^N_{\mathbb{R}}\subset\mathbb{R}^{N+1}$, with the convention that the point which is added is $\infty=(1,0,\ldots,0)$, via the stereographic projection. That is, we need the precise formulae of two inverse maps, as follows:

$$\Phi:\mathbb{R}^N\to S^N_{\mathbb{R}}-\{\infty\}\quad,\quad \Psi:S^N_{\mathbb{R}}-\{\infty\}\to\mathbb{R}^N$$

In one sense, according to our conventions above, we must have a formula as follows for our map Φ, with the parameter $t\in(0,1)$ being such that $||\Phi(v)||=1$:

$$\Phi(v)=t(0,v)+(1-t)(1,0)$$

The equation for the parameter $t\in(0,1)$ can be solved as follows:

$$\begin{aligned}(1-t)^2+t^2||v||^2=1 &\iff t^2(1+||v||^2)=2t\\ &\iff t=\frac{2}{1+||v||^2}\end{aligned}$$

We conclude that the formula of the map Φ is as follows:

$$\Phi(v)=(1,0)+\frac{2}{1+||v||^2}(-1,v)$$

In the other sense now we must have, for a certain $\alpha \in \mathbb{R}$:

$$(0, \Psi(c, x)) = \alpha(c, x) + (1 - \alpha)(1, 0)$$

But from $\alpha c + 1 - \alpha = 0$ we get the following formula for the parameter α:

$$\alpha = \frac{1}{1 - c}$$

We conclude that the formula of the map Ψ is as follows:

$$\Psi(c, x) = \frac{x}{1 - c}$$

Here, as before, we use the convention in the statement, namely $\mathbb{R}^{N+1} = \mathbb{R} \times \mathbb{R}^N$, with the coordinate of $\mathbb{R}$ denoted x_0, and with the coordinates of $\mathbb{R}^N$ denoted $x_1, \ldots, x_N$. □

Many other things can be said about smooth manifolds. As a second idea now, in order to generalize the conics, we can look as well at zeros of arbitrary polynomials:

$$P(x, y) = 0 \quad , \quad P \in \mathbb{R}[x, y]$$

More generally, we can look at zeros of polynomials in arbitrary N dimensions:

$$P(x_1, \ldots, x_N) = 0 \quad , \quad P \in \mathbb{R}[x_1, \ldots, x_N]$$

Observe that, at $N \geq 3$, what we have is not exactly a curve, but rather some sort of $(N - 1)$-dimensional surface, called algebraic hypersurface. Due to this, in order to have a full collection of beasts, of all possible dimensions, we must intersect such algebraic hypersurfaces. We are led in this way to zeros of families of polynomials, as follows:

Definition 5.8. *An algebraic manifold is a space of the form*

$$X = \left\{ (x_1, \ldots, x_N) \in \mathbb{R}^N \middle| P_i(x_1, \ldots, x_N) = 0, \forall i \right\}$$

with $P_i \in \mathbb{R}[x_1, \ldots, x_N]$ being a family of polynomials.

And, good news, this is the good definition, and with the branch of mathematics studying such manifolds being called algebraic geometry. In what follows we will discuss a bit what can be done with this definition, as a continuation of our previous work on conics, at the elementary level. All this will lead us into the conclusion that we must first develop commutative algebra, and come back to algebraic geometry afterwards.

As a first observation, what we have in Definition 5.8 does not guarantee the smoothness of our manifold. The non-smooth points are called singularities, and there are many interesting things that can be said about them. Also, we can replace $\mathbb{R}$ by $\mathbb{C}$, in Definition 5.8, and again, there are many interesting things that can be said here.

5c. Commutative algebra

Getting back to algebraic geometry and manifolds, in general, it is in fact possible to do even more generally, by looking at algebraic manifolds defined over an arbitrary field F, by using a family of polynomials $P_i \in F[x_1, \ldots, x_N]$, as follows:

$$X = \left\{(x_1, \ldots, x_N) \in F^N \middle| P_i(x_1, \ldots, x_N) = 0, \forall i\right\}$$

Such ideas are very old, again going back to the ancient Greeks, and there are many things that can be said about algebraic geometry in its "arithmetic" version, over arbitrary fields F as above. In fact, this is where algebraic geometry really shines, with many known advanced results in number theory having been obtained in this way.

5d. Projective manifolds

We have already seen a bit of projective geometry in chapter 1, when talking about the projective planes over the finite fields, notably with the Fano plane.

There is some linear algebra to be done here too, by identifying the lines in $\mathbb{R}^N$ with the corresponding rank 1 projections, along with many other things, and we have:

THEOREM 5.9. *The projective space $P_{\mathbb{R}}^{N-1}$ can be thought of as being the space of rank 1 projections in the matrix algebra $M_N(\mathbb{R})$, given by*

$$P_x = \frac{1}{||x||^2}(x_i x_j)_{ij}$$

by identifying the lines in $\mathbb{R}^N$ passing through the origin with the corresponding rank 1 projections in $M_N(\mathbb{R})$, in the obvious way.

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

(1) The main assertion is more or less clear from definitions, the point being that the lines in $\mathbb{R}^N$ passing through the origin are obviously in bijection with the corresponding rank 1 projections. Thus, we obtain the interpretation of $P_{\mathbb{R}}^{N-1}$ in the statement.

(2) Regarding now the formula of the rank 1 projections, which is a must-know, for this, and in everyday life, consider a vector $y \in \mathbb{R}^N$. Its projection on $\mathbb{R}x$ must be a certain multiple of x, and we are led in this way to the following formula:

$$P_x y = \frac{< y, x >}{< x, x >} x = \frac{1}{||x||^2} < y, x > x$$

(3) But with this in hand, we can now compute the entries of P_x, as follows:

$$\begin{aligned}(P_x)_{ij} &= < P_x e_j, e_i > \\ &= \frac{1}{||x||^2} < e_j, x >< x, e_i > \\ &= \frac{x_j x_i}{||x||^2}\end{aligned}$$

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

Regarding now embeddings of $P_{\mathbb{R}}^{N-1}$ into Euclidean spaces $\mathbb{R}^n$, many things can be said, with a straightforward construction here being as follows:

THEOREM 5.10. *The projective space $P_{\mathbb{R}}^{N-1}$ is a smooth manifold, with charts*

$$(x_1, \ldots, x_N) \to \left(\frac{x_1}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_N}{x_i}\right)$$

where $x_i \neq 0$. This manifold is compact, and of dimension $N-1$.

PROOF. We know that $P_{\mathbb{R}}^{N-1}$ appears as the space of lines in $\mathbb{R}^N$ passing through the origin, so we have the following formula, with $\sim$ being the proportionality of vectors, given as usual by $x \sim y$ when $x = \lambda y$, for some scalar $\lambda \neq 0$:

$$P_{\mathbb{R}}^{N-1} = \mathbb{R}^N - \{0\}/\sim$$

Alternatively, we can restrict if we want the attention to the vectors on the unit sphere $S_{\mathbb{R}}^{N-1} \subset \mathbb{R}^N$, and this because any line in $\mathbb{R}^N$ passing through the origin will certainly cross this sphere. Moreover, it is clear that our line will cross the sphere in exactly two points $\pm x$, and we conclude that we have the following formula, with $\sim$ being now the proportionality of vectors on the sphere, given by $x \sim y$ when $x = \pm y$:

$$P_{\mathbb{R}}^{N-1} = S_{\mathbb{R}}^{N-1}/\sim$$

With this discussion made, let us get now to what is to be proved. Obviously, once we fix an index $i \in \{1, \ldots, N\}$, the condition $x_i \neq 0$ on the vectors $x \in \mathbb{R}^N - \{0\}$ defines an open subset $U_i \subset P_{\mathbb{R}}^{N-1}$, and the open subsets that we get in this way cover $P_{\mathbb{R}}^{N-1}$:

$$P_{\mathbb{R}}^{N-1} = U_1 \cup \ldots \cup U_N$$

Moreover, the map in the statement is injective $U_i \to \mathbb{R}^{N-1}$, and it is clear too that the changes of charts are C^∞. Thus, we have our smooth manifold, as claimed. □

Many other things can be said about $P_{\mathbb{R}}^{N-1}$, and we will be back to this. Importantly, most of the above results extend to the complex setting, and we have:

THEOREM 5.11. *We can define the complex projective space $P_{\mathbb{C}}^{N-1}$ as being the space of complex lines in $\mathbb{C}^N$ passing through the origin. Alternatively, we can say that $P_{\mathbb{C}}^{N-1}$ is the space of rank 1 projections in the matrix algebra $M_N(\mathbb{C})$, given by*

$$P_x = \frac{1}{||x||^2}(x_i\bar{x}_j)_{ij}$$

by identifying the lines in $\mathbb{C}^N$ passing through the origin with the corresponding rank 1 projections in $M_N(\mathbb{C})$, in the obvious way. The complex projective space $P_{\mathbb{C}}^{N-1}$ is a smooth compact manifold, having complex dimension $N-1$.

PROOF. All this follows indeed via the same arguments as in the real case. □

Next, we can talk about Grassmannians, flag manifolds, and Stiefel manifolds.

5e. Exercises

Exercises:

EXERCISE 5.12.

EXERCISE 5.13.

EXERCISE 5.14.

EXERCISE 5.15.

EXERCISE 5.16.

EXERCISE 5.17.

EXERCISE 5.18.

EXERCISE 5.19.

Bonus exercise.

CHAPTER 6

Elliptic curves

6a. Elliptic curves

Elliptic curves.

6b. Abelian varieties

Abelian varieties.

6c. Rational points

Rational points.

6d. Further results

Further results.

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

Hasse principle

7a. p-adic numbers

We discuss here 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 is 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 7.1. *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 1, 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 7.2. *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 this chapter, 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.

Before starting, now that we are getting into subtle arithmetic, some advice too. You have surely noticed that the present book is written from an all-around scientific perspective, which is quite geometric and intuitive, as it should, but with as drawback, often trading simple and beautiful but very abstract mathematical arguments for some more geometric and intuitive explanations, which can be, after all, more complicated. So, we will keep doing this, that is how I concieved my book, but now that we are getting at the advanced level, do not hesitate to disagree with me, and have sometimes a look at other books, in case you feel not fully happy with my mathematics here.

I would especially recommend the books of Serre, [79], [80], [81] and many more, which are a masterpiece of uncompromising pure mathematics, with the shortest possible proofs for everything, "being short no matter what that takes" being the idea there. By the way, talking such references, let me recommend too Lang [63] for general algebra, complemented by Atiyah and MacDonald [7] for more commutative algebra, and Shafarevich [82] and Hartshorne [46] for algebraic geometry. Or at least these were the popular books when I was your age, doing my studies, some 30 years ago, in the early 90s, and believe me, these are all excellent books, and personally, I survived them.

Getting to work now, let us further reformulate our dreams and hopes, as follows:

QUESTION 7.3. *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 7.4. *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 7.4 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 7.5. *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 7.4 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 7.6. *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 7.5.

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 7.5, 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 7.5. 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 7.5, namely the modified norm axioms there, we can formulate, as a simple consequence:

PROPOSITION 7.7. *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 7.8. *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 7.9. *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 7.5, 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 7.10. *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 nice 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 7.11. *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 7.12. *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 chapter 4, and with this being quite similar to the construction $\mathbb{R} \to \mathbb{C}$. □

7b. Hasse-Minkowski

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 7.13. *Hasse local-global principle, and Hasse-Minkowski theorem.*

PROOF. Many things can be said here, especially in continuation of our study of elliptic curves, from chapter 6. The proofs, however, use a lot of non-trivial algebra. We will present here the main ideas, behind these proofs, with some details missing. □

Finally, on a personal note, let me mention that this whole subject is dear to me, with the Hasse-Minkowski theorem having been the subject of my master thesis dissertation, long years ago. For whatever reasons, I chose afterwards to do quantum physics instead. And, although I sometimes regret that, I must admit that doing quantum physics with a strong pure mathematics and number theory background is something quite fun. There are plently of good questions there, requiring, besides of course love for physics, chemistry, engineering, and real life and science in general, this sort of background.

In case you are in a similar situation, and hesitating too, who knows, what I can tell you is that this kind of choice is not very important. What matters is to know and love both number theory and quantum physics, and have a look at thermodynamics too, and then go for whatever problems you would like to have solved. And, you will see, over the time these problems will start to solve themselves, with minimal input from you.

7c. Algebraic groups

We have already seen some geometry over the p-adics, with a number of elementary results, obtained from definitions when introducing the p-adics, and then with various advanced results, centered around the Hasse principle, and obtained the hard way.

Our purpose now will be that of doing more geometry, sometimes spilling into analysis too, by looking at the representation theory invariants, such as the spectral measures, of the basic compact Lie groups, regarded over the field $\mathbb{Q}_p$ of the p-adic numbers.

Formally, we will include as well in our discussion some beasts called "quantum groups", whose study is interesting too, and is often simpler than that of the classical groups. We will be a bit sloppy here on the exact formalism of these quantum groups, but we will be back to this with more details in Part IV of the present book, where we will navigate in the icy waters of quantum algebra, in relation with number theory.

Generally speaking, the problem that we want to solve is as follows:

QUESTION 7.14. *What is the basic geometry and analysis, say in relation with the computation of the law of the main character, whose moments are the numbers*

$$M_k = \int_G \chi^k$$

for the basic compact Lie groups, and perhaps for some basic compact quantum Lie groups too, regarded over the field $\mathbb{Q}_p$ of the p-adic numbers?

Let us start our study with a brief discussion regarding probability, over arbitrary fields. Basic probability and free probability can be developed by using a pair (A, tr), with A being a complex algebra, and with $tr : A \to \mathbb{C}$ being a trace, $tr(ab) = tr(ba)$. At a more advanced level, we need an involution $*$ on the algebra A. At an even more advanced level, we need a norm $||.||$ on the algebra A. And at the most advanced level of them all, we need to assume that A is a von Neumann algebra, $A \subset B(H)$.

In order to do probability and free probability over an arbitrary field F, we will need here, as starting objects, a pair (A, tr) as before, with A being this time a F-algebra, and with $tr : A \to F$ being a trace, $tr(ab) = tr(ba)$.

The simplest example of a F-algebra is the algebra $F(X) = \{f : X \to F\}$ of functions on a finite set X. We can endow this algebra with the following trace:

$$tr(f) = \frac{1}{|X|} \sum_{x \in X} f(x)$$

Observe that, in order for tr to be well-defined, the number $N = |X|$ must be prime to $p = char(F)$. In view of this, we will assume in what follows $char(F) = 0$.

We will be interested in the algebras of functions $F(G) = \{f : G \to F\}$ on various matrix groups $G \subset GL_N(F)$. Observe that the standard coordinates, $u_{ij}(g) = g_{ij}$, satisfy $u_{ij} \in F(G)$. In particular the main character, $\chi = \sum_i u_{ii}$, satisfies $\chi \in F(G)$. Thus, once we have a Haar measure on G, or rather the corresponding integration functional $tr : F(G) \to F$, we have a pair $(F(G), tr)$, that we can use for studying the main character $\chi = \sum_i u_{ii}$. It is possible to talk as well about quantum groups, with the algebraic theory going well under the same assumption as before, $char(F) = 0$.

Let (A, tr) as above, over a characteristic 0 field F. Normally the standard definitions of the independence and freeness make sense, as follows:

$$tr(a) = tr(b) = 0 \implies tr(ab) = 0$$

$$tr(a_i) = tr(b_i) = 0 \implies tr(a_1 b_1 a_2 b_2 \ldots) = 0$$

Also, the theory of the classical and free cumulants, which is purely combinatorial, and so "scalarless", modulo some integers coming from the Möbius inversion formula, normally works over any field of characteristic 0. Thus, we have probabilistic tools.

7d. Analytic aspects

The very first thing would be to compute the p-adic norms of the moments of the main limiting laws, coming from CLT, PLT, FCLT, FPLT. In what regards CLT, FCLT, FPLT, where the laws are the normal, Wigner and Marchenko-Pastur ones, the moments appear as simple modifications of the central binomial coefficients:

$$\binom{2n}{n} = \frac{(2n)!}{n!n!}$$

Now observe that the power of p inside $N!$ is the following number:

$$v_p(N) = \left[\frac{N}{p}\right] + \left[\frac{N}{p^2}\right] + \left[\frac{N}{p^3}\right] + \ldots$$

Thus, in what regards the central binomial coefficients, the p-adic norm is something quite complicated, unless we are in the case $p = 2$, where things simplify, and we obtain 2^{-n}. As for the limiting theorem left, PLT, here the law is the Poisson law, the moments are the Bell numbers, and once again things seem to be complicated at $p \geq 3$.

Summarizing, for more advanced theory, the field that we should have in mind is probably the dyadic numbers, $F = \mathbb{Q}_2$. There are many computations to be done here: Bell and Stirling numbers, and computations with Gram and Weingarten matrices.

We are interested in computing the p-adic norm $|\,.\,|$, or simply the p-adic valuation $v(.)$ for the moments of the main limiting laws in classical and free $\mathbb{Q}_p$-probability, namely CLT, PLT, FCLT, FPLT. Assuming of course that these theorems exist indeed.

Before starting, let us comment on the case $F = \mathbb{C}$. Combinatorially speaking, the CLT, FCLT, FPLT are "trivial", from the point of view of the moments of the limiting laws, because these moments appear as very simple modifications of some very basic quantities in combinatorics, namely the central binomial coefficients:

$$\binom{2n}{n} = \frac{(2n)!}{n!n!}$$

However, in what regards the PLT, here the moments are the highly non-trivial Bell numbers. Thus, in connection with the present p-adic questions regarding moments, we are much more into classical probability, than into free probability.

As a final comment, this is not surprising from the quantum group viewpoint, where S_N^+ is known to be something much simpler, combinatorially speaking, than S_N.

CLT. The even moments of the normal law are the following numbers:

$$M_{2k} = 1 \cdot 3 \cdot 5 \ldots (2k-1) = \frac{(2k)!}{2^k k!}$$

Now observe that the power of p inside $N!$ is the following number:

$$v(N)=\left[\frac{N}{p}\right]+\left[\frac{N}{p^2}\right]+\left[\frac{N}{p^3}\right]+\ldots$$

In the case $p=2$ we obtain the following formula, which is something very nice:

$$\begin{aligned}v(M_{2k})&=v((2k)!)-k-v(k!)\\&=\sum_{r\geq 1}\left[\frac{2k}{2^r}\right]-k-\sum_{r\geq 1}\left[\frac{k}{2^r}\right]\\&=[k]-k\\&=0\end{aligned}$$

In the case $p\geq 3$ the formula is more complicated, as follows:

$$\begin{aligned}v(M_{2k})&=v((2k)!)-v(k!)\\&=\sum_{r\geq 1}\left[\frac{2k}{p^r}\right]-\sum_{r\geq 1}\left[\frac{k}{p^r}\right]\end{aligned}$$

As a conclusion, things here are massively simpler at $p=2$.

FCLT. The even moments of the Wigner semicircle law are the Catalan numbers:

$$M_{2k}=\frac{(2k)!}{k!(k+1)!}$$

At $p=2$ we have the following formula, which is quite reasonable:

$$\begin{aligned}v(M_{2k})&=v((2k)!)-v(k!)-v((k+1)!)\\&=k-v((k+1)!)\end{aligned}$$

At $p\geq 3$ the simplification at the end does not appear.

FPLT. No new computations needed there, the moments of the Marchenko-Pastur law being the Catalan numbers.

PLT. Tough problems here, the moments in question being the Bell numbers B_k. Things here are known to be quite complicated in general, with this being a popular problem in p-adic analysis, according to work of Barsky et al. However, the $p=2$ choice, not considered by Barsky et al., might simplify a few things. To be confirmed.

As another trick here, we can try to look into Stirling numbers instead, which might be simpler. These numbers appear by looking at the Poisson(t) laws, and actually the work here is definitely to be done, if by "CLT, PLT, FCLT, FPLT" we mean the parametric versions of these theorems. By the way, in what regards the CLT, FCLT, FPLT, no new computations needed here, the moments being simple rescalings of those at $t=1$.

There are probably many interesting summation formulae over S_N to be tried as well, as well as various computations with Gram and Weingarten matrices, and their traces and their determinants, with all this waiting to be evaluated in p-adic norm.

7e. Exercises

Exercises:

Exercise 7.15.

Exercise 7.16.

Exercise 7.17.

Exercise 7.18.

Exercise 7.19.

Exercise 7.20.

Exercise 7.21.

Exercise 7.22.

Bonus exercise.

CHAPTER 8

About Fermat

8a. Fermat equation

Fermat equation.

8b. Basic results

Basic results.

8c. General strategy

General strategy.

8d. Wiles proof

Wiles proof.

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

Analytic methods

On the second day I brought her a flower
She was more beautiful than any woman I'd seen
I said, do you know where the wild roses grow
So sweet and scarlet and free

CHAPTER 9

Primes, again

9a. Euler estimates

Welcome to analysis. We have seen in Part II how number theory leads us into quite complicated algebra, and the truth of course is, nothing can replace that type of complicated algebra, for understanding numbers, and equations relating them.

So, what about our beloved calculus, can that be of help? We will see in this third part of the present book that yes, calculus can indeed help. Help with understanding primes, help with understanding other questions regarding numbers, and especially help with understanding the Riemann zeta function, which is a joint algebraic and analytic business, at a core of number theory at large, both algebraic and analytic.

Finally, let us mention that both algebraic and analytic number theory are as old as mankind, and there has been always a bit of competition between them. In ancient times, algebra was the way. Then came analysis, then algebra again, and so on. So, hard to judge, but to the question "which of them first produced results which are hard to understand for the modern scientist", the answer is analysis, with the Prime Number Theorem proved by Hadamard and de la Vallée Poussin in 1896, by using quite tough analysis techniques. So, from this perspective, analytic number theory comes first, and what we are doing in this book is a bit upside down. But, more on this later.

Getting started now, save for all sorts of interesting computations with Gauss sums, and with that sign still to be computed, what we know so far analysis reduces to the Euler formula from chapter 1. So, let us start by improving that formula. We have:

THEOREM 9.1. *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. Since we are now stepping into analysis, which looks a bit like an alien science, after all the algebra that we did in this book, it is quite instructive to go slow, and not necessarily look for the best estimates that we can get, right away. That is, our purpose, to start with, is to understand how various tools of analysis, starting with the simplest

ones, fit into our algebraic story of numbers. With this idea in mind, the story with the estimate in the statement, told with several detours and details, is as follows:

(1) By using the unique factorization $n = p_1^{a_1} \dots p_k^{a_k}$, we have:

$$\begin{aligned}\prod_{p<N}\left(1-\frac{1}{p}\right)^{-1} &= \prod_{p<N}\left(1+\frac{1}{p}+\frac{1}{p^2}+\frac{1}{p^3}+\dots\right)\\ &> \sum_{n=1}^{N-1}\frac{1}{n}\\ &> \int_1^N \frac{1}{x}\,dx\\ &= \log N\end{aligned}$$

(2) But the product on the left can be estimated by using log, as follows:

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

(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 9.2. *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. □

As a comment now on Theorem 9.1, this suggests looking as well into reverse inequalities, of the same type, but for this purpose the Euler method does not really work, or at least there is nothing that can be done quickly, the point being as follows:

COMMENT 9.3. *In the context of the Euler estimate, we also have, trivially*

$$\log\left[\prod_{p<N}\left(1-\frac{1}{p}\right)^{-1}\right]>\sum_{p<N}\frac{1}{p}$$

but in order to exploit this, we need an upper bound for the following sum, with $S_N\subset\mathbb{N}$ standing for the set of positive integers having all prime factors $<N$,

$$\sum_{n\in S_N}\frac{1}{n}=\prod_{p<N}\frac{1}{p}$$

and this bound, wanted of $\log\log N$ *type, does not look easy to get, with bare hands.*

To be more precise, the above sum of $1/n$ values can be estimated by using the writing $n = m^2 s$, with s square-free, and we obtain in this way the following estimate:

$$\begin{aligned}
\sum_{p<N}\frac{1}{p} &< \log\left[\prod_{p<N}\left(1-\frac{1}{p}\right)^{-1}\right]\\
&= \log\left[\prod_{p<N}\left(1+\frac{1}{p}+\frac{1}{p^2}+\frac{1}{p^3}+\ldots\right)\right]\\
&= \log\left(\sum_{n\in S_N}\frac{1}{n}\right)\\
&< \log\left[\sum_{m=1}^{\infty}\frac{1}{m^2}\prod_{p<N}\left(1+\frac{1}{p}\right)\right]\\
&= \sum_{p<N}\log\left(1+\frac{1}{p}\right)+\log\left(\frac{\pi^2}{6}\right)\\
&< \sum_{p<N}\frac{1}{p}+\log\left(\frac{\pi^2}{6}\right)
\end{aligned}$$

And isn't this sweet. Normally serious books are not supposed to contain such things, but this being an introductory book, I have to introduce you, among others, to mathematical research too. And here, believe me, what you usually get at the end of your research day, after a lot of sweat, blood and tears, are estimates of the above type.

9b. 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 9.4. *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 $Re(s)>1$. And then, with a bit of complex analysis, we can have the zeta function working in the whole complex plane $\mathbb{C}$, as a meromorphic function there, by analytic continuation. But more on this, complex zeta, in chapter 10 below.

Here we will just use zeta at $s>1$, and why not its truncations too, at any $s\in\mathbb{R}$:

$$S_N=\sum_{n=1}^N\frac{1}{n^s}$$

As a first observation, the Basel formula, from Theorem 9.2, reformulates as follows:

THEOREM 9.5. *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 9.2, 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 9.2. 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 10 below. What we want to discuss here is what happens to the Euler estimate from Theorem 9.1, when adding an exponent $s \in \mathbb{R}$ there. Let us start with:

PROPOSITION 9.6. *The Euler estimate can be generalized into*

$$\sum_{p<N} \frac{1}{p^s} > \log\left(\int_1^N \frac{1}{x^s}\,dx\right) - \frac{1}{2}\sum_{n=2}^{N-1} \frac{1}{n^s(n^s-1)}$$

with the above integral given by the formula

$$\int_1^N \frac{1}{x^s}\,dx = \begin{cases} \frac{N^{1-s}-1}{1-s} & \text{if } s \neq 1 \\ \log N & \text{if } s = 1 \end{cases}$$

involving now a real parameter $s \in \mathbb{R}$, with exactly the same proof.

PROOF. By using the unique factorization $n = p_1^{a_1} \ldots p_k^{a_k}$, 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 9.7. *We have the following Euler type estimate*

$$\sum_{p<N}\frac{1}{p^s}>\log\left(\frac{1-N^{1-s}}{s-1}\right)-\frac{\zeta(2s)}{2}$$

valid for any value of the parameter $s>1$.

PROOF. In the case $s>1$ the estimate that we found in Proposition 9.6 gives:

$$\begin{aligned}
\sum_{p<N}\frac{1}{p^s} &> \log\left(\frac{1-N^{1-s}}{s-1}\right)-\frac{1}{2}\sum_{n=2}^{N-1}\frac{1}{n^s(n^s-1)}\\
&> \log\left(\frac{1-N^{1-s}}{s-1}\right)-\frac{1}{2}\sum_{n=2}^{\infty}\frac{1}{n^s(n^s-1)}\\
&> \log\left(\frac{1-N^{1-s}}{s-1}\right)-\frac{1}{2}\sum_{n=2}^{\infty}\frac{1}{(n-1)^{2s}}\\
&> \log\left(\frac{1-N^{1-s}}{s-1}\right)-\frac{\zeta(2s)}{2}
\end{aligned}$$

Here we have used the following inequality, with $\varepsilon = 1/n < 1$, which is true:

$$\begin{aligned}
\frac{1}{n^s(n^s-1)} < \frac{1}{(n-1)^{2s}} &\iff (n-1)^{2s} < n^s(n^s-1)\\
&\iff \left(1-\frac{1}{n}\right)^{2s} < 1-\frac{1}{n^s}\\
&\iff (1-\varepsilon)^{2s} < 1-\varepsilon^s\\
&\iff (1-\varepsilon)^{2s-1} < \frac{1-\varepsilon^s}{1-\varepsilon}
\end{aligned}$$

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

It is possible to futher build along the above lines, but we will leave this discussion for later, in chapter 10, when talking more in detail about the Riemann zeta function.

9c. 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 9.8. *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 9.9. *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 9.10. *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 9.1 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 9.11. *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:

$$N! = \int_0^\infty x^N e^{-x} dx$$

(5) Since the integrand is sharply peaked at $x = N$, as you can see by computing the derivative of $\log(x^N e^{-x})$, this suggests writing $x = N + y$, and we obtain:

$$\begin{aligned}\log(x^N e^{-x}) &= N\log x - x \\ &= N\log(N+y) - (N+y) \\ &= N\log N + N\log\left(1 + \frac{y}{N}\right) - (N+y) \\ &\simeq N\log N + N\left(\frac{y}{N} - \frac{y^2}{2N^2}\right) - (N+y) \\ &= N\log N - N - \frac{y^2}{2N}\end{aligned}$$

(6) By exponentiating, we obtain from this the following estimate:

$$x^N e^{-x} \simeq \left(\frac{N}{e}\right)^N e^{-y^2/2N}$$

(7) Now by integrating, we obtain from this the following estimate:

$$\begin{aligned} N! &= \int_0^\infty x^N e^{-x} dx \\ &\simeq \int_{-N}^N \left(\frac{N}{e}\right)^N e^{-y^2/2N}\, dy \\ &\simeq \left(\frac{N}{e}\right)^N \int_{\mathbb{R}} e^{-y^2/2N}\, dy \\ &= \left(\frac{N}{e}\right)^N \sqrt{2N} \int_{\mathbb{R}} e^{-z^2}\, dz \\ &= \left(\frac{N}{e}\right)^N \sqrt{2\pi N} \end{aligned}$$

(8) Here we have used at the end the following key formula, due to Gauss:

$$\begin{aligned} \left(\int_{\mathbb{R}} e^{-z^2} dz\right)^2 &= \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-x^2-y^2} dxdy \\ &= \int_0^{2\pi}\int_0^\infty e^{-r^2} r dr dt \\ &= 2\pi \int_0^\infty \left(-\frac{e^{-r^2}}{2}\right)' dr \\ &= 2\pi \left[0 - \left(-\frac{1}{2}\right)\right] \\ &= \pi \end{aligned}$$

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 9.10, 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 9.12. *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:

$$\gamma=\lim_{s\to1^+}\sum_{n=1}^{\infty}\frac{1}{n^s}-\frac{1}{s^n}$$

In terms of the zeta function, this latter formula simply reads:

$$\gamma=\lim_{s\to1}\zeta(s)-\frac{1}{s-1}$$

Alternatively, still in terms of the zeta function around 1, this reads:

$$\gamma=\lim_{s\to0}\frac{\zeta(1+s)+\zeta(1-s)}{2}$$

(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 9.13. *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 9.10, 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 9.10, the most famous one, being his second theorem. His first theorem is a related formula, as follows:

THEOREM 9.14. *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 9.15. *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: □

There are of course many other things that can be said, in relation with the above, and for more on all this, you can check any advanced number theory book. In what concerns us, we will be back to this at various places, in what follows.

9d. 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 9.16. *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 9.17. *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 9.16, 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 9.17:

PROPOSITION 9.18. *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 9.19. *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 9.20. *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 9.19, 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 9.21. *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 9.20, 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 9.22. *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 9.23. *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 9.21 and Theorem 9.22, 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.

In what follows, we will explain in chapter 11 a modern proof of Selberg, which is somewhat elementary, and we will explain afterwards as well the original proof, by Hadamard and de la Vallée Poussin, by summing over the zeroes of the Riemann zeta function.

9e. Exercises

We had a lot of interesting mathematics in this chapter, all good old things, going back to the greats, and as exercises on all this, we have:

EXERCISE 9.24. *Clarify all the details in the proof of the Basel formula.*

EXERCISE 9.25. *Try getting upper estimates on* $\sum_{p<N} 1/p$*, using Euler products.*

EXERCISE 9.26. *Learn about Bernoulli numbers, and the formula of* $\zeta(2k)$*.*

EXERCISE 9.27. *Find the next term in the Stirling formula, with* $N \to \infty$*.*

EXERCISE 9.28. *Learn more about the Euler-Mascheroni constant.*

EXERCISE 9.29. *Clarify the exact computation of the Mertens constant.*

EXERCISE 9.30. *Work out the details for the proofs of Mertens 1 and 3.*

EXERCISE 9.31. *Clarify how Chebycheff implies the Bertrand postulate.*

As bonus exercise, get a calculus black belt, this being obviously needed for analytic number theory. The best is to train with some probability, PDE or physics.

CHAPTER 10

Zeta function

10a. Real zeta

We have already met the Riemann zeta function on several occasions, in chapter 9, 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 10.1. *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 10.2. *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 10.3. *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 10.4. *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 10.5. *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 10.6. *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. □

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 in chapter 9:

THEOREM 10.7. *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 chapter 9, 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 in chapter 9, so in short, not a path to be followed.

(2) Instead, we have the following luminous computation, based on Theorem 10.6:

$$\begin{aligned}
\zeta(2k) &= \frac{1}{\Gamma(2k)} \int_0^\infty \frac{x^{2k-1}}{e^x - 1}\, dx \\
&= \frac{1}{(2k-1)!} \int_0^\infty \frac{x^{2k-1}}{e^x - 1}\, dx \\
&= \frac{1}{(2k-1)!} \int_0^\infty \frac{(2\pi t)^{2k-1}}{e^{2\pi t} - 1}\, 2\pi dt \\
&= \frac{(2\pi)^{2k}}{(2k-1)!} \int_0^\infty \frac{t^{2k-1}}{e^{2\pi t} - 1}\, dt
\end{aligned}$$

(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:

$$\begin{aligned}\sum_{k=0}^{n-1}f(x) &\simeq \int_0^n f(x)dx-\frac{1}{2}(f(n)-f(0))\\ &+\frac{1}{6}\cdot\frac{f'(n)-f'(0)}{2!}-\frac{1}{30}\cdot\frac{f^{(3)}(n)-f^{(3)}(0)}{4!}\\ &+\frac{1}{42}\cdot\frac{f^{(5)}(n)-f^{(5)}(0)}{6!}-\frac{1}{30}\cdot\frac{f^{(7)}(n)-f^{(7)}(0)}{8!}+\ldots\end{aligned}$$

(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 10.7, let us record as well:

THEOREM 10.8. *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 10.7, 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 10.9. *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 10.6:

$$\begin{aligned}
\int_0^1 \int_0^1 \frac{1}{1-xy}\, dxdy &= \int_0^1 \left[-\frac{\log(1-xy)}{y} \right]_0^1 dy \\
&= -\int_0^1 \frac{\log(1-y)}{y}\, dy \\
&= -\int_0^\infty \frac{\log(e^{-t})}{1-e^{-t}}\, e^{-t} dt \\
&= \int_0^\infty \frac{t}{e^t - 1}\, dt \\
&= \zeta(2)\Gamma(2) \\
&= \zeta(2)
\end{aligned}$$

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:

$$\begin{aligned}
\int_0^1 \frac{1}{1-x}\, dx &= \left[-\log(1-x) \right]_0^1 \\
&= -\log(1-1) + \log(1-0) \\
&= \infty + 0 \\
&= \zeta(1)
\end{aligned}$$

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.

10b. 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 10.10. *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 10.11. *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:

$$\gamma=\lim_{s\to1^+}\sum_{n=1}^{\infty}\frac{1}{n^s}-\frac{1}{s^n}$$

But in terms of the zeta function, this latter formula simply reads:

$$\gamma=\lim_{s\to1^+}\zeta(s)-\frac{1}{s-1}$$

(3) Thus, we are led to the formula in the statement. Note that we have as well:

$$\gamma=\lim_{s\to0}\frac{\zeta(1+s)+\zeta(1-s)}{2}$$

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 $t\neq0$, of the boundary line $Re(s)=1$, between known and unknown. We have here:

THEOREM 10.12. *The Riemann zeta function, namely*

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

converges at any $s=1+it$ with $t\neq0$.

PROOF. We have the following computation, to start with:

$$\begin{aligned}\zeta(1+it)&=\sum_{n=1}^{\infty}\frac{1}{n^{1+it}}\\&=\sum_{n=1}^{\infty}\frac{1}{ne^{it\log n}}\\&=\sum_{n=1}^{\infty}\frac{e^{-it\log n}}{n}\\&=\sum_{n=1}^{\infty}\frac{\cos(t\log n)-i\sin(t\log n)}{n}\end{aligned}$$

Normally the convergence at $t\neq0$ can be proved via some calculus. We will leave this as an instructive exercise, and come back to it with details in a moment. □

With this discussed, let us get now into the true unknown, $Re(s)<1$, with our first objective being that of understanding what happens in the strip $0<Re(s)<1$. And here the idea is very simple, coming from the following very basic fact:

PROPOSITION 10.13. *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 10.14. *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 1.13 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 [74] 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 10.12 as well. Thinking retrospectively, we were in need there precisely of a Dirichlet type idea. □

All the above is quite interesting, and as usual when is comes to the zeta function, countless more formulae are available. For intance, in analogy with what we did in Theorem 10.12 on the vertical $Re(s) = 1$, we can try to understand what happens on the other vertical, $Re(s) = 0$, or why not on the middle vertical too, $Re(s) = 1/2$. And, regarding this latter problem, a lot of theory can be developed here, notably with a formula as follows, with a, b being called the Riemann-Siegel functions:

$$\zeta\left(\frac{1}{2} + it\right) = a(t)e^{-ib(t)}$$

But more on this later, do not worry, we will certainly come back very soon to the strip $0 < Re(s) < 1$, that we will see to be of key importance, in the theory of ζ.

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 10.15. *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 10.14, 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 10.14, 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 10.16. *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 10.14 and 10.15 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 10.14. □

Very nice all this, we have now zeta up and working, over the whole $\mathbb{C}$. We should mention that there are several other possible ways to reach to this, that is, to reach from Theorem 10.14 to Theorem 10.16, without getting into the Riemann symmetry formula from Theorem 10.15, which is something quite complicated. We will be back to this.

Getting back now to the Riemann formula from Theorem 10.15, 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 10.17. *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 10.15, and is in fact what comes out from computations, when proving Theorem 10.15. □

In practice, the quantity in Theorem 10.17 is best normalized as follows:

THEOREM 10.18. *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 10.15. □

10c. Further formulae

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, as follows:

(1) Are there alternative formulae for the analytic continuation?

(2) What about other general formulae satisfied by zeta?

(3) What values of zeta can we explicitely compute?

(4) What are the zeroes of zeta located?

These are all excellent questions, and we will discuss them all, in this section. 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.

(6) And there are many other formulae, for instance using contour integrals, over suitably chosen curves. For more on all this, check any advanced number theory book.

Getting now to the 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:

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

There are many other known formulae satisfied by zeta, along this lines. Again, for all this, you can check any advanced number theory book.

Getting now to the 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:

$$\sum_{k=2}^{\infty}(\zeta(k)-1)=1\qquad,\qquad\sum_{k=1}^{\infty}(\zeta(2k)-1)=\frac{3}{4}$$

$$\sum_{k=1}^{\infty}(\zeta(2k+1)-1)=\frac{1}{4}\qquad,\qquad\sum_{k=2}^{\infty}(-1)^k(\zeta(k)-1)=\frac{1}{2}$$

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 10.15, we have:

THEOREM 10.19. *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 10.15 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. □

Observe that the above formula has nothing to do with the original formula of the zeta function, namely $\zeta(s)=\sum_{n=1}^{\infty}\frac{1}{n^s}$, which is only valid at $Re(s)>1$. However, as a matter of having some fun, let us write the following formula, in a formal sense:

$$\sum_{n=1}^{\infty}\frac{1}{n^{-2k}}=0$$

Observe that this reminds a bit the weird p-adic formulae from chapter 7, as for instance the following formula, which was something true, over the dyadic numbers:

$$\sum_{n=0}^{\infty}2^n=-1$$

We will keep this remark in our pocket, and pull out zeta, when p-adics do not work, or even when they work, for all sorts of dirty tricks, in advanced math, or physics.

10d. Riemann hypothesis

As explained above, the zeta function has trivial zeroes at $-2,-4,-6,\ldots$, and it is also clear that the nontrivial zeroes must lie in the closed critical strip, namely:

$$0\leq Re(s)\leq 1$$

The Riemann hypothesis states that the nontrivial zeroes must satisfy:

$$Re(s)=\frac{1}{2}$$

And such nontrivial zeroes are known to exist indeed, an infinity of them.

The Riemann hypothesis is important, among others because many questions in arithmetic, as we will soon see, reformulate in terms of sums at the zeroes of zeta.

In order to get an idea of the difficulty of the Riemann hypothesis, the Prime Number Theoem, that we will prove in chapter 11 the hard way, is equivalent to the fact that there are no zeroes of zeta with $Re(s)=1$, or equivalently, with $Re(s)=0$.

10e. Exercises

Exercises:

Exercise 10.20.

Exercise 10.21.

Exercise 10.22.

Exercise 10.23.

Exercise 10.24.

Exercise 10.25.

Exercise 10.26.

Exercise 10.27.

Bonus exercise.

CHAPTER 11

Prime distribution

11a. Zero summing

Let us go back to the main result from chapter 9, 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 9, 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. Then, we will explain other proofs as well, notably with the Selberg proof, not using zeta, and also with the modern 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 11.1. *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 11.2. *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 11.3. *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 11.4. *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 11.1 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 11.3, 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 11.2, 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 11.2 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 11.4, 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 11.5. *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. □

Regarding now the non-trivial zeroes of zeta, we know from chapter 10 that these lie inside the strip $0 \leq Re(s) \leq 1$, and as a first observation, we have:

PROPOSITION 11.6. *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 11.5. □

We are now getting to the core of the proof, with the key ingredient being:

THEOREM 11.7. *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 11.6, but which of course has great theoretical interest, follows from the $Re(s) = 1$ result, via the Riemann reflection formula from chapter 10.

(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 11.8 (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 11.2 that $\pi(x)\sim x/\log x$ is equivalent to $\psi(x)\sim x$.

(2) We have in Theorem 11.4 a formula for $\psi(x)$, in terms of the zeroes of zeta.

(3) Most of this zeroes are taken care of by Proposition 11.5 and Proposition 11.6.

(4) As for the remaining zeroes, there are none, as shown by Theorem 11.7. □

11b. Selberg method

Selberg method.

11c. Other proofs

Other proofs.

11d. Further results

Further results.

11e. Exercises

Exercises:

Exercise 11.9.

Exercise 11.10.

Exercise 11.11.

Exercise 11.12.

Exercise 11.13.

Exercise 11.14.

Exercise 11.15.

Exercise 11.16.

Bonus exercise.

CHAPTER 12

Progressions, gaps

12a. Erdős conjecture

Erdős conjecture.

12b. Combinatorics

Combinatorics.

12c. Green-Tao

Green-Tao.

12d. Further results

Further results.

12e. Exercises

Exercises:

Exercise 12.1.

Exercise 12.2.

Exercise 12.3.

Exercise 12.4.

Exercise 12.5.

Exercise 12.6.

Exercise 12.7.

Exercise 12.8.

Bonus exercise.

Part IV

Some physics

If Joan of Arc had a heart
Would she give it as a gift
To such as me who longs to see
How an angel ought to be

CHAPTER 13

Quantum groups

13a.

13b.

13c.

13d.

13e. Exercises

Exercises:

Exercise 13.1.

Exercise 13.2.

Exercise 13.3.

Exercise 13.4.

Exercise 13.5.

Exercise 13.6.

Exercise 13.7.

Exercise 13.8.

Bonus exercise.

CHAPTER 14

Random matrices

14a.

14b.

14c.

14d.

14e. Exercises

Exercises:

Exercise 14.1.

Exercise 14.2.

Exercise 14.3.

Exercise 14.4.

Exercise 14.5.

Exercise 14.6.

Exercise 14.7.

Exercise 14.8.

Bonus exercise.

CHAPTER 15

Geometric aspects

15a.

15b.

15c.

15d.

15e. Exercises

Exercises:

EXERCISE 15.1.

EXERCISE 15.2.

EXERCISE 15.3.

EXERCISE 15.4.

EXERCISE 15.5.

EXERCISE 15.6.

EXERCISE 15.7.

EXERCISE 15.8.

Bonus exercise.

CHAPTER 16

Absolute spaces

16a.

16b.

16c.

16d.

16e. Exercises

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

Bibliography

[1] E. Abe, Hopf algebras, Cambridge Univ. Press (1980).

[2] T.M. Apostol, Introduction to analytic number theory, Springer (1976).

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

[4] V.I. Arnold, Catastrophe theory, Springer (1974).

[5] M.F. Atiyah, K-theory, CRC Press (1964).

[6] M.F. Atiyah, The geometry and physics of knots, Cambridge Univ. Press (1990).

[7] M.F. Atiyah and I.G. MacDonald, Introduction to commutative algebra, Addison-Wesley (1969).

[8] T. Banica, Invitation to Hadamard matrices (2023).

[9] T. Banica, Principles of operator algebras (2023).

[10] T. Banica, A guide to quantum algebra (2024).

[11] R.J. Baxter, Exactly solved models in statistical mechanics, Academic Press (1982).

[12] N. Berline, E. Getzler and M. Vergne, Heat kernels and Dirac operators, Springer (2004).

[13] B. Blackadar, K-theory for operator algebras, Cambridge Univ. Press (1986).

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

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

[16] V. Chari and A. Pressley, A guide to quantum groups, Cambridge Univ. Press (1994).

[17] A. Connes, Noncommutative geometry, Academic Press (1994).

[18] A. Connes and M. Marcolli, Noncommutative geometry, quantum fields and motives, AMS (2008).

[19] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).

[20] H.S.M. Coxeter, Regular polytopes, Dover (1948).

[21] H. Davenport, Multiplicative number theory, Springer (1980).

[22] W. de Launey and D. Flannery, Algebraic design theory, AMS (2011).

[23] P.A.M. Dirac, Principles of quantum mechanics, Oxford Univ. Press (1930).

[24] M.P. do Carmo, Differential geometry of curves and surfaces, Dover (1976).

[25] M.P. do Carmo, Riemannian geometry, Birkhäuser (1992).

[26] S.K. Donaldson, Riemann surfaces, Oxford Univ. Press (2004).

[27] R. Durrett, Probability: theory and examples, Cambridge Univ. Press (1990).

[28] A. Einstein, Relativity: the special and the general theory, Dover (1916).

[29] P. Etingof, S. Gelaki, D. Nikshych and V. Ostrik, Tensor categories, AMS (2016).

[30] L.C. Evans, Partial differential equations, AMS (1998).

[31] B. Eynard, Counting surfaces, Birkhäuser (2016).

[32] W. Feller, An introduction to probability theory and its applications, Wiley (1950).

[33] E. Fermi, Thermodynamics, Dover (1937).

[34] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics, Caltech (1963).

[35] P. Flajolet and R. Sedgewick, Analytic combinatorics, Cambridge Univ. Press (2009).

[36] W. Fulton, Algebraic topology, Springer (1995).

[37] W. Fulton and J. Harris, Representation theory, Springer (1991).

[38] M.B. Green, J.H. Schwarz and E. Witten, Superstring theory, Cambridge Univ. Press (2012).

[39] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).

[40] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).

[41] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).

[42] P. Griffiths and J. Harris, Principles of algebraic geometry, Wiley (1994).

[43] A. Grothendieck and J. Dieudonné, Éléments de géométrie algébrique, IHES (1967).

[44] A. Grothendieck et al., Séminaire de géométrie algébrique, IHES (1972).

[45] G.H. Hardy and E.M. Wright, An introduction to the theory of numbers, Oxford Univ. Press (1938).

[46] J. Harris, Algebraic geometry, Springer (1992).

[47] R. Hartshorne, Algebraic geometry, Springer (1977).

[48] K.J. Horadam, Hadamard matrices and their applications, Princeton Univ. Press (2007).

[49] L. Hörmander, The analysis of linear partial differential operators, Springer (1983).

[50] R.A. Horn and C.R. Johnson, Matrix analysis, Cambridge Univ. Press (1985).

[51] J.E. Humphreys, Introduction to Lie algebras and representation theory, Springer (1972).

[52] J.E. Humphreys, Linear algebraic groups, Springer (1975).

[53] K. Ireland and M. Rosen, A classical introduction to modern number theory, Springer (1982).

[54] H. Iwaniec and E. Kowalski, Analytic number theory, AMS (2004).

[55] N. Jacobson, Basic algebra, Dover (1974).

[56] V.F.R. Jones, Subfactors and knots, AMS (1991).

[57] L.P. Kadanoff, Statistical physics: statics, dynamics and renormalization, World Scientific (2000).

[58] M. Karoubi, K-theory: an introduction, Springer (1978).

[59] C. Kassel, Quantum groups, Springer (1995).

[60] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).

[61] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).

[62] G. Landi, An introduction to noncommutative spaces and their geometry, Springer (1997).

[63] S. Lang, Algebra, Addison-Wesley (1993).

[64] S. Lang, Abelian varieties, Dover (1959).

[65] P. Lax, Linear algebra and its applications, Wiley (2007).

[66] P. Lax, Functional analysis, Wiley (2002).

[67] F. Lusztig, Introduction to quantum groups, Birkhäuser (1993).

[68] S. Majid, Foundations of quantum group theory, Cambridge Univ. Press (1995).

[69] Y.I. Manin, Quantum groups and noncommutative geometry, Springer (2018).

[70] M.L. Mehta, Random matrices, Elsevier (2004).

[71] J. Neukirch, Algebraic number theory, Springer (1999).

[72] P. Petersen, Riemannian geometry, Springer (2006).

[73] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[74] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[75] W. Rudin, Fourier analysis on groups, Dover (1974).

[76] H.J. Ryser, Combinatorial mathematics, Wiley (1963).

[77] W. Schlag, A course in complex analysis and Riemann surfaces, AMS (2014).

[78] J. Seberry and M. Yamada, Hadamard matrices, Wiley (2020).

[79] J.P. Serre, A course in arithmetic, Springer (1973).

[80] J.P. Serre, Linear representations of finite groups, Springer (1977).

[81] J.P. Serre, Local fields, Springer (1979).

[82] I.R. Shafarevich, Basic algebraic geometry, Springer (1974).

[83] J.H. Silverman, The arithmetic of elliptic curves, Springer (1986).

[84] J.H. Silverman and J.T. Tate, Rational points on elliptic curves, Springer (2015).

[85] B. Singh, Basic commutative algebra, World Scientific (2011).

[86] D.R. Stinson, Combinatorial designs: constructions and analysis, Springer (2006).

[87] M.E. Sweedler, Hopf algebras, W.A. Benjamin (1969).

[88] T. Tao, Topics in random matrix theory, AMS (2012).

[89] T. Tao and V.H. Vu, Additive combinatorics, Cambridge Univ. Press (2016).

[90] C.H. Taubes, Differential geometry, Oxford Univ. Press (2011).

[91] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[92] D.V. Voiculescu, K.J. Dykema and A. Nica, Free random variables, AMS (1992).

[93] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[94] L.C. Washington, Introduction to cyclotomic fields, Springer (1982).

[95] A. Weil, Basic number theory, Springer (1967).

[96] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[97] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[98] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).

[99] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).

[100] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).

Index